Quantitative Finance

ISSN: 1469-7688 (Print) 1469-7696 (Online) Journal homepage: http://www.tandfonline.com/loi/rquf20

Canonical sectors and evolution of firms in the US stock markets

Lorien X. Hayden, Ricky Chachra, Alexander A. Alemi, Paul H. Ginsparg & James P. Sethna

To cite this article: Lorien X. Hayden, Ricky Chachra, Alexander A. Alemi, Paul H. Ginsparg & James P. Sethna (2018): Canonical sectors and evolution of firms in the US stock markets, Quantitative Finance, DOI: 10.1080/14697688.2018.1444278

To link to this article: https://doi.org/10.1080/14697688.2018.1444278

Quantitative Finance, 2018
https://doi.org/10.1080/14697688.2018.1444278

© 2018 iStockphoto LP

Canonical sectors and evolution of firms in the US stock markets

LORIEN X. HAYDEN, RICKY CHACHRA, ALEXANDER A. ALEMI,
PAUL H. GINSPARG and JAMES P. SETHNA*

Department of Physics, Cornell University, Ithaca, NY 14853, USA

(*Received 17 November 2016; accepted 19 January 2018; published online 8 June 2018*)

Unsupervised machine learning can provide an objective and comprehensive broad-level sector decomposition of stocks

1. Main text

Stock market performance is measured with aggregated quantities called indices that represent a weighted average price of a basket of stocks. Market-wide indices such as Russell 3000® (Russell 3000®Index 2015) and the S&P 500® (S&P 500®Index 2014) consist of stocks from diverse companies reflecting a broad cross-section of the market. Sector-specific indices such as the Dow Jones® Financials Index (Dow Jones®US Indices 2015), CBOE® Oil Index (CBOE®Oil Index 2013) and the Morgan Stanley® High-Tech 35 Index (Morgan Stanley®High-Tech 35 Index 2005), etc., are more granular and their composition requires a classification of companies into sectors. Major industrial classification schemes classify firms into sectors, albeit with many ambiguities (Nadig and Crigger 2011). It is not clear, for example, how to assign a sector to conglomerates or diversified companies such as General Electric®. Conversely, non-conglomerates with exposure to firms outside their own sector (for example, an investment bank exclusively serving pharmaceutical firms) also blur the boundaries of sector-identification. Moreover, as companies and their economic environments evolve, neither the industrial sectors nor the firms' sector association remains static, necessitating updates to sector assignments and addition of new sectors.

A significant number of studies have previously aimed at identifying categories of stocks in financial markets with a variety of approaches. Recent numerical techniques have included extensive use of random matrix theory, principal component analysis or associated eigenvalue decomposition of the correlation matrix (Plerou *et al.* 2002, Coronnello *et al.* 2005, Kim and Jeong 2005, Eom *et al.* 2007, Conlon *et al.* 2009, Fenn *et al.* 2011), specialized clustering methods (Mantegna 1999, Bonanno *et al.* 2000, 2003, Kullmann *et al.* 2000, Basalto *et al.* 2005, Heimo *et al.* 2009, Musmeci *et al.* 2014) or time series analysis (Martins 2007, Podobnik and Stanley 2008), pairwise coupling analysis (Bury 2013), and even topic-modeling of returns (Doyle and Elkan 2009). Indeed, relevant prior work analyzing historical stock price returns (Fama and French 1993, Laloux *et al.* 1999, Plerou *et al.* 2002) elucidated that the high-dimensional space of stock price returns has a low-dimensional representation.

In parallel with this, there is a long tradition of style analysis in finance in which time series can be selected which serve as useful benchmarks for the performance of other stocks or

*Corresponding author. Email: sethna@lassp.cornell.edu

Table 1. Canonical sectors and major business lines of primary constituent firms. The eight canonical sectors identified by the analysis described here are listed in the column on the left; these were named in accord with the business lines (middle column) of firms that show strong association with these sectors. Some examples are provided in the right column; a full list is available on companion website (Chachra *et al.* 2013).

Canonical sector	Business lines	Prototypical examples
c-cyclical	general and specialty retail, discretionary goods	Gap, Macy's, Target
c-energy	oil and gas services, equipment, operations	Halliburton, Schlumberger
c-financial	banks, insurance (except health)	US Bancorp., Bank of America
c-industrial	capital goods, basic materials, transport	Kennametal, Regal–Beloit
c-non-cyclical	consumer staples, healthcare	Pepsi, Procter & Gamble
c-real estate	realty investments and operations	Post Properties, Duke Realty
c-technology	semiconductors, computers, comm. devices	Cisco, Texas Instruments
c-utility	electric and gas suppliers	Duke Energy, Wisconsin Energy

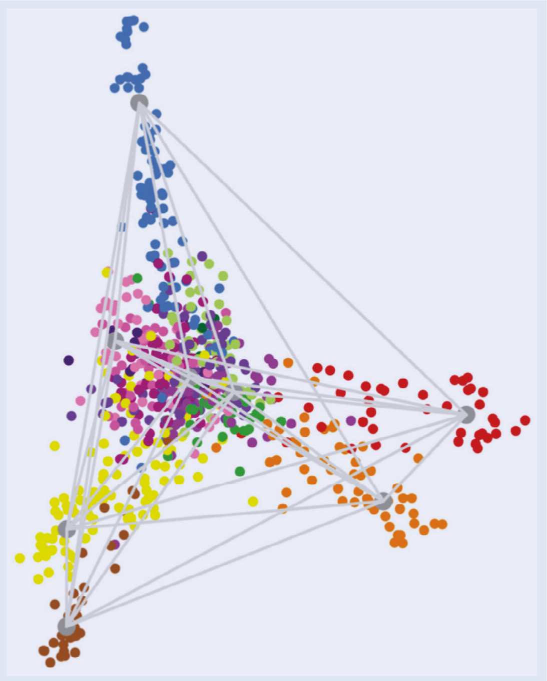

Figure 1. Low-dimensional projection of the stock price return data. Stock price returns are projected onto a plane spanned by two stiff vectors from the SVD of the emergent simplex corners as described in appendix E. Each coloured circle corresponds to one of the 705 stocks in the dataset used in the analysis. Colours denote the sectors assigned to companies by Scottrade® (2015) with the colour scheme of figure B1. The grey corners of the simplex correspond to sector-defining prototype stocks, whereas all other circles are given by a suitably weighted sum of these grey corners. Projections along other singular vectors are shown in Figure E1.

indices. The three-factor model of Fama and French (1993) is one such example. Recently, Vistocco and Conversano (2009) proposed that Archetypal Analysis (AA) (Cutler and Breiman 1994) could provide these benchmark time series while also providing a way to plot this data in a meaningful way. In particular, they provide a triangular plot for Italian mutual funds and suggest parallel coordinate plots or asymmetric maps for higher dimensional representations. The positive decomposition of mutual funds into sectors using standard benchmarks (not derived using AA) was later studied by the same authors (Conversano and Vistocco 2010).

Here, we demonstrate a new, holistic way of classifying stocks into industrial sectors by utilizing the emergent structure of price returns in data space. Beyond the proposal of Vistocco and Conversano, we provide an interpretation of the archetypes of AA as sectors of the economy. This structure is purely contained in the geometry of the time series. Other methods, such as SVD, can discern that there is some such structure but are not well suited to a clean description. AA, on the other hand, determines the convex hull of the data-set making it uniquely suited to creating a quantitative analysis of the data. In particular, if we take the log price returns of individual stocks, remove the overall market return, normalize to zero mean and unit s.d., then stock returns are well-approximated by a hyper-tetrahedral structure. Each lobe of the hyper-tetrahedron is populated by stocks of similar or related businesses (figure 1); the lobe-corners (*canonical sectors*) approximate the returns of companies that are prototypical of individual sectors (table 1). Returns of each stock can be decomposed into a weighted sum (figure 2) of the canonical sector returns (figure 3). Lastly, the canonical sector weights for a given company are dynamic and lead to insights into its evolution (figure 5).

The matrix of daily log returns of a stock s are defined as $r_{ts} = \log P_{ts} - \log P_{(t-1)s}$ where P_{ts} are adjusted closing prices (*i.e.* corrected for stock splits and dividend issues) and t is in trading days. In the present analysis, we used normalized returns, $R'_{ts} = (r_{ts} - \langle r_{ts}\rangle_t)/\sigma_s$, where $\sigma_s^2 = \langle r_{ts}^2\rangle_t - \langle r_{ts}\rangle_t^2$ is the variance (squared volatility) and $\langle\rangle_t$ represents the average over time (trading days). Overall market returns from each stock were also removed, yielding what we shall call the log price returns $R_{ts} = R'_{ts} - \langle R'_{ts}\rangle_s$. (The two degrees of freedom we remove from each stock — the variance and the overall return — are of practical interest elsewhere, but obscure the classification into sectors.) The hyper-tetrahedron, or simplex, which emerges (figure 1) is a self-organized structure: it has prototypical firms in corners (table 1), closely related firms clumped together in each lobe, diversified companies (GE®, Walt Disney®, 3M®, etc.) close to the centre, and the number of lobes denoting how many distinct sectors are exhibited by the data. This suggests a natural way to decompose stocks into canonical sectors: for convex sets, each interior point is representable as a unique weighted sum of corner points, implying here that every stock's return is approximated by a weighted sum of returns from the canonical sectors. Conversely, the weights for a given stock quantify its exposure to the canonical sectors.

We applied an in house python implementation of the AA algorithm described by Mørup and Hansen (Mörup and Hansen 2012). The dataset consisted of 705 US firms' stocks with a minimum $1 billion June 2013 market capitalization and with continuous 20 years (1993–2013) of listing on major exchanges (appendix A). Analysis of this dataset (appendices B and C) revealed eight emergent sectors which were named in accordance with the companies they comprised (prefix *c-* denotes 'canonical'): *c-cyclical* (including retail), *c-energy* (including oil and gas), *c-industrial* (including capital goods and basic materials), *c-financial*, *c-non-cyclical* (including healthcare and consumer non-cyclical goods), *c-real estate*, *c-technology*, and *c-utility*. Calculated participation weights for a sample of 12 firms in figure 2 show a decomposition of their stocks into the canonical sectors with resulting insights discussed in the

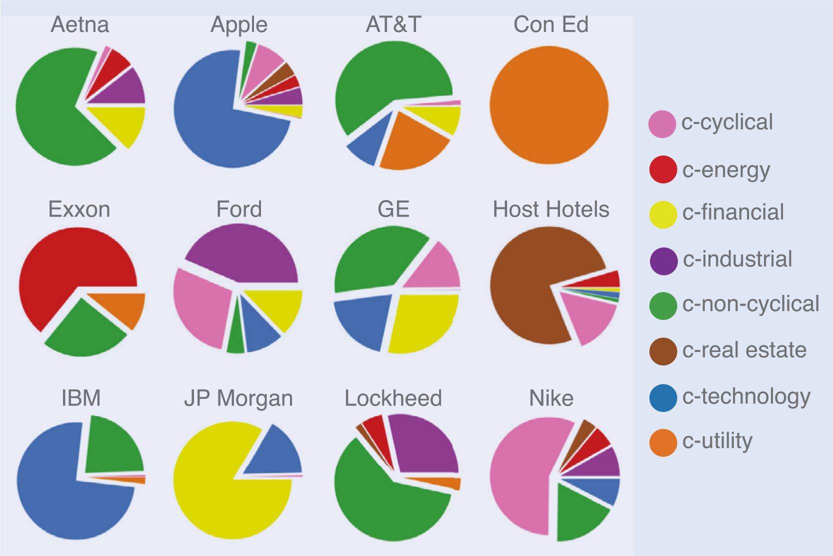

Figure 2. Canonical sector decomposition of stocks of selected companies. A complete set of all 705 stocks is provided on the companion website (Chachra *et al*. 2013); the color scheme is shown on the right. Conglomerates like GE® decompose roughly into their core business lines. Tech firms such as Apple® that sell mass-market consumer goods have an important fraction in *c-cyclical*, whereas IBM® has a significant portion of *c-non-cyclical* returns presumably due to its government contracts. Telecom companies like AT&T® are generally classified under a separate telecom category by major classification systems, yet analysis shows their returns are described by a combination of *c-non-cyclical* and *c-utility* sectors. Health insurance providers like Aetna® are commonly classified as financial services firms, but their returns consist of a major part *c-non-cyclical* and only a minor part of *c-financial*—the healthcare sector is generally less prone to economic downturns. Defense contractors like Lockheed® are listed as capital goods companies, but their returns are seen to be majority *c-non-cyclical* and only a smaller share of *c-industrial* sector.

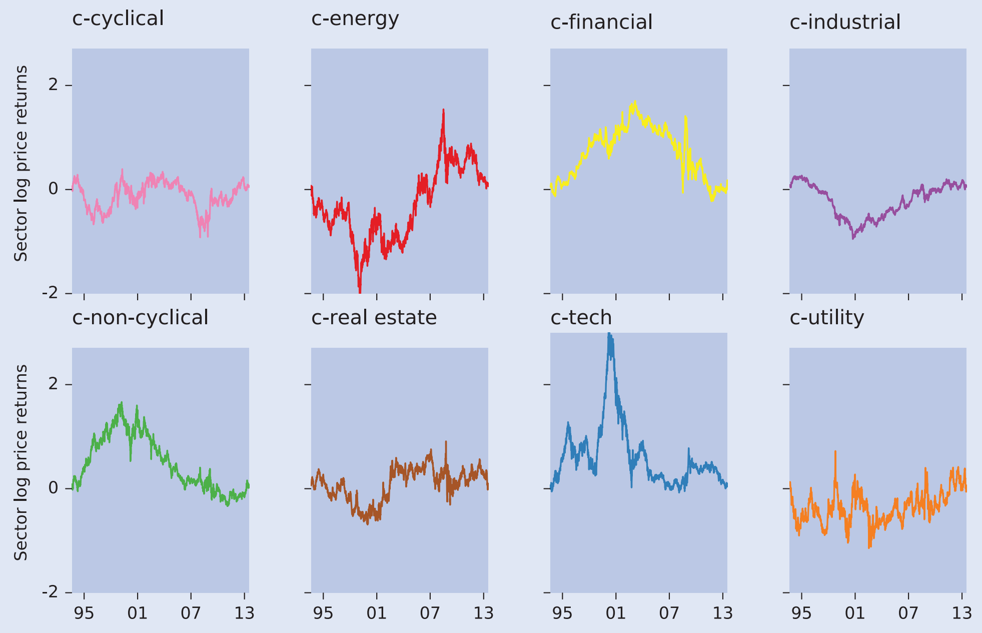

Figure 3. Emergent sector time series. Annualized cumulative log price returns of the eight emergent sectors are shown. The time series capture all important features affecting different sectors: building-up of the dot-com bubble (c. 2000) followed by a burst, the soaring energy valuations (2003–2008) followed by a crash, and the financial crisis of 2008. We note that the dot-com bubble was confined to the c-tech sector whereas the financial crisis effects were spread throughout the sectors. Precise definition of the cumulative returns plotted here is given in equation (C1); other measures of sector dynamics are in figure C1.

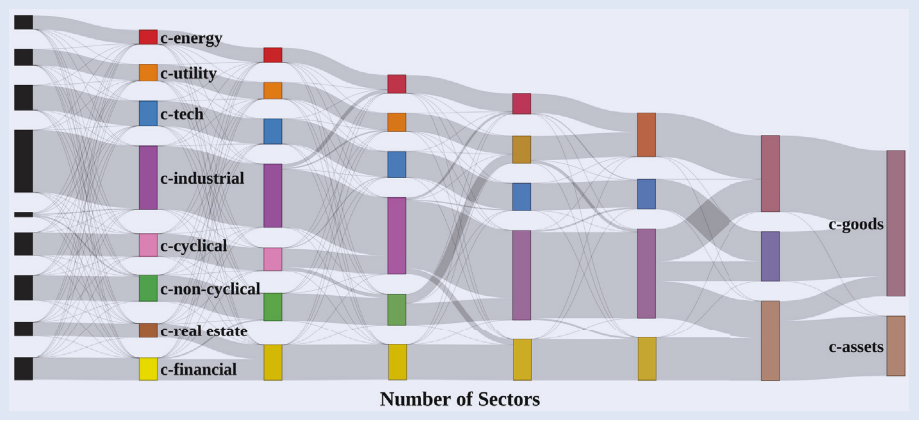

Figure 4. Changes in the decomposition with dimensionality. A Sankey diagram (generated using D3 (Bostock *et al.* 2011)) displaying the relationships between sector decompositions with $n = N + 1$ and $n = N$. Relative node sizes correspond roughly to the amount of the market participating in the sector. Connection width depicts how strongly the sectors for decompositions with different n relate. For details, see appendix G.1.

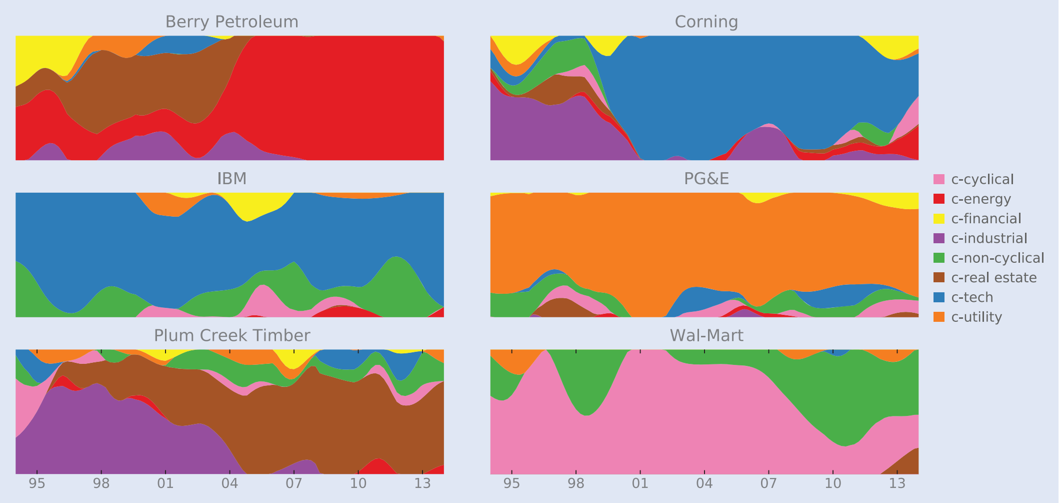

Figure 5. Evolving sector participation weights. Results from the sector decomposition made with rolling two-year Gaussian windows are shown for selected stocks. A complete set of 705 charts is provided on the companion website (Chachra *et al.* 2013). For stable and focused companies such as Pacific Gas & Electric® or IBM®, one sees no significant shifts in sector weights; changes in time agree with errors expected from unresolved fluctuations (Chachra *et al.* 2013). Wal-Mart®'s returns, on the other hand, have moved significantly from *c-cyclical* to *c-non-cyclicals* (consumer staples) in the post-financial crisis years as shown; this is also true of other low-price consumer commodities retailers such as Costco®, but not true of higher price retailers such as Whole Foods®, Macy's®, etc. Corning®, previously an *industrial* firm with a huge presence in optical fibre, suffered in the aftermath of the dot-com crisis and now is classified as a *tech* firm presumably due to its Gorilla® glass used in cellphones, laptop displays, and tablets. Berry Petroleum grew within its home state of California in the early 1990s through development on properties that were purchased in the earlier part of 20th century. In 2003, the company embarked on a transformation (Berry Petroleum Company History 2013) by direct acquisition of light oil and natural gas production facilities outside California. The figure shows a clear shift in the distribution of sector weights as the company has moved toward *c-energy* and away from *c-real estate*. Similarly, as Plum Creek® Timber converted to a real estate investment trust (REIT) in the late 1990s (Plum Creek®History 2014), its sector weights have significantly shifted toward *c-real estate* sector.

caption. Associated with each canonical sector f is a time series of returns. As expected, these series show hallmark historical events of individual sectors (figure 3): the dot-com bubble, the energy crisis, and the financial crisis being the major events in the last two decades.

Determining the correct number of canonical sectors that appropriately describe the space of stock market returns is akin to the more general issue of selecting a signal-to-noise ratio cut-off, or a truncation threshold in the dimensional-reduction of data. The choice of this threshold is generally sensitive to sampling, yet the results presented here are reasonably robust with different choices leading to meaningful and similar decompositions. Figure 4 depicts the changes in the decomposition with dimension. Details of how the figure was generated as well as more information on the two and three dimensional decompositions are available in appendix G.

In addition to the full data-set of 20 years × 705 firms, we also applied the algorithm to overlapping, two-year Gaussian windows to study how the sector weights for firms have evolved in time (figure 5, see also appendix C). As expected, the sector decomposition of firms is dynamic. Mergers, acquisitions, spin-offs, new products, effect of competitive environments or shifting consumer preferences can change the business foci of firms and hence alter the sector association of firms. External events affecting companies in an idiosyncratic manner also show clear signature in this analysis.

The eight-factor decomposition presented here explains 11.1% of the total variation (r^2) in the normalized returns with the market mode removed, and 56% of the random matrix theory explainable variation defined in appendix F. For comparison, the classic three-factor decomposition of portfolio returns by Fama and French (Fama and French 1993) into market mode, market capitalization, and growth vs. value yields an r^2 value of only 4.75%. Indeed, if only three factors are used instead of the eight for the decomposition presented here, the regression yields a comparable r^2 value (5.61%) but there appears to be no correspondence between three factors found by our unsupervised model, and those of Fama and French (figure F1). Carrying out a similar comparison with Fama and French's analysis applied to model portfolio returns, the regression on the S&P 500® yields an r^2 value of 99.4% for Fama and French compared to 93.5% for our eight-factor decomposition (market mode reintroduced). Our decomposition was optimized without concern for market capitalization, which appears to be the key difference: For an equal weighted index of the 338 stocks in the S&P 500® with current tickers and a complete data series in our time of interest, we obtain an r^2 value of 99.0% (97.0% for 3 factors) compared to 95.8% for Fama and French. We conclude that a sector decomposition like the one presented here, perhaps weighted by market capitalization, should be an improved guide to investors, compared to the widespread value/growth and large-cap/small-cap stock characterizations currently used.

Future work remains to address survivorship bias, effects of sampling at different frequencies, and incorporating market capitalization. Investors, analysts, and governments alike would benefit from the development of new investable sector indices (appendix H) that measure the health of our industrial sectors just like the macroeconomic indicators (GDP, housing starts, unemployment rate, etc.) measure the health of our broader economy. Tracing the sectors back in time could elucidate the incorporation of science and technology into our economic system. Finally, our unsupervised decomposition could provide data suitable for quantitative modelling of the internal and external dynamics of our economic system.

Acknowledgements

We thank Jean–Philippe Bouchaud, Ming Huang and Janet Gao for helpful discussions.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was partially supported by NSF [grant number DMR-1312160], [grant number DMR-1719490], [grant number IIS-1247696] and [grant number DGE-1144153].

ORCID

Lorien X. Hayden http://orcid.org/0000-0002-0047-5140
James P. Sethna http://orcid.org/0000-0001-9126-0892

References

Basalto, N., Bellotti, R., Carlo, F.D., Facchi, P. and Pascazio, S., Clustering stock market companies via chaotic map synchronization. *Phys. A: Stat. Mech. Appl.*, 2005, **345**, 196–206.

Berry Petroleum Company History, 2013. Available online at: http://www.bry.com/pages/history.html (accessed 1 January 2015).

Bonanno, G., Caldarelli, G., Lillo, F. and Mantegna, R.N., Topology of correlation-based minimal spanning trees in real and model markets. *Phys. Rev. E*, 2003, **68**, 046130.

Bonanno, G., Vandewalle, N. and Mantegna, R.N., Taxonomy of stock market indices. *Phys. Rev. E*, 2000, **62**, R7615–R7618.

Bostock, M., Ogievetsky V. and Heer J., D3: Data-driven documents. *IEEE Trans. Vis. Comput. Graph. (Proc. InfoVis)*, 2011, **17**, 2301–2309.

Burda, Z., Görlich, A., Jarosz, A. and Jurkiewicz, J., Signal and noise in correlation matrix. *Physica A*, 2004, **343**, 295–310.

Burda, Z., Görlich, A., Jurkiewicz, J. and Wacaw, B., Correlated Wishart matrices and critical horizons. *Eur. Phys. J. B*, 2006, **49**, 319–323.

Bury, T., Market structure explained by pairwise interactions. *Phys. A: Stat. Mech. Appl*, 2013, **392**, 1375–1385.

CBOE®Oil Index, 2013. Available online at: http://www.cboe.com/products/IndexComponentsAuto.aspx?PRODUCT=OIX (accessed 01 January 2015).

Chachra, R., Alemi, A.A., Hayden, L., Ginsparg, P.H. and Sethna, J.P., 2013. Project Website with additional figures and analyses [online]. Available online at:www.lassp.cornell.edu/sethna/Finance (accessed 1 January 2015).

Conlon, T., Ruskin, H. and Crane, M., Cross-correlation dynamics in financial time series. *Phys. A: Stat. Mech. Appl*, 2009, **388**, 705–714.

Conversano, C. and Vistocco, D., Analysis of mutual funds management styles: A modeling, ranking and visualizing approach. *J. Appl. Stat*, 2010, **37**, 1825–1845.

Coronnello, C., Tumminello, M., Lillo, F., Micciche, S. and Mantegna, R., Sector identification in a set of stock return time series traded at the London stock exchange. *Acta Phys. Pol. B*, 2005, **36**, 2653–2679.

Cutler, A. and Breiman, L., Archetypal analysis. *Technometrics*, 1994, **36**, 338–347.

Ding, C. and He, X., K-means clustering via principal component analysis. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML'04, Banff, Alberta, Canada, pp. 29, 2004 (ACM: New York).

Ding, C.H.Q., Li, T. and Jordan, M.I., Convex and semi-nonnegative matrix factorizations. *IEEE Trans. Pattern Anal. Mach. Intell*, 2010, **32**, 45–55.

Dow Jones®US Indices, Industry Indices, 2015. Available online at:www.djindexes.com/mdsidx/downloads/fact_info/Dow_Jones_US_Indices_Industry_Indices_Fact_Sheet.pdf (accessed 1 January 2015).

Doyle, G. and Elkan, C., Financial topic models. In *Proceedings of the NIPS Workshop on Applications for Topic Models: Text and Beyond*, Whistler, Canada, 2009.

Eom, C., Oh, G., Jeong, H. and Kim, S., Topological properties of stock networks based on random matrix theory in financial time series. *Papers*, arXiv.org, 2007.

Fama, E.F. and French, K.R., Common risk factors in the returns on stocks and bonds. *J. Financ. Econ.*, 1993, **33**, 3–56.

Fenn, D.J., Porter, M.A., Williams, S., McDonald, M., Johnson, N.F. and Jones, N.S., Temporal evolution of financial-market correlations. *Phys. Rev. E*, 2011, **84**, 026109.

Heimo, T., Kaski, K. and Saramñki, J., Maximal spanning trees, asset graphs and random matrix denoising in the analysis of dynamics of financial networks. *Phys. A: Stat. Mech. Appl.*, 2009, **388**, 145–156.

Hyvärinen, A. and Oja, E., Independent component analysis: Algorithms and applications. *Neural Netw.*, 2000, **13**, 411–430.

Kersting, K., Wahabzada, M., Thurau, C. and Bauckhage, C., Hierarchical convex NMF for clustering massive data. *J. Mach. Learn. Res. Proc. Track*, 2010, **13**, 253–268.

Kim, D.H. and Jeong, H., Systematic analysis of group identification in stock markets. *Phys. Rev. E*, 2005, **72**, 046133.

Kullmann, L., Kertész, J. and Mantegna, R.N., Identification of clusters of companies in stock indices via Potts super-paramagnetic transitions. *Phys. A: Stat. Mech. Appl*, 2000, **287**, 412–419.

Laloux, L., Cizeau, P., Bouchaud, J.P. and Potters, M., Noise dressing of financial correlation matrices. *Phys. Rev. Lett.*, 1999, **83**, 1467–1470.

Lee, D.D. and Seung, H.S., Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999, **401**, 788–791.

Li, T. and Ding, C., The relationships among various nonnegative matrix factorization methods for clustering. In *Sixth International Conference on Data Mining, 2006. ICDM'06*, pp. 362–371, 2006.

Livan, G., Alfarano, S. and Scalas, E., Fine structure of spectral properties for random correlation matrices: An application to financial markets. *Phys. Rev. E*, 2011, **84**, 016113.

Mantegna, R.N., Hierarchical structure in financial markets. *Eur. Phys. J. B. Conden. Matter Complex Syst.*, 1999, **11**, 193–197.

Martins, A.C., Random, but not so much a parameterization for the returns and correlation matrix of financial time series. *Phys. A: Stat. Mech. Appl.*, 2007, **383**, 527–532.

Mehta, M.L., *Random Matrices*, Vol. 3, 2004 (Academic Press: Boston, MA).

Morgan Stanley®High-Tech 35 Index, 2005. Available online at: www.nasdaq.com/options/indexes/msh.aspx (accessed 1 January 2015).

Mörup, M. and Hansen, L.K., Archetypal analysis for machine learning and data mining. *Neurocomputing*, 2012, **80**, 54–63.

Musmeci, N., Aste, T. and Di Matteo, T., Relation between financial market structure and the real economy: Comparison between clustering methods. 2014. Available online at SSRN: https://ssrn.com/abstract=2525291.

Nadig, D. and Crigger, L., Signal from noise. *J. Indexes*, 2011, **14**, 40–43.

Pastor, L., Heaton, J. and Foss, A., The index is dead long live the index. *J. Indexes*, 2013, **16**(16–21), 55.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., Guhr, T. and Stanley, H.E., Random matrix approach to cross correlations in financial data. *Phys. Rev. E*, 2002, **65**, 066126.

Plum Creek®History, 2014. Available online at: http://www.plumcreek.com/AboutPlumCreek/History/tabid/55/Default.aspx (accessed 1 January 2015).

Podobnik, B. and Stanley, H.E., Detrended cross-correlation analysis: A new method for analyzing two nonstationary time series. *Phys. Rev. Lett.*, 2008, **100**, 084102.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P., *Numerical Recipes: The Art of Scientific Computing*, 3rd ed., Vol. 3, 2007 (Cambridge University Press: New York, NY).

Russell 3000®Index, 2015. Available online at:www.russell.com/indexes/data/fact_sheets/us/russell_3000_index.asp (accessed 1 January 2015).

Scottrade®, 2015. Available online at: www.scottrade.com (accessed 1 January 2015).

S&amp;P 500®Index, 2014. Available online at: us.spindices.com/indices/equity/sp-500 (accessed 1 January 2015).

Tagiliani, M., *The Practical Guide to Wall Street*, Vol. 1, 2009 (John Wiley & Sons: Hoboken, NJ).

Thurau, C., Kersting, K. and Bauckhage, C., Convex non-negative matrix factorization in the wild. In *Proceedings of the Ninth IEEE International Conference on Data Mining ICDM '09*, pp. 523–532, 2009.

Thurau, C., Kersting, K. and Bauckhage, C., Yes we can–simplex volume maximization for descriptive web scale matrix factorization. In *Proceedings of the CIKM*, edited by J. Huang, N. Koudas, G.J.F. Jones, X. Wu, K. Collins-Thompson and A. An, pp. 1785–1788, 2010 (ACM: New York, NY).

Thurau, C., Kersting, K., Wahabzada, M. and Bauckhage, C., Convex non-negative matrix factorization for massive datasets. *Knowl. Inform. Syst.*, 2011, **29**, 457–478.

Tsalmantza, P. and Hogg, D.W., A Data-driven model for spectra: Finding double redshifts in the sloan digital sky survey. *Astrophys. J.*, 2012, **753**, 122.

Vistocco, D. and Conversano, C., Visualizing and clustering financial portfolios using internal compositions. In *Presented at Statistical Methods for the Analysis of Large Data-Sets Pescara*, Italy, 23–25 September, 2009. Available online at: http://new.sis-statistica.org/wp-content/uploads/2013/10/CO09-Visualizing-and-clustering-financial-portfolios-using.pdf.

Wang, Y.X. and Zhang, Y.J., Nonnegative matrix factorization: A comprehensive review. *IEEE Trans. Knowl. Data Eng.*, 2013, **25**, 1336–1353.

Yahoo!®Finance, 2015. Available online at: finance.yahoo.com (accessed 1 January 2015).

Zhang, Z., Li, T., Ding, C. and Zhang, X., Binary matrix factorization with applications. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, ICDM '07*, pp. 391–400, 2007 (IEEE Computer Society: Washington, DC).

Appendix A. Data-set particulars

Company names, tickers, listed-sectors and market caps of US-based firms used in this analysis were obtained from Scottrade® (Scottrade® 2015). Daily closing prices adjusted for stock splits and dividend issues were obtained from Yahoo® Finance (Yahoo!®Finance 2015). The rare cases of missing prices in the time series were replaced with linearly interpolated values. A brief summary of listed sectors and number of companies in each is provided in Table A1 and a full list of company names, tickers, market caps and listed-sector info is available on the companion website (Chachra *et al.* 2013).

Table A1. Listed sectors and number of companies dataset analyzed. Tickers for each company were obtained from (Scottrade® 2015).

Listed sector	Companies
Basic materials	58
Capital goods	61
Consumer cyclical	41
Consumer non-cyclical	40
Energy	42
Financial (+Real estate)	138
Healthcare	53
Services (+Retail)	101
Technology	93
Telecom	6
Utility	57
Transport	15
TOTAL	705

Appendix B. Returns factorization and sector decomposition

A variety of factorization algorithms have been developed in recent years for dimensional reduction, classification or clustering. Examples include archetypal analysis (AA) (Cutler and Breiman 1994), heteroscedastic matrix factorization (Tsalmantza and Hogg 2012), binary matrix factorization (Zhang *et al.* 2007), K-means clustering (Ding and He 2004), simplex volume maximization (Thurau *et al.* 2010), independent component analysis (Hyvärinen and Oja 2000), non-negative matrix factorization (NMF) (Lee and Seung 1999; Wang and Zhang 2013) and its variants such as the semi- and convex-NMF (Ding *et al.* 2010), convex hull NMF (Thurau *et al.* 2011) and hierarchical convex NMF (Kersting *et al.* 2010), among others. Each method has a unique interpretation (Li and Ding 2006) and therefore, a successful application of any of these methods is contingent upon the underlying structure of the data.

The hyper-tetrahedral structure of log price returns seen in our analysis motivates a decomposition so that each stock's return is a weighted mixture of canonical sectors, constrained to lie in the convex hull of the data. Hence we employ AA factorization which is defined as:

$$\begin{aligned} R_{ts} &\sim R_{ts'}C_{s'f}W_{fs} \\ C_{s'f} &\geq 0, \sum_{s'} C_{s'f} = 1, \\ W_{fs} &\geq 0, \sum_{f} W_{fs} = 1. \end{aligned} \tag{B1}$$

Columns of $R_{ts}C_{sf} = E_{tf}$ are the emergent sector time series (basis vectors) representing the n corners of the hyper-tetrahedron, and W_{fs} are the participation weights ($W_{fs} \geq 0$) in sector f so that $\sum_f W_{fs} = 1$ for each stock s. The sector matrix E_{tf} is within the convex hull ($C > 0$, $\sum_s C_{sf} = 1$) of the data R_{ts}. It can be found by either minimizing the squared error with convex constraints in factorization as originally proposed (Cutler and Breiman 1994), or by making a convex hull of the dataset and choosing one or more of its vertices to be basis vectors, or by making a convex hull in low-dimensions and choosing one or more of its vertices to be basis vectors (Thurau *et al.* 2009), or by minimizing after initializing with candidate archetypes that are guaranteed to lie in the minimal convex set of the data (Mörup and Hansen 2012). The columns of the C matrix are shown in figure B1.

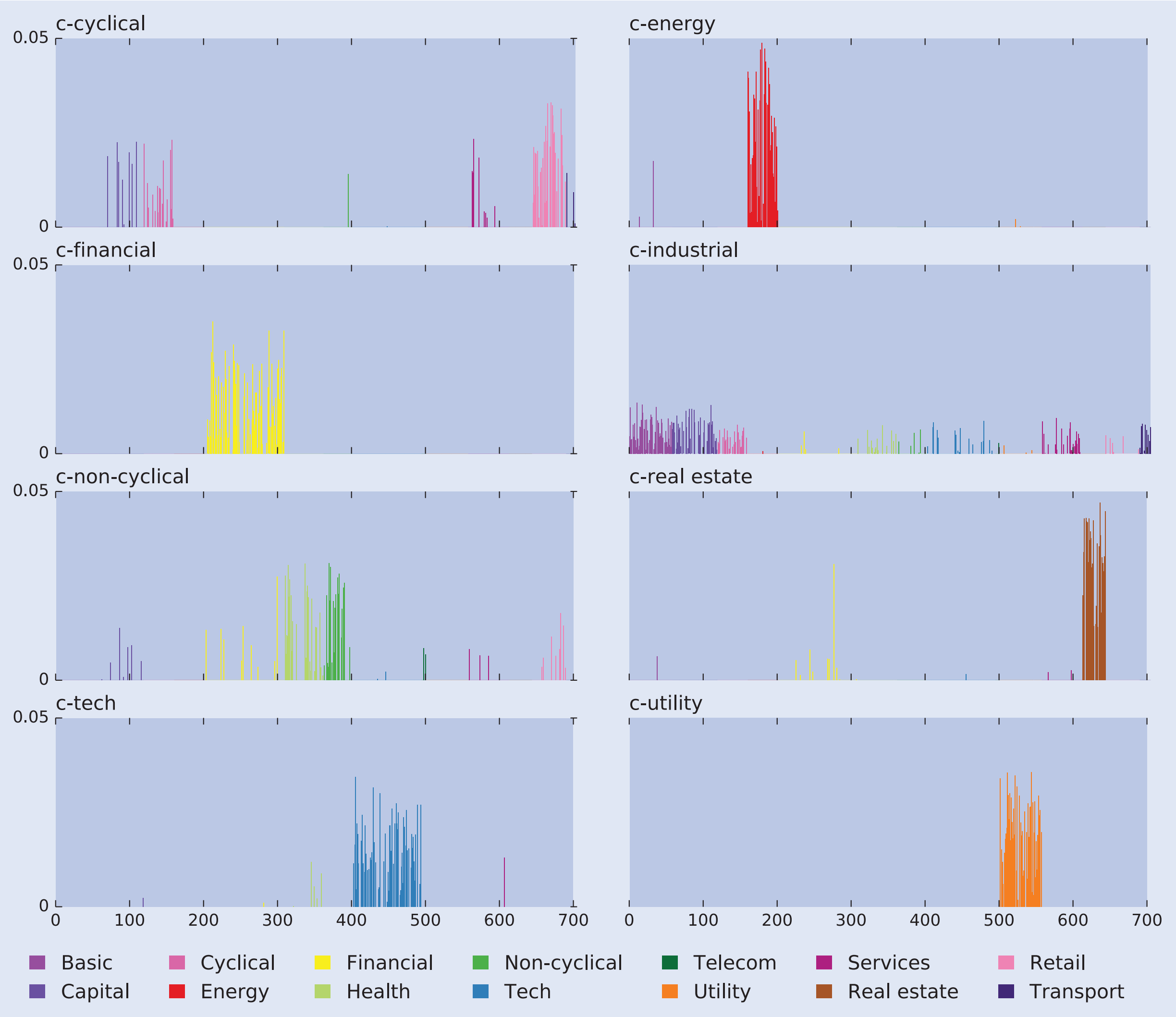

Figure B1. Canonical Sector Constituents (shown as columns of the C_{sf}). C_{sf} represents a weighted combination of stocks that defines the canonical sector each of which has a time series represented by E_{tf} that is given by $E_{tf} = R_{ts}C_{sf}$. The eight subplots show the constituent participation component of stocks in each canonical sector f. Canonical sectors are labeled on the plot; their names were chosen according to the listed sectors of firms that comprise them. Noteworthy features seen above include the co-association of listed sectors: basic, capital, transport and part of cyclicals into *industrial goods*. Similarly, healthcare and non-cyclicals are coupled together in what we call *non-cyclicals*. Canonical *retail* goes primarily with listed retail and cyclicals. Stocks are colored by listed sectors as shown at the bottom. Listed sector information was obtained from Scottrade® (2015).

Appendix C. Calculations and convergence

Numerical computations were performed using an in-house Python language implementation of the principal convex hull analysis (PCHA) algorithm as described in (Mörup and Hansen 2012). For the full dataset, the factorization $R = EW$, with $E = RC$ as defined in equation (B1) converged in 35 iterations to a predefined tolerance value of $\Delta_{SSE} < 10^{-7}$, where Δ_{SSE} is the average difference in the sum of squared error per matrix element in $R - EW$ from one iteration to the next. The resulting columns of E_{tf} are shown in figure C1 (top row). Annualized cumulative log returns are obtained by summing rows of E_{tf}:

$$Q_f(\tau) = \frac{1}{\sqrt{250}} \sum_{t=0}^{t=\tau} E_{tf} \tag{C1}$$

The time series $Q_f(\tau)$ are shown in figure 3 and the middle row of figure C1. Weights W_{fs} for selected stocks are shown in figure 2, the remainder are available on the companion website (Chachra *et al*. 2013). In each canonical sector f, the component of weights for companies are shown in figure C2.

The analysis of evolving sector weights was performed similarly, but with a sliding Gaussian time window. We decomposed the local normalized log returns for each stock into the canonical sectors determined from the entire time series. Each column (time series) of the returns matrix R_{ts} was multiplied with a Gaussian, $G_\mu(\tau) = \exp(-(\tau - \mu)^2/(2 \times 250^2))$ of standard deviation 250 centered at μ to obtain R^μ_{ts}. We use $C_{s'f}$ found using the full dataset (equation (B1)) (corresponding to keeping the sector-defining simplex corners fixed). R^μ_{ts} is factorized to obtain new weights W^μ_{fs} that describe sector decomposition of stocks in that period focused at $t = \mu$: $R^\mu = R^\mu_{ts'} C_{s'f} W^\mu_{fs}$. μ is increased in steps of 50 starting at $\mu = 0$ and ending at $\mu = 5000$, and W^μ is calculated at each μ with the corresponding R^μ. These results are plotted in figure 5 for a select group of companies; the remainder are available on the companion website (Chachra *et al*. 2013).

To address the challenge of distinguishing signal from noise in the evolving sector weights, we emulated the effect of noise for each of the companies from figure 5. For each of these companies, we took its sector weights, $\vec{\omega_f}$, and multiplied by E_{tf} to obtain a time series for the company with weights that are constant in time. We then added gaussian random noise with standard deviation one and replaced these companies by this simulated data. Figure C3 shows the comparison between the real flows and the simulated constant data with noise added. General features are shown to be signal while small fluctuations are consistent with noise.

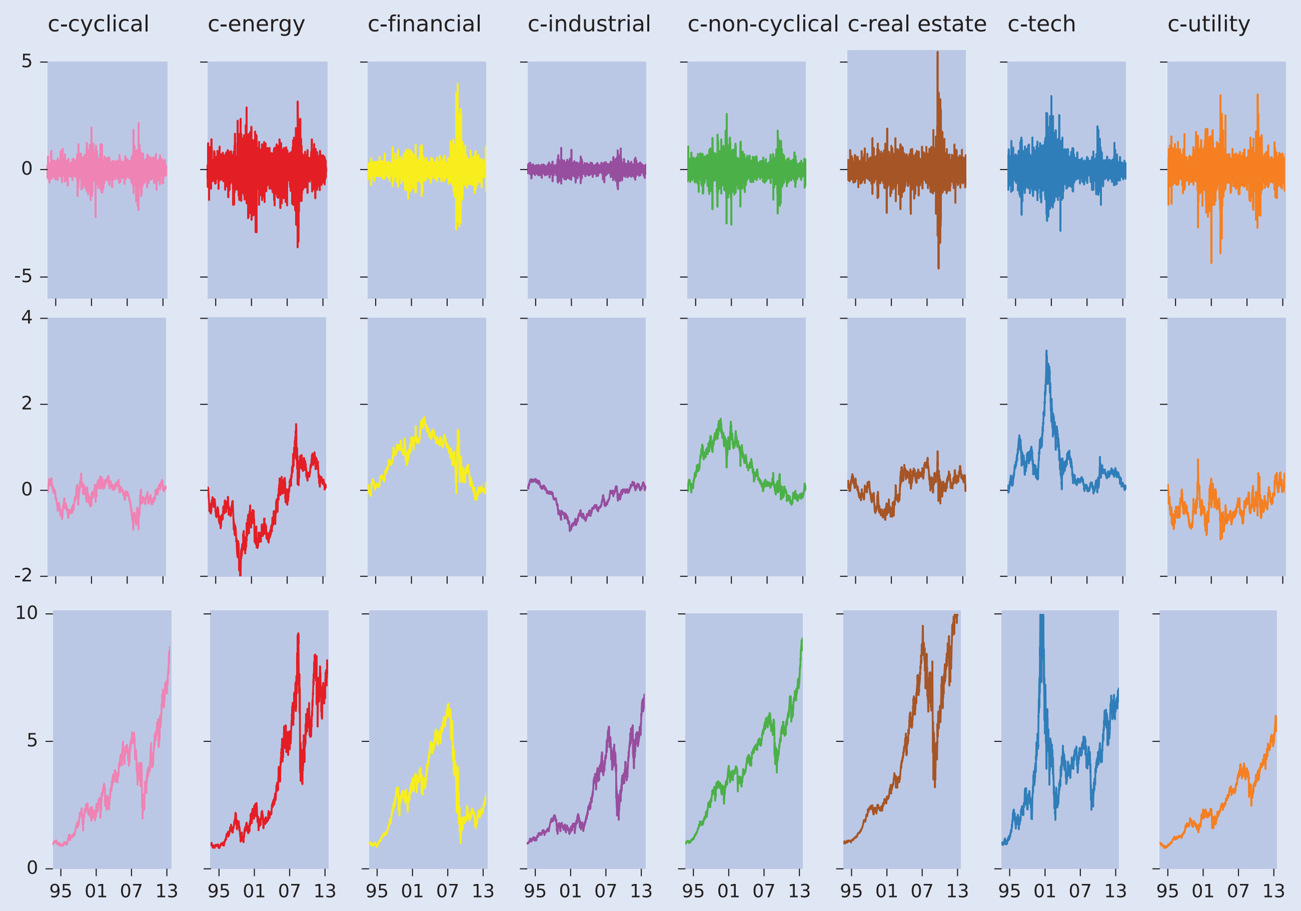

Figure C1. Canonical sector time series. Top row: normalized log returns (columns of E_{tf}), middle row: cumulative log returns (same as figure 3 and defined in equation (C1)), and bottom row: unweighted price index of canonical sectors (equation (H1)).

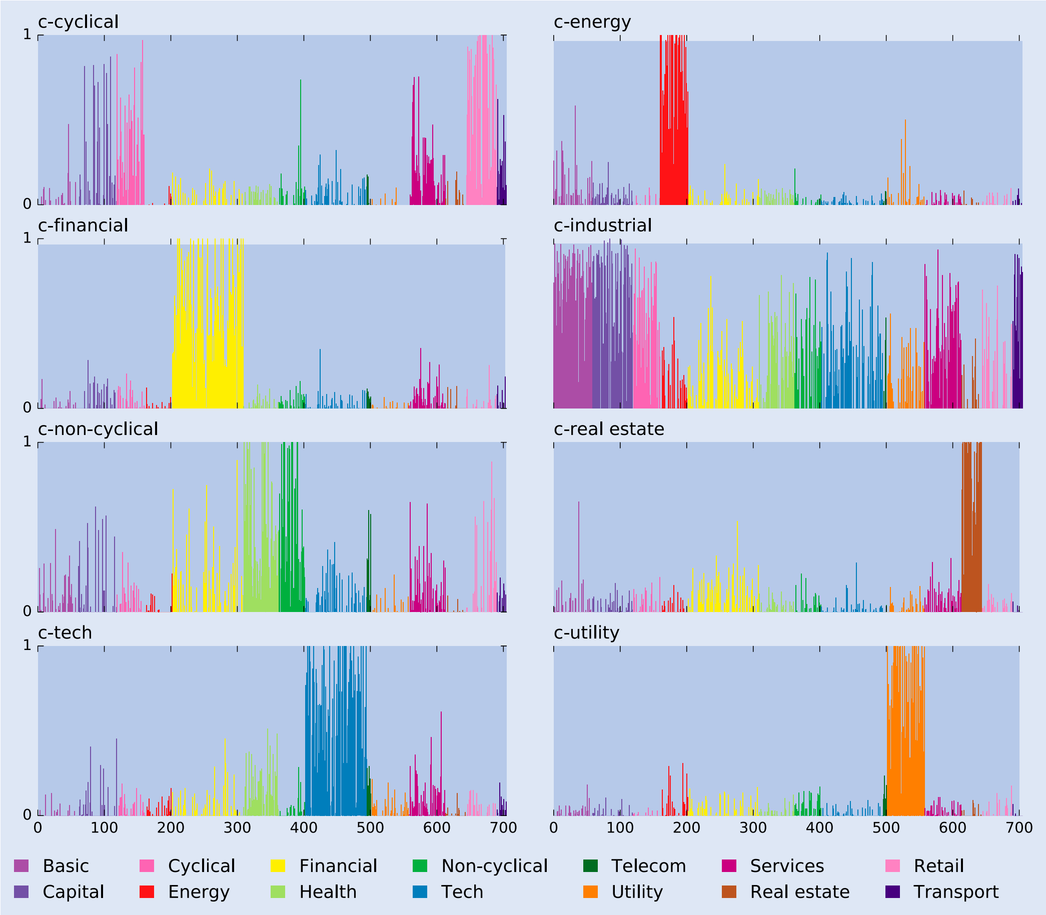

Figure C2. Weight distribution in canonical sectors. Each of the eight subplots shows the constituent participation weights of all 705 companies in a canonical sector (rows of W_{fs}). Stocks are colored by listed sectors as shown at the bottom. Listed sector information was obtained from (Scottrade® 2015).

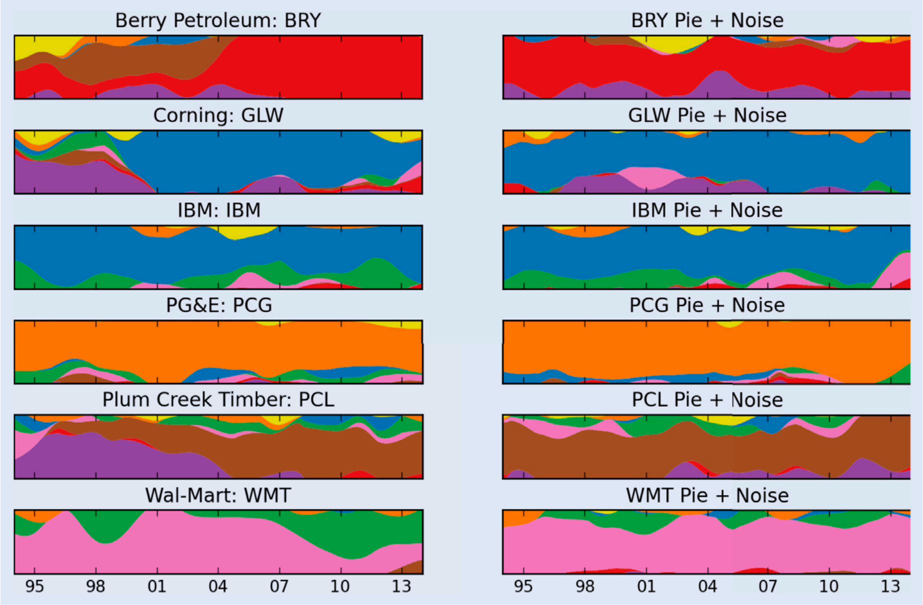

Figure C3. Comparison between flow diagrams presented in figure 5 with simulated data. The simulated data is created from the dot product of the weight vector of the company with the corner time series as described in this section. This yields a version of the company with constant weights in time. To this we add gaussian noise with standard deviation one and repeat the analysis to generate the flows in time. In the left column are the actual flows for companies, on the right is their constant in time counterpart with added noise. We see that key features are in fact signal while small fluctuations correspond to noise. Colour scheme as in figures 2 and 5.

Appendix D. Dimensionality of the space of price returns

It is often the case with large data-sets that the effective dimensionality of the data space is much lower when one filters out the noise. Of the many dimensional reduction methods, the most commonly used is singular value decomposition (SVD) (Press *et al.* 2007), a deterministic matrix factorization. We discuss SVD in more detail in order to draw a contrast with previous SVD results, and to apply it for quantifying the explainable variation in the returns data.

An SVD of R_{ts} is a matrix factorization (Press *et al.* 2007) $R_{ts} = U_{tf}\Sigma_{ff'}V^T_{f's}$ such that matrices U and V are orthogonal; Σ is a diagonal matrix of ‘singular values’. If the goal were purely rank-reduction, n entries of Σ chosen to lie above ‘noise threshold’ are retained and the rest truncated so that $0 \leq f,\ f' \leq n$. This effectively reduces the dimension of R to n. The choice of n can be informed by the distribution of singular values as discussed later. The rows of V^T are precisely the eigenvectors of the stock-stock returns correlation matrix, $\xi_{ss'} \sim R^T_{st}R_{ts}$. It was previously reported that some components of the stiff eigenvectors of this stock-stock correlation matrix loosely corresponded to firms belonging to the same conventionally identified business sector (Plerou *et al.* 2002) (but see figure D1).

After normalizing the log returns, the returns matrix R has entries of unit variance. If the entries were uncorrelated random variables drawn from a standard normal distribution, their singular values (which are also the positive square roots of the eigenvalues of $R^T R$) would be

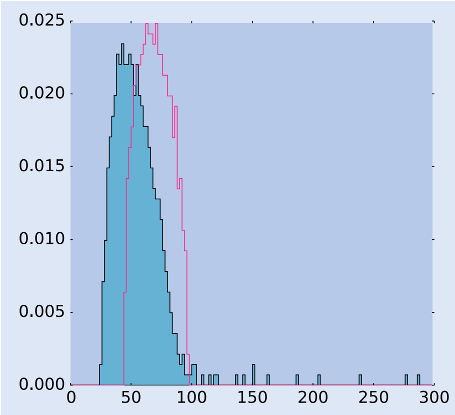

Figure D2. Normalized distribution of singular values. Filled blue histogram corresponds to distribution of singular values of returns from the dataset R_{ts}—one notices a clear separation of the hump-shaped bulk of singular values, and about 20 stiff singular values (the largest singular value ~952, corresponding to the *market mode* is not shown). Pink line histogram outline shows the distribution of singular values of a matrix of the same shape as R but containing purely random Gaussian entries.

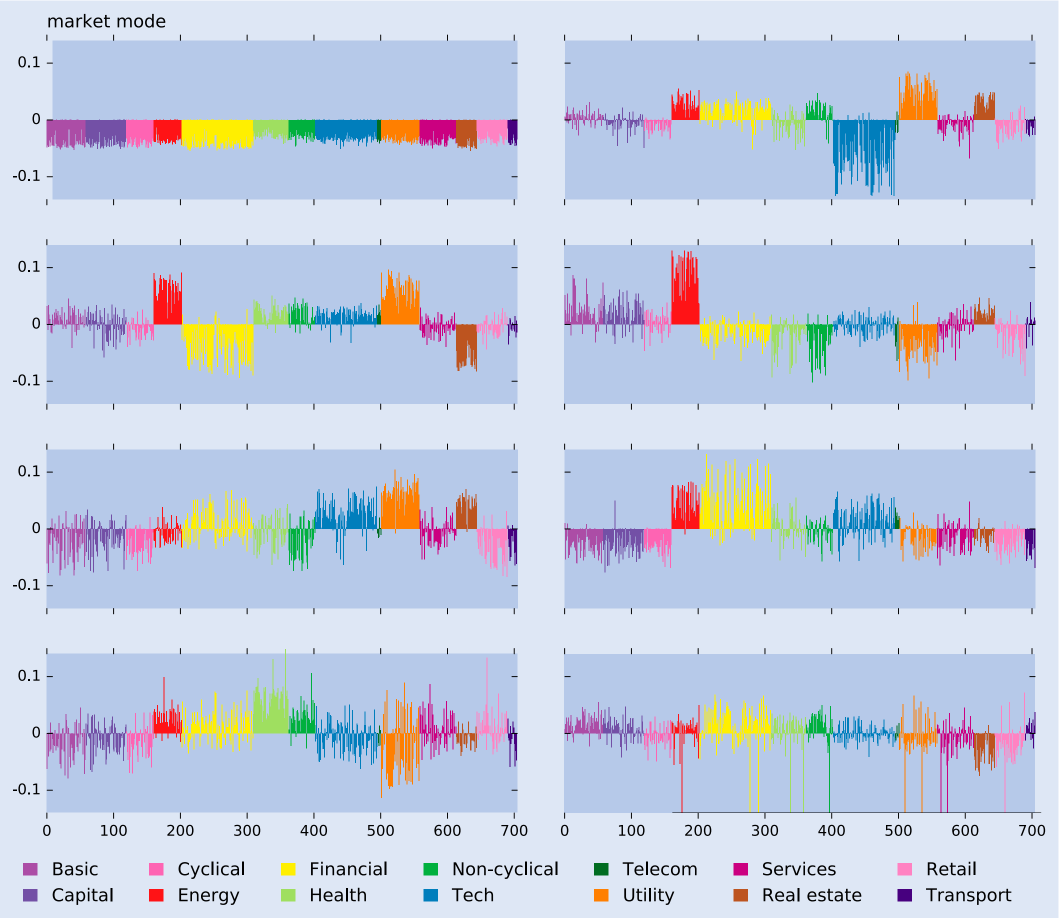

Figure D1. Singular vectors V^T_{fs} of the SVD of returns R_{ts}. The orthonormal right singular vectors (rows of V^T_{fs}) of SVD of R_{ts} are equivalent to the eigenvectors of the stock-stock correlation matrix $\xi_{ss'} \sim R^T R$. Eight of these stiffest eigenvectors including the *market mode* are shown in rows of two at a time. Each has 705 components corresponding to stocks in the dataset. The *market mode* with all components in the same direction describes overall fluctuations in the market; it was excluded from the analysis described in the paper. Previous work (Plerou *et al.* 2002) has suggested that each eigenvector of the stock-stock correlation matrix describes a listed sector, however as seen above, a more correct interpretation is that each eigenvector is a mixture of listed sectors with opposite signs in components. For example, the stiffest direction (after market mode) has positive components in real estate and utility, but negative in tech. Less stiff eigenvectors (including the last one shown here), do not contain sector-relevant information. Stocks are coloured by listed sectors as shown at the bottom. Listed sector information was obtained from (Scottrade® 2015).

described by Wishart statistics (Mehta 2004). The Wishart ensemble for a matrix of size $\alpha \times \beta$ predicts a distribution of singular values with a characteristic shape (Mehta 2004), bounded for large matrices by $\sqrt{\alpha} \pm \sqrt{\beta}$. Comparing the stock correlations with Wishart statistics has been previously used to filter noise from financial datasets (Laloux *et al.* 1999). As shown in figure D2, most singular values of the returns matrix R lie in the bulk below the bound set by the Wishart ensemble, whereas only ∼20 fall outside that cut-off (The singular value bounds of a random Gaussian rectangular matrix of size $\alpha \times \beta$ can be shown to be $\sqrt{\alpha} \pm \sqrt{\beta}$ for large matrices.) Historically, this has served as indication that singular values within the bulk correspond to noise (Laloux *et al.* 1999). Recently, however, much progress has been made in the development of techniques to extract signal from the bulk (Burda *et al.* 2004, 2006, Livan *et al.* 2011). Our method does not claim to capture this information. Rather, we measure its ability to capture variation in the data above the cutoff by means of random matrix theory explainable variation as defined in section F. The largest singular value of R_{ts} corresponds to what we will refer to as the 'market mode' as this represents overall simultaneous rise and fall of stocks. In the analysis presented in this paper, this mode has been filtered from the returns matrix by projecting the R matrix into the subspace spanned by all non-market mode eigenvectors. This is nearly equivalent to filtering the market mode using simple linear regression (as done commonly (Plerou *et al.* 2002)), although more convenient.

Appendix E. Low-dimensional projections of price returns

The emergent low-dimensional, hyper-tetrahedral (simplex) structure of stock price returns can be seen by projecting the dataset into stiff 'eigenplanes'. Eigenplanes are formed by pairs of right singular vectors from a SVD. Here, we construct an SVD of the simplex corners, $E_{tf} = X_{tk} Y Z^T_{kf}$; simplex corners are mapped to columns of YZ^T because $YZ^T_{kf} = X^T_{kt} E_{tf}$ (in other words, X^T_{kt} is a projection operator). The plots in figure E1 are the projections of the dataset, $X^T_{kt} R_{ts} = v_{ks}$. The rows of v taken in pairs form the axes of the projections in figures 1 and E1. With those plots, it becomes clear that the eigenplanes represent projections of a simplex-like data into two-dimensions. Secondly, we note that the simplex structure becomes less clear as one looks at planes corresponding to smaller singular value directions; the signal eventually becomes buried in the noise.

Similarly, the results of the factorization can be seen in eigenplanes from the SVD of $E_{tf} W_{sf} = L_{tk} M N^T_{ks}$. These results (rows of MN^T_{ks}) are shown in figure E2, where we notice that the data is now perfectly resides in simplex region as expected due to constraints.

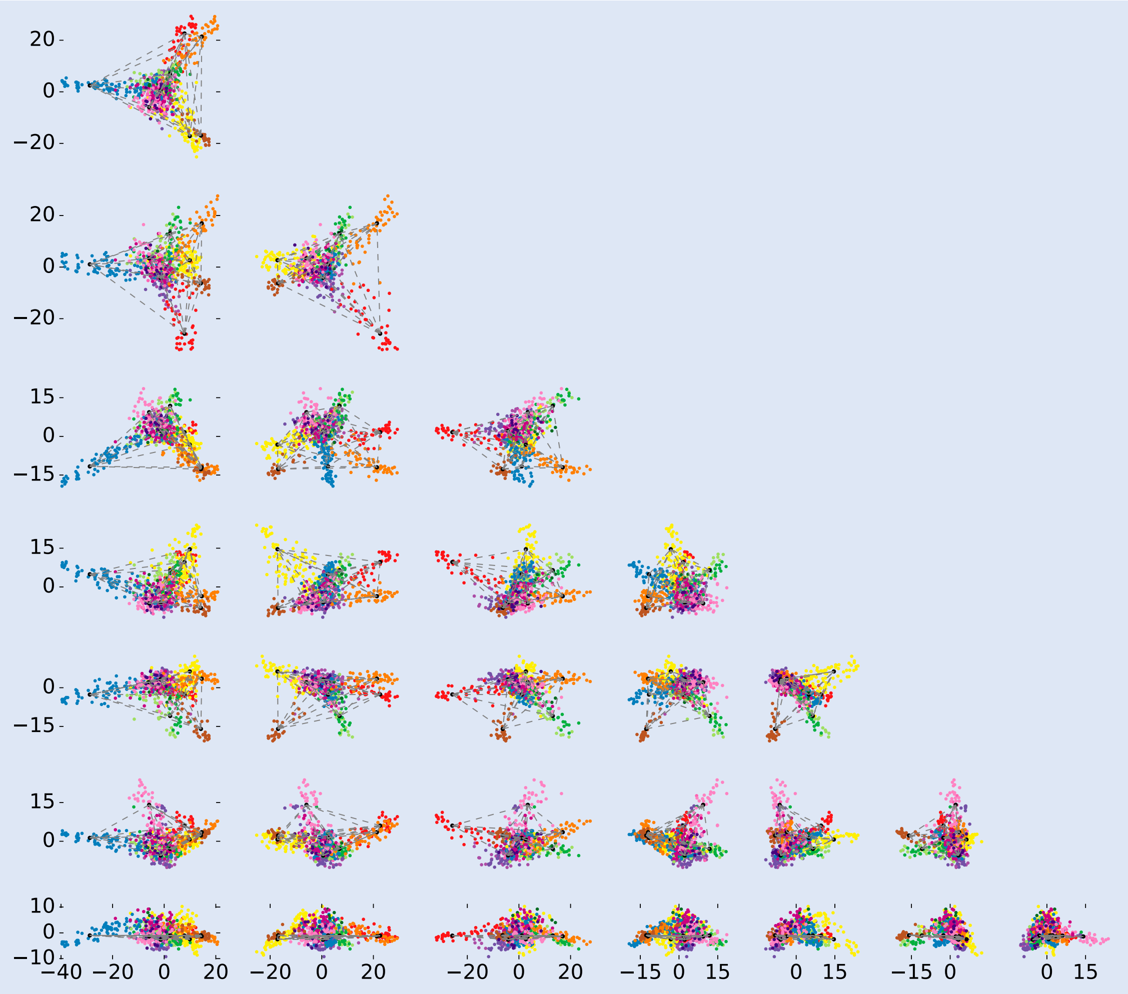

Figure E1. Low-dimensional projections of stock returns data, coloured by Scottrade® sector. Each coloured circle represents a stock in our dataset and is coloured according to sectors assigned by Scottrade® (Scottrade® 2015) as indicated in figure D1. The first row is equivalent to figure 1. Black circles represent the archetypes found with our analysis. The (i, j)th figure in the grid is a plane spanned by singular vectors i and $j + 1$ (rows of $X^T R$) from the calculations described earlier. Projections after the factorization are shown in figure E2.

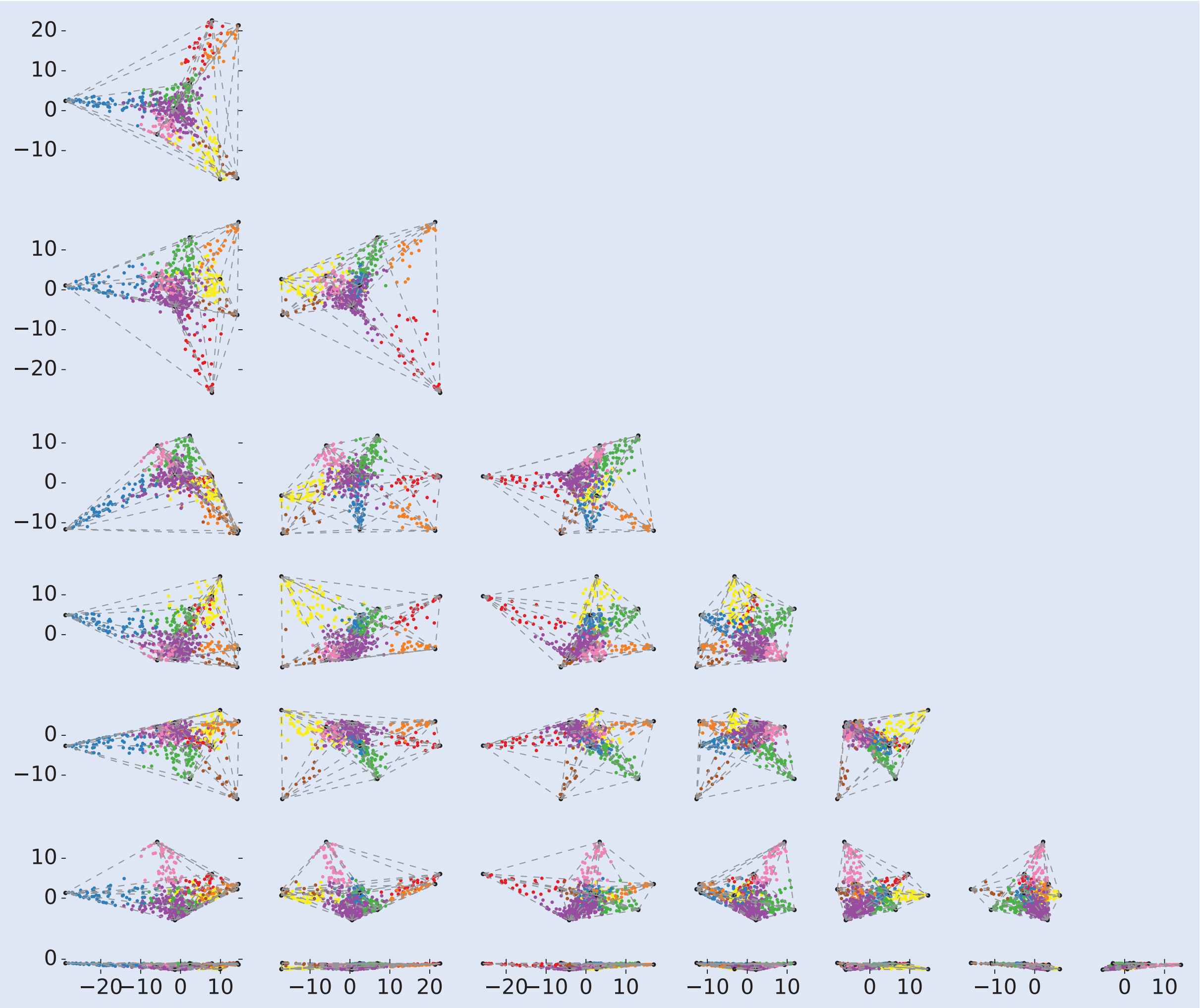

Figure E2. Cross-sections along eigenplanes of the factorized returns. Each coloured circle represents a stock in our dataset and is coloured according to the primary canonical sector association with the colour scheme in figure 2. Black circles represent the archetypes found with our analysis. The (i, j)th figure in the grid is a plane spanned by singular vectors i and $j + 1$ (rows of MN^T) from the calculations described earlier. Projections of raw data (before the factorization) are shown in figure E1. Note that the colours are very similar to those of the traditional Scottrade® classification shown in figure E1; the colour schemes were designed to roughly match. Note that here all points have been projected into the hyper-tetrahedron by our factorization.

Appendix F. Coefficient of determination (r^2)

We measured the goodness of the returns decomposition $R = EW$ by measuring the coefficient of determination (r^2) as follows:

$$r^2 = 1 - SSE/SST \qquad \text{(F1)}$$

Here, SSE denotes the sum of square errors $||R - EW||_F^2$, and SST is the total sum of squares $||R||_F^2$. This is also known as the *proportion of variance explained* (PVE). For the factorization of the full dataset, normalized with the market mode removed, the calculated r^2 value is 11.1%. The SVD of R with singular values shown in figure D2 provides a convenient way to put this number in context for the returns dataset. Only 20 singular values (excluding the market mode) were above the cut-off that was predicted by random matrix theory for a matrix of purely random Gaussian entries. For any matrix M with elements m_{ij}, the norm $||M||_F^2 = \sum_{i,j} m_{ij}^2 = \sum_i s_i^2$, where s_i are the singular values (Press *et al.* 2007). Thus, the fraction of intrinsic variation in R above the cutoff is the sum of squares of the 20 singular values (not including market mode) divided by SST, $\sum_{i=1}^{i=20} s_i^2/||R||_F^2 = 19.8\%$. Therefore, as a first approximation, the factorization explains $11.1/19.8 = 56\%$ of the *random matrix theory (RMT) explainable variation.*

For reference we provide the RMT explainable variation for the factor decomposition of Fama and French, the classification by Scottrade®, and the top 8 singular vectors given by SVD. The percentage of the RMT explainable variation for different numbers of factors compared to the 3 factor decomposition of Fama and French is shown in table F1. Fama and French have the benefit of allowing factors to have positive or negative weights. In order to compare with another non-negative decomposition, we fix the weight matrix according to the Scottrade® labels and run archetypal analysis for this $n = 14$ factor version. The r^2 value for this decomposition is 10.7% with a corresponding RMT explainable variance of 54.2% compared to 56% for our 8 factors. For completeness, we also note that if R is rank-reduced to the eight stiffest components found by SVD (not including market mode), then the factorization explains 85% of the the RMT explainable variation in R with overall results in good accord with the analysis presented here. This implies that sector decomposition information was already contained in the stiff modes from the SVD of R, however SVD is not the appropriate tool for the decomposition. Figure F1 further shows that our unsupervised 3-factor decomposition appears quite distinct from Fama and French's hand-created one.

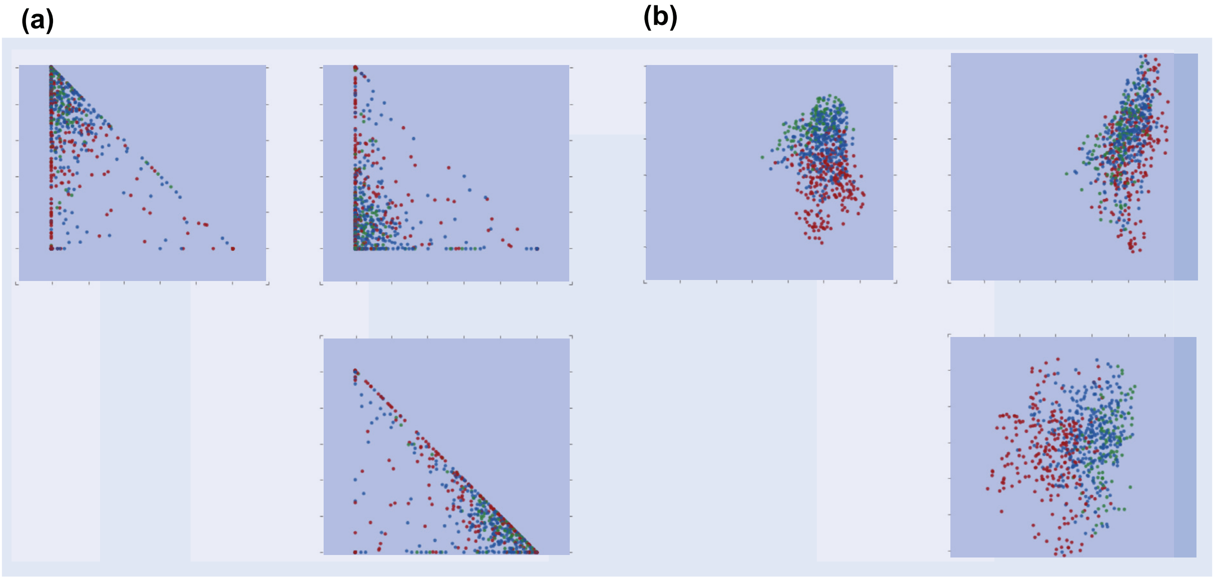

Figure F1. Three Factor Model vs. Fama and French. 2D projections of the weights for each company in the SP500 with current tickers and data in the date range we consider. Red denotes companies with large market caps (market cap >10 billion), blue denotes medium (market cap 2-10 billion) and green denotes small (market cap < 2 billion). For our decomposition (a), there is no separation distinguishable by size of company. In comparison, for the Fama and French decomposition (b), there appears a gradation from large to small companies consistent with a factor of the model being related to size. (This is natural, since one of Fama and French's factors explicitly is the difference between large and small-cap returns). Thus our unsupervised 3-factor decomposition appears quite distinct from Fama and French's hand-created one.

Table F1. Percentage of the Explainable Variance captured by our model compared with the Fama and French factor model. Regression is done on the normalized dataset of 705 stocks without the market mode removed. To capture this, we add the market mode to factors obtained by our decomposition.

Bulk Variation	80.2%
Explainable Variation	19.8%
Factors	Percent of Explainable Variation
Market Mode (MM)	8.0%
2 factors + MM	26.0%
3 factors + MM	36.1%
4 factors + MM	42.8%
5 factors + MM	48.9%
6 factors + MM	55.3%
7 factors + MM	59.4%
8 factors + MM	63.7%
9 factors + MM	68.1%
Fama and French	24.0%

Appendix G. The number *n* of canonical sectors

It is an open problem to determine the effective dimensionality (optimal rank) of a general dataset (matrix). One could select among models of different dimensions using statistical tests such as the r^2 discussed above, or information theory based criteria such as Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), but the choice of the selection criterion is itself generally made on an *ad hoc* basis. Therefore, a direct observation of the comprehensibility of results is often the most reliable criterion. In the data-set used for analysis described here, a factorization with $n > 8$ yielded results where both the emergent time series E_{tf} and weights in W_{fs} showed qualitative signs of overfitting. For example, with $n = 9$ the results were in good agreement with $n = 8$ except for an additional resulting sector involving participation from only 11 seemingly unrelated stocks (table G1 and figure 4). The high-level results of factorization with different values of n may be explored in a number of ways, several of which are described below.

G.1. *Sector changes with dimensionality*

One approach to investigating how the sector decomposition changes with dimension is to produce a flow diagram. To do this, we performed the fit $||E_{t,f} - E_{t,f'}S_{f',f}||_F^2$ with the constraint $\sum_{f'} S_{f',f} = 1$. Hence the sectors for $n = 9$ can be expressed as a linear combination of sectors for $n = 8$, $n = 8$ as a linear combination of $n = 7$, and so forth. The results of these fits are presented in figure 5. The figure represents these relationships though connections between the decompositions for $n = N + 1$ and $n = N$ weighted according to the matrix $S^{(N,N+1)}$. More precisely, we create a node corresponding to each of the 9 sectors whose size is proportional to $\sum_s W_{f,s}$ where $W_{f,s}$ is the weight matrix for the 9 sector decomposition. Hence, the relative node sizes represent the amount of the market particpating in the sector. Multiplying this vector by $S^{(8,9)}$ gives the approximate size for each node in $n = 8$. Multiplying this vector by $S^{(7,8)}$ gives the approximate size for each node in $n = 7$, and so on. In this way, we generate a Sankey diagram whose node sizes correspond roughly to the amount of the market in the sector and whose connections depict how strongly the sectors for decompositions with different n overlap. In the image, we see that the $n = 9$ decomposition gives the 8 sector version with an additional small sector whose companies were listed in table G1. We also see that for $n = 7$ *c-finance* and *c-real estate* merge. At $n = 6$, *c-industrial* and *c-cyclical* merge. For $n = 5$, the new sector containing *c-industrial* and *c-cyclical* merges with *c-non-cyclical*. For $n = 4$, *c-utility* and *c-energy* merge. Finally, for $n = 3$ and $n = 2$, no clear pattern emerges given this image alone.

G.2. *Two and three sector decompositions*

We further explore the two and three sector decompositions by examining their constituent companies and looking at pie charts describing the relationship between our 8 sector decomposition and those with $n = 2$ and $n = 3$ respectively. Recall that each archetype is constrained to be a linear combination of companies, or in other words to lie in the convex hull of the data. Using this information, we list the 20 companies which contribute the most to each sector in the two and three factor decompositions (tables G2–G4). For the two sector decomposition, we find the sectors divide roughly into *c-assets* (e.g. financial and real estate companies) and *c-goods* (e.g. companies which provide goods and services). For $n = 3$,

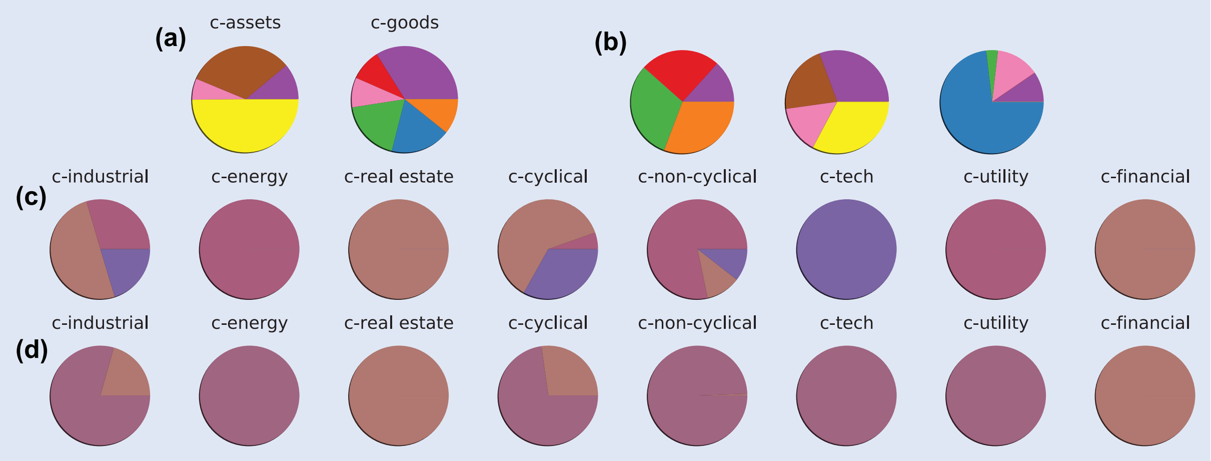

Figure G1. Pie charts depicting sectors as linear combinations of other sector decompositions having a different value of the dimensionality n. (a) Two sector decomposition with respect to the eight sector version (b) Three with respect to eight (c) eight with respect to two (d) eight with respect to three. For (a) and (b), the color scheme is the same as used throughout for the eight sector decomposition. For (c) and (d) colors correspond to those in figure 4 for the two and three sector nodes. Through these charts it is evident that the two sector decompositions corresponds to an *c-assets* sector containing *c-finance* and *c-real estate*. and a *c-goods* sector containing companies which provide goods and services. In (c) and (d) we see *c-industrial*, *c-cyclical* and *c-non-cyclical* which merge by $n = 5$ split between the two and three factor decompositions respectively, consistent with figure 4.

Table G1. Companies which form a new sector when the dimensionality of the decomposition is increased from $n = 8$ to $n = 9$. The labels given are those indicated by Scottrade®.

Ticker	Company name	Label
EQT	EQT Corporation	Energy
RDN	Radian Group Inc.	Financials
STT	State Street Corporation	Financials
LH	Laboratory Corp. of America Holdings	Healthcare
UHS	Universal Health Services Inc.	Healthcare
STZ	Constellation Brands Inc.	Non-Cyclicals
CNL	Cleco Corporation	Utilities
OKE	ONEOK Inc.	Utilities
CAKE	The Cheesecake Factory Incorporated	Cyclicals
EFX	Equifax Inc.	Industrials
ESRX	Express Scripts Holding Company	Non-Cyclicals

Table G2. Top 20 contributing companies to each sector in the two sector decomposition. Ranking is determined by the martix $C_{s,f}$ which describes each sector as a linear combination of stocks. Labels are those given by Scottrade® and percentage describes the percentage of the sector attributable to the company.

C-assets	Label	Percent	Full name	C-goods	Label	Percent	Full name
DDR	real estate	1.77%	DDR Corp.	HON	tech	0.53%	Honeywell International Inc.
ONB	financial	1.7%	Old National Bankcorp.	TMO	health	0.51%	Thermo Fisher Scientific Inc.
BRE	real estate	1.66%	Brookfield Real Estate Serv.	NAV	cyclical	0.49%	Navistar International Corp.
PEI	real estate	1.54%	Pennsylvania RIT	CSL	basic	0.47%	Carlisle Companies Inc.
FMBI	financial	1.5%	First Midwest Bancorp. Inc.	IRF	tech	0.47%	International Rectifier Corp.
PRK	financial	1.5%	Park National Corp.	APD	basic	0.46%	Air Products & Chemicals Inc.
BAC	financial	1.42%	Bank of America Corp.	PCP	basic	0.43%	Precision Castparts Corp.
STI	financial	1.41%	SunTrust Banks Inc.	OMC	misc services	0.43%	Omnicom Group Inc.
DRE	real estate	1.29%	Duke Realty Corp.	MXIM	tech	0.43%	Maxim Integrated Products, Inc.
UBSI	financial	1.28%	United Bankshares Inc.	TFX	health	0.41%	Teleflex Inc.
CPT	real estate	1.28%	Camden Property Trust	NSC	transport	0.41%	Norfolk Southern Corp.
PPS	real estate	1.28%	Post Properties Inc.	NBL	energy	0.4%	Noble Energy Inc.
WABC	financial	1.26%	Westamerica Bancorp.	SM	energy	0.4%	SM Energy Company
FMER	financial	1.26%	FirstMerit Corp.	WMT	retail	0.39%	Wal-Mart Stores Inc.
CNA	financial	1.26%	CNA Financial Corp.	CR	basic	0.38%	Crane Co.
VLY	financial	1.25%	Valley National Bancorp.	ADI	tech	0.38%	Analog Devices Inc.
MTB	financial	1.24%	M&T Bankcorp.	ITW	cyclical	0.38%	Illinois Tool Works Inc.
WRI	real estate	1.23%	Weingarten Realty Investors	PPG	basic	0.38%	PPG Industries Inc.
BDN	real estate	1.21%	Brandywine Realty Trust	BA	capital	0.38%	The Boeing Company
ZION	financial	1.2%	Zions Bancorp.	AME	tech	0.38%	Ametek Inc.
Total		27.54%		Total		8.53%	

Table G3. Top 20 contributing companies to each sector in the three sector decomposition. Ranking is determined by the martix $C_{s,f}$ which describes each sector as a linear combination of stocks. Labels are those given by Scottrade® and percentage describes the percentage of the sector attributable to the company.

Sector 1	Label	Percent	Sector 2	Label	Percent	Sector 3	Label	Percent
XOM	energy	1.29%	BRE	real estate	2.16%	IRF	tech	1.29%
HP	energy	1.22%	PEI	real estate	2.08%	EMC	tech	1.22%
CVX	energy	1.21%	BWS	retail	1.99%	ADI	tech	1.21%
ETR	utility	1.2%	CNA	financial	1.79%	CSCO	tech	1.2%
APD	basic	1.2%	ONB	financial	1.73%	TXN	tech	1.2%
OXY	energy	1.19%	DDR	real estate	1.63%	BMC	tech	1.19%
NFG	utility	1.18%	PRK	financial	1.59%	SNPS	tech	1.18%
PX	basic	1.17%	CBSH	financial	1.59%	PLXS	tech	1.17%
CL	non-cyclical	1.16%	BC	cyclical	1.56%	CPWR	tech	1.16%
NBL	energy	1.15%	FMER	financial	1.55%	AVT	tech	1.15%
OII	energy	1.11%	RDN	financial	1.54%	SWKS	tech	1.11%
LNT	utility	1.11%	MAS	capital	1.54%	HPQ	tech	1.11%
D	utility	1.08%	DDS	retail	1.47%	PMCS	tech	1.08%
DTE	utility	1.07%	FMBI	financial	1.47%	MXIM	tech	1.07%
SCG	utility	1.06%	ALK	transport	1.46%	ARW	tech	1.06%
WEC	utility	1.04%	WABC	financial	1.43%	TER	tech	1.04%
APA	energy	0.99%	PCH	real estate	1.42%	ATML	tech	0.99%
BAX	health	0.98%	VLY	financial	1.41%	MCHP	tech	0.98%
MUR	energy	0.98%	BAC	financial	1.41%	LRCX	tech	0.98%
CPB	non-cyclical	0.98%	STI	financial	1.37%	CGNX	tech	0.98%
Total		22.38%	Total		19.14%	Total		32.18%

Table G4. Top 20 contributing companies to each sector in the three sector decomposition. Ranking is determined by the martix $C_{s,f}$ which describes each sector as a linear combination of stocks.

Sector 1	Full name	Sector 2	Full name	Sector 3	Full name
XOM	Exxon Mobil Corp.	BRE	Brookfield Real Estate Serv.	IRF	International Rectifier Corp.
HP	Helmerich & Payne Inc.	PEI	Pennsylvania RIT	EMC	EMC Corp.
CVX	Chevron Corp.	BWS	Brown Shoe Co. Inc.	ADI	Analog Devices Inc.
ETR	Entergy Corp.	CNA	CNA Financial Corp.	CSCO	Cisco Systems Inc.
APD	Air Products & Chemicals Inc.	ONB	Old National Bancorp.	TXN	Texas Instruments Inc.
OXY	Occidental Petroleum	DDR	DDR Corp.	BMC	BMC Software Inc.
NFG	National Fuel Gas Company	PRK	Park National Corp.	SNPS	Synopsys Inc.
PX	Praxair Inc.	CBSH	Commerce Bancshares Inc.	PLXS	Plexus Corp.
CL	Colgate-Palmolive Co.	BC	Brunswick Corp.	CPWR	Compuware Corp.
NBL	Noble Energy Inc.	FMER	FirstMerit Corp.	AVT	Avnet Inc.
OII	Oceaneering International Inc.	RDN	Radian Group Inc.	SWKS	Skyworks Solutions Inc.
LNT	Alliant ENergy Corp.	MAS	Masco Corp.	HPQ	Hewlett-Packard Company
D	Dominion Resources Inc.	DDS	Dillard’s Inc.	PMCS	PMC-Sierra Inc.
DTE	DTE Energy Corp.	FMBI	First Midwest Bancorp. Inc.	MXIM	Maxim Integrated Products Inc.
SCG	SCANA Corp.	ALK	Alaska Air Group Inc.	ARW	Arrow Electronics Inc.
WEC	Wisconsin Energy Corp.	WABC	Westamerica Bancorp.	TER	Teradyne Inc.
APA	Apache Corp.	PCH	Potlatch Corp.	ATML	Atmel Corp.
BAX	Baxter International Inc.	VLY	Valley National Bancorp.	MCHP	Microchip Technology Inc.
MUR	Murphy Oil Corp.	BAC	Bank of America Corp.	LRCX	Lam Research Corp.
CPB	Campbell Soup Company	STI	SunTrust Banks Inc.	CGNX	Cognex Corp.

the division is less clear. Another way to look at the constituents of these sectors is by examining pie chart representations of these decompositions. Again consider the fit $||E_{t,f} - E_{t,f'}S_{f',f}||_F^2$ with the constraint $\sum_{f'} S_{f',f} = 1$. Applying this, we can express the two sector archetypes as linear combinations of the 8 sector archetypes and vice versa. Additionally, we can do the same for the three factor decomposition. The pie charts these fits produce are shown in figure G1. The results are consistent with the sector breakdowns described from examining the constituent companies.

G.3. *Robustness*

In general, a factorization analysis of the returns dataset would be sensitive to number of stocks in the dataset, criteria applied for picking stocks, period over which historical prices are obtained, and frequency at which returns are computed. A robust macroeconomic analysis would therefore require a large number of stocks chosen without sampling bias, with returns calculated over the period of interest and sensitivity checked for frequency of returns calculation. On the other hand, an equity fund manager faces a less daunting task for an analysis that is limited to the universe of her portfolio of stocks: either to find its canonical sectors, or to analyse the exposure of her holdings to the core sectors of the economy.

Appendix H. Canonical sector indices

The matrix C_{sf} in decomposition $R = RCW$ represents how returns R of stocks s must be combined to make canonical sector returns $E_{tf} = R_{ts}C_{sf}$. Since a canonical sector is defined as a combination of stocks, an investment in the sector f can made via buying a basket of constituent stocks s in proportions given by C_{sf} or through an index I_{tf}:

$$I_{tf} = p_{ts'}C_{s'f} \tag{H1}$$

where, p are stocks prices suitably weighted by market cap or other divisor as common practice for common indices (Tagiliani and Guide 2009). An unweighted index of this kind is shown in the bottom row of figure C1 for results corresponding to the analysis described in this paper. Conversely, a pre-defined basket of stocks such as the S&P 500® can be unbundled to find its exposure to the canonical sectors. With an investment strategy employing longs and shorts at the same time in correct proportions, it is conceivable to invest in, for example, the *c-tech* component of S&P 500®.

The desirable features of an index include completeness, objectivity and investability (Pastor *et al*. 2013). The *c-indices* constructed using the ideas outlined here would not only be of value to investors through investment vehicles such as exchange-traded funds, Futures, etc., but also serve as important economic indicators.