Financial Analysts Journal

Financial Analysts Journal

ISSN: 0015-198X (Print) 1938-3312 (Online) Journal homepage: www.tandfonline.com/journals/ufaj20

Thematic Investing: A Risk-Based Perspective

Emmanuel Candès, Trevor Hastie, Ked Hogan, Ronald N. Kahn, Robert Luo & Asher Spector

To cite this article: Emmanuel Candès, Trevor Hastie, Ked Hogan, Ronald N. Kahn, Robert Luo & Asher Spector (2025) Thematic Investing: A Risk-Based Perspective, Financial Analysts Journal, 81:4, 103-120, DOI: 10.1080/0015198X.2025.2526483

To link to this article: https://doi.org/10.1080/0015198X.2025.2526483

Published online: 01 Aug 2025.

Financial Analysts Journal | A Publication of **CFA Institute**
https://doi.org/10.1080/0015198X.2025.2526483

RESEARCH

Thematic Investing: A Risk-Based Perspective

OPEN ACCESS

Emmanuel Candès, Trevor Hastie, Ked Hogan, Ronald N. Kahn, Robert Luo and Asher Spector

Emmanuel Candès is Barnum-Simons Chair in Mathematics and Statistics, Stanford University, Stanford, California. Trevor Hastie is John A. Overdeck Professor of Mathematical Sciences and Professor of Statistics and Biomedical Data Science, Stanford University, Stanford, California. Ked Hogan is a Managing Director, Senior Researcher, BlackRock, San Francisco, California. Ronald N. Kahn is a Managing Director, Global Head of Systematic Investment Research, BlackRock, San Francisco, California. Robert Luo is a Vice-President, Portfolio Manager, and Researcher, BlackRock, New York, New York. Asher Spector is a Graduate Student in the Department of Statistics, Stanford University, Stanford, California. Send correspondence to Ronald N. Kahn at ron.kahn@blackrock.com.

Thematic investing has grown in popularity even without a clear definition. We propose a risk-based definition of a theme and focus on themes that involve significant transient correlations of residual returns. We present a bootstrapping style approach to determine the statistical significance of the average pairwise correlation among stocks in a thematic basket. Analyzing thematic baskets provided by an investment bank, we find evidence of statistically significant correlations. The thematic baskets with statistically significant average pairwise correlation will have risk levels above predictions. Furthermore, they exhibit statistically significant trending. Baskets with insignificant average pairwise correlation do not trend on average.

Keywords: average pairwise correlation; bootstrap test; mosaic permutations; residual returns; risk models; thematic investing

Disclosure: Most of the authors work for BlackRock or consult to BlackRock, and BlackRock offers many thematic products. We undertook this research to better understand thematic investing, not to promote any particular product.

PL Credits: 2.0

Introduction

Thematic investing has become increasingly popular. According to Morningstar (Lamont et al. 2024), worldwide assets and number of thematic funds have nearly doubled over the last five years (ending June 30, 2024), reaching USD 562 billion and 2,776 funds, respectively. However, these figures have declined from their pandemic-era peaks, reflecting notable performance challenges. Over the year from July 1, 2023, to June 30, 2024, only 18% of funds outperformed the Morningstar Global Target Market Exposure Index. This figure drops even further to just 9% over the past 15 years (ending June 30, 2024). Successful thematic investing will require identifying outperforming themes.

Beyond dedicated thematic funds, for example, mutual funds and exchange-traded funds, many asset managers allocate some risk to themes in their funds. Most sell-side firms and many buy-side firms now offer thematic ideas and customized baskets to their clients. Themes offered range from artificial intelligence, cybersecurity, renewable energy, and healthcare innovations to the shutdown and subsequent re-opening of the economy because of the COVID pandemic. Events somewhat like these have occurred in the past, but not for the span of most investors' lives or over the historic data commonly used to model returns. These events are too infrequent for accurate statistical modeling or are absent altogether.

We thank Robert Tibshirani for discussions and input at the beginning of this project. This paper reflects the views of the authors and not necessarily those of their employers.
This work was supported by the National Science Foundation Graduate Research Fellowship Program; Citadel GQS PhD Fellowship; Two Sigma Graduate Fellowship Fund; and BlackRock.

Background

Research studies of thematic investing have accompanied their recent strong growth. Somefun et al. (2023) recognize the lack of consensus on what exactly is a theme and, equally importantly, what is not a theme. According to the authors, "Themes are structural trends expected to significantly impact economies and redefine business models." Importantly, they highlight themes as something unique: "Thematic investing can be seen as an additional dimension in portfolios."

In a similar vein, Morningstar (Choy et al. 2022) states that, " ... themes may pertain to macroeconomic or structural trends that transcend the traditional business cycle." They are also careful to focus on what makes themes different. "As a rule of thumb, we have excluded funds that either track standard sector, industry, or subindustry indexes or closely resemble mainline sector funds from our definition of thematic funds." These definitions are useful in understanding the intuition behind themes.

Thematic investing also relates to the growing field of "narratives" developed by Robert Shiller (2017). This literature argues that oversimplified, easy-to-transmit word phrases in the press can influence human behavior and ultimately impact the economy and the markets over and beyond normal transmission mechanisms of economic shocks. Several empirical studies have followed this path, extracting topics, themes, and narratives from popular news-based corpora and empirically testing for their impact on markets. For example, Bhargava et al. (2023) find a link between a carefully selected set of 73 themes and future market returns. Blanqué et al. (2022) quantify narratives with natural language processing techniques to create volume and sentiment metrics over the Global Database of Events, Language, and Tone, which improve upon purely macroeconomic forecasts of S&P 500 returns.

We make three contributions to the above-mentioned literature. We define themes in the context of a risk model framework. We test our definition based on a statistical methodology. We link our definition of themes to a novel dataset of broker-generated theme topics and baskets. Our focus is on analyzing thematic baskets ex post after release and we show that correlations within baskets persist for at least 60 calendar days. We also demonstrate that our definition allows us to identify actionable implications. We leave for future research whether we can use the insights described here to identify emerging new themes.

Defining Themes.

Before we can identify and analyze investment themes, we need a clear definition of what we mean by a theme.

We often view investment behavior through the lens of a fundamental factor model. These models take the form:

$$\mathbf{r} = \mathbf{X} \cdot \mathbf{b} + \mathbf{u}. \tag{1}$$

This models the vector of excess returns, **r**, based on exposures **X** to common factors **b** plus residual, or idiosyncratic returns, **u**. (We will use the terms *residual* and *idiosyncratic* interchangeably in this paper.) In the case of equities, Equation (1) could decompose an individual stock return into the return to its industry, the return resulting from its exposure to various common style factors (e.g., Value or Size), and finally the return idiosyncratic to that stock. We typically estimate the factor returns, **b**, and hence the idiosyncratic returns, **u**, via cross-sectional regression. We observe the returns and the factor exposures directly but must estimate the factor returns and the idiosyncratic returns.

This framework leads to the asset-by-asset covariance matrix **V** taking the form:

$$\mathbf{V} = \mathbf{X} \cdot \mathbf{F} \cdot \mathbf{X}^{\mathrm{T}} + \Delta. \tag{2}$$

The matrix **F** is the covariance of the factor returns, **b**, and Δ is the diagonal variance matrix of the idiosyncratic returns, **u**. One implication of Equation (2) is that two stocks are correlated if they are exposed to the same factors or if they are exposed to different but correlated factors. The idiosyncratic returns are not correlated across assets. That's why we call them idiosyncratic.

Equations (1) and (2) will help provide a risk modeling perspective into factors and themes. Of course, we also care about the return implications of themes. We will start with this risk modeling perspective and later come back to the return implications.

We have often referred to factors as broad and persistent sources of risk and return. In contrast, we assert that themes are more narrow and transient sources of risk and return. Let's make that more precise using this factor model lens.

By saying that factors are broad and persistent, we mean that if we research a risk model over, say, 10 years of data, these factors often help us understand the cross-section of asset returns. If we run daily cross-sectional regressions as in Equation (1), testing out a particular choice of factors, **X**, we

should observe that the absolute value of the *t* statistics for each factor will be greater than two in much more than 5% of the periods. An additional implicit assumption here is that if a factor significantly explains cross-sectional returns in much more than 5% of days over the past 10 years, commercial risk models have identified them. These models have been around and in broad use since the 1970s.

Connecting the 5% cutoff for statistical significance with *t* statistics greater than two requires assuming that errors are normally distributed. When one of the authors worked at Barra in the 1980s and 1990s building risk models, they would add random factors to such testing. Empirically, using monthly equity return data, they would observe significant *t* statistics for these random factors in roughly 6% or 7% of all periods, a bit higher than the normal distribution result. Broad and persistent factors should outperform such random factors and help explain cross-sectional returns.

When we say that a factor is persistent, we are stating that, beyond what we have already said, the mean factor return over time is positive and statistically significant. Over much of the past 10 years, some factors, including Value in particular, have not exhibited consistent positive returns (even though their returns were persistent prior to that period). That said, even over that recent period, the Value factor has consistently helped explain cross-sectional returns.

We have asserted that themes, in contrast to broad and persistent factors, are narrower and more transient. How would they alter Equation (1)? Let's say that over a short period of time, we can describe returns as:

$$\mathbf{r} = \mathbf{X} \cdot \mathbf{b} + \mathbf{Y} \cdot \mathbf{g} + \mathbf{u}'. \tag{3}$$

In Equation (3), we have added new components of returns: exposures **Y** to themes **g**. We have also allowed for the idiosyncratic returns to be different once we have accounted for these themes. Note that if we estimate **b**, **g**, and **u'** via cross-sectional regression, our estimates of **b** will change due to the presence of the theme exposures **Y**.

According to Equation (3), themes provide additional sources of correlation across stocks, creating correlations across what we had thought were uncorrelated idiosyncratic returns. What we thought were idiosyncratic returns may in fact be correlated and follow a factor model, at least over some limited period. It is in this sense that we refer to such themes as *coherent themes*, in loose analogy to a laser providing coherent (identical frequency and phase) light. Stocks in a coherent theme basket have correlated idiosyncratic returns like the waves in a laser light source have related—and in fact identical—frequency and phase.

Note that Equation (3) implies risk implications even for investors who do not explicitly include themes in their approach.

How do we distinguish themes from broad and persistent factors then? We can distinguish them in at least two ways. First, consider the persistent versus transient description. Factors significantly explain cross-sectional returns in much more than 5% of all periods over long-term intervals. Themes significantly explain cross-sectional returns only over shorter periods of time. Without being very precise, factors explain cross-sectional returns over years and decades, while themes explain cross-sectional returns over weeks and months.

Second, consider the broad versus narrow description. Some persistent risk factors, for example, industry factors for equities, can be narrow. There may be only a few stocks in some industries. That said, factor investing focuses mainly on style risk factors like Value, Momentum, Size, or Quality. We can calculate an exposure to style factors for every single stock. In contrast, a theme may only apply to a small subset of our investment universe.

In the themes view of the world, we describe returns as in Equation (3). Over time, the themes driving returns will change. If the same theme explains cross-sectional returns over long periods of time, we should more properly describe it as a factor.

Testing for Coherent Themes: Test Statistics

Let us say we start with a risk model (Equations (1) and (2)) and ask whether, for any two assets *n* and *m*, $Corr(u_m, u_n) = 0$ for $m \neq n$. Empirically these will not be exactly zero. We need to ask whether deviations from zero are statistically significant. How do we do that? Our approach is to develop a bootstrap-style method that enables us to test various hypotheses with respect to these types of questions. This will allow us to formalize a test over

any given horizon, without ad hoc distributional assumptions, for any possible test statistic.[1] This paper will focus on the average idiosyncratic return pairwise correlation *(APC or* $\overline{\rho}$*)*. There are, of course, many other ways to detect additional correlations, and we list some additional possible statistics in Appendix A. Here, we focus on the average pairwise correlation in our analysis of theme baskets.

For simplicity, we will mainly drop the explicit reference to residual returns, as all references to correlations in this paper from now on refer to residual return correlations. We compute the $N \times N$ correlation matrix of the risk model residuals, $\{u_n\}$ over a period from T_1 to T_2. Letting ρ_{nm} be the entries of this correlation matrix, our test statistic is:

$$\overline{\rho} = \frac{2}{N(N-1)} \sum_{n>m} \rho_{nm}. \tag{4}$$

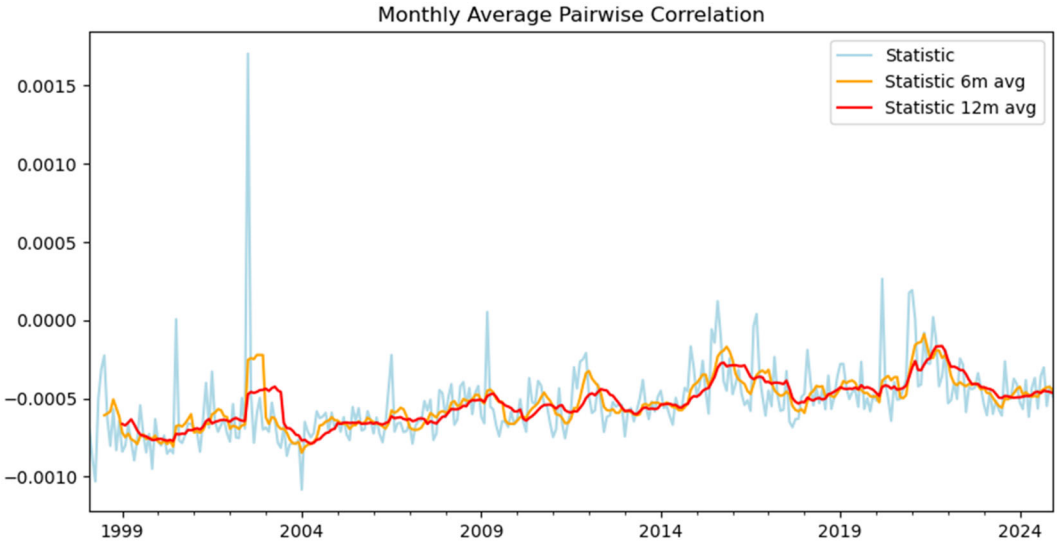
A large value may indicate that the residuals are correlated. Over the full risk model estimation universe, we find that this average pairwise correlation is very small (see Figure 1 below). We can calculate this test statistic over just the set of assets in a proposed theme basket. There we can find statistically significant average correlations.

Empirical Results: Broad Estimation Universe

We estimate our factor-based model (Equation (1)) by running cross-sectional ordinary least square regressions using daily excess returns for liquid large-cap US equities and the Barra GEM3 (Global Equity Model, third edition) industry and style factors from 1999 to 2024.[2] Our universe contains about 1,100 stocks on average. We calculate the daily average residual return correlation (APC) as the average of the approximately 600,000 pairwise residual correlations computed using a trailing 30-day window. Figure 1 plots the result.

We expected a very small average pairwise correlation across the entire estimation universe and Figure 1 confirms that intuition. We estimate the residual returns via regression, and the regression mathematics force the average residual return to zero. (In ordinary least squares, the average is zero. In generalized least squares, the weighted average is zero.) The constraint that the average residual return is zero, or equivalently that the (possibly weighted) sum of all the residuals is zero, induces a slight negative average correlation across the residual returns which we can see in the exhibit. We only expect to see significant pairwise correlations among much smaller groups of stocks (20 to 100

Figure 1. Time-Series of the Average Pairwise Correlation of Residual Returns Across Estimation Universe



Source: BlackRock Systematic.

stocks rather than 1,100 stocks). While there is a small but statistically significant trend in average pairwise correlations over time, it does not appear to be economically meaningful.

Statistical Tests

How can we formally test for significant correlations in residual returns? There are several challenges that arise due to applying bootstrap methods to residuals estimated via regression:

- The regression induces correlations among the estimated residuals, for example, because the estimated (possibly weighted) mean residual is zero. The regression also biases downward the estimated variances of the residuals as described in Appendix B.
- The exposure matrix **X** in Equation (1) varies over time.
- The residual variances change over time.
- The residuals are orthogonal to the factor exposures, and we want to retain that property as we generate bootstrap samples.

These challenges inspired the development of an improved *mosaic permutation test* that we will call the mosaic + bootstrap method (Spector et al. 2024). Here we will describe the basic approach which is subject to the objections we have just mentioned. Our analysis, though, relies on the mosaic + bootstrap method and is statistically rigorous.

Motivation: The Basic Idea.

We start with a *T x N* matrix of residuals from Equation (1) run daily for *N* stocks over *T* days:

$$\mathbf{H} = \begin{bmatrix} u_{1,1} & \cdots & u_{1,N} \\ \vdots & \ddots & \vdots \\ u_{T,1} & \cdots & u_{T,N} \end{bmatrix}. \tag{5}$$

The rows of **H** are the residual returns of each of the *N* assets on days 1, ... *T*. The columns of **H** are the time series of the asset residuals. The first complication is that the variances of residuals from a regression are biased downward.

From the matrix **H**, we create a new matrix $\mathbf{H}^*$ by independently shuffling each column of **H**. This scrambles the time series of each asset's residuals. This randomization creates a new matrix of quasi residuals that are uncorrelated with each other. Imagine we are trying to understand if the residual returns to, for example, IBM and Alphabet are correlated over a particular month.[3]

We do not expect to see any such correlations if we scramble the order of those two stocks' residuals over that month so we can use empirical correlations after shuffling as calibration. Importantly, if individual residual returns are independent draws from a possibly unknown distribution, randomizing the columns still yields returns with the same distribution (the mean, variance, skewness, kurtosis, and so on do not change). This is, of course, assuming that the distributions of idiosyncratic returns do not change over time. Unfortunately, our risk model data typically violate that assumption even over one-to-two-month periods.

This scrambling of the residuals means they are no longer uncorrelated with the factor exposures, **X**. We can restore this property by first backing out a new set of total returns based on the scrambled residuals, denoted as $\mathbf{u}^*$, and original regression parameters:

$$\mathbf{r}_t^* = \mathbf{X}_t \cdot \mathbf{b}_t + \mathbf{u}_t^*. \tag{6}$$

For clarity, we include the time subscript in Equation (6).

Starting now with Equation (6), we can reestimate the model to form new residuals with new estimated factor returns:

$$\mathbf{r}_t^* = \mathbf{X}_t \cdot \mathbf{b}_t^* + \mathbf{u}_t^{**}. \tag{7}$$

In Equation (7), we are running new cross-sectional regressions using the returns, $\mathbf{r}^*$ we generated in Equation (6) and our same factor exposure matrix, **X**. This leads to new estimated factor returns, $\mathbf{b}^*$ and new residual returns, $\mathbf{u}^{**}$. Those new simulated residuals maintain the same marginal distributions of the original data under the null that they are uncorrelated with each other and traditional industry and style factors.

We can repeat the steps outlined by Equations (5)–(7) above multiple times to create a bootstrapped sample null distribution of outcomes. For example, for each daily average pairwise correlation, we can compute its *z* score by comparing the observed value to the mean and standard deviation of the null distribution generated through the bootstrap methodology:

$$z_t = \frac{\overline{\rho}_t - Mean_i(\overline{\rho}_{i,t}^*)}{StDev_i(\overline{\rho}_{i,t}^*)}. \tag{8}$$

In Equation (8), the subscript *i* refers to a bootstrap replicate. We can interpret these *z* scores in the

usual sense, that is, statistically significant if greater than two in magnitude.

The bootstrap approach breaks the correlations in the residuals but does not deal with the reduced variance. Also, in addition to reduced variance, the residuals that are randomized from the initial regression have complicated (and artificial) structure that is created by the regression (see Appendix B). Recently Spector et al. (2024) showed that this naive bootstrap approach can lead to overestimation of the *z* scores. They propose a mosaic + bootstrap variation of this method to mitigate these problems. This approach partitions the dataset into a number of non-overlapping panels (both across stocks and across time) and performs the same residualization via regression *separately* within each panel.[4] We compute average correlation using these residuals and apply the bootstrap scrambling to blocks of time within panels of stocks, breaking the pairwise correlations between residuals in different blocks, and recomputing the average pairwise correlations. Because we run the regressions separately *within* blocks of stocks, there are no artifacts created *between* blocks. The blocking in time also preserves autocorrelation in the residuals. This approach largely eliminates the complications mentioned above and is what we have used in this paper.

We have researched this approach to analyzing statistically significant residual return correlations due to our interest in thematic investing and our hypothesis connecting themes to those correlations. We should point out though that this approach has other applications, for example, in risk model testing or other fields that utilize factor models.

From Statistics to the Economics of Themes

We sometimes motivate fundamental risk models by stock characteristics that appeal to some level of economic intuition, namely industries and investment styles.[5] Similarly, we would like to attach some economic intuition to coherent themes identified by significant pairwise correlations of idiosyncratic returns. Focusing on their transient nature suggests a sudden unexpected shock or change in events that captures investor attention. This motivates us to consider popular investment topics published by investment banks. Investment banks can access their substantial pool of analysts to identify topics that resonate with their extensive client base and identify specific companies most related to a particular topic.

For this paper, we will investigate theme portfolios provided by Goldman Sachs but note that most major investment banks and asset managers offer such portfolios.

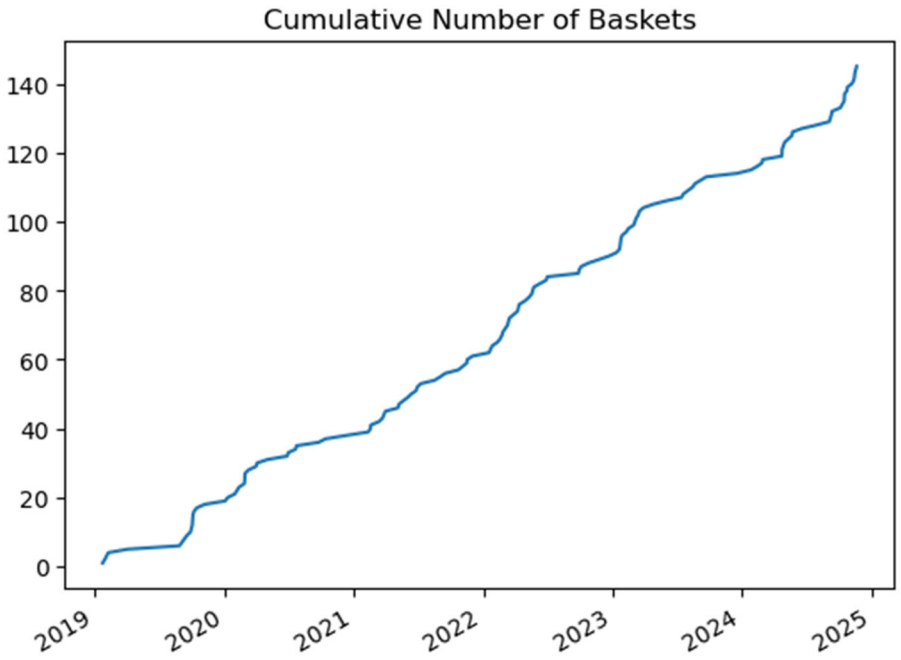
Goldman Sachs has a dedicated custom baskets team responsible for thematic topics. That team determines the stocks most related to a thematic topic jointly with Goldman Sachs sector specialists. The custom baskets team can update and remove thematic baskets at their discretion. Goldman Sachs provides an extensive list of themes, including ones directly related to sources of factor risk such as value, duration, and specific industries. Fortunately, they also provide portfolios they claim capture *macro themes* less likely defined by either industry or style, narrowing the list to 145 baskets of US liquid large-cap stocks as of March 2025.[6] Baskets date back to 2010; however, Figure 2 illustrates the growth in the number of thematic baskets since 2020. Since it displays the cumulative number of baskets, the steep but fairly uniform slope from mid-2019 through 2024 implies roughly equal numbers of new theme baskets (around 35) each year over that period.

To further describe these theme baskets, Figure 3 provides some summary statistics regarding numbers of holdings, residual returns, and residual risk.

For each period, we chose all the macro thematic baskets released in that year and in the prior year (e.g., 2019 and 2020 for the 2020 period). Figure 3 shows the distribution of holdings for those baskets as well as the distribution of annualized residual return and risk for that year. For the 2020 column, we display statistics regarding residual return and risk over 2020. One thing to note is that the average residual return in each period is negative but not statistically significant, roughly consistent with theme fund performance statistics described in the introduction. This argues against an investment strategy that simply invests in all the thematic baskets. We will need to be more discerning than that.

As an alternative to analyzing these Goldman Sachs thematic baskets, we could have applied text analysis to business news to identify topical themes. See Bybee et al. (2024) for a description of this approach. We expect that thematic baskets

Figure 2. Growth in Numbers of Goldman Sachs Theme Baskets



Source: Goldman Sachs.

produced this way will exhibit similar properties to the baskets we analyze in this paper; however, analyzing alternative sets of thematic baskets went beyond the scope of this research.

Our analysis is explicitly ex post in nature. We start with thematic baskets identified by Goldman Sachs and then analyze average residual return pairwise correlations within them. As we will show based on results in Figure 7, we believe investors can effectively utilize this analysis to identify compelling theme baskets. We plan future research on whether we can ex ante analyze individual stock correlations to identify emerging themes.

Returning to the Goldman Sachs theme baskets, we hypothesize that these topics fit our definition of themes, with portfolios with a high degree of idiosyncratic average pairwise correlation. Often the name of the theme itself suggests this to be the case, like re-opening of the economy after COVID. It is also possible that some Goldman Sachs macro theme portfolios do not exhibit correlated residual returns. To illustrate this, Figure 4 displays the top and bottom macro thematic baskets ranked by their residual return average pairwise correlation for the first half of 2020, as well as the correlation z scores from the mosaic + bootstrap methodology. Note that many of these thematic baskets exhibit average pairwise correlations much higher than we see across the full investment universe as Figure 1 displays.

Not only do we see significant residual correlations within theme baskets, we do not see such correlations across baskets. To analyze this, we calculated pairwise correlations across all stocks over a 60-day window ending in June 2023 (1,523,990 observations) and regressed them against four dummy variables to indicate the following:

- Neither stock is in any basket
- Both have at least one basket in common
- Both are in baskets but different baskets
- Only one member of the pair is in any basket

We see no significant correlation if neither stock belongs to any basket, a significantly positive correlation if both belong to the same basket, and then statistically but not economically significant correlations in the other two cases, with estimated correlations of −0.0013, roughly consistent with

Figure 3. Summary Statistics for Goldman Sachs Theme Baskets

		Period	
		2020	2023
Number of Holdings	Mean	64.4	71.0
	Standard Deviation	44.0	42.5
Basket Residual Return	Mean	-0.0088	-0.0264
	Standard Deviation	0.1185	0.0907
Basket Residual Risk	Mean	0.0746	0.0601
	Standard Deviation	0.0416	0.0419
Number of Baskets		50	50

Source: Goldman Sachs and BlackRock Systematic, March 2025.
Note: These statistics will change over time.

Figure 1.[7] We also examined random baskets rather than theme baskets and observed z scores centered on zero and ranged from about −2 to 2.

We can see from Figure 4 that COVID was a dominant news narrative in 2020. Not surprisingly, brokers launched several COVID-related thematic baskets, first with respect to "Stay-at-Home" and later "Re-opening" in anticipation of vaccines.

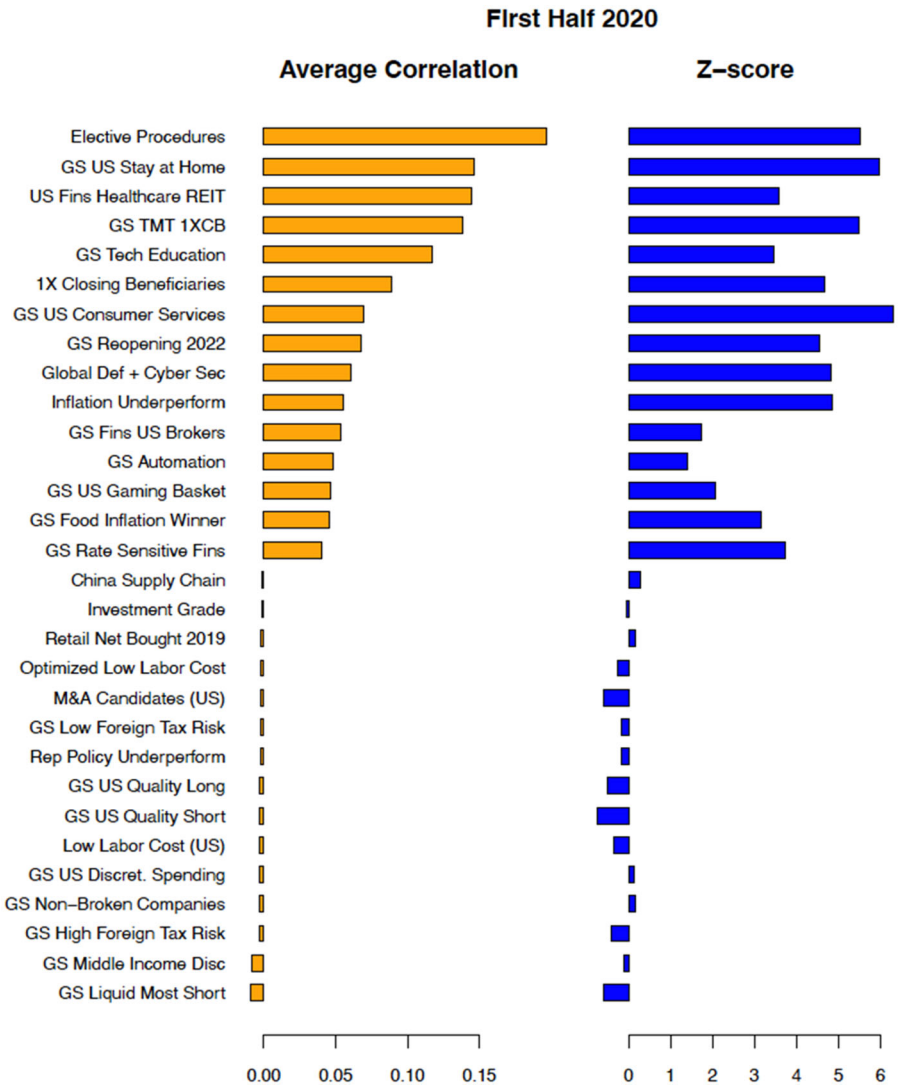
These themes score near the top of Figure 4 as sources of idiosyncratic correlation. We would classify these as coherent themes, especially if the correlations dissipate over time. Other broker themes during this period do not meet our definition of coherent themes, including popular themes related to variants of traditional styles. Arguably, the risk model already captures many of these. Interestingly, other baskets like the "GS China Supply Chain" basket, which seem like reasonable candidates for being coherent themes, do not exhibit significant return correlations at least in this particular time period and using our test statistic. Different test statistics may identify somewhat different coherent theme baskets.

To illustrate the importance of timing and the emergence of a coherent theme, let us focus on two themes: Stay-at-Home and Re-opening. The Stay-at-Home theme attempts to capture the shifting consumer expenditures to home office materials, furniture, and computer equipment that enabled us to work without being in the office. The impact of Stay-at-Home on consumer and investor behavior was remarkable. For example, Wayfair,[8] a home furnishings online retailer, increased in value 10-fold from March 18, 2020, to August 28, 2020. Importantly, the collection of stocks in the Stay-at-Home basket belonged to a diverse set of industries with exposures to many different traditional factors. This is a perfect candidate for our definition of coherent themes. Similarly, the Re-opening basket anticipated recovery in several diverse industries like transportation, leisure, and brick and mortar retail.

Given the evolving information about COVID, the Stay-at-Home and Re-opening coherent themes played out in interestingly different ways. For each, Figure 5 plots the average pairwise correlation of the theme basket from June 2019 through December 2022. We estimate these correlations based on daily residual returns over rolling six-month windows. For the Stay-at-Home theme, the basket constituents showed little or no correlation before the onset of COVID lockdown in March of 2020, dramatically increased, then return to normal within the course of roughly six months. In contrast, the Re-opening theme correlations ebbed and flowed, but consistently remained elevated over the period.

We can assess the statistical significance of the time variation of each Goldman Sachs theme basket's average residual return pairwise correlation with the same mosaic + bootstrap methodology we described above but focusing on the residual correlation matrix of only the members of their

Figure 4. Top and Bottom Average Pairwise Correlation of Residual Returns of Stocks in Goldman Sachs Thematic Baskets



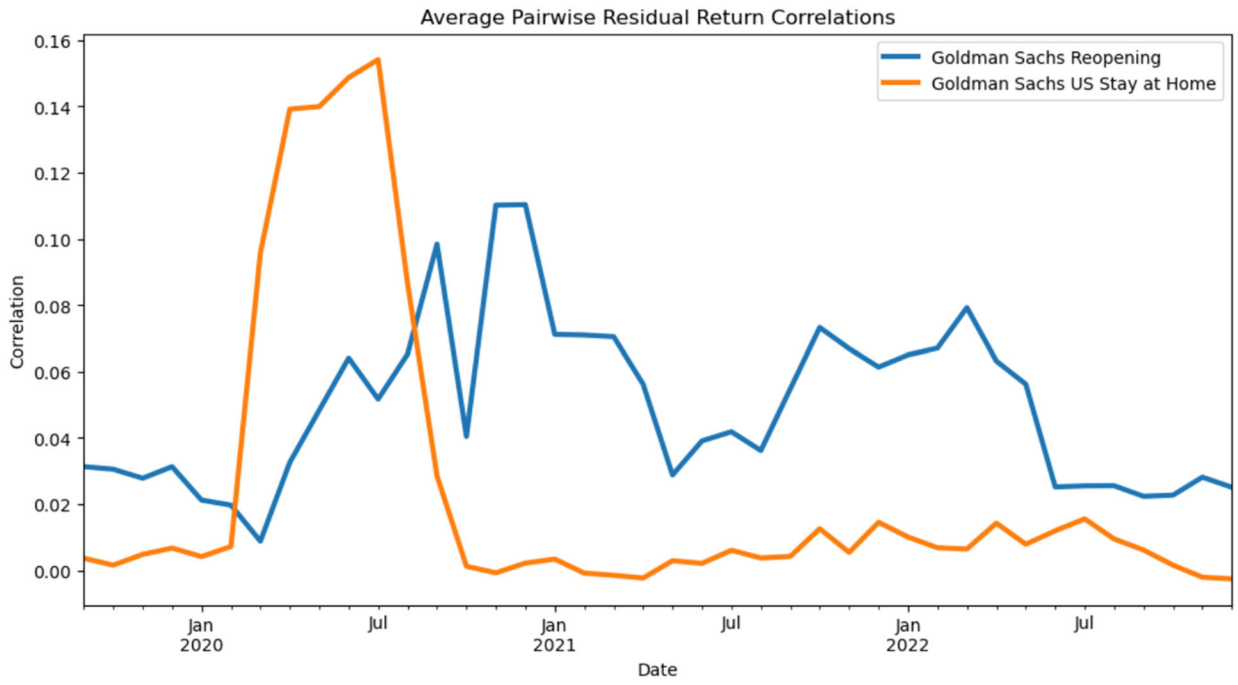
Source: Goldman Sachs and BlackRock Systematic.

respective baskets. Figure 6 plots the null distributions (blue) relative to the basket's average residual return pairwise correlation (red) for the Stay-at-Home basket for six different points over the COVID window. (The charts in Figure 6 display correlations over different six-month periods and so are calculated a bit inconsistently with the results shown in Figure 5.)

As the narrative of COVID evolved over the year, so did the significance of the average pairwise correlation for the "Stay-at-Home" theme basket, rising sharply at first, retracting, and re-surfacing as new variants emerged.

Themes appear regularly over time, not just during the pandemic. Figure 7 applies our analysis to all the Goldman Sachs theme baskets produced from January 2020 through December 2024, where we report the name, Bloomberg ticker, number of stocks, and its average residual return pairwise correlation z scores (before and after release)

Figure 5. Time-Series Plots of Average Residual Return Pairwise Correlations for the "Stay-at-Home" and "Re-opening" Theme Baskets



Source: Goldman Sachs and BlackRock Systematic.

computed using the mosaic + bootstrapping methodology discussed above. The bootstrap in this instance involves just the stocks of a given broker basket, using daily data over the 60 calendar days before and after the release date. Note that Figure 7 wraps around for display purposes: It displays the theme baskets with smaller pre-release *z* scores on the right-hand side.

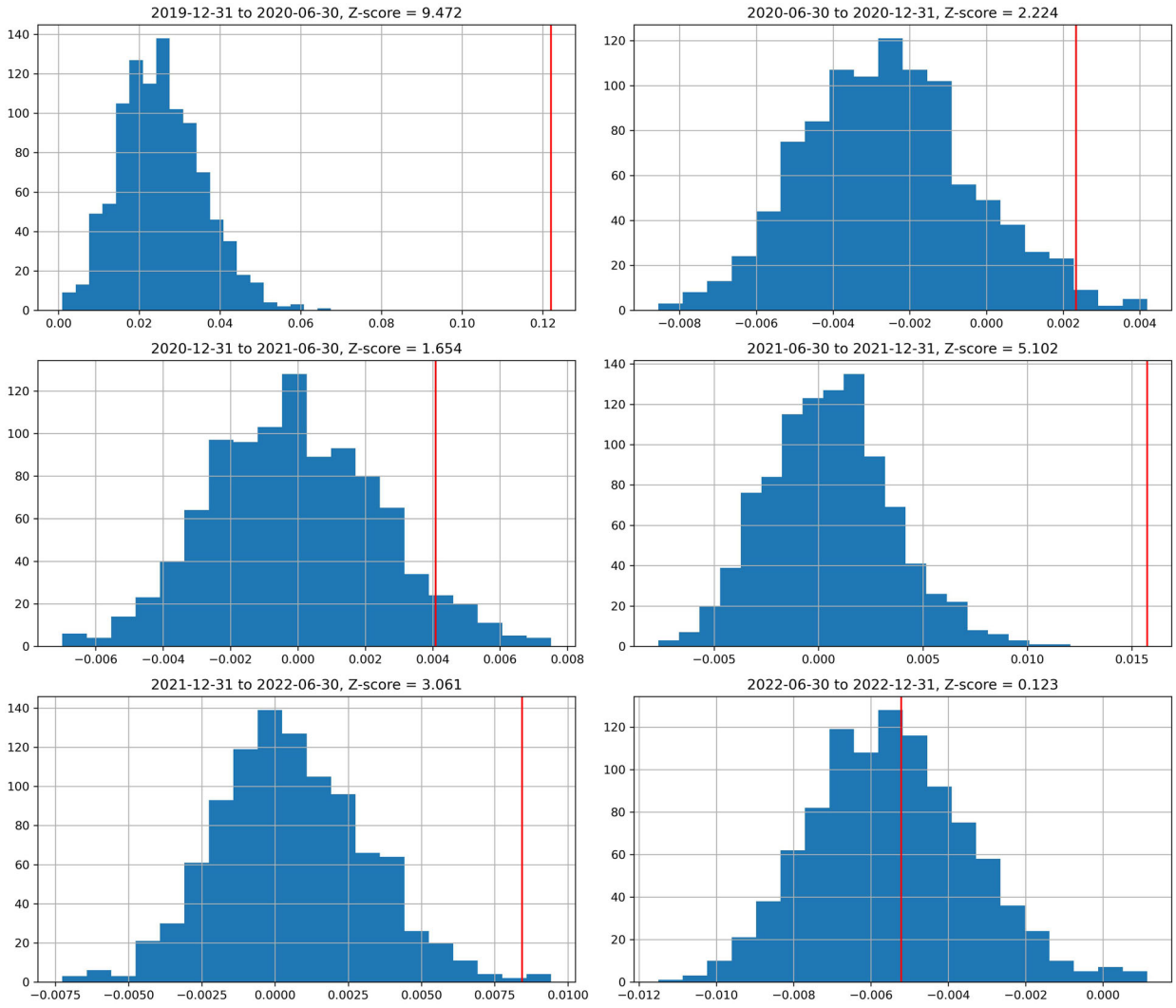
Designating a *z* score of two or more as statistically significant, we find that 61 of the 145 thematic baskets (42%) have statistically significant average residual pairwise correlation over the 60 calendar days prior to basket release. Similarly, we find that 60 of the 145 thematic baskets (41%) have statistically significant average residual pairwise correlation over the 60 calendar days after release. The statistically significant correlations after release implies that we can effectively utilize this ex post analysis. We want to identify baskets with significant average pairwise correlation of residual returns, and these correlations persist for at least some period after release. In Figure 7, we designate all *z* scores greater than two in red. For the period both directly before and after the launch date, many Goldman Sachs themes have statistically significant average residual pairwise correlation consistent with our definition of coherent themes. The correlation between pre- and post-announcement *z* scores is 49%, so there is a high degree of persistence in which baskets have significant correlation.

We also test whether the degree of significance is higher before the launch than after and do not find that to be the case. Importantly, the baskets reflect both what has been happening to basket risks and also what is likely to happen in the future. This suggests that investment strategies based on trends identified before release might persist after release, creating an alpha opportunity which we investigate below.[9]

Alternative Explanations for Observed Residual Correlations

We have shown that many of the Goldman Sachs thematic baskets exhibit statistically significant average pairwise correlation. Are changes in flows

Figure 6. "Stay-at-Home" Basket Average Residual Return Pairwise Correlation Compared to the Null Distribution



Source: Goldman Sachs and BlackRock Systematic.

or media attention the source of these correlations? Regarding flows, we regressed post-release average pairwise correlation *z* scores on percentage increase in flows, aggregating buyer- and seller-initiated flows from Hazeltree at the stock level to the thematic basket level. Hazeltree calculate signed order flow by multiplying the volume of each trade in the market by +1 if the trade ocurred closer to the ask than the bid and by −1 if the the trade occurred closer to the bid. The higher the aggregated signed order flow, the greater the pressure on prices to rise to meet buyers' demand.[10] If this is common to the stocks within a thematic basket, it could induce correlations across those stocks. We also ran similar regressions controlling for pre-release average pairwise correlation *z* scores. We do not see any significant results.

We then used similar regressions to see whether media attention can help explain the average pairwise correlations. Narrative economics suggest that media attention on a given event may create its own impact on human behavior, beyond the simple dissemination of news. We measure media attention by the count of theme-related words that appear in print.[11] We use the Thomson Reuters broker reports corpus, which captures market commentary and research of all major sell-side banks

Figure 7. Goldman Sachs Theme Baskets and Their Average Residual Return Pairwise Correlation *z* Scores

			APWC z-score						APWC z-score		
	Theme	Symbol	pre-release	post release	N		Theme	Symbol	pre-release	post release	N
1	Global Def + Cyber Sec	GSXGDFCS	6.58	9.14	30	74	GS HC Drug Pricing	GSHLCDRP	1.66	1.11	14
2	GS US Solar	GSCBDSOL	6.41	5.17	26	75	US Health Care AI Basket	GSHLCAIP	1.64	0.11	15
3	GS Rate Sensitive Fins	GSXURATE	6.39	1.77	30	76	GS Delivery & Rideshare	GSIDLDEL	1.62	1.26	11
4	GS Liq Opt Sec Growth	GSCMSGRO	6.09	6.72	68	77	GS US Junk Food Exposed	GSCNSJNK	1.60	2.71	10
5	GS US Consumer Goods	GSXUGOOD	5.83	7.00	34	78	GS US Labor Risk	GSXULABR	1.58	1.27	20
6	GS Stagflation Outprfrm.	GSXUSTGL	5.59	1.08	76	79	GS US Health Risk	GSXUMAHA	1.57	2.22	28
7	GS Food Inflation Winner	GSXUFOOD	5.56	3.33	14	80	GS Stable Growers	GSXUSTGR	1.54	1.13	50
8	Industrials Labor Risk	GSXUILAB	5.49	2.47	22	81	GS DeRegulation Benefit.	GSXUDREG	1.54	0.37	60
9	GS Medtech Tools	GSHLCMDT	5.00	4.23	43	82	GS US AI Semis	GSCBSMHX	1.54	2.01	16
10	GS US Liquid Services	GSCBLSVC	4.95	4.56	77	83	SDG 12 Circular Economy	GSCBS12C	1.50	1.72	3
11	GS Software Small Bus.	GSCBSFSB	4.61	0.29	20	84	GS US Profitable Buyback	GSXUHQBB	1.49	0.74	55
12	GS HC Utilization	GSHLCELE	4.58	4.60	13	85	GS Global Tariff Risk	GSXGTRFS	1.45	2.08	35
13	GS US Expensive Software	GSXUSF8X	4.51	3.96	37	86	GS Food Inflation Losers	GSXUFOOL	1.44	4.78	26
14	GS Healthcare Therapeuti	GSHLCTHE	4.38	1.30	20	87	GS EV & Battery Basket	GSXGEVBA	1.44	-0.80	7
15	GS US Soft Landing	GSXUSOFL	4.21	2.90	71	88	GS High Retail Sentiment	GSCBHRSB	1.37	0.96	14
16	GS Stagflation Undrprfrm	GSXUSTGS	4.06	1.92	70	89	LT AI Beneficiaries	GSTHLTAI	1.35	1.03	49
17	GS Fintech Basket	GSFINTEC	4.02	2.88	22	90	Retail Net Bought 2019	GSCBRFTS	1.34	0.48	25
18	GS Fins Regime Benefic.	GSFINWIN	3.70	1.80	33	91	GS Global HLC GLP Risk	GSHLCGLP	1.33	2.75	20
19	GS US Global Health Risk	GSXUPAND	3.64	4.37	33	92	GS US Analog Semis	GSCBASMC	1.27	1.65	11
20	US Fins Healthcare REIT	GSFINHCR	3.49	3.29	8	93	GS US PCs AI Upgrades	GSXUPCAI	1.26	0.72	18
21	High Profit Russell	GSCBHPRS	3.48	1.11	77	94	Global Ukraine Rebuild	GSXGUKRA	1.25	-0.09	18
22	Global Tariff Immune	GSXGTRIM	3.32	2.08	44	95	GS HC IT & Dig Disrupt.	GSHLCITD	1.20	1.86	12
23	GSXUINOV	GSXUINOV	3.28	7.63	32	96	GS Low Tax	GSXXLTAX	1.14	-0.92	86
24	GS US Gov Exposure Risk	GSXUDOGE	3.27	5.25	21	97	High Yield Equities	GSCBHYEQ	1.13	-0.30	57
25	GS US Inflation Comeback	GSXU1970	3.27	3.64	86	98	GS HC Genomics	GSHLCGMC	1.08	3.05	7
26	GS24 Dem Pol Outperform	GS24DEML	3.23	3.28	59	99	GS New Tech	GSXUNEXT	1.06	2.56	31
27	GS Liquid Consumer Goods	GSCBLGOD	3.18	3.03	78	100	GS Secular Growth	GSXUSGRO	1.06	2.61	51
28	Dem Policy Underperform	GSXUDEMS	3.16	1.08	56	101	GS Equities w/ IG Credit	GSCBEQUY	1.06	1.42	40
29	GS Healthcare Prog Policy	GSHLCPRO	3.15	4.40	17	102	AI Data Centers	GSTMTDAT	1.02	1.52	10
30	Vol Adjusted GSXUINOV	GSCBINV1	3.05	2.74	27	103	Global Onshoring & Benef	GSXGSHOR	1.01	0.23	25
31	GS US High Gov. Exposure	GSXUGOVT	3.05	3.34	52	104	GS Brand Loyalty Basket	GSXULYTY	0.97	0.11	26
32	GS AI Software	GSTMTAIS	3.05	2.58	23	105	GS Low Foreign Tax Risk	GSXUFTAL	0.96	0.58	81
33	GS Hi Growth Lo Margin	GSCBHGLM	3.04	5.17	42	106	GS US Renewables	GSXURNEW	0.94	0.98	18
34	US 5y Inflation Exp	GSQI5YIL	3.00	0.89	49	107	GS MF Tax Loss 2022	GSCBMF22	0.88	1.81	35
35	GS US Wage Growth Basket	GSXUWAGE	2.99	2.53	44	108	GS Global AI Basket	GSXGGLAI	0.88	1.47	22
36	GS GIR AI Phase 2	GSCBAIP2	2.97	1.34	72	109	US Nonprofitable Biotech	GSHLCNPB	0.86	2.16	6
37	GS Republican ex Commods	GS24REP2	2.89	5.47	54	110	GS HC Oncology	GSHLCONC	0.84	0.16	8
38	GS Vol Opt Sec Growth	GSCBSGRL	2.84	1.55	34	111	GS US M&A Candidates	GSCBMNAT	0.77	0.63	20
39	GS US Uranium	GSXURANI	2.83	2.28	10	112	GS Over-Earning Cyclical	GSXUOECY	0.75	0.49	36
40	GS Bond Proxies	GSXUBOND	2.79	0.87	49	113	GS US High Margin Risk	GSXUMRSK	0.74	0.31	82
41	US SMB Exposed ex Tech	GSCBSMBB	2.78	-0.52	20	114	GS Financial Progressive	GSTMTPRO	0.64	2.12	26
42	GS High Foreign Tax Risk	GSXUFTAH	2.76	2.78	51	115	High Stable Gross Margin	GSXUSHGM	0.63	0.52	56
43	GS TMT Secular Growth	GSCBOSQT	2.72	1.18	34	116	GS Pensions	GSXUPENS	0.55	0.67	58
44	GS Reflation	GSXUREFL	2.66	1.70	62	117	GS Power Up America	GSENEPOW	0.55	0.84	9
45	GS Remote/Home HC	GSHLCHOM	2.58	0.25	14	118	GSXUHTAX	GSXUHTAX	0.47	-0.46	54
46	GS Lending Sensitive	GSXULEND	2.57	3.13	45	119	GSXULTAX	GSXULTAX	0.42	0.88	38
47	GS Fins De-Regulation	GSFINREG	2.53	4.27	28	120	GS Memes Stocks	GSXUMEME	0.40	3.18	22
48	GS US Commodity Basket	GSXUCOMO	2.53	0.10	42	121	GS High Floating Rt Debt	GSXUHIFL	0.37	0.78	55
49	GS Republican Long exCom	GS24RLXC	2.50	3.21	54	122	Low TMT Stock Based Cmp	GSXUSBCL	0.35	2.30	38
50	GS Small Biz Exposure	GSXUSMBB	2.36	0.27	17	123	GS US Margin Risk ex Fin	GSXUMREF	0.29	0.27	78
51	GS Margin Risk Hedge	GSXUMRSH	2.32	0.49	95	124	GS Hi Growth Hi Margin	GSCBHGHM	0.28	0.43	36
52	Rep Policy Underperform	GSXUREPS	2.24	1.49	41	125	GS GIR AI Phase 3	GSCBAIP3	0.26	1.90	24
53	Banks w/ Steady Deposits	GSFINDBK	2.24	2.12	25	126	GS TMT AI Basket	GSTMTAIP	0.18	1.18	22
54	GS ACA Expansion	GSHLCACA	2.17	4.00	20	127	GS US Stay at Home	GSXUSTAY	0.18	6.85	28
55	GS Biodiversity Basket	GSCBIODV	2.17	0.07	25	128	Low Profitability R2K	GSCBNPR2	0.14	2.31	22
56	GS TMT Cyber Security	GSTMTCYB	2.16	1.60	12	129	GS HLC Qty SMID Biotech	GSHLCSMQ	0.10	0.92	19
57	GS 2024 Tariff Immune	GS24TRIM	2.13	3.17	67	130	GS24 Dem Pol Undrperform	GS24DEMS	0.10	1.91	66
58	GS US Electric Vehicles	GSXUELCV	2.13	-0.36	8	131	Hi Value Hi Margin	GSCBHVHM	0.07	0.94	33
59	GS Democrat ex Renewable	GS24DEM2	2.10	4.48	36	132	GS 2024 Tariff Risk	GS24TRFS	0.05	-0.44	24
60	GS Progressive Policy	GSXUPROG	2.08	1.73	117	133	GS IRA Beneficiaries	GSXUIRAB	-0.10	1.77	36
61	GS Vol Opt Renewables	GSCBRNE1	2.03	4.73	22	134	GS Green Cap Ex	GSCBGCAP	-0.18	-0.26	30
62	GS HY Debt Sensitivity	GSXUDEBT	1.98	1.87	47	135	Banks w/ Volatile Deposi	GSFINLRB	-0.20	0.35	22
63	Theme of Themes Basket	GSXUTHEM	1.96	3.42	18	136	GS US Serial Acquirers	GSXUSACQ	-0.21	0.93	32
64	2020 Dem Policy Outperf	GSXUDEML	1.95	2.71	65	137	GS AI At Risk	GSTMTAIR	-0.38	1.98	20
65	GS Global Renewables	GSXGRNEW	1.90	2.05	17	138	GS US Emergency	GSCBMRGC	-0.41	3.33	12
66	China Supply Chain	GSXUCSUP	1.89	2.81	36	139	GS Infrastructure	GSXUINFS	-0.43	4.02	19
67	Onshore & Beneficiaries	GSXUSHOR	1.88	1.54	37	140	GS Unsustainable Buyback	GSXULQBB	-0.44	0.87	53
68	GS US Offshore	GSXUOFFS	1.84	1.09	51	141	GS Gene Therapy/Editing	GSHLCGEN	-0.57	0.10	5
69	GS Reflation ex Energy	GSXUREF1	1.81	2.68	51	142	GS Oil Input Cost	GSXUOILX	-0.65	2.34	29
70	Non-Buyback	GSCBNREP	1.80	-0.45	58	143	GS Fins Reg Bank M&A	GSFINRMA	-0.73	0.46	21
71	GS Consumer Tariff	GSCNSTAR	1.79	1.06	10	144	GS SUSTAIN E&S Bot Qntle	GSSUSES5	-0.85	1.39	90
72	High TMT Stock Based Cmp	GSXUSBCH	1.71	0.98	41	145	GS HC GLP-1 Exposure	GSHLCBMI	-1.28	-0.33	4
73	GS High 23 vs 22 Margins	GSCBHIMG	1.68	0.89	100						

Source: Goldman Sachs and BlackRock Systematic.

including their economists, strategists, and sector analysts. Typically, there are more than 8,000 broker reports per day.

For each of the Goldman Sachs thematic baskets, we select "key words" based on their titles and count the number of instances they appear each day among all broker reports. We regress the post-release correlation *z* score of each Goldman Sachs thematic basket against the percentage change in media key word counts over the six-month window centered on the release date. We control for the pre-release *z* score in that regression. We do not see any statistically significant impact for the changes in media mentions. If we drop the pre-release *z* scores, the changes in media mentions are statistically significantly related to the post-release *z* scores.

We are defining coherent themes as baskets of stocks that exhibit statistically significant average pairwise correlation over some time. The driver of these correlations could be a transient factor driving a set of stocks that previously were not related. Changes in flows or changes in media attention could contribute to that effect, as could other explanations. In this paper, we are agnostic about the proximate drivers of the increased correlations.

Themes as a Source of Risk and Return

We have characterized coherent themes based on the transient appearance of statistically significant average pairwise correlations of returns residual to risk model factors. This has implications for risk and return. We start with the risk implications and then move on to returns.

Risk model forecasts generally assume uncorrelated residual returns and hence will underestimate forecast risk in the presence of these significant transient correlations. For example, they will underestimate the residual risk of coherent theme baskets by ignoring the correlations. How big an effect is this? To get a rough order of magnitude, assume an equal-weighted theme basket where each of *N* component stocks has annual residual risk of 25%, a typical number. In the absence of any correlations, the theme basket residual risk will be:

$$\psi_p = \frac{25\%}{\sqrt{N}}. \tag{9}$$

Accounting for pairwise correlations, which we assume for this example are the same value, ρ, for every pair of stocks, this becomes:

$$\psi_p \Longrightarrow 25\% \sqrt{\frac{1+\rho\cdot(N-1)}{N}}. \tag{10}$$

Figure 8 shows the dependence of portfolio residual risk on average pairwise correlations for equal-weighted theme baskets of 20 stocks and 50 stocks.

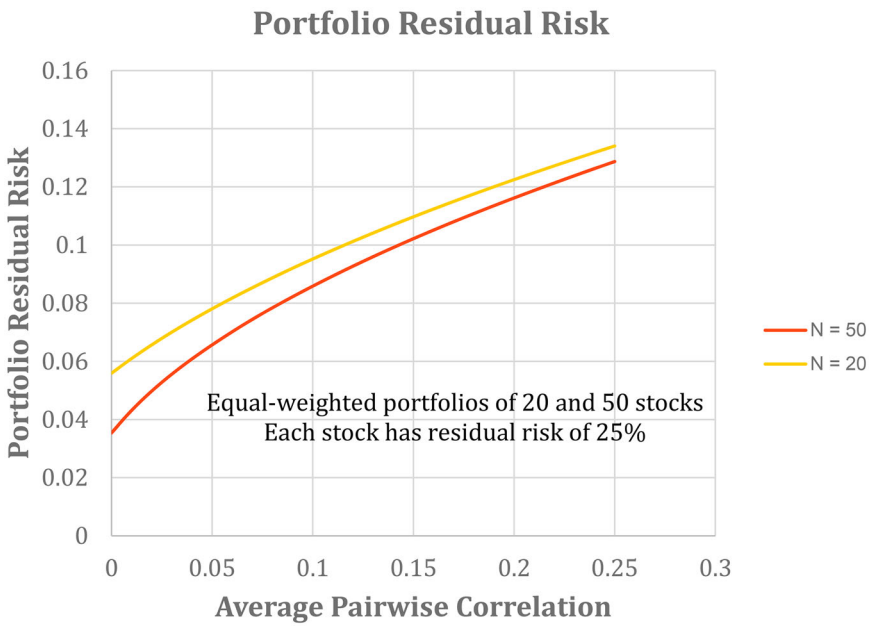
The error is larger the more stocks there are in the portfolio. For example, if average pairwise correlations hit 0.06, the risk model forecasts will undershoot by a factor of two (3.5% when the actual residual risk is 7.0%) for the 50-stock portfolio and a factor of about 1.5 (5.6% when the actual residual risk is 8.2%) for the 20-stock portfolio. As the average pairwise correlation exceeds 0.06, the forecast error will be even larger.

One mitigating issue for risk-forecasting is that the average pairwise correlations are transient. They may be 0.1 today but indistinguishable from zero in three months. Investors may not want to see risk forecasts bounce around that much, especially if their investment horizon is longer than a few months. In that case, they probably want to avoid trading based solely on short-term fluctuations in risk forecasts. It is also the case that risk implications for investors will depend on their exposure to coherent themes.

Figure 7 shows that significant correlations persist for at least 60 days after release, so we could implement improved risk-forecasting using this ex post analysis.

What about the return implications of coherent themes? There is a common view that themes create transitory (and possibly exploitable) trends capturable with simple momentum strategies. In that case, we should observe trending behavior, especially for coherent themes with significant average pairwise correlation among member stocks. To analyze that, we looked at the persistence of basket residual returns. We started by examining this persistence for 60-day cumulative residual returns before and after each basket's release date. However, we were concerned that this magnified the importance of the release date when that date may only loosely relate to exactly

Figure 8. Portfolio Residual Risk as a Function of Average Pairwise Correlation



Source: BlackRock Systematic.

when investor attention focuses on the theme and flows occur in the theme baskets. To account for that, we also looked at persistence of 60-day basket residual returns around four additional dates spread from two years prior to release to two years after release. Figure 9 displays the results.

Every point on this chart corresponds to a particular basket at a particular time within two years of the release date. We have distinguished the significant average pairwise correlations by showing those points in orange while displaying themes with insignificant average pairwise correlations in blue. For each point, we estimate the *z* score of the average pairwise correlation calculated over the 60-day period prior to the analysis date. The shaded regions display pointwise 95% confidence intervals produced by the bootstrap. These are confidence intervals around the assumed linear relationship between the future residual returns and past residual returns. Since each basket appears five times in the plot, we also implemented a stratified bootstrap, sampling symbols with replacement. The results were essentially the same.

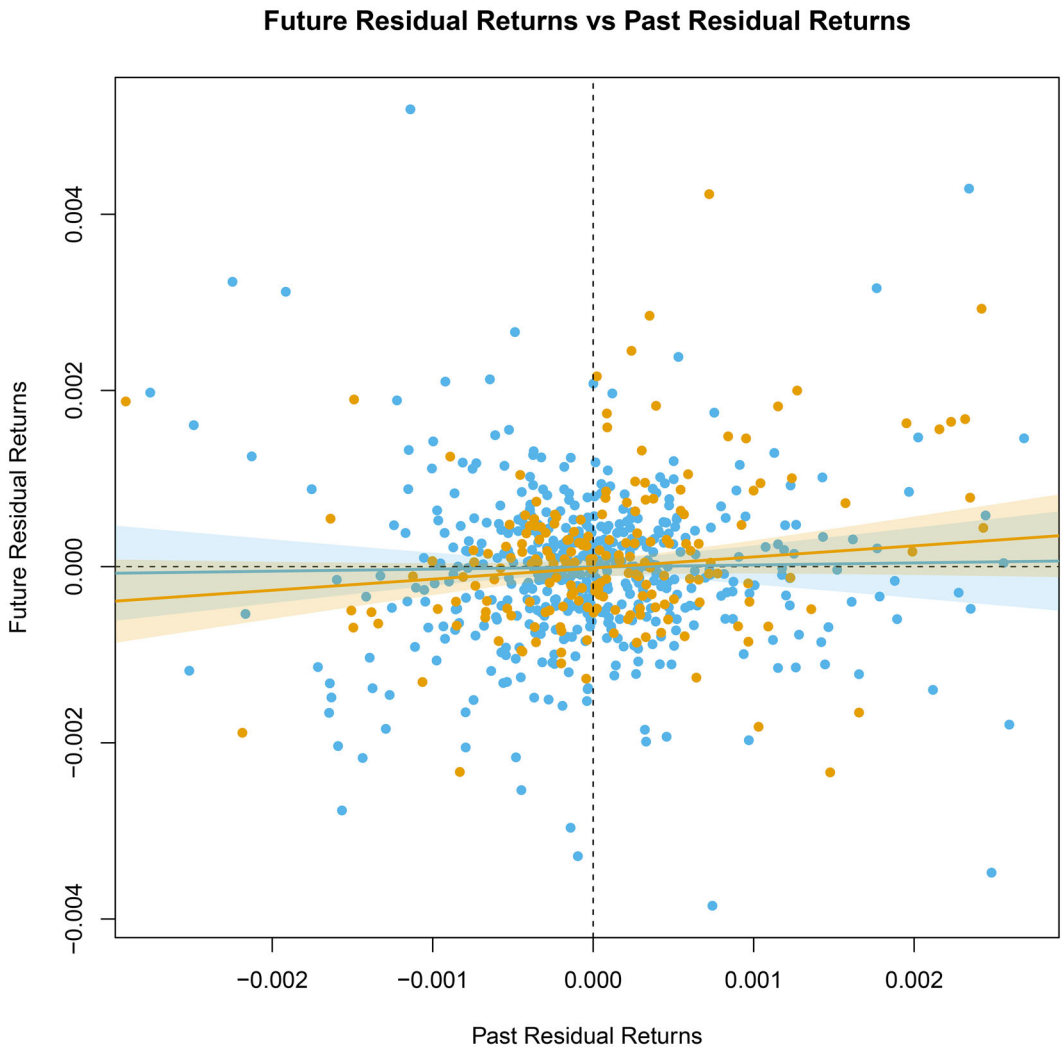
We also show the trend lines (in the same respective colors) for significant and insignificant average pairwise correlations. We estimate the slopes of these regression lines by regressing future residual returns against past residual returns separately for the baskets exhibiting significant and insignificant *z* scores. Figure 10 displays the result.

Consistent with Figure 9, if the average pairwise correlation is statistically significant, we see a statistically significant trend line for basket idiosyncratic returns and no significant trend otherwise. This leads to an implementable thematic investing strategy based on ex post analysis of baskets after release. Average pairwise residual return correlation should inform our decision to invest in a given coherent theme basket.

Summary and Conclusions

We define coherent themes by the transient correlation among a basket of stocks beyond what traditional risk models capture. We introduce a test statistic, the average pairwise correlation, and a new

Figure 9. Persistence of Theme Basket Idiosyncratic Returns.



Note: This analysis of performance may not hold in the future.
Source: Goldman Sachs and BlackRock Systematic, June 2025.

Figure 10. Regressing Future Residual Returns on Past Residual Returns

Average PWC z-score	Estimated coefficient	t-statistic	Observations	R-Squared
z>=2	0.126	2.11	238	0.02
z<2	0.028	0.47	458	0.000

mosaic + bootstrap methodology for testing its significance for any set of stocks. We observe only economically insignificant average pairwise correlations of residual returns across the entire estimation universe. In contrast, we show that many Goldman Sachs thematic baskets meet our definition of coherent themes. Risk models will underestimate risk for these baskets and coherent theme baskets; that is, baskets with statistically significant residual return correlations exhibit significant trending of their residual returns. Even though our analysis is ex post, utilizing published thematic baskets, it leads to investible risk and return implications because the correlations persist for at least 60 days after release. Backtest performance of a strategy based on this idea could further demonstrate this point; however, we do not show any such analysis in this paper.

In the case of themes, a second-order statistic (correlation) can help predict a first-order statistic (return). We plan future research to see whether we can use observed residual return correlations to identify newly emerging coherent themes and determine whether their behavior is consistent with our observations in the paper.

Editor's Note

Submitted 18 November 2024

Accepted 25 June 2025 by William N. Goetzmann

Notes

1. We prefer using the mosaic + bootstrapping method (described later) to invoking random matrix theory because random matrix theory results can depend delicately on distributional assumptions about the errors (e.g., Gaussian or elliptical distributions) or the precise choice of asymptotic regime. Our methodology relies only on a transparent local exchangeability assumption (see Spector et al. [2024] for discussion).

2. We estimated this using the mosaic tiling approach discussed later in the paper. That approach breaks the dataset of stocks and dates into tiles and runs separate cross-sectional regressions in each tile. We will need this for our analyses of statistical significance. It has limited impact on the estimated factor returns and residuals. Our liquid large-cap universe of US equities consists of Russell 1000 stocks plus additional stocks that satisfy market cap, volume, and price restrictions. We apply grandfathering to the criteria to avoid stocks bouncing in and out of the universe. This leads to roughly 100 additional US equities though the exact number varies over time. We have used the Barra GEM3 model for the analysis throughout this paper. This model includes 34 industry dummy variables and 11 style variables (including Momentum, Size, Value, Liquidity, and Volatility). We have also repeated the analysis using the BlackRock Fundamental Equity Risk model, BFRE. Our results and conclusions are unchanged by using a different commercially available risk model. For example, the correlation of GEM3 and BFRE prerelease z scores is 84% and the correlation of post-release z scores is 86%, where these are z scores of the observed average residual return correlations. We also observe the same trending results (shown in Figures 9 and 10 at the end of this paper).

3. We chose these two stocks only as a clarifying example as they are large, well-known companies. This is not investment advice.

4. To choose the panels and blocks, we use the default approach from Spector et al. (2024). This involves (i) splitting the data across time into disjoint batches of 10 consecutive observations and (ii) splitting each 10-day batch into K disjoint, equal-sized groups of stocks, where K is the largest number so that in each group there are at least 5 times as many stocks as there are factors. These groups are chosen uniformly at random within each batch, so the groups change from batch to batch. This ensures that most assets can be separately permuted at most times.

5. See Kahn (2018) for a discussion of how the choice of intuitive factors in the original Barra model was critical in encouraging investment managers to use an approach much more mathematically sophisticated than the investment industry at that time.

6. We used ChatGPT to sort the Goldman Sachs theme portfolios into macro themes on the one hand and industry, sector, geography, or style portfolios on the other hand. Early in this research, we performed this sorting by hand. The ChatGPT results are very similar though applied to a larger set of Goldman Sachs theme portfolios (because the number of theme baskets is growing).

7. One way to understand economic significance here is to see the implications of assuming correlations of zero when they are actually −0.0013. Figure 8 shows the example of a 50-stock equal-weighted portfolio where each stock has residual risk of 25%. If all the stocks are uncorrelated, the

portfolio residual risk is 3.5%. If they are actually correlated at −0.0013, the portfolio residual risk is 3.4%. That is not an economically significant difference.

8. We mention Wayfair only as an interesting historical example to illustrate our point. This reference is not investment advice.

9. In a subsequent analysis, we also investigated average pairwise correlation for 30 BlackRock megatrend ETF portfolios quarterly from 2019 (or fund inception) through 2024 and find them statistically significant 38% of the time, roughly consistent with our results on the Goldman Sachs theme baskets. In most cases, these portfolios exhibit statistically significant pairwise correlations over particular sub-periods, that is, not over the full five-year period.

10. See Warther (1995) or Barberis and Shleifer (2003).

11. Our approach is similar to Shiller (2017), Shiller (2019), Bhargava et al. (2023), Bybee et al. (2024), and Hirshleifer et al. (2025).

References

Barberis, N., and A. Shleifer. 2003. "Style Investing." *Journal of Financial Economics* 68 (2): 161–199. https://doi.org/10.1016/S0304-405X(03)00064-3.

Bhargava, R., X. Lou, G. Ozik, R. Sadka, and T. Whitmore. 2023. "Quantifying Narratives and Their Impact on Financial Markets." *The Journal of Portfolio Management* 49 (5): 82–95. https://doi.org/10.3905/jpm.2023.1.472.

Blanqué, P., M. Ben Slimane, A. Cherief, T. Le Guenedal, T. Sekine, and L. Stagnol. 2022. "The Benefit of Narratives for Prediction of the S&P 500 Index." *The Journal of Financial Data Science* 4 (4): 72–94. https://doi.org/10.3905/jfds.2022.1.107.

Bybee, L., B. Kelly, A. Manela, and D. Xiu. 2024. "Business News and Business Cycles." *Journal of Finance* 79 (5): 3105–3147.

Choy, J., M. Dutt, B. Johnson, A. Jung, K. Lamont, Z. Sanzgiri, L. Tran, and Y. Wu. 2022. "Morningstar Global Thematic Funds Landscape 2022." *Morningstar*, March.

Hirshleifer, D., D. Mai, and K. Pukthuanthong. 2025. "War Discourse and Disaster Premium: 160 Years of Evidence from the Stock Market." *The Review of Financial Studies* 38 (2): 457–506. https://doi.org/10.1093/rfs/hhae081.

Kahn, R. N. 2018. *The Future of Investment Management*. Charlottesville, VA: CFA Institute Research Foundation.

Lamont, K., M. Caley, D. Motori, and M. Black. 2024. "Morningstar Global Thematic Funds Landscape 2024." *Morningstar*, October.

Shiller, R. J. 2017. "Narrative Economics." *American Economic Review* 107 (4): 967–1004. https://doi.org/10.1257/aer.107.4.967.

Shiller, R. J. 2019. *Narrative Economics*. Princeton, NJ: Princeton University Press.

Somefun, K., R. Perchet, C. Yin, and R. Leote de Carvalho. 2023. "Allocating to Thematic Investments." *Financial Analysts Journal* 79 (1): 18–36. https://doi.org/10.1080/0015198X.2022.2112895.

Spector, A., R. F. Barber, E. Candes, T. Hastie, and R. N. Kahn. 2024. "The Mosaic Permutation Test: An Exact and Nonparametric Goodness-of-Fit Test for Factor Models." https://arxiv.org/abs/2404.15017.

Warther, V. 1995. "Aggregate Mutual Fund Flows and Security Returns." *Journal of Financial Economics* 39 (2-3): 209–235. https://doi.org/10.1016/0304-405X(95)00827-2.

Appendix A

Here are several additional test statistics we considered, though, in the end, we opted for the intuitive simplicity of average pairwise correlation:

- Largest eigenvalue: *We compute the $N \times N$ covariance matrix of the residuals. Letting $\lambda_1 \geq \ldots \geq \lambda_N$ be the eigenvalues of this covariance matrix, we set $T = \lambda_1$. The rationale is that a large eigenvalue may be a sign of a missing factor.*
- Scaled largest eigenvalue: *This measures the fraction of variance captures by the first eigenvalue:*

$$\frac{T = \lambda_1}{\sum_{j=1}^{N} \lambda_j.} \tag{A1}$$

Once again, a large value may indicate that we are missing factors.

- *Average maximum correlation: For each stock, n, calculate the maximum of its correlations with all other stocks $m \neq n$. Then average those maximum correlations over all N stocks.*
- Deviance: *We compare the statistical likelihood of observing our historical returns with the risk model covariance* $\mathbf{V}$*, as specified in Equation (2), versus a risk model with empirical estimates of every element of the covariance matrix* $\mathbf{V}^*$*. Our deviance test assumes returns are Gaussian with corresponding log likelihoods:*

$$l(\mathbf{V}) = -\frac{1}{2}\sum_{t=1}^{T}\left\{\mathbf{r}_t^T\mathbf{V}^{-1}\mathbf{r}_t + log|\mathbf{V}_t| + N \cdot log\{2\pi\}\right\}, \tag{A2}$$

where $\mathbf{V}$ is a candidate covariance matrix. The deviance test statistic is:

$$T = -2\left\{l(\mathbf{V}) - l(\mathbf{V}^*)\right\}. \tag{A3}$$

A large value would indicate the risk model is missing important factors/themes.

Appendix B

Here are some additional mathematical details on how regression deflates the variances of the

residuals. We believe that our returns follow a factor model:

$$\mathbf{r} \sim \mathbf{X} \cdot \mathbf{b} + \mathbf{u}. \tag{B1}$$

We will further assume that the returns have covariance matrix **V**:

$$V = \mathbf{X} \cdot \mathbf{F} \cdot \mathbf{X}^T + \Delta, \tag{B2}$$

with the residuals having a covariance matrix, Δ. For our purposes here, we will make no assumptions that Δ is diagonal.

Now assume that we estimate those residuals via linear regression. In that case:

$$\varepsilon = \mathbf{r} - \mathbf{X} \cdot (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{r} = (\mathbf{I} - \mathbf{K}) \cdot \mathbf{r}. \tag{B3}$$

Here we are using ε to denote the regression-estimated values of **u**. According to Equation (B3), the covariance matrix of the estimated residuals is:

$$Cov\{\varepsilon\} = (\mathbf{I} - \mathbf{K}) \cdot Cov\{\mathbf{r}\} \cdot (\mathbf{I} - \mathbf{K}). \tag{B4}$$

Note, however, that $(\mathbf{I} - \mathbf{K}) \cdot \mathbf{X} = 0$. Hence:

$$Cov\{\varepsilon\} = (\mathbf{I} - \mathbf{K}) \cdot \Delta \cdot (\mathbf{I} - \mathbf{K}). \tag{B5}$$

We can see in Equation (B5) that the covariances of our estimated residuals are shrunk by the two factors of (**I** – **K**).