Alternative Risk Premium: Workhorse or Trojan Horse?

by

Stephen Gorman

A dissertation submitted in partial fulfillment
Of the requirements for the degree of

Doctor of Philosophy
[EDHEC Business School]

October 2021

Dissertation Committee

Nikolaos Tessaromatis, PhD
Hossein Kazemi, PhD
Emmanuel Jurczenko, PhD
Enrique Schroth, PhD

EDHEC BUSINESS SCHOOL

APPROVAL FORM

CANDIDATE'S NAME: Stephen GORMAN

TITLE OF DISSERTATION: **"Alternative Risk Premium: Workhorse or Trojan Horse?"**

APPROVAL:

Nikolaos Tessaromatis		*Chair*
Hossein B. Kazemi	Hossein Kazemi	*External Examiner*
Frank. J. Fabozzi		*Member*
Emmanuel Jurczenko		*Member*
Enrique Schroth		*Member*

☑ with minor revisions

☐ with major revisions

DATE: 25 October 2021

Abstract

Alternative risk premium (ARP) is an investment category consisting of a wide range of rules-based trading strategies targeting returns representing compensation either for bearing risk or behavioral biases among market participants. These systematic strategies span all the major asset classes, trading equity indices, government bonds, currencies, commodities, credit spreads, volatility, and individual stocks. ARP constituents generally share the following three characteristics: (1) clear economic rationale supported by empirical research, (2) persistent risk-adjusted return distinct from that of traditional beta, and (3) liquid (scalable), rules-based and transparent, with a predominantly long-short trading profile.

Assets under management in ARP increased significantly in the wake of the Global Financial Crisis through 2017. Poor performance by diversified ARP funds over the 2018-2020 period abruptly reversed this trend, producing considerable soul searching regarding the role of this category in institutional portfolios. Frustrated investors attributed the recent outcome to many causes, ranging from a brutal style headwind to myopia by the quantitative investing community.

To ground this debate, the first paper addresses ARP benchmarks, which remain elusive, making performance evaluation challenging. Focus on this topic understandably intensified with recent disappointing performance. This paper introduces comprehensive categorical and statistical families of ARP benchmarks, using a proprietary database of tradable bank indices. The exercise includes a detailed and overdue discussion of the many nuances of ARP data, including classification, curation and interpretation.

Specifically, this research applies agglomerative hierarchical cluster analysis, partial least squares, elastic net regularization and principal component analysis to a database of 2,000 tradable bank indices to supplement a partially-nested family of categorical benchmarks with a fully-nested family of statistical benchmarks. Given the difficulty of ARP performance evaluation, the benchmarks introduced here represent an important methodological complement to the small number of benchmarks currently available and facilitate analysis at different levels of granularity.

The second paper utilizes the statistical benchmarks to analyze ARP performance between 2018 and 2020. Little research focuses upon this three-year period for systematic investing, with recent papers investigating the quantitative equity space. No comprehensive study of multi-asset ARP returns during this window exists, so this paper fills an important gap and provides a foundation for subsequent studies.

This empirical paper approaches the topic by questioning what the investment community missed given the information available at the end of 2017. The focus is identifying the deviations from expectations most responsible for the ARP performance problems between 2018 and 2020. This investigation involves establishing appropriate expectations for Sharpe ratios, cross-correlations, auto-correlations, skewness, kurtosis, and state-based relative returns to serve as the basis for evaluating outcomes during the period in question. The results reveal four strategy groups principally responsible for the poor performance of diversified ARP portfolios — equity sensitive, volatility sensitive, diversified stocks and value oriented. The problem is predominantly one of average returns, with successive market crises weighing on the first two groups and an historic lack of breadth wreaking havoc on the latter two.

Table of Contents

Alternative Risk Premium: Workhorse or Trojan Horse?

Acknowledgments

The journey to a PhD, particularly for someone in my position, requires the support of so many people. I first must thank my advisor, Frank Fabozzi, for the patience he displayed, the guidance and support he provided, and the wisdom he shared over many years of working together. I am honored to be his final doctoral student and fortunate to depart the program with both a mentor and friend.

I want to thank EDHEC Business School, specifically Nikolaos Tessaromatis, for accommodating me when work and family responsibilities forced me to sideline my research for almost two years. The EDHEC PhD program was a perfect fit for me – such a great combination of interesting courses, high quality professors, and opportunities to interact with students from around the globe. I am grateful that I was afforded the opportunity to complete my research and to earn my degree.

I want to recognize my employer, Wellington Management Company, for supporting my academic journey. I want to acknowledge the numerous banks that worked with me over the years to provide survey data. I am grateful to Don Tunnell and Chris Grohe for sharing their quantitative equity perspective and data, and to Varun Dubey, Aaron Nahabedian, and Francesco Fabozzi for their assistance with data management.

Finally, I want to thank my wife, Tammi, for encouraging me to pursue a PhD, knowing full well that I have a demanding job, that we have a very busy household with five children, and that classes and research would consume many nights and weekends. I also am grateful to my father, Raymond, for his consistent interest and encouragement over the years.

List of Tables

List of Figures

The Data Dilemma in Alternative Risk Premium: Why Is a Benchmark So Elusive?

Stephen Gorman

Abstract

Alternative risk premium (ARP) is an investment category consisting of a wide range of rules-based trading strategies targeting returns representing compensation either for bearing risk or behavioral biases among market participants. These strategies span all the major asset classes, trading equity indices, government bonds, currencies, commodities, credit spreads, volatility, and individual stocks. ARP constituents generally share the following three characteristics: (1) clear economic rationale supported by empirical research, (2) persistent risk-adjusted return distinct from that of traditional beta, and (3) liquid (scalable), rules-based and transparent, with a predominantly long-short trading profile. Assets under management in ARP have increased significantly over the past decade, but benchmarks remain elusive, making performance evaluation challenging. Focus on this topic has intensified with recent disappointing performance. This paper introduces comprehensive categorical and statistical families of ARP benchmarks, using a proprietary database of tradable bank indices. The exercise includes a detailed and overdue discussion of the many nuances of ARP data, including classification, curation and interpretation. These benchmarks mark an important foundational milestone for analysis in this evolving space.

Keywords: Alternative risk premium, multi asset, benchmarks, tradable indices, data imputation, partial least squares, elastic net, principal components analysis

1.1 Introduction

Alternative risk premium (ARP) emerged as an absolute return solution for investors seeking refuge from crashes in traditional asset classes and greater transparency, better liquidity and lower fees than hedge funds – an *alternative* to traditional beta and to hedge funds. ARP includes a broad spectrum of systematic trading strategies incorporating multiple investment styles (carry, trend, convergence, and risk anomaly) and covering all the major asset classes (equity indices, government bonds, currencies, commodities, credit spreads, volatility, and individual stocks). Gorman (2019) provides a foundational exploration of ARP and its deep academic roots, positioning these strategies as the confluence of investor supply-demand dynamics; redeployment of quantitative equity and systematic macro and tactical asset allocation investment processes; and decades of research on empirical anomalies, hedge fund replication, multi-factor models and data snooping.

As an alternative investment, the initial focus was on the performance objective rather than benchmarks. ARP portfolios fell into the cash-plus category of investments, meaning returns in excess of cash defined success. Given the diversifying nature of the underlying investments, practitioners typically assigned a Sharpe ratio expectation of 0.7 to 1.0 to ARP portfolios. Supplemental performance evaluation entailed reviewing return contributions by style and asset class within the ARP portfolio and making peer relative comparisons at the aggregate portfolio level. Given the significant heterogeneity across solutions in both defining and weighting individual ARP strategies, this initial approach was adequate, but it was contingent upon performance generally being in line with expectations.

As the amount of money invested in ARP has increased and performance in recent years has disappointed, the pressure has intensified to refine performance evaluation. The focus has shifted from the performance objective to benchmarks. Because ARP targets "factors" and trades systematically, the topic of benchmarks is not new, but current circumstances are driving demand for a solution to this vexing problem. The two candidates are primitive strategy and composite strategy benchmarks.

1.1.1 Primitive Strategy Benchmark

A primitive strategy benchmark relies upon a simple, reductionist rule base to represent a given alternative risk premium. This position-based approach attempts to meet the benchmark criteria of Maginn *et al*. (2007) listed in Table 1. While commonly applied to long-only investments in stocks and bonds, this philosophy does not extend neatly to ARP. Unlike a capitalization-weighted equity benchmark, an ARP primitive strategy benchmark has no theoretical foundation, no buy-and-hold profile, no canonical methodology with respect to factor specification or portfolio construction. In short, no truly passive alternative exists for a given ARP.

Maginn *et al*. (2007) anticipate the ARP challenge as they consider different types of benchmarks. The authors struggle with the large 5.4% return difference in 1999 between the S&P Large Value Index and the Russell Large Value Index, acknowledging that each target the same investment style but are not necessarily equally applicable as a benchmark for a given manager. Moving from style to factor benchmarks, they highlight a similar challenge. One can build a series of benchmarks having identical factor exposures, but the associated returns may be very different. Finally, the authors discuss the applicability of a custom security-based benchmark, emphasizing that such an approach must be tailored to the investment process of a given manager. The message is that the properties in Table 1 may define a valid benchmark but do not ensure a useful benchmark. A single, valid benchmark may not apply to seemingly similar portfolios, and potential return variability due to methodological variation across valid benchmarks complicates performance evaluation.

Ideally, the primitive strategy benchmark provides a common reference point for market participants. However, the absence of a standard methodology muddies this objective as competing index vendors, seeking differentiation, eventually offer equally defensible primitive strategy benchmarks for a given alternative risk premium. A common reference point becomes elusive in the presence of multiple possibilities that necessitate benchmark selection or blending protocols.

Even if a common reference point exists, will consultants and plan sponsors accept responsibility for imposing a primitive strategy benchmark on an ARP manager, in the process becoming accountable for the benchmark performance and incenting portfolio managers to focus upon benchmark-relative returns and tighter tracking risk? Committing fully to a simple, debatable benchmark methodology could be a step too far for asset owners and gatekeepers.

Despite the concept of primitive strategy benchmarks being well-established and often employed by asset managers in research, index vendors are in the very early stages of making such benchmarks available. As a result, the breadth of offerings does not exist yet to support the full spectrum of ARP styles, and the investment management industry does not yet have the applied experience base to understand fully the marginal insights, complexities and behavioral consequences of living with such benchmarks.

The appeal of primitive strategy benchmarks is clear – convenience, methodological parsimony and (potentially) a common frame of reference. The purpose of a benchmark is to facilitate understanding of and dialogue regarding portfolio performance. Primitive strategy benchmarks provide useful perspective and contribute to this process for ARP. However, ARP is not a simple investment category so expectations regarding the role of a simple benchmark should be consistent with this reality.

1.1.2 Composite Strategy Benchmark

Investment banks offer a plethora of tradable indices, representing the full gamut of ARP styles and approaches. As with primitive strategy benchmarks, specific rules govern these indices, ensuring transparency. Daily returns are available in Bloomberg for both types of index. Composite strategy benchmarks aggregate the performance of indices sharing similar characteristics. Each underlying index is position-based, so this approach produces a blended benchmark that is distinct from typical manager universe reference points. Therefore, primitive and composite strategy benchmarks essentially register the same desirable benchmark properties in Table 1. Unlike the methodological reductionism of a primitive strategy benchmark, a composite strategy benchmark provides a methodological spanning approach, diversifying away the idiosyncrasies of competing specifications to reveal the core profile of an ARP style.

Due to the different economic models underlying primitive and composite strategy benchmarks, slight differences do exist with respect to transparency. The profit opportunity for primitive strategy benchmarks is licensing to other parties that will convert index positions into products. Conversely, tradeable bank indices generate revenue via charges on invested assets.[1] As a result, index vendors provide T-1 positions (i.e. one day prior to the trade date) to those paying the licensing fee for the primitive strategy benchmark. Investment banks share T+1 positions (i.e. one day after the trade date) with those invested in the tradeable index. Access to complete index details is available in both cases, just at a price.

The challenge with composite strategy benchmarks is that targeting broad methodological representation requires a significant amount of data. No well-vetted, standardized, widely available database, such as Compustat or CRSP, exists for alternative risk premium. In fact, the financial literature has not provided a complete picture of the nuances and challenges of gathering and managing ARP data.

Composite strategy benchmarks also face the challenge of different aggregation approaches yielding different return profiles – the same issue confronting primitive strategy benchmarks. The number of contributing banks coupled with variation in ARP strategy classification, exclusion criteria, and weighting scheme contribute to potentially sizable return differences among ostensibly similar benchmarks.

Composite strategy benchmarks make an indispensable contribution to understanding ARP portfolio performance. As with primitive strategy benchmarks, no common reference point exists, and much work remains to be done. ARP performance evaluation

---

[1] Of course, investment banks could choose to license methodologies for indices on which they do not manage assets.

is necessarily a triangulation exercise, with a role for both benchmark approaches. Appreciating the benefits and limitations of the two benchmark conventions requires a thorough understanding of the underlying ARP data.

This paper fills a void by introducing a proprietary database of 2,000 tradable bank indices compiled by the author, in the process detailing the challenges of working with ARP data. The first objective of this paper is to enumerate, at a level of detail not previously available with a database unique in its comprehensiveness, the nuances of tradable index data, including taxonomy, specification variability, quality, redundancy, access, costs, and survivorship. Such perspective is a prerequisite for any ARP benchmarking exercise. The second objective is to introduce two families of composite strategy benchmarks, highlighting technical considerations and best practices. Three motivations anchor the approach of this paper to the ARP benchmarking problem.

1. Represent comprehensively the competing strategy specifications traded by ARP investors to provide a strategy spanning approach as part of the performance evaluation mosaic.
2. Facilitate the triangulation exercise required for ARP performance evaluation due to the absence of canonical strategy specifications.
3. Provide tiered ARP performance perspective to address performance questions ranging from narrow to broad.

Specifically, this research applies agglomerative hierarchical cluster analysis, partial least squares, elastic net regularization and principal component analysis to the proprietary database to supplement a partially-nested family of categorical benchmarks with a fully-nested family of statistical benchmarks. The nesting is comparable to a global equity benchmark rolling up various underlying region, country and sector indices. These results facilitate analysis at different levels of granularity and represent a unique contribution to ARP performance evaluation. Given the difficulty of ARP performance evaluation, the benchmarks introduced here represent an important methodological complement to the small number of benchmarks currently available.

The paper proceeds as follows. Section two contextualizes ARP data. Section three provides a detailed review of the proprietary database structure. Section four discusses best practices with respect to data curation. Section five analyzes the ARP metadata. Section six constructs the categorical and statistical benchmarks. Section seven provides comparative analytics on the benchmarks, and section eight summarizes.

1.2 ARP Data Preamble

1.2.1 Echoes of the 1990's

The current state of data in the alternative risk premium space is reminiscent of that confronted by academics and investors working with hedge fund returns in the late 1990's. At that time, analyzing hedge fund performance represented a stark departure from the preceding studies of mutual fund returns that benefited from the structure of a regulated environment and the availability of established databases such as the one offered by Morningstar. Because hedge funds are exempt from the Investment Company Act of 1940 contributing to a database is voluntary – no repository exists for returns and no industry association ensures that information is accurate. In the mid-1990's, several organizations commenced efforts to fill this data void. By the late 1990's academics began publishing papers summarizing hedge fund performance and highlighting the biases introduced by databases dependent upon discretionary submissions.

Ackerman, McNally and Ravenscraft (1999), Brown, Goetzmann, Ibbotson and Ross (1999), Fung and Hsieh (1997, 2000, 2002a), and Liang (1999, 2000) investigate hedge fund performance using some subset of the databases of Managed Account Reports (MAR), Hedge Fund Research (HFR), and TASS as well as hand-collected data from the U.S. Offshore Funds Directory. The authors encountered numerous unique data considerations (e.g., onshore versus offshore fund distinctions, incentive fees and high-water marks) and wrestled with performance measurement issues extending beyond the familiar survivorship bias shared with mutual fund data – namely, selection bias, instant history (or backfill) bias and end-of-life reporting bias.

The emergence of additional hedge fund databases (e.g., Altvest, CISDM) only added to the data complexities. A number of relatively new data vendors in a voluntary reporting environment resulted in a variety of fund classification systems, no standard fund identification codes, limited or non-existent graveyard fund databases, inconsistent fund representation across databases, and data discrepancies for funds represented in multiple databases. Other issues such as return smoothing and a preponderance of relatively short track records exacerbated the analytical challenges. The different databases and hedge fund classification schemes combined to produce a plethora of composites summarizing hedge fund style performance. Compared to stock and mutual fund records, hedge fund databases represented a veritable 'Wild West' that supported years of academic studies.

The availability of multiple hedge fund databases also precipitated the introduction of a financial technology solution to facilitate user access to and analysis of this information. In the spirit of the Ibbotson Associates EnCorr suite of software products (now owned by Morningstar) offered years earlier to support asset allocation work with a large cross-section of traditional market indices, the PerTrac Analytical Platform (now owned by eVestment) appeared in the late 1990's to provide access to the various hedge fund databases and analytical capabilities. PerTrac did not tackle the database consolidation problem but did represent a useful manager research tool.

Alternative risk premium data currently sits at a late-1990's-like juncture. Despite an abundance of research on the factor universe, no comprehensive (multi-asset), regularly updated, widely utilized data library exists – in the spirit of the Kenneth French Data Library for the Fama/French factors. Even if such a factor repository did exist, the returns would not represent implemented approaches. Because investors have been moving significant assets into alternative risk premium, an expanding universe of offerings exists among asset managers, but these funds generally are diversified (similar to fund-of-funds in the hedge fund databases so underlying strategy granularity is lacking) and offer relatively short performance track records. For example, the Societe Generale Multi Alternative Risk Premia Index provides the equally weighted performance of funds diversified across multiple asset classes and alternative risk premia. Tradable bank indices represent an intriguing data solution but also introduce a number of challenges, some familiar and others new.

During the past decade, investment banks have provided access to an increasing number of alternative risk premium by creating an index and delivering the associated returns to investors via a total return swap structure. A regulated financial institution therefore sponsors each index. A published, rules-based methodology supports each index. Almost all indices offer daily dealing terms (and pricing). Responsibility for index calculation has evolved from an internal function to a third-party agent that coordinates with the bank to an arms-length index administrator. Tradable indices today represent a seasoned, widely utilized means for institutional asset managers to invest in alternative risk premium for the following reasons.

- Breadth of truly alternative (long-short) offerings
- Transparency of process – a comprehensive rulebook accompanies each strategy
- Access to bank research resources and execution capabilities
- Capital efficiency and flexibility in targeting volatility afforded by a swap structure
- Desirability in certain regulatory environments -- e.g., commodity investments in UCITS (Undertakings for Collective Investment in Transferable Securities Directive) portfolios
- Execution ease – assuming an ISDA (International Swaps and Derivatives Association) master agreement is in place

Although these indices reside in Bloomberg, assembling a cross-section of data is extremely difficult. Index names often are generic, index descriptions are lacking, the universe evolves constantly, no screening tool is available, no uniform index classification framework exists, accessing indices may require bank permission, and obtaining detailed index information may involve a non-disclosure agreement. Tradable index design is an unregulated, highly competitive space devoid of an industry association -- banks simply have no incentive to coordinate on the data front.

Tradable index data is necessarily survey-based. Creating a true point-in-time database is not feasible. Survivorship and voluntary reporting biases are unavoidable. No complete record exists of indices that were launched years ago, performed poorly and were discontinued – any graveyard data is limited to tracking decommissioned indices over the survey history. Similarly, banks may not disclose the ticker for an existing index that has performed poorly but must remain accessible in Bloomberg until the last investor departs. Cost structure varies across indices.

Return histories for tradable indices are relatively short (typically 15 years) at a daily frequency. Each index represents a blend of live and back-tested history so back-test bias accompanies backfill bias. Banks provide no data on the number of trials underpinning a given strategy or the indices that never made it out of the lab. Tradable indices present many of the same challenges encountered in hedge fund data, with simulated return histories and data mining risk replacing smoothed performance and illiquidity considerations. Once harnessed, this alternative risk premium data represents the next frontier for voluntarily reported investment vehicle databases and a fertile ground for research.

As occurred during the hedge fund data evolution in the late 1990's, commercial data offerings are beginning to appear. Among the hedge fund index and database providers, HFR offers its Bank Systematic Risk Premia (BSRP) Index. The BSRP reports equal volatility-weighted post-publication returns and breaks down into approximately three dozen bank-classified asset class style sub-indices. HFR also maintains statistically grouped asset class style sub-indices (unpublished) and is developing simple tradable representations of styles to provide position versus performance-based indices.

Bloomberg and Goldman Sachs Asset Management (GSAM) recently released a family of alternative risk premia indices providing gross returns for macro and stock strategies. The position-based nature of these indices is an evolution that will enable index vendors to compete with investment banks for ARP product development opportunities.

Eurekahedge offers the Multi-Factor Risk Premia Index, a composite of systematic bank strategies. However, this index includes traditional beta strategies and therefore is not exclusively an ARP offering. On the financial technology side, PremiaLab, LumRisk, and Quantilia maintain databases of tradable indices to support an analytical front-end. Unfortunately, harmonization, regulatory, marketing and cost considerations continue to complicate access to tradable index data.

Harmonizing or consolidating the indices of different banks is an inconsistent process. The number of banks polled and the breadth of indices requested varies by surveyor. Banks manually map each index to the classification structure of the surveyor, with each surveyor having its own marginally different taxonomy. Graveyard (i.e., discontinued) indices, if included at all, depend upon the survey inception date.

With a distinct raw universe of indices in hand, each surveyor then proceeds to a filtering stage. Adjustments for closely related strategies, length of track record, strategy complexity, style constraints (e.g., multi-style indices), currency numeraire, dealing terms, and the balance of pre and post-publication returns reduce the index universe to its final size. Such data curation is an inherently discretionary process that varies by database administrator.

In addition to data processing variability, EU Benchmarks Regulation (governing the provision of, contribution to and use of benchmarks) may influence the willingness of a given bank to participate in a survey. Know Your Customer (KYC) regulations and the desire of banks to track sales opportunities may impede anonymous access to individual return series. Technology solutions can be expensive. For all these reasons, the tradable index data landscape continues to evolve.

1.2.2 Related Research

Research incorporating a cross-section of investable alternative risk premium data is both sparse and recent. A small number of papers utilizing proprietary tradable index data sets investigate a handful of ARP topics. Table 2 summarizes this research and punctuates the abundant opportunity for additional analysis in this space.

Hamdan *et al.* (2016) utilize a broad internal database of ETF's, bank strategies, and indices from index providers to provide a survey of alternative risk premium. After separating long-only constructs and excluding strategies blending risk premia or incorporating engineered trading rules, they produce a composite for each alternative risk premium using a statistical filtering technique, dropping strategies until the R-square among the remaining strategies exceeds a 70% threshold. The index return is a simple average of the surviving strategy returns. The authors then produce a range of summary statistics for their 59 alternative risk premium indices and 17 traditional risk premium indices, including Sharpe ratio, volatility-adjusted maximum drawdown, and the relationship between Sharpe ratio and skewness. The authors provide a lengthy motivation for alternative risk premium, highlighting the diversification potential but cautioning against naïve portfolio construction given the non-normality of ARP. Hamdan *et al.* conclude by using the alternative risk premium and 12 traditional indices within a lasso method to explain hedge fund performance, finding alternative risk premium to be useful in both in-sample and out-of-sample analyses.

Suhonen, Lennkh and Perez (2017) work with 215 strategies from a proprietary database of bank-sponsored alternative risk premium strategies and, in the spirit of McLean and Pontiff (2016), focus upon deterioration in the back-tested versus live Sharpe ratio across strategy groups. Their results raise the specter of overfitting. They report out-of-sample Sharpe ratio declines exceeding 50% by asset class and 60% by strategy (excluding volatility trades) and note relatively similar deterioration for strategies incepting before and after the GFC (so the GFC is not driving the finding). The authors introduce a simple, manual, categorical complexity score to demonstrate that larger Sharpe ratio declines accompany strategies that are more complicated. They also run pooled panel regressions on four strategies to assess the consistency of live versus back-test factor exposures, finding unintuitive factor instability in value equity strategies but recovering the anticipated significant exposure to the equity volatility risk premium, interest rate factors, and a naïve currency carry representation respectively in equity volatility, fixed income curve, and foreign exchange carry strategies.

Vatanen and Suhonen (2019) focus upon the risk profile of ARP strategies, contending that this in more stable pre and post-launch than returns. The authors organize an internal universe of bank indices into 8 style groups and 28 underlying composites based upon asset class. They use cluster and principal component analysis to show that the bank strategies fall broadly into offensive and defensive cohorts. The authors report a generally low beta to equity and commodity markets across the composites over the full 2007-2018 period, but a more positive beta to bond markets. They conclude by questioning the level of true stock and bond diversification offered by bank strategies due to the positive relationship between ARP and the stock or bond market in the lowest quintile of market returns.

Naya and Tuchschmid (2019) use an internal universe of bank strategies to focus upon data mining risk and the homogeneity of ARP strategies across index providers over the 2010 to 2017 period. The authors emphasize the importance of strategy selection given that the average strategy cross-correlation varies meaningfully across ARP style and over time, with homogeneity increasing as correlation with a benchmark rises. They warn about significant overfitting bias, highlighting a large average reduction in post-launch strategy performance and advocating an 80% discount of back test results.

Also using an internal database of bank strategies, Baltas and Scherer (2019) highlight the heterogeneity of index performance within both asset class and ARP style groups over the 2008 to 2018 period. The authors point to poor ARP performance coinciding with the worst return quintile for stock and bond markets and argue that an extension of the multifactor downside risk CAPM of Lettau, Maggiori and Weber (2014) better explains the cross-section of ARP returns than the simple CAPM. Paradoxically, they find weak evidence of downside risk compensation, speculating that the limited return history or data mining might be confounding their results.

1.3 Proprietary Database Structure

This section introduces a comprehensive, proprietary tradable index database and details the statistical characteristics of and important considerations regarding data that underpins an emerging category of alternative risk premium studies. Since only a handful of recent papers reference tradable indices, this section provides a thorough discussion of the nuances of this data. This paper endeavors to fill this void.

This paper leverages an annual survey conducted by the author over the past five years of the 16 investment banks providing almost the entirety of tradable indices to the world investing community underpins the database. These banks are Morgan Stanley, Goldman Sachs, J.P. Morgan, Deutsche Bank, Credit Suisse, UBS, RBC, Macquarie Bank, BNP Paribas, HSBC, Nomura, Citibank, Barclays, Societe Generale, CIBC, and Bank of America Merrill Lynch. To respect redistribution agreements with these institutions, this paper will genericize bank names, referencing only Bank 1, Bank 2, etc. going forward.

The database supporting this paper focuses upon an internally consistent universe of tradable ARP strategies and, as a result, does not include the following:

- ETF's -- predominantly long-only positions
- Enhanced beta -- 130/30 or factor-tilted long-only structures
- Hedge funds – inconsistent with ARP criteria
- Market-neutral indices from traditional index providers -- benchmark versus implemented methodologies
    - e.g., Dow Jones U.S. Thematic Market Neutral Indices, MSCI Market Neutral Barra Factor Indexes, iSTOXX Europe Single Factor Market Neutral Indices

In response to the annual survey, each bank supplies the 18 data items per index listed in Table 3. This collection of metadata represents an effort to paint as complete a picture as possible for each tradable index in a space that is rife with nuance. Given the absence of data standardization across banks, the survey defines permissible responses and reflects considerable collaborative engagement with the banks to ensure that survey responses reflect a consistent interpretation of each field. Based upon feedback from the banks, this database may represent the most comprehensive database undertaking to date in the tradable index arena.

Five additional data items (average gross and net exposure, strategy AUM and capacity, and UCITS eligibility) are part of the five-year history of conducting the survey; however, incomplete and inconsistent population renders these fields unusable. Enforcing a consistent, representative exposure reporting convention across indices trading a variety of instruments (e.g., multi-asset class options) in very different ways proved challenging. Strategy AUM and capacity are competitively sensitive items, so responses were either non-existent or vague. Some banks are reluctant to provide an indication of UCITS suitability, as they seek to avoid a survey response being misconstrued as a legal opinion regarding the usability of an index by regulated investors.

The following section summarizes the primary considerations accompanying each data item.

Bloomberg Ticker

Each bank provides a representative ticker for each strategy in its inventory of indices. A representative ticker is necessary because banks often manage slightly different versions of a strategy – an alternative laddering of option positions, a variation in futures roll methodology, a different volatility target, etc. These strategies are highly correlated and practically redundant. Because banks exercise judgment in selecting a representative strategy, data curation prior to any statistical analysis must check for and eliminate highly correlated return series to ensure fair representation of a given methodology in the overall universe.

Banks provide complete coverage of available tickers. This list does not include "desk strategies" (available on swap but limited in capacity or not cleanly represented in a rulebook) or custom strategies developed for a specific client. All the banks make the simulated index price history available in Bloomberg. These tickers typically have no intuitive structure (e.g., AQPEECSP, BXIIMMJE) and access to the price history may be restricted without permission from the bank. Typically, Bloomberg provides no descriptive information for these indices, instead directing interested parties to contact the bank directly for details. The database retains tickers that banks discontinue. This graveyard obviously exists only for the survey history and therefore is an incomplete record of retired strategies.

Index Name

Tradable index naming conventions vary greatly with banks conveying simple purpose for some strategies (e.g., FX G10 Carry, Cross Asset Trend, US Volatility Carry) while emphasizing unique identity for others (e.g., Volemont, GAINS, Gravity, AIR, ComBATS). This variation in naming practices makes it impossible to use Bloomberg's name search functionality to build a complete list of indices targeting similar outcomes.

Objective

The objective stratifies indices in a manner consistent with the taxonomy of Gorman (2019). Specifically, *enhanced beta* identifies long-oriented refinement of traditional beta. This enhancement can include factor tilts, unique portfolio construction or option overlays but remains a long-biased, traditional beta sensitive investment. *Traditional beta* denotes equity, bond, and commodity exposure in a conventional index fashion. *Other* covers eclectic strategies, predominantly long volatility structures but also distinct approaches such as the Credit Suisse Liquid Alternative Beta Index (CSLAB) which focuses on replicating an asset-weighted hedge fund index, targeting three hedge fund groups -- long-short equity, event driven, and global strategies (all remaining hedge fund types).

Systematic alpha is the most subjective objective, capturing situations in which the trade incorporates incremental insight that moves beyond simple harvesting of an alternative risk premium. This might involve blending risk premia or adding thresholds, downside risk mitigation or factor timing. A short volatility trade with a risk-off trigger falls in this category, as does a term premia trade that adjusts positioning depending upon short-rate momentum.

This definition implies that *alternative risk premium* targets a return source in a direct, less complex way than systematic alpha, but this distinction is too simple. Every strategy incorporates multiple decisions regarding signal specification and portfolio construction and no widely accepted baseline “recipe” exists so such a black-and-white characterization, while reasonable at the extremes, is challenging for many indices. What one bank classifies as alternative risk premium another might reasonably classify as systematic alpha.

The market positioning and ARP philosophy of the various banks also might influence this classification decision as some advocate simplicity or purity, adopting a quasi-passive orientation with an emphasis on execution platform strength, while others favor refining methodologies to enhance return and to mitigate risk, taking a quasi-active approach with an emphasis on intellectual property and research platform. Since establishing complexity criteria could result in artificial distinctions and would remain open to interpretation by banks, the survey did not follow this path. Noise and nuance represent inescapable realities with ARP, so this paper focuses upon the union of systematic alpha and alternative risk premium objectives, erring on the side of inclusivity at the outset to allow subsequent statistical analysis to determine distinctions.

Style

Each bank has its own style classification system, so the survey asks banks to map their internal choices to a standard set of options. This self or bank-classified system differs from a statistically specified structure appearing in subsequent pages. Because this exercise requires judgment, the potential exists for inconsistency and data entry errors, so a quality control process is important (e.g., reconciling the style with the description). The style classification system used in this paper strikes a balance among specificity, critical mass, parsimony, and necessity. The 14 options do not cover every possible strategy explicitly (e.g., a buyback index) but, with very few exceptions, provide a reasonable alternative.

Table 4 compares this classification system to the ones used in five recent ARP studies. While much common ground not surprisingly exists, the lack of uniformity also is apparent and complicates comparisons across studies. The taxonomy deployed in this database effectively spans the others and therefore provides a reasonable basis for analysis. The following provides a brief description of each of the 14 styles. Bear in mind that many specifications of a given strategy exist and that the orientation of the strategies within a given style can be *cross-sectional* (relative attractiveness of a position) or *time series* (stand-alone appeal of a position).

Carry (spread) includes strategies seeking income-like return (under a status quo market assumption) through *cross-asset* positions. In commodities, the strategy might have long positions in the most backwardated energy futures contracts and a short position in a broad energy index. In currencies, the positioning might be long the highest yielding emerging market currencies and short the lowest yielding emerging market currencies. In rates, a typical strategy has long positions in futures markets with the most attractive coupon plus roll-down combination and short positions in the least attractive markets. All three are example of cross-sectional carry (spread) strategies. In credit, a short position in the Markit CDX North America High Yield Index combined with a beta-adjusted long position in the Markit CDX North America Investment Grade Index targets the high yield premium. This is an example of a time series carry (spread) strategy.

Carry (curve) includes strategies seeking income-like return (under a status quo market assumption) with a *single-asset* focus. In commodities, this typically involves a long position in a deferred contract for a contangoed commodity combined with a short position in a nearby contract. In equities, this strategy might have a long position in the front end of the STOXX dividend futures curve hedged by a short position in a longer maturity contract. Both are calendar or time spread strategies. In rates, the structure could be a long position in a 12-month Euribor contract to profit from rolling down the yield curve, seeking to take advantage of the forward rate bias. In credit, a short position in the Markit CDX North America High Yield Index effectively rolls down the US high yield credit spread curve. All three are examples of time series carry (curve) strategies and are vulnerable to curve reshaping.

Congestion (rebalance, month-end) in the most basic form makes money by providing liquidity to passive indexes during rebalance periods, taking advantage of the pressure this pre-specified trading exerts on prices. In commodities, the strategy takes a long position in the contracts the Bloomberg Commodity Index (BCOM) will be buying during the monthly rebalance period and a short position in the contracts the BCOM will be selling in this window (i.e. a short time-spread position). The strategy closes this trade out during the BCOM rebalance window, ideally generating return via movement in the time-spread as compensation for providing liquidity to BCOM investors. In bonds and equities, congestion strategies aim to anticipate whether indexes or institutions will be buying or selling in their rebalance windows, trade ahead of the indexes in the anticipated direction, and then close the position as expected trading occurs. For example, one rates strategy takes a short-term, month-end long position in Treasury bond futures to benefit from a calendar effect. These strategies have liquidity provision and fundamental trend attributes and generally a time series orientation.

Merger Arbitrage harvests the spread associated with cash and stock-based company acquisitions, in the process assuming the possibility the deal collapses from those looking to offload this risk. The strategy involves a long position in the stock of the target company and a short position in the stock of the acquiring company (and no short position in a cash deal). This is a distinct class of time series carry (spread) trades with a predominantly stock-specific risk profile.

Multi-Style combines several ARP, in some cases across asset classes. In equities, this might involve combining value, quality, low beta and momentum strategies in a cross-sectional, long-short stock strategy. Similarly, in rates, commodities and currencies, this strategy often involves a blend of carry, momentum and value positions in a cross-sectional, long-short portfolio. The amalgamated nature of this style makes it a candidate for exclusion from statistical analysis.

Other represents an eclectic set of strategies not fitting clearly in one of the other styles. Strategies tracking hedge fund stock holdings (13F and 13D strategies) or seeking to replicate broad hedge fund indices reside here. A hedged stock portfolio tilted toward companies with appealing environmental, social & governance (ESG) profiles also falls in this category. The unique nature of this style also makes it a candidate for exclusion from statistical analysis (after confirming that none of the strategy return histories correlates with a different style, suggesting a possible classification error).

Reversal generally pursues short-term technical or sentiment-oriented retracement opportunities. In commodities, this strategy might have long (short) positions in the most oversold (overbought) contracts as indicated by the CFTC's Commitment of Traders (COT) report. In currencies, the positioning could be long (short) developed market currencies manifesting significant recent volatility-adjusted depreciation (appreciation). In equities, the strategy might be long (short) S&P 500 futures contracts if recent volatility adjusted performance was materially negative (positive). Reversal strategies may be cross-sectional (commodity example) or time series (currency and equity examples) in nature.

Risk Anomaly (quality, low volatility/beta) targets situations in which a portfolio structure with an attractive risk profile generates a higher Sharpe ratio than a less desirable alternative. In commodities, this strategy might take leveraged long positions in a set of low volatility contracts hedged with a short position in the BCOM. In equities, the leveraged long positions might be in relatively low volatility or high-quality (or profitability) stocks with the beta-neutral short positions in relatively high volatility or low-quality (or profitability) names. Risk anomaly (quality, low volatility/beta) is a predominantly stock-based and inherently cross-sectional style.

Size attempts to extract the return spread between small and large companies via long positions in the former and short positions in the latter (beta matched in some structures). This is a stock-based and inherently cross-sectional style.

Trend (cross-sectional momentum) leans into (away from) relatively strong (weak) price action. Similar approaches exist for commodities, currencies, rates, equity indices and stocks. For example, a commodity strategy might focus on 12-month excess return of the front-month contract, taking long (short) positions in the best (worst) performing tertile of the universe.

Trend (time-series momentum) leans into (away from) strong (weak) price action. As indicated by the name, the orientation of this style is time series so the emphasis is on absolute versus relative price action. As a result, the style, while uncorrelated with traditional assets over time due to its ability to take long and short positions, exhibits directionality (i.e. pro or counter-cyclicality) at points in time. Trend (time-series momentum) typically is the marginal determinant of beta in an ARP portfolio. Similar approaches exist for commodities, currencies, rates, equity indices and credit indices. For example, an equity strategy might focus on 12-month excess return and take long (short) positions in all futures contracts with a positive (negative) return, with the net position representing the appetite for pro-cyclicality.

Value seeks deviations from some notion of fundamental worth, taking long (short) positions in cheap (expensive) assets. Fair value in stocks and equity indices typically is a function of earnings or cash flow while inflation often is the focus in rates and currencies and default risk in credit. Fair value in commodities is more elusive and could be a function of futures term structure, marginal cost, or simply price reversion. This style almost exclusively executes relative value trades and therefore manifests a cross-sectional portfolio orientation.

Volatility (arbitrage) groups volatility trades with dynamic or non-short risk profiles and therefore is more general than a strict definition of volatility arbitrage as trading a delta neutral portfolio of an option and its underlying asset – with the common ground being the consideration of the attractiveness of volatility. In equities, this strategy might allocate between long and short positions in short-term VIX futures depending upon the shape of the VIX futures term structure. Alternatively, this structure might attempt to monetize realized dispersion of single stocks relative to index-implied correlation. This style is predominantly equity focused and time series oriented.

Volatility (short) harvests the variance risk premium, compensation from option market participants desiring to transfer the risk of a significant market event. The strategy predominantly sells delta-hedged straddles or strangles (or in some cases variance swaps). Volatility (short) is a time-series-oriented style, spanning commodities, credit, currencies, equities and rates.

Asset Class

Seven categories exist – *equity: index-based, equity: stock-based*, *commodity*, *credit*, *currency*, *multi-asset* and *rates*. The only nuance here relates to equities. The index-based equity strategies fit naturally with the other asset classes under a macro or top-down ARP umbrella. A time-series trend strategy trading global equity futures contracts falls in this group. The stock-based equity strategies represent a distinct subset of bottom-up ARP strategies, most aligned with traditional quantitative equity investing. For example, a value strategy purchasing cheap European stocks and hedging the beta with a short Euro Stoxx 50 futures position resides in this category. Despite the strategy trading an index, the driver of returns is stock selection. The stock-based equity group also includes strategies such as merger arbitrage and dispersion (trading long individual stock variance against short index variance).

Directionality

The vast majority of ARP strategies incorporate *long-short* positions. The *long-only* classification appears on carry strategies such as credit, dividend futures and (rates) term premium, but it also indicates gray-area strategies (candidates for Enhanced Beta or Other objectives) such as a time-series trend portfolio limited to long positions. Directionality therefore has potential filtering value for statistical analysis. The *short-only* classification predominantly applies to short volatility strategies, despite the delta hedge not limited technically to a short position. This flag also signals a potential issue with the objective classification – e.g., a time-series trend strategy limited to short positions. Finally, representation by the banks of risk exposure versus positioning occasionally generates some inconsistency in this field (e.g., a credit carry strategy is long credit risk but carries a short position in a credit default swap).

Region

Regional assignment options include *North America*, *Europe*, *Asia-Pacific*, *Emerging Markets* and *Multi-Region*. The classification is self-explanatory for all assets except commodities, for which the assignment reflects the regional orientation of the market. Brent crude and gasoil fall under Europe. Aluminum, cocoa, coffee, copper, gold, lead, nickel, platinum, silver, sugar and zinc are multi-region markets. Corn, cotton, WTI crude oil, feeder cattle, gasoline, heating oil, Kansas wheat, lean hogs, live cattle, natural gas, soybean meal, soybean oil, soybeans and wheat receive a North America flag.

Index Description

Given the limited information conveyed by index names, this brief explanation of the strategy provides valuable perspective. While this field often is light on details, it can be useful in confirming strategy metadata, understanding index return behavior and highlighting potential redundancy.

History Start Date

This technically represents the inception date for the back test supporting the tradable index, although supplemental history not conforming entirely to the published rule base may be available upon request for some strategies. Occasionally, flat filling of index values exists at the start and/or end of the price history in Bloomberg so one should check such occurrences against the incidence of static data over the complete index history.

Live Start Date

The live start date corresponds to the publication date for the tradable index, the point at which the rulebook was finalized and available to prospective investors and a calculation agent assumes responsibility for pricing the index. This date may or may not coincide with the initial funding of the index. Note that responsibility for index maintenance continues to evolve and varies across banks. A separate internal function or an external agent (e.g., Standard & Poor's, STOXX) traditionally priced the index and coordinated with the bank research department to address any rulebook issues. Regulatory pressure may increase usage of an index administrator, thereby completely outsourcing management of both index pricing and methodology.

Return Type

Almost all tradable indices are *excess return* vehicles, providing index total returns net of a local cash return. A few *total return* indices do exist, requiring the subtraction of the local LIBOR rate to render them comparable to the index universe.

FX Denomination

Tradable indices listed as *USD* dominate the universe. A limited number of indices have alternative currency denominations -- *EUR*, *JPY*, *GBP*, *AUD*, *CAD*, *CHF* or *Other*. If the non-USD denominated index is in excess return space, no comparability problem exists as excess return represents a standardized (or effectively currency hedged) return format. If the non-USD denominated index is in total return space, the FX denomination indicates the appropriate LIBOR rate to subtract to convert the index to its excess return form to facilitate comparison with the broad index universe.

Dealing Terms

Given the liquidity objective of ARP and the implication of the tradable index name, investors can transact *daily* in almost the entirety of this universe. A handful of indices trade on a *weekly* or *monthly* cycle and none lists the *other* frequency.

US Availability

Nearly the entire tradable index universe is available to US investors. For those indices listed as not being available in the US, this often indicates that the index is not *yet* available in the US but that the exact strategy or a very close approximation would be available upon request. As a result, US availability is not particularly useful as a screening variable.

1.3.1 The Four Cost Levers of Tradable Indices

In a portfolio context, investors historically treated tradable indices as representing net returns. This made it possible for a fund to appear to be a low-cost ARP provider by charging a low management fee while embedding various costs in the swap returns; however, the march in recent years toward greater clarity regarding the drivers of portfolio returns is facilitating more-informed fund comparisons. MiFID II (Markets in Financial Instruments Directive II) in Europe and RG97 (Regulatory Guide 97) in Australia legislated transparency regarding fund trading costs. Asset managers now must report direct and indirect transaction costs. Methodological inconsistency in the representation of costs, particularly for derivatives, remains a problem, but the increased focus on costs and execution efficiency is here to stay. An important contribution of this paper is framing the cost component of tradable indices – both the specific drivers and fungibility, profitability and negotiability considerations.

Tradable indices introduce four expense considerations split between internal and external cost buckets. The internal cost, index fee and trading costs, impact the net return calculation and exist within the quoted index values. The external cost, swap spread and in/out cost, represent additional expenses that reduce the final return experience of the investor – index values do not reflect these costs. The combination of index fee and swap spread represent the headline cost, conceptually analogous to a management fee. However, a lofty back-test and small headline cost could distract from significant embedded trading costs, so it is important to consider all four potential sources of realized return give-up.

Fee structures vary considerably among the banks, with most opting to use only two or three cost levers depending upon the strategy. This complicates tradable index return comparisons, as the emphasis on internal versus external costs is inconsistent. Reported index returns technically are scattered across the gross to net return spectrum. Further complicating matters is the fact that "costs" represent the confluence of execution realities and profit considerations. Banks are in the tradable index business to earn a profit. The index rulebook commits a bank to deliver the specified return stream. As a result, a bank has an incentive to be conservative regarding trading costs to avoid having to subsidize an index for which it underestimates execution costs. The profit margin could reside in the headline cost, be part of the trading cost calculation or be some combination of both. Such fungibility is another consideration when comparing net returns.

Finally, the costs associated with a tradable index are negotiable. A bank eager to fund an index will price it more aggressively than a bank with a mature, capacity-constrained index. The overall trading relationship between investor and bank and the size of an investment also influence these negotiations. An investor can negotiate any of the four cost components, with a reduction in the internal costs typically captured by a rebate to the swap spread. Therefore, negotiability represents yet another consideration when evaluating the 'rack' rates or list prices captured in this database.

Index Fee

This annual flat fee embedded within the published returns of some tradable indices represents compensation for operational oversight and/or strategy design. The index fee and the more prevalent swap spread constitute the *headline cost*, effectively the tradable index management fee. Given the wide variety of cost structures, there are instances in which the index fee also folds in executions costs.

Swap Spread

The typical implementation of a tradable index is an excess return swap with a one-year term and monthly resets (opportunities to upsize or downsize the notional exposure). The investor effectively pays LIBOR plus a spread to receive the index total return. The swap spread is similar in purpose to the index fee but represents a cost external to the published index returns. As such, it is the usual point of adjustment following cost negotiations. Changing internal cost structures requires publishing a new index so working with the swap spread (via a rebate) is an efficient alternative. Some banks charge no swap spread, preferring other cost levers. Given that a profit margin must exist after factoring in all costs, the choice among levers is mostly a function of business strategy. The survey asks banks to provide a representative or "rack-rate" (pre-negotiation) swap spread, but the possibility exists that some banks may be relatively aggressive in their database submissions.

In/Out Costs

In and out costs effectively represent commissions, transaction charges on any incremental change in the notional exposure of the swap that discourage investors from over-trading the position and cover banks in initiating and closing positions. This is an external cost, driven entirely by the trading decisions of the investor. When present, in/out costs apply almost universally to both sides of the trade (entering and exiting the swap). A few indices levy only an exit charge. Occasionally, indices include a swap break fee, a penalty for exiting a swap on a non-reset date. A break fee is distinct from in/out costs.

Trading Costs

This represents the cost of executing the strategy and is an internal cost, embedded within the published index returns. Because costs may vary over time, banks provide a single indicative annual cost estimate per strategy. These costs are a function of turnover and the instruments traded, with volatility strategies clearly bearing the heaviest implementation burden. The trading cost methodology for options varies across banks and its vega orientation makes it different from more familiar calculations for stocks and bonds. Vega indicates the change in an option price per a 1% change in the volatility of the underlying asset. A bank might calculate trading costs for a volatility strategy as follows. The current option implied volatility relative to a reference volatility provides a cost scalar to apply to the product of a base transaction cost expressed in vega and current vega, with the number of option contracts and the associated multiplier rounding out the transaction cost calculation.

From a back-test perspective, incorporating trading costs adds a level of robustness and represents a departure from the gross return orientation of academia. While most indices include trading costs, many do not. These costs are an important consideration for every strategy, but they may not be explicit, instead residing within the swap or index fee.

1.4 Data Curation

Given the unique nature of ARP data, numerous preparatory steps to ensure consistency and accuracy must occur prior to conducting any analysis of the proprietary database. The database design facilitates accomplishing the five primary objectives of this pre-processing stage.

1. *Confirm internal consistency of metadata*. Mistakes and misinterpretations by the banks are possible in a survey-driven data gathering process. Cross-checking responses and clarifying bank intent are important quality assurance exercises. Having multiple classification fields makes this possible. For example, long-only directionality might indicate an enhanced beta offering that does not belong in an ARP study. The index description might not support the style choice. The reported history start date might not align with the earliest available price in Bloomberg because additional data is available that is not completely consistent with the index rulebook.

2. *Align costs across indices*. Published tradable index returns incorporate costs on an inconsistent basis due to the variety of cost structures employed by banks. The more the pricing model of a bank leans on external (internal) costs, the closer the published numbers will be to gross (net) returns. To maximize comparability, gross returns add back internal cost whereas net returns subtract external cost. Within external cost, the in/out cost component assumes a three-year, fixed-size investment – a shorter (longer) holding period or more dynamic sizing would increase (decrease) the contribution from this source. Within internal cost, trading costs are constant and thus reflect an average versus point-in-time experience.

   The possibility of negotiating a cost reduction suggests that applying a discount factor to the total strategy cost might be reasonable. This process begins with a discount assumption of zero in the interest of conservatism, particularly given the upward bias in back-test returns. Increasing this scalar for sensitivity analysis is a small matter. Finally, a vintage effect may exist within the cost estimates, with older survey responses predating downward fee pressure in the space. This represents a more important consideration for some individual indices than the full cross-section.

3. *Convert total to excess returns*. For the small number of total return indices, the FX denomination provides the appropriate 1-month LIBOR series to subtract to generate an excess return history.

4. *Correct index start and end dates*. Banks sometimes provide static index values at the beginning (bank not yet pricing the index) and end (bank discontinued the

index) of the time series appearing in Bloomberg. Some of these indices include dynamic position sizing features that, at times, can eliminate positions for valid reasons so simply excluding flat-filled data is not appropriate. The objective is to balance the probability of discarding useful data and retaining incorrect data. Ignoring runs of constant index values at the start and end of each index time series, the flat-fill correction process determines the distribution of flat-filled sequences over the history of a given index. If the run length at the start (end) of the series exceeds the 90th percentile of this distribution, the start (end) date shifts to the first (last) change in the index value. This systematic process helps to ensure consistency between the index history available in Bloomberg and bank-supplied index inception and decommission dates.

This step also includes basic index-level integrity checks for outsized changes, restatements, negative or missing values, and sequences of static values inconsistent with the overall index history. Quality control is a critical consideration. The availability of ARP index data in Bloomberg is no guarantee that it is accurate and complete.

5. *Eliminate redundant indices*. The survey asks banks to provide a representative index for each strategy, but banks often include variations of a given strategy that over-represent a particular methodology in the database. These variations include differences in the following strategy parameters.
    a. Investable universe
    b. Volatility or leverage target

c. Weighting scheme -- equal weight versus risk parity, vega versus theta weighting, beta or volatility-adjusted versus equal-notional position sizing
d. Rebalance cycle and roll frequency
e. Portfolio constraints and exclusions – beta or gamma ceiling, elimination of a position (e.g., agriculture futures) or factor (e.g., equity size), liquidity requirements, seasonality adjustments, conditional filters
f. Volatility control and de-risking mechanisms
g. Hedging methodology -- time of day, laddering, instrument
h. Futures tenor or option expiry

Different start and end dates for related strategies complicate visual identification of redundancies and index descriptions do not address uniqueness sufficiency so a returns-based approach is necessary. Given a null hypothesis that two indices are distinct, this process attempts to manage the Type 1 error, concluding that two indices belong to the same methodological family when in fact they do not. The three-step elimination process first identifies indices *published by a given bank* having a correlation exceeding 0.90 (~80% R^2) for the longest available overlapping period of returns.

Next, metadata (asset class, style, region, description) confirms redundancy. The index with the latest flat-fill adjusted end date survives. Ceteris paribus, return type (excess return preferred), FX denomination (US dollar preferred), years of return history (more preferred), and dealing terms (daily preferred) determine the surviving index. One surviving index may represent multiple redundant indices. The live start date for the surviving index represents the earliest date listed among the related strategies.

Finally, the process maximizes the retention of return history. In certain instances, this requires splicing the standardized (volatility-adjusted) return history from a related (dropped) index to the beginning of the surviving index history. To ensure materiality, back-filled history must exceed six months or extend the existing return history by at least 5%.

The proprietary database contains approximately 2,500 bank indices reporting systematic alpha or alternative risk premium as the objective. The flat-fill correction process adjusts start dates for 10% of this raw universe. End date adjustments apply only to graveyard indices, those no longer priced and representing just under 20% of this universe. Since few index return histories extend further back than the late 1990's, this paper focuses upon weekly returns, to manage the synchronicity issues accompanying daily returns, over the period 12/31/1999 through 8/31/2020.

The intra-bank redundancy check eliminates 20% of the raw universe. Among the affected indices, 80% of the surviving indices eliminate a single index, implying that 100 surviving indices each displace multiple indices from the same strategy family. This process surfaces very few debatable redundancies, consistent with the desired low Type 1 error. Potentially undesirable eliminations include indices trading different, but highly correlated, instruments such as WTI versus Brent crude oil futures or 10-year versus 5-year US Treasury futures. In other words, this process effectively distinguishes between very similar strategy construction rules and high correlation among underlying assets.

Redundancy is reasonably consistent across banks and ARP styles, with an index attrition rate of 15-30% applying to the underlying constituents of each.

Approximately 5% of the surviving indices inherit return history (back-fill) from a dropped index. A three-year minimum return history requirement eliminates only a handful of indices. The result of the data curation process is a working universe of almost 2,000 bank indices.

1.5 Tradable Bank Index Universe Metadata Review

1.5.1 Universe Characteristics

The working universe for this paper contains 1,932 tradable bank indices sourced from 16 investment banks. Figure 1 shows the variation in the size of each bank's index inventory, with some banks having broad, well-established ARP businesses while others occupy niches or are relative newcomers to the space. Regardless, bank proportional representation (p) is not overly concentrated, with the universe carrying a Herfindahl-Hirschman Index (*HHI*) of 8.4 against a possible range of 6.3 to 100.

$$HHI = \sum_{i=1}^{n} p_i^2 * 100 \qquad \text{Equation 1}$$

The data set sufficiently represents most ARP styles, with only merger arb and size sparsely populated. Volatility (short) includes many single-asset volatility carry trades (S&P 500, Japanese yen, copper, US 10-year Treasury, etc.) that inflate the number of indices in this style. From an asset class perspective, the largest number of indices trade equities, split between index-based and stock-based strategies. Indices trading commodities represent the single largest block. Ignoring volatility strategies materially reduces the proportional representation of equity (index-based) indices while only slightly decreasing that of rates and commodities. Equity (stock-based) indices absorb much of this reduction.

Figure 1 also shows that most indices span regions, with the global nature of many commodities significantly influencing this result. Equity (stock-based) indices exist in a variety of regional and global forms whereas commodity, rates and currency typically are broadly diversified strategies. Finally, most of the universe consists of USD denominated excess return indices offering daily dealing terms. Graveyard indices (those no longer priced) represent 17% of the universe.

Table 5 provides a crosstab of the working universe by asset class and style. Of the 98 possible combinations, approximately a third are unpopulated, a third are thinly populated, and a third show some critical mass (green shading). Among the asset classes, the relatively limited number of strategies in credit reflects current implementation realities. These indices trade a handful of liquid CDX and iTraxx credit default swap (CDS) indices. Expanding the number of credit strategies depends upon finding a way to trade CDS efficiently at the company level.

Among the styles, merger arb and size are specific to stocks and therefore do not extend to other asset classes. Curve strategies are not relevant in stocks and currencies. Congestion is predominantly a commodity strategy while risk anomaly is primarily stock oriented. Carry, trend and volatility generally apply broadly across asset classes. The green shaded areas in Table 5 highlight a logical, representation-driven starting point for creating ARP performance benchmarks, setting aside the infeasibility or irrelevance of unpopulated boxes and the lack of consensus implicit in the thinly populated boxes.

1.5.2 Universe Return Availability

Figure 2 summarizes the index return availability across the universe. Index returns are relatively plentiful from a frequency perspective (daily availability) but relatively limited from a historical perspective (exposure to a narrow set of economic cycles). Over 80% of indices have returns predating the Great Recession but only 30% have returns predating the 2001 recession. The median index return availability is approximately 16 years. Figure 2 also highlights the variability in return history by style. Trend and carry strategies offer the longest average return histories while volatility, merger arbitrage and reversal offer the shortest.

A distinguishing feature of bank index returns is that the history represents a blend of live and back-tested returns. Figure 3 illustrates that these returns predominantly consist of pre-publication data, reflecting the recent emergence of ARP as an investment category. The median live index return availability is approximately four years, representing only a quarter of available returns. This duality in index return history poses a considerable challenge when analyzing performance – specifically, disentangling the potential influence of sampling error, data mining, and environmental headwinds on live index returns.

1.5.3 Universe Cost Data

As introduced previously, cost is an important and largely unreported tradable index consideration. Variability, profitability, and negotiability are the key dimensions. Figure 4 summarizes the variety of ways investment banks use the four cost levers – index fee, swap spread, trading costs, and in/out fees. The top left panel shows that banks use index fees infrequently, instead relying upon the other three costs with similar regularity. The top right panel punctuates the variability in cost structure, revealing the preferred combination of swap spread, in/out charge and trading costs to represent only a third of the cost combinations. The bottom panel shows modest cost structure variability across index style. Strategies facing more significant implementation hurdles (e.g., stocks and options) generally make greater use of the trading cost and in/out fee levers.

Profitability and negotiability are related elements. Banks participate in the tradable index space to earn a profit. To do so, they must include charges in excess of the cost of delivering the return stream promised in the index rule book. The levers used by a given bank will depend upon index execution realities and the business model of the bank, including messaging and competitor considerations. Profitability considerations also lead to fungibility among cost levers. Transaction costs could be relatively aggressive if this is the only cost component or potentially less so in the presence of an index fee providing an additional profit buffer. If transaction costs do not appear explicitly, they exist implicitly within another cost item. Bottom-line profitability, and not any single cost component, is most important to a bank.

Profitability is an aggregate business consideration for banks, which leads to a willingness (like any asset manager) to negotiate fees. Reducing the profit margin on a given index may be acceptable to support the broader business. Perhaps the investor is a large institution with the potential to invest in additional indices or with a large existing trading relationship with the bank. Perhaps the index is a new release and the bank is seeking early adopters. Perhaps the bank is attempting to retain assets in an underperforming index. Perhaps the bank is trying to grab market share from a competitor. The important takeaway is that all costs are negotiable. The costs in this database generally represent rack rates, list prices, or pre-negotiation levels. Effective or executed cost information is closely held and, therefore, not a tenable database item.

Analysis in the tradable index space focuses upon reported index returns. Because investment banks charge varying combinations of internal and external costs, comparing reported index returns can be a bit like comparing book value across geographies, balancing the consequence of inconsistency with the benefit of convenience. Reported index returns exist along the gross to net return spectrum. This is a first moment issue, meaning average return (or Sharpe ratio) comparisons represent the concern. Covariance and higher moment comparisons do not encounter the same problem given the nature of the cost data. While the database provides a conservative estimate of tradable index costs, the comprehensive information enables one to create a consistent set of returns and to frame reported numbers. Given the situation-specific nature of discounts, the reader then can ponder appropriate reductions to these ranges.

A final consideration regarding costs is the vintage effect. Fee pressure is unrelenting in the investment management industry and tradeable indices are not immune to this. Evolution in financial markets, changes in competitive positioning, infrastructure upgrades, and accumulated experience guaranteeing spreads may reduce costs over time.

As a result, some older database entries (e.g., graveyard indices) could reflect a prior cost regime. Banks review all data items during each update so cost datedness should not be a significant issue, but the possibility exists for some indices.

Figure 5 summarizes the costs in the database. The first panel presents headline costs (index fee and swap spread) by index style. For indices reporting this cost, the median is 40 basis points with 90% of observations between 15 and 100 basis points. The median headline cost does vary by style, with merger arb at the high end and reversal, size and congestion at the low end.

The second panel presents in/out costs. The median charge is 5 basis points with 90% of observations between 2 and 45 basis points. This median is relatively constant across styles, excepting volatility strategies which levy transaction charges several times higher than those of other strategies. Investment banks have a comparative advantage in volatility trading infrastructure over all but the largest or most focused buy-side firms, certainly a reason for the broad range of index offerings in a space with significant implementation hurdles. Note too that in/out costs may apply only to trades on non-roll dates for an index or may be higher on these off-cycle dates. Therefore, the possibility exists that banks report these charges inconsistently, with some reporting lower numbers for standard execution dates and others reporting higher costs representing break fees and non-standard trade date penalties. This potential inconsistency is an index level consideration and not a bias concern in index aggregations.

The third panel highlights trading costs. Across the universe reporting trading cost, the median cost is 95 basis points with 90% of observations between 4 and 685 basis points. Volatility strategies significantly influence these findings due to the high execution cost of volatility positions, commonly exacerbated by the impact of delta hedging and leverage. Excluding volatility, the median cost is 54 basis points with 90% of observations between 4 and 260 basis points. Among non-volatility index styles, reversal strategies, given relatively high trading volume, are an outlier to the high end of the execution cost spectrum while risk anomaly and size reside on the low end.

The fourth panel shows total costs, the summation of all four components. Unlike the three preceding charts, this one includes the entire index universe because all indices report some form of cost. The median total cost is 98 basis points with 90% of observations between 27 and 561 basis points. For the non-volatility universe, the median total cost is 80 basis points with 90% of observations between 24 and 246 basis points. Among the non-volatility index styles, carry and congestion strategies are on average relatively less expensive while merger arb and reversal bear higher costs.

The fifth panel shows total costs adjusted by strategy volatility. Considering costs per unit of volatility is a simple way to standardize costs incurred in the pursuit of index leverage and greater return generating power. Total cost represents on average 19% of index volatility, with 90% of observations between 4 and 85%. For the non-volatility universe, the median is 15% with 90% of indices between 4 and 43%. Standardization does reduce the gap between volatility and other strategies but does not change the fact that the former bears the greatest cost burden. The scaling also changes the relative cost standing of numerous index styles. Congestion, a relatively low volatility strategy, is on average among the most expensive indices on a standardized basis -- a stark contrast to its low-cost profile in panel four. Conversely, trend is a relatively low-cost strategy when considered relative to index volatility. Merger arb remains relatively costly.

As mentioned previously, the cost data permits both the calculation of consistent index returns and the framing of reported returns in a space characterized by a variety of cost structures. Net returns subtract the estimated external cost (swap spread and in/out charges) from reported returns while gross returns add the estimated internal cost (index fee and trading costs) to reported returns. The former provides a conservative representation of the return experience of end investors while the latter indicates raw strategy performance -- a unique element of this paper. To manage outliers, the data curation process caps internal and external costs such that total cost does not exceed 10%. This winsorization affects 2.5% of the index universe, primarily volatility strategies.

1.6 Composite Strategy Benchmark Design

1.6.1 Structural Considerations

Three decisions underpin benchmark construction: universe definition, pruning criteria, and constituent weighting. Universe definition is a function of classification and focus preferences. For example, an equity benchmark methodology relies upon a certain industry classification scheme for stocks and may be relatively narrow (e.g., sector specific) or broad. Pruning criteria eliminate members of the candidate constituent pool that violate certain conditions. For example, an equity benchmark might apply a liquidity or public ownership hurdle to each stock. Finally, a methodology for weighting the returns of index members determines benchmark performance. Equity benchmarks conventionally rely upon capitalization weighting, but much has been written on alternative weighting approaches – equal weighting (1/n or minimum HHI), fundamental weighting, volatility weighting, equal contribution-to-risk, minimum variance, maximum diversification, and the efficient weight method. Amenc *et al*. (2011), Choueifaty and Coignard (2008), and Clarke *et al*. (2013) discuss many of these methods.

This paper approaches the ARP benchmarking problem with three objectives.

1. Represent comprehensively the competing strategy specifications traded by ARP investors to provide a strategy spanning approach as part of the performance evaluation mosaic.
2. Facilitate the triangulation exercise required for ARP performance evaluation due to the absence of canonical strategy specifications.
3. Provide tiered ARP performance perspective to address performance questions ranging from narrow to broad.

These objectives have direct implications for the pruning criteria, favoring more rather than less accommodative strategy exclusion policies. For example, one might discard strategies due to methodological complexity related to signal generation and/or portfolio construction. Such moderately aggressive pruning is inherently subjective, so multiple equally defensible approaches exist, but may be consistent with benchmark objectives and structures different than those underpinning this exercise. Similarly, primitive strategies like the recently released Bloomberg GSAM Risk Premia Indices represent extreme pruning, selecting a single, basic strategy specification as the benchmark. Of course, the lack of theoretical basis and text-book definitions for trading strategies means that a range of specifications could warrant this singular distinction. A primitive strategy is no panacea, but rather a piece of the ARP performance puzzle attempting to address a specific set of questions.

1.6.2 Benchmark Classification and Pruning

This paper uses the proprietary database to introduce two ARP composite strategy benchmark families. The first is a three-tiered, categorical structure. The most granular tier stratifies the tradable bank index universe along previously defined style, asset class and regional dimensions. Value Equity (stock-based) in North America and Carry (curve) Rates in Europe are two examples of this approach. Pruning is relatively light, with a minimum requirement of five indices per benchmark yielding 85 foundational, asset-style-region benchmarks representing approximately 90% of the index population. These benchmarks roll up to 46 asset-style, 14 style and 7 asset benchmarks. (Because style and asset are non-nested aggregations, the categorical approach has four benchmark groups but only three nested tiers.) Appendix A details the categorical benchmark family. This approach has the benefit of simplicity and intuitive appeal, leveraging transparent benchmark inclusion logic and a robust taxonomy proposal in a space with no standard strategy classification system.

The second ARP benchmark family is a four-tiered, fully-nested statistical structure. This approach eschews metadata and focuses entirely on the return structure of bank indices to establish foundational benchmarks. Working exclusively with index returns requires decisions regarding missing data treatment, clustering and pruning methodologies.

1.6.2.1 Data Imputation

Determining the treatment of missing data is a prerequisite for statistical classification. The scattered nature of missing ARP data precludes a simple elimination (complete case) strategy. Given the relatively limited return history, maximizing use of available data is important, subject to imputed returns representing a reasonable proportion of the overall data set. Considering the nature of the missing data and the imputation process, the 9-to-1 actual-imputed data ratio associated with a December 2004 start date strikes an appropriately conservative balance. Figure 6 illustrates the ARP missing data problem.

Rubin (1976) emphasizes the need first to understand the process resulting in missing data. The choice of imputation method should consider whether data is missing completely at random (MCAR, missingness unrelated to observed or missing values), missing at random (MAR, missingness may be related to some factor but not to the missing value), or not missing at random (NMAR, missingness could be a function of the missing value). Missing ARP returns fall in the MAR category. Underlying input availability, not the strategy returns, predominantly dictates missingness. While poor performance could result in an index being discontinued, this represents a small proportion of the missing data and is a function of observed data, not the missing data.

A variety of methods, falling broadly into deletion and imputation approaches, exist to deal with MAR data. The overarching reality is that no dominant approach exists -- the method must fit the situation. Deletion is straightforward but discards useful information and may bias subsequent analysis. As Figure 6 indicates, this paper drops 24% of the available weeks in the database (2000-2004) but only 13% of the available returns given the index start date distribution in Figure 3. The imputation requirement increases non-linearly by dropping fewer weeks, so the December 2004 start date balances the marginal benefit and cost of additional data retention.

Imputation methods fall generally into discrete and predictive categories. The former includes mean, mode, forward-fill, and linear interpolation. These approaches are easy to implement but underestimate variance and ignore correlation. Predictive imputation methods include expectation-maximization (EM), K-nearest neighbor (KNN), ordinary least squares (OLS), singular value decomposition (SVD), and support vector regression (SVR). See Bertsimas *et al.* (2018) and Molenberghs *et al.* (2015) for discussions of the many predictive alternatives.

A final consideration relates to single versus multiple imputation. The former replaces missing data to form a single, complete data set. The latter passes numerous complete data sets to the analysis stage, the results of which then must be pooled. Multivariate imputation by chained equations (MICE) is one example. Multiple imputation has the benefit of capturing imputation uncertainty but may not apply to every situation. Pooling of results may be problematic or the marginal improvement in results versus single imputation may be small relative to the incremental computation time.

This paper employs a stochastic regression blend for data imputation. Specifically, the process combines recursive estimates from elastic net and partial least squares, each including a scaled disturbance term. As mentioned previously, context dictates the appropriate treatment of missing data and several unique considerations apply to the ARP data set.

- The return data supports a classification exercise, the clustering of indices into benchmark groups within a tree structure. The covariance, not the mean or variance, of missing data is the focus.
- For most indices, missing data comprises a small proportion of available data, so observable data drives covariance estimates.
- Peer groups exist within the index universe, meaning that index offerings from competitors in a given strategy often anchor missing data predictions, in the process conveying strategy-specific non-normality.
- The ARP data set for imputation includes a maximum of 817 weekly observations for 1,932 bank indices and 104 supplemental regressors from the Bloomberg GSAM Risk Premia Indices, Fama-French Factor Library, Barra Global Equity Model, and the various market indices listed in Appendix B. This creates a dimensionality problem for OLS, as the number of regressors exceeds the number of observations. OLS does not produce a unique solution.

The data imputation process proceeds through the index universe, from least missingness to most, combining a dimension reduction method (partial least squares) and a shrinkage method (elastic net). The intent is to balance the bias-variance tradeoff for missing data estimates in a setting with many possible regressors, subsets of which are likely to be highly correlated.

Partial least squares (PLS) represents a union between principal component analysis (PCA) and OLS, achieving dimension reduction by collapsing the regressors into a specified number of principal components. Dimension reduction and regression occur simultaneously. PLS is a supervised alternative to principal component regression (PCR), since latent factor design targets high covariance with the regressand. Boulesteix and Strimmer (2006), Frank and Friedman (1993), and De Jong (1993) provide helpful perspective on applying PLS to high-dimension data.

Equation 2 provides the PLS objective function and *Equation 3* the p-by-1 PLS regressor loading vector, $\boldsymbol{\beta_P}$. $\boldsymbol{X}$ is the n-by-p (centered) matrix of independent variables. $\boldsymbol{Y}$ is the n-by-1 (centered) response variable. $\boldsymbol{W}$ is the p-by-c matrix of loadings transforming $\boldsymbol{X}$ into an n-by-c matrix of latent components ($\boldsymbol{T}$) -- with $\boldsymbol{T}$ equal to the product of $\boldsymbol{X}$ and $\boldsymbol{W}$, c representing the number of latent components, and $\boldsymbol{w_i}$ being a column of $\mathbf{W}$.

$$\boldsymbol{w}_i = \underset{w}{argmax}\ \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{w} \qquad \textit{Equation 2}$$

$$for\ i = 1, \dots, c\ and\ j = 1, \dots, i-1$$

$$s.t.\ \boldsymbol{w}_i^T\boldsymbol{w}_i = 1\ and\ \boldsymbol{w}_i^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}_j = 0$$

$$\widehat{\boldsymbol{\beta_P}} = \boldsymbol{W}(\boldsymbol{T}^T\boldsymbol{T})^{-1}\boldsymbol{T}^T\boldsymbol{Y} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{X}^T\boldsymbol{Y} \qquad \textit{Equation 3}$$

The data imputation process sets c equal to 10 and uses 10-fold cross-validation to compute the mean squared error (MSE).[2] $\boldsymbol{Y_P}$ in *Equation 4* is an m-by-1 vector of missing data estimates. $\boldsymbol{X_m}$ is the m-by-p matrix of regressors. $\boldsymbol{\varepsilon_P}$ is the PLS stochastic disturbance -- an m-by-1 zero-mean normally distributed error vector, orthogonalized to $\boldsymbol{X_m \beta_P}$ with a variance proportional to the explanatory power of $\boldsymbol{X\beta_P}$.

$$\widehat{\boldsymbol{Y}_P} = \boldsymbol{X}_m \widehat{\boldsymbol{\beta}_P} + \boldsymbol{\varepsilon}_P \qquad \textit{Equation 4}$$

Elastic net (EN) is a coefficient shrinkage or regularization technique combining least absolute shrinkage and selection operator (LASSO) regression and ridge regression (RR). LASSO solves the L_1-norm penalized OLS problem, with the penalty being the sum of the absolute regressor loadings. RR solves the L_2-norm penalized OLS problem, with the penalty being the sum of the squared regressor loadings. As a result, LASSO is a shrinkage and factor selection technique while RR is solely a shrinkage technique. RR balances the loadings on correlated regressors whereas LASSO may select a single representative. LASSO emphasizes the most compelling regressors whereas RR may temper the importance. EN attempts to balance the benefits and drawbacks of LASSO and RR. Zou and Hastie (2005), Tibshirani (2011), and Waldmann *et al.* (2013) discuss the merits of the approach.

---

[2] The choice of 10 components balances parsimony and explanatory power across a diverse range of ARP strategies. For example, changing *c* from 20 to 10 reduces the median strategy R^2 by a modest 6%. This reduction grows to 9% and 16% respectively for the 5th and 25th percentile strategy. A smaller *c* becomes overly punitive on the more heterogenous part of the ARP universe.

Equation 5 provides the EN objective function. $\boldsymbol{\beta_E}$ represents the 1-by-p EN loadings on the independent variables, $\boldsymbol{X}$, to predict $\boldsymbol{Y}$ – respectively, n-by-p and n-by-1 matrices. N is the number of observations and p the number of predictors. α is a mixing parameter between 0 and 1 and λ is a non-negative regularization (penalty) parameter. EN approaches LASSO for α equal to 1 and is equivalent to RR for □ equal to 0. This paper uses a value of 0.5.[3]

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{E}} = \underset{\beta_0, \beta}{argmin} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \boldsymbol{\beta} \boldsymbol{x}_i^T)^2 + \lambda \sum_{j=1}^{p} \left(\frac{(1-\alpha)}{2} \boldsymbol{\beta}_j^2 + \alpha |\boldsymbol{\beta}_j| \right) \quad \text{Equation 5}$$

The data imputation process sets the maximum number of predictors to 50 and uses 10-fold cross-validation to compute the mean squared error (MSE).[4] $\boldsymbol{Y_E}$ in *Equation 6* is an m-by-1 vector of missing data estimates. $\boldsymbol{X_m}$ is the m-by-p matrix of regressors. $\boldsymbol{\varepsilon_E}$ is the EN stochastic disturbance -- an m-by-1 zero-mean normally distributed error vector, orthogonalized to $\boldsymbol{X_m \beta_E}$ with a variance proportional to the explanatory power of $\boldsymbol{X \beta_E}$.

$$\widehat{\boldsymbol{Y}}_E = \boldsymbol{X}_m \widehat{\boldsymbol{\beta}_E} + \boldsymbol{\varepsilon}_E \quad \text{Equation 6}$$

[3] An α of 0.5 is standard in EN applications.

[4] The existence of both distinct cohorts in a large data set and specification variance within those cohorts justifies a degree of inclusiveness regarding the number of predictors. Setting p equal to 50 represents the point at which this constraint is binding for 10% of strategy fits, an indication the parameter choice is becoming overly restrictive. As additional support for the parameter choice, increasing p to 75 produces no meaningful improvement in overall explanatory power.

The data imputation process is recursive, repeating the above steps and averaging the results to manage the path-specific nature of each iteration. As the final, complete set of returns supports a strategy classification exercise, multiple imputation is not additive in this context. Pooling entire classification structures is challenging (e.g., a matching matrix approach has a sparse distributional profile), and the combination of focus on strategy association and relatively small proportion of missing data limits the potential benefit.

1.6.2.2 Strategy Classification

Because classifying ARP strategies is an unsupervised learning problem (i.e. no training set of "correct" benchmark groups exists), cluster analysis is the choice to obtain a statistical taxonomy. The objective is to separate strategies into benchmark groups (clusters) that minimize the intergroup returns-based similarity while maximizing the intragroup similarity. Of the two most common types of clustering algorithms, partitional and hierarchical, hierarchical aligns best with the current objectives. Partitional methods (k-means, k-medoids) produce a single set of non-nested clusters that is inconsistent with a benchmark structure including subgroups. Of the two types of hierarchical clustering algorithms, agglomerative and divisive, this paper employs the more frequently used agglomerative method to classify ARP strategies. This is a bottom-up process, starting with N singleton clusters and successively merging the two most similar clusters according to the inter-cluster distance measure (or linkage method) until only one cluster exists. Alternatively, divisive clustering is a top-down approach, starting with the aggregate data set and iteratively splitting the groups into the next, least similar cluster until reaching a stopping criterion (or N singleton clusters remain).

The agglomerative hierarchical clustering algorithm used in this paper works on a return matrix that is double-standardized (time-series and cross-sectional) and winsorized to manage large outliers. In *Equation 7*, $\boldsymbol{Z}$ is an n-by-p matrix of time-series z-scores, with $\boldsymbol{X}$ representing the matrix of underlying data, $\boldsymbol{\mu_X}$ the 1-by-p mean row vector and $\boldsymbol{\sigma_X}$ the standard deviation row vector. In *Equation 8*, z_{lim} is the absolute maximum z-score for the purpose of winsorization, which occurs asymptotically between z_{lim} and (z_{lim} – 1). z_{lim} is set to 3.5. $\boldsymbol{I_a}$ is an indicator matrix equal to 1 if $|\boldsymbol{Z}| \geq (z_{lim} - 1)$ and 0 otherwise. $\boldsymbol{Z_w}$ in *Equation 9* is the winsorized time-series z-score matrix. The cross-sectional z-score calculation works identically, operating latitudinally on $\boldsymbol{Z_w}$.

$$\boldsymbol{Z} = (\boldsymbol{X} - \boldsymbol{\mu}_X) \bullet \frac{1}{\boldsymbol{\sigma}_X} \qquad \textit{Equation 7}$$

$$\boldsymbol{Z}_a = sgn(\boldsymbol{Z})(z_{lim} - 1) + tanh[\boldsymbol{Z} - sgn(\boldsymbol{Z})(z_{lim} - 1)] \qquad \textit{Equation 8}$$

$$\boldsymbol{Z}_w = \boldsymbol{I}_a \bullet \boldsymbol{Z}_a + (1 - \boldsymbol{I}_a) \bullet \boldsymbol{Z} \qquad \textit{Equation 9}$$

The cophenetic correlation coefficient, c, measures the extent to which a classification reflects the original data and can be helpful in selecting the linkage method, or algorithm calculating the distance between clusters. Specifically, c is the linear correlation coefficient between the dendrogrammatic distance and the pairwise distance in the underlying data. x_{ij} is the distance between observations i and j in the original data.

$\bar{x}$ is the average of x_{ij}. z_{ij} is the distance between points i and j in the dendogram. $\bar{z}$ is the average of z_{ij}.

$$c = \frac{\sum_{i<j}(x_{ij} - \bar{x})(z_{ij} - \bar{z})}{\sqrt{\sum_{i<j}(x_{ij} - \bar{x})^2 \sum_{i<j}(z_{ij} - \bar{z})^2}}$$

Equation 10

As with determining the appropriate data imputation approach, the choice of linkage method is data and objective dependent. In this paper, Ward's method, Ward (1963), determines the distance between clusters. This approach manages the merging cost of combining clusters, limiting the marginal increase in the sum of squared deviations from the cluster mean. Numerous studies cite the usefulness of Ward's method.[5] The tendency of this approach to produce a relatively balanced distribution of cluster sizes aligns well with the benchmarking exercise. Some methods (e.g., the average method) may deliver a higher cophenetic correlation but do so via a highly skewed distribution of cluster sizes. Such emphasis on a small subset of strategies has less intuitive appeal and less applicability to this universe classification exercise. Hence, the choice of linkage method is not simply a cophenetic correlation maximization exercise.

This classification process yields a four-tiered statistical structure. To be comparable with the categorical benchmark family, 85 base groupings exist. The colors in Figure 7 summarize the base benchmark composition. The base benchmarks roll up into 40 super-base, 20 hypo-broad and 10 broad benchmarks. Consistent with the tree structure, all

[5] See Kuiper and Fisher (1975), Blashfield (1976), Hands and Everitt (1987), Milligan and Cooper (1988), and Ferreira and Hitchcock (2009).

benchmarks are fully nested. The colors in Figure 8 depict the broad benchmarks. Appendix C details the statistical benchmark family. This statistical approach to benchmarking has the appeal of data rationality, trading the transparency and descriptive convenience of the categorical approach for tighter performance-based alignment of benchmark constituents.

1.6.2.3 Strategy Pruning

Cluster analysis does not yield the optimal number of clusters and includes every member of the ARP universe. Hierarchical clustering provides N – 1 groupings of the underlying data, leaving the user to determine the appropriate pruning of the tree structure. Numerous authors propose statistics and decision-making heuristics to aid in this decision.[6] Ultimately, however, the pruning strategy must align with the purpose of the analysis.

This paper utilizes PCA to govern the pruning process. PCA is a dimension reduction technique identifying latent factors (principal components) of decreasing variance that preserve the total variance of the underlying data. PCA solves *Equation 11*, with $\boldsymbol{v}$ being the eigenvector of the underlying n-by-n covariance matrix $\boldsymbol{\Sigma}$ and λ the associated eigenvalue (scalar). The first principal component is the linear combination of the original data, with the eigenvector $\boldsymbol{v_1}$ providing the weights, that explains the maximum variance (eigenvalue λ_1) among all linear combinations. The n principal components are

[6] See, among others, Thorndike (1953), Calinski and Harabasz (1973), Davies and Bouldin (1979), Rouseeuw (1987), Krzanowski and Lai (1988), and Tibshirani *et al.* (2001)

orthogonal to one another, with each successively explaining as much of the remaining variance as possible.

$$\boldsymbol{\Sigma v} - \lambda \boldsymbol{v} = 0 \qquad \textit{Equation 11}$$

Because the principal components are orthogonal, ω_{ij} in *Equation 12* represents the percent of strategy j variance explained by principal component i. $\boldsymbol{S}$ is the n-by-p matrix of index returns comprising a benchmark group and $\boldsymbol{v}$ is an eigenvector.

$$\omega_{ij} = corr\left(\boldsymbol{S}\boldsymbol{v}_i, \boldsymbol{S}_j\right)^2 \qquad \textit{Equation 12}$$

To facilitate comparison, the pruning algorithm begins with the same number of base benchmarks as the categorical ARP benchmark family (85). The algorithm then establishes a ω threshold, $\boldsymbol{PCT_b}$, for each benchmark group, b, and drops strategies that do not clear the threshold. $\boldsymbol{PCT_b}$ is a 3-by-1 vector containing the thresholds for the proportion of variance explained by the first three principal components. The exponentially weighted hurdles in *Equation 13* ensure that three (h) components explain at least 50% of the total variance of each surviving strategy, with the first component explaining the largest proportion.[7] The thresholds apply a higher acceptability standard

[7] Pruning is not particularly sensitive to alternative specifications of h, indicating the clusters are reasonably tight. Setting h equal to 3 requires the first three principal components to explain at least 50% of strategy variance and eliminates 247 strategies in the first pass of the pruning algorithm. Fixing h at 2 (4) requires the first three components to explain at least 40% (65%) of variance and eliminates 212 (346) strategies.

to the first three components of a benchmark with five members (n) than one with 20 constituents.

$$PCT_b = \begin{bmatrix} \frac{1-\Upsilon}{1-\Upsilon^n} \\ \frac{\Upsilon-\Upsilon^2}{1-\Upsilon^n} \\ \frac{\Upsilon^2-\Upsilon^3}{1-\Upsilon^n} \end{bmatrix} \qquad \textit{Equation 13}$$

$$where \quad \Upsilon = 0.5^{\frac{1}{h}}$$

The algorithm proceeds iteratively (thresholds change as the benchmark constituent count changes) until all remaining strategies clear the threshold. The first pass evaluates the first principal component, the sum of the first two components, and the sum of the first three components relative to the corresponding threshold. Subsequent passes focus upon the first principal component. To be consistent with the categorical benchmarks, a minimum of five strategies per base benchmark applies.

As with the categorical family, pruning of the base statistical benchmarks is relatively minor, with the process retaining approximately 85% of the index universe. As discussed previously, more aggressive pruning is not the objective of this broad-based benchmarking exercise.

1.6.3 Benchmark Weighting

The unique nature of the ARP space narrows the weighting alternatives. Capitalization and fundamental weighting approaches do not apply. The fungible nature of tradeable bank indices, that leverage is available within the swap structure or via the notional allocation to the swap, necessitates a risk-based approach to weighting. A simple, equally weighted approach would produce representation disparities within a benchmark due to differences in baseline leverage across ARP indices. Among the risk-based alternatives, minimum variance and maximum diversification tend to exclude some constituents and to concentrate in others. This runs counter to the objective of balanced representation within the ARP benchmarks and arguably relies too heavily on covariance estimates combining live and simulated returns.

This paper applies a volatility weighting scheme to calculate benchmark returns, implying that each tradeable bank index has the same standard deviation. In this context, volatility and 1/n weighting produce the same result. Volatility and equal contribution-to-risk also produce the same result, assuming all correlations are identical – essentially, a shrinkage approach for cohorts of strategies represented by different blends of simulated and live returns.

This paper scales all indices and benchmarks to a 7% annual volatility, the median strategy volatility across the entire database of tradeable bank indices. (Asset managers generally offer ARP portfolios trading in the 6-10% volatility range.) Because the purpose of this adjustment is to facilitate comparison and to address structural (leverage) inconsistency among bank indices, the volatility estimate should be slow moving. With only 20 years of return data, using, for example, a 5-year trailing volatility estimate sidelines a material amount of the early data for start-up purposes and produces results comparable to scaling by the full-sample volatility, so this paper adopts the latter convention in the interest of clarity and data presentation.

Utilizing a dynamic, ex ante volatility estimate for scaling purposes potentially moves the benchmark toward a constant volatility profile but also introduces estimation noise and benchmark turnover considerations into a setting already rife with strategy specification and classification variation. Optimizing the volatility estimation window represents a possible area for future investigation.

A final ARP benchmark weight consideration relates to strategy redundancy. As discussed previously, the data curation process includes a step to eliminate clearly duplicative indices within the same strategy family at a given bank. The survey asks banks to provide only a single representative for each strategy family, but some banks are more inclusive than others when defining a unique strategy.

To balance the risk of strategy overrepresentation in a benchmark with the reality that some banks offer more than one distinct strategy family within a given ARP style, this paper calculates the weights for individual bank indices using *Equation 14*. $\boldsymbol{w}$ is a 1-by-N weight vector. N is the total number of strategies. $\boldsymbol{I_b}$ is a B-by-N indicator matrix of bank ownership (1 if a given bank owns strategy i and 0 otherwise). B is the total number of banks. $\boldsymbol{I_1}$ is a B-by-N matrix of ones, and s is a shrinkage factor (0 to 1) capturing the degree of concern regarding overrepresentation. This paper sets s equal to 0.5.[8] $\boldsymbol{\sigma_a}$ is a

[8] s is essentially a confidence parameter. Absolute conviction in intra-bank strategy methodological independence (overlap) warrants setting s equal to zero (one). Since the data pre-processing specifically targets redundancy, very large values of s are irrelevant. Very small values for s either have an immaterial impact or overly discount the likelihood of information sharing within a bank. Therefore, a practical range for s is 0.3 to 0.6, and this paper applies 0.5 in the interest of conservatism.

1-by-N volatility adjustment vector containing the ratio of target volatility to strategy volatility. With redundancy risk addressed at the base benchmark level, combining ARP benchmarks is straightforward (*Equation 15*). $\boldsymbol{w_g}$ is a benchmark aggregation weight vector, with M indicating the total number of benchmarks being combined in benchmark group, g. $\boldsymbol{\sigma_b}$ is a 1-by-M volatility adjustment vector containing the ratio of target volatility to constituent benchmark volatility, represented by the full-period standard deviation of benchmark returns.

$$\boldsymbol{w} = \boldsymbol{\sigma}_a \bullet \left(\frac{s}{B(\boldsymbol{I}_b \boldsymbol{I}_1^T)^T \boldsymbol{I}_b} + \frac{1-s}{N} \right) \qquad \textit{Equation 14}$$

$$\boldsymbol{w}_g = \boldsymbol{\sigma}_b \frac{1}{M} \qquad \textit{Equation 15}$$

The benchmark weights are set to zero if fewer than three strategies are available for a given date. While each benchmark group includes a minimum of five strategies, the underlying indices may have different start and end dates.

1.7 Categorical and Statistical Benchmark Comparison

The categorical benchmark provides a comparative base for the statistical benchmark. This section compares the two approaches on three dimensions – structure, performance and factor sensitivity. The findings underscore the appeal of the statistical approach and punctuate the general challenges in ARP benchmarking.

1.7.1 Benchmark Structure Comparison

Several authors propose metrics for comparing the similarity of two groups of classifications, notably Sokal and Michener (1958), Rand (1971), Fowlkes and Mallows (1983), Hubert and Arabie (1985), Warrens (2008), and Morlini and Zani (2010). While most of these authors focus upon comparing the results of two hierarchical procedures at a specific pruning point, Morlini and Zani (2010) consider comparing two dendrograms. Following the approach of Morlini and Zani (2010), Table 6 provides a set of measures comparing the assignment of each of the 1,932 tradable bank indices to the 85 base benchmarks in the categorical and statistical families.

The similarity indices of Rand (1971), Fowlkes and Mallows (1983) and Morlini and Zani (2010) all leverage two sources of information, a binary matrix for each classification family identifying every pairing of tradable indices in the same base group and a matching matrix cross-tabulating the pairs in the categorical and statistical base groups. This approach is attractive because working with tradable index pairs eliminates the need for benchmark labels – a numerical identifier suffices for each benchmark. *Equation 16* provides the calculation of the indices in Table 6, with k representing the number of clusters (85 categorical, c, and statistical, s, base benchmarks) and n indicating the number of underlying data items (1,932 strategies).

$$Morlini \,\&\, Zani\ Index = 1 - \frac{P_k + Q_k - 2T_k}{P_k + Q_k} \qquad \text{Equation 16}$$

$$Adjusted\ Morlini\ \&\ Zani\ Index = \frac{2T_k - \frac{2P_kQ_k}{N}}{P_k + Q_k - \frac{2P_kQ_k}{N}}$$ *Equation 16a*

$$Fowlkes\ \&\ Mallows\ Index = \frac{T_k}{\sqrt{P_kQ_k}}$$ *Equation 16b*

$$Rand\ Index = \frac{N - P_k - Q_k + 2T_k}{N}$$ *Equation 16c*

Where…

Matching Matrix $\mathbf{M_k}$

	1	…	s	…	k	Total
1	m_{11}	…	m_{1s}	…	m_{1k}	$m_{1\sim}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
c	m_{c1}	…	m_{cs}	…	m_{ck}	$m_{c\sim}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	m_{k1}	…	m_{ks}	…	m_{kk}	$m_{k\sim}$
Total	$m_{\sim 1}$	…	$m_{\sim s}$	…	$m_{\sim k}$	n

$$T_k = \frac{1}{2}\left[\sum_{c=1}^{k}\sum_{s=1}^{k} m_{cs}^2 - n\right]$$

$$P_k = \frac{1}{2}\left[\sum_{c=1}^{k} m_{c\sim}^2 - n\right]$$

$$Q_k = \frac{1}{2}\left[\sum_{s=1}^{k} m_{\sim s}^2 - n\right] \quad and \quad N = \frac{n(n-1)}{2}$$

The similarity indices, which exist on a unit scale, reveal a significant number of tradable index assignment differences between the two classification schemes. The Rand index value is comparatively high in this setting because, unlike the competing methodologies, its calculation deems pairs not linked within either group to be indicative of similarity. With 85 base benchmarks parsing the sizable bank index universe, this cohort is very large. The important implication for ARP benchmarking is that the statistical approach to strategy taxonomy represents a material departure from conventional categorical classification.

Comparing Appendix A and C unveils the factors driving these classification differences. The statistical taxonomy recognizes the diversification inherent within commodities, for example, distinguishing short volatility strategies in precious metals from energy. The statistical approach also frequently acknowledges methodological differences within a categorical base group, for instance, splitting North American risk anomaly stock strategies into three sub-classes. Finally, the statistical classification consolidates regional strategies with a common factor footprint (e.g., North American, European, and multi-region credit time series trend strategies).

By design, the statistical approach delivers greater homogeneity within base benchmark groups than the categorical alternative. Table 7 reveals that the combination of hierarchical clustering and PCA-based pruning increases the variance explained by the first principal component by 43% versus the categorical approach. The first principal component explains at least 40% of return variation for 83% of statistical benchmark constituents. For half of these constituents, variance explained exceeds 65%. This result is simultaneously validating and indicative of the lack of uniformity among closely related ARP strategies.

Understanding the lack of uniformity begins with the correlation heatmap in Figure 9. The matrix is sorted to highlight highest correlation between categorical and statistical base benchmarks along the diagonal. Ostensibly, many categorical benchmarks have a clear statistical analog while the off-diagonal correlations are predominantly low to very low. This is not surprising since the intuitive appeal of the categorical groupings exists for a reason. However, correlation paints an incomplete picture for benchmarks ultimately used in performance evaluation. Tracking error is an important consideration, particularly when comparing ARP benchmarks.

The tracking error (*TE*) between two benchmarks, 1 and 2 (*Equation 17*), is a function of the associated standard deviations (σ) and correlation (ρ). *TE* increases nonlinearly with a reduction in correlation. Figure 10 illustrates this dynamic as it applies the base benchmark comparison. The tracking risk between the highest correlation categorical and statistical base benchmarks generally represents 35% to 75% of benchmark volatility. For context, comparing a traditional US versus Europe equity benchmark or EM bond versus US high yield bond benchmark produces tracking within this range. Yes, the benchmarks are highly correlated due to common risk factors; however, using a European benchmark to evaluate a US equity manager or a US high yield benchmark for an EM bond manager obviously would create performance assessment problems. Such variation is consequential in the context of assessing the active contribution of an ARP asset manager, particularly in the absence of de facto style definitions. Selecting the most defensible benchmark methodologies and triangulating performance evaluation is the prudent response.

$$TE_{12} = \sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho_{12}\sigma_1\sigma_2} \qquad \text{Equation 17}$$

1.7.2 Benchmark Performance Comparison

The objective, particularly in the context of simulated returns, is not to recommend a benchmark methodology based upon performance – either as a fast or slow rabbit. The structural underpinnings and accompanying rationale are far more important. Understanding the performance history is necessary to appreciate the consequences of the ARP data nuances and alternative benchmark methodologies. Given the challenge of summarizing performance for the large number of benchmarks defined in the preceding pages, this section focuses upon the top-level benchmarks. The observations apply to all benchmark tiers.

Figure 11 displays the three-year rolling Sharpe ratio for the 14 categorical style benchmarks. This chart structure enables the discussion of three points. First, the shaded area incorporates the gross and net tradable index returns introduced in section 5. Recall that reported bank index returns treat costs inconsistently. This paper is the first to contextualize reported ARP returns in this fashion. Figure 11 illustrates the significant costs embedded in reported index returns for volatility and multi-style benchmarks, in contrast to more modest costs in benchmarks for size and trend. Transaction costs account for most of the gap between reported and gross performance and, because these costs vary only to a limited degree over time, indicate the relative vulnerability of a style to signal efficacy decay.

Second, Figure 11 highlights a general decline over time in the performance of ARP strategies. Carry(spread), merger arbitrage, multi-style, other, risk anomaly, size, trend (cross-sectional), value and volatility (short) all post a negative reported Sharpe ratio over the past three years. On a net basis, the Sharpe ratio is negative for reversal and volatility (arbitrage) and nil for carry (curve) and trend (time-series). These results mark a significant departure from previous, primarily back-tested results. The extent to which this deterioration is a byproduct of crowding, overfitting, environmental factors, or some other influence is beyond the scope of this paper but represents an important and fertile area for future research. The statistical benchmark family introduced here will facilitate investigation of the extent to which ARP is facing temporary or structural headwinds.

Finally, the ARP style benchmarks register very high historical three-year Sharpe ratios, with the vast majority exceeding one, often by a wide margin. Back-tested returns undoubtedly fuel this result. Aggregating a collection of strategy returns with the benefit of hindsight ensures an inflated Sharpe ratio. However, the Sharpe ratio associated with underlying strategies typically is much lower than that of the benchmark. The combination of equal weighting within benchmarks and modest correlation among strategies also contributes to this inflation dynamic.

Figure 12 highlights, for a fixed and comparatively modest constituent Sharpe ratio of 0.5, the nonlinear impact of relatively low correlation among ARP strategies on the benchmark-level Sharpe ratio. The median interquartile range of correlations among base benchmarks within each style benchmark is 0.0 to 0.4, well within the region of significant Sharpe ratio benefit from linear summation in the numerator outstripping nonlinear aggregation in the denominator. Investigating whether a shift in the correlation structure among base benchmarks exacerbated recent ARP strategy underperformance is another research opportunity.

Figure 13 displays the three-year rolling Sharpe ratio for the seven categorical asset benchmarks. This chart reinforces the themes discussed previously through a different lens. The commodity complex registers the highest historical Sharpe ratios, with a variety of volatility strategies contributing to the gross-net spread. The decline in Sharpe ratios over time, coincident with moving from simulated to live performance, is apparent. Figure 13 reveals stock-based strategies to be the outlier in terms of weak performance, posting a net Sharpe ratio of -1.7 over the past three years. Only rates strategies report a positive net Sharpe ratio during the most recent window.

Figure 14 shows the three-year rolling Sharpe ratio for the 10 statistical broad benchmarks. The distinct benchmark methodology still reflects the performance of the underlying strategy universe, so the general Sharpe ratio observations echo those for the categorical benchmarks. Value oriented and commodity curve strategies print the highest historical Sharpe ratios. Stock-based strategies lead the recent drop-off in performance, followed by commodity trend and spread, volatility sensitive and equity sensitive strategies. Rates carry and commodity curve post solidly positive results, even on a net basis, over the past three years.

Table 8 and Table 9 provide supporting detail for the preceding Sharpe ratio charts. These tables show the reported calendar year Sharpe ratio for the categorical style and asset and statistical broad benchmarks. The additional granularity emphasizes previously discussed points and concludes the performance comparison between the two benchmark methodologies.

1.7.3 Benchmark Factor Sensitivity

This final section evaluates the factor footprint of the top tier categorical and statistical ARP benchmarks introduced in this paper. The objective here, using composite benchmarks that should diversify the specification variability, is to reveal the distinct sensitivity of the various ARP benchmarks to both traditional market factors and primitive ARP strategies. The analysis utilizes the supplemental factor set in Appendix B. The 104 factors tap the Bloomberg GSAM Risk Premia Indices, Fama-French Factor Library, Barra Global Equity Model (GEM3), and various market indices. Recall that the Bloomberg GSAM indices represent ARP primitive strategy benchmarks – the logical limit of strategy pruning via the selection of a single, simple methodology. While the current universe of Bloomberg GSAM indices does not span all the ARP strategies represented in the proprietary database, these indices do cover the core ARP approaches.

The analysis employs elastic net regularization (α of 0.5, 10-fold cross validation for MSE) to select a maximum of six explanatory variables for each benchmark. The marginal benefit from a larger number of regressors is small, yielding no incremental insights. The results include the factor loadings, Newey-West (1987) adjusted p-values, univariate R^2 and adjusted R^2 for the model. Appendix B includes complete names for the factor abbreviations.

Table 10 presents the regression results for categorical benchmarks in Panel A and statistical benchmarks in Panel B. Consider the results for the categorical carry (spread) style benchmark. The Bloomberg GSAM FX Carry Index, Bloomberg GSAM Cross Asset Carry Index, MSCI Europe Gross Total Return Local Index, and J.P. Morgan Emerging Markets Bond Total Return Index all show p-values significant at the 1% level. The results are reasonably intuitive. The two Bloomberg indices specifically target carry. The European equity sensitivity reflects the risk-on nature of carry (spread) and the EM bond index captures the credit spread exposure resident in the style. The ICE BofA US High Yield Total Return Index is significant at the 10% level, reinforcing further the credit spread and limited duration sensitivity. The MSCI EM Gross Total Return USD Index, helpful in the bias-variance tradeoff of a regularized setting, does not have a significant p-value. The overall 64% R^2 speaks to the explanatory power of the factors and to the significant idiosyncratic risk present in the carry (spread) benchmark. The individual factors have a univariate R^2 in the 30-40% range, indicating that the carry (spread) benchmark does not anchor disproportionately on any single factor – most notably, the Bloomberg GSAM Cross Asset Carry Index specifically targeting carry trades.

The categorical asset benchmarks generate a relatively similar adjusted R^2 of approximately 60%, a byproduct of the diversified nature of asset-based aggregation. Interestingly, the lowest R^2 applies to stock-based strategies. This runs counter to the perception that quantitative equity is a relatively standard combination of value, momentum and quality factors. Among the categorical style benchmarks, trend (time series) aligns closely with the Bloomberg GSAM Cross Asset Trend Index. The high R^2 and factor loading reinforce the notion that time series trend is among the most homogenous strategies in the ARP space. The size benchmark echoes the previous point regarding quantitative equity. Despite loading on an understandable combination of size (the orientation of the Barra size factor is opposite that of the other factors), negative momentum (size has been a longstanding relative performance loser), and value, plenty of idiosyncratic risk remains – a reminder of the specification challenge confronting primitive strategy benchmarks. Not surprisingly, the unavailability of a primitive strategy index results in weak explanatory power for the regressions on benchmarks for niche, dynamic strategies such as congestion, merger arbitrage and reversal.

The regressions on the statistical broad benchmarks generally post the highest adjusted R^2, reflecting the returns-based construction of this approach. For example, the equity sensitive benchmark shows p-values significant at the 1% level for the Bloomberg GSAM Equity Trend Index, MSCI Europe Gross Total Return Local Index, CBOE VIX, and Monthly S&P 500 Variance Swap. Note that the negative loading on the volatility factors is due to the long volatility profile of the factors whereas ARP strategies sell volatility to pick up carry. The p-value on the MSCI EM Gross Total Return USD Index is significant at the 5% level. The 83% adjusted R^2 signals both the clear footprint of the benchmark and that the benchmark represents more than traditional equity market beta. As expected, the regressions on nuanced, dynamic or specification rich benchmarks (crude oil volatility, volatility sensitive, value oriented) have the weakest explanatory power.

Because the Bloomberg GSAM indices appear often in the previous regressions, Table 11 provides some additional analysis. The table identifies the categorical and statistical benchmark (from all available tiers) having the highest correlation with each relatively focused Bloomberg GSAM index. The table also provides the associated tracking error and tracking error standardized by the benchmark volatility.

Among the categorical benchmarks in Panel A, the alignment is intuitive. The lowest correlation applies to the US equity quality and bond futures value indices, suggesting significant specification variation. The FX carry and FX trend indices register the highest correlation. The alignment is the same among the statistical benchmarks in Panel B, with the same Bloomberg GSAM indices showing the highest and lowest correlation. Overall, the relationship between the statistical benchmarks and Bloomberg GSAM indices is slightly tighter based upon the average of the columns in Table 11.

As discussed previously, the appeal of a primitive strategy benchmark is its simplicity. But, in the absence of a de facto methodology for every trading strategy, distinguishing the active contribution of an asset manager from benchmark specification noise remains difficult. The spanning approach of the benchmarks introduced in this paper resides at the other end of the methodological spectrum, seeking to average a universe of credible, traded methodologies. A relatively high correlation can distract from the practical tracking error consequence of this choice.

Table 12 uses traditional indices to contextualize this matter. For example, debating the choice of the primitive US stock-based cross-sectional momentum index and its statistical benchmark counterpart is tantamount to choosing between a North American and Pacific equity index for a traditional stock manager. Identifying the regional affiliation of a stock (global companies notwithstanding) is straightforward, so the index choice for the traditional manager is clear. With ARP, defining a trading strategy is not black and white. Without exercising caution, one could end up with the equivalent of evaluating a US stock manager with a Pacific equity index.

The largest differences between primitive and composite strategy benchmarks, US equity quality and bond futures value, are helpful in raising topics for additional investigation. These disconnects stem from some combination of the lack of a distinct base cluster post-pruning and variation in strategy specification. These two examples highlight areas in which benchmark selection requires particular care.

The unique nature of ARP necessitates a multi-pronged approach to performance evaluation. The composite strategy benchmarks introduced here provide an important and comprehensive complement to primitive strategy indices, which currently do not cover all strategies in the ARP universe.

1.8 Summary

Performance evaluation in ARP is uniquely challenging and no simple answer exists. Benchmarking in this space has received insufficient attention, particularly given recent disappointing performance. Data availability complicates matters, and parallels exist between the evolution of hedge fund databases in the late 1990s and the state of ARP data today. This paper adds to the limited existing literature by introducing a proprietary database of tradable bank indices, including both returns and comprehensive metadata - a recommended ARP strategy taxonomy, costs, live start dates, etc. Details regarding data curation frame best-practice requirements.

Using the custom database, this paper introduces a partially-nested family of categorical ARP composite strategy benchmarks. Such an approach has the benefit of simplicity and intuitive appeal, leveraging transparent benchmark inclusion logic and a robust taxonomy proposal. This paper also introduces a fully-nested family of statistical ARP benchmarks, emphasizing the return structure of the underlying strategies over descriptive information. The statistical process employs a blend of PLS and elastic net to handle data imputation. Hierarchical clustering underpins strategy classification, and PCA governs the strategy pruning process. Consistent with its objective, the statistical approach improves the homogeneity of the benchmark structure, increasing the variance explained by the first principal component by over 40% versus the categorical baseline.

The paper concludes with a factor analysis of the benchmarks, revealing sensitivities reflective of the underlying trading strategies, and a comparison with primitive strategy benchmarks, highlighting important tracking considerations. The comprehensive review of the proprietary database and the introduction of two families of composite strategy benchmarks represent important contributions to the ARP performance evaluation quandary. This foundational work will facilitate a variety of future research initiatives, including refining benchmark methodologies, characterizing specification noise, analyzing the post-publication behavior of strategies, mapping asset manager ARP offerings to benchmarks, and investigating recent ARP performance weakness.

Alternative Risk Premium: Workhorse or Trojan Horse?

Stephen Gorman

Abstract

Diversified alternative risk premium (ARP) portfolios seek to generate absolute returns using a broad range of systematic trading strategies incorporating multiple investment styles covering all the major asset classes. Following a period of rapid adoption, disappointing performance over the 2018-2020 period has produced considerable soul searching regarding the role of ARP in institutional portfolios. To examine this very topical issue, this paper utilizes a unique array of benchmarks designed using a proprietary database of 2,000 tradable bank indices. The following pages evaluate whether recent returns are consistent with long-term expectations, in the process considering the extent to which data mining, unique environmental headwinds, capacity pressure, or a lack of true breadth across ARP strategies contributed to this outcome.

Keywords: Alternative risk premium, multi asset, benchmarks, tradable indices, Sharpe ratio, multiple testing problem, correlation, nonnormality, conditional returns, turbulence, elastic net

2.1 Introduction

Spurred by the pain of recent financial market turmoil (Global Financial Crisis and European Debt Crisis), frustration with the performance of hedge funds, and a strong educational campaign by asset managers and investment banks, the nascent alternative risk premium (ARP) category grew rapidly over the past decade. Gorman (2019) details the evolution of ARP, including the primary appeal to investors.

1. clear economic rationale supported by years of empirical research by both academia and practitioners
2. persistent risk-adjusted return distinct from that of traditional beta and priced more reasonably than hedge funds
3. liquid (scalable), rules-based, transparent, predominantly long-short trading profile

Suhonen *et al.* (2019) estimated assets under management in the space to be approximately $150b by the end of 2017, a likely materially understated figure due to the absence of large plan sponsors, hedge funds and investment banks trading ARP strategies separately from the surveyed asset managers. The SG Multi Alternative Risk Premia Index, introduced at the end of 2017 to track the performance of diversified ARP asset managers, showed strong results for 2016 and 2017 (its only history). Entering 2018, the table was set for the ARP category to cement its role within diversified portfolios.

The subsequent three years dramatically altered this trajectory. The 2018 to 2020 period witnessed poor ARP performance and retreating investors. By late 2020, Bloomberg articles with titles such as “A $200 Billion Exotic Quant Trade Is Facing

Existential Doubts" and "Fast-Money Quants Get Schooled as Markets Get Faster and Wilder" reflected the shift in investor sentiment. Disillusionment and confusion replaced high expectations. Were the disappointing returns driven by unique environmental factors, crowding, overfitting, a regime shift, randomness, unconventional monetary policy, or some other factor? What exactly happened?

Little research exists yet on the 2018 to 2020 period for systematic investing. Recent papers focus upon the quantitative equity space -- primarily the value factor, for which crowding does not provide a convenient explanation since value spreads widened in recent years. Arnott *et al.* (2021) divide the value premium into a revaluation component and a structural component (profitability differences supporting growth and mean reversion in multiples favoring value). They attribute the underperformance of value predominantly to the now historic undervaluation of value versus growth, contending that the structural drivers of the value premium have not changed significantly. The authors also highlight that failing to capitalize intangibles biases down the denominator of price-to-book, understating performance of the traditional Fama-French HML factor.

Blitz (2021) focuses upon the narrowness of the stock market during the recent quant crisis, emphasizing that investing in large growth stocks represented the only relative performance path, with returns from factors such as momentum, profitability and low volatility being highly conditional on mega-cap growth exposure. He also considers recent performance of multi-factor stock strategies to be an unusual combination of events that falls within a range of reasonable outcomes.

Bellone *et al.* (2020) juxtapose the recent poor performance of value against better performance by quality, low volatility and momentum factors, portraying the recent drawdown in multi-factor stock strategies as significant but not exceptional. The authors also highlight the positive impact of using multiple definitions of each style factor and incorporating portfolio construction discipline -- targeting beta and sector neutrality, an equal risk contribution from factors, and constant volatility for the portfolio.

Lev and Srivastava (2020) attribute poor value factor performance partially to a failure to capitalize intangibles and to economic developments (contracting bank lending and declining consumer demand) that have slowed mean reversion between value and growth companies. Israel *et al.* (2020) highlight that the explanatory power of fundamentals is time varying. When investors place less importance on this information, value strategies suffer. The authors also find little empirical support for common criticisms of value related to share repurchase activity, accounting effects, low interest rates and crowding.

Pagano *et al.* (2020) and Baig *et al.* (2021) focus upon the role of Robinhood retail traders during the COVID-19 pandemic. The former authors find evidence of this group responding quickly to short-term news and overnight returns via both contrarian and momentum approaches. The latter researchers find a negative impact of this growing retail constituency on financial market stability. While not focused upon returns during the past few years, the recent paper by Koeppel (2021) introduces a sentiment factor leveraging the Refinitiv-MarketPsych social media-based sentiment indicator (RMI) that enhances the ability of the Fama-French five-factor model to explain the cross-section of US stock returns. Collectively, these papers highlight the possible impact of emerging technology platforms upon recent stock factor performance.

No comprehensive study of recent returns in the multi-asset ARP space exists, so this paper fills an important gap and provides a foundation for subsequent research. The recent plight of systematic multi-asset portfolios, this "quant winter", is certain to be the subject of many case studies. This empirical paper approaches the topic by questioning what the investment community missed given the information available at the end of 2017. Specifically, *the focus is identifying the deviations from expectations most responsible for the ARP performance problems between 2018 and 2020*. This investigation involves establishing appropriate expectations for Sharpe ratios, cross-correlations, auto-correlations, skewness, kurtosis, and state-based relative returns to serve as the basis for evaluating outcomes during the period in question. The results reveal four strategy groups principally responsible for the poor performance of diversified ARP portfolios — equity sensitive, volatility sensitive, diversified stocks and value oriented. The problem is predominantly one of average returns, with successive market crises weighing on the first two groups and an historic lack of breadth wreaking havoc on the latter two.

The paper proceeds as follows. Section two establishes the context, reviewing the recent financial market backdrop relative to history. Section three introduces the proprietary ARP benchmarks and conducts a detailed evaluation of the strategy return properties. Section four assesses the conditional return structure of ARP strategies. Section five utilizes the ARP benchmarks to analyze the effective exposure of an array of ARP fund managers during the 2018-2020 period. Section six considers extended return histories for select ARP strategies. Section seven concludes.

2.2 Financial Markets Backdrop between 2018 and 2020

On the heels of historically low equity market volatility in 2017, three successively larger crises buffeted financial markets during the 2018-2020 period. The low volatility regime ended abruptly on February 5, 2018 as a massive percentage spike in the VIX index punished short volatility investors and drove the closure (XIV) or restructuring (SVXY) of inverse volatility exchange traded products. The S&P 500 experienced a 10% drawdown around this relatively localized event, much of which reversed in a couple weeks despite the full retracement requiring six months.

In the fourth quarter of 2018, equity markets succumbed to angst over global economic growth resulting from Fed rate hikes, US-China trade tensions and Brexit. The S&P 500 dropped 20%, recovering most of the loss within two months and all of it within four months. Implied equity and bond market volatility returned to February levels. Crude oil prices abruptly ended a year of increases and crashed over 40%. This window also marked the end of over two years of gradual Fed rate hikes and upward drift in US bond yields.

After a particularly calm year for financial markets in 2019 characterized by strong gains in stock and bond market indices, the COVID pandemic struck in the first quarter of 2020, unleashing unprecedented economic calamity, erasing trillions of dollars from global output, and creating stark differences between industry winners and losers. The S&P 500 plummeted over 30% between February and March and the VIX reached levels last posted during the Global Financial Crisis. Crude oil prices plunged 70%, with Saudi Arabia's decision to increase supply to discipline Russia exacerbating the demand shock. The US 10-year government bond yield dropped 140bps in two months as global central banks flooded the economy with stimulus. Incredibly, the S&P 500 again retraced most of the drawdown in two months and all of it in four months.

This paper utilizes 22 reference benchmarks across equities (6), fixed income (7), commodities (4), currencies (3) and volatility (2) to contextualize ARP performance. This data provides the necessary market backdrop and permits analysis of the interaction between ARP strategies and traditional sources of risk. Two of these benchmarks do not have history back to December 1999 and serve situational purposes. All returns are excess of applicable one-month cash returns. Appendix D provides the complete list of reference benchmarks along with methodological notes.

Two exhibits summarize the market backdrop. To identify market-level departures from the validation window available to ARP strategies at the end of 2017, Panel A in Figure 15 compares reference benchmark annual excess return and volatility over the 2018-2020 period to the preceding 18 years. Panel B consolidates this information in a Sharpe ratio comparison. The fungibility of ARP strategies, due to the availability of both material leverage and derivative-based implementation, supports the use of standardized performance metrics like the Sharpe ratio. Figure 16 provides a heatmap summary of the 157 weekly data points constituting the 2018-2020 window to highlight the asset-level ebb and flow of market volatility and the profound relative impact of the COVID crisis.

These figures introduce numerous potential implications of the market backdrop for ARP strategies.

- **Equities:** Equity market performance was similar across the two periods for non-US markets, but conspicuously stronger in the US over the 2018-2020 period. Strong performance despite three crises within the 2018-2020 window could indicate performance headwinds for trend-oriented ARP strategies and persistently strong relative US performance could highlight a challenge for rotational strategies.

- **Treasuries:** Government bond markets posted a very high Sharpe ratio over both periods, underscoring a two-decade central bank effort to support real growth while controlling deflationary forces. Although the Sharpe ratio is lower during the recent period due to an increase in volatility, hedged non-US government bonds still delivered a Sharpe ratio of almost one – roughly three times long-term expectations. Such profound performance has clear implications for long duration leaning ARP strategies.
- **Credit:** Credit markets also experienced strong performance in both periods, with lower Sharpe ratios during the recent period reflecting the volatility impact of the pandemic. As with equities, the juxtaposition of solid returns and elevated volatility could portend whipsaw risk for ARP strategies.
- **Commodities:** Commodity markets experienced considerable rotation in risk-adjusted returns between the two windows. The Sharpe ratio in energy was decidedly negative in the recent period, much lower than that for the 2000-2017 period. A similar, albeit less dramatic, fate befell industrial metals. Conversely, the Sharpe ratio in agriculture and particularly precious metals improved in the recent period. The relatively strong performance of traditionally weaker carry commodities could be a headwind for ARP strategies.

- **Currencies:** Performance in currency markets also flipped during the most recent period, with the risk adjusted return in EUR and AUD slipping negative while that for JPY moved positive. This relative performance is inconsistent with the traditional positioning of ARP currency carry strategies.
- **Volatility:** Mirroring the reward to other risk-off assets (bonds, precious metals, JPY) during the 2018-2020 period, long variance printed a positive Sharpe ratio, a notable departure from the negative expected return for such a hedge and a significant problem for ARP strategies targeting the volatility risk premium. Additionally, the heatmap illustrates the profound relative volatility of the pandemic crisis window. An event of this magnitude can exert broad pressure on ARP strategies due to systemic deleveraging and the consequence of modest underlying positive correlations with reference benchmarks being fully evident during extreme market moves. The relatively rapid rebounds highlighted in this chart by blue shading also flag headwinds for both divergent ARP strategies and those incorporating volatility-based position sizing.
- **Value:** The world value spread, the specification here focusing upon undervalued companies returning capital relative to expensive companies consuming capital, posted the most dramatic reversal of 2000-2017 performance among reference benchmarks, a Sharpe ratio swing of positive to negative one in the recent window. The embodiment of the brutal short side of the value trade, the Goldman Sachs Non-Profitable Tech Company Index, produced a staggering annual excess return of 50% and a Sharpe ratio of 1.5 over the 2018-

2020 period. Such results have dire implications for stock-based ARP strategies and echo the findings of the value papers discussed in the introduction.

2.3 ARP Benchmark Return Structure Analysis

2.3.1 Performance Review

This paper utilizes the statistical composite ARP benchmarks introduced in Gorman (2020). This benchmark methodology leverages a proprietary database of 2,000 tradable bank index strategies. An agglomerative hierarchical clustering algorithm classifies each strategy, with a combination of partial least squares (PLS) and elastic net (EN) governing data imputation to maximize the assignment window. Principal component analysis (PCA) controls strategy pruning -- dropping strategies with weaker intra-group affiliations. Given the derivative-based nature of ARP strategies, all benchmark position sizes reflect a volatility weighting scheme to ensure equal representation and all benchmarks share a common volatility target.

The ARP statistical benchmark structure includes four, fully-nested tiers consisting of 155 benchmarks -- 85 base, 40 super-base, 20 hypo-broad and 10 broad benchmarks. In other words, the most granular benchmark tier includes 85 groupings of the 2,000 underlying ARP strategies. The 85 cohorts roll up successively into 40, then 20 and finally 10 groupings, with the latter being the coarsest benchmark tier. The range of benchmarks facilitates micro and macro analysis. Appendix C includes the full taxonomy.

Three return series are available for each benchmark. Reported returns reflect the index values reported on Bloomberg. While these indices typically are net of costs, practices vary across indices and banks in terms of incorporating costs. Therefore, the gross and net return series increase consistency by respectively eliminating and including all costs. This paper focuses upon the net return series. Given the negotiable nature of costs, net returns likely provide a conservative representation of the return experience of end investors -- a potentially lower, but structurally comparable, hurdle for the net returns reported by ARP fund managers.

This paper updates the ARP benchmark data from Gorman (2020) through December 2020. Panel A of Figure 17 displays the three-year rolling Sharpe ratio history for the 10 broad ARP benchmarks, with the shaded area reflecting the spread between gross and net tradable index returns and the dark line indicating reported returns. The disconcerting level and trajectory of recent Sharpe ratios encapsulates the quandary confronting ARP investors. Panel B spotlights the 2018-2020 period that is the focus of this paper.

Only rates carry and commodity curve carry post a positive net Sharpe ratio. Relative to history, the recent Sharpe ratio effectively is at a nadir for every benchmark except FX carry and rates carry. Of course, history for ARP benchmarks is a combination of pre and post-publication results, so the inherent overfitting creates an inflated comparative base. (Note that the 2018-2020 period is almost entirely out of sample.) Nevertheless, results for 2018-2020 materially undershoot even a short-hand 50% haircut of the 2000-2017 Sharpe ratio for all but one broad ARP benchmark.

Table 13 and Table 14 provide a more granular picture of trailing three-year ARP performance by disaggregating the broad benchmarks into the hypo-broad and super-base tiers. (Appendix E provides results for skewness, kurtosis and autocorrelation in the same format.) Lo (2002) provides the formula for the annualization factor (A_f) for the weekly Sharpe ratios in these tables:

$$A_f = \frac{f}{\sqrt{f + 2\sum_{k=1}^{f-1}(f-k)\rho_k}} = \sqrt{f} \; if \; IID \qquad \textit{Equation 18}$$

where f is the annualization frequency (52 for weekly) and ρ is the k^{th}-lag autocorrelation. The significance of autocorrelation estimates, particularly at high lags, is a consideration. Autocorrelation in ARP returns is modest given the liquidity profile, so this paper applies just a single-lag adjustment.

The sobering message of these exhibits is that 80% of broad, 75% of hypo-broad and 70% of super-base ARP benchmarks print a negative net Sharpe ratio for the 2018-2020 period. Quantitative stock selection strategies clearly represent the epicenter of recent performance woes, but most macro strategies reinforce, rather than counterbalance, these losses.

Among the broad ARP benchmarks, stocks, value light (also referenced in this paper as diversified stock strategies) post the worst results, as contributions from risk anomaly and trend factors reinforce or only marginally offset the performance drag from value exposure in multi-style strategies. (The "value light" moniker here simply indicates that pure value stock selection strategies reside in the value-oriented benchmark – "value light" and "diversified" are interchangeable.) Commodity spread carry/trend, volatility sensitive, value oriented and equity sensitive benchmarks all register conspicuously disappointing performance during the 2018-2020 period.

At the hypo-broad benchmark level, stocks value not surprisingly deliver the lowest Sharpe ratio, with the other stock selection benchmarks also posting weak results. This tier also identifies poor performance in equity trend and across a variety of short volatility strategies (commodities, rates, equities). The positive performance of reversal strategies is consistent with the market behavior discussed previously.

Turning to the super-base benchmarks, the worst of the stock selection strategies produce a Sharpe ratio approaching negative two – statistically plausible but still a staggering degree of underperformance and the impetus for the recent research focusing upon quantitative equity. Commodity trend and equity/credit trend also struggle significantly. At the other end of the spectrum, equity trend dynamic (short-term, predominantly intraday, strategies), FX value and rates trend round out the small array of positive performers.

The dearth of positively performing strategies creates a serious loss-stacking problem for diversified ARP portfolios. The negative skew in the Sharpe ratio distribution punctuates the lack of return air cover available to underperforming strategies during the 2018-2020 period.

The results in Table 13 for the 2000 to 2017 period are in stark contrast to the corresponding data in Table 14 for 2018 to 2020. Two dynamics are at work – post-publication return decay and a uniquely challenging period for many systematic strategies. Naya and Tuchschmid (2019), Suhonen, Lennkh and Perez (2017), and McLean and Pontiff (2016) discuss the tendency of strategy returns to deteriorate after going “live” — a byproduct of overfitting and profit erosion (implementation realities and crowding). The ARP database provides a uniquely broad manifestation of this phenomenon.

The 2000-2017 ARP data set is split 82/18% between simulated and post-publication results. The 2018-2020 sample is 91% post-publication. (For reference, the three years ending in 2017 is split approximately 50/50, reflecting the recent emergence of this investment category.) Therefore, the early (recent) window essentially is in-sample (out-of-sample). An appropriate representation of the Sharpe ratio expectations prevailing at the start of 2018 must acknowledge this reality.

Harvey and Liu (2015), Harvey and Liu and Zhu (2016), and Bailey and López de Prado (2012 and 2014) introduce methodologies for shrinking a back-test Sharpe ratio. The primary motivation is addressing the multiple testing problem, that reported Sharpe ratios reflect many iterations thereby increasing the likelihood of the result being a false discovery (exacerbated by the selection bias of reporting only successful strategies). Harvey *et al.* consider several multiple testing methods, ultimately preferring the more accommodative false discovery rate control for financial settings. Their simulation-based approach provides a p-value that controls for multiple testing, and the corresponding t-statistic provides the haircut Sharpe ratio (HSR). The result is a function of the estimated Sharpe ratio for a strategy, the sample size underpinning the Sharpe ratio, and the number of tests required to produce the Sharpe ratio. The HSR incorporates an estimate of the p-value distribution for tested strategies that accommodates different degrees of dependency among these strategies but does not account for non-normality in strategy returns.

Bailey and López de Prado propose an alternative probability measure, the deflated Sharpe ratio (DSR). As with the HSR, this probability has a corresponding Sharpe ratio. Whereas Harvey and Liu leaned on the multiple testing statistics literature, Bailey and López de Prado employ extreme value theory. The following equation provides the DSR calculation:

$$\widehat{DSR} = Z\left[\frac{(\widehat{SR} - \widehat{SR}_0)\sqrt{T-1}}{\sqrt{1 - \hat{\gamma}_3\widehat{SR} + \frac{\hat{\gamma}_4 - 1}{4}\widehat{SR}^2}}\right] \qquad \text{Equation 19}$$

$$where \quad \widehat{SR}_0 = \sqrt{V\left[\{\widehat{SR}_n\}\right]}\left((1-\varphi)Z^{-1}\left[1 - \frac{1}{N}\right] + \varphi Z^{-1}\left[1 - \frac{1}{Ne}\right]\right)$$

and

SR is the observed Sharpe ratio, SR_0 the expected maximum Sharpe ratio (assuming a mean of zero), T the sample size, γ_3 (γ_4) skewness (kurtosis) of observed strategy returns, φ the Euler-Mascheroni constant, e Euler's number, N the number of independent trials, V the Sharpe ratio variance across the trials, and Z the normal cumulative distribution function

HSR and DSR are complementary methodologies, employing different thresholds in assessing the significance of an observed Sharpe ratio. Each approach assumes certain underlying conditions and requires key assumptions regarding the strategy return generation process. Table 15 summarizes the inputs to the two calculations. HSR and DSR require an estimate of the number of trials underpinning an observed Sharpe ratio.

This information is unavailable for tradable bank indices. DSR includes a Sharpe ratio variance input and HSR a representation of the correlation among trials, both of which also are unavailable for bank indices. The remaining parameters, while not uniform, are observable.

Figure 18 illustrates the sensitivity of the Sharpe ratio reduction to the three inputs requiring assumptions. The observed Sharpe ratio exerts considerable influence on both the HSR and DSR adjustments, with a lower Sharpe ratio, ceteris paribus, receiving a larger penalty. The number of trials exerts a diminishing influence on the result, with most of the Sharpe ratio decrease in both approaches occurring prior to 150 trials. Increasing the correlation among trials shrinks the numbers of independent tests and reduces the HSR Sharpe ratio. This parameter becomes more consequential as the observed Sharpe ratio drops below one. Lastly, the DSR is very sensitive to the Sharpe ratio variance assumption. An observed Sharpe ratio less than one is not significantly different than zero for all but the most nominal variance inputs.

Figure 18 also highlights that the DSR is a more conservative than the HSR, producing a larger Sharpe ratio reduction. Harvey and Liu (2015) explore the relative aggressiveness of adjusting the p-value using the family-wise error rate versus the false discovery rate. For financial applications, the authors prefer the latter, while the DSR leans toward the punitiveness of the former. Finally, certain combinations of inputs also produce problematic discontinuities in the output. As a result, HSR and DSR represent helpful and complementary frameworks, not definitive solutions, for tempering an in-sample Sharpe ratio.

Acknowledging the appeal of these frameworks and the various application considerations, this paper introduces the following blended approach to determine an appropriate adjusted Sharpe ratio:

$$ASR_b = SR_b(1 + \Delta_b) \qquad \textit{Equation 20}$$

$$where\ \ \Delta_b = \frac{1}{n}\sum_{i=1}^{n} min[\delta_i, -0.2]$$

$$and\ \ \boldsymbol{\delta} = \begin{bmatrix} \frac{HSR_b}{SR_b} - 1 \\ \frac{DSR_b}{SR_b} - 1 \\ -0.5 \end{bmatrix}$$

and

ASR is the adjusted Sharpe ratio, *SR* the in-sample Sharpe ratio, and *Δ* the proportional Sharpe ratio adjustment for each broad ARP benchmark *b*

The *n* proportional Sharpe ratio adjustments (δ) include *HSR*, *DSR* and the practitioner 50% rule-of-thumb referenced by Harvey and Liu (2015). The rule-of-thumb and the 20% floor represent shrinkage parameters, jointly tempering the large (negligible) reduction in modest (high) observed Sharpe ratios to reflect input uncertainty and the possibility that a high historical Sharpe ratio may not persist due to competitive pressure. The shrinkage parameters result in an in-sample Sharpe ratio of 3.0 translating into an expected Sharpe ratio of 2.0.

Table 16 presents the Sharpe ratio adjustments for the 10 broad ARP benchmarks.[9] Consistent with Figure 18, a clear inverse nonlinear relationship exists between the level of the observed, in-sample Sharpe ratio and the size of the adjustment. Due to the lack of available information, inputs for Sharpe ratio variance, the number of trials, and the correlation among trials require assumptions. These parameters essentially become scalars that interact with the calculable inputs (observed Sharpe ratio, number of observations, skewness and kurtosis). The results in the table reflect a Sharpe ratio variance of 0.1, 150 trials, and 0.75 correlation among trials. The assumption regarding the number of trials reflects a point of diminishing marginal impact in Figure 18. The relatively high assumed correlation among trials also attempts to capture a meaningful degree of the possible parameter impact. Conversely, the choice of a modest Sharpe ratio variance offsets the overall conservatism of the DSR. This input calibration process balances parameter sensitivity, information gaps, and false discovery susceptibility to establish a reasonable set of Sharpe ratio expectations.

The ASR is one way to approach expectations, focusing on the back tests and differentiating among strategies. Establishing a Sharpe ratio hurdle (SRH) is an alternative. Consider the conventional formulation of a portfolio Sharpe ratio:

$$SR_P = \frac{\boldsymbol{w}\boldsymbol{r}^T}{\boldsymbol{w}\boldsymbol{\Sigma}\boldsymbol{w}^T} \quad and \quad SR_s = \frac{r_s}{\sigma_s} \qquad \text{Equation 21}$$

[9] Return aggregates linearly. Assuming average return is the source of Sharpe ratio adjustment, the adjustment to the broad benchmark Sharpe ratio is a weighted average of the underlying strategy Sharpe ratio reductions.

where $\boldsymbol{w}$ is the vector of strategy weights, $\boldsymbol{r}$ the excess return vector and $\boldsymbol{\Sigma}$ the covariance matrix. The Sharpe ratio per underlying strategy (SR_s) is simply the ratio of strategy excess return (r_s) to strategy volatility (σ_s). The SRH represents the required Sharpe ratio per ARP strategy to produce a target portfolio Sharpe ratio ($\overline{SR}_P$). In other words, the inclusion of a strategy in the portfolio implies investor confidence, after considering all the caveats regarding the in-sample results, in at least this level of strategy efficacy (i.e. $SR_s \geq \mathrm{SRH}$).

Working with a diversified ARP portfolio permits some simplifying assumptions in deriving the SRH. Because these portfolios employ leverage and can weight underlying strategy groups equally by volatility contribution, the strategy weight and volatility inputs in *Equation 21* reduce to constants — respectively, $\frac{1}{n}$ and $\bar{\sigma}$, with n representing the number of strategies. Some algebraic manipulation yields the following result:

$$SRH = \overline{SR}_P \sqrt{\frac{n}{n^2}(1 + \bar{\rho}n - \bar{\rho})} \qquad \textit{Equation 22}$$

with $\bar{\rho}$ representing the (average) correlation among strategies. *Equation 22* reveals that *SRH* for an ARP portfolio is basically a correlation-adjusted target portfolio Sharpe ratio. Assuming an average correlation of 0.2 among 10 ARP strategy groups, a portfolio Sharpe ratio of 0.6 corresponds to a strategy-level hurdle of 0.3.[10]

[10] The 0.2 correlation assumption is conservative relative to realized data and the 0.6 Sharpe ratio target is on the low end of expectations commonly referenced by practitioners.

Lo (2002), Mertens (2002), Memmel (2003), Opdyke (2007), Ledoit and Wolf (2008), and Auer and Schuhmacher (2013) expand upon the foundational work of Jobson and Korkie (1981) evaluating the significance of the Sharpe ratio. Lo (2002) and Mertens (2002) respectively address the impact of serial correlation and non-normality upon standard error estimates. Ledoit and Wolf (2008) propose a studentized time series bootstrap confidence interval for a Sharpe ratio difference. However, this test requires two return series for comparison, so the p-values in Table 16 reflect the parametric approach.

Table 16 contextualizes the recent poor returns of the broad ARP benchmarks relative to two measures of expectations — the ASR and SRH. Half the benchmarks register highly significant shortfalls versus the ASR. Value-oriented strategies, while middling in terms of relative 2018-2020 risk-adjusted performance, represent the most bitter disappointment due to the high expectation. Stock-based strategies contribute conspicuously to this result, but a collection of reversal trades meaningfully underdeliver. Volatility sensitive, equity sensitive, commodity spread carry/trend and diversified stocks all post similar, significant deviations from the ASR. Despite positive performance, commodity curve suffers from the burden of a high expectation, with the negative deviation significant at the 10% level.

The more conservative SRH comparison identifies four underperformers, with only diversified stocks and volatility sensitive significant at the 5% level. The deviation from the minimum requirement is significant at the 10% level for equity sensitive and commodity spread carry/trend. Despite the broadly disappointing returns, Table 16 reveals that as few as two ARP strategy groups deliver statistically significant departures from expectations for the 2018-2020 period.

The preponderance of negative Sharpe ratios highlights the dearth of positive return offsets to significantly underperforming strategies. Piling unexceptional upon exceptional losses has obvious implications for diversified ARP portfolios. The objective here is to frame the recent Sharpe ratios and to identify outliers. Further research could explore the median shortfall versus expectations.

2.3.2 Nonnormality Considerations

While weak average returns represent the clear epicenter of the recent ARP crisis, the possibility exists that departures from expectations in terms of distributional profile or dependency structure exacerbated the problem. Non-normality characterizes many ARP return distributions. Vatanen and Suhonen (2019), Baltas and Scherer (2019), Hamdan *et al.* (2016), and Lempérière *et al.* (2014a) discuss this profile, conditional market betas, and the possibility that returns to numerous ARP strategies may be compensation for negative skewness.

Table 17 provides skewness and kurtosis for the 10 broad ARP benchmarks. Appendix E provides additional detail for the hypo-broad and super-base tiers. Values for 2000-2017 reflect the historical information available to investors entering 2018 and represent a reasonable representation of expectations prevailing at that time. All benchmarks manifest some degree of leptokurtosis, but skewness varies from negative to minimal to positive.

This paper uses the following conventional calculations of skewness and kurtosis:

$$sk = \frac{\sqrt{n(n-1)}}{n-2}\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^3}\right]$$

Equation 23

$$and \quad ku = \frac{n-1}{(n-2)(n-3)}\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}(n+1)-3(n-1)\right]$$

where sk is sample skewness, ku is excess sample kurtosis, x is a return series, and n is the number of observations.

Strategies exposed to short volatility positions unsurprisingly register high negative skewness given the insurance provision nature of the trade. Volatility sensitive, crude oil volatility and equity sensitive groups fall in this class. The dependence of FX carry upon pro-risk currency exposure results in a risk profile akin to that of short volatility and a similar negative skew. The negative skewness for diversified stock strategies reflects the negative skew (blow-up or reversal risk) accompanying underlying cross-sectional momentum trades. The negative skew in rates carry is a byproduct of the secular bull market in bonds, shrinking yields and investor apprehension regarding a reversal in monetary policy.

In contrast, value-oriented strategies register large positive skewness due to a combination of the historical payoff to value stocks and dynamic (intraday) equity trend strategies. The scarceness of, global appetite for, and theoretical incompatibility with an expectation of positive return and positive skewness invariably raises questions regarding the sustainability of such a result. Finally, FX/multi-asset trend, often characterized as a long volatility or convex strategy, displays the negligible skewness and relatively low kurtosis consistent with a long-short beta timing approach.

The difference line in Table 17 highlights inconsistency between the 2018-2020 results and historical expectations, using bootstrapped standard errors. Four broad benchmarks experience significant deterioration in skewness – volatility sensitive and equity sensitive strategies due to equity market turbulence and diversified stock and value-oriented strategies due to factor-based performance problems at the stock level. The common denominator in terms of distributional surprise during the recent window is the equity market.

Kurtosis also changes meaningfully in four broad benchmarks during the 2018-2020 window. Leptokurtosis decreases in rates carry and value-oriented strategies and increases in volatility sensitive commodity curve trades. Generally, these results do not suggest a problematic distributional surprise.

For context, Table 18 provides skewness and kurtosis for traditional benchmarks, showing that ARP benchmarks reflect the variably negatively skewed, broadly leptokurtotic profile of the underlying assets. The high negative skewness of credit is consistent with that of ARP strategies having an insurance provision orientation. The less extreme negative skewness of equities, while not significant, moved more negative during the recent window, as did the skewness of ARP strategies focused upon the asset class. Government bonds (with the recent exception of US linkers) and commodities print the most modest skewness, comparable to readings for rates and commodity ARP strategies.

A point regarding the p-values referenced in Table 17 and Table 18 warrants mention. The number of data points is the only input to the conventional standard error calculation for skewness and kurtosis. This raises test power considerations as the sample size increases. As a result, some researchers employ heuristic approaches to assess the significance of skewness and kurtosis, varying the significance level or relying upon absolute thresholds depending upon the number of data points.

Table 17 and Table 18 calculate the standard error via bootstrap. This has a marginal impact upon the larger 2000-2017 sample but is more consequential in the recent window given the smaller number of data points and underlying market turbulence. This nonparametric approach samples with replacement from the underlying data, with the standard error capturing the variability across 5,000 calculations of each statistic.

A variety of general normality tests exist to incorporate the strong connection between absolute changes in skewness and kurtosis. See Lilliefors (1967), Massey (1951), Jarque and Bera (1987), Anderson and Darling (1954), D'Agostino *et al.* (1990), and Shapiro and Wilk (1965). Running the D'Agostino, Shapiro-Wilk, Lilliefors, Kolmogorov-Smirnov, Jarque-Bera, and Anderson-Darling tests on the ARP data yields very similar results. The p-value for all ARP broad benchmarks over the 2000-2017 period is less than 1% (rejecting the null hypothesis of normality). Over the 2018-2020 period, the p-value is less than 5% for all benchmarks and tests, with three exceptions within the Kolmogorov-Smirnov results. The six tests also reject normality for the traditional benchmarks (with the recent exception of agriculture) for both time periods.

Rejecting normality does not necessarily indicate that the deviation is sufficiently material to invalidate insights from standard hypothesis testing. The implications may be different for a marginally versus heavily skewed return distribution (e.g., FX/multi-asset trend versus volatility sensitive). Regardless, the important take-away is that, while ARP broad benchmark returns register as non-normal, the distributional profile varies considerably and the material deviations from expectations during the 2018-2020 period generally have ties to the equity market – primarily volatility and stock-based trades.

Finally, this paper reports the Sharpe ratio despite the non-normality of returns for most ARP strategy groups. The rationale is twofold. First, while alternative measures of risk-adjusted performance statistics exist that consider skewness and kurtosis, using the conventional method preserves reader familiarity with the scale of a statistic that is ubiquitous in both practice and the literature. For example, the modified Sharpe ratio uses value-at-risk in the denominator, the Calmar ratio maximum drawdown, and the Sortino ratio downside deviation. In addition to direct comparability among such measures being an issue, none represents a broadly accepted alternative due to the various considerations accompanying each metric.

Second, this paper incorporates skewness and kurtosis into hypothesis testing and analysis as appropriate, and generally focuses on the Sharpe ratio difference for recent versus historical results for a given strategy, narrowing the issue to the marginal change in non-normality over a relatively short and volatile sample period. Because no single performance metric is a substitute for understanding the return generation process, this paper emphasizes full contextualization of the Sharpe ratio.

2.3.3 Correlation Structure Stability

2.3.3.1 Cross-Correlation

An event as consequential as the onset of the COVID pandemic can impact the correlation structure materially within a relatively short measurement window. While the 2000-2017 period contains two recessions, the second of which being the Global Financial Crisis, the possibility exists that deviations from correlation assumptions rooted in history exacerbated strategy return headwinds during the 2018-2020 window.

Figure 19 compares the correlations among and between ARP and traditional benchmarks available to investors entering the 2018-2020 period with the subsequent realizations. Figure 20 provides additional perspective on the relationships among the ARP benchmarks. Table 19 evaluates the significance of the correlation differences. Given the considerable market turbulence, the correlations among broad ARP benchmarks are remarkably stable. The average correlation difference between the two periods is zero, with an average absolute change of 0.06. Only three correlations (7% of the population) register as significantly higher during the recent window, each reflecting different dynamics. Significance tests of pairwise correlation differences incorporate the Fisher (1921) transformation, bootstrapped standard errors and a conservative alpha of 10%.

The tighter connection between the volatility and equity sensitive benchmarks over the 2018-2020 period reflects the broad underlying pressure generally on risk assets and particularly on the volatility market. The rise in correlation between volatility sensitive and diversified stock strategies, albeit to a modest level, represents a disappointing lack of air cover from the historically independent quantitative equity complex. March 2020 included a deleveraging impulse and the disappointing preceding performance of stock-based strategies made these trades particularly vulnerable to capitulation-oriented de-risking.

Lastly, the increase in correlation between commodity spread/trend and commodity curve reflects a time-varying, structural relationship between the two strategy groups. Declining (rebounding) energy returns coincide with upward (downward) pressure on the correlation. Curve trades generally have a short position in the front-month contract, funding a long position in a more distant contract for a given commodity. Trend strategies typically traffic in the front-month contract, so weak energy returns result in position alignment and positive correlation between commodity trend and curve strategies. Cross-commodity spread carry strategies can reinforce this situational dynamic if falling energy prices reduce the relative roll yield of crude oil and distillates, making these contracts an attractive short versus other commodities.

Four significant decreases in ARP benchmark correlations over the 2018-2020 period offset the three significant increases. Commodity curve strategies registered negative correlations with the equity sensitive and FX carry groups, reflecting the resilience of the former during the COVID sell-off. The value-oriented benchmark posted negative correlations with rates carry and diversified stock strategies, in both cases a byproduct of returns moving in different directions for two years as opposed to a localized event.

In contrast, recent changes in the correlation matrix for traditional benchmarks are more consequential than those for ARP benchmarks. A systemic shock measured over a three-year window amplifies the inherent connectivity among risk assets. The sample correlation among equities, credit and energy (in the absence of an inflationary event) rises significantly. In this instance, the correlation between risk assets and precious metals also increases significantly, a reflection of gold prices following the inverse trajectory of US real yields and aligning with stock market behavior.

The few significant decreases in traditional benchmark correlations in 2018-2020 versus 2000-2017 are equally unsurprising given the environment. The typical relationship between government bonds and investment grade corporate bonds breaks down during a credit event. The relationship between the idiosyncratic, weather-sensitive agricultural sector and the other commodity blocks is prone to instability. In total, using 2000-2017 as the expectational base, almost 40% of the correlations among traditional benchmarks moved significantly (primarily higher) during the 2018-2020 period.

The northeast (or southwest) quadrant of Figure 19 indicates that the forces impacting correlations among traditional benchmarks between 2018 and 2020 also affect the relationship among ARP and traditional benchmarks, albeit to a lesser extent. The significant correlation decreases occur between FX carry and EM equity and certain commodities and between value-oriented strategies and government bonds and agriculture, the former a consequence of changes in global relative monetary policy and rate levels and the latter due to distinctly different return patterns in 2019 and 2020.

The significant correlation increases concentrate in volatility sensitive, crude oil volatility, and diversified (value light) stock strategies with traditional risk assets – equity, credit and certain commodities. The result for volatility sensitive strategies is understandable since these approaches have asymmetric sensitivity to the market crises characterizing the 2018-2020 window. The correlations between diversified stock strategies and traditional benchmarks moved from negative/low in 2000-2017 to low/modest in 2018-2020 period. These strategies remain diversifying to most traditional benchmarks (the highest correlations are with risk-off assets), and variation in the annual correlation profile within the 2018-2020 window suggests that idiosyncratic outcomes in 2020 may be skewing results.

Despite the correlation changes being larger between ARP and traditional benchmarks than among ARP benchmarks, the significant differences during the 2018-2020 period are limited to select strategies and not surprising given the market turmoil and short sample window. Table 20 introduces a simple ARP portfolio to provide some numerical context. The relationships among ARP benchmarks is sufficiently stable that realized portfolio volatility and the strategy-specific contributions to that volatility are consistent with expectations. Misestimation of the ARP benchmark covariance matrix leading to excess portfolio leverage during the 2018-2020 period is not meaningfully exacerbating the impact of the underlying strategy Sharpe ratios. Conversely, the connection with traditional benchmarks is a source of disappointment, as the combination of elevated realized equity market correlation and lack of participation in equity market rebounds points to ARP underdelivering on its role as a diversifier.

Using return data available through 2017, this portfolio targets an equal volatility contribution from each of the broad ARP benchmarks. The results assume the portfolio maintains the underlying notional strategy allocations established at the outset through 2020. The realized behavior among ARP strategies, summarized in the contribution to portfolio volatility section, is consistent with expectations. Portfolio volatility exceeds the target only marginally and proportional contribution deviations are modest, excepting the comparatively large offsetting differences from equity sensitive and rates carry strategies. Environmental qualifications notwithstanding, the realized portfolio correlation with the equity market (0.6) is a disappointment relative to expectation (0.1). Approximately, half of this gap is attributable to two non-diversifying strategies (volatility and equity sensitive) and half to two diversifying strategies (rates carry and stocks, value light).

Finally, the focus here has been upon pairwise correlations. Mantel (1967) introduces a statistic to measure the correlation between two symmetric proximity (similarity or dissimilarity) matrices. The nonparametric Mantel test involves permuting the rows and columns of one matrix to obtain a distribution of correlations to determine the p-value.[11] This test facilitates a comparison of the entire correlation matrix between the two time periods for the ARP and traditional benchmarks. Because each of the columns in Table 19 includes at least a handful of significant pairwise differences, the Mantel test consistently rejects the null hypothesis of matrix equivalence. Focusing upon pairwise correlations permits a more robust exploration of the topic.

2.3.3.2 Autocorrelation

Given the liquid and dynamic nature of the strategies, weekly ARP returns should be reasonably independent. Table 21 displays first-order autocorrelation for the 10 broad ARP benchmarks. Appendix E provides additional detail for the hypo-broad and super-base tiers. For comparison, Table 22 provides the same information for 15 traditional benchmarks. Recall that the conventional time-based adjustment understates (overstates) annual volatility for a return series with positive (negative) autocorrelation.

[11] See Glerean *et al.* (2016) for an application.

The traditional benchmarks provide little evidence of significant autocorrelation. High yield bonds are the conspicuous exception, manifesting very high positive autocorrelation historically due to the liquidity of the underlying bonds. Interestingly, the significant risk reversals during the 2018-2020 period temper the recent reading. US equities exhibit evidence of negative autocorrelation during the longer window and precious metals experience strong autocorrelation as a safe-haven trade during the recent crises.

ARP broad benchmarks also offer few indications of significant autocorrelation. Volatility sensitive strategies show significant positive autocorrelation in both periods, with the COVID crisis clearly impacting the recent return generating process for multi-asset and FX short volatility trades. Stock-based strategies (diversified, trend and value) show long-term evidence of significant positive autocorrelation that reverses in recent years. Commodity spread/trend strategies indicate long-term negative autocorrelation that dissipates recently. FX/multi-asset trend shows significant negative autocorrelation within the 2018-2020 window, a troubling profile for a trend strategy and a reflection of the choppy market conditions confronting the approach.

In aggregate, the absolute level of first-order autocorrelation is not particularly high for most ARP benchmarks, and the subsequent three years represent a very limited departure from expectations prevailing at the end of 2017. Volatility sensitive strategies again are in focus, reflecting the impact of the historic COVID crisis on a three-year window. Given this sampling consideration, these results do not raise serious statistical concerns and are reflected in volatility annualization throughout this paper.

2.4 ARP Conditional Return Analysis

2.4.1 State-Dependent Performance

As discussed previously, the average correlation between ARP strategies and traditional benchmarks is low. However, certain ARP strategies exhibit more positive market beta exposure than others. For example, the equity sensitive group manifests a clear positive connection with global stock markets, while rate carry strategies register a decidedly positive correlation with bond markets. A mix of procyclical, countercyclical, and low macro sensitivity strategies clearly populate the ARP universe. Together with the position dynamism and higher moment profile, these considerations indicate that ARP strategies likely accrue returns in distinct ways over a market cycle – an important acknowledgement given the relatively short and unique nature of the 2018-2020 period.

Two recent studies investigate the returns of ARP strategies during weak equity and bond markets. Vatanen and Suhonen (2019) focus upon the conditional beta of ARP strategies for the lowest quintile of market returns versus that for the remaining 80% of the data. They then run a PCA on conditional ARP strategy returns (scaled by full-period volatility) to explain variation across market return quintiles. The authors use the results to distinguish offensive and defensive strategies and to highlight increasing ARP strategy sensitivity to the worst bond markets.

Baltas and Scherer (2019) also question the market neutrality of ARP strategies, highlighting negative average ARP strategy returns coincident with bottom quintile stock and bond market performance. The authors employ a downside risk CAPM framework, first estimating the traditional and downside equity and fixed income betas per strategy via time-series regression and then deriving the equilibrium risk premia via cross-sectional regression of average strategy returns on the beta estimates. The authors find that including downside betas increases the explained proportion of cross-sectional average ARP strategy return variance, but they struggle with the counter-intuitive negative sign on these coefficients.

This paper approaches the state-dependent evaluation of ARP strategies with a different data set, a focus on in-sample versus out-of-sample consistency, and an alternative framework. Specifically, event definition, as opposed to market index quantiles, underpins the subsequent analysis. Quantiles require return frequency and group count assumptions. Event definition involves slightly different decisions regarding change magnitude and time. Essentially, the former focuses on periodic returns and the latter on drawdowns of a certain magnitude. The approaches are complementary, with considerable overlap. This paper proceeds with an event orientation to supplement the previous research.

Event identification is an example of rules-based, or defined, regime classification and represents a transparent baseline for statistical alternatives that could be the subject of future research. Also, the approach may produce some overlap in states around inflection points -- the incidence is higher in the volatile 2018-2020 window. The implication is that certain data points receive extreme, but ambiguous, designation and appear in both states. This is a byproduct of characterizing an event versus an individual data point, and potentially provides incremental perspective.

Figure 21 illustrates the process for two market barometers – the CBOE VIX index and the Bloomberg US Financial Conditions index. This paper identifies an event as a signed minimum change of x in the n-day moving average over a t-day period. The specification for all state indicators is a 60-day change in the three-day average of the target data set. The choice of a 60-day window represents a balance between an indicator that changes more rapidly than underlying ARP positioning responds to market conditions and one that is too static. A state measured in weeks rather than days or months achieves this balance for the overall ARP strategy universe. The choice of 40, 50 or 60 days does not meaningfully impact results. The change parameter is unique to each data set, with time-series stationarity dictating whether a change in level or a proportional change applies. The process employed here determines the change in each data series necessary to produce a state occurrence of 20-25% across the full 2000-2020 period. This quintile-quartile orientation attempts to avoid having states too sparsely populated to support analysis.

Figure 21 raises two important considerations. First, rising volatility and deteriorating financial conditions both correspond to market stress; however, the shaded areas do not align perfectly as that stress manifests in different ways. Second, state incidence may not be equal in the 2000-2017 and 2018-2020 periods. Rising equity volatility is the outlier with more than 40% of 2018-2020 data characterized as extreme, a significant departure from the 20% occurrence during the preceding 18 years.

This paper leverages 12 underlying data series to introduce 24 rising/falling state indicators. Appendix G contains the complete list and provides the change specification for each state. These data series clearly are not independent; however, the nuances among ARP strategies justifies a comprehensive exploration of sensitivities.

The state indicators facilitate the comparison of conditional mean returns within and across ARP benchmarks. The analysis here targets the broad ARP benchmarks to convey the universe profile and the process dynamics. The nested nature of the benchmark structure permits more granular strategy evaluation. The following is the state-based conditional mean return calculation:

$$\mu_s = \frac{1}{\sum_{t=1}^{T} I_{s,t}} \sum_{t=1}^{T} r_t I_{s,t} \qquad \text{Equation 24}$$

where *s* is the state, *I* the binary state indicator, and *r* the strategy return for week *t*.

Figure 22 presents the conditional weekly mean return profile for value-oriented strategies for rising/falling/neutral states across 12 market measures and the 2000-2017 and 2018-2020 time periods. Appendix H contains this information for all 10 broad ARP benchmarks. A missing line segment indicates no state representation during that time period. Importantly, the unconditional mean is very different across the two time periods, so the shape of each conditional payoff line (i.e. the deviation from the neutral state) is of primary interest. Benchmark returns have a consistent target volatility, so the conditional means do not require standardization.

The objective here is to discern whether changes relative to expectations in the state-based behavior of ARP strategies contribute to recent performance woes. For example, Figure 22 reveals considerable consistency between the in-sample and out-of-sample relative return structure of value-oriented ARP strategies. In all market states except for gold, this strategy group generates higher returns in both pro and counter-cyclical extreme states. The V-shaped, straddle-like payoff profile is evident in both time periods. The result for gold warrants a caveat. The falling gold state in the 2018-2020 period is the most thinly populated of all states posting a mean. Further, the correlation between gold and both equity and bonds increased meaningfully during the recent window. The interaction of small sample size and shifting dependencies could be impacting this finding. Overall, Figure 22 reveals a payoff profile that is consistent both over time and with intuition regarding the dynamics of the underlying value trades.

Payoff profile uniformity varies among the remaining ARP broad benchmarks. The pro-risk sub-class of benchmarks (equity sensitive, volatility sensitive, FX carry, and crude oil volatility) generally register a consistent northwest-to-southeast payoff profile between 2000-2017 and 2018-2020. The defensive sub-class (rates carry) also prints broadly consistent relative returns – in this case, a southeast to northwest orientation. The result is consistent with these strategies incorporating a certain amount of market beta.

Except for value-oriented strategies, the diversifier ARP sub-class (FX/multi-asset trend, commodity spread carry/trend, and commodity curve) does not deliver a uniform payoff profile. Similarly, the remaining defensively oriented ARP strategy, diversified stocks, shows little consistency over the two time periods, except for a tendency to produce relatively strong performance when equity volatility rises. This finding is not particularly surprising. Within a short evaluation window, positioning entering a crisis, the duration of the crisis and the velocity of the rebound will impact results for the two trend strategies. In terms of economic significance, these four strategies carry the lowest average absolute relative returns, reflecting low state sensitivity due to the nature of the underlying position taking. Such considerations temper expectations for in-sample versus out-of-sample payoff profile consistency.

2.4.1.1 Role-Based ARP Strategy Designation

Figure 23 uses a minimum spanning tree (MST) to support grouping the broad ARP benchmarks by general portfolio construction role – risk seeking, diversifying and defensive. The MST is an edge-weighted undirected graph. Mantegna (1999), Mantegna and Stanley (2000), Djauhari (2013), and Wang *et al.* (2013) provide examples of using the MST in equity and currency settings to reduce the dimensionality of the correlation matrix. The graphing process begins by converting correlations to Euclidean distances using the following formula:

$$d_{ij} = \sqrt{2\left(1 - \rho_{ij}\right)} \qquad \textit{Equation 25}$$

where ρ is the correlation between return series *i* and *j*.

From the network of connections, an algorithm finds the subset of edges minimizing the total distance while including every vertex. See Kruskal (1956), Prim (1957), and Sammon (1969) for examples of these algorithms. This paper uses Sammon's nonlinear mapping criterion to produce the graph.

Figure 23 depicts an intuitive arrangement of ARP strategies, relative to both one another and traditional benchmarks. The following three groups emerge for both the 2000-2017 and 2018-2020 periods.

- **Risk Seeking**: Crude Oil Volatility (short), Volatility Sensitive, FX Carry, Equity Sensitive
- **Defensive**: Rates Carry, Diversified Stocks

- **Diversifying**: FX/Multi-Asset Trend, Commodity Curve Carry, Commodity Spread Carry & Trend, Value Oriented

Table 23 evaluates the statistical significance of the conditional mean return spreads. The Welch (1947) unequal variances t-test with bootstrapped standard errors produces the p-values. For the 2000-2017 period, the spread for the risk seeking and defensive ARP sub-classes is significant in most states. Within the diversifier sub-class, mean spreads generally are insignificant. For all ARP strategies, statistical significance is elusive in the 2018-2020 period due to the combination of sample size and volatility that renders insignificant even spreads larger than those for the in-sample period.

Of course, all states do not correspond equally to the return generation process of the various strategies, so the expectation is not that significant spreads exist in every state. Across the three volatility states, volatility sensitive ARP strategies post the largest mean spreads and lowest p-values. Equity sensitive strategies print similar results across the stock, credit and breakeven states. Across the Fed, yield, gold and real yield states, rates carry strategies register the largest spreads and lowest p-values. The results all square with intuition.

In terms of noteworthy departures from expectations, diversified stock strategies top the list. During the 2000-2017 period, these strategies demonstrate significant conditional mean spreads reflecting relative strength in rising equity volatility and falling stock price regimes. This valuable performance offset essentially is non-existent within the 2018-2020 window. Generally, however, the evidence does not support changes in the state-based payoff profile of ARP strategies being a material driver of recent performance problems.

2.4.2 Performance during Turbulent Periods

The event identification process is a mosaic approach, gathering insight through a panel of indicators. The following targeted methodology represents an appealing alternative for stated-based evaluation of ARP returns. This approach focuses on the tendency of many ARP strategies to perform reasonably well in orderly market environments (prices moving sideways or up/down at a modest rate or for a sustained time period) and to struggle during periods of disruption (inflection points, price jumps, significant market chop). Carry strategies thrive on the status quo, with large price moves potentially undoing many months of return accumulation. Trend strategies require time to react to events and wrestle with whipsaw risk in thrashing markets. Value-oriented strategies operate with an effective investment horizon and can struggle during periods of significant misalignment between market oscillation and position rotation.

Chow *et al.* (1999) propose using the following multivariate distance measure of Mahalanobis (1936) to identify financial market turbulence.

$$d_t = (\boldsymbol{y}_t - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t - \boldsymbol{\mu})^T \qquad \textit{Equation 26}$$

For a T-by-n matrix of returns ($\boldsymbol{Y}$), *Equation 26* defines the distance (d) at time t, with $\boldsymbol{y}$ the 1-by-n return vector, $\boldsymbol{\mu}$ the 1-by-n mean vector for $\boldsymbol{Y}$, and $\boldsymbol{\Sigma}$ the covariance matrix for $\boldsymbol{Y}$. d is a chi-square statistic, so a multivariate outlier exceeds the critical value for a specified tolerance.

This paper proposes using the interaction among global stock, bond and precious metals markets to identify the market turbulence relevant to ARP strategies. Figure 24 depicts the anatomy of the turbulence indicator, with the red dots outside the tolerance ellipsoid denoting turbulent weeks over the 2000 to 2020 period. The focus with this measure is discrete data points versus short-term regimes. Turbulence, an ex post environmental classification, characterizes approximately 20% of the data points in both the 2000-2017 and 2018-2020 sub-periods.

Table 24 evaluates the conditional mean returns across the broad ARP benchmarks using the turbulence indicator. The analysis yields four important results.

- Between 2000 and 2017, all ARP strategies, except value-oriented, deliver lower average returns during turbulent versus non-turbulent times. Eight of the 10 mean spreads are significant.
- Over the 2018-2020 period, all but two ARP strategies post lower average returns during turbulent states. Five of the spreads are significant, which is noteworthy given sample size and volatility considerations.
- The sign of the turbulence/non-turbulence mean spread is consistent across the two time periods for eight of the 10 ARP benchmarks.
- Comparing the conditional turbulence mean for the 2018-2020 and 2000-2017 periods produces only two significant differences – the lower (higher) recent mean return for equity sensitive (commodity curve). This cross-period

comparison arguably is a very conservative equivalence test given that an inflated unconditional mean buoys the in-sample results.

These findings reveal a significant amount of consistency between the two time periods with respect to the behavior of ARP strategies during market turbulence. The evidence does not support departures from expected state-based payoffs playing a significant role in recent poor ARP performance.

2.5 ARP Fund Analysis

2.5.1 Universe Performance

Reference market benchmarks indicate likely performance headwinds for ARP strategies between 2018 and 2020. The composite ARP benchmarks reveal the sobering consequences of this environment. Such broad-based negative performance across lowly correlated strategies portends a loss-stacking problem for diversified ARP fund managers. However, performance assessment of ARP portfolios is a challenging proposition. Strategy definition, inclusion and allocation vary across managers. This section provides a case study in the application of ARP benchmarks to fund performance analysis, highlighting important considerations while examining the alignment between benchmark returns and the pain experienced by end investors.

The same challenges that complicate ARP strategy comparison and benchmark construction apply to ARP fund managers. Each firm defines its investable universe independently, specifies strategies differently, allocates to and constrains strategies uniquely, incorporates varying degrees of dynamism and downside risk control, and employs different amounts of leverage. Such heterogeneity in a single investment category can create material performance dispersion and makes relative manager skill assessment very challenging.

This paper assembles a universe of 22 diversified ARP managers with weekly prices and distributions quoted in Bloomberg between December 2017 and December 2020. Appendix F provides the full list. Only 80% (50%) of this universe carries a four-year (five-year) track record, reflecting the nascency of the category. The SG Multi Alternative Risk Premia Index, an equally weighted blend of ARP managers with multi-asset and multi-style exposures, provides an aggregate performance measure.

Figure 25 summarizes standardized cumulative performance for this group of managers over the 2018-2020 period. Assuming normally distributed returns and an expected 0.6 Sharpe ratio on 8% fund volatility (a conservative benchmark relative to the typical assumptions entering 2018), the shaded cones grow as a function of time and volatility and frame the likelihood of the realized outcomes. The following formula determines the performance expectations or cone portion of the chart:

$$C_{pt} = \left(e^{R_T \pm Z * \frac{V_T}{\sqrt{t}}}\right)^t - 1 \qquad \textit{Equation 27}$$

where

$$V_T = ln\left[1 + \left(\frac{v}{1+r}\right)^2\right]^{0.5} \qquad R_T = ln[1+r] - \frac{V_T^2}{2}$$

and C is the cone given probability p and time t in years. The discrete target volatility (v) is 0.08, target return (r) is cash + 0.6 * 0.08, and Z is $|\phi^{-1}(p/2)|$, where ϕ^{-1} indicates the normal inverse cumulative distribution function.

The results are dismal. The funds produced a median Sharpe ratio of -0.6, relative to an expectation of 0.6, with an interdecile range of -1.5 to 0.4. No fund exceeded the objective between 2018 and 2020. Only three funds posted returns within the 50% probability region, and most delivered results statistically inconsistent with expectations. The dark line representing the SG index falls in the latter category. The chart highlights the vulnerability of most funds to the acute volatility inflection in February 2018 and particularly February-March 2020, as well as the sideways to downward drift (i.e. no significant rebound) during much of the remainder of the three-year period. Along with Figure 17, Figure 25 captures the profound disappointment of investors, catalyzing widespread soul searching within the ARP space. These two exhibits also prompt the natural follow-on question regarding the relationship between ARP benchmarks and funds. What strategies weigh most heavily on fund performance?

2.5.2 ARP Fund Performance Analysis via Base Benchmarks

As detailed in Gorman (2020), performance attribution for ARP funds is a challenging and nuanced exercise under complete information conditions – i.e. daily base strategy weights and risk allocations are known, leaving benchmark related considerations as the sole focus. For obvious reasons, funds do not make such information publicly available, so performance analysis for non-clients entails imputing strategy exposures. Such an exercise is valuable, but inevitably introduces estimation error.

To determine fund strategy exposures across 85 base ARP composite strategy and 21 investable reference benchmarks (to identify any long beta bias), this paper utilizes elastic net (EN), a coefficient shrinkage or regularization technique combining least absolute shrinkage and selection operator (LASSO) regression and ridge regression (RR). The EN objective function is:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{E}} = \underset{\beta_0,\beta}{argmin} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \boldsymbol{\beta x}_i^T)^2$$
$$+ \lambda \sum_{j=1}^{p} \left(\frac{(1-\alpha)}{2} \boldsymbol{\beta}_j^2 + \alpha \left| \boldsymbol{\beta}_j \right| \right)$$

Equation 28

where $\boldsymbol{\beta_E}$ is the 1-by-p EN loadings on the independent variables, $\boldsymbol{X}$, to predict $\boldsymbol{Y}$ – respectively, n-by-p and n-by-1 matrices. N is the number of observations and p the number of predictors. α is a mixing parameter between 0 and 1 and λ is a non-negative regularization (penalty) parameter. EN approaches LASSO for α equal to 1 and is equivalent to RR for α equal to 0. This paper uses a value of 0.5.

This analysis sets the maximum number of predictors to 10 and uses 10-fold cross-validation to compute the mean squared error (MSE). The choice of 10 predictors reflects the small marginal impact of additional latitude. The following equation produces the fitted returns:

$$\widehat{\boldsymbol{Y}}_E = \boldsymbol{X}_m \widehat{\boldsymbol{\beta}_E} + \boldsymbol{\varepsilon}_E \qquad \textit{Equation 29}$$

$\boldsymbol{Y_E}$ in is an *m*-by-1 vector of fitted fund returns. $\boldsymbol{X_m}$ is the *m*-by-*p* matrix of regressors, the benchmark returns. $\boldsymbol{\varepsilon_E}$ is the EN stochastic disturbance -- an *m*-by-1 zero-mean normally distributed error vector, orthogonalized to $\boldsymbol{X_m \beta_E}$ with a variance proportional to the explanatory power of $\boldsymbol{X\beta_E}$.

To ensure that the EN coefficients explain the maximum possible percent of the variance of $\boldsymbol{Y}$ and to facilitate consistent comparison of funds, this paper scales the fitted returns as follows:

$$\widehat{\boldsymbol{S}} = \left(\widehat{\boldsymbol{Y}}_{1E}^T \widehat{\boldsymbol{Y}}_{1E}\right)^{-1} \widehat{\boldsymbol{Y}}_{1E}^T \boldsymbol{Y} \qquad \textit{Equation 30}$$

where the scalar $\boldsymbol{S}$ is a 2-by-1 vector — $S_{2,1}$ applies to $\boldsymbol{\beta_E}$ and $S_{1,1}$ is an intercept adjustment. The subscript $_{1E}$ denotes the inclusion of a vector of ones. Because the possibility exists that *Equation 5* reaches a local minimum, the strategy exposure estimation process repeats the calculation 1,000 times, selecting the coefficients producing the maximum adjusted R^2. Finally, the process does not impose a positive sign constraint upon coefficients since ARP is a long-short strategy that could manifest modest relative tilts among strategy benchmarks.

Table 25 provides a broad summary of the regression results and conveys four important points.

1. **Fund heterogeneity** – The median correlation among this group of diversified ARP funds is 0.4, with an interdecile range of 0.2 to 0.6. As previewed, ARP

fund managers make independent decisions on a wide range of portfolio parameters. The correlation profile contrasts with broadly consistent performance.

2. **Benchmark applicability** – The median adjusted R^2 among these funds is 54%, with an interdecile range of 34 to 78%. This result indicates that the ARP strategy benchmarks explain a significant proportion of fund return variance and that material specific variance exists for many funds.
3. **Weak benchmark relative return** – The median intercept is negative (a statistically insignificant 0.04% weekly or 2.3% annualized). Seven funds post a negative intercept significant at the 90% level while one fund posts a positive intercept with this significance. The negative average intercept is interesting given the use of what should be reasonably conservative net return ARP benchmarks. The possibility exists that some fund costs exceed those embedded in the benchmarks. Exposure misestimation, out-of-favor strategy specification, inopportune de-risking, or ill-timed tactical strategy allocation changes also may contribute to a negative intercept for a given fund. (Strategy timing was not a focus of most ARP fund investment processes due to the inherent difficulty, but recent poor performance unsurprisingly has increased interest in this capability.) Generally, the insignificant nature of most intercept estimates points to the suitability of the ARP benchmark structure and the ambiguous contribution of these intercept considerations to fund performance.
4. **Strong connection between idiosyncratic return and Sharpe ratio** – The rank correlation between intercept and Sharpe ratio is a very high 0.9. While not

particularly surprising given the average residual variance, this result punctuates the ironic reality that, in an investment category predicated upon factor footprint, performance during the 2018-2020 window requires being distinct from an expansive ARP benchmark universe -- perhaps through dynamic capital allocation, niche strategy use, or unique strategy specification.

Figure 26 summarizes the base ARP benchmarks capturing the majority of variance across the fund universe over the 2018-2020 period. The following formula produces the variance contribution:

$$\boldsymbol{\varphi} = \frac{\boldsymbol{w} \bullet (\boldsymbol{w}\boldsymbol{\Sigma})}{\boldsymbol{w}\boldsymbol{\Sigma}\boldsymbol{w}^T} \qquad \textit{Equation 31}$$

with $\boldsymbol{\varphi}$ being the proportional variance contribution, n the number of regressors, the 1-by-n vector $\boldsymbol{w}$ representing the loadings and $\boldsymbol{\Sigma}$ indicating the n-by-n covariance matrix. Benchmarks explain almost 60% of overall variance for this period. Panel A shows that 10, 20 and 30 benchmarks respectively account for 60, 80 and 90% of this total. Multi-asset trend, North American stock, and short equity volatility ARP benchmarks are the primary variance drivers.

Panel B aggregates the explained variance by broad benchmark. Multi-asset trend, equity sensitive, volatility sensitive and multi-factor stock strategies are most prominent. ARP benchmarks account for 82% of explained variance, reinforcing the unique nature of the investment category. This total arguably exceeds 90%. The contributions from the two long volatility reference benchmarks, in all cases carrying negative loadings, belong with the short equity volatility ARP benchmarks. Eliminating these benchmarks does not materially impact overall explanatory power – including them highlights some implementation nuance. The same dynamic applies to the much smaller contributions from FX reference benchmarks.

When interpreting risk contributions, bear in mind that diversified ARP fund managers target strategy balance, typically defined in multi-faceted terms. Realized volatility exceeding expectations in (or correlation with) a group such as equity sensitive strategies could skew 2018-2020 results. Also, a modest risk contribution from a highly diversifying strategy group such as rates carry may conceal a material underlying position size.

Lastly, the three funds with returns closest to target in Figure 25 are those most peripheral to the diversified ARP fund universe. Traditional long beta exposure explains almost 70% of total variance for one fund and approximately half of the explained variance for another. Both are significant outliers. The intercept for all three funds is positive, but the third fund carries the only significant positive reading in the universe – another outlier. These funds arguably do not belong in the universe, particularly if the ARP category exists principally to provide access to factors distinct from the traditional beta and idiosyncratic positions (alpha) available elsewhere. Regardless, the clear message is that printing positive returns during the 2018-2020 period requires maintaining a profile as distinct as possible from ARP.

2.5.3 ARP Index Performance Analysis via Broad Benchmarks

Figure 27 delivers additional fund manager perspective via the SG Multi Alternative Risk Premia Index. Such a peer universe index diversifies away some of the idiosyncrasies of individual managers and therefore provides an alternative lens through which to examine fund performance. Panel A uses the 10 broad ARP composite strategy benchmarks as explanatory variables in a 52-week rolling regression analysis. This exercise incorporates an additional year of returns (2017) to provide a baseline for exposures entering the 2018-2020 period. The exhibit shows the index variance explained by each ARP benchmark, with yellow highlighting the residual variance and dotted fill indicating a negative loading.

At a high level, Figure 27 reinforces the applicability of the ARP benchmarks, which explain on average 82% of index variance between 2018 and 2020. Few benchmarks receive negative loadings and those that do contribute very little variance. A rolling regression with up to 10 regressors naturally exhibits variability in results due to the evolution of the underlying covariance matrix. Therefore, given the volatile market backdrop, the ARP benchmark representation appears reasonable. Additionally, over the three-year window, the higher moments of the benchmark blend almost match those of the SG index – skewness of -2.3 versus -2.1 and kurtosis of 9.8 versus 9.3.

FX/multi-asset trend, equity sensitive, FX carry, and diversified stocks have relatively consistent footprints. Volatility sensitive, commodity spread/trend, and rates carry have meaningful, but more variable, representation. Commodity curve, value oriented and crude oil volatility do not register. The volatility crisis in February 2018 has an immediate and fleeting impact on the variance explained by equity sensitive strategies, while the COVID crisis in 2020 has a more persistent impact on the risk contribution from volatility sensitive strategies, which subsume the commodity spread footprint.

Note that the negligible loading on short crude oil volatility is not surprising. Practically, this category is not a point of emphasis for ARP fund managers, with any exposure typically nested within a diversified volatility program. The statistical benchmarking process identifies crude oil volatility as being sufficiently distinct and populated to warrant broad benchmark classification. However, if the broad benchmark tier included nine rather than ten constituents, crude oil volatility would collapse into volatility sensitive strategies.

Of course, the regression-based approach has limitations. Fund managers likely hold value-oriented and commodity curve positions. They do not abandon rates carry or commodity spread and trend positions in 2020. The regression attempts to disentangle estimates of manager positioning from shocks to the covariance structure, against a structural backdrop of ARP strategy groups with correlations ranging between -0.5 and 0.5. Further complicating matters, strategies such as time-series trend have a time-varying correlation structure depending upon the pro or countercyclicality of positioning -- crises catalyze defensive trend positioning, whereas recoveries have the opposite effect. Conversely, carry and relative value strategies have a comparatively static profile, with the former generally manifesting greater market sensitivity than the latter.

Principal component analysis (PCA) illustrates one facet of the exposure estimation challenge. Kritzman *et al.* (2010) propose the following systemic risk measure, termed the absorption ratio:

$$AR_n = \frac{\sum_{i=1}^{n} \sigma_{E_i}^2}{\sum_{j=1}^{m} \sigma_{B_j}^2} \qquad \textit{Equation 32}$$

The absorption ratio (*AR*) simply represents the fraction of variance explained by the first n eigenvectors (E) given m benchmarks (B) and σ^2 representing the return variance. Panel B of Figure 27 applies the absorption ratio concept to the broad ARP benchmarks (with n equal to three) and performs the PCA using both the covariance and correlation matrix to highlight the extent to which unexpected volatility or strategy connectivity confront managers.

ARP managers generally size positions based upon covariance estimates, so materially higher realizations of volatility or correlation result in ex post positions being higher than intended. Panel B shows the 2018 crisis to be a volatility event whereas the 2020 crisis produces both a volatility surprise and a reduction in strategy breadth. Importantly, the implication is that the connection among strategies was tighter entering the 2020 crisis, making the vulnerability to an event greater than in 2018.

Given these challenges and available data, the *effective* average manager positions in Figure 27 provide a reasonable characterization of ARP funds during the 2018-2020 period. Figure 28 punctuates the suitability of the ARP benchmarks for this exercise versus traditional long-only benchmarks. The traditional benchmarks explain approximately half of the SG index variance, a 40% reduction versus the ARP benchmarks. A degree of traditional benchmark explanatory power is not surprising given the macro orientation of many ARP strategies. To generate even this level of explanatory power, however, loadings on traditional benchmarks vary significantly and often are negative – conveying the dynamism of the ARP fund universe. The contrast between Figure 27 and Figure 28 is stark.

Figure 29 shows the annual return contributions accompanying the risk positioning estimates in Figure 27. Value-oriented, equity sensitive and FX/multi-asset trend drove the 2018 ARP underperformance. 2019 was a positive performance year, with positive idiosyncratic return being the swing factor. Value-oriented, diversified stocks and commodity spread/trend were significant detractors – more than offsetting positive contributions from rates and FX carry. Except for FX/multi-asset trend, rates carry and commodity curve, every sub-component of 2020 return was negative, with the primary performance detractors being volatility sensitive, FX carry, equity sensitive, diversified stocks (i.e. value light), and a large idiosyncratic element.

While the idiosyncratic contribution over the entire 2018-2020 period is zero, the large, likely overstated, negative impact in 2020 warrants additional discussion. The regression constant is an amalgamation of many factors, making generalization challenging. Structural considerations, such as a difference between manager fees and those embedded in net benchmarks, do not drive intercept time variation. The explanation resides in some combination of the following three sources.

1. *Exposure misestimation* – The lack of publicly available, ex ante strategy exposure data for ARP funds necessitates regression-based exposure estimation. This introduces the risk, particularly when financial market volatility is high, that effective positioning differs from actual exposure. Material performance dispersion among ARP strategy benchmarks then can lead to intercept inflation. Specifically, a large negative constant may reflect underestimation

(overestimation) of actual exposures to strategies generating below (above) average returns. This may occur because the underlying strategy covariance matrix makes it difficult for a regression to distinguish among strategies or because fund managers vary position sizes during the estimation window. The latter may be a byproduct of active strategy views or, more typically, volatility-based risk management and performance-driven drawdown control.

2. *Strategy misalignment* – The benchmarks used to estimate exposures may not reflect the strategies targeted by fund managers. The use of broad ARP composite benchmarks may miss intra-group allocation nuances or include a range of strategies outside the focus of fund managers. A large negative intercept could result from poor performance by "core" strategies or strong performance by peripheral strategies.
3. *Strategy specification* – A negative intercept could reflect average fund manager strategy implementations underperforming the average specification of investment banks captured by the ARP benchmarks.

Exposure misestimation likely accounts for most of the overstatement of the 2020 idiosyncratic return. Fund manager specifications systematically and significantly underperforming those of their investment banking counterparts is unlikely given the information sharing that occurs within the industry. Rerunning the analysis using the 85 base benchmarks instead of the 10 broad benchmarks does not meaningfully alter the intercept profile, so a lack of ARP strategy selection latitude (i.e. strategy misalignment) also is not driving the result.

The 2020 crisis was historic, and the profound impact on ARP fund performance almost certainly precipitated a range of portfolio management responses. Fund managers may have reduced relative position sizes to maintain target risk contributions, de-risked broadly to manage portfolio volatility, implemented stop-losses, or shifted strategy allocations due to tactical views. The rapid reversal in equity markets during Q2 makes whipsaw risk a likely source of negative idiosyncratic return. But, misestimation of strategy exposures appears to be the primary contributor to the large negative constant.

Shortening the length of the 52-week rolling window sheds light upon the underlying dynamics. The largest negative intercepts correspond to the periods spanning the V-shaped equity market move during Q1 and Q2. Volatility sensitive strategies absorb an outsized proportion of the estimated risk exposure as the regression wrestles with the market turbulence – a result indicative of narrow market breadth but not reflective of manger holdings. These strategies rebounded quicker than other ARP strategies as equity markets rallied in the second quarter. Overestimation of exposure in an outperforming strategy is a recipe for a negative intercept. Given the historically poor performance in stock-based strategies, any underestimation of exposure will reinforce this effect.

This observation does not undermine the usefulness of the overall fund analysis. Rather, the crowding out of exposure, for example, in commodity spread and trend by volatility sensitive strategies is a reminder of the challenges introduced by a highly charged environment and the need to treat attribution of ARP portfolios as a nuanced, triangulation exercise.

2.6 Extending ARP Return History

2.6.1 Important Considerations

To address the full breadth of ARP strategies, this paper uses the available 2000-2017 tradable bank index return history represented in statistical composite benchmarks to develop expectations for the recent three-year period. This approach raises a couple possible questions. Would additional history cast the 2018-2020 period in a different light by materially changing the distributional baseline assumptions based upon 2000-2017 ARP returns? Would using simple rather than composite benchmarks lead to different conclusions regarding disappointing ARP performance over the past three years?

A handful of researchers create extended return histories for a subset of ARP strategies. Baltussen *et al* (2019) investigate value, trend, momentum, seasonality, carry and risk anomaly returns across equity indices, government bonds, commodities and currencies over a 200-year period. Lempérière *et al* (2014b) explore trend returns in equity, bond, commodity and currency markets over a history of similar length. Doskov and Swinkels (2015) consider the carry trade across 20 currencies between 1900 and 2012. The Kenneth R. French Data Library provides almost a century of returns for the cross-sectional stock factors detailed in Fama and French (1993). These authors report results for the additional history broadly supporting those for more recent windows.

As no database like the Kenneth R. French Data Library exists for macro ARP strategies, reconstituting all these return histories is beyond the scope of this paper. To address the two questions raised above, this section instead focuses upon two important contributors to recent disappointing ARP portfolio performance — US (long-short) cross-sectional stock value and equity time-series trend.

Creating simple strategies to extend the return history necessarily represents a departure from the composite benchmarks utilized in this paper. As discussed in Gorman (2020), two classes of ARP benchmarks exist – composite and primitive. The former combines numerous tradable indices for a given ARP strategy to diversify the idiosyncrasies of any single specification. The latter, essentially a mimicking portfolio, employs a single, relatively simple strategy definition. Because no canonical specification exists for a given ARP strategy, oftentimes material performance dispersion among equally defensible primitive methodologies produces a practical benchmark selection dilemma. As this section highlights, no shortcut exists in ARP performance evaluation. Composite and primitive benchmarks play a complementary role in what ultimately is a triangulation exercise.

Designing a primitive benchmark with a long return history is a two-part challenge.

1. Data management – assembling historical inputs, applying data quality checks, and creating reasonable proxies for missing data
2. Strategy specification -- balancing parsimony and real-world applicability in both factor design and portfolio construction

Important caveats accompany an extended return history. Financial markets function very differently today than 50 or 100 years ago. Data quality is tenuous given the information was collected manually many decades ago and often is single sourced and may be smoothed or recorded at a low frequency. Therefore, the resultant returns play an indicative role, providing a qualified and approximate view of historical ARP strategy performance. None of the benchmarks in this section include trading costs.

2.6.1.1 US Cross-Sectional Stock Value (long-short)

This analysis includes six primitive representations of US stock cross-sectional value. Lakonishok *et al* (1994) advocate using different measures of value -- earnings yield, book yield, cash flow yield, and sales growth. Consistent with this research and the multi-factor approach of practitioners, the first four benchmarks specify value as an equally weighted combination of seven z-scores (applying a normal distribution to ranks for robustness) -- cashflow yield, free cashflow yield, trailing earnings yield, forward earnings yield (post-1985 given data availability), dividend yield, sales to price, and book to price. These calculations leverage the CRSP, S&P Compustat and I/B/E/S databases. The universe includes stocks with a market capitalization exceeding 80% of the median for the Dow Jones Industrial Average to produce an institutionally investable set of names and to ensure consistency over time. The benchmark rebalances monthly and is dollar-neutral, with long positions in the top quintile of value scores and short positions in the bottom quintile. Weekly return history begins in December 1954.

The first benchmark (Simple) equally weights the stocks within each quintile formed by ranking universe-wide composite z-scores. The second benchmark (Simple Constr) applies basic portfolio construction rigor to the first benchmark, constructing quintiles by ordering the residuals from a regression of composite z-score on 252-day trailing equity market beta, 252-day trailing volatility and market capitalization. This refinement tempers hitchhiking factor exposures typically managed in practice. The second benchmark weights stocks within each quintile by the log of market cap within each quintile to introduce a liquidity preference within an investable universe, another practical acknowledgement. The third benchmark (Simple Sec Neut) equally weights the stocks within each quintile formed by ranking sector-relative z-scores, recognizing fundamental differences among companies in different sectors and moderating incidental factor tilts. The fourth benchmark (Simple Sec Neut Constr) applies the z-score orthogonalization of the second benchmark, including a set of sector dummy variables.

The fifth primitive benchmark is the Bloomberg GSAM US Equity Value L/S Index (Bloomberg GSAM). This benchmark calculates an equally-weighted average of z-scores for (winsorized) book to price, sales to price, earnings to price, cash flow to price, forward earnings to price, and dividend to price for the 500 largest US securities. The benchmark takes long positions in the 150 highest ranked securities and a short position in the benchmark index futures contract, weighting stock positions by the square root of market capitalization and rebalancing quarterly. Weekly return history begins in January 2000.

The final value benchmark is the Fama-French high-minus-low factor (FF HML) from The Kenneth R. French Data Library. HML is the average return on the two Fama-French value portfolios (big and small as determined by the median NYSE market cap) minus the average return on the two growth portfolios. Value includes stocks above the 70th NYSE percentile for previous year book equity to market cap, while growth includes those below the 30th percentile. Portfolios include all NYSE, AMEX, and NASDAQ stocks, value weight positions, and reconstitute annually. Weekly return history begins in July 1926.

2.6.1.2 Equity Time-Series Trend

This analysis includes three primitive benchmarks for equity time-series trend. The first two utilize a daily total return history for the S&P 500 starting in December 1911. While equity trend typically includes the major global equity markets, this simple specification focuses on the US partly due to data considerations and partly because time-series trend is among the most homogenous ARP strategies, so including only the S&P 500 may be simultaneously reductionist and representative. The US-only approach also acknowledges that country inclusion varies among equity trend indices, with some strategies focusing on only three or four markets and others including all liquid (with some thresholds more accommodating than others) futures.

Constructing a daily history of S&P 500 futures returns highlights the data challenges accompanying long-term analysis. The Chicago Mercantile Exchange (CME) introduced the S&P 500 futures contract in 1982. Considering contract uptake and data availability, this analysis uses a roll-adjusted, front-month futures return series beginning in 1990. Prior to that date, returns represent a synthetic futures return – the total return on the S&P 500 minus a LIBOR financing rate.

Daily price returns for the S&P 500 are available back to 1928 (although fewer than 500 names comprise the index prior to 1957). The New York Times Combined Average extends the daily price return series for US equities back to 1911. Daily income returns are extrapolated from a monthly US equity dividend yield series from Global Financial Data (GFD). The LIBOR history begins in 1971. Prior to that, the short-term treasury bill return is a function of US treasury yield data from Federal Reserve Economic Data, Ibbotson Associates and GFD, and a spread-adjusted commercial paper yield prior to

1919. The LIBOR spread over treasury bills is a loglinear estimate based upon trailing S&P 500 price return volatility, the treasury bill rate, the BAA-AAA credit spread, and the commercial paper spread.

The first primitive trend benchmark (Simple) takes a long or short position in the S&P 500 futures consistent with the sign of the trailing 260-day return. Implementation occurs with a one-day lag. The second primitive benchmark (Simple Confirm) includes a basic confirmation signal to avoid taking a position in a weak trend. An absolute trailing 260-day z-score (assuming a mean of zero) less than 0.4 corresponds to no position. This cutoff eliminates the bottom quintile of least compelling trends over the full history.

The third primitive benchmark is the Bloomberg GSAM Equity Trend Index (Bloomberg GSAM). The signal for each of the 11 equity index futures is the average of twelve binary directional indicators corresponding to the sign of the excess return from each of the previous twelve months. The signal aims to capture both trend direction and strength. Positions rebalance weekly, with a leverage factor, targeting an equal contribution to benchmark volatility, scaling the signal for each futures position.

2.6.2 Comparing Returns from Recent and Preceding Decades

Figure 30 places the 2018-2020 Sharpe ratio for the simple US value (left) and trend (right) benchmarks in long-term historical context. The recent results for all five value benchmarks effectively represent a 65-year low. Using the return history through 2017 to create a bootstrapped distribution of 5,000 three-year Sharpe ratios for each benchmark reveals the 2018-2020 observation uniformly to be in the 99th percentile. Given the amount of history, this is consistent with the empirical result. Regardless of the benchmark methodology, the past three years represent an extraordinary failing for cross-sectional value in US stocks.

Time-series equity trend also fared poorly over the past few years, just not to the same extent as value. The 2018-2020 three-year Sharpe ratio for the simple strategy is in the 89th (93rd) percentile of the bootstrapped (empirical) distribution. The simple strategy with confirmation better weathered the recent market environment by remaining idle for almost 40% of the period — twice the historical level of inactivity. The three-year Sharpe ratio for this benchmark is in the 72nd (81st) percentile of the bootstrapped (empirical) distribution.

Table 26 provides a statistical summary for each of the three time periods of interest for the value benchmarks — the 1955-1999 early history, 2000-2017 distributional baseline window, and 2018-2020 crisis. For the latter two periods, this exhibit includes the Bloomberg GSAM index and the relevant base statistical composite used in this paper. The Sharpe ratio, skewness and correlation profile among the five simple benchmarks is very similar for the 1955-1999 and 2000-2017 periods, indicating that the additional data would not change meaningfully the distributional assumptions for value in this paper. The correlation with the S&P 500 during the early period is slightly more negative, but this only reinforces the expectation here that value strategies function as a market diversifier in ARP portfolios.

The Bloomberg GSAM index is less correlated with the other value representations over the 2000-2017 period, likely a byproduct of using a market hedge as opposed to shorting individual stocks. The statistical composite is comparatively more correlated with the primitive benchmarks including portfolio construction constraints, as tradeable indices incorporate a variety of risk control measures.

Results for the past few years across the benchmarks confirm the findings in this paper for value in stocks — very negative Sharpe ratios, increased correlation among strategies and with the equity market, and less-than-expected positive skewness. Of note, the Sharpe ratio for the statistical composite is much lower than that of the primitive benchmarks. The tradeable indices often increase portfolio construction rigor and narrow the universe of stocks to respect institutional liquidity and short sale considerations. The 2018-2020 period clearly did not reward such methodological refinements.

Table 27 provides the same statistical summary for the equity trend benchmarks. The Sharpe ratio for the primitive benchmarks during the 1912-1999 period is relatively low, but this is typical for a single-market time-series trend strategy – the diversification from combining the approach across multiple markets is necessary to boost the Sharpe ratio. As with value, the results for the early window and the 2000-2017 baseline period are consistent. The Sharpe ratios are similar, with the confirmation strategy posting a slightly higher Sharpe ratio. The correlation between the simple strategies is very high and the correlation with the equity market very low. The skewness is negative in the 1912-1999 period, reflecting the significant volatility of the 1930's and contrasting the slightly positive result for the baseline window.

The Bloomberg GSAM index and the statistical composite benchmark are highly correlated during the 2000-2017 and 2018-2020 periods, although the correlation is marginally lower during the latter period. The Sharpe for the statistical composite is slightly higher (lower) during the baseline (recent) window; however, the profile for both benchmarks is quite similar during the crisis period — negative Sharpe ratios, negative skewness, and higher correlation with the equity market during the 2018-2020 window. This profile is consistent with that of the two primitive strategies, with the confirmation strategy faring relatively well by avoiding trend trading. The choice of benchmark does not change the conclusions in this paper regarding the role of equity trend in recent disappointing performance across ARP portfolios.

2.6.2.1 Benchmark Return Dispersion

If the creation of additional history does not change the distributional expectations used in this paper and if simple benchmarks lead to similar conclusions regarding the 2018-2020 period (for the two ARP strategies under consideration), does the choice of primitive or composite benchmark matter? The important point here is that this is not an either-or proposition. Primitive and composite benchmarks are complementary, with both playing an important role. While the two approaches may lead to similar conclusions regarding the general trajectory of ARP strategy performance, such a finding is only possible with both sets of benchmarks. This dual perspective becomes increasingly important as the focus shifts to studying the returns of an individual ARP fund manager.

Recall that the composites aggregate essentially all the individual indices traded by institutional investors. As such, they capture the breadth of *implemented* indices across the ARP space. Primitive benchmarks represent a single, relatively simple approach among numerous candidates. The tracking risk among these possible methodologies depends upon the consistency of a given ARP strategy — here, equity trend is a relatively homogenous strategy while US stock value is much less so. Statistical composites have the benefit of breadth, availability and investor relevance. Building a primitive benchmark for every ARP strategy, including a long-term return series, would be valuable and represents yet another possible research project. (Bloomberg GSAM currently offers 13 primitive benchmarks to align with the 85 base statistical composites, so work remains to be done.)

Return variation among primitive ARP benchmarks with equally defensible rules and no theoretical arbiter necessitates a multi-faceted approach to performance evaluation. The left panel in Figure 31 highlights the annual return variation among the simple US stock value benchmarks since 2000. The median spread among annual returns is six percent for benchmarks targeting seven percent volatility. The annual spread often exceeds 10 percent, with 2020 being the latest example. The variation among the trend returns in the right panel, while non-trivial, is smaller and arguably inflated by the US orientation of the simple benchmarks. Such return spreads are extremely relevant when evaluating the performance of a specific ARP portfolio manager in a given year. Context is essential.

Figure 32 illustrates the complementary roles of primitive and statistical composite benchmarks for the 2018-2020 period. Competing simple methodologies should contextualize any primitive benchmark. The alternative specifications discussed here frame the Bloomberg GSAM indices for US stock value and equity trend. Similarly, the individual tradable indices comprising the statistical composite provide the methodological return spread. Understanding the methodological drivers of return differences is a critical step in understanding the performance of a given fund manager.

(Tackling the databasing challenge of both primitive and composite benchmark positions represents the next frontier of performance analysis but will not replace the first contextual step.)

Only with both composite and primitive benchmarks and the necessary context can one properly explore ARP fund manager performance. For example, the statistical US value composite underperformed the various primitive benchmarks over the 2018-2020 period. Is this a case of style headwind for the composite (the market temporarily not rewarding conventional methodologies), overspecification by tradable index providers (noise accompanying unnecessarily complex rules), underspecification in primitive benchmarks (excluding important implementation considerations), or the composite including strategies with very similar statistical footprints but different opportunity sets (e.g. global value and US value)?

Over the past three years, each of these considerations is relevant. Value strategies employing portfolio optimization and shorting individual stocks (particularly larger growth stocks) suffered greatly, but this observation does not invalidate these approaches as reasonable comparators for fund manager implementations. The combination of primitive and composite benchmarks reveals such strategy dynamics and facilitates more robust performance insights.

The 2018-2020 annualized return spread among alternative value indices is more than 10% within the composite and four percent across the simple candidates. Return variation among the equity trend approaches is comparatively modest, particularly if one acknowledges that the two upside outliers incorporate confirmation signals and macroeconomic indicators. These two examples are indicative of the varying heterogeneity that characterizes the broad universe of ARP strategies. There is no shortcut in ARP performance analysis, no incontrovertible benchmark. Only with a mosaic like that in Figure 32 can one properly frame the questions and establish the appropriate context to assess the strategy execution of a given fund manager.

2.7 Summary

The momentum supporting ARP entering 2018, the sobering reality of subsequent returns, the breadth of back-tested research, the true information content of factors, the impact of an historic crisis, the resilience of traditional benchmarks – the 2018-2020 alternative risk premium experience abounds with questions, controversies, and apparent contradictions. This paper lays the foundation for the soul searching and analysis that undoubtedly will continue regarding the role of ARP in diversified portfolios.

The performance of ARP over the 2018-2020 period was extremely disappointing. Fund returns in 2018 were weak, the rebound in 2019 modest, and the results in 2020 terrible. Traditional benchmarks rebounded quickly after each crisis while ARP strategies shared the pain but not the recovery. Frustration and disillusionment mounted, resulting in reported AUM across the ARP funds analyzed in the preceding pages halving by the end of the period.

Using proprietary composite ARP benchmarks to provide the necessary strategy grouping and performance granularity, this paper seeks to answer the following question. *What distributional expectations prevailed among ARP portfolio managers at the end of 2017 (reflecting relevant data and data mining considerations), and what deviations from those expectations during 2018-2020 were most responsible for the disappointing result?*

The analysis proceeds methodically through the following fundamental ARP portfolio construction inputs, applying a battery of tests to assess the significance of recent deviations from the distributional baseline.

- Returns & Volatility (Sharpe ratio)
- Correlation (among ARP strategies, with traditional benchmarks)
- Non-normality (skewness, kurtosis)
- Conditional Returns (payoffs in extreme states, turbulent period

Table 28 summarizes the findings of this investigation. The analysis supports most assumptions regarding ARP return distributions for this three-year period being reasonable despite the financial market turmoil. Four strategy groups surface repeatedly as exceptions — equity sensitive, volatility sensitive, diversified stocks and value oriented. The first two are from the risk-seeking ARP sub-class and were overwhelmed by successive equity market crises. The latter two, diversified stocks from the defensive sub-class and value oriented from the diversifier sub-class, wilted under historic losses from well-established quantitative stock selection strategies.

These are the strategies with the most significant deviations from Sharpe ratio and skewness expectations and the most problematic increases in correlation with both other strategies and traditional risk assets. These strategies factor prominently in the ARP fund analysis results, playing a recurring role in the realized experience of investors. Importantly, diversified stock strategies register the most disappointing departure from state-based payoff expectations and arguably represent the most consequential breakdown of the 2018-2020 period given the size and timing of the drawdown.

While drawing inferences from a relatively short window impacted by an historic crisis and rebound is extremely challenging, the risk-seeking strategy group appears to be an environmental casualty. As highlighted in Figure 21, elevated volatility states represented 40% of the 2018-2020 period, twice the 20% prevalence over the 2000-2017 window. Two volatility crises in the span of two years create an incredible headwind for volatility sensitive strategies dependent upon market stability to accrue the variance risk premium. Whether more frequent volatility storms are period specific or the new norm could be a future research topic.

Repeated V-shaped equity market recoveries undermine equity trend strategies. Appendix I shows that the velocity of significant US equity drawdowns in 2018-2020 was 2.5 times that in the preceding two decades (larger losses over shorter durations). The rate of significant recovery was almost three times that in the earlier period. Significant drawdown cycles account for almost 75% of weekly observations, and the acceleration of return realization increases whipsaw risk for a variety of ARP strategies, most notably trend. Whether more intense drawdown cycles are unique to the period or indicative of a structural shift should be the subject of future research, with consequences for the parameterization and portfolio role of vulnerable strategies. Given the unique environmental backdrop, however, a referendum on the prospective Sharpe ratio of volatility and equity sensitive strategies appears premature.

The more significant problem resides with the stock-based strategies. As detailed by the authors cited in the introduction to this paper, a narrowly-driven large cap growth market created the equivalent of a 100-year flood for quantitative equity managers. The Sharpe ratio of the global value factor in Figure 15 plunged from +1.0 between 2000 and

2017 to -1.0 in the recent window. The embodiment of the brutal short side of the value trade, the Goldman Sachs Non-Profitable Tech Company Index, produced a staggering annual excess return of 50% and a Sharpe ratio of 1.5 over the 2018-2020 period. Many questions remain regarding the causes and likely persistence of this quantitative equity crisis. The evolving nature of value stocks, possible crowding by fundamental investors fixated on growth stocks, new retail trading platforms, and whether large growth companies now represent “safe-harbor” stocks are just a few of the topics that warrant investigation.

These stock-based strategies play an integral diversifying role in ARP portfolios. While diversification is not a hedge (i.e. low correlation certainly does not guarantee an offsetting return), experiencing a multi-year, historic drawdown in stock-based strategies concurrent with successive historic volatility crises is a tortuous alignment of possible outcomes. The confluence of events was crushing for ARP strategies, particularly with so few strategies thriving that were not in the environmental crosshairs.

As this paper demonstrates, ARP applies a wide range of strategies and defies simple performance explanations. Crowding is one such all-too-convenient narrative. While short volatility trades were crowded in early 2018, the events of February of that year rebalanced the market. The results for volatility sensitive strategies in 2020 were not a byproduct of crowding, but rather the consequence of an exogenous global health shock. This experience is the justification for, not an invalidation of, the economic rationale supporting the variance risk premium. Unfortunately, the timing of the COVID crisis, arriving so soon after the events of 2018, created significant headwinds for ARP strategies.

Similarly, equity/credit trend strategies did not get whipsawed repeatedly due to crowding. Crowding should push the returns of a divergent strategy higher. Value stock strategies did not languish due to crowding. Crowding in a convergent strategy should close valuation gaps and increase turnover — arguably, the crowding existed among those on the other side of the trade pushing the prices of expensive stocks higher. Liquidation of quantitative equity strategies during 2020 may have been a marginal contributor to, but not the primary driver of, demand for growth stocks. The results in this paper indicate that environmental forces were more responsible for recent ARP performance than crowding among ARP fund managers. Of course, a rigorous exploration of crowding in this space during the past few years represents yet another research opportunity.

This paper sets the table, framing in detail what transpired over the past few years, but much work remains to be done in the ARP space. For example, with the specter of the COVID crisis looming so large, evaluating the below-target Sharpe ratio for most strategies over the 2018-2020 period represents both a test of a single data point versus its distributional expectation and an event study. Confirming the expected Sharpe ratio of sub-strategies requires further attention, since the investigation here identifies limited damage from distributional assumptions other than the first moment – and the small sample size of the recent window obviously creates statistical inference challenges for the first moment.

The recent drawdown should not be a death knell for ARP strategies, but rather a research catalyst. Interestingly, these strategies have rebounded during the first few months of 2021, despite previously resilient rate carry trades being stung by the spike in bond yields — perhaps the beginning of the next chapter of this story.

References

Ackerman, C., R. McNally, and D. Ravenscraft. "The Performance of Hedge Funds: Risk, Return, and Incentives." *Journal of Finance*, vol. 54, no. 3 (June 1999), pp. 833–874.

Amenc, N., L. Martellini, F. Goltz and S. Ye. "Improved Beta? A Comparison of Index-Weighting Schemes." EDHEC-Risk Institute Publication, September 2011.

Anderson, T. and D. Darling. "A Test of Goodness-of-Fit." *Journal of the American Statistical Association*, vol. 49, no. 268 (December 1954), pp. 765–769.

Arnott, R., C. Harvey, V. Kalesnik and J. Linnainmaa. "Reports of Value's Death May Be Greatly Exaggerated." Working Paper, February 2021.

Auer, B. and F. Schuhmacher. "Performance Hypothesis Testing with the Sharpe Ratio: The Case of Hedge Funds." *Finance Research Letters*, vol. 10, no. 4 (December 2013), pp. 196–208.

Baig, A., B. Blau, H. Butt and A. Yasin. "Do Retail Traders Destabilize Financial Markets? An Investigation Surrounding the COVID-19 Pandemic." Working Paper, February 2021.

Bailey, D. and M. López de Prado. "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality." *Journal of Portfolio Management*, vol. 40 no. 5 (40th Anniversary 2014), pp. 94-107.

Bailey, D. and M. López de Prado. "The Sharpe Ratio Efficient Frontier." *Journal of Risk*, vol. 15, no. 2 (December 2012), pp. 3-44.

Baltas, N. and B. Scherer. "Tail Risk in the Cross Section of Alternative Risk Premium Strategies." *Journal of Portfolio Management*, vol. 45, no. 2 (Multi-Asset Special Issue 2019), pp. 93-104.

Baltussen, G., L. Swinkels, and P. van Vliet. "Global Factor Premiums." Working Paper, January 2019.

Bellone, B., T. Heckel, F. Soupé and R. Carvalho. "Equity Factor Investing: Historical Perspective of Recent Performance." Working Paper, October 2020.

Bertsimas, D., C. Pawlowski and Y. Zhuo. "From Predictive Methods to Missing Data Imputation: An Optimization Approach." *Journal of Machine Learning Research*, vol. 18, no. 196 (March 2018), pp. 1-39.

Blashfield, R. "Mixture Model Tests of Cluster Analysis: Accuracy of Four Agglomerative Hierarchical Methods." *Psychological Bulletin*, vol. 83, no. 3 (1976), pp. 377-388.

Blitz, D. "The Quant Crisis of 2018-2020: Cornered by Big Growth." Working Paper, January 2021.

Boulesteix, A. and K. Strimmer. "Partial Least Squares: A Versatile Tool for the Analysis of High-Dimensional Genomic Data." *Briefings in Bioinformatics*, vol. 8, no. 1 (May 2006), pp. 32-44.

Brown, S., W. Goetzmann, R. Ibbotson, and S. Ross. "Offshore Hedge Funds: Survival and Performance, 1989–95." *Journal of Business*, vol. 72, no. 1 (January 1999), pp. 91-117.

Calinski, T. and J. Harabasz. "A Dendrite Method for Cluster Analysis." *Communications in Statistics*, vol. 3, no. 1 (December 1973), pp. 1–27.

Choueifaty, Y. and Y. Coignard. "Toward Maximum Diversification." *Journal of Portfolio Management*, vol. 35, no. 1 (Fall 2008), pp. 40-51.

Chow, G., E. Jacquier, M. Kritzman, and K. Lowry. "Optimal Portfolios in Good Times and Bad." *Financial Analysts Journal*, vol. 55, no. 3 (May/June 1999), pp. 65-73.

Clarke, R., H. De Silva and S. Thorley. "Risk Parity, Maximum Diversification, and Minimum Variance: An Analytic Perspective." *Journal of Portfolio Management*, vol. 39, no. 3 (Spring 2013), pp.39-53.

D'Agostino, R., A. Belanger and R. D'Agostino, Jr. "A Suggestion for Using Powerful and Informative Tests of Normality." *The American Statistician*, vol. 44, no. 4 (November 1990), pp. 316-321.

Davies, D. and D. Bouldin. "A Cluster Separation Measure." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 2 (April 1979), pp. 224–227.

De Jong, S. "SIMPLS: An Alternative Approach to Partial Least Squares Regression." *Chemometrics and Intelligent Laboratory Systems*, vol. 18, no. 3 (March 1993), pp. 251–263.

Djauhari, M. "Minimal Spanning Tree Problem in Stock Networks Analysis: An Efficient Algorithm." *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 9 (May 2013), pp. 2226–2234.

Doskov, N. and Swinkels, L. "Empirical Evidence on the Currency Carry Trade, 1900-2012." *Journal of International Money and Finance*, vol. 51 (March 2015), pp. 370-389.

Fama, E. and K. French. “Common Risk Factors in the Returns of Stocks and Bonds.” *Journal of Financial Economics*, vol. 33, no.1 (February 1993), pp. 3-56.

Ferreira, L. and D. Hitchcock. "A Comparison of Hierarchical Methods for Clustering Functional Data." *Communications in Statistics - Simulation and Computation*, vol. 38, no. 9 (December 2009), pp. 1925-1949.

Fisher, R. "On the 'Probable Error' of a Coefficient of Correlation Deduced from a Small Sample." *Metron*, vol, 1, no. 4 (September 1921), pp. 3-32.

Fowlkes, E. and C. Mallows. "A Method for Comparing Two Hierarchical Clusterings." *Journal of the American Statistical Association*, vol. 78, no. 383 (September 1983), pp. 553-569.

Frank, I. and J. Friedman. "A Statistical View of Some Chemometrics Regression Tools." *Technometrics*, vol. 35, no. 2 (May 1993), pp. 109-135.

Fung, W. and D. Hsieh. "Hedge-Fund Benchmarks: Information Content and Biases." *Financial Analysts Journal*, vol. 58, no. 1 (January/February 2002), pp. 22-34.

Fung, W. and D. Hsieh. "Performance Characteristics of Hedge Funds and Commodity Funds: Natural vs. Spurious Biases." *Journal of Financial and Quantitative Analysis*, vol. 35, no. 3 (September 2000), pp. 291–307.

Fung, W. and D. Hsieh. “Empirical Characteristics of Dynamic Trading Strategies: The Case of Hedge Funds.” *Review of Financial Studies*, vol. 10, no. 2 (April 1997), pp. 275–302.

Glerean, E., R. Pan, J. Salmi, R, Kujala, J. Lahnakoski, U. Roine, L. Nummenmaa, S. Leppämäki, T. Nieminen-von Wendt, P. Tani, J. Saramäki, M. Sams, and I. Jääskeläinen. "Reorganization of Functionally Connected Brain Subnetworks in High-Functioning Autism." *Human Brain Mapping*, vol. 37, no. 3 (March 2016), pp. 1066-1079.

Gorman, S. “The ABC’s of ARP: The Genesis and Anatomy of Alternative Risk Premium.” Working Paper, December 2019.

Gorman, S. “The Data Dilemma in Alternative Risk Premium: Why Is a Benchmark So Elusive?” Working Paper, December 2020.

Hamdan, R., F. Pavlowsky, T. Roncalli, and B. Zheng. “A Primer on Alternative Risk Premia.” Lyxor Research Paper, April 2016.

Hands, S. and B. Everitt. "A Monte Carlo Study of the Recovery of Cluster Structure in Binary Data by Hierarchical Clustering." *Multivariate Behavioral Research*, vol. 22, no. 2 (April 1987), pp. 235-243.

Harvey, C. and Y. Liu. "Backtesting." *Journal of Portfolio Management*, vol. 42, no. 1 (Fall 2015), pp. 12-28.

Harvey, C., Y. Liu, and H. Zhu. ". . . and the Cross-Section of Expected Returns." *Review of Financial Studies*, vol. 29, no. 1 (January 2016), pp. 5-68.

Hubert, L. and P. Arabie. "Comparing Partitions." *Journal of Classification*, vol. 2, no. 1 (December 1985), pp. 193-218.

Israel, R., K. Laursen and S. Richardson. "Is (Systematic) Value Investing Dead?", Working Paper, October 2020.

Jarque, C. and A. Bera. "A Test for Normality of Observations and Regression Residuals." *International Statistical Review*, vol. 55, no. 2 (August 1987), pp. 163–172.

Jobson, J. and M. Korkie. "Performance Hypothesis Testing with the Sharpe and Treynor Measures." *Journal of Finance*, vol. 36, no. 4 (September 1981), pp. 889-908.

Koeppel, C. "Does Social Media Sentiment Matter in the Pricing of U.S. Stocks?" Working Paper, January 2021.

Kritzman, M. and Y. Li. "Skulls, Financial Turbulence, and Risk Management." *Financial Analysts Journal*, vol. 66, no. 5 (September/October 2010), pp. 30-41.

Kritzman, M., Y. Li, S. Page and R. Rigobon. "Principal Components as a Measure of Systemic Risk." Working Paper, June 2010.

Kruskal, J. "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem." *Proceedings of the American Mathematical Society*, vol. 7, no. 1 (February 1956), pp. 48-50.

Krzanowski, W. and Y. Lai. "A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering." *Biometrics*, vol. 44, no. 1 (March 1988), pp. 23-34.

Kuiper, F. and L. Fisher. "A Monte Carlo Comparison of Six Clustering Procedures." *Biometrics*, vol. 31, no. 3 (September 1975), pp. 777-783.

Lakonishok, J., A. Shleifer, and R. Vishny. "Contrarian Investment, Extrapolation, and Risk." *Journal of Finance*, vol. 49, no. 5 (December 1994), pp.1541-1578.

Ledoit, O. and M. Wolf. "Robust Performance Hypothesis Testing with the Sharpe Ratio." *Journal of Empirical Finance*, vol. 15, no. 5 (December 2008), pp. 850-859.

Lempérière, Y., C. Deremble, P. Seager, M. Potters, and J. P. Bouchaud. "Two Centuries of Trend Following." *Journal of Investment Strategies*, vol. 3, no. 3 (April 2014), pp. 41–61.

Lempérière, Y., C. Deremble, T. T. Nguyen, P. Seager, M. Potters, and J. P. Bouchaud. "Risk Premia: Asymmetric Tail Risks and Excess Returns." *Quantitative Finance* (September 2014), pp. 1-25.

Lettau, M., M. Maggiori, and M. Weber. "Conditional Risk Premia in Currency Markets and Other Asset Classes." *Journal of Financial Economics*, vol. 114, no. 2 (November 2014), pp. 197-225.

Lev, B. and A. Srivastava. "Explaining the Recent Failure of Value Investing." Working Paper, March 2020.

Liang, B. "Hedge Funds: The Living and the Dead." *Journal of Financial and Quantitative Analysis*, vol. 35, no. 3 (September 2000), pp. 309–326.

Liang, B. "On the Performance of Hedge Funds." *Financial Analysts Journal*, vol. 55, no. 4 (July/August 1999), pp. 72-85.

Lilliefors, H. "On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown." *Journal of the American Statistical Association*, vol. 62, no 318 (June 1967), pp. 399–402.

Lo, A. "The Statistics of Sharpe Ratios." *Financial Analysts Journal*, vol. 58, no. 4 (July/August 2002), pp. 36-52.

Maginn, J., Tuttle, D., McLeavey, D. and Pinto, J. (2007), *Managing Investment Portfolios: A Dynamic Process*, Hoboken, NJ: John Wiley & Sons, Inc.

Mahalanobis, P. "On the Generalized Distance in Statistics." *Proceedings of the National Institute of Sciences of India*, vol. 2, no. 1 (January 1936), pp. 49–55.

Mantegna, R. "Hierarchical Structure in Financial Markets." *European Physical Journal B - Condensed Matter and Complex Systems*, vol. 11, September 1999, pp. 193–197.

Mantegna, R. and H. Stanley (2000), *An Introduction to Econophysics: Correlations and Complexity in Finance*, Cambridge, UK: Cambridge University Press.

Mantel, N. "The Detection of Disease Clustering and a Generalized Regression Approach." *Cancer Research*, vol 27, no. 2 (February 1967), pp. 209-220.

Massey, F. “The Kolmogorov-Smirnov Test for Goodness of Fit.” *Journal of the American Statistical Association*, vol. 46, no. 253 (March 1951), pp. 68–78.

McLean, D., and J. Pontiff. “Does Academic Research Destroy Stock Return Predictability?” *Journal of Finance*, vol. 71, no. 1 (February 2016), pp. 5-32.

Memmel, C. "Performance Hypothesis Testing with the Sharpe Ratio." Working Paper, June 2003.

Mertens, E. "Comments on Variance of the IID Estimator in Lo (2002)." Working Paper, 2002.

Milligan, G. and M. Cooper. "A Study of Standardization of Variables in Cluster Analysis." *Journal of Classification*, vol. 5, no. 2 (September 1988), pp. 181-204.

Molenberghs, G., G. Fitzmaurice, M. Kenward, A. Tsiatis and G. Verbeke (2015), *Handbook of Missing Data Methodology*, Boca Raton, FL: CRC Press, Taylor & Francis Group.

Morlini, I. and S. Zani. "An Overall Index for Comparing Hierarchical Clusterings." Conference Paper: Challenges at the Interface of Data Analysis, Computer Science, and Optimization, Proceedings of the 34th Annual Conference of the Gesellschaft für Klassifikation e. V., Karlsruhe, July 21 - 23, 2010.

Naya, F. and N. Tuchschmid. "Alternative Risk Premia: Is the Selection Process Important?" *Journal of Wealth Management*, vol. 22, no. 1 (Summer 2019), pp. 25-38.

Newey, W. and K. West. “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix.” *Econometrica*, vol. 55, no. 3 (May 1987), pp. 703-708.

Opdyke, J. "Comparing Sharpe Ratios: So Where Are the p-values?" *Journal of Asset Management*, vol. 8, no. 5 (December 2007), pp. 308–336.

Pagano, M., J. Sedunov and R. Velthuis. "How Did Retail Investors Respond to the COVID-19 Pandemic? The Effect of Robinhood Brokerage Customers on Market Quality." Working Paper, November 2020.

Prim, R. "Shortest Connection Networks and Some Generalizations." *Bell System Technical Journal*, vol. 36, no. 6 (November 1957), pp. 1389-1401.

Rand, W. "Objective Criteria for the Evaluation of Clustering Methods." *Journal of the American Statistical Association*, vol. 66, no. 336 (December 1971), pp. 846-850.

Rouseeuw, P. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics*, vol. 20, no. 1 (November 1987), pp. 53–65.

Rubin, D. "Inference and Missing Data." *Biometrika*, vol. 63, no. 3 (December 1976), pp. 581–592

Sammon, J. "A Nonlinear Mapping for Data Structure Analysis." *IEEE Transactions on Computers*, vol. 18, no. 5 (May 1969), pp. 401-409.

Shapiro, S. and M. Wilk. "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika*, vol. 52, no. 3/4 (December 1965), pp. 591–611.

Shea, Y. and E. Radatz. "Searching for Inner Peace with Value Factors." *Risk & Reward*, 4th issue, September 2020, pp. 14-19.

Sokal, R. and C. Michener. "A Statistical Method for Evaluating Systematic Relationships." *University of Kansas Science Bulletin*, vol. 38, no. 22 (1958), pp. 1409-1438.

Suhonen, A., J. Bryant, S. Gao, and A. Cheung. "The MJ Hudson Allenbridge Alternative Risk Premia Fund Review 2019." February 2019.

Suhonen, A., M. Lennkh, and F. Perez. "Quantifying Backtest Overfitting in Alternative Beta Strategies." *Journal of Portfolio Management*, vol. 43, no. 2 (Winter 2017), pp. 90-104.

Thorndike, R. "Who Belongs in the Family?" *Psychometrika*, vol. 18, no. 4 (December 1953). pp. 267–276.

Tibshirani, R. "Regression Shrinkage and Selection Via the Lasso: A Retrospective." *Journal of the Royal Statistical Society*, Series B, vol. 73, no. 3 (May 2011), pp. 273-282.

Tibshirani, R., G. Walther and T. Hastie. "Estimating the Number of Clusters in a Data Set via the Gap Statistic." *Journal of the Royal Statistical Society*, Series B (Statistical Methodology), vol. 63, no. 2 (2001), pp. 411-423.

Vatanen, K. and A. Suhonen. "A Framework for Risk Premia Investing: Anywhere to Hide?" Working Paper, July 2019.

Waldmann, P., G. Meszaros, B. Gredler, C. Fuerst and J. Solkner. "Evaluation of the Lasso and the Elastic Net in Genome-wide Association Studies." *Frontiers in Genetics*, vol. 4, no. 270 (December 2013), pp. 1-11.

Wang, G., C. Xie, Y. Chen, and S. Chen. "Statistical Properties of the Foreign Exchange Network at Different Time Scales: Evidence from Detrended Cross-Correlation Coefficient and Minimum Spanning Tree." *Entropy*, vol. 15, no. 5 (May 2013), pp. 1643-1662.

Ward, J. "Hierarchical Grouping to Optimize an Objective Function." *Journal of the American Statistical Association*, vol. 58, no. 301 (March 1963), pp. 236-244.

Warrens, M.J. "On the Equivalence of Cohen’s Kappa and the Hubert-Arabie Adjusted Rand index." *Journal of Classification*, vol. 25, no. 2 (February 2008), pp. 177–183.

Welch, B. “The Generalization of `Student's' Problem When Several Different Population Variances Are Involved.” *Biometrika*, vol. 34, no. 1/2 (January 1947), pp. 28-35.

Zou, H. and T. Hastie. "Regularization and Variable Selection Via the Elastic Net." *Journal of the Royal Statistical Society*, Series B, vol. 67, no. 2 (April 2005), pp. 301-320.

Appendices

Appendix A

The ARP categorical benchmark structure includes three tiers consisting of 152 benchmarks. The 85 base benchmarks stratify the ARP universe by asset, style and region. The tier-two benchmarks roll up the base benchmarks into 46 asset-style benchmarks. The tier-three benchmarks roll up the asset-style benchmarks into 14 style benchmarks and 7 asset benchmarks. Columns provide benchmark codes indicating membership for each row.

Categorical Benchmark	Asset-Style-Region	Asset-Style	Style	Asset
Carry (curve) Equity (index-based) Europe	1	101	201	302
Carry (curve) Commodity Multi-Region	2	102	201	303
Carry (curve) Credit Multi-Region	3	103	201	304
Carry (curve) Rates North America	4	104	201	307
Carry (curve) Rates Europe	5	104	201	307
Carry (curve) Rates Multi-Region	6	104	201	307
Carry (spread) Commodity Multi-Region	7	105	202	303
Carry (spread) Credit Multi-Region	8	106	202	304
Carry (spread) Currency Multi-Region	9	107	202	305
Carry (spread) Rates Multi-Region	10	108	202	307
Congestion (rebalance, month-end) Equity (index-based) North America	11	109	203	302
Congestion (rebalance, month-end) Commodity Multi-Region	12	110	203	303
Merger Arbitrage Equity (stock-based) Multi-Region	13	111	204	301
Multi-Style Equity (stock-based) North America	14	112	205	301
Multi-Style Equity (stock-based) Europe	15	112	205	301
Multi-Style Equity (stock-based) Asia-Pacific	16	112	205	301
Multi-Style Equity (stock-based) Multi-Region	17	112	205	301
Multi-Style Commodity North America	18	113	205	303
Multi-Style Commodity Multi-Region	19	113	205	303
Multi-Style Currency Multi-Region	20	114	205	305
Multi-Style Multi-Asset Multi-Region	21	115	205	306
Multi-Style Rates Multi-Region	22	116	205	307
Other Equity (stock-based) North America	23	117	206	301
Other Multi-Asset Multi-Region	24	118	206	306
Reversal Equity (index-based) North America	25	119	207	302
Reversal Equity (index-based) Europe	26	119	207	302
Reversal Equity (index-based) Emerging Markets	27	119	207	302
Reversal Equity (index-based) Multi-Region	28	119	207	302
Reversal Commodity Multi-Region	29	120	207	303
Reversal Currency Multi-Region	30	121	207	305
Risk Anomaly (quality, low vol/beta) Equity (stock-based) North America	31	122	208	301
Risk Anomaly (quality, low vol/beta) Equity (stock-based) Europe	32	122	208	301
Risk Anomaly (quality, low vol/beta) Equity (stock-based) Asia-Pacific	33	122	208	301
Risk Anomaly (quality, low vol/beta) Equity (stock-based) Multi-Region	34	122	208	301

Appendix A continued

Categorical Benchmark	Asset-Style-Region	Asset-Style	Style	Asset
Size Equity (stock-based) North America	35	123	209	301
Size Equity (stock-based) Europe	36	123	209	301
Size Equity (stock-based) Asia-Pacific	37	123	209	301
Size Equity (stock-based) Multi-Region	38	123	209	301
Trend (cross-sectional momentum) Equity (stock-based) North America	39	124	210	301
Trend (cross-sectional momentum) Equity (stock-based) Europe	40	124	210	301
Trend (cross-sectional momentum) Equity (stock-based) Asia-Pacific	41	124	210	301
Trend (cross-sectional momentum) Equity (stock-based) Multi-Region	42	124	210	301
Trend (cross-sectional momentum) Commodity Multi-Region	43	125	210	303
Trend (cross-sectional momentum) Currency Multi-Region	44	126	210	305
Trend (cross-sectional momentum) Multi-Asset Multi-Region	45	127	210	306
Trend (cross-sectional momentum) Rates Multi-Region	46	128	210	307
Trend (time-series momentum) Equity (index-based) North America	47	129	211	302
Trend (time-series momentum) Equity (index-based) Asia-Pacific	48	129	211	302
Trend (time-series momentum) Equity (index-based) Multi-Region	49	129	211	302
Trend (time-series momentum) Commodity North America	50	130	211	303
Trend (time-series momentum) Commodity Multi-Region	51	130	211	303
Trend (time-series momentum) Credit North America	52	131	211	304
Trend (time-series momentum) Credit Europe	53	131	211	304
Trend (time-series momentum) Credit Multi-Region	54	131	211	304
Trend (time-series momentum) Currency Emerging Markets	55	132	211	305
Trend (time-series momentum) Currency Multi-Region	56	132	211	305
Trend (time-series momentum) Multi-Asset Multi-Region	57	133	211	306
Trend (time-series momentum) Rates Multi-Region	58	134	211	307
Value Equity (stock-based) North America	59	135	212	301
Value Equity (stock-based) Europe	60	135	212	301
Value Equity (stock-based) Asia-Pacific	61	135	212	301
Value Equity (stock-based) Multi-Region	62	135	212	301
Value Commodity Multi-Region	63	136	212	303
Value Currency Multi-Region	64	137	212	305
Value Rates Multi-Region	65	138	212	307

Appendix A continued

Categorical Benchmark	Asset-Style-Region	Asset-Style	Style	Asset
Volatility (arbitrage) Equity (stock-based) North America	66	139	213	301
Volatility (arbitrage) Equity (index-based) North America	67	140	213	302
Volatility (arbitrage) Equity (index-based) Multi-Region	68	140	213	302
Volatility (arbitrage) Commodity Multi-Region	69	141	213	303
Volatility (short) Equity (index-based) North America	70	142	214	302
Volatility (short) Equity (index-based) Europe	71	142	214	302
Volatility (short) Equity (index-based) Asia-Pacific	72	142	214	302
Volatility (short) Equity (index-based) Emerging Markets	73	142	214	302
Volatility (short) Equity (index-based) Multi-Region	74	142	214	302
Volatility (short) Commodity North America	75	143	214	303
Volatility (short) Commodity Europe	76	143	214	303
Volatility (short) Commodity Multi-Region	77	143	214	303
Volatility (short) Credit North America	78	144	214	304
Volatility (short) Currency Europe	79	145	214	305
Volatility (short) Currency Asia-Pacific	80	145	214	305
Volatility (short) Currency Multi-Region	81	145	214	305
Volatility (short) Rates North America	82	146	214	307
Volatility (short) Rates Europe	83	146	214	307
Volatility (short) Rates Asia-Pacific	84	146	214	307
Volatility (short) Rates Multi-Region	85	146	214	307
Carry (curve) Equity (index-based)		101	201	302
Carry (curve) Commodity		102	201	303
Carry (curve) Credit		103	201	304
Carry (curve) Rates		104	201	307
Carry (spread) Commodity		105	202	303
Carry (spread) Credit		106	202	304
Carry (spread) Currency		107	202	305
Carry (spread) Rates		108	202	307
Congestion (rebalance, month-end) Equity (index-based)		109	203	302
Congestion (rebalance, month-end) Commodity		110	203	303
Merger Arbitrage Equity (stock-based)		111	204	301
Multi-Style Equity (stock-based)		112	205	301
Multi-Style Commodity		113	205	303

Appendix A continued

Categorical Benchmark	Asset-Style-Region	Asset-Style	Style	Asset
Multi-Style Currency		114	205	305
Multi-Style Multi-Asset		115	205	306
Multi-Style Rates		116	205	307
Other Equity (stock-based)		117	206	301
Other Multi-Asset		118	206	306
Reversal Equity (index-based)		119	207	302
Reversal Commodity		120	207	303
Reversal Currency		121	207	305
Risk Anomaly (quality, low volatility/beta) Equity (stock-based)		122	208	301
Size Equity (stock-based)		123	209	301
Trend (cross-sectional momentum) Equity (stock-based)		124	210	301
Trend (cross-sectional momentum) Commodity		125	210	303
Trend (cross-sectional momentum) Currency		126	210	305
Trend (cross-sectional momentum) Multi-Asset		127	210	306
Trend (cross-sectional momentum) Rates		128	210	307
Trend (time-series momentum) Equity (index-based)		129	211	302
Trend (time-series momentum) Commodity		130	211	303
Trend (time-series momentum) Credit		131	211	304
Trend (time-series momentum) Currency		132	211	305
Trend (time-series momentum) Multi-Asset		133	211	306
Trend (time-series momentum) Rates		134	211	307
Value Equity (stock-based)		135	212	301
Value Commodity		136	212	303
Value Currency		137	212	305
Value Rates		138	212	307
Volatility (arbitrage) Equity (stock-based)		139	213	301
Volatility (arbitrage) Equity (index-based)		140	213	302
Volatility (arbitrage) Commodity		141	213	303
Volatility (short) Equity (index-based)		142	214	302
Volatility (short) Commodity		143	214	303
Volatility (short) Credit		144	214	304
Volatility (short) Currency		145	214	305
Volatility (short) Rates		146	214	307

Appendix A continued

Categorical Benchmark	Asset-Style-Region	Asset-Style	Style	Asset
Carry (curve)			201	
Carry (spread)			202	
Congestion (rebalance, month-end)			203	
Merger Arbitrage			204	
Multi-Style			205	
Other			206	
Reversal			207	
Risk Anomaly (quality, low volatility/beta)			208	
Size			209	
Trend (cross-sectional momentum)			210	
Trend (time-series momentum)			211	
Value			212	
Volatility (arbitrage)			213	
Volatility (short)			214	
Equity (stock-based)				301
Equity (index-based)				302
Commodity				303
Credit				304
Currency				305
Multi-Asset				306
Rates				307

Appendix B

The supplemental factor set taps four data sources for additional return context – the Bloomberg GSAM Risk Premia Indices, Fama-French Factor Library, Barra Global Equity Model (GEM3), and various market indices.

#	Abbrev	Reference Factor Group	Name
1	DIVRP6	Bloomberg GSAM Risk Premia Indices	Cross Asset Risk Premia 6% Volatility Target
2	DIVRP	Bloomberg GSAM Risk Premia Indices	Cross Asset Risk Premia Index
3	MACRP	Bloomberg GSAM Risk Premia Indices	Macro Risk Premia Index
4	CAVALRP	Bloomberg GSAM Risk Premia Indices	Cross Asset Carry and Value Index
5	UEEQMO	Bloomberg GSAM Risk Premia Indices	US Equity Momentum Long-Short Index
6	USEQVAL	Bloomberg GSAM Risk Premia Indices	US Equity Value Long-Short Index
7	USEQLR	Bloomberg GSAM Risk Premia Indices	US Equity Low Risk Long-Short Index
8	USEQQU	Bloomberg GSAM Risk Premia Indices	US Equity Quality Long-Short Index
9	USEQDIV	Bloomberg GSAM Risk Premia Indices	US Equity Multi Factor Long-Short Index
10	FXCARR	Bloomberg GSAM Risk Premia Indices	FX Carry Index
11	BDCARR	Bloomberg GSAM Risk Premia Indices	Bond Futures Carry Index
12	COMCARR	Bloomberg GSAM Risk Premia Indices	Commodity Carry Index
13	DIVCARR	Bloomberg GSAM Risk Premia Indices	Cross Asset Carry Index
14	FXVAL	Bloomberg GSAM Risk Premia Indices	FX G10 Value Index
15	BDVAL	Bloomberg GSAM Risk Premia Indices	Bond Futures Value Index
16	DIVVAL	Bloomberg GSAM Risk Premia Indices	Cross Asset Value Index
17	FXTRD	Bloomberg GSAM Risk Premia Indices	FX Trend Index
18	BDTRD	Bloomberg GSAM Risk Premia Indices	Bond Futures Trend Index
19	EQTRD	Bloomberg GSAM Risk Premia Indices	Equity Trend Index
20	COMTRD	Bloomberg GSAM Risk Premia Indices	Commodity Trend Index
21	DIVTRD	Bloomberg GSAM Risk Premia Indices	Cross Asset Trend Index

Appendix B continued

#	Abbrev	Reference Factor Group	Name
22	USSMB	Fama-French Factor Library	US Small minus Big Capitalization factor
23	USHML	Fama-French Factor Library	US High minus Low Book-to-Market factor
24	USRMW	Fama-French Factor Library	US Robust minus Weak Operating Profitability factor
25	USCMA	Fama-French Factor Library	US Conservative minus Aggressive Investment factor
26	USWML	Fama-French Factor Library	US Winners minus Losers Price Momentum factor (2-12m return)
27	USSTREV	Fama-French Factor Library	US Short-term Reversal Factor (1m return)
28	USLTREV	Fama-French Factor Library	US Long-term Reversal factor (13-60m return)
29	EURSMB	Fama-French Factor Library	Europe Small minus Big Capitalization factor
30	EURHML	Fama-French Factor Library	Europe High minus Low Book-to-Market factor
31	EURRMW	Fama-French Factor Library	Europe Robust minus Weak Operating Profitability factor
32	EURCMA	Fama-French Factor Library	Europe Conservative minus Aggressive Investment factor
33	EURWML	Fama-French Factor Library	Europe Winners minus Losers Price Momentum factor (2-12m return)
34	JPNSMB	Fama-French Factor Library	Japan Small minus Big Capitalization factor
35	JPNHML	Fama-French Factor Library	Japan High minus Low Book-to-Market factor
36	JPNRMW	Fama-French Factor Library	Japan Robust minus Weak Operating Profitability factor
37	JPNCMA	Fama-French Factor Library	Japan Conservative minus Aggressive Investment factor
38	JPNWML	Fama-French Factor Library	Japan Winners minus Losers Price Momentum factor (2-12m return)

Appendix B continued

#	Abbrev	Reference Factor Group	Name
39	BETA	Barra Global Equity Model (GEM3)	Beta style factor
40	MOM	Barra Global Equity Model (GEM3)	Momentum factor
41	SIZE	Barra Global Equity Model (GEM3)	Size style factor
42	EARNYLD	Barra Global Equity Model (GEM3)	Earnings Yield style factor
43	RESVOL	Barra Global Equity Model (GEM3)	Residual Volatility style factor
44	GROWTH	Barra Global Equity Model (GEM3)	Growth style factor
45	DIVYLD	Barra Global Equity Model (GEM3)	Dividend Yield style factor
46	BTOP	Barra Global Equity Model (GEM3)	Book-to-Price style factor
47	LEV	Barra Global Equity Model (GEM3)	Leverage style factor
48	LIQ	Barra Global Equity Model (GEM3)	Liquidity style factor
49	SIZENL	Barra Global Equity Model (GEM3)	Non-linear Size style factor
50	ENERGY	Barra Global Equity Model (GEM3)	Energy Equipment and Services industry factor
51	OILGAS	Barra Global Equity Model (GEM3)	Oil Gas and Consumable Fuels industry factor
52	OILEXPL	Barra Global Equity Model (GEM3)	Oil and Gas Exploration and Production industry factor
53	CHEM	Barra Global Equity Model (GEM3)	Chemicals industry factor
54	CONSTR	Barra Global Equity Model (GEM3)	Construction Containers Paper industry factor
55	DIVMET	Barra Global Equity Model (GEM3)	Aluminum Diversified Metals industry factor
56	PRECMET	Barra Global Equity Model (GEM3)	Gold and Precious Metals industry factor
57	STEEL	Barra Global Equity Model (GEM3)	Steel industry factor
58	CAPGD	Barra Global Equity Model (GEM3)	Capital Goods industry factor
59	PROFSVC	Barra Global Equity Model (GEM3)	Commercial and Professional Services industry factor
60	TRANSP	Barra Global Equity Model (GEM3)	Transportation Non-Airline industry factor
61	AIRLINE	Barra Global Equity Model (GEM3)	Airlines industry factor

Appendix B continued

#	Abbrev	Reference Factor Group	Name
62	AUTO	Barra Global Equity Model (GEM3)	Automobiles and Components industry factor
63	CONSDUR	Barra Global Equity Model (GEM3)	Consumer Durables and Apparel industry factor
64	CONSVCS	Barra Global Equity Model (GEM3)	Hotels Restaurants and Leisure industry factor
65	MEDIA	Barra Global Equity Model (GEM3)	Media industry factor
66	RETAIL	Barra Global Equity Model (GEM3)	Retailing industry factor
67	FOODRTL	Barra Global Equity Model (GEM3)	Food and Staples Retailing industry factor
68	FOODPRD	Barra Global Equity Model (GEM3)	Food Beverage and Tobacco industry factor
69	HSHLD	Barra Global Equity Model (GEM3)	Household and Personal Products industry factor
70	HEALTH	Barra Global Equity Model (GEM3)	Health Care Equipment and Services industry factor
71	BIOTECH	Barra Global Equity Model (GEM3)	Biotechnology industry factor
72	PHARMA	Barra Global Equity Model (GEM3)	Pharmaceuticals and Life Sciences industry factor
73	BANKS	Barra Global Equity Model (GEM3)	Banks industry factor
74	DIVFINL	Barra Global Equity Model (GEM3)	Diversified Financials industry factor
75	INSUR	Barra Global Equity Model (GEM3)	Insurance industry factor
76	REALEST	Barra Global Equity Model (GEM3)	Real Estate industry factor
77	INTERNT	Barra Global Equity Model (GEM3)	Internet Software and Services industry factor
78	SOFTWAR	Barra Global Equity Model (GEM3)	IT Services and Software industry factor
79	COMM	Barra Global Equity Model (GEM3)	Communications Equipment industry factor
80	COMPUT	Barra Global Equity Model (GEM3)	Computers Electronics industry factor
81	SEMICON	Barra Global Equity Model (GEM3)	Semiconductors industry factor
82	TELECOM	Barra Global Equity Model (GEM3)	Telecommunication Services industry factor
83	UTILITY	Barra Global Equity Model (GEM3)	Utilities industry factor

Appendix B continued

#	Abbrev	Reference Factor Group	Name
84	NAEQ	Market Indices	MSCI North America Gross Total Return Local Index
85	EUREQ	Market Indices	MSCI Europe Gross Total Return Local Index
86	PACEQ	Market Indices	MSCI Pacific Gross Total Return Local Index
87	EMEQ	Market Indices	MSCI EM Gross Total Return USD Index
88	GOVEXUS	Market Indices	FTSE Non-USD World Govt Bond Ten-Markets Total Return Index FX-Hedged USD
89	GOVUS	Market Indices	FTSE USBIG Treasury Total Return Index
90	ILBUS	Market Indices	Bloomberg Barclays US Govt Inflation-Linked All Maturities Total Return Index
91	ILBEXUS	Market Indices	Bloomberg Barclays World Govt ex-US Inflation-Linked Bonds All Maturities Total Return FX-Hedged USD
92	ENCOMM	Market Indices	Bloomberg Energy Commodity Subindex Total Return
93	IMCOMM	Market Indices	Bloomberg Industrial Metals Commodity Subindex Total Return
94	PMCOMM	Market Indices	Bloomberg Precious Metals Commodity Subindex Total Return
95	AGCOMM	Market Indices	Bloomberg Agriculture Commodity Subindex Total Return
96	USDEUR	Market Indices	USDEUR Spot Exchange Rate - Price of 1 USD in EUR (% change)
97	USDAUD	Market Indices	USDAUD Spot Exchange Rate - Price of 1 USD in AUD (% change)
98	USDJPY	Market Indices	USDJPY Spot Exchange Rate - Price of 1 USD in JPY (% change)
99	EMBDS	Market Indices	J.P. Morgan Emerging Markets Bond Total Return Index
100	USHYBD	Market Indices	ICE BofA US High Yield Total Return Index
101	USCORBD	Market Indices	ICE BofA US Corporate Total Return Index
102	VIX	Market Indices	Chicago Board Options Exchange Volatility Index (% change)
103	USBREV	Market Indices	Bloomberg Generic 10-year US Breakeven (% change)
104	USVAR	Market Indices	Long Monthly S&P 500 Variance Swap (synthetic) Return Index

Appendix C

The fully-nested ARP statistical benchmark structure includes four tiers consisting of 155 benchmarks. The 85 base benchmarks are the result of agglomerative hierarchical clustering and PCA-based pruning. The 40 super-base, 20 hypo-broad and 10 broad benchmarks reflect the cluster tree structure. The benchmark names are generalizations of the constituents. A statistical approach obviously may combine strategies with different categorical profiles. Columns provide benchmark codes indicating membership for each row.

Statistical Benchmark	Base	Super-Base	Hypo-Broad	Broad
FX Trend (T/S) EM Focus	1	136	208	306
FX Trend (T/S) Developed Mkt Focus	2	136	208	306
FX Reversal Developed Mkt Focus	3	118	202	310
Stocks Multi-Style N. America Approach 1	4	122	206	307
Equity Volatility (short) N. America VIX Focus	5	116	204	302
Equity Volatility (short)	6	116	204	302
Rates Value	7	118	202	310
Commodity Carry (curve) Approach 1	8	121	219	308
Stocks Risk Anomaly N. America Approach 1	9	125	207	307
Stocks Multi-Style Approach 1	10	125	207	307
Commodity Congestion Approach 1	11	118	202	310
Equity Congestion	12	118	202	310
Rates Volatility (short) Asia-Pac	13	116	204	302
Stocks Merger Arbitrage	14	116	204	302
FX Volatility (short) Europe	15	138	210	302
FX Volatility (short)	16	138	210	302
Rates Volatility (short) N. America Approach 1	17	139	212	302
Rates Volatility (short) N. America Approach 2	18	139	212	302
Equity Reversal Approach 1	19	106	201	310
Equity Reversal Approach 2	20	106	201	310
Credit Carry (curve)	21	116	204	302
Commodity Trend (C/S)	22	104	220	309
Commodity Trend (T/S) Approach 1	23	104	220	309
Stocks Multi-Style Approach 2	24	116	204	302
Stocks Trend (C/S) N. America	25	123	206	307
Stocks Trend (C/S) Europe	26	123	206	307
Stocks Multi-Style Asia-Pac Approach 1	27	119	211	307
Stocks Multi-Style Asia-Pac Approach 2	28	119	211	307
Equity Volatility (arbitrage) N. America Approach 1	29	118	202	310
Rates Carry (curve) Long Rate Focus	30	127	218	305
Rates Carry (curve) Short Rate Focus Approach 1	31	127	218	305
Rates Carry (spread) Approach 1	32	128	218	305
Rates Multi-Style	33	128	218	305

Appendix C continued

Statistical Benchmark	Base	Super-Base	Hypo-Broad	Broad
Equity Trend (T/S)	34	129	214	304
Credit Trend (T/S)	35	129	214	304
Equity Carry (curve) Europe Dividend Focus	36	116	204	302
Stocks Multi-Style N. America Approach 2	37	125	207	307
Commodity Volatility (short) Soybean	38	134	203	302
Commodity Volatility (short) Wheat & Corn	39	134	203	302
Rates Carry (spread) Approach 2	40	128	218	305
Commodity Trend (T/S) Natural Gas Intraday	41	118	202	310
FX Value Approach 1	42	112	202	310
Stocks Risk Anomaly N. America Approach 2	43	108	211	307
Stocks Risk Anomaly Europe	44	108	211	307
Commodity Carry (spread) Approach 2	45	132	220	309
Commodity Carry (spread) Approach 3	46	132	220	309
Multi-Asset Multi-Style	47	116	204	302
Rates Congestion	48	118	202	310
Commodity Congestion Approach 2	49	121	219	308
Commodity Congestion Approach 3	50	121	219	308
FX Value Approach 2	51	112	202	310
Commodity Multi-Style Crude Oil	52	118	202	310
Rates Volatility (short)	53	139	212	302
Stocks Value Europe	54	101	215	310
Stocks Value N. America Approach 1	55	101	215	310
Stocks Value N. America Approach 2	56	102	215	310
Stocks Risk Anomaly N. America Approach 3	57	102	215	310
Rates Carry (curve) Short Rate Focus Approach 2	58	127	218	305
Credit Carry (spread)	59	130	214	304
Equity Multi-Style N. America	60	130	214	304
Commodity Volatility (short) Sugar	61	118	202	310
Stocks Value Asia-Pac	62	101	215	310
FX Volatility (short) Asia-Pac	63	138	210	302
Stocks Risk Anomaly Approach 1	64	118	202	310
Equity Volatility (short) Asia-Pac	65	116	204	302

Appendix C continued

Statistical Benchmark	Base	Super-Base	Hypo-Broad	Broad
Multi-Asset Trend (T/S) Approach 1	66	137	209	306
Multi-Asset Trend (T/S) Approach 2	67	137	209	306
Equity Volatility (arbitrage) N. America Approach 2	68	130	214	304
Commodity Trend (T/S) Approach 2	69	103	220	309
Equity Trend (T/S) N. America Dynamic	70	105	201	310
Stocks Risk Anomaly Approach 2	71	107	211	307
Equity Volatility (short) Europe	72	109	213	304
Equity Volatility (short) N. America	73	110	213	304
Commodity Reversal	74	111	202	310
FX Carry (spread) Developed Mkt Focus	75	113	217	303
FX Carry (spread) EM Focus	76	114	217	303
Commodity Volatility (short) Industrial Metals	77	115	204	302
Equity Volatility (arbitrage) N. America Approach 3	78	117	202	310
Commodity Carry (curve) Approach 2	79	120	219	308
Stocks Multi-Style Europe	80	124	207	307
Rates Trend (T/S)	81	126	218	305
Commodity Carry (spread) Approach 1	82	131	220	309
Commodity Volatility (short) Gold	83	133	204	302
Commodity Volatility (short) Natural Gas	84	135	205	302
Crude Oil Volatility (short)	85	140	216	301
Stocks Value 1		101	215	310
Stocks Value 2		102	215	310
Commodity Trend (T/S)		103	220	309
Commodity Trend		104	220	309
Equity Trend (T/S) N. America Dynamic		105	201	310
Equity Reversal		106	201	310
Stocks Risk Anomaly 1		107	211	307
Stocks Risk Anomaly 2		108	211	307
Equity Volatility (short) Europe		109	213	304
Equity Volatility (short) N. America		110	213	304
Commodity Reversal		111	202	310

Appendix C continued

Statistical Benchmark	Base	Super-Base	Hypo-Broad	Broad
FX Value		**112**	202	310
FX Carry (spread) Developed Mkt Focus		**113**	217	303
FX Carry (spread) EM Focus		**114**	217	303
Commodity Volatility (short) Industrial Metals		**115**	204	302
Multi-Asset Volatility Sensitive		**116**	204	302
Equity Volatility (arbitrage) N. America		**117**	202	310
Multi-Asset Reversal Oriented		**118**	202	310
Stocks Multi-Style Asia-Pac		**119**	211	307
Commodity Carry (curve) 1		**120**	219	308
Commodity Carry (curve) 2		**121**	219	308
Stocks Multi-Style N. America		**122**	206	307
Stocks Trend (C/S)		**123**	206	307
Stocks Multi-Style Europe		**124**	207	307
Stocks Multi-Style		**125**	207	307
Rates Trend (T/S)		**126**	218	305
Rates Carry (curve)		**127**	218	305
Rates Carry (spread)		**128**	218	305
Equity/Credit Trend (T/S)		**129**	214	304
Equity Multi-Style		**130**	214	304
Commodity Carry (spread) 1		**131**	220	309
Commodity Carry (spread) 2		**132**	220	309
Commodity Volatility (short) Gold		**133**	204	302
Commodity Volatility (short) Grains		**134**	203	302
Commodity Volatility (short) Natural Gas		**135**	205	302
FX Trend (T/S)		**136**	208	306
Multi-Asset Trend (T/S)		**137**	209	306
FX Volatility (short) Plus		**138**	210	302
Rates Volatility (short) Plus		**139**	212	302
Crude Oil Volatility (short)		**140**	216	301

Appendix C continued

Statistical Benchmark	Base	Super-Base	Hypo-Broad	Broad
Equity Reversal Plus			**201**	310
Multi-Asset Reversal Oriented Plus			**202**	310
Commodity Volatility (short) Grains			**203**	302
Multi-Asset Volatility Sensitive Plus			**204**	302
Commodity Volatility (short) Natural Gas			**205**	302
Stocks Trend (C/S) Plus			**206**	307
Stocks Multi-Style Plus			**207**	307
FX Trend (T/S)			**208**	306
Multi-Asset Trend (T/S)			**209**	306
FX Volatility (short) Plus			**210**	302
Stocks Risk Anomaly			**211**	307
Rates Volatility (short) Plus			**212**	302
Equity Volatility (short)			**213**	304
Equity Trend			**214**	304
Stocks Value			**215**	310
Crude Oil Volatility (short)			**216**	301
FX Carry			**217**	303
Rates Carry			**218**	305
Commodity Curve Carry			**219**	308
Commodity Spread Carry & Trend			**220**	309
Crude Oil Volatility (short)				**301**
Volatility Sensitive				**302**
FX Carry				**303**
Equity Sensitive				**304**
Rates Carry				**305**
FX/Multi-Asset Trend				**306**
Stocks, Value Light				**307**
Commodity Curve Carry				**308**
Commodity Spread Carry & Trend				**309**
Value Oriented				**310**

Appendix D

The reference benchmarks consist primarily of standard market indices. These weekly indices all are gross of implementation and access costs. (The MSCI net indices adjust only for dividend withholding taxes.)

#	Abbrev	Source	Name
1	NAEQ	Bloomberg	MSCI North America Net Total Return USD Index
2	EUREQ	Bloomberg	MSCI Europe Net Total Return USD Index
3	PACEQ	Bloomberg	MSCI Pacific Net Total Return USD Index
4	EMEQ	Bloomberg	MSCI EM Gross Total Return USD Index
5	GOVEXUS	Bloomberg	FTSE Non-USD World Govt Bond Ten-Markets Total Return Index FX-Hedged USD
6	GOVUS	Bloomberg	FTSE USBIG Treasury Total Return Index
7	ILBUS	Bloomberg	Bloomberg Barclays US Govt Inflation-Linked All Maturities Total Return Index
8	ILBEXUS	Bloomberg	Bloomberg Barclays World Govt ex-US Inflation-Linked Bonds All Maturities Total Return FX-Hedged USD
9	ENCOMM	Bloomberg	Bloomberg Energy Commodity Subindex Total Return
10	IMCOMM	Bloomberg	Bloomberg Industrial Metals Commodity Subindex Total Return
11	PMCOMM	Bloomberg	Bloomberg Precious Metals Commodity Subindex Total Return
12	AGCOMM	Bloomberg	Bloomberg Agriculture Commodity Subindex Total Return
13	USDEUR	Bloomberg	Bloomberg EURUSD Currency Carry Return Index [1]
14	USDAUD	Bloomberg	Bloomberg AUDUSD Currency Carry Return Index [1]
15	USDJPY	Bloomberg	Bloomberg JPYUSD Currency Carry Return Index [1]
16	EMBDS	Bloomberg	J.P. Morgan Emerging Markets Bond Total Return Index
17	USHYBD	Bloomberg	ICE BofA US High Yield Bond Total Return Index
18	USCORBD	Bloomberg	ICE BofA US Corporate Bond Total Return Index
19	NPTECH	Bloomberg	Goldman Sachs Non-Profitable Tech Company Index [2]
20	VALFSPD	Internal	World Value & Financing Appetite Stock Return Spread Index [3]
21	USVIX	Bloomberg	Credit Suisse Long 1-month VIX 1% Vega Index
22	USVAR	Internal	Long Monthly S&P 500 Variance Swap Return Index [4]

Appendix D continued

[1] The three currency carry indices provide the return from borrowing in USD to fund buying EUR, AUD or JPY. The return adds the spot FX change to the interest rate differential.

[2] The Goldman Sachs Non-Profitable Tech Company Index includes non-profitable, US-listed companies in innovative industries. Tech is defined broadly to include new-economy companies across GICS industry groupings. The index excludes hard-to-borrow names, names with trading restrictions, and names with pending mergers. The index is equally weighted, with an ADV cap of 10% on a notional of $100m and no name initially weighted more than 5%. The index has a September 2014 inception date, so it does not appear in full-period analyses in this paper.

[3] The World Value & Financing Appetite Stock Return Spread Index reflects two independent sorts of the MSCI World Index universe on cashflow to enterprise value and net external financing to enterprise value. The index represents the return spread between the intersection of the top and bottom quintile (by name count) of the two sorts – the return to undervalued companies returning capital minus the return to expensive companies consuming capital. The index rebalances monthly and equally weights names within quintiles.

[4] The Long Monthly S&P 500 Variance Swap Return Index approximates the return to a monthly variance swap using the following formula:

$$r_t = e^{\frac{(ln[P_t]-ln[P_{t-1}])^2 * 252 - V_{m-1}^2}{N_m}} - 1 \qquad \text{Equation 33}$$

with P representing the S&P 500 index value on day t, V representing the previous month-end VIX level and N representing the number of trading days in month m.

Given the exposure to the difference between implied and realized volatility in ARP portfolios, the Credit Suisse Long 1-month VIX 1% Vega Index provides a complementary perspective by rolling 1-month VIX futures. VIX futures began trading in 2004 and returns for the index commence in 2006, so the index does not appear in the full-period analyses in this paper.

Appendix E

These tables summarize the skewness, kurtosis and autocorrelation for the 40 super-base, 20 hypo-broad and 10 broad ARP composite strategy benchmarks. Negative (positive) values appear in red (black) font. The calculations use weekly data for the 2000-2017 and 2018-2020 periods, with the former window underpinning expectations for the latter.

2018-2020 Skewness

Broad	Skew	Hypo-Broad	Skew	Super-Base	Skew
Stocks, Value Light	-1.8	Stocks Trend (C/S) Plus	-2.3	Stocks Multi-Style N. America	-1.9
				Stocks Trend (C/S)	-1.8
		Stocks Multi-Style Plus	0.1	Stocks Multi-Style	-0.4
				Stocks Multi-Style Europe	0.8
		Stocks Risk Anomaly	-1.1	Stocks Multi-Style Asia-Pac	0.0
				Stocks Risk Anomaly 1	-1.8
				Stocks Risk Anomaly 2	-1.8
Commodity Spread Carry & Trend	-0.3	Commodity Spread Carry & Trend	-0.3	Commodity Trend	-0.2
				Commodity Carry (spread) 2	-1.5
				Commodity Carry (spread) 1	-0.2
				Commodity Trend (T/S)	0.2
Volatility Sensitive	-3.0	Commodity Volatility (short) Grains	-0.4	Commodity Volatility (short) Grains	-0.4
		Rates Volatility (short) Plus	-1.8	Rates Volatility (short) Plus	-1.8
		Commodity Volatility (short) Natural Gas	-1.6	Commodity Volatility (short) Natural Gas	-1.6
		Multi-Asset Volatility Sensitive Plus	-3.5	Commodity Volatility (short) Gold	-2.8
				Multi-Asset Volatility Sensitive	-2.5
				Commodity Volatility (short) Industrial Metals	-2.0
		FX Volatility (short) Plus	-2.3	FX Volatility (short) Plus	-2.3
Value Oriented	0.6	Stocks Value	0.7	Stocks Value 1	0.5
				Stocks Value 2	0.8
		Equity Reversal Plus	2.5	Equity Reversal	-2.7
				Equity Trend (T/S) N. America Dynamic	6.4
		Multi-Asset Reversal Oriented Plus	0.2	Equity Volatility (arbitrage) N. America	0.0
				Commodity Reversal	-0.3
				Multi-Asset Reversal Oriented	1.9
				FX Value	-0.5
Equity Sensitive	-2.7	Equity Trend	-2.5	Equity/Credit Trend (T/S)	-2.8
				Equity Multi-Style	-1.6
		Equity Volatility (short)	-2.2	Equity Volatility (short) N. America	-3.1
				Equity Volatility (short) Europe	-1.2
Crude Oil Volatility (short)	-2.3	Crude Oil Volatility (short)	-2.3	Crude Oil Volatility (short)	-2.3
FX/Multi-Asset Trend	-0.9	FX Trend (T/S)	0.5	FX Trend (T/S)	0.5
		Multi-Asset Trend (T/S)	-1.9	Multi-Asset Trend (T/S)	-1.9
FX Carry	-1.0	FX Carry	-1.0	FX Carry (spread) Developed Mkt Focus	-0.7
				FX Carry (spread) EM Focus	-1.2
Commodity Curve Carry	1.3	Commodity Curve Carry	1.3	Commodity Carry (curve) 2	2.0
				Commodity Carry (curve) 1	0.7
Rates Carry	-0.4	Rates Carry	-0.4	Rates Carry (spread)	-0.7
				Rates Trend (T/S)	0.0
				Rates Carry (curve)	1.4

Appendix E continued

2000-2017 Skewness

Broad	Skew	Hypo-Broad	Skew	Super-Base	Skew
Stocks, Value Light	-0.6	Stocks Trend (C/S) Plus	-1.1	Stocks Multi-Style N. America	-0.3
				Stocks Trend (C/S)	-1.0
		Stocks Multi-Style Plus	-0.2	Stocks Multi-Style	-0.3
				Stocks Multi-Style Europe	-0.1
		Stocks Risk Anomaly	-0.2	Stocks Multi-Style Asia-Pac	-0.7
				Stocks Risk Anomaly 1	-0.1
				Stocks Risk Anomaly 2	-0.5
Commodity Spread Carry & Trend	0.0	Commodity Spread Carry & Trend	0.0	Commodity Trend	-0.3
				Commodity Carry (spread) 2	0.0
				Commodity Carry (spread) 1	-0.2
				Commodity Trend (T/S)	0.5
Volatility Sensitive	-0.8	Commodity Volatility (short) Grains	-0.6	Commodity Volatility (short) Grains	-0.6
		Rates Volatility (short) Plus	-1.3	Rates Volatility (short) Plus	-1.3
		Commodity Volatility (short) Natural Gas	-0.9	Commodity Volatility (short) Natural Gas	-0.9
		Multi-Asset Volatility Sensitive Plus	-0.6	Commodity Volatility (short) Gold	-2.6
				Multi-Asset Volatility Sensitive	-0.7
				Commodity Volatility (short) Industrial Metals	-1.4
		FX Volatility (short) Plus	-1.7	FX Volatility (short) Plus	-1.7
Value Oriented	1.6	Stocks Value	1.2	Stocks Value 1	0.9
				Stocks Value 2	0.4
		Equity Reversal Plus	1.6	Equity Reversal	-0.2
				Equity Trend (T/S) N. America Dynamic	2.6
		Multi-Asset Reversal Oriented Plus	0.7	Equity Volatility (arbitrage) N. America	0.5
				Commodity Reversal	0.0
				Multi-Asset Reversal Oriented	0.7
				FX Value	0.1
Equity Sensitive	-1.5	Equity Trend	-0.6	Equity/Credit Trend (T/S)	0.2
				Equity Multi-Style	-0.7
		Equity Volatility (short)	-2.0	Equity Volatility (short) N. America	-2.4
				Equity Volatility (short) Europe	-1.7
Crude Oil Volatility (short)	-1.5	Crude Oil Volatility (short)	-1.5	Crude Oil Volatility (short)	-1.5
FX/Multi-Asset Trend	-0.3	FX Trend (T/S)	0.1	FX Trend (T/S)	0.1
		Multi-Asset Trend (T/S)	-0.6	Multi-Asset Trend (T/S)	-0.6
FX Carry	-0.9	FX Carry	-0.9	FX Carry (spread) Developed Mkt Focus	-0.9
				FX Carry (spread) EM Focus	-0.7
Commodity Curve Carry	0.1	Commodity Curve Carry	0.1	Commodity Carry (curve) 2	0.6
				Commodity Carry (curve) 1	-0.2
Rates Carry	-0.7	Rates Carry	-0.7	Rates Carry (spread)	-0.3
				Rates Trend (T/S)	-0.4
				Rates Carry (curve)	-0.6

Appendix E continued

2018-2020 Kurtosis (excess)

Broad	Kurtosis	Hypo-Broad	Kurtosis	Super-Base	Kurtosis
Stocks, Value Light	**6.5**	Stocks Trend (C/S) Plus	**9.9**	Stocks Multi-Style N. America	**7.2**
				Stocks Trend (C/S)	**7.5**
		Stocks Multi-Style Plus	**4.3**	Stocks Multi-Style	**2.6**
				Stocks Multi-Style Europe	**3.2**
		Stocks Risk Anomaly	**3.3**	Stocks Multi-Style Asia-Pac	**1.7**
				Stocks Risk Anomaly 1	**9.4**
				Stocks Risk Anomaly 2	**7.5**
Commodity Spread Carry & Trend	**2.4**	Commodity Spread Carry & Trend	**2.4**	Commodity Trend	**0.4**
				Commodity Carry (spread) 2	**7.6**
				Commodity Carry (spread) 1	**0.6**
				Commodity Trend (T/S)	**7.1**
Volatility Sensitive	**17.8**	Commodity Volatility (short) Grains	**1.1**	Commodity Volatility (short) Grains	**1.1**
		Rates Volatility (short) Plus	**9.7**	Rates Volatility (short) Plus	**9.7**
		Commodity Volatility (short) Natural Gas	**5.2**	Commodity Volatility (short) Natural Gas	**5.2**
		Multi-Asset Volatility Sensitive Plus	**19.5**	Commodity Volatility (short) Gold	**15.1**
				Multi-Asset Volatility Sensitive	**13.6**
				Commodity Volatility (short) Industrial Metals	**8.4**
		FX Volatility (short) Plus	**10.7**	FX Volatility (short) Plus	**10.7**
Value Oriented	**1.9**	Stocks Value	**6.0**	Stocks Value 1	**3.0**
				Stocks Value 2	**7.0**
		Equity Reversal Plus	**20.0**	Equity Reversal	**27.1**
				Equity Trend (T/S) N. America Dynamic	**56.2**
		Multi-Asset Reversal Oriented Plus	**0.5**	Equity Volatility (arbitrage) N. America	**6.9**
				Commodity Reversal	**0.6**
				Multi-Asset Reversal Oriented	**6.1**
				FX Value	**5.1**
Equity Sensitive	**12.3**	Equity Trend	**11.8**	Equity/Credit Trend (T/S)	**14.4**
				Equity Multi-Style	**5.5**
		Equity Volatility (short)	**8.8**	Equity Volatility (short) N. America	**13.8**
				Equity Volatility (short) Europe	**4.2**
Crude Oil Volatility (short)	**11.2**	Crude Oil Volatility (short)	**11.2**	Crude Oil Volatility (short)	**11.2**
FX/Multi-Asset Trend	**3.2**	FX Trend (T/S)	**3.7**	FX Trend (T/S)	**3.7**
		Multi-Asset Trend (T/S)	**7.0**	Multi-Asset Trend (T/S)	**7.0**
FX Carry	**3.9**	FX Carry	**3.9**	FX Carry (spread) Developed Mkt Focus	**2.3**
				FX Carry (spread) EM Focus	**5.3**
Commodity Curve Carry	**11.1**	Commodity Curve Carry	**11.1**	Commodity Carry (curve) 2	**17.0**
				Commodity Carry (curve) 1	**4.9**
Rates Carry	**2.6**	Rates Carry	**2.6**	Rates Carry (spread)	**5.2**
				Rates Trend (T/S)	**3.9**
				Rates Carry (curve)	**10.7**

Appendix E continued

2000-2017 Kurtosis (excess)

Broad	Kurtosis	Hypo-Broad	Kurtosis	Super-Base	Kurtosis
Stocks, Value Light	2.9	Stocks Trend (C/S) Plus	4.1	Stocks Multi-Style N. America	1.7
				Stocks Trend (C/S)	3.6
		Stocks Multi-Style Plus	1.8	Stocks Multi-Style	1.9
				Stocks Multi-Style Europe	2.1
		Stocks Risk Anomaly	2.5	Stocks Multi-Style Asia-Pac	3.8
				Stocks Risk Anomaly 1	2.9
				Stocks Risk Anomaly 2	4.0
Commodity Spread Carry & Trend	2.0	Commodity Spread Carry & Trend	2.0	Commodity Trend	0.9
				Commodity Carry (spread) 2	1.2
				Commodity Carry (spread) 1	0.5
				Commodity Trend (T/S)	7.5
Volatility Sensitive	6.5	Commodity Volatility (short) Grains	3.7	Commodity Volatility (short) Grains	3.7
		Rates Volatility (short) Plus	8.4	Rates Volatility (short) Plus	8.4
		Commodity Volatility (short) Natural Gas	7.7	Commodity Volatility (short) Natural Gas	7.7
		Multi-Asset Volatility Sensitive Plus	8.7	Commodity Volatility (short) Gold	21.3
				Multi-Asset Volatility Sensitive	9.6
				Commodity Volatility (short) Industrial Metals	5.1
		FX Volatility (short) Plus	9.5	FX Volatility (short) Plus	9.5
Value Oriented	9.3	Stocks Value	8.3	Stocks Value 1	6.4
				Stocks Value 2	3.0
		Equity Reversal Plus	38.9	Equity Reversal	37.2
				Equity Trend (T/S) N. America Dynamic	22.9
		Multi-Asset Reversal Oriented Plus	3.4	Equity Volatility (arbitrage) N. America	13.7
				Commodity Reversal	0.2
				Multi-Asset Reversal Oriented	3.3
				FX Value	4.0
Equity Sensitive	6.6	Equity Trend	3.0	Equity/Credit Trend (T/S)	5.5
				Equity Multi-Style	5.5
		Equity Volatility (short)	12.6	Equity Volatility (short) N. America	13.3
				Equity Volatility (short) Europe	13.4
Crude Oil Volatility (short)	6.5	Crude Oil Volatility (short)	6.5	Crude Oil Volatility (short)	6.5
FX/Multi-Asset Trend	2.3	FX Trend (T/S)	4.7	FX Trend (T/S)	4.7
		Multi-Asset Trend (T/S)	1.6	Multi-Asset Trend (T/S)	1.6
FX Carry	3.0	FX Carry	3.0	FX Carry (spread) Developed Mkt Focus	3.8
				FX Carry (spread) EM Focus	2.6
Commodity Curve Carry	2.4	Commodity Curve Carry	2.4	Commodity Carry (curve) 2	4.0
				Commodity Carry (curve) 1	1.7
Rates Carry	4.0	Rates Carry	4.0	Rates Carry (spread)	1.3
				Rates Trend (T/S)	3.5
				Rates Carry (curve)	4.6

Appendix E continued

2018-2020 Autocorrelation (lag 1)

Broad	Autocorr	Hypo-Broad	Autocorr	Super-Base	Autocorr
Stocks, Value Light	-0.06	Stocks Trend (C/S) Plus	-0.09	Stocks Multi-Style N. America	-0.03
				Stocks Trend (C/S)	-0.09
		Stocks Multi-Style Plus	-0.05	Stocks Multi-Style	-0.10
				Stocks Multi-Style Europe	0.03
		Stocks Risk Anomaly	-0.16	Stocks Multi-Style Asia-Pac	-0.16
				Stocks Risk Anomaly 1	-0.13
				Stocks Risk Anomaly 2	-0.11
Commodity Spread Carry & Trend	-0.04	Commodity Spread Carry & Trend	-0.04	Commodity Trend	-0.06
				Commodity Carry (spread) 2	0.03
				Commodity Carry (spread) 1	-0.12
				Commodity Trend (T/S)	0.22
Volatility Sensitive	0.21	Commodity Volatility (short) Grains	-0.07	Commodity Volatility (short) Grains	-0.07
		Rates Volatility (short) Plus	0.07	Rates Volatility (short) Plus	0.07
		Commodity Volatility (short) Natural Gas	-0.05	Commodity Volatility (short) Natural Gas	-0.05
		Multi-Asset Volatility Sensitive Plus	0.37	Commodity Volatility (short) Gold	0.28
				Multi-Asset Volatility Sensitive	0.14
				Commodity Volatility (short) Industrial Metals	0.04
		FX Volatility (short) Plus	0.35	FX Volatility (short) Plus	0.35
Value Oriented	0.07	Stocks Value	-0.10	Stocks Value 1	0.00
				Stocks Value 2	-0.17
		Equity Reversal Plus	0.28	Equity Reversal	-0.29
				Equity Trend (T/S) N. America Dynamic	0.52
		Multi-Asset Reversal Oriented Plus	0.02	Equity Volatility (arbitrage) N. America	-0.11
				Commodity Reversal	0.08
				Multi-Asset Reversal Oriented	0.22
				FX Value	-0.05
Equity Sensitive	0.08	Equity Trend	0.13	Equity/Credit Trend (T/S)	-0.05
				Equity Multi-Style	0.06
		Equity Volatility (short)	0.02	Equity Volatility (short) N. America	0.18
				Equity Volatility (short) Europe	-0.14
Crude Oil Volatility (short)	-0.11	Crude Oil Volatility (short)	-0.11	Crude Oil Volatility (short)	-0.11
FX/Multi-Asset Trend	-0.18	FX Trend (T/S)	-0.21	FX Trend (T/S)	-0.21
		Multi-Asset Trend (T/S)	-0.16	Multi-Asset Trend (T/S)	-0.16
FX Carry	0.10	FX Carry	0.10	FX Carry (spread) Developed Mkt Focus	0.05
				FX Carry (spread) EM Focus	0.12
Commodity Curve Carry	-0.11	Commodity Curve Carry	-0.11	Commodity Carry (curve) 2	-0.21
				Commodity Carry (curve) 1	0.00
Rates Carry	0.00	Rates Carry	0.00	Rates Carry (spread)	0.02
				Rates Trend (T/S)	0.05
				Rates Carry (curve)	0.09

Appendix E continued

2000-2017 Autocorrelation (lag 1)

Broad	Autocorr	Hypo-Broad	Autocorr	Super-Base	Autocorr
Stocks, Value Light	0.06	Stocks Trend (C/S) Plus	0.08	Stocks Multi-Style N. America	0.03
				Stocks Trend (C/S)	0.07
		Stocks Multi-Style Plus	0.05	Stocks Multi-Style	0.01
				Stocks Multi-Style Europe	-0.03
		Stocks Risk Anomaly	0.02	Stocks Multi-Style Asia-Pac	-0.04
				Stocks Risk Anomaly 1	0.00
				Stocks Risk Anomaly 2	0.03
Commodity Spread Carry & Trend	-0.06	Commodity Spread Carry & Trend	-0.06	Commodity Trend	-0.04
				Commodity Carry (spread) 2	-0.04
				Commodity Carry (spread) 1	0.00
				Commodity Trend (T/S)	-0.08
Volatility Sensitive	0.08	Commodity Volatility (short) Grains	-0.08	Commodity Volatility (short) Grains	-0.08
		Rates Volatility (short) Plus	0.05	Rates Volatility (short) Plus	0.05
		Commodity Volatility (short) Natural Gas	-0.11	Commodity Volatility (short) Natural Gas	-0.11
		Multi-Asset Volatility Sensitive Plus	0.20	Commodity Volatility (short) Gold	-0.04
				Multi-Asset Volatility Sensitive	0.17
				Commodity Volatility (short) Industrial Metals	0.08
		FX Volatility (short) Plus	-0.10	FX Volatility (short) Plus	-0.10
Value Oriented	0.08	Stocks Value	0.22	Stocks Value 1	0.15
				Stocks Value 2	0.19
		Equity Reversal Plus	-0.19	Equity Reversal	-0.24
				Equity Trend (T/S) N. America Dynamic	-0.08
		Multi-Asset Reversal Oriented Plus	0.06	Equity Volatility (arbitrage) N. America	-0.19
				Commodity Reversal	-0.01
				Multi-Asset Reversal Oriented	0.17
				FX Value	0.02
Equity Sensitive	-0.02	Equity Trend	-0.07	Equity/Credit Trend (T/S)	0.01
				Equity Multi-Style	-0.08
		Equity Volatility (short)	0.10	Equity Volatility (short) N. America	0.07
				Equity Volatility (short) Europe	0.07
Crude Oil Volatility (short)	0.01	Crude Oil Volatility (short)	0.01	Crude Oil Volatility (short)	0.01
FX/Multi-Asset Trend	-0.01	FX Trend (T/S)	-0.01	FX Trend (T/S)	-0.01
		Multi-Asset Trend (T/S)	0.01	Multi-Asset Trend (T/S)	0.01
FX Carry	-0.03	FX Carry	-0.03	FX Carry (spread) Developed Mkt Focus	-0.03
				FX Carry (spread) EM Focus	-0.01
Commodity Curve Carry	-0.01	Commodity Curve Carry	-0.01	Commodity Carry (curve) 2	0.00
				Commodity Carry (curve) 1	-0.03
Rates Carry	-0.02	Rates Carry	-0.02	Rates Carry (spread)	-0.06
				Rates Trend (T/S)	0.02
				Rates Carry (curve)	-0.01

Appendix F

The diversified ARP funds include those listed in Bloomberg with a minimum of three years of weekly returns. Fund excess returns subtract the cash return associated with the Bloomberg currency code (EUR, GBP or USD) from the reported fund return.

#	Abbrev	Source	Name
1	QRPIX	Bloomberg	AQR ALTERNAT RISK PREMIA-I
2	QSPIX	Bloomberg	AQR STYLE PREMIA ALT-I
3	CHMPZIA	Bloomberg	AXAWF-CHORUS MS-ZI USD
4	BCVLAUI	Bloomberg	BCV LIQUID ALTERN BETA-IUSD
5	BSSAA2U	Bloomberg	BLACKROCK SF STYLE ADV-A2USD
6	CFMISDC	Bloomberg	CFM INSTIT SYS DIVER FD-C
7	CLAAX	Bloomberg	COL MULTI STRAT ALTER FD-A
8	FULABEU	Bloomberg	FULCRUM RISK PREMIA-E USD
9	GSBDMAI	Bloomberg	GAM STAR DIVER ALT- A-I
10	GSARPIU	Bloomberg	GSLIF-ALT RSK PRM PF-IACC
11	JPMSAAE	Bloomberg	JPMORGAN SYSTEMATIC ALPH-AA
12	LFABRIU	Bloomberg	LFIS VISION-PREMIA OPP-I-USD
13	LGTAXGC	Bloomberg	LGT-A GENERIX UCITS-C USD
14	LARPUIA	Bloomberg	LO FUNDS-ALT RISK PR-USDNA
15	MANABST	Bloomberg	MAN ALT RISK PREMIA SP-A USD
16	NNMAICU	Bloomberg	NN L MULTI AST FAC OPP-ICUSD
17	NMAPBIE	Bloomberg	NORDEA 1-ALPHA 15MA-BI EUR
18	QFARPEA	Bloomberg	QUONIAM-ALT RISK PREMIA-EAD
19	SCHTSIU	Bloomberg	SCHRODER GAIA TWO SIG DVF-IU
20	SPRP2MU	Bloomberg	SERVICED PL-AA RPE-M
21	SARPCUN	Bloomberg	SYSTEMATICA ALT RSK P-CUSDND
22	UGARPRA	Bloomberg	UNI-GB ALT RISK PRE-RA USD

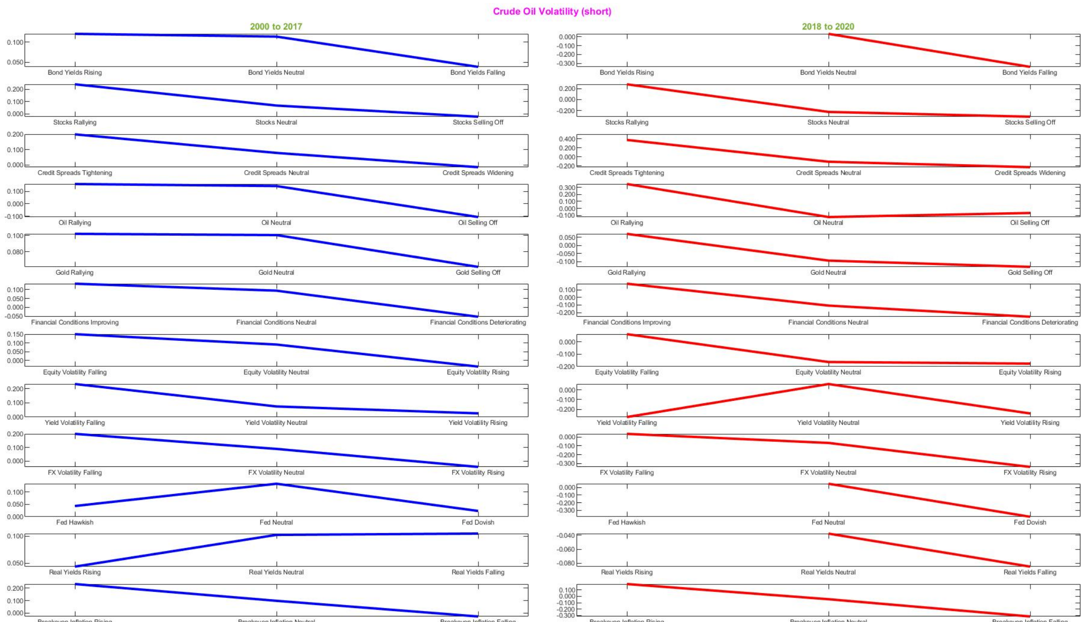
Appendix G

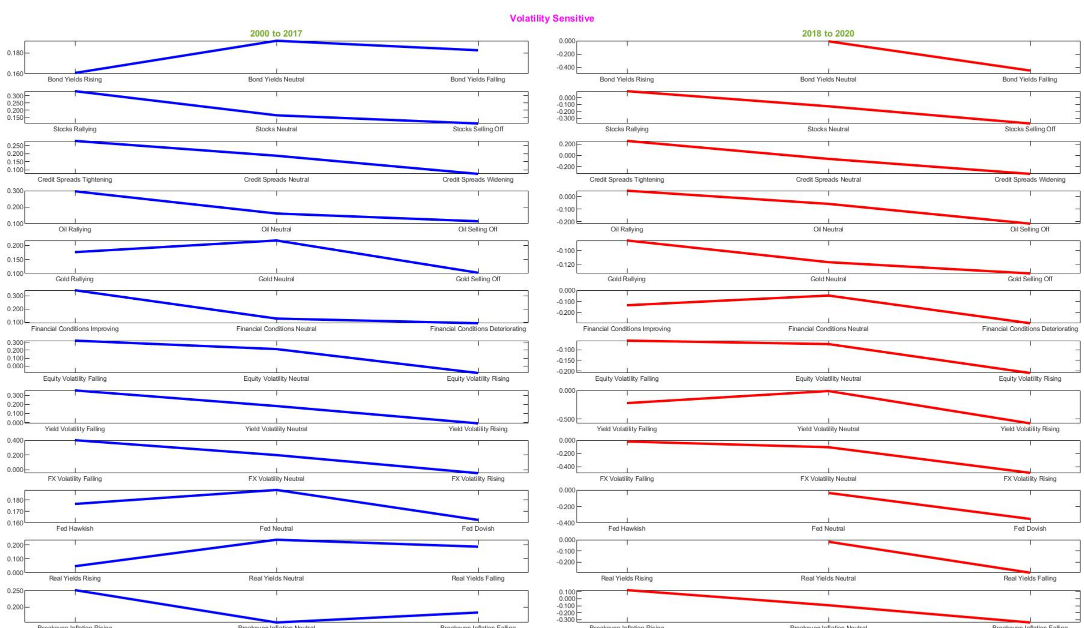
This table provides data sources for the 14 daily financial market states considered in this paper and the 60-day change in the 3-day moving average necessary to produce a state representing 20-25% of the 2000-2020 data history. The US dollar and yield curve states (the last four rows in this table) contain zero or very few extreme observations during the 2018-2020 period and therefore provide no basis for comparison with the preceding window.

State	60d Change	Source
Rising Treasury Yields Falling Treasury Yields	60 bps -70 bps	FRED 10-Year US Treasury Constant Maturity Yield
Hawkish Fed Dovish Fed	35 bps -35 bps	FRED Effective Federal Funds Rate
Rising Equity Volatility Falling Equity Volatility	11 pts -11pts	CBOE VIX Index
Rising Treasury Volatility Falling Treasury Volatility	30 pts -28 pts	ICE BofA MOVE Index
Rising FX Volatility Falling FX Volatility	2.0 pts -2.2 pts	J.P. Morgan G7 Currency Volatility Index
Deteriorating Financial Conditions Improving Financial Conditions	-1.25 pts 1.00 pts	Bloomberg United States Financial Conditions Index
Rising Real Treasury Yields Falling Real Treasury Yield	40 bps -50 bps	FRED 10-Year US Treasury Inflation-Indexed Constant Maturity Yield
Rising Crude Oil Price Falling Crude Oil Price	28% -20%	Generic 1st WTI Crude Oil Future
Equity Sell-Off Equity Rally	-9% 11%	S&P 500
Falling Gold Price Rising Gold Price	-8% 14%	Generic 1st Gold Future
Widening Credit Spreads Tightening Credit Spreads	45 bps -45 bps	Moody's US Corporate BAA 10 Year Spread
Rising Inflation Expectations Falling Inflation Expectations	35 bps -35 bps	Bloomberg US TIPS 10-Year Inflation Breakeven
Steepening Yield Curve Flattening Yield Curve	55 bps -55 bps	FRED 10-Year minus 1-Year US Treasury Constant Maturity Yield
Strengthening US Dollar Weakening US Dollar	6.5% -6.5%	Average of CHF, JPY and EUR

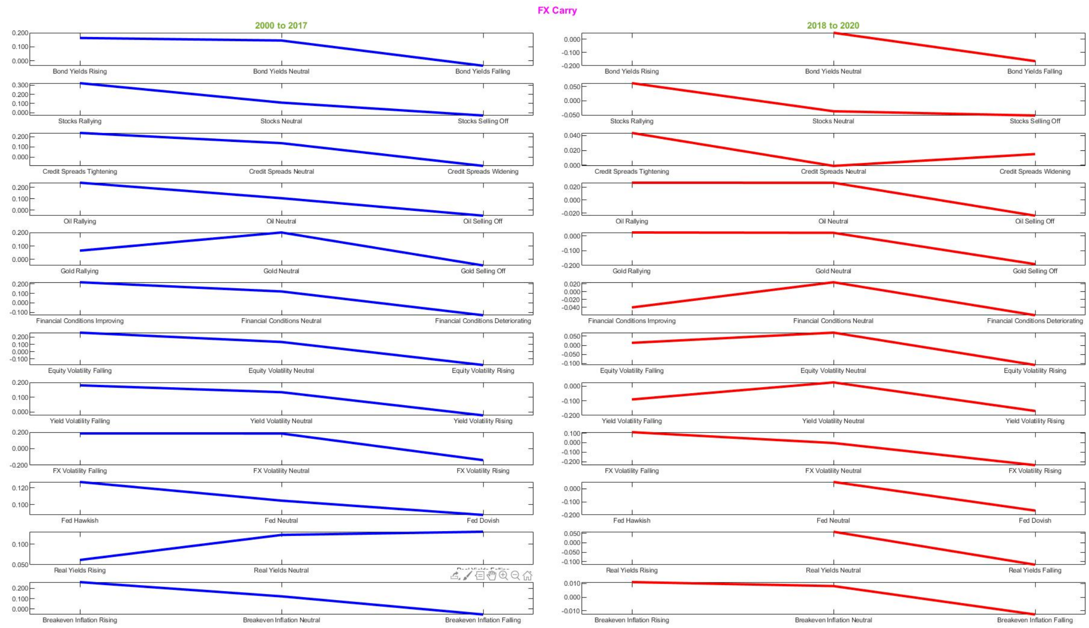
Appendix H

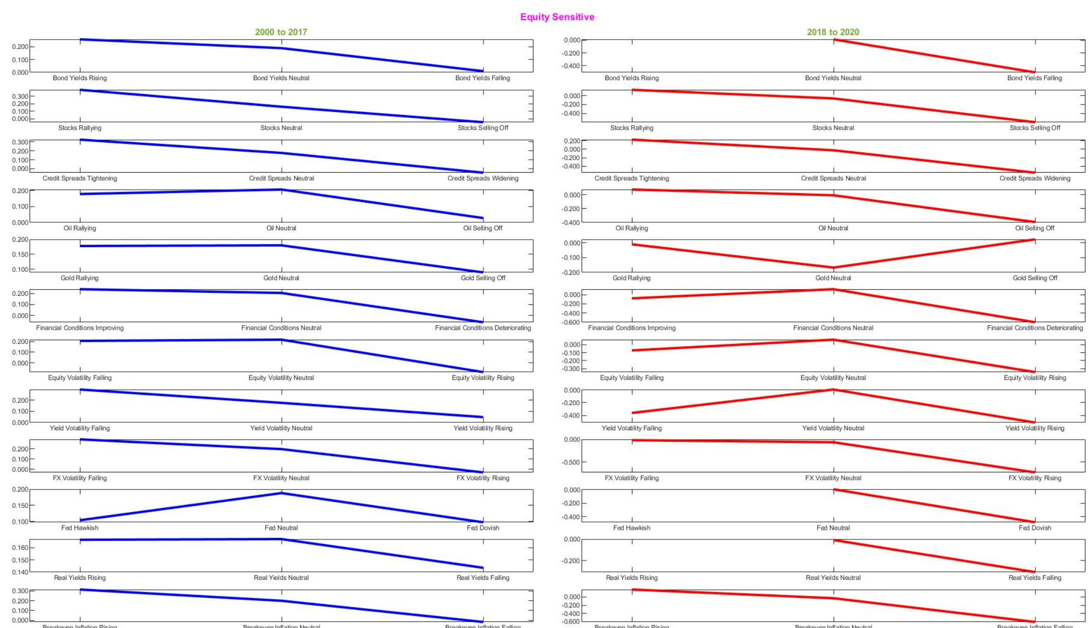
These exhibits compare the in-sample and out-of-sample conditional means for the 10 broad ARP benchmarks across 12 market states. The regimes are non-independent but also non-identical, offering different points of emphasis. The blue lines on the left represent the 2000-2017 period and the red lines on the right the 2018-2020 period. Pro-risk states appear on the left of each column. Three states, hawkish Fed and rising nominal and real yields, did not occur during the 2018-2020 period.



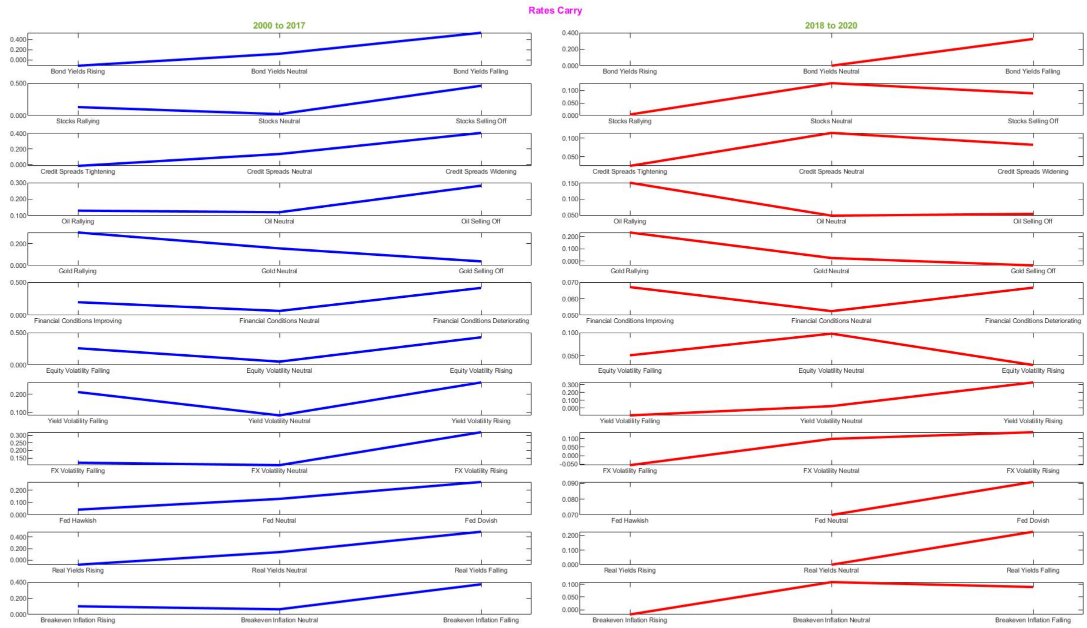


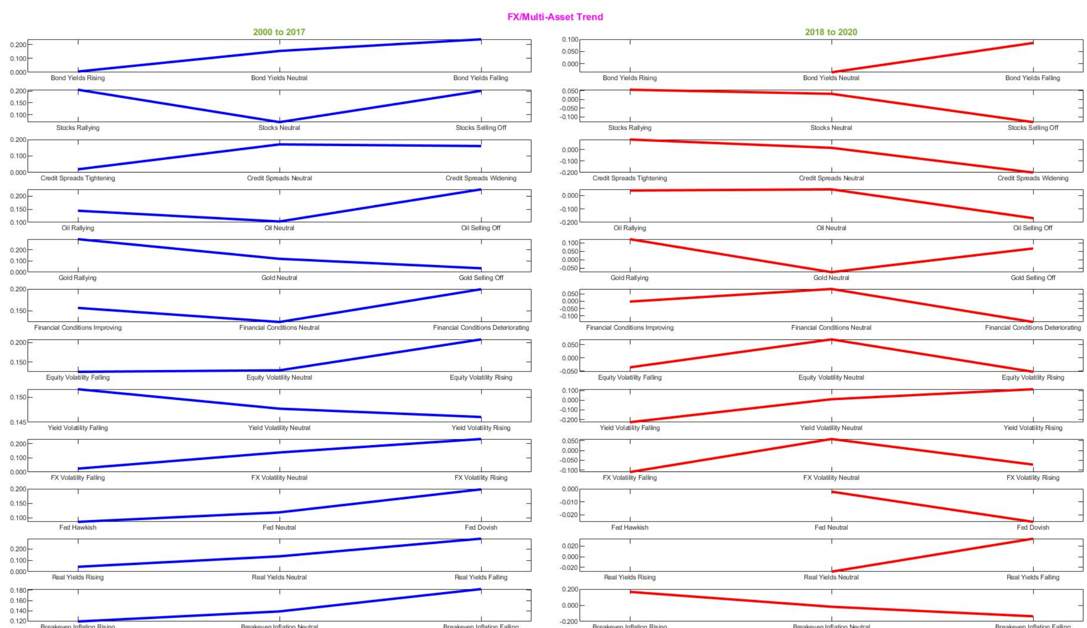
Appendix H continued



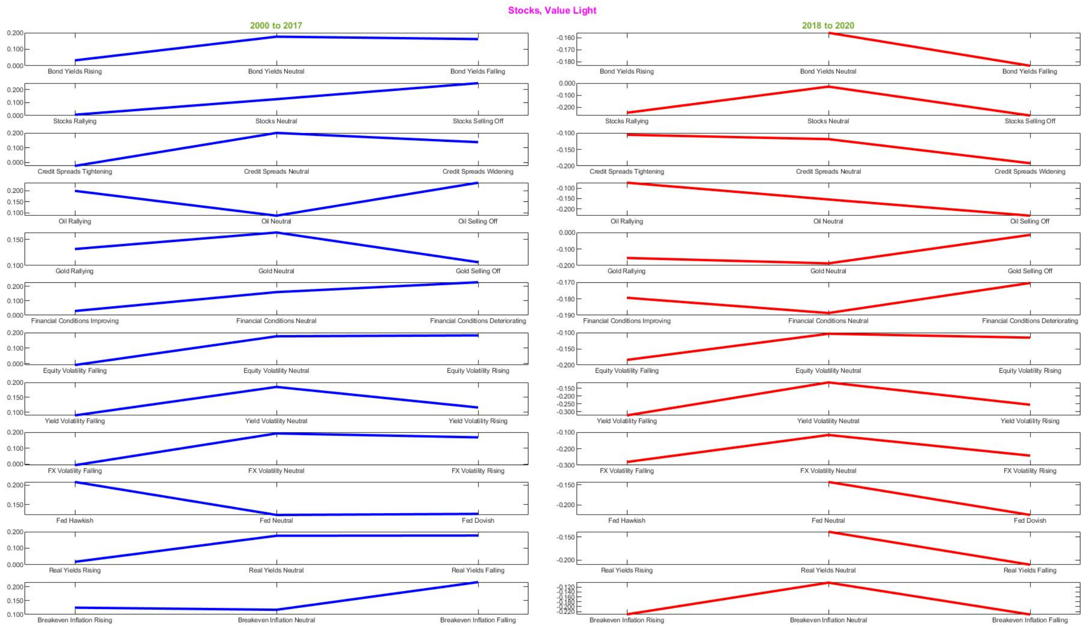


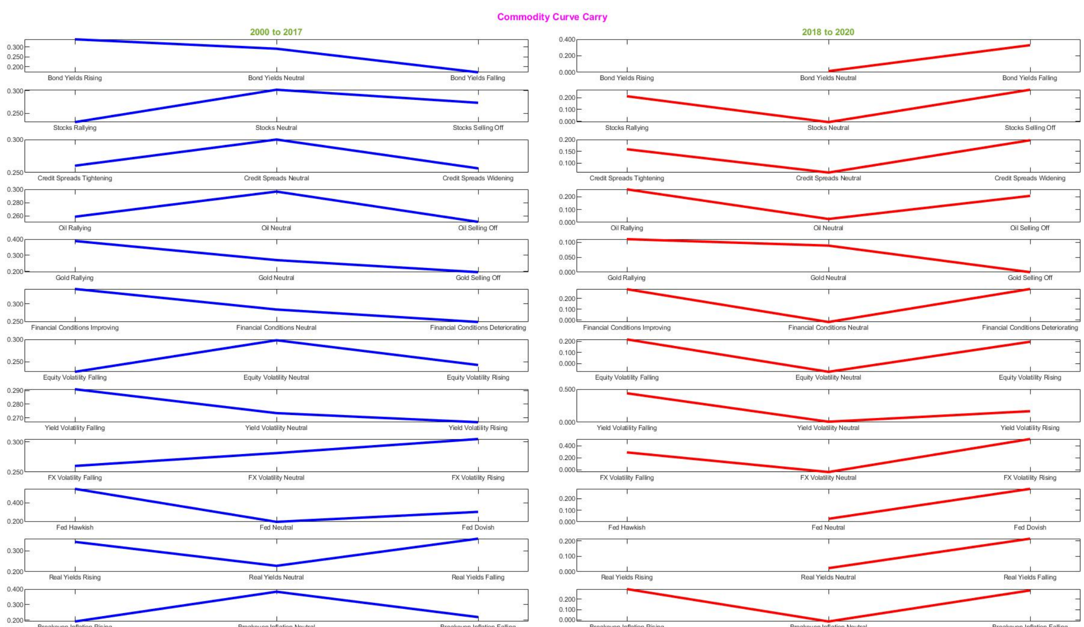
Appendix H continued



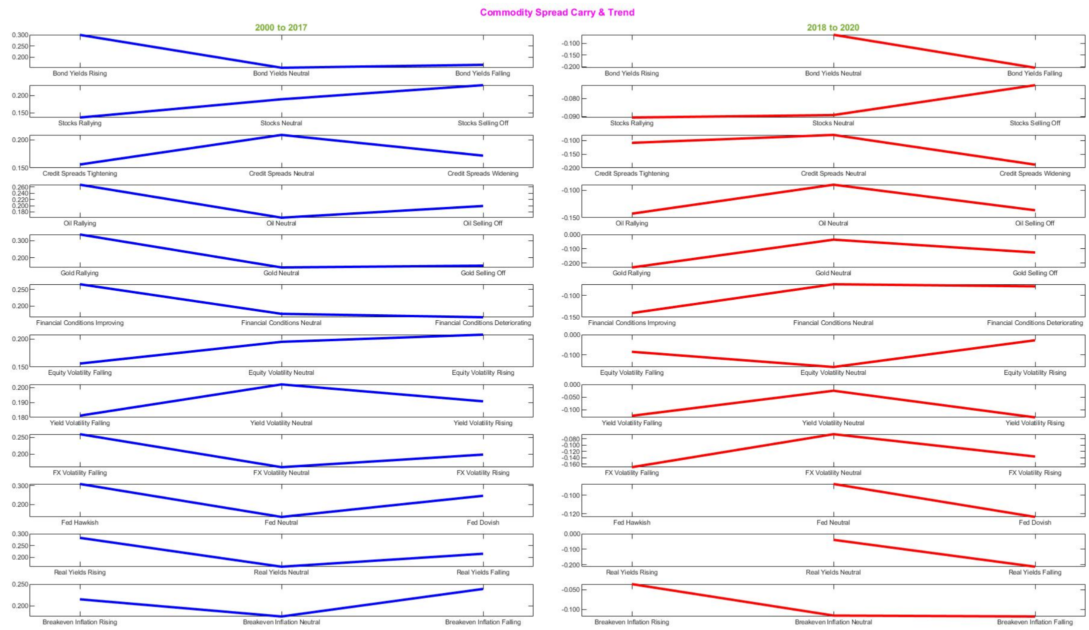


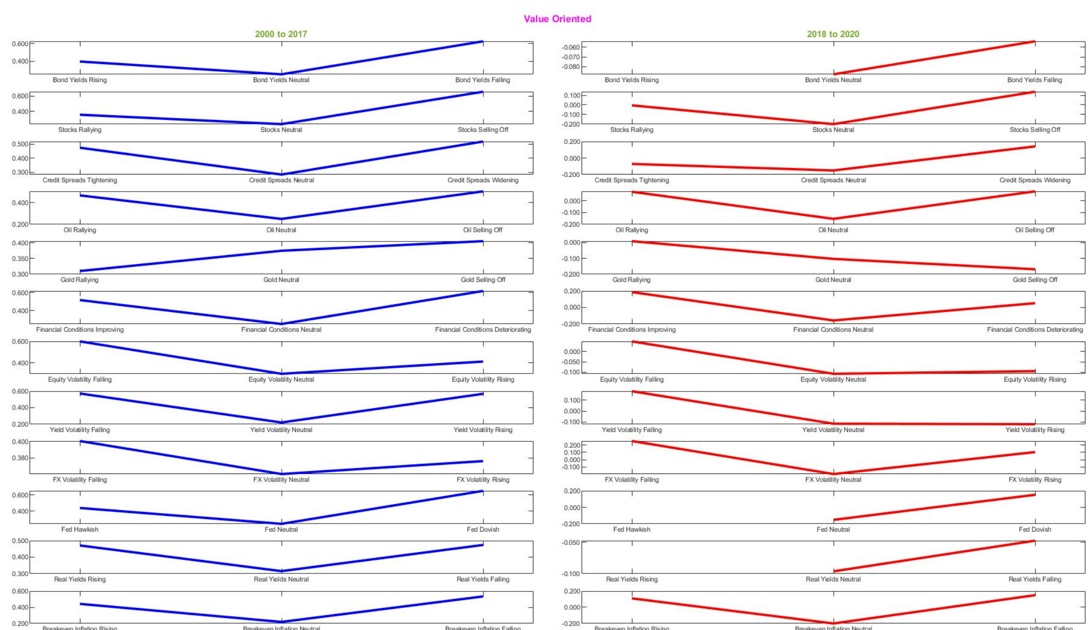
Appendix H continued





Appendix H continued





Appendix I

Market Whiplash in 2018-2020

This table uses weekly total returns for the S&P 500 between January 2000 and December 2020 to summarize the return generation process, distinguishing significant drawdown cycles from frictional drawdown cycles and accumulation periods. The numbers in parentheses indicate the number of drawdowns. The return ratio is the 2018-2020 weekly return divided by the 2000-2017 return. Because the table presents results in terms of weeks, the return for the growth period is inflated to account for stub returns within recovery periods (i.e. returns within a recovery week in excess of the trough-to-peak retracement). The total return decomposition is accurate.

	% of Weeks	Weekly Return	Return Ratio
Drawdown greater than 5%			
Drawdown			
2000-2017 (8)	25.5%	-0.81%	
2018-2020 (6)	27.4%	-1.98%	*2.4x*
Recovery			
2000-2017 (8)	47.2%	0.44%	
2018-2020 (6)	44.6%	1.24%	*2.8x*
Drawdown less than 5%			
Drawdown			
2000-2017 (50)	11.5%	-0.66%	
2018-2020 (10)	7.0%	-0.72%	*1.1x*
Recovery			
2000-2017 (50)	8.6%	0.89%	
2018-2020 (10)	7.0%	0.72%	*0.8x*
Growth (outside drawdown cycle)			
2000-2017	7.2%	1.40%	
2018-2020	14.0%	1.82%	*1.3x*
Total			
2000-2017	100.0%	0.10%	
2018-2020	100.0%	0.25%	*2.5x*

Tables

Table 1 ***Properties of a Valid Benchmark***
This table provides the frequently referenced benchmark characteristics of Maginn *et al*. (2007).

Property	Explanation
Specified in advance	The benchmark is specified prior to the start of an evaluation period and known to all interested parties.
Appropriate	The benchmark is consistent with the manager's investment style or area of expertise.
Measurable	The benchmark's return is readily calculable on a reasonably frequent basis.
Unambiguous	The identities and weights of securities or factor exposures constituting the benchmark are clearly defined.
Reflective of current investment opinions	The manager has current investment knowledge (be it positive, negative, or neutral) of the securities or factor exposures within the benchmark.
Owned	The investment manager should be aware of and accept accountability for the constituents and performance of the benchmark. It is encouraged that the benchmark be embedded in and integral to the investment process and procedures of the investment manager
Investable	It is possible to forgo active management and simply hold the benchmark.

Table 2 ***ARP Research based upon Tradable Indices***

This table summarizes the small set of papers utilizing a database of tradable indices from a cross-section of investment banks. The table highlights the paucity of research, limited historical data, different ARP classification systems, and the absence of focus on 2018 through 2020 (a broadly disappointing performance period).

Author	Data Set	Time Period	Strategy Composites	ARP Focus
Hamdan *et al.* (2016)	624 indices from 11 banks plus many equity long-only ETF's and benchmarks	Jun 2000 to Dec 2015	• 59 statistically classified across 4 asset classes • 46 bank-classified across 5 asset classes and 11 styles (25 separately deemed relevant)	• Broad overview • Diversification potential • Non-normality • Usefulness in explaining hedge fund returns
Suhonen, Lennkh and Perez (2017)	215 indices from 15 banks	Dec 1999 to Mar 2015	• 32 bank-classified across 4 asset classes (plus multi-asset) and 11 styles	• Post-publication return deterioration (overfitting risk) • Complexity risk
Vatanen and Suhonen (2019)	Undisclosed number of indices from 7 banks	Jan 2007 to May 2018	• 28 bank-classified across 5 asset classes and 8 styles	• Offensive and defensive profile of styles • Diversification potential • Vulnerability during very weak stock and bond markets
Naya and Tuchschmid (2019)	234 indices from 12 banks, varies within paper	Jun 2010 to Apr 2017	• 32 bank-classified across 5 asset classes and 11 styles	• Heterogeneity within styles • Post-publication return deterioration (overfitting risk)
Baltas and Scherer (2019)	262 indices from 6 banks	Feb 2008 to Jan 2018	• 6 bank-classified multi-style asset class groups • 8 bank-classified multi-asset style groups	• Vulnerability during very weak stock and bond markets

Table 3 ***Metadata Requested in Recurring Tradable Index Survey***

This table provides the 18 data fields requested from 16 investment banks offering tradable index products as of May 2020. Given the absence of data standardization across banks, the survey reflects considerable collaborative engagement with the banks to ensure that survey responses reflect a consistent interpretation of each field.

Field	Description
Bloomberg Ticker	Identifier to access historical index levels
Index Name	Bank-assigned name
Objective	5 Options: Alternative Risk Premium, Enhanced Beta, Systematic Alpha, Traditional Beta, Other
Style	14 Options: Carry (spread), Carry (curve), Congestion (rebalance, month-end), Merger Arbitrage, Multi-Style, Other, Reversal, Risk Anomaly (quality, low volatility/beta), Size, Trend (cross-sectional momentum), Trend (time-series momentum), Value, Volatility (arbitrage), Volatility (short)
Asset Class	7 Options: Equity: Index-based, Equity: Stock-based, Commodity, Credit, Currency, Multi-Asset, Rates
Directionality	3 Options: Long-Only, Long-Short, Short-Only (reflects positioning, not beta)
Region	5 Options: North America, Europe, Asia-Pacific, Emerging Markets, Multi-Region
Index Description	Brief explanation of index structure (e.g., sells 1mo S&P 500 straddles and buys a 5-delta put as downside insurance)
Index Fee	Some banks charge an index construction/operation fee -- basis points per annum, single fee or range and explanatory note
Swap Spread	Typical structure is to receive the index return and to pay LIBOR plus a spread -- basis points per year, single spread or range and explanatory note
In/Out Costs	Entry and/or exit fee on certain strategies (effectively a commission or transaction cost supplement) - basis points per trade and explanatory note
Trading Costs	Transaction costs embedded in the index return calculation – basis points per annum, historical average or projection, single estimate or range and explanatory note
History Start Date	Inception date for the back-tested index returns
Live Start Date	Formal publication date – defines out-of-sample index returns, may precede funding date
Return Type	2 Options: Excess Return (excludes a cash return), Total Return (includes a cash return)
FX Denomination	8 Options: USD, EUR, JPY, GBP, AUD, CAD, CHF, Other
Dealing Terms	4 Options: Daily, Weekly, Monthly, Other
US Availability	2 Options: Yes, No -- indicates whether US investors can access the index

Table 4 ***Tradable Index Classification Systems***

This table compares the top-level style classification used in five recent papers (8 to 11 options) with that appearing in this paper (14 options). Note that the authors in the first two columns also included a second level in their classification systems. While a reasonable amount of consistency exists, variation in both terminology and views on strategy independence complicates comparing results across studies.

Hamdan, Pavlowsky, Roncalli and Zheng (2016)	Vatanen and Suhonen (2019)	Naya and Tuchschmid (2019)	Suhonen, Lennkh and Perez (2017)	Baltas and Scherer (2019)	ARP Style in This Paper
			Asset Allocation		—
Carry	Cross-Sectional Carry	Carry	Carry	Carry	Carry (spread)
	Curve Carry		Curve	Curve	Carry (curve)
	Equity Specific			Equity Factor	—
Event			Event-Driven		Merger Arbitrage
Growth					—
Liquidity	Flow Based		Liquidity		Congestion (rebalance, month-end)
			Macro		—
		Merger Arbitrage			Merger Arbitrage
				Multifactor	Multi-Style
Low Volatility		Low Vol/Beta			Risk Anomaly (quality, low volatility/beta
Momentum	Cross-Sectional Momentum	Momentum		Momentum	Trend (cross-sectional momentum)
			Other		Other
		Profitability			Risk Anomaly (quality, low volatility/beta
Quality		Quality			Risk Anomaly (quality, low volatility/beta)
Reversal		Mean Reversion	Mean reversion		Reversal
Size		Size			Size
	Time-Series Momentum	Trend	Trend Following	Trend	Trend (time-series momentum)
Value	Value	Value	Value	Value	Value
Volatility	Short Volatility Carry	Volatility Carry	Volatility	Volatility	Volatility (short)
					Volatility (arbitrage)
11	8	11	11	8	14

Table 5 ***Tradable Index Asset Class and Style Crosstab***

This contingency table shows the intersection between asset class and style. The green shading highlights concentration at the joint level while the red shading highlights concentration at the style or asset class level. Equity index and commodity short volatility indices are the largest offerings, followed by stock-based risk anomaly, commodity curve carry, rates short volatility, commodity congestion, and stock-based multi-style.

		Asset Class							
		Equity (stock-based)	Equity (index-based)	Commodity	Credit	Currency	Multi-Asset	Rates	*Style Total*
Style	Carry (curve)	0	15	90	16	0	0	54	*175*
	Carry (spread)	5	3	60	14	63	3	27	*175*
	Congestion (rebalance, month-end)	0	9	74	0	5	1	9	*98*
	Merger Arbitrage	8	0	0	0	0	0	0	*8*
	Multi-Style	73	3	28	0	12	21	10	*147*
	Other	13	5	3	0	3	10	0	*34*
	Reversal	4	57	30	0	7	5	0	*103*
	Risk Anomaly (quality, low volatility/beta)	131	2	4	0	0	0	1	*138*
	Size	27	1	0	0	0	0	0	*28*
	Trend (cross-sectional momentum)	60	2	34	0	14	8	6	*124*
	Trend (time-series momentum)	1	40	54	27	40	35	29	*226*
	Value	62	7	20	1	34	0	15	*139*
	Volatility (arbitrage)	23	36	10	0	0	0	1	*70*
	Volatility (short)	3	177	151	8	49	0	79	*467*
	Asset Class Total	*410*	*357*	*558*	*66*	*227*	*83*	*231*	*1,932*

Table 6 ***Comparison of Categorical and Statistical Base Benchmark Assignments via a Similarity Index***

This table provides several comparisons of the two benchmark partitions of the tradable index universe into 85 base groups. The metrics are standardized on a unit scale. The universe size inflates the Rand index because pairs not combined in either benchmark group contribute to similarity. The other two approaches exclude such pairs from the similarity calculation. The adjusted index includes a correction for agreement due to chance.

Similarity Index	Value
Morlini & Zani	0.21
Adjusted Morlini & Zani	0.18
Fowlkes & Mallows	0.22
Rand	0.95

Table 7 ***Principal Component Summary for the ARP Benchmark Families***

This table summarizes the percent of base benchmark variance explained by the first three principal components. The first data column shows the percent of underlying strategies within each of the 85 base benchmarks for which the percent of variance attributable to the first principal component exceeds 40%. The remaining columns repeat this calculation for the first two and first three components, respectively versus 55% and 65% thresholds. The statistical benchmark family offers materially greater homogeneity within its base benchmark tier than the categorical approach. Weekly data between December 2004 and August 2020 underpins the PCA.

	PC 1	PC 1 & 2	PC 1, 2 & 3
Benchmark Family	**≥ 40%**	**≥ 55%**	**≥ 65%**
Categorical	58%	58%	59%
Statistical	83%	80%	78%

Table 8 Annual Sharpe Ratio for Categorical ARP Benchmarks

Using weekly reported tradable index return between December 1999 and August 2020, Panel A provides the calendar year Sharpe ratio for the 14 categorical style benchmarks. The top portion of the table shows the effect of aggregating many strong back tests. The bottom part of the table highlights a markedly different realized experience. Panel B repeats this exercise for the 7 categorical asset benchmarks. Dark blue shading indicates the highest ratios and red the lowest. For the cross-sectional and time-series medians, dark green shading designates the highest ratios and yellow the lowest.

Panel A

Calendar Year	Carry (curve)	Carry (spread)	Congestion (rebalance, month-end)	Merger Arbitrage	Multi-Style	Other	Reversal	Risk Anomaly (quality, low volatility/beta)	Size	Trend (cross-sectional momentum)	Trend (time-series momentum)	Value	Volatility (arbitrage)	Volatility (short)	Style Median
2000	1.4	2.5	1.4	N/A	2.7	0.9	2.4	2.8	N/A	1.2	2.4	2.2	2.4	2.6	2.4
2001	4.1	3.3	1.9	N/A	6.7	-0.1	0.9	1.2	N/A	0.7	1.4	2.9	-0.5	1.8	1.6
2002	2.5	4.7	2.5	N/A	5.5	1.6	3.2	2.3	N/A	2.8	3.6	6.5	0.4	2.8	2.8
2003	1.7	3.4	1.3	N/A	5.3	3.4	3.6	0.6	2.1	1.3	1.9	8.0	0.0	3.1	2.1
2004	5.1	4.5	6.1	N/A	6.4	2.2	1.0	3.9	1.9	2.3	1.9	4.6	2.4	2.1	2.4
2005	5.4	3.0	4.4	2.6	5.6	2.0	2.3	0.0	0.5	2.1	1.1	5.2	1.3	0.9	2.2
2006	6.7	4.6	5.5	4.3	4.6	1.6	1.9	3.7	-0.2	1.9	1.8	2.4	1.1	0.5	2.2
2007	5.5	2.6	5.3	0.5	2.6	0.9	1.7	0.4	-2.0	2.1	1.8	1.1	-0.4	0.1	1.4
2008	1.4	-0.2	3.1	-0.4	3.3	0.3	3.1	1.4	0.0	1.9	2.8	3.1	1.6	-0.7	1.5
2009	5.8	3.6	3.7	3.3	7.5	3.6	3.1	-0.4	1.9	0.1	1.9	4.4	4.1	8.3	3.6
2010	2.9	4.1	3.0	3.3	4.4	2.5	2.1	1.2	1.6	2.0	1.1	3.1	1.2	2.5	2.5
2011	1.7	1.4	3.0	1.2	4.5	2.9	3.3	2.8	-0.6	1.8	1.5	1.7	1.6	2.6	1.8
2012	3.0	3.1	2.6	0.7	2.1	1.2	2.0	0.1	-0.3	0.5	1.1	2.9	3.9	5.1	2.0
2013	2.3	0.8	3.3	2.0	2.2	2.5	2.1	0.7	0.9	2.2	1.5	3.0	1.2	1.4	2.0
2014	2.4	1.4	2.9	-0.6	4.6	1.7	2.8	2.7	0.3	2.8	3.6	1.3	0.2	0.5	2.1
2015	2.7	0.5	2.4	0.1	1.3	-0.3	2.1	1.7	0.9	0.3	0.8	1.2	0.2	0.4	0.8
2016	2.4	1.8	1.8	0.3	1.4	0.0	2.6	-0.5	0.4	-0.3	0.5	2.5	0.9	0.1	0.7
2017	1.9	1.0	2.4	0.1	1.4	1.8	2.1	1.0	-1.1	0.6	1.4	3.5	1.9	3.5	1.6
2018	0.4	0.5	0.4	0.1	-0.8	-0.8	-0.6	0.5	-1.8	-1.2	-0.3	-0.1	-0.9	-1.1	-0.4
2019	2.2	1.0	0.6	0.8	0.2	0.7	1.3	-0.4	-0.7	-0.2	0.3	-0.8	-0.6	-0.3	0.2
Aug-2020	0.1	-1.2	2.1	-0.4	-1.3	-0.5	-0.3	-1.1	-1.4	-0.6	-0.2	-0.4	0.8	-0.6	-0.5
Median	2.4	2.5	2.6	0.6	3.3	1.6	2.1	1.0	0.1	1.3	1.5	2.9	1.1	1.4	2.0

Panel B

Calendar Year	Equity (stock-based)	Equity (index-based)	Commodity	Credit	Currency	Multi-Asset	Rates	Asset Median
2000	4.9	2.3	1.9	N/A	2.2	1.3	3.3	2.3
2001	1.8	1.7	3.6	N/A	1.8	0.5	1.7	1.8
2002	2.7	6.0	3.4	N/A	3.4	2.5	3.8	3.4
2003	3.1	5.0	6.0	N/A	5.1	2.8	0.6	4.0
2004	5.4	2.6	10.3	N/A	1.1	2.8	2.2	2.7
2005	2.6	2.9	7.2	0.4	1.4	2.5	1.8	2.5
2006	2.8	1.9	7.4	2.5	0.3	2.1	0.6	2.1
2007	-0.3	1.4	6.2	0.1	0.5	1.4	2.7	1.4
2008	0.9	3.2	5.5	-1.3	0.7	1.9	2.4	1.9
2009	3.4	6.7	5.6	2.7	4.0	3.0	2.5	3.4
2010	3.2	2.3	3.4	1.2	1.4	2.2	2.7	2.3
2011	2.2	2.8	4.7	0.8	1.0	2.5	3.1	2.5
2012	0.8	3.3	3.3	3.9	1.4	1.4	2.5	2.5
2013	3.9	2.5	3.5	3.5	0.4	1.5	0.2	2.5
2014	1.6	0.8	4.9	0.8	2.7	2.5	2.3	2.3
2015	1.7	1.1	2.5	0.7	0.2	0.1	0.8	0.8
2016	0.0	1.0	1.4	1.1	2.1	0.7	0.7	1.0
2017	0.7	4.4	3.6	3.1	0.6	2.4	-0.1	2.4
2018	-1.4	-1.0	0.1	-0.5	-0.3	-1.0	1.1	-0.5
2019	-2.5	0.0	-0.6	0.8	1.4	1.6	1.2	0.8
Aug-2020	-1.6	0.6	0.2	-0.5	-0.9	-0.5	-0.6	-0.5
Median	1.8	2.3	3.6	0.8	1.4	1.9	1.8	2.3

Table 9 ***Annual Sharpe Ratio for Statistical ARP Benchmarks***

Using weekly reported tradable index return between December 1999 and August 2020, this table provides the calendar year Sharpe ratio for the 10 statistical broad benchmarks. The top portion of the table shows the effect of aggregating many strong back tests. The bottom part of the table highlights a markedly different realized experience. Dark blue shading indicates the highest ratios and red the lowest. For the cross-sectional and time-series medians, dark green shading designates the highest ratios and yellow the lowest.

Calendar Year	Crude Oil Volatility	Volatility Sensitive	FX Carry	Equity Sensitive	Rates Carry	FX Trend	Stocks, Value Light	Commodity Curve	Commodity Spread & Trend	Value Oriented	Broad Median
2000	N/A	5.8	2.0	0.3	1.7	1.0	N/A	-0.5	1.3	4.6	1.5
2001	0.1	0.6	1.7	2.3	1.6	1.1	0.3	3.3	2.8	2.8	1.6
2002	2.6	3.0	4.0	1.5	3.2	3.4	1.7	0.9	0.7	6.9	2.8
2003	0.9	1.5	4.7	4.8	0.5	2.4	0.9	1.3	1.1	6.3	1.4
2004	-1.6	2.1	1.9	2.7	1.1	1.4	4.1	6.5	3.2	4.6	2.4
2005	1.5	1.2	2.5	2.0	0.7	0.7	1.9	5.2	2.3	3.9	1.9
2006	1.7	0.9	0.4	1.9	0.6	0.9	2.7	7.0	3.9	4.6	1.8
2007	0.8	1.0	0.9	0.4	2.1	0.9	0.1	5.6	2.3	1.7	1.0
2008	-1.4	-0.8	-1.5	-0.3	2.1	1.5	0.8	5.8	3.0	6.0	1.1
2009	4.5	5.2	2.4	4.2	0.8	0.4	-0.2	2.3	1.8	4.8	2.3
2010	2.7	2.4	0.7	1.6	2.4	1.1	2.3	2.1	1.9	2.3	2.2
2011	0.9	2.8	-0.1	0.6	1.8	0.9	2.3	3.4	1.6	3.7	1.7
2012	4.1	3.4	1.0	3.4	2.0	-0.1	0.3	0.9	0.9	3.2	1.5
2013	2.0	0.7	-0.6	2.3	-0.7	0.9	2.0	1.1	1.5	2.8	1.3
2014	-1.1	0.2	0.2	0.3	2.3	2.4	1.8	2.3	1.8	3.8	1.8
2015	-0.4	0.6	-1.0	-0.4	0.7	0.3	1.8	1.6	1.1	1.7	0.7
2016	0.2	0.1	1.5	0.9	0.7	0.1	-0.6	1.3	-0.5	2.4	0.5
2017	2.3	2.3	-0.5	4.3	-0.7	0.8	1.6	2.6	0.7	0.3	1.2
2018	-1.6	-0.7	-0.1	-1.1	0.6	-0.8	-1.0	0.2	0.0	-1.1	-0.7
2019	0.0	-0.7	2.0	0.7	1.7	0.6	-1.4	0.5	-1.6	-0.8	0.3
Aug-2020	-0.3	-0.7	-0.7	-0.7	-0.2	-0.6	-1.4	1.4	-0.8	0.6	-0.6
Median	0.8	1.0	0.9	1.5	1.1	0.9	1.2	2.1	1.5	3.2	1.5

Table 10 *ARP Benchmark Factor Sensitivity*

Using weekly data between December 2008 and August 2020, Panel A uses elastic net to select a maximum of six factors from the universe in Appendix B to explain the returns of the categorical style and asset benchmarks. p-values reflect Newey-West (1987) corrected standard errors and ρ^2 represents the stand-alone coefficient of determination. The blue, green and red shading of p-values indicates significance respectively at the 1%, 5% and 10% levels. The shading for model adjusted R^2 illustrates simple quartiles, with blue indicating 75-100% explanatory power, gray 50-75%, orange 25-50% and brown 0-25%. Panel B repeats the exercise for the statistical broad benchmarks.

Panel A

Carry (curve)

factor	coeff	p-value	rho^2
constant	0.00	0.99	--
DIVCARR	0.25	0.00	18%
BETA	0.24	0.00	20%
EMEQ	0.01	0.86	21%
EMBDS	0.05	0.33	25%
USHYBD	0.13	0.02	30%
USCORBD	0.26	0.00	18%
Model Adj R^2	--	--	43%

Carry (spread)

factor	coeff	p-value	rho^2
constant	0.00	0.87	--
FXCARR	0.21	0.00	39%
DIVCARR	0.50	0.00	43%
EUREQ	0.13	0.01	25%
EMEQ	-0.02	0.72	29%
EMBDS	0.13	0.00	31%
USHYBD	0.10	0.10	28%
Model Adj R^2	--	--	64%

Congestion (rebalance, month-end)

factor	coeff	p-value	rho^2
constant	0.02	0.62	--
BDCARR	-0.11	0.01	2%
COMCARR	0.33	0.00	13%
BETA	0.10	0.05	3%
USBREV	0.20	0.28	1%
USVAR	0.64	0.00	14%
Model Adj R^2	--	--	26%

Merger Arbitrage

factor	coeff	p-value	rho^2
constant	0.00	0.98	--
BETA	0.09	0.32	16%
RESVOL	0.07	0.23	7%
NAEQ	0.06	0.55	18%
EMEQ	0.01	0.94	12%
USHYBD	0.12	0.14	18%
VIX	-0.14	0.05	8%
Model Adj R^2	--	--	15%

Multi-Style

factor	coeff	p-value	rho^2
constant	0.00	0.96	--
DIVRP6	0.04	0.64	28%
DIVRP	0.22	0.01	30%
DIVCARR	0.35	0.00	31%
DIVTRD	0.26	0.00	11%
EMBDS	0.26	0.00	18%
Model Adj R^2	--	--	51%

Other

factor	coeff	p-value	rho^2
constant	0.00	0.98	--
DIVRP6	0.10	0.26	23%
DIVRP	0.26	0.00	25%
NAEQ	0.31	0.00	28%
EMBDS	0.31	0.00	29%
Model Adj R^2	--	--	46%

Reversal

factor	coeff	p-value	rho^2
constant	0.01	0.73	--
USEQLR	0.11	0.08	8%
USSTREV	0.25	0.00	9%
BANKS	-0.14	0.01	1%
EMBDS	0.11	0.06	10%
VIX	-0.16	0.00	8%
USBREV	0.20	0.13	5%
Model Adj R^2	--	--	16%

Risk Anomaly (quality, low volatility/beta)

factor	coeff	p-value	rho^2
constant	-0.01	0.78	--
USEQLR	0.49	0.00	27%
EURHML	-0.17	0.00	37%
EURRMW	0.26	0.00	33%
BETA	-0.35	0.00	33%
Model Adj R^2	--	--	70%

Size

factor	coeff	p-value	rho^2
constant	0.00	0.98	--
USEQVAL	0.15	0.00	21%
USSMB	0.42	0.00	25%
USWML	-0.28	0.00	15%
EURSMB	0.18	0.00	7%
SIZE	-0.27	0.00	19%
SIZENL	0.41	0.00	3%
Model Adj R^2	--	--	62%

Trend (cross-sectional momentum)

factor	coeff	p-value	rho^2
constant	0.00	0.87	--
DIVRP6	0.19	0.00	18%
DIVTRD	0.44	0.00	45%
EURWML	0.11	0.04	33%
JPNWML	0.11	0.00	23%
MOM	0.18	0.00	30%
GOVEXUS	0.17	0.00	19%
Model Adj R^2	--	--	65%

Trend (time-series momentum)

factor	coeff	p-value	rho^2
constant	0.00	0.97	--
DIVRP6	0.04	0.09	10%
FXTRD	0.07	0.03	41%
EQTRD	0.11	0.00	42%
DIVTRD	0.69	0.00	75%
ILBUS	0.18	0.00	4%
VIX	-0.15	0.00	5%
Model Adj R^2	--	--	80%

Value

factor	coeff	p-value	rho^2
constant	0.00	0.90	--
USEQVAL	0.26	0.00	24%
FXVAL	0.20	0.00	15%
DIVVAL	0.17	0.00	10%
USWML	-0.09	0.28	25%
EURWML	-0.27	0.00	25%
Model Adj R^2	--	--	40%

Volatility (arbitrage)

factor	coeff	p-value	rho^2
constant	-0.01	0.76	--
USEQDIV	-0.14	0.00	3%
FXTRD	0.08	0.02	6%
EQTRD	0.13	0.03	16%
VIX	-0.42	0.00	24%
USVAR	-0.18	0.30	2%
Model Adj R^2	--	--	32%

Volatility (short)

factor	coeff	p-value	rho^2
constant	-0.02	0.51	--
EUREQ	0.05	0.43	28%
EMEQ	0.04	0.48	24%
EMBDS	0.11	0.04	21%
VIX	-0.38	0.00	34%
USVAR	-0.62	0.00	25%
Model Adj R^2	--	--	47%

Equity (stock-based)

factor	coeff	p-value	rho^2
constant	0.00	0.93	--
DIVRP	0.16	0.01	17%
USEQDIV	0.21	0.00	23%
USSMB	0.27	0.00	13%
SIZE	-0.32	0.00	17%
SIZENL	0.33	0.00	3%
Model Adj R^2	--	--	44%

Equity (index-based)

factor	coeff	p-value	rho^2
constant	0.01	0.67	--
EQTRD	0.17	0.00	23%
NAEQ	0.04	0.59	32%
EUREQ	0.04	0.56	32%
PACEQ	0.06	0.11	22%
EMEQ	0.09	0.01	25%
VIX	-0.46	0.00	54%
Model Adj R^2	--	--	54%

Commodity

factor	coeff	p-value	rho^2
constant	-0.03	0.32	--
COMCARR	0.66	0.00	45%
DIVCARR	0.10	0.01	24%
COMTRD	0.09	0.06	8%
OILEXPL	-0.13	0.05	7%
ENCOMM	-0.04	0.56	9%
USVAR	-0.35	0.00	0%
Model Adj R^2	--	--	54%

Credit

factor	coeff	p-value	rho^2
constant	0.01	0.86	--
BETA	0.13	0.12	31%
NAEQ	-0.16	0.09	37%
EUREQ	0.21	0.00	45%
EMEQ	0.08	0.13	36%
USHYBD	0.33	0.00	35%
VIX	-0.29	0.00	33%
Model Adj R^2	--	--	51%

Currency

factor	coeff	p-value	rho^2
constant	0.00	0.86	--
MACRP	0.08	0.10	10%
FXCARR	0.51	0.00	41%
DIVCARR	0.05	0.20	13%
FXTRD	0.45	0.00	26%
EUREQ	0.04	0.52	20%
VIX	-0.10	0.01	20%
Model Adj R^2	--	--	67%

Multi-Asset

factor	coeff	p-value	rho^2
constant	0.00	0.97	--
DIVRP6	0.23	0.00	24%
DIVTRD	0.53	0.00	32%
EMBDS	0.51	0.00	24%
Model Adj R^2	--	--	66%

Rates

factor	coeff	p-value	rho^2
constant	0.00	0.96	--
BDCARR	0.18	0.00	22%
BDTRD	0.34	0.00	47%
GOVEXUS	0.19	0.00	47%
GOVUS	0.11	0.26	49%
ILBUS	0.05	0.26	32%
USCORBD	0.17	0.01	33%
Model Adj R^2	--	--	64%

Table 10 *continued*

ARP Benchmark Factor Sensitivity

Using weekly data between December 2008 and August 2020, Panel A uses elastic net to select a maximum of six factors from the universe in Appendix B to explain the returns of the categorical style and asset benchmarks. p-values reflect Newey-West (1987) corrected standard errors and ρ^2 represents the stand-alone coefficient of determination. The blue, green and red shading of p-values indicates significance respectively at the 1%, 5% and 10% levels. The shading for model adjusted R^2 illustrates simple quartiles, with blue indicating 75-100% explanatory power, gray 50-75%, orange 25-50% and brown 0-25%. Panel B repeats the exercise for the statistical broad benchmarks.

Panel B

Crude Oil Volatility			
factor	coeff	p-value	rho^2
constant	0.00	0.98	--
NAEQ	0.21	0.01	19%
EUREQ	0.00	0.98	17%
PACEQ	0.05	0.44	14%
EMEQ	0.10	0.09	17%
USHYBD	0.10	0.34	18%
USVAR	-0.38	0.01	8%
Model Adj R^2	--	--	28%

Volatility Sensitive			
factor	coeff	p-value	rho^2
constant	-0.01	0.86	--
DIVCARR	0.17	0.00	8%
EUREQ	0.09	0.13	19%
EMEQ	0.01	0.87	19%
EMBDS	0.13	0.06	22%
VIX	-0.22	0.00	18%
USVAR	-0.35	0.00	20%
Model Adj R^2	--	--	28%

FX Carry			
factor	coeff	p-value	rho^2
constant	0.00	0.90	--
FXCARR	0.86	0.00	90%
DIVCARR	0.03	0.06	21%
EUREQ	0.00	0.89	38%
PACEQ	0.06	0.01	27%
EMEQ	-0.01	0.83	45%
USDAUD	-0.07	0.02	39%
Model Adj R^2	--	--	90%

Equity Sensitive			
factor	coeff	p-value	rho^2
constant	-0.02	0.32	--
EQTRD	0.30	0.00	36%
NAEQ	-0.01	0.79	36%
EUREQ	0.09	0.00	42%
EMEQ	0.05	0.03	29%
VIX	-0.53	0.00	68%
USVAR	-0.52	0.00	17%
Model Adj R^2	--	--	83%

Rates Carry			
factor	coeff	p-value	rho^2
constant	0.00	1.00	--
BDCARR	0.26	0.00	32%
BDTRD	0.35	0.00	52%
GOVEXUS	0.18	0.00	52%
GOVUS	0.20	0.01	59%
ILBUS	0.04	0.27	37%
USCORBD	0.09	0.07	34%
Model Adj R^2	--	--	76%

FX Trend			
factor	coeff	p-value	rho^2
(Intercept)	0.00	0.95	--
FXTRD	0.45	0.00	63%
DIVTRD	0.47	0.00	69%
ILBUS	0.20	0.00	5%
Model Adj R^2	--	--	80%

Stocks, Value Light			
factor	coeff	p-value	rho^2
(Intercept)	-0.01	0.83	--
USEQLR	0.40	0.00	26%
EURHML	-0.01	0.89	23%
EURRMW	0.25	0.00	25%
EURWML	0.18	0.00	33%
MOM	0.25	0.00	33%
SIZE	-0.27	0.00	16%
Model Adj R^2	--	--	64%

Commodity Curve			
factor	coeff	p-value	rho^2
(Intercept)	-0.01	0.61	--
COMCARR	0.58	0.00	43%
DIVCARR	0.04	0.25	17%
EURWML	-0.13	0.00	0%
ENCOMM	-0.32	0.00	22%
Model Adj R^2	--	--	55%

Commodity Spread & Trend			
factor	coeff	p-value	rho^2
(Intercept)	-0.02	0.52	--
COMCARR	0.57	0.00	39%
DIVCARR	0.05	0.22	17%
COMTRD	0.45	0.00	33%
ENERGY	-0.09	0.02	3%
USVAR	-0.25	0.00	0%
Model Adj R^2	--	--	62%

Value Oriented			
factor	coeff	p-value	rho^2
(Intercept)	0.01	0.78	--
USEQVAL	0.20	0.00	17%
FXVAL	0.22	0.00	12%
USWML	-0.20	0.00	22%
EURWML	-0.07	0.25	17%
JPNWML	-0.16	0.00	17%
USVAR	0.42	0.01	7%
Model Adj R^2	--	--	37%

Table 11 ***Maximum Correlation between Primitive Strategy and ARP Benchmark***

Using weekly data between December 2008 and August 2020, Panel A shows the maximum correlation (ρ) between the 13 relatively focused Bloomberg GSAM indices and the 152 categorical benchmarks. Volatility (σ) is scaled to 7% for each component of the annual tracking error (TE) calculation. TE/σ contextualizes tracking error using underlying volatility. This ratio generally falls in the 60-100% range, indicating material tracking error. Panel B repeats the exercise with the 155 statistical benchmarks.

Panel A

Bloomberg GSAM Index	Categorical Benchmark	ρ	TE	TE/σ
US Equity Momentum Long-Short	Trend (cross-sectional momentum) Equity (stock-based) North America	0.8	4.7	71%
US Equity Value Long-Short	Value Equity (stock-based) North America	0.8	3.8	57%
US Equity Low Risk Long-Short	Risk Anomaly (quality, low volatility/beta) Equity (stock-based) North America	0.7	5.4	81%
US Equity Quality Long-Short	Trend (cross-sectional momentum) Equity (stock-based) North America	0.3	8.2	122%
FX Carry	Carry (spread) Currency Multi-Region	0.9	2.2	31%
Bond Futures Carry	Carry (spread) Rates Multi-Region	0.6	5.5	93%
Commodity Carry	Carry (curve) Commodity Multi-Region	0.7	5.4	74%
FX G10 Value	Value Currency Multi-Region	0.7	5.6	83%
Bond Futures Value	Value Rates Multi-Region	0.4	7.5	118%
FX Trend	Trend (time-series momentum) Currency	0.9	2.4	36%
Bond Futures Trend	Trend (time-series momentum) Rates Multi-Region	0.9	3.2	46%
Equity Trend	Trend (time-series momentum) Equity (index-based) Multi-Region	0.9	3.8	56%
Commodity Trend	Trend (time-series momentum) Commodity Multi-Region	0.8	4.0	57%

Panel B

Bloomberg GSAM Index	Statistical Benchmark	ρ	TE	TE/σ
US Equity Momentum Long-Short	Stocks Trend (C/S) N. America	0.8	4.8	73%
US Equity Value Long-Short	Stocks Value	0.8	4.3	64%
US Equity Low Risk Long-Short	Stocks Risk Anomaly N. America Approach 2	0.9	3.2	48%
US Equity Quality Long-Short	Stocks Multi-Style Approach 2	0.3	7.9	117%
FX Carry	FX Carry	0.9	2.2	31%
Bond Futures Carry	Rates Carry (spread) Approach 1	0.7	5.0	84%
Commodity Carry	Commodity Carry (curve) Approach 3	0.7	5.6	77%
FX G10 Value	FX Value Approach 2	0.9	3.0	45%
Bond Futures Value	Rates Value	0.2	8.6	134%
FX Trend	FX Trend (T/S)	0.9	2.9	42%
Bond Futures Trend	Rates Trend (T/S)	0.9	2.6	37%
Equity Trend	Equity Trend (T/S)	0.9	3.8	56%
Commodity Trend	Commodity Trend (T/S) Approach 1	0.9	3.7	53%

Table 12 ***Tracking of Primitive Strategy and Statistical ARP Benchmark in Traditional Index Terms***
Using weekly data between December 2008 and August 2020 and the list of market indices in Appendix B, this table contextualizes the tracking error between the Bloomberg GSAM index and the best fit statistical ARP benchmark.

Bloomberg GSAM Index	Statistical Benchmark	TE/σ Equivalent Pair in Traditional Index Space	
US Equity Momentum Long-Short	Stocks Trend (C/S) N. America	MSCI Pacific Equity	MSCI North America Equity
US Equity Value Long-Short	Stocks Value	MSCI Pacific Equity	MSCI Europe Equity
US Equity Low Risk Long-Short	Stocks Risk Anomaly N. America Approach 2	MSCI Europe Equity	MSCI North America Equity
US Equity Quality Long-Short	Stocks Multi-Style Approach 2	J.P. Morgan EM Bond	Bloomberg Precious Metals Commodity
FX Carry	FX Carry	MSCI Europe Equity	MSCI North America Equity
Bond Futures Carry	Rates Carry (spread) Approach 1	MSCI Europe Equity	ICE BofA US High Yield Bond
Commodity Carry	Commodity Carry (curve) Approach 3	ICE BofA US Corporate Bond	J.P. Morgan EM Bond
FX G10 Value	FX Value Approach 2	MSCI Europe Equity	MSCI North America Equity
Bond Futures Value	Rates Value	ICE BofA US Corporate Bond	MSCI Pacific Equity
FX Trend	FX Trend (T/S)	MSCI Europe Equity	MSCI North America Equity
Bond Futures Trend	Rates Trend (T/S)	MSCI Europe Equity	MSCI North America Equity
Equity Trend	Equity Trend (T/S)	MSCI Europe Equity	MSCI North America Equity
Commodity Trend	Commodity Trend (T/S) Approach 1	MSCI Europe Equity	MSCI North America Equity

Table 13 ***ARP Benchmark Performance Summary (period one)***

This table presents the annual Sharpe ratio using weekly net returns between December 2017 and December 2020 for the top three tiers of the statistical ARP benchmark taxonomy (broad, hypo-broad and super-base). Multi-colored shading separates the 10 broad benchmarks. Red font highlights the preponderance of negative Sharpe ratios for the recent three-year period. 80% of broad, 75% of hypo-broad and 70% of super-base Sharpe ratios are negative, with the median Sharpe ratio for each benchmark tier between -0.3 and -0.5.

Broad	Sharpe	Hypo-Broad	Sharpe	Super-Base	Sharpe
Stocks, Value Light	-1.1	Stocks Trend (C/S) Plus	-1.0	Stocks Multi-Style N. America	-1.7
				Stocks Trend (C/S)	0.0
		Stocks Multi-Style Plus	-0.9	Stocks Multi-Style	-1.7
				Stocks Multi-Style Europe	0.3
		Stocks Risk Anomaly	-0.8	Stocks Multi-Style Asia-Pac	-0.9
				Stocks Risk Anomaly 1	-0.7
				Stocks Risk Anomaly 2	-0.3
Commodity Spread Carry & Trend	-0.7	Commodity Spread Carry & Trend	-0.7	Commodity Trend	-0.9
				Commodity Carry (spread) 2	-0.9
				Commodity Carry (spread) 1	-0.5
				Commodity Trend (T/S)	0.0
Volatility Sensitive	-0.7	Commodity Volatility (short) Grains	-0.9	Commodity Volatility (short) Grains	-0.9
		Rates Volatility (short) Plus	-0.6	Rates Volatility (short) Plus	-0.6
		Commodity Volatility (short) Natural Gas	-0.6	Commodity Volatility (short) Natural Gas	-0.6
		Multi-Asset Volatility Sensitive Plus	-0.5	Commodity Volatility (short) Gold	-0.7
				Multi-Asset Volatility Sensitive	-0.4
				Commodity Volatility (short) Industrial Metals	-0.3
		FX Volatility (short) Plus	-0.1	FX Volatility (short) Plus	-0.1
Value Oriented	-0.6	Stocks Value	-1.3	Stocks Value 1	-1.8
				Stocks Value 2	-0.8
		Equity Reversal Plus	0.3	Equity Reversal	-0.2
				Equity Trend (T/S) N. America Dynamic	0.4
		Multi-Asset Reversal Oriented Plus	0.5	Equity Volatility (arbitrage) N. America	-0.1
				Commodity Reversal	0.2
				Multi-Asset Reversal Oriented	0.2
				FX Value	0.7
Equity Sensitive	-0.6	Equity Trend	-0.6	Equity/Credit Trend (T/S)	-0.9
				Equity Multi-Style	-0.3
		Equity Volatility (short)	-0.5	Equity Volatility (short) N. America	-0.5
				Equity Volatility (short) Europe	-0.4
Crude Oil Volatility (short)	-0.3	Crude Oil Volatility (short)	-0.3	Crude Oil Volatility (short)	-0.3
FX/Multi-Asset Trend	-0.1	FX Trend (T/S)	-0.3	FX Trend (T/S)	-0.3
		Multi-Asset Trend (T/S)	0.1	Multi-Asset Trend (T/S)	0.1
FX Carry	0.0	FX Carry	0.0	FX Carry (spread) Developed Mkt Focus	-0.1
				FX Carry (spread) EM Focus	0.0
Commodity Curve Carry	0.5	Commodity Curve Carry	0.5	Commodity Carry (curve) 2	0.4
				Commodity Carry (curve) 1	0.7
Rates Carry	0.8	Rates Carry	0.8	Rates Carry (spread)	0.3
				Rates Trend (T/S)	0.7
				Rates Carry (curve)	0.7

Table 14 ***ARP Benchmark Performance Summary (period two)***

This table presents the annual Sharpe ratio using weekly net returns between December 1999 and December 2017 for the top three tiers of the statistical ARP benchmark taxonomy (broad, hypo-broad and super-base). Multi-colored shading separates the 10 broad benchmarks. All Sharpe ratios are positive for the 18 years preceding the recent three-year period. The median Sharpe ratio for the broad, hypo-broad and super-base benchmark tiers are respectively 1.2, 1.0 and 0.9.

Broad	Sharpe	Hypo-Broad	Sharpe	Super-Base	Sharpe
Stocks, Value Light	1.0	Stocks Trend (C/S) Plus	0.6	Stocks Multi-Style N. America	0.8
				Stocks Trend (C/S)	0.4
		Stocks Multi-Style Plus	1.0	Stocks Multi-Style	0.8
				Stocks Multi-Style Europe	0.9
		Stocks Risk Anomaly	0.8	Stocks Multi-Style Asia-Pac	0.4
				Stocks Risk Anomaly 1	0.4
				Stocks Risk Anomaly 2	1.0
Commodity Spread Carry & Trend	1.5	Commodity Spread Carry & Trend	1.5	Commodity Trend	0.7
				Commodity Carry (spread) 2	1.4
				Commodity Carry (spread) 1	1.5
				Commodity Trend (T/S)	0.9
Volatility Sensitive	1.3	Commodity Volatility (short) Grains	0.3	Commodity Volatility (short) Grains	0.3
		Rates Volatility (short) Plus	0.8	Rates Volatility (short) Plus	0.8
		Commodity Volatility (short) Natural Gas	0.5	Commodity Volatility (short) Natural Gas	0.5
		Multi-Asset Volatility Sensitive Plus	1.4	Commodity Volatility (short) Gold	0.5
				Multi-Asset Volatility Sensitive	1.7
				Commodity Volatility (short) Industrial Metals	0.6
		FX Volatility (short) Plus	0.5	FX Volatility (short) Plus	0.5
Value Oriented	2.8	Stocks Value	1.0	Stocks Value 1	0.9
				Stocks Value 2	0.8
		Equity Reversal Plus	1.4	Equity Reversal	1.1
				Equity Trend (T/S) N. America Dynamic	1.1
		Multi-Asset Reversal Oriented Plus	2.4	Equity Volatility (arbitrage) N. America	0.7
				Commodity Reversal	0.9
				Multi-Asset Reversal Oriented	2.5
				FX Value	0.6
Equity Sensitive	1.4	Equity Trend	1.1	Equity/Credit Trend (T/S)	0.6
				Equity Multi-Style	1.2
		Equity Volatility (short)	1.1	Equity Volatility (short) N. America	1.2
				Equity Volatility (short) Europe	0.9
Crude Oil Volatility (short)	0.7	Crude Oil Volatility (short)	0.7	Crude Oil Volatility (short)	0.7
FX/Multi-Asset Trend	1.0	FX Trend (T/S)	0.5	FX Trend (T/S)	0.5
		Multi-Asset Trend (T/S)	1.3	Multi-Asset Trend (T/S)	1.3
FX Carry	0.8	FX Carry	0.8	FX Carry (spread) Developed Mkt Focus	0.5
				FX Carry (spread) EM Focus	1.0
Commodity Curve Carry	2.3	Commodity Curve Carry	2.3	Commodity Carry (curve) 2	2.6
				Commodity Carry (curve) 1	1.7
Rates Carry	1.1	Rates Carry	1.1	Rates Carry (spread)	1.0
				Rates Trend (T/S)	0.7
				Rates Carry (curve)	0.9

Table 15 ***Sharpe Ratio Adjustment Frameworks***

This table summarizes the inputs to the haircut Sharpe ratio (HSR) and deflated Sharpe ratio (DSR) calculations. Bold, blue font indicates parameters requiring assumptions for tradable bank indices, and pink shading indicates variation in parameterization.

	Observed Sharpe Ratio	Sharpe Ratio Variance	Return Frequency	Number of Trials	Size of Data Sample	Correlation among Trials	Return Skewness	Return Kurtosis
HSR	yes	**no**	yes	**yes**	yes	**yes**	no	no
DSR	yes	**yes**	yes	**yes**	yes	**no**	yes	yes

Table 16 ***Framing 2018-2020 ARP Sharpe ratios***

This table evaluates the 2018-2020 Sharpe ratio for 10 broad ARP benchmarks against two expectations, the 2000-2017 Sharpe ratio adjusted for multiple testing and the Sharpe ratio hurdle consistent with a 0.6 Sharpe ratio for a diversified ARP portfolio. p-values reflect Mertens (2002) standard errors and red, orange or blue shading indicates a significant departure from expectation respectively at the 1, 5 or 10% level.

	Crude Oil Volatility (short)	Volatility Sensitive	FX Carry	Equity Sensitive	Rates Carry	FX/Multi-Asset Trend	Stocks, Value Light	Commodity Curve Carry	Commodity Spread Carry & Trend	Value Oriented
2000-2017 Sharpe ratio	0.7	1.3	0.8	1.4	1.1	1.0	1.0	2.3	1.5	2.8
Adjustment	*-70%*	*-46%*	*-67%*	*-46%*	*-53%*	*-55%*	*-55%*	*-36%*	*-42%*	*-31%*
Adjusted Sharpe ratio	0.2	0.7	0.3	0.7	0.5	0.5	0.5	1.5	0.9	2.0
2018-2020 Sharpe ratio	-0.3	-0.7	0.0	-0.6	0.8	-0.1	-1.1	0.5	-0.7	-0.6
Difference	*-0.6*	*-1.4*	*-0.3*	*-1.3*	*0.3*	*-0.6*	*-1.6*	*-0.9*	*-1.6*	*-2.5*
p-value	29%	**0%**	63%	**1%**	58%	32%	**0%**	**9%**	**0%**	**0%**
Sharpe ratio Hurdle	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
Adjustment	*-60%*	*-78%*	*-62%*	*-78%*	*-72%*	*-71%*	*-71%*	*-87%*	*-80%*	*-89%*
Difference	*-0.6*	*-1.0*	*-0.3*	*-0.9*	*0.5*	*-0.4*	*-1.4*	*0.2*	*-1.0*	*-0.9*
p-value	24%	**4%**	58%	**9%**	38%	48%	**0%**	68%	**7%**	13%

Table 17 ***ARP Benchmark Return Normality***

This table uses weekly returns to provide estimates of skewness and (excess) kurtosis for the 10 broad ARP benchmarks over the 2000-2017 and 2018-2020 periods (939 and 157 data points). p-values reflect bootstrapped standard errors and red, orange or blue shading indicates a significant departure from zero respectively at the 1, 5 or 10% level.

	Crude Oil Volatility (short)	Volatility Sensitive	FX Carry	Equity Sensitive	Rates Carry	FX/Multi-Asset Trend	Stocks, Value Light	Commodity Curve Carry	Commodity Spread Carry & Trend	Value Oriented
Skewness										
2000-2017	-1.5	-0.8	-0.9	-1.5	-0.7	-0.3	-0.6	0.1	0.0	1.6
p-value	0%	5%	0%	0%	2%	18%	2%	57%	99%	0%
2018-2020	-2.3	-3.0	-1.0	-2.7	-0.4	-0.9	-1.8	1.3	-0.3	0.6
p-value	0%	0%	5%	0%	37%	3%	0%	24%	44%	6%
Difference	-0.8	-2.2	-0.1	-1.2	0.3	-0.6	-1.2	1.2	-0.3	-1.0
p-value	26%	3%	79%	6%	38%	16%	1%	29%	43%	0%
Kurtosis										
2000-2017	6.5	6.5	3.0	6.6	4.0	2.3	2.9	2.4	2.0	9.3
p-value	0%	1%	0%	0%	2%	0%	2%	0%	7%	0%
2018-2020	11.2	17.8	3.9	12.3	2.6	3.2	6.5	11.1	2.4	1.9
p-value	0%	0%	5%	1%	0%	1%	1%	2%	0%	3%
Difference	4.6	11.3	0.9	5.7	-1.4	0.9	3.6	8.6	0.4	-7.4
p-value	26%	6%	68%	22%	9%	49%	16%	8%	69%	0%

Table 18 ***Traditional Benchmark Return Normality***

This table uses weekly returns to provide estimates of skewness and (excess) kurtosis for 15 traditional benchmarks (equities, government bonds, commodities and credit) over the 2000-2017 and 2018-2020 periods (939 and 157 data points). p-values reflect bootstrapped standard errors and red, orange or blue shading indicates a significant departure from zero respectively at the 1, 5 or 10% level.

	NA Stocks	Eur Stocks	Pac Stocks	EM Stocks	Non-US Govt	US Govt	US Linker	Non-US Linker	Energy	Ind Metals	Prec Metals	Agriculture	EM Bonds	High Yld Bds	Corp Bds
Skewness															
2000-2017	-0.6	-0.8	-0.5	-0.4	-0.2	-0.4	-0.3	-0.3	-0.3	-0.3	-0.5	-0.2	-1.7	-2.0	-0.9
p-value	20%	13%	13%	42%	7%	0%	28%	26%	3%	6%	2%	29%	24%	5%	4%
2018-2020	-0.9	-1.5	-0.5	-0.8	-0.7	0.2	-2.3	-0.2	-0.8	-0.2	0.2	0.1	-3.8	-2.8	-3.3
p-value	25%	17%	70%	9%	1%	74%	26%	75%	1%	50%	81%	69%	3%	13%	14%
Difference	-0.3	-0.8	0.1	-0.4	-0.5	0.6	-2.0	0.0	-0.6	0.1	0.6	0.3	-2.2	-0.8	-2.5
p-value	70%	50%	96%	37%	6%	27%	33%	97%	7%	72%	38%	5%	21%	66%	28%
Kurtosis															
2000-2017	6.5	7.3	3.7	6.1	0.7	1.2	3.3	2.4	1.1	2.0	2.2	2.2	29.5	22.5	5.4
p-value	1%	2%	9%	0%	1%	1%	1%	0%	0%	0%	0%	0%	0%	1%	4%
2018-2020	6.6	11.1	10.3	3.1	1.4	3.7	23.7	5.0	2.1	1.1	5.3	-0.4	27.1	24.2	29.1
p-value	0%	2%	0%	1%	4%	0%	2%	0%	3%	18%	0%	3%	1%	1%	1%
Difference	0.1	3.8	6.6	-3.0	0.8	2.5	20.4	2.6	0.9	-0.9	3.1	-2.7	-2.4	1.7	23.7
p-value	96%	43%	5%	2%	25%	2%	4%	9%	33%	27%	7%	0%	82%	84%	2%

Table 19 ***ARP Benchmark Correlation Change Summary***

This exhibit provides a tabular summary of the correlation comparison (2000-2017 versus 2018-2020) in Figure 19. The first data column evaluates correlations among ARP benchmarks in the two northwest quadrants. The middle column focuses upon the traditional benchmarks in the southeast quadrants. The final column considers the correlation between ARP and traditional benchmarks in the remaining two quadrants. A bootstrapped standard error and 10% alpha (in the interest of conservatism) determine the significance of each pairwise correlation difference.

	Among ARP Broad Benchmarks		Among Traditional Benchmarks		Across ARP and Traditional Benchmarks	
Average Correlation Change	0.00		0.15		0.09	
Average Absolute Correlation Change	0.06		0.12		0.08	
	#	%	#	%	#	%
Increases	22	49%	82	78%	107	71%
Decreases	23	51%	23	22%	43	29%
Significant Increases	3	7%	36	34%	36	24%
Significant Decreases	4	9%	4	4%	6	4%
Significant Changes	7	16%	40	38%	42	28%
Insignificant Changes	38	84%	65	62%	108	72%
Total Correlations	45	100%	105	100%	150	100%

Table 20 ***Simple ARP Portfolio Risk Profile***

This exhibit uses the weekly returns from 2000 to 2017 to build a portfolio with an equal volatility contribution from nine broad ARP benchmarks (excluding Crude Oil Volatility, the most niche benchmark). The portfolio targets a volatility of 8% and carries a correlation of 0.1 with the US equity market. The realized lines show the results for the 2018-2020 window.

	Volatility Sensitive	FX Carry	Equity Sensitive	Rates Carry	FX/Multi-Asset Trend	Stocks, Value Light	Commodity Curve Carry	Commodity Spread Carry & Trend	Value Oriented	Portfolio
Contribution to Portfolio Volatility (%)										
Expected	11%	11%	11%	11%	11%	11%	11%	11%	11%	8.0
Realized	15%	7%	23%	2%	8%	12%	14%	13%	7%	8.6
Difference	4%	-4%	12%	-9%	-3%	0%	3%	2%	-5%	0.6
Contribution to Portfolio Correlation w/ NA Stocks										
Expected	0.06	0.12	0.13	-0.08	-0.02	-0.08	-0.01	-0.04	0.00	0.1
Realized	0.14	0.08	0.22	0.00	0.02	0.04	0.02	0.02	0.02	0.6
Difference	0.09	-0.04	0.09	0.09	0.04	0.12	0.03	0.05	0.02	0.5

Table 21 *ARP Benchmark Weekly Return Independence*

This table provides the autocorrelation between time t and t-1 weekly returns for the 10 broad ARP benchmarks for the 2000-2017 and 2018-2020 periods (939 and 157 data points). Red, orange or blue shading of p-value indicates a significant departure from zero respectively at the 1, 5 or 10% level.

	Crude Oil Volatility (short)	Volatility Sensitive	FX Carry	Equity Sensitive	Rates Carry	FX/Multi-Asset Trend	Stocks, Value Light	Commodity Curve Carry	Commodity Spread Carry & Trend	Value Oriented
2000-2017										
Autocorr (lag 1)	0.01	0.08	-0.03	-0.02	-0.02	-0.01	0.06	-0.01	-0.06	0.08
p-value	69%	2%	39%	63%	53%	76%	8%	79%	6%	2%
2018-2020										
Autocorr (lag 1)	-0.11	0.21	0.10	0.08	0.00	-0.18	-0.06	-0.11	-0.04	0.07
p-value	15%	1%	20%	33%	95%	2%	44%	19%	58%	38%
Difference										
Autocorr (lag 1)	-0.13	0.13	0.13	0.09	0.02	-0.17	-0.12	-0.10	0.02	-0.01
p-value	11%	10%	11%	24%	85%	3%	14%	22%	83%	93%

Table 22 ***ARP Benchmark Weekly Return Independence***

This table provides the autocorrelation between time t and t-1 weekly returns for 15 traditional benchmarks (equities, government bonds, commodities and credit) for the 2000-2017 and 2018-2020 periods (939 and 157 data points). Red, orange or blue shading of p-value indicates a significant departure from zero respectively at the 1, 5 or 10% level.

	NA Stocks	Eur Stocks	Pac Stocks	EM Stocks	Non-US Govt	US Govt	US Linker	Non-US Linker	Energy	Ind Metals	Prec Metals	Agriculture	EM Bonds	High Yld Bds	Corp Bds
2000-2017															
Autocorr (lag 1)	-0.07	-0.04	-0.04	0.02	0.00	-0.05	-0.04	-0.08	0.02	-0.02	-0.01	0.02	-0.01	0.40	0.03
p-value	3%	17%	23%	56%	99%	11%	24%	1%	60%	59%	86%	56%	65%	0%	33%
2018-2020															
Autocorr (lag 1)	-0.08	-0.07	-0.06	0.04	-0.05	-0.06	-0.08	-0.14	0.06	0.03	-0.26	0.00	0.11	0.09	0.09
p-value	29%	35%	48%	59%	52%	46%	34%	7%	46%	68%	0%	99%	15%	26%	26%
Difference															
Autocorr (lag 1)	-0.01	-0.03	-0.02	0.02	-0.05	-0.01	-0.04	-0.06	0.04	0.05	-0.25	-0.02	0.13	-0.31	0.06
p-value	87%	72%	83%	76%	52%	94%	64%	43%	60%	53%	0%	82%	11%	0%	46%

Table 23 *ARP State-Based Conditional Mean Spreads*

Using weekly returns between January 2000 and December 2020 and 24 market environment indicators, the table provides the conditional weekly mean spread between the two opposing states. For example, the first ↓ — ↑ Equity Vol row subtracts the mean ARP benchmark return for periods of falling equity volatility within the 2000-2017 window from the average return for periods of rising volatility. The p-value reflects a bootstrapped standard error. To facilitate directional comparison across rows, the mean spread appears in green (red) font if positive (negative). Red, orange or blue shading of p-value indicates a significant mean difference respectively at the 1, 5 or 10% level. The three blank rows for the 2018-2020 period reflect no rising rate regimes -- the state indicator vector is empty.

	State Difference	Period	Crude Oil Volatility (short)	Volatility Sensitive	FX Carry	Equity Sensitive	Rates Carry	FX/Multi-Asset Trend	Stocks, Value Light	Commodity Curve Carry	Commodity Spread Carry & Trend	Value Oriented
Mean Spread	↓ — ↑	2000 to	0.187	0.411	0.443	0.294	-0.169	-0.084	-0.190	-0.015	-0.051	0.188
p-value	Equity Vol	2017	10%	0%	0%	0%	17%	49%	9%	87%	66%	15%
Mean Spread	↓ — ↑	2018 to	0.236	0.153	0.122	0.272	0.021	0.017	-0.069	0.021	-0.057	0.143
p-value	Equity Vol	2020	39%	53%	37%	33%	84%	93%	74%	94%	79%	49%
Mean Spread	↓ — ↑	2000 to	0.206	0.363	0.203	0.245	-0.053	0.006	-0.026	0.024	-0.010	0.004
p-value	Bond Vol	2017	1%	0%	5%	0%	62%	95%	77%	77%	91%	97%
Mean Spread	↓ — ↑	2018 to	-0.037	0.355	0.079	0.150	-0.426	-0.340	-0.068	0.272	0.006	0.301
p-value	Bond Vol	2020	93%	38%	73%	72%	5%	15%	81%	39%	98%	34%
Mean Spread	↓ — ↑ FX	2000 to	0.246	0.453	0.324	0.320	-0.202	-0.209	-0.174	-0.045	0.060	0.024
p-value	Vol	2017	1%	0%	0%	0%	5%	3%	5%	56%	52%	82%
Mean Spread	↓ — ↑ FX	2018 to	0.372	0.475	0.345	0.714	-0.193	-0.038	-0.039	-0.222	-0.034	0.150
p-value	Vol	2020	54%	38%	21%	17%	42%	90%	91%	67%	94%	72%
Mean Spread	↑ — ↓ Fin	2000 to	0.187	0.250	0.349	0.301	-0.217	-0.043	-0.199	0.093	0.099	-0.100
p-value	Cond	2017	8%	6%	0%	0%	6%	71%	9%	33%	37%	47%
Mean Spread	↑ — ↓ Fin	2018 to	0.438	0.160	0.020	0.523	0.000	0.139	-0.009	-0.002	-0.061	0.134
p-value	Cond	2020	28%	66%	91%	13%	100%	55%	97%	100%	84%	64%
Mean Spread	↑ — ↓	2000 to	0.266	0.226	0.354	0.431	-0.329	0.005	-0.243	-0.043	-0.093	-0.300
p-value	Stock Prc	2017	0%	3%	0%	0%	0%	96%	2%	63%	32%	0%
Mean Spread	↑ — ↓	2018 to	0.595	0.468	0.114	0.724	-0.085	0.185	0.023	-0.055	-0.018	-0.144
p-value	Stock Prc	2020	11%	16%	55%	3%	56%	36%	93%	88%	95%	59%
Mean Spread	↑ — ↓ Oil	2000 to	0.264	0.182	0.290	0.150	-0.152	-0.081	-0.036	0.008	0.068	-0.037
p-value	Prc	2017	1%	10%	1%	7%	15%	43%	70%	94%	53%	75%
Mean Spread	↑ — ↓ Oil	2018 to	0.417	0.261	0.051	0.469	0.097	0.205	0.160	0.051	-0.006	-0.004
p-value	Prc	2020	29%	42%	78%	11%	51%	31%	54%	90%	98%	99%
Mean Spread	↓ — ↑	2000 to	-0.041	-0.074	-0.111	-0.088	-0.267	-0.261	-0.025	-0.193	-0.184	0.095
p-value	Gold Prc	2017	60%	40%	27%	29%	0%	1%	74%	2%	6%	29%
Mean Spread	↓ — ↑	2018 to	-0.202	-0.048	-0.214	0.034	-0.270	-0.055	0.140	-0.111	0.104	-0.177
p-value	Gold Prc	2020	51%	84%	36%	88%	10%	80%	56%	71%	74%	46%
Mean Spread	↓ — ↑	2000 to	0.215	0.204	0.324	0.372	-0.421	-0.141	-0.161	0.004	-0.016	-0.044
p-value	Credit Sprd	2017	3%	7%	0%	0%	0%	18%	11%	96%	88%	72%
Mean Spread	↓ — ↑	2018 to	0.602	0.580	0.029	0.774	-0.056	0.290	0.086	-0.038	0.079	-0.212
p-value	Credit Sprd	2020	14%	6%	88%	2%	68%	20%	78%	93%	81%	47%
Mean Spread	↑ — ↓	2000 to	0.261	0.068	0.311	0.334	-0.272	-0.063	-0.091	-0.028	-0.024	-0.092
p-value	Breakeven	2017	0%	46%	0%	0%	0%	44%	32%	73%	78%	35%
Mean Spread	↑ — ↓	2018 to	0.507	0.465	0.024	0.790	-0.108	0.301	0.001	0.013	0.082	-0.040
p-value	Breakeven	2020	26%	22%	91%	3%	48%	20%	100%	98%	81%	90%
Mean Spread	↑ — ↓	2000 to	0.020	0.014	0.039	0.006	-0.225	-0.112	0.084	0.247	0.062	-0.211
p-value	Fed Fds	2017	85%	91%	71%	95%	2%	30%	43%	3%	56%	9%
Mean Spread	↑ — ↓	2018 to										
p-value	Fed Fds	2020										
Mean Spread	↑ — ↓	2000 to	0.082	-0.022	0.198	0.248	-0.641	-0.233	-0.129	0.166	0.134	-0.231
p-value	Yields	2017	40%	86%	6%	1%	0%	2%	21%	10%	18%	5%
Mean Spread	↑ — ↓	2018 to										
p-value	Yields	2020										
Mean Spread	↑ — ↓	2000 to	-0.061	-0.140	-0.069	0.023	-0.575	-0.249	-0.160	-0.015	0.068	-0.003
p-value	Real Yields	2017	50%	16%	52%	78%	0%	2%	6%	86%	50%	97%
Mean Spread	↑ — ↓	2018 to										
p-value	Real Yields	2020										

Table 24 ***ARP Turbulence-Based Conditional Mean Spreads***

Using weekly returns between January 2000 and December 2020 and an indicator of market turbulence, the table provides the conditional weekly mean spread between the turbulent and non-turbulent observations within the 2000-2017 and 2018-2020 periods and between the turbulent observations across the two windows. The p-value reflects a bootstrapped standard error. To facilitate directional comparison across rows, the mean spread appears in green (red) font if positive (negative). Red, orange or blue shading of p-value indicates a significant mean difference respectively at the 1, 5 or 10% level.

	Crude Oil Volatility (short)	Volatility Sensitive	FX Carry	Equity Sensitive	Rates Carry	FX/Multi-Asset Trend	Stocks, Value Light	Commodity Curve Carry	Commodity Spread Carry & Trend	Value Oriented
Turbulent - Non-Turbulent State										
2000-2017 Mean Spread	-0.330	-0.665	-0.456	-0.430	-0.173	-0.418	-0.295	-0.030	-0.235	0.267
p-value	**0%**	**0%**	**0%**	**0%**	18%	**0%**	**0%**	72%	**2%**	**2%**
2018-2020 Mean Spread	-0.503	-0.941	-0.589	-1.697	-0.107	-0.797	-0.441	0.960	0.136	0.487
p-value	16%	**2%**	**1%**	**0%**	64%	**2%**	18%	**1%**	56%	14%
2018-2020 - 2000-2017 Turbulent State										
Mean Spread	-0.292	-0.525	-0.222	-1.321	-0.020	-0.462	-0.424	0.628	0.019	-0.254
p-value	40%	20%	37%	**1%**	94%	17%	20%	**7%**	94%	44%

Table 25 ***ARP Fund Performance Summary***

Using weekly excess returns between December 2017 and December 2020 across 22 diversified ARP funds in an EN regression on 85 base ARP and 21 investable reference benchmarks, the upper half of this table reveals the heterogeneity within this investment category via the median and interdecile fund correlation, adjusted R^2, intercept level, and Newey-West (1987) intercept p-value. The lower half of the table shows the rank correlation among Sharpe ratio, residual variance and intercept, highlighting the very strong relationship between Sharpe ratio and intercept.

Fund Percentile	Fund Correlation	Adj R^2	Intercept	p-value
10%	0.17	34%	-0.16%	2%
50%	0.37	54%	-0.04%	37%
90%	0.57	78%	0.04%	72%
Fund Rank Correlation	**Sharpe ratio**	**Residual Variance**	**Intercept**	
Sharpe	1.00			
Residual	-0.32	1.00		
Intercept	0.90	-0.05	1.00	

Table 26 ***Historical Context for US Cross-Sectional Value***

This exhibit provides summary data (Sharpe ratio, skewness, correlation among benchmarks, and correlation with the S&P 500) for three time periods – the 1955-1999 early history, 2000-2017 distributional baseline window, and 2018-2020 crisis. Five primitive benchmarks provide the historical context -- the first measuring the composite value score relative to the investable universe (Simple), the second adding size and market risk constraints to the first, the third utilizing the composite value score on a sector relative basis (Simple Sec Neut), and the fourth adding size and market risk constraints to the third, and the fifth being the Fama-French high-minus-low factor (FF HML). Gross returns for the Bloomberg GSAM US Equity Value L/S Index and North American value stocks statistical base composite benchmark are available only for the two most recent periods. Green (red) font indicates a desirable (undesirable) sign from an ARP portfolio construction perspective. Shading in the correlation matrix highlights greater (less) similarity in green (red).

——— 1955 to 1999 ———	Simple	Simple Constr	Simple Sec Neut	Simple Sec Neut Constr	FF HML	Bloomberg GSAM	Statistical Base Composite
Sharpe ratio	0.6	0.7	1.0	1.1	0.5	—	—
Skewness	1.0	0.7	0.7	0.8	0.2	—	—
Correlation	1.0	0.7	0.8	0.7	0.7	—	—
	0.7	1.0	0.9	1.0	0.7	—	—
	0.8	0.9	1.0	0.9	0.7	—	—
	0.7	1.0	0.9	1.0	0.6	—	—
	0.7	0.7	0.7	0.6	1.0	—	—
	—	—	—	—		—	—
	—	—	—	—		—	—
S&P 500 Corr	-0.4	-0.3	-0.4	-0.3	-0.5	—	—

——— 2000 to 2017 ———	Simple	Simple Constr	Simple Sec Neut	Simple Sec Neut Constr	FF HML	Bloomberg GSAM	Statistical Base Composite
Sharpe ratio	0.7	0.7	0.8	0.9	0.4	0.6	0.9
Skewness	1.3	1.0	1.6	1.0	0.6	0.5	1.0
Correlation	1.0	0.7	0.8	0.7	0.6	0.4	0.6
	0.7	1.0	0.9	0.9	0.7	0.6	0.7
	0.8	0.9	1.0	0.9	0.7	0.5	0.8
	0.7	0.9	0.9	1.0	0.6	0.5	0.7
	0.6	0.7	0.7	0.6	1.0	0.6	0.5
	0.4	0.6	0.5	0.5	0.6	1.0	0.5
	0.6	0.7	0.8	0.7	0.5	0.5	1.0
S&P 500 Corr	-0.2	0.1	-0.1	0.0	0.1	0.2	0.0

——— 2018 to 2020 ———	Simple	Simple Constr	Simple Sec Neut	Simple Sec Neut Constr	FF HML	Bloomberg GSAM	Statistical Base Composite
Sharpe ratio	-0.9	-1.2	-0.7	-1.1	-1.1	-1.0	-2.0
Skewness	0.7	0.5	1.4	0.5	0.3	0.4	0.3
Correlation	1.0	0.7	0.7	0.7	0.7	0.6	0.7
	0.7	1.0	1.0	1.0	0.9	0.8	0.9
	0.7	1.0	1.0	1.0	0.9	0.8	0.9
	0.7	1.0	1.0	1.0	0.9	0.8	0.9
	0.7	0.9	0.9	0.9	1.0	0.8	0.8
	0.6	0.8	0.8	0.8	0.8	1.0	0.8
	0.7	0.9	0.9	0.9	0.8	0.8	1.0
S&P 500 Corr	0.3	0.3	0.3	0.3	0.3	0.2	0.2

Table 27 ***Historical Context for Equity Time-Series Trend***

This exhibit provides summary data (Sharpe ratio, skewness, correlation among benchmarks, and correlation with the S&P 500) for three time periods – the 1955-1999 early history, 2000-2017 distributional baseline window, and 2018-2020 crisis. Two primitive benchmarks provide the historical context -- the first taking a long or short position based upon the sign of the trailing 260-day S&P 500 return (Simple) and the second adding a basic confirmation to the first in the form of a z-score threshold to avoid taking positions based upon a weak trend signal (Simple Confirm). Gross returns for the Bloomberg GSAM Equity Trend Index and equity trend (time-series) statistical base composite benchmark are available only for the two most recent periods. Green (red) font indicates a desirable (undesirable) sign from an ARP portfolio construction perspective. Shading in the correlation matrix highlights greater (less) similarity in green (red).

	1912 to 1999				2000 to 2017				2018 to 2020			
	Simple	Simple Confirm	Bloomberg GSAM	Statistical Base Composite	Simple	Simple Confirm	Bloomberg GSAM	Statistical Base Composite	Simple	Simple Confirm	Bloomberg GSAM	Statistical Base Composite
Sharpe ratio	0.2	0.3	—	—	0.2	0.3	0.5	0.6	-0.5	-0.1	-0.5	-0.8
Skewness	-1.1	-0.6	—	—	0.4	0.8	0.0	0.3	-0.9	1.4	-2.3	-3.0
Correlation	1.0	0.9	—	—	1.0	0.9	0.6	0.7	1.0	0.8	0.4	0.6
	0.9	1.0	—	—	0.9	1.0	0.7	0.7	0.8	1.0	0.6	0.6
	—	—	—	—	0.6	0.7	1.0	0.9	0.4	0.6	1.0	0.8
	—	—	—	—	0.7	0.7	0.9	1.0	0.6	0.6	0.8	1.0
S&P 500 Corr	-0.1	0.0	—	—	-0.1	-0.2	-0.1	-0.1	0.2	0.0	0.1	0.4

Table 28 ***Surprises in ARP Portfolio Construction Inputs***

This table summarizes the deviations from expectations for the primary ARP portfolio construction inputs over the 2018-2020 period. Expectations reflect data available at the end of 2017. Red shading indicates a result significantly and problematically different than the expectation. Yellow shading denotes a realization nominally and problematically different than the expectation. Green shading highlights an outcome consistent with or beneficially different than the expectation.

		Correlation		Tails		Conditional Returns	
ARP Broad Strategy	Sharpe Ratio	ARP Strategies	Traditional Beta	Skewness	Kurtosis	States	Turbulence
Crude Oil Volatility (short)							
Volatility Sensitive							
FX Carry							
Equity Sensitive							
Rates Carry							
FX/Multi-Asset Trend							
Stocks, Value Light							
Commodity Curve Carry							
Commodity Spread Carry & Trend							
Value Oriented							

Figures

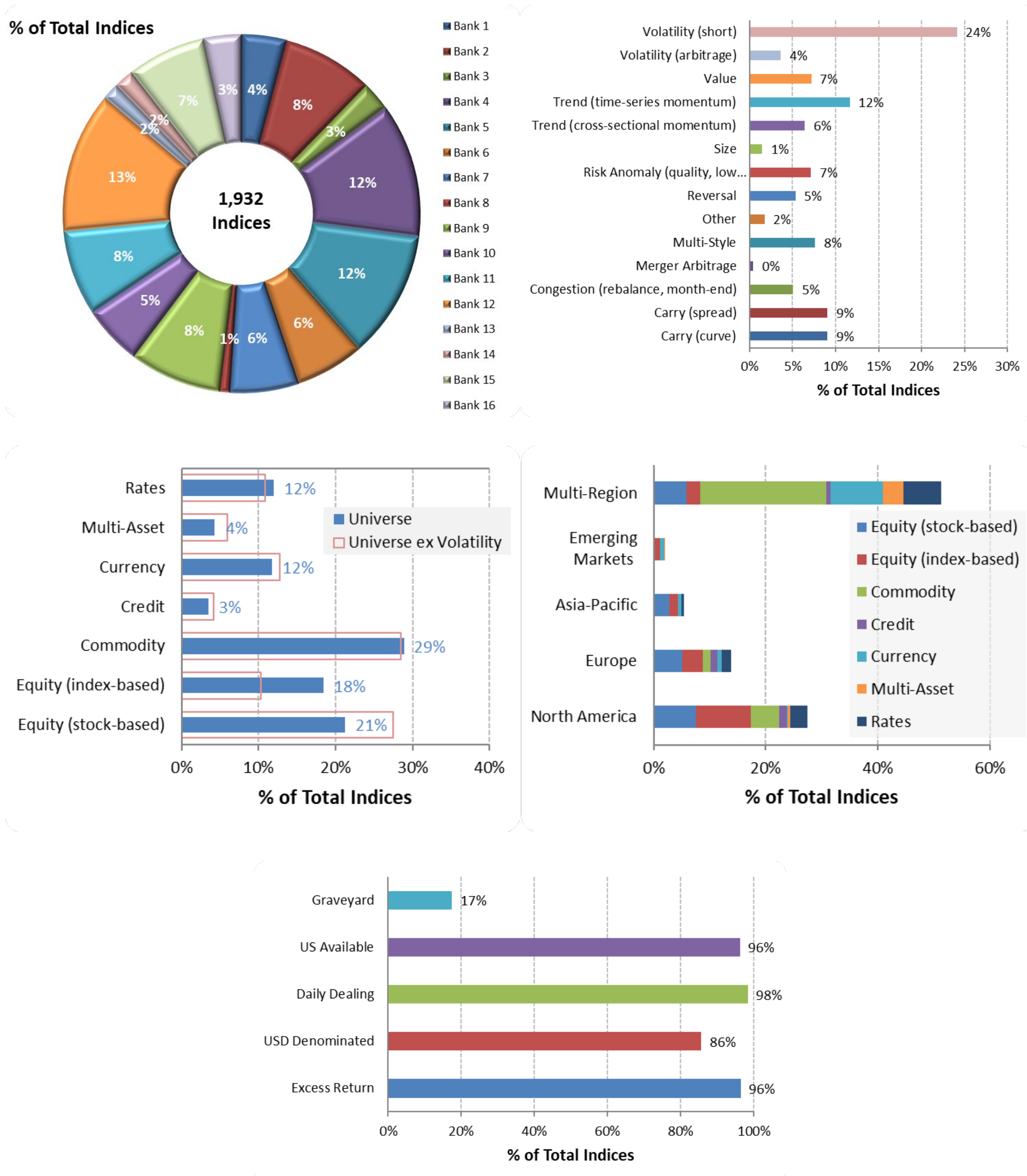


Figure 1 ***Tradable Index Database Characteristics Summary***

The top left panel provides the index distribution by financial institution, highlighting the wide range in strategy inventory across banks. The top right panel shows the index distribution by style. Short volatility strategies represent a relatively large proportion of the universe as banks regularly offer indices on individual assets. The middle left panel focuses upon asset class, displaying the universe skew toward commodity and equity strategies. The middle right panel reveals the broad regional footprint of most strategies. The bottom center panel shows that graveyard indices (no longer priced) account for 17% of the universe and that USD denominated excess return indices, available in the US with daily dealing, dominate the universe.

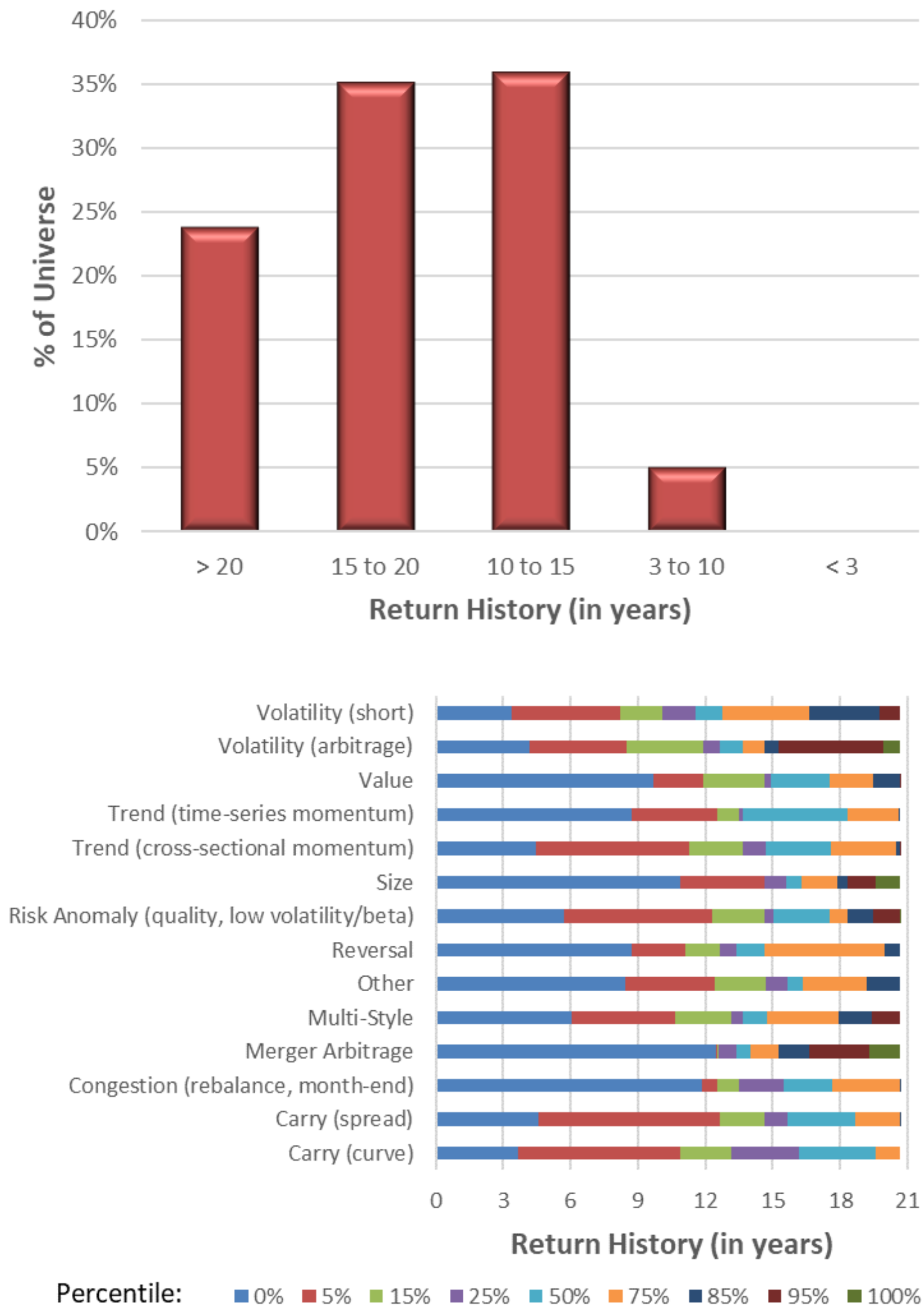


Figure 2 ***Tradable Index Return Availability***

The top panel indicates that most of the universe has more than 15 years of daily returns available as of August 2020. The data curation process eliminates indices with less than three years of history. The bottom panel provides the return history distribution by style, highlighting that the median availability is shortest for volatility, reversal, merger arb and multi-style strategies. The 0 and 100th percentile respectively represent the minimum and maximum performance history.

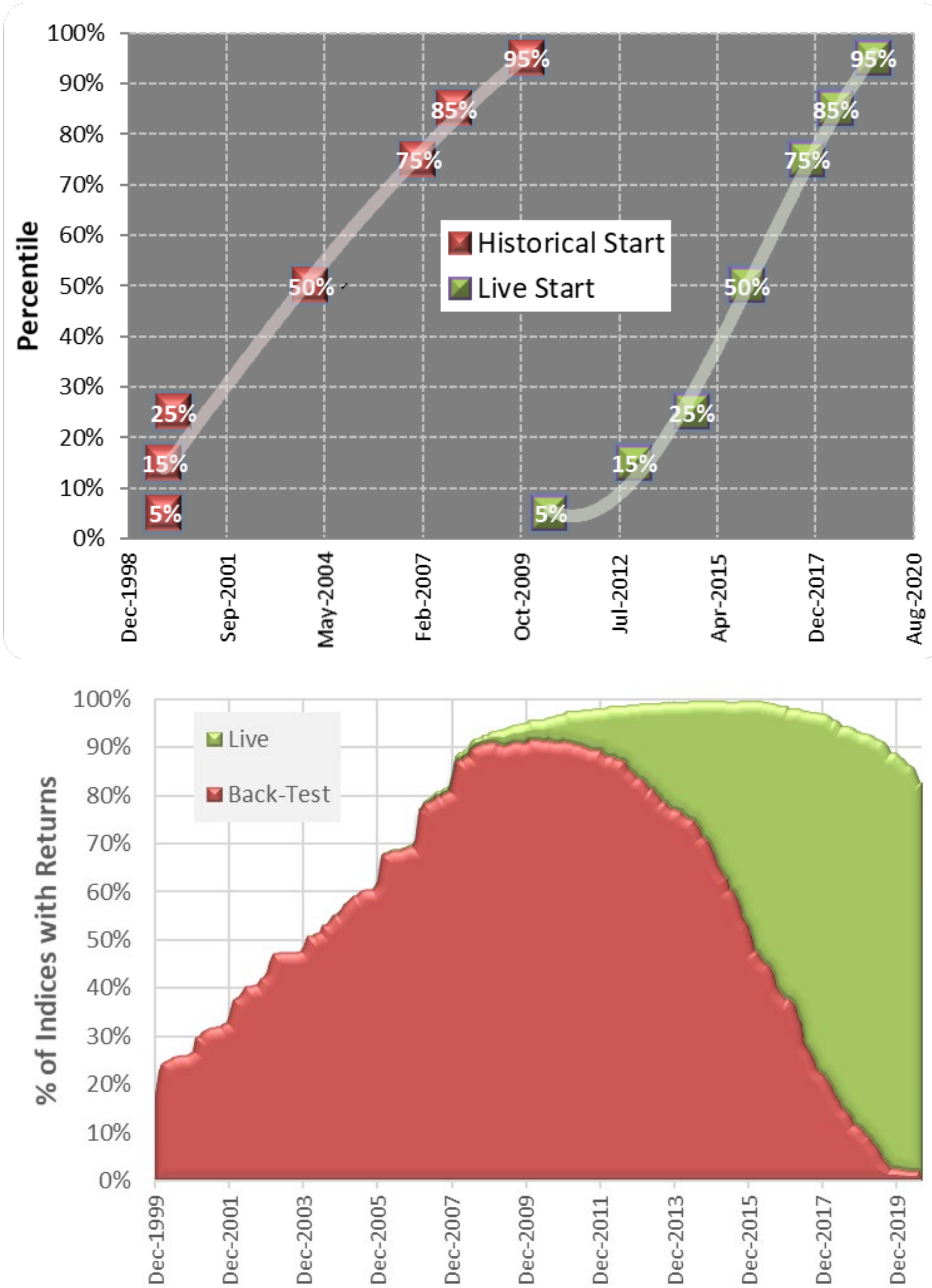
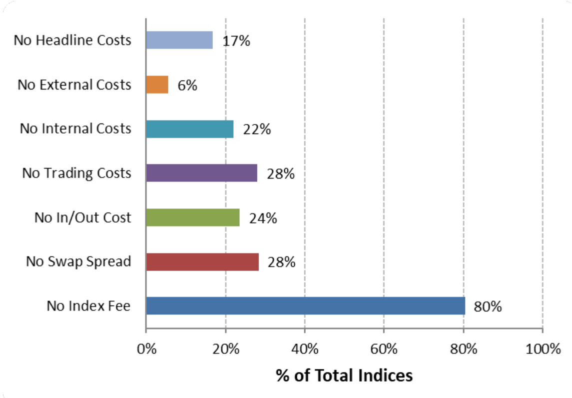
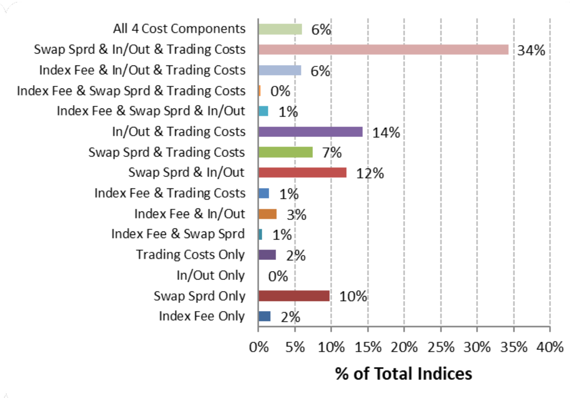


Figure 3 ***Duality of Tradable Index Return History***

A distinguishing feature of bank index returns is that the history represents a blend of live and back-tested returns. The top panel displays the percentile distribution of start dates for each return subset, with an accompanying third-order polynomial trend line. Given the dearth of earlier start dates, this paper truncates return histories at 12/31/1999, an adjustment affecting only a small fraction of the universe. Live history generally represents a small fraction of available returns, with the median years of live history being four, the median live proportion of total history being 26%, and 11 years generally separating the two lines. The bottom panel summarizes return availability at any point in time, with the peak being almost 100% in early 2016 and the decline since then representing the cumulative effect of graveyard indices.





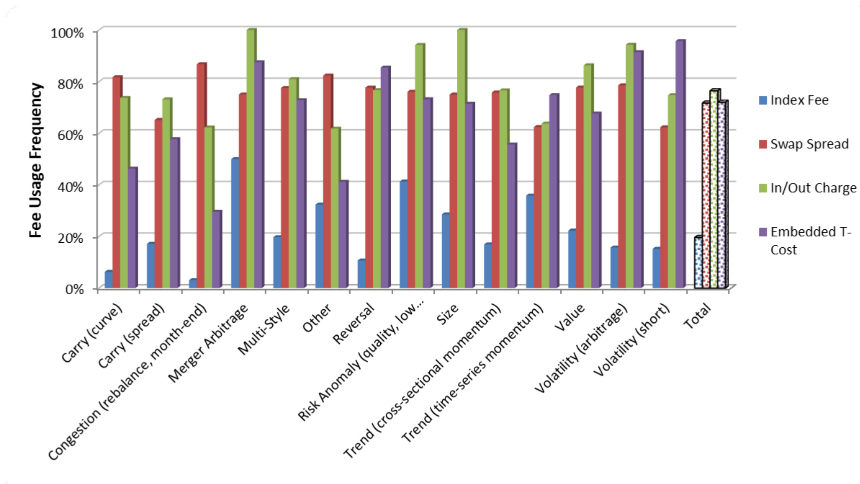
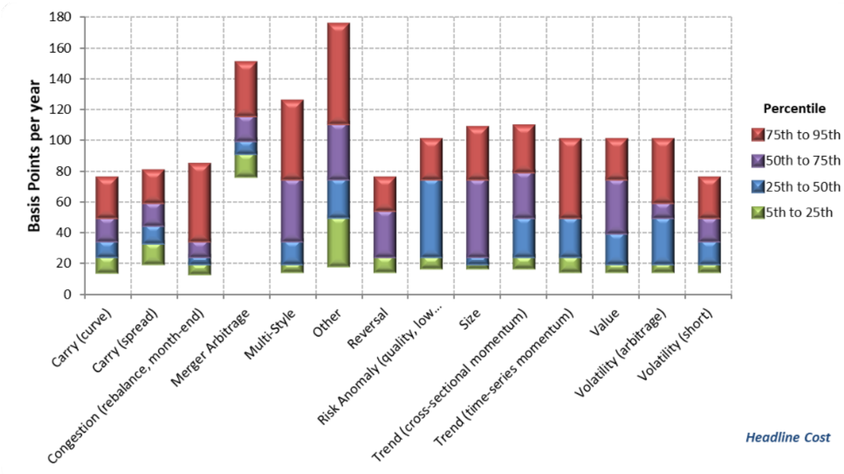


Figure 4 ***Tradable Index Cost Structure Variation***

Bank indices employ a wide range of cost structures, with an index fee and trading costs potentially embedded in the index return series and a swap spread and in/out charges possibly sitting on top. The top left panel indicates the extent to which banks do not use specific cost levers. Index fees, encountered regularly at one time, are no longer typical whereas external costs (swap spread and transaction charges) are very common. The top right panel summarizes the distribution of cost structures in the database. The most prevalent structure combines a swap spread with in/out charges and embedded trading costs. A variety of cost combinations appear with relatively similar frequency, highlighting an important consistency consideration when comparing index returns. The bottom panel shows modest variation in the prominence of cost levers across index styles. For example, volatility strategies tend to incorporate trading costs while both stock and volatility strategies make liberal use of in/out charges – a byproduct of execution realities in these spaces.



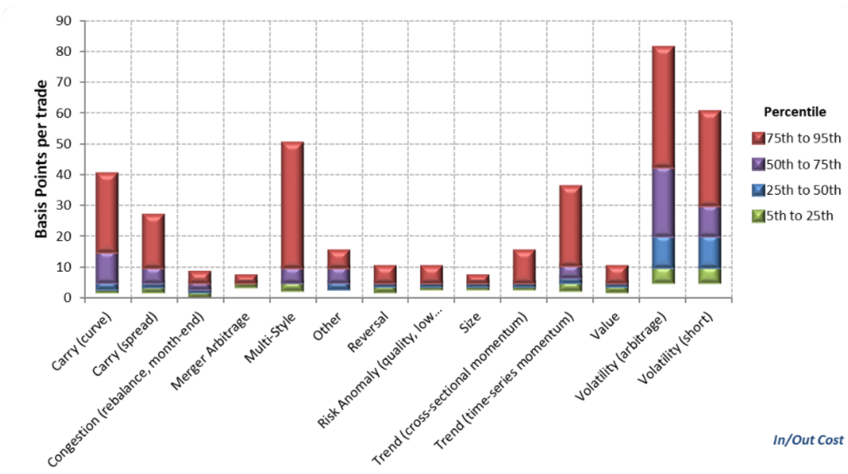


Figure 5 ***Tradable Index Cost Summary***

Costs vary greatly across index style. The five panels in this exhibit present headline cost, in/out cost, trading cost, total cost and total cost relative to index volatility. Each chart summarizes the distribution of cost by index style, focusing on the 5th to 95th percentile to limit potential distortion from outliers. These distributions exclude zero costs to capture the profile of a specific cost lever when in effect. Headline cost combines the index fee and swap spread. Total cost combines all external and internal costs, assuming a three-year, fixed-size investment for in/out cost. Because leverage impacts total cost and volatility provides a rough indication of leverage, volatility-adjusted total cost represents a standardized metric. Standardization narrows the gap between volatility and other strategies (panel five versus four), but the former clearly are the highest cost strategies due to significant execution costs.

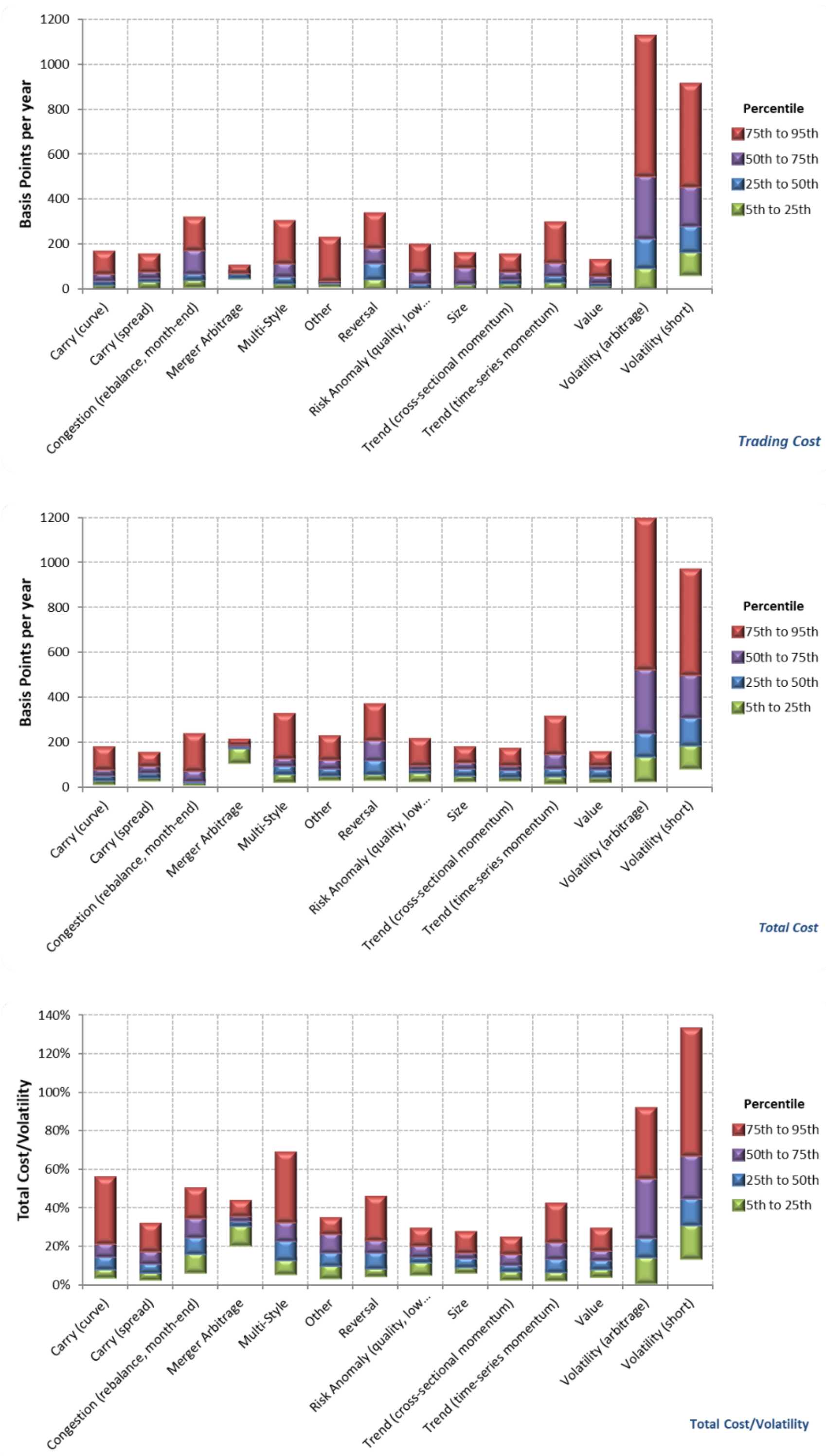


Figure 5 ***continued***
Tradable Index Cost Summary

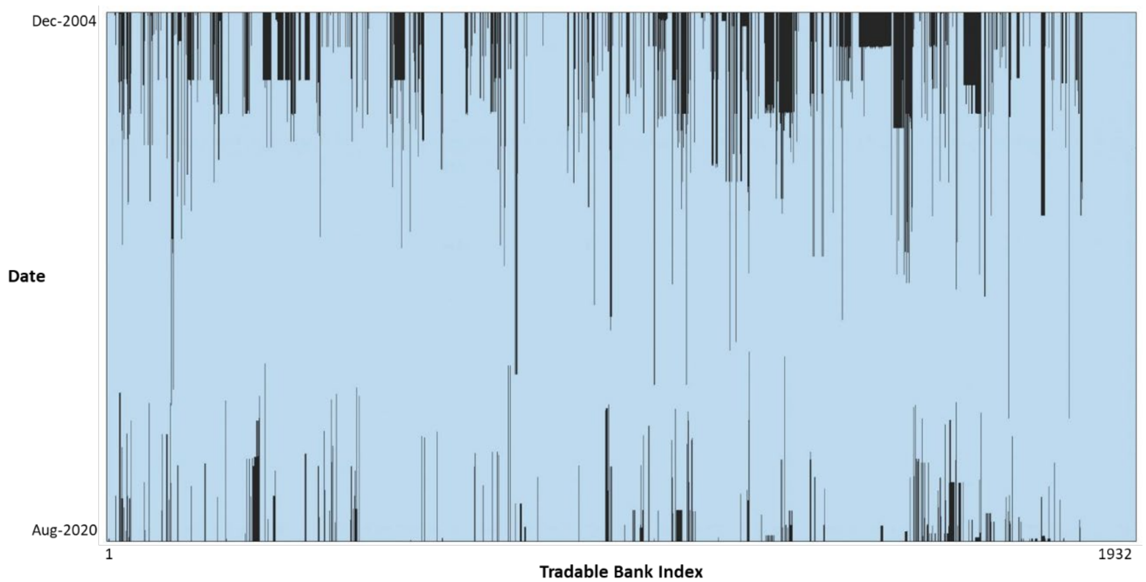


Figure 6 ***Tradable Index Missing Data***
With a December 2004 start date, missing data (dark shading) impacts about 50% of the indices, with a median 14% of return history missing for the affected indices. Missing data represents 10% of the overall return history -- 8% of which exists in early, pre-index inception years and 2% of which resides in recent, post-index termination years.

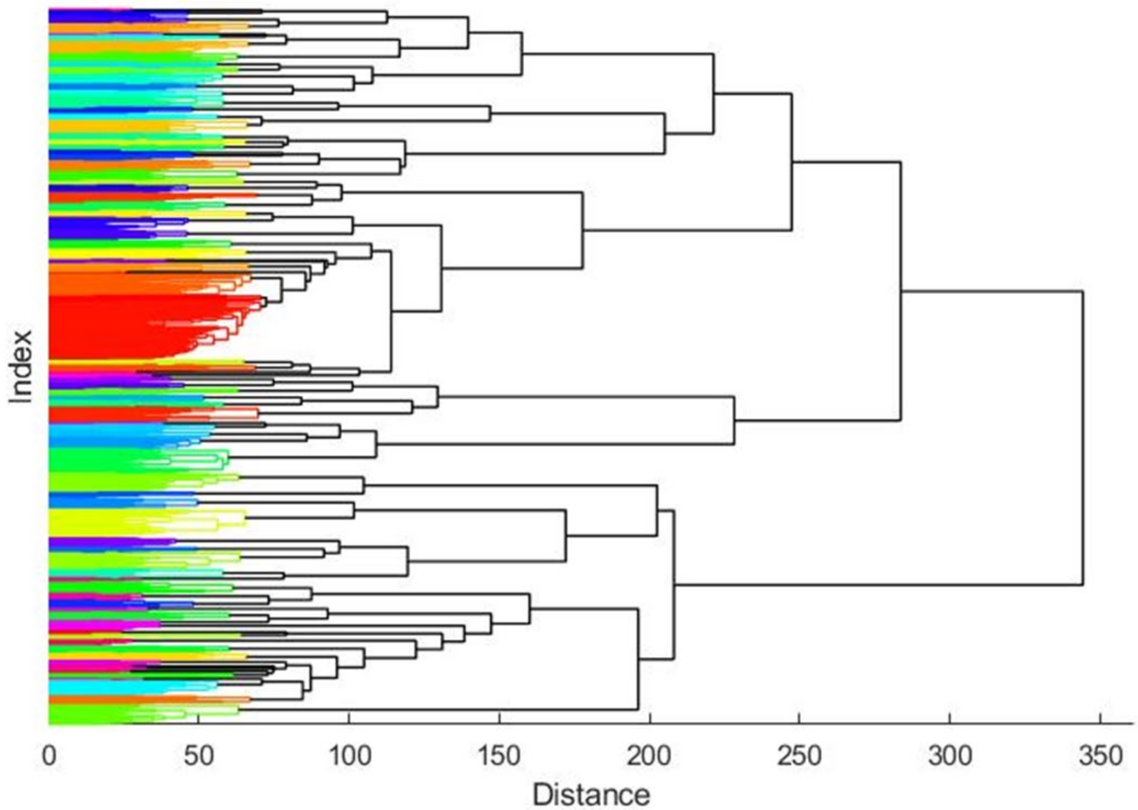


Figure 7 ***Hierarchical Clustering of Tradable Bank Indices***
This dendogram illustrates the grouping of the 1,932 indices in the proprietary database, using weekly returns between December 2004 and August 2020. Colors highlight 85 clusters, equal to the base number of categorical benchmarks.

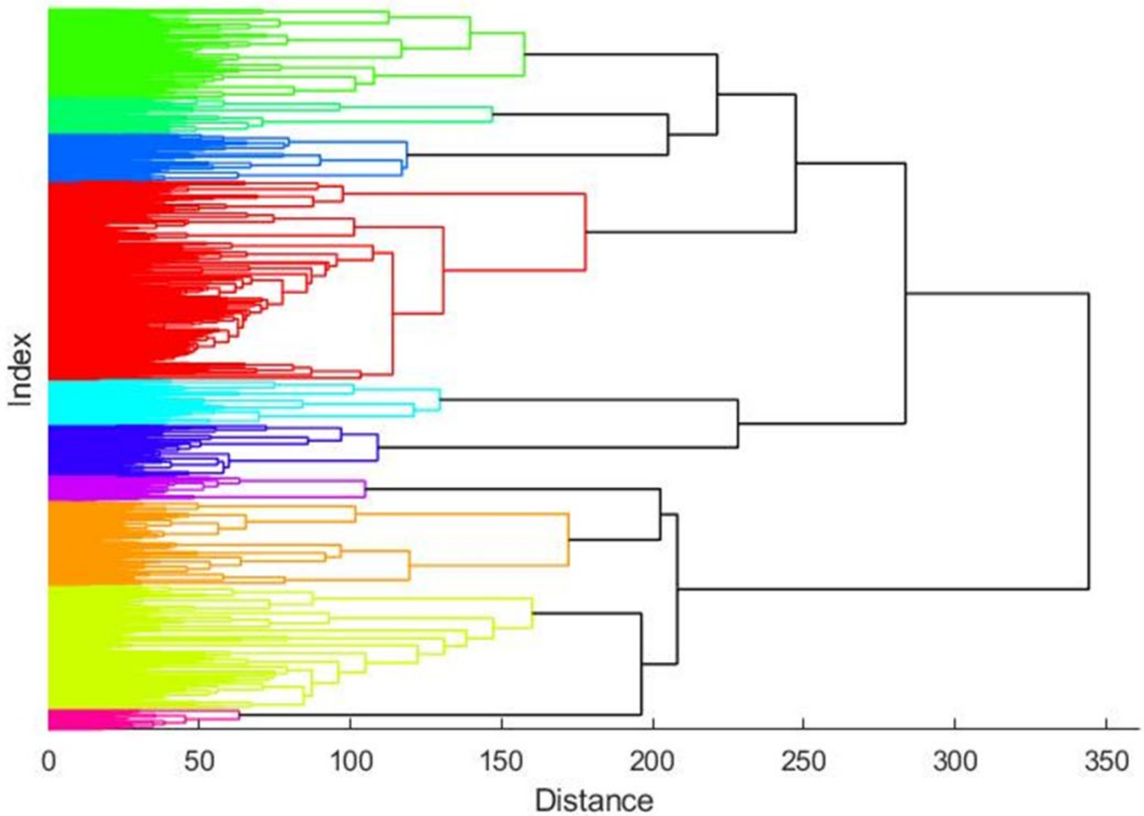


Figure 8 ***Hierarchical Clustering of Tradable Bank Indices***
This dendogram illustrates the grouping of the 1,932 indices in the proprietary database, using weekly returns between December 2004 and August 2020. Colors highlight 10 broad clusters and the greater distance between clusters relative to that of the narrower groupings in **Error! Reference source not found.**.

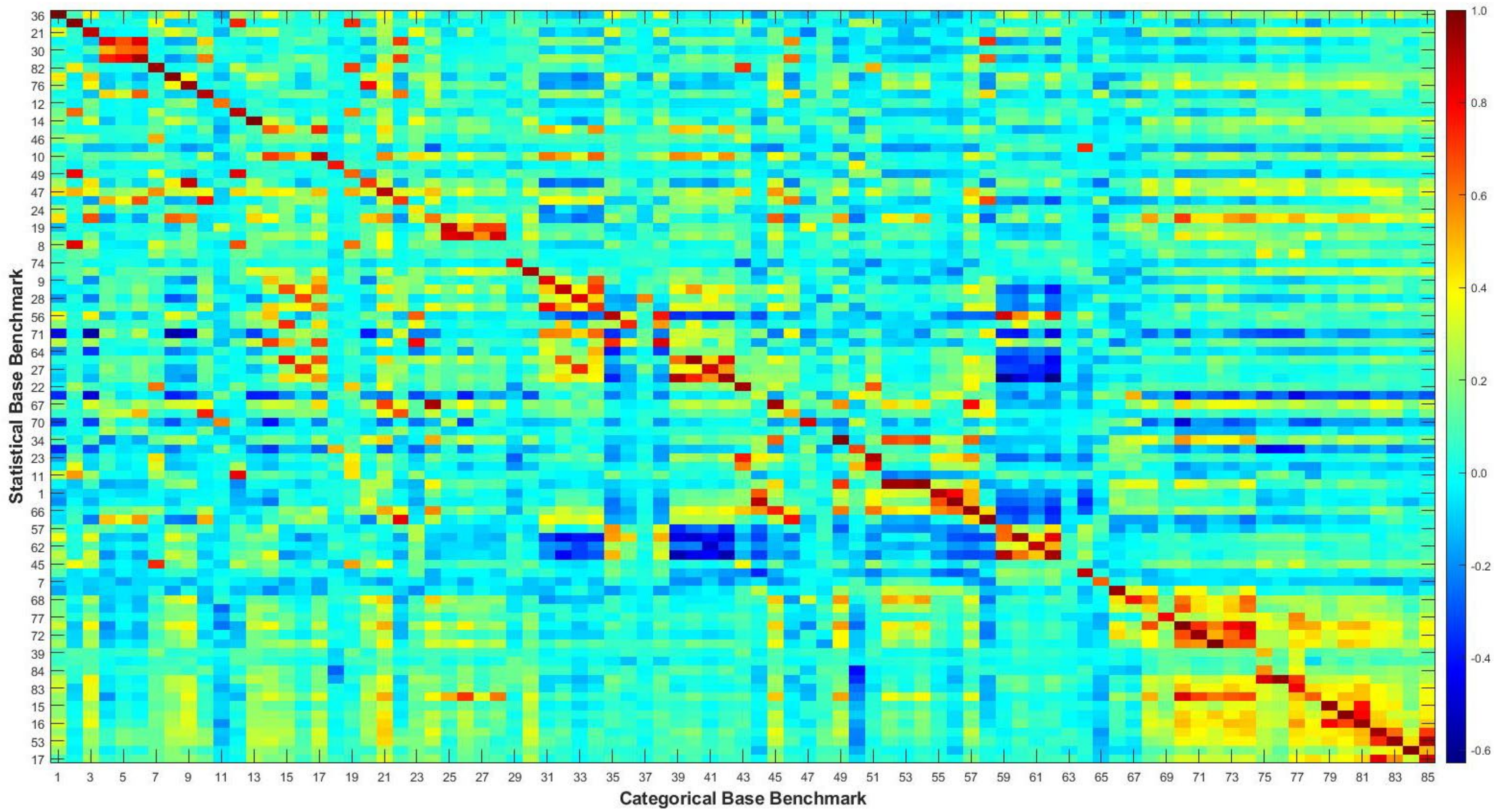


Figure 9 *Correlation of Categorical and Statistical Base Benchmarks*
This heatmap summarizes the correlation between the categorical and statistical base benchmark returns between December 2008 and August 2020. Dark red indicates the highest and dark blue the lowest correlations. An algorithm maximizing the correlations along the diagonal determines the order of the statistical benchmarks (vertical axis).

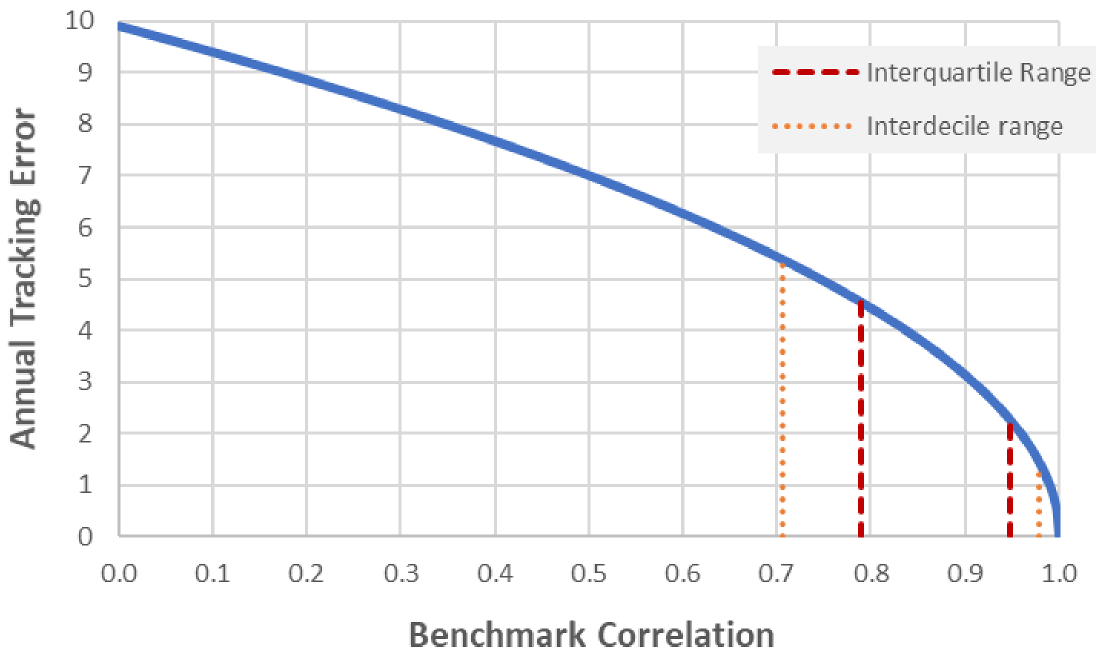


Figure 10 ***Correlation and Tracking Error between ARP Benchmarks***
This chart depicts the nonlinear relationship between tracking error and correlation. The red lines highlight the interquartile range for the maximum correlation between categorical and statistical base benchmarks over the December 2008 to August 2020 period. Even relatively high correlations translate into 2.3% to 4.5% annual tracking error given the 7% scaled volatility of the benchmarks. Tracking error for the interdecile range is 1.4% to 5.4%.

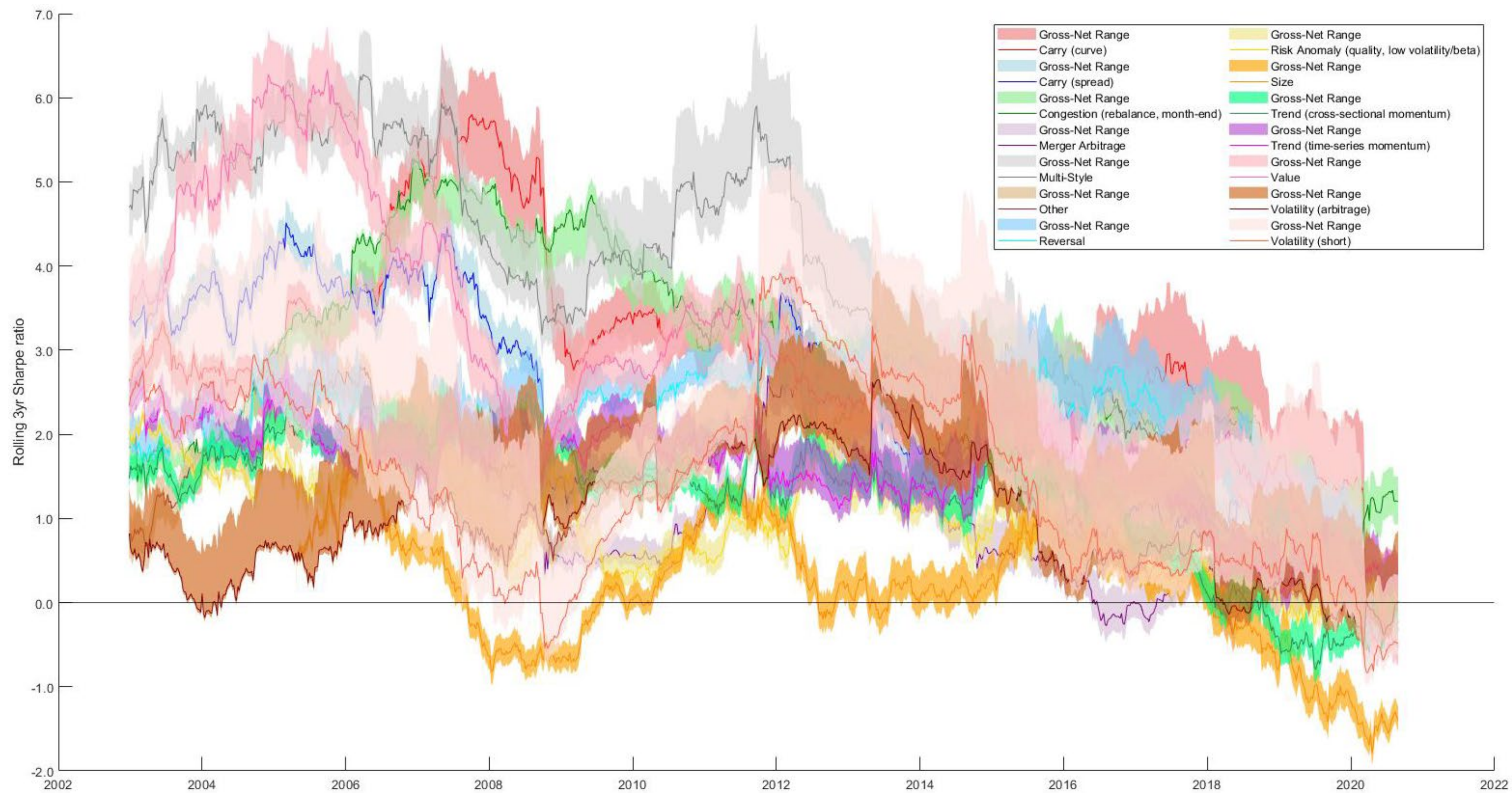


Figure 11 ***Sharpe Ratio Range for Categorical Style Benchmarks***

This chart shows the rolling three-year Sharpe ratio for the 14 categorical style benchmarks. The dark line in each range represents the reported returns. The shaded range denotes the gap between gross and net returns. A wider range indicates a combination of higher trading costs and/or fees. The realized experience of most investors resides between the dark line and the bottom of the range. The chart uses weekly returns between December 1999 and August 2020.

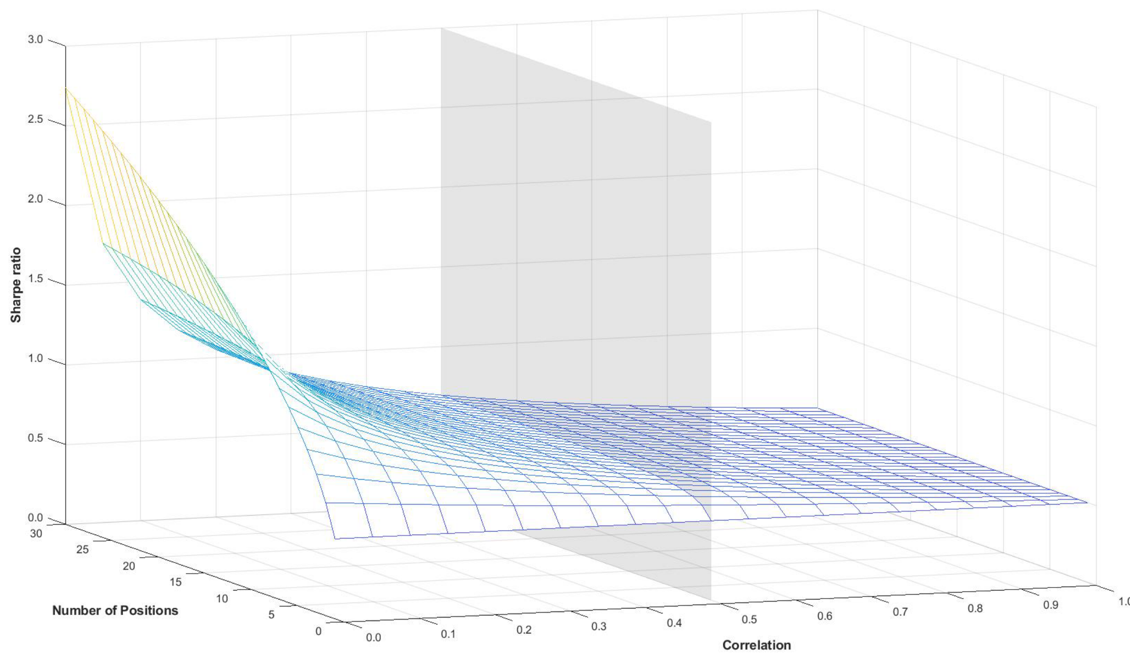


Figure 12 ***Sharpe Ratio Dynamic within ARP Benchmarks***

This chart shows the Sharpe ratio as a function of the number of benchmark constituents and the correlation among them. Per *Equation 18*, the Sharpe ratio (SR_{ij}) is a function of the weight ($\boldsymbol{w}$) and return ($\boldsymbol{r}$) vector for j positions and the covariance matrix ($\boldsymbol{\Sigma}$) for j positions with a constant correlation of i. $\boldsymbol{r}$ assumes that underlying strategies have a 7% volatility and a 0.5 Sharpe ratio. The gray plane separates the sub-0.5 correlation region within which the Sharpe ratio increases exponentially. This dynamic is driving the Sharpe ratio in many of the higher tier ARP benchmarks. Based upon data from December 2008 to August 2020, the median interquartile correlation among base benchmarks within each categorical style, categorical asset and statistical broad benchmark is 0.0 to 0.4, 0.0 to 0.4 and 0.2 to 0.6, respectively – well within the high impact region.

$$SR_{ij} = \frac{\boldsymbol{r}_j \boldsymbol{w}_j^T}{\sqrt{\boldsymbol{w}_j \boldsymbol{\Sigma}_{ij} \boldsymbol{w}_j^T}} \qquad \textit{Equation 34}$$

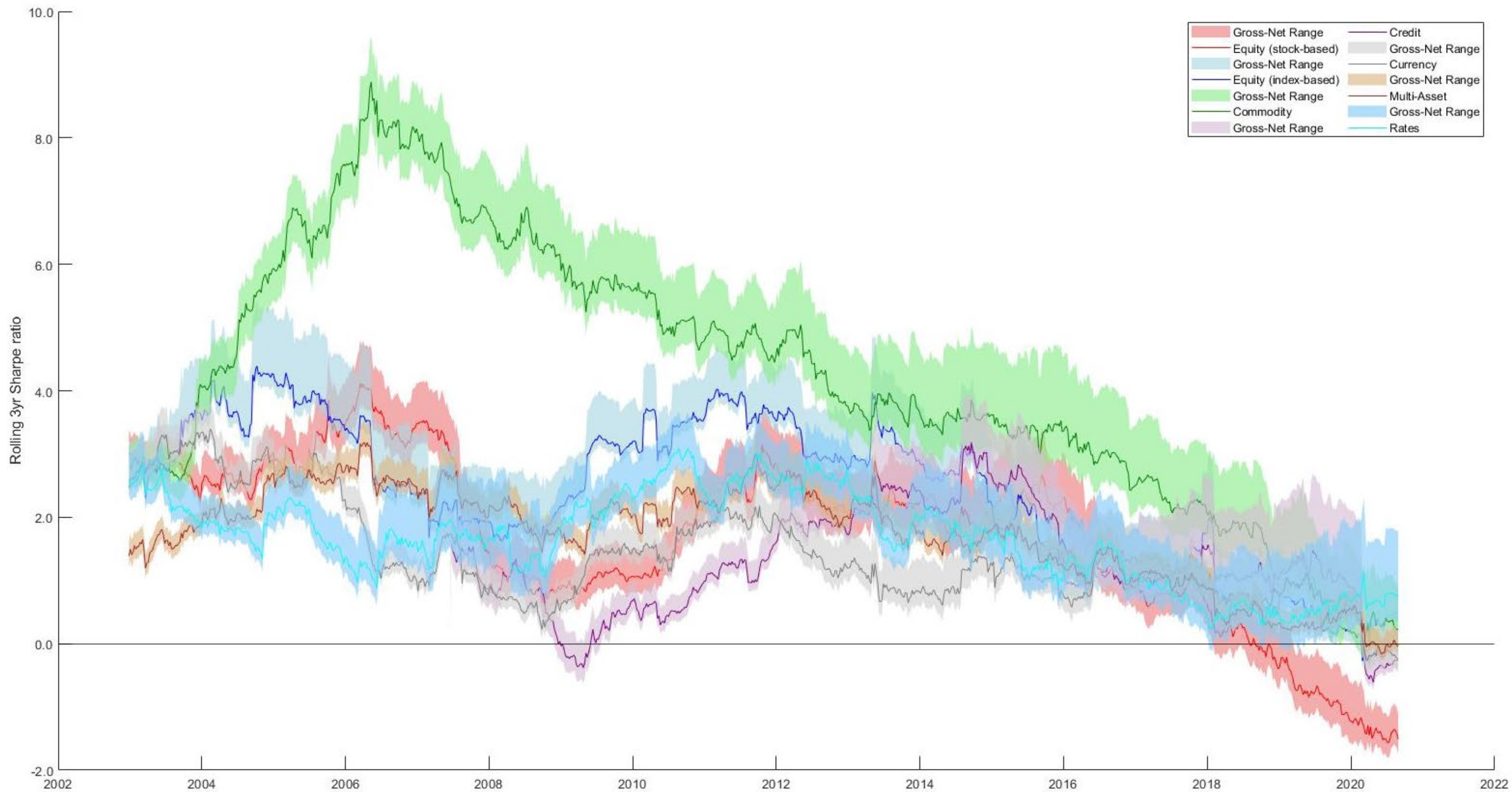


Figure 13 ***Sharpe Ratio Range for Categorical Asset Benchmarks***

This chart shows the rolling three-year Sharpe ratio for the seven categorical asset benchmarks. The dark line in each range represents the reported returns. The shaded range denotes the gap between gross and net returns. A wider range indicates a combination of higher trading costs and/or fees. The realized experience of most investors resides between the dark line and the bottom of the range. The chart uses weekly returns between December 1999 and August 2020.

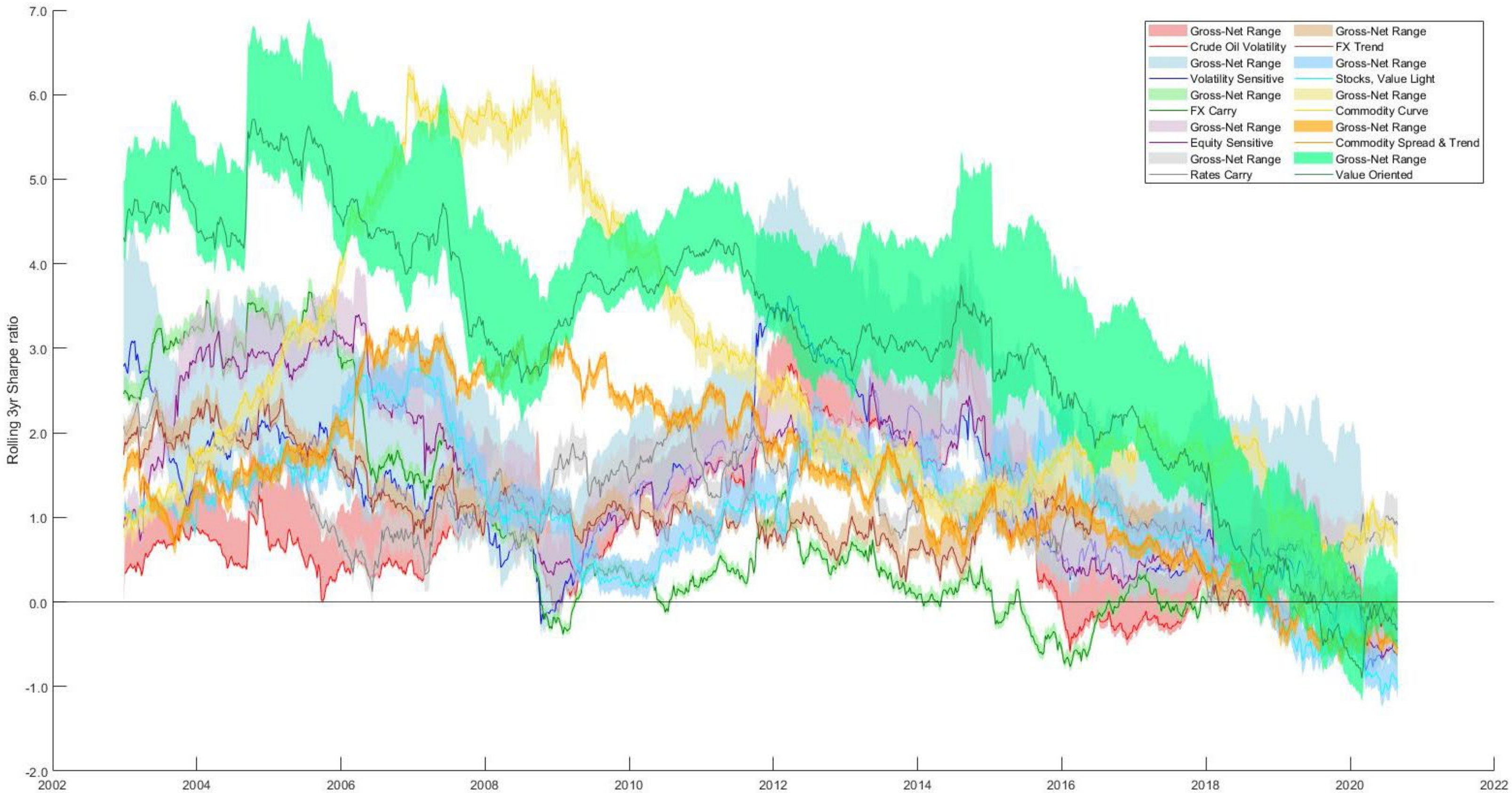
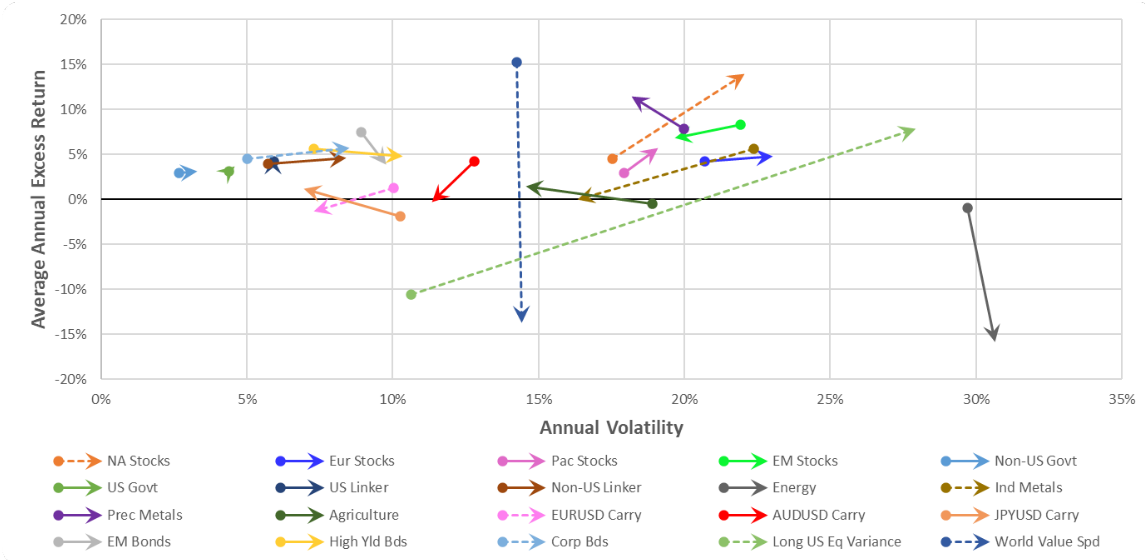


Figure 14 ***Sharpe Ratio Range for Statistical Broad Benchmarks***

This chart shows the rolling three-year Sharpe ratio for the 10 statistical broad benchmarks. The dark line in each range represents the reported returns. The shaded range denotes the gap between gross and net returns. A wider range indicates a combination of higher trading costs and/or fees. The realized experience of most investors resides between the dark line and the bottom of the range. The chart uses weekly returns between December 1999 and August 2020.

Panel A



Panel B

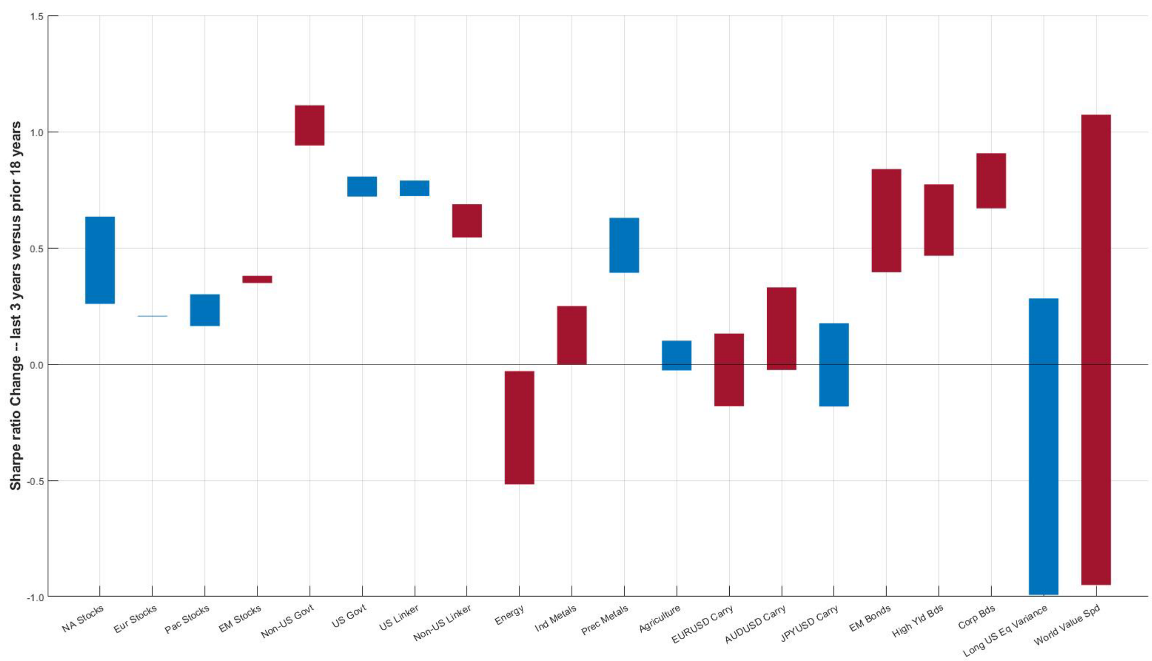


Figure 15 ***Reference Benchmark 2018-2020 Return and Volatility Footprint***

Panel A compares the average annual excess return and standard deviation of 20 benchmarks listed in Appendix A between the 2018-2020 and 2000-2017 periods. Calculations utilize weekly data from December 1999 through December 2020. The circle indicates results for the preceding 18-year period and the arrow denotes performance during the recent 3-year window. Panel B highlights the change in Sharpe ratio. A blue bar indicates a higher Sharpe ratio for the recent 3-year window (top of the bar) than the preceding 18-year period (bottom of the bar). Conversely, a red bar indicates a lower 2018-2020 than 2000-2017 Sharpe ratio. The bar length represents the Sharpe ratio difference.

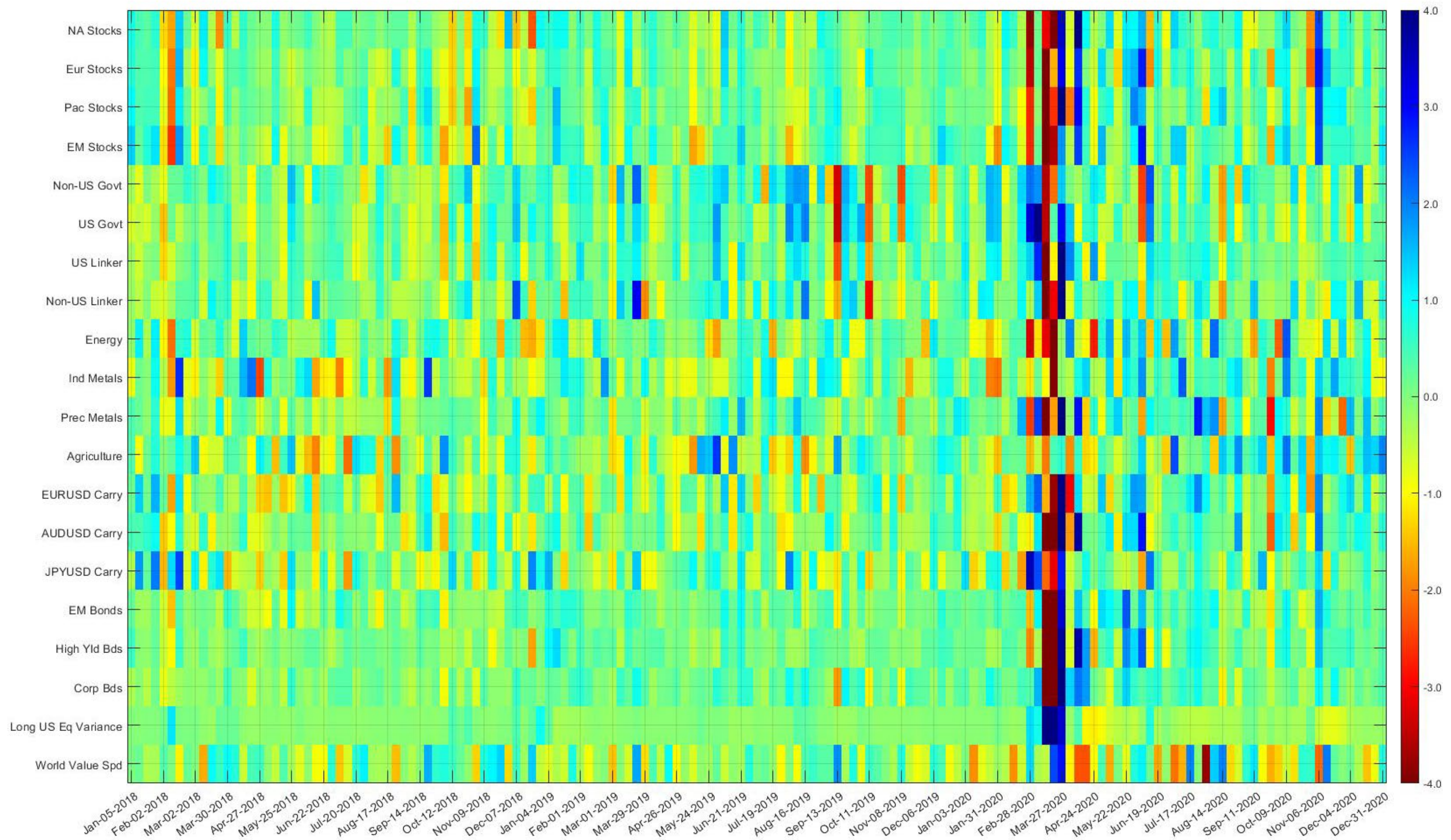


Figure 16 ***Reference Benchmark 2018-2020 Weekly Return Extremes***
This figure presents weekly excess returns, standardized by the 2018-2020 volatility, for 20 benchmarks between 2018 and 2020. Dark blue and dark red respectively indicate extreme positive and negative weekly returns. The standardized data is winsorized at +/-4.0, tempering the March 2020 outliers for graphical clarity.

Panel A

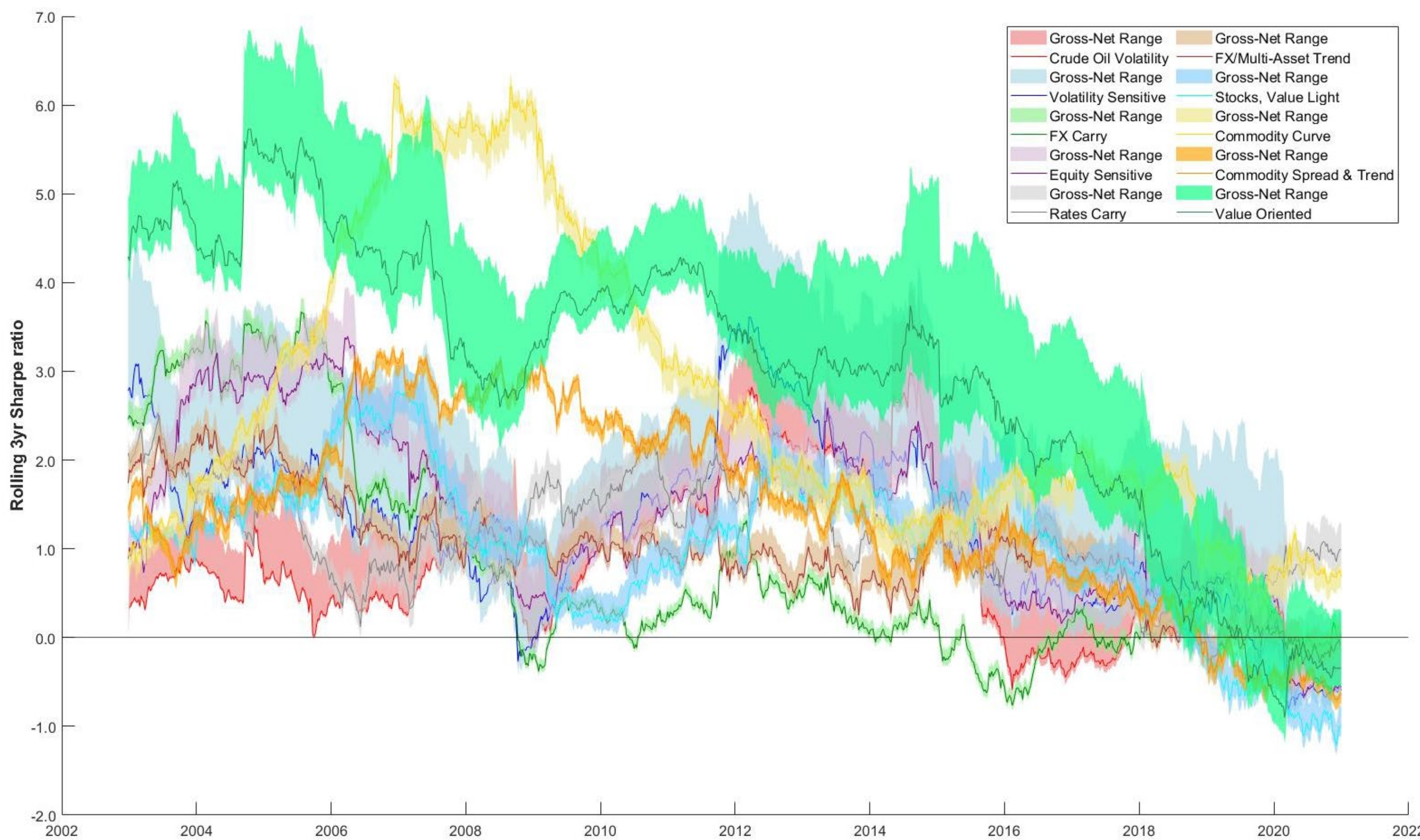


Figure 17 ***Sharpe Ratio Deterioration for Broad ARP Benchmarks***

Panel A shows the rolling three-year Sharpe ratio for the 10 statistical broad benchmarks. The dark line in each range represents the reported returns. The shaded range denotes the gap between gross and net returns. A wider range indicates a combination of higher trading costs and/or fees. The realized experience of most investors resides between the dark line and the bottom of the range. Panel B spotlights the latest observations in Panel A, punctuating that only three ARP broad benchmarks register a positive 2018-2020 Sharpe ratio or are not effectively at a Sharpe ratio nadir (no red shading). Red shading indicates a negative Sharpe ratio or a bottom decile ending Sharpe ratio relative to history. The chart uses weekly returns between December 1999 and December 2020.

Panel B

	Crude Oil Volatility (short)	Volatility Sensitive	FX Carry	Equity Sensitive	Rates Carry	FX/Multi-Asset Trend	Stocks, Value Light	Commodity Curve Carry	Commodity Spread Carry & Trend	Value Oriented
December 2020 3yr Sharpe ratio										
Reported	-0.3	-0.6	0.0	-0.6	1.0	-0.1	-1.0	0.7	-0.6	-0.4
Net	-0.3	-0.7	0.0	-0.6	0.8	-0.1	-1.1	0.5	-0.7	-0.6
Gross	0.0	0.3	0.1	-0.3	1.2	0.2	-0.8	0.8	-0.6	0.3
December 2020 3yr Sharpe ratio Percentile versus 2000-2020										
Reported	6%	3%	22%	1%	42%	5%	2%	3%	1%	3%
Net	6%	3%	21%	2%	39%	6%	1%	5%	0%	5%
Gross	3%	3%	24%	1%	41%	6%	0%	3%	0%	2%

Figure 17 *continued*

Sharpe Ratio Deterioration for Broad ARP Benchmarks

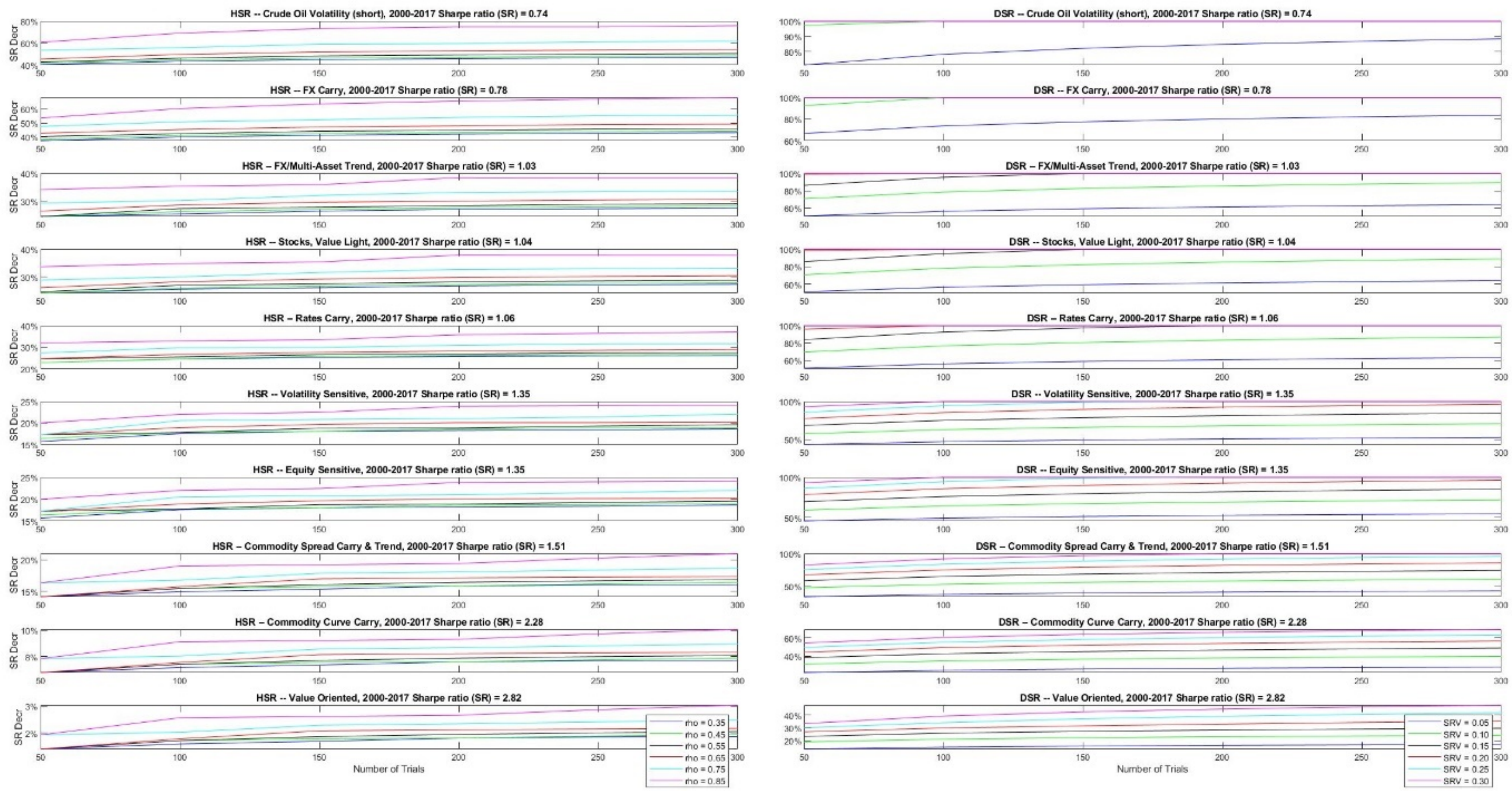


Figure 18 ***Sensitivity of Sharpe ratio Adjustment to Parameters Requiring Assumptions***

This figure highlights the sensitivity of the Sharpe ratio reduction (SR Decr on the vertical axes) to the three inputs for which information is unavailable for tradable bank indices – number of trials (horizontal axes), correlation among trials (rho, different colored lines for HSR), and Sharpe ratio variance (SRV, different colored lines for DSR). The haircut Sharpe ratio (HSR) appears in the left column and the deflated Sharpe ratio (DSR) the right. The different historical Sharpe ratios for the 10 broad ARP benchmarks illustrates the interaction among these parameters.

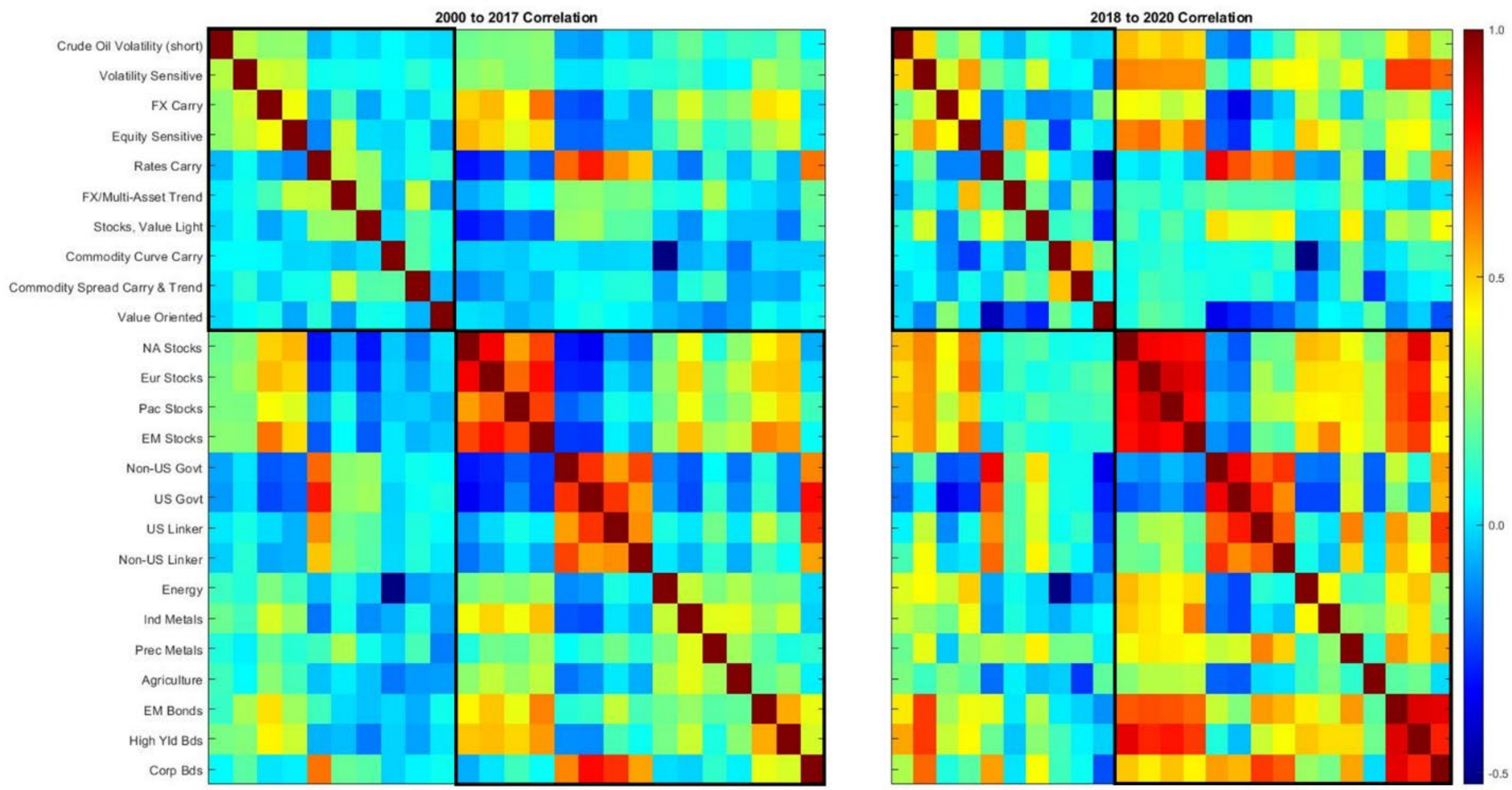
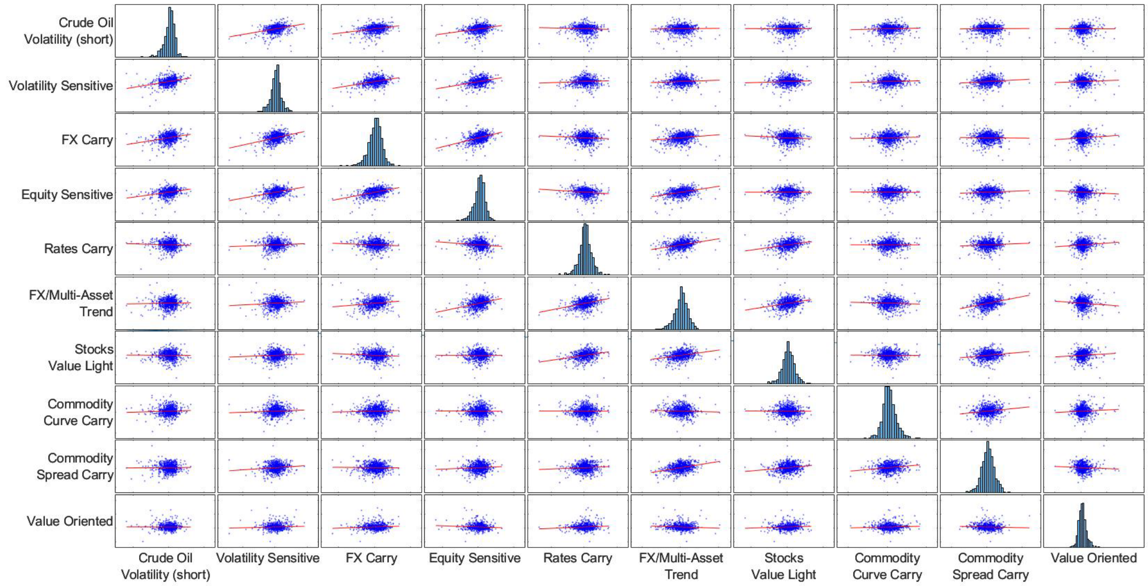


Figure 19 ***ARP Broad Benchmark Correlation Comparison***
This heatmap compares the correlations among weekly returns for 10 broad ARP and 15 traditional benchmarks over the 2000-2017 and 2018-2020 periods. Dark red (blue) indicates a very high (low) correlation. The diagonal provides a label point of reference and the box in the northwest (southeast) quadrant demarcates intra-ARP (intra-traditional) benchmark correlations.

Panel A: 2000 to 2017



Panel B: 2018 to 2020

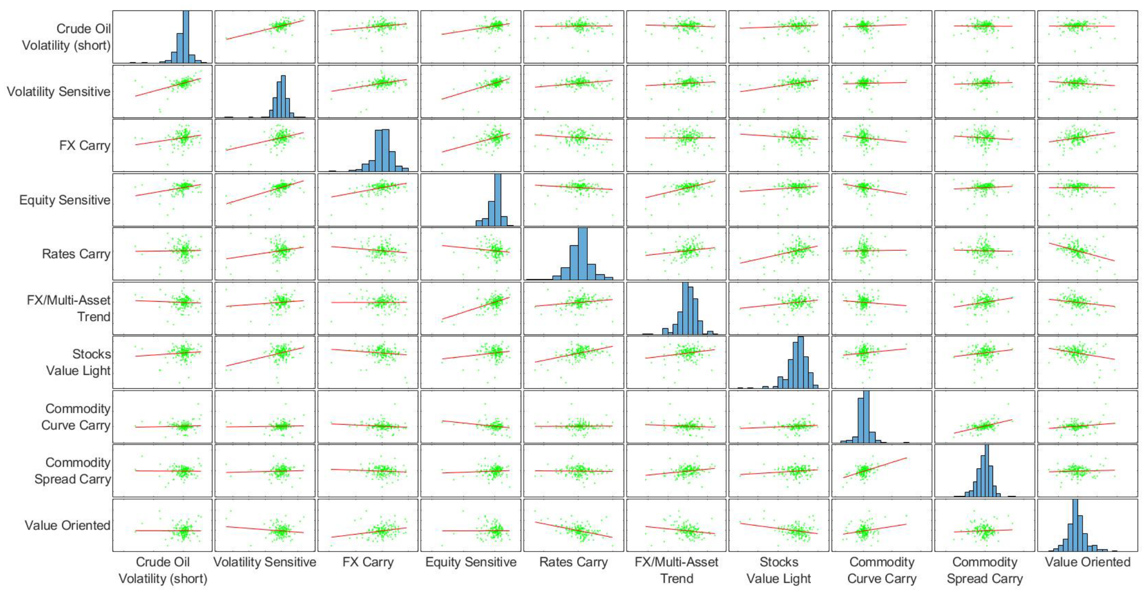


Figure 20 ***ARP Broad Benchmark Scatterplot Comparison***

The matrix plots below supplement the correlation heatmap, summarizing the distribution of returns for the ARP broad benchmarks. Panel A (blue) uses weekly returns between January 2000 and December 2017, while Panel B (green) targets January 2018 to December 2020. The diagonal of each panel contains histograms for each benchmark. The off-diagonal contains scatterplots, with the column (row) label being the X (Y) coordinates. The red line represents the OLS fit, indicating the linear relationship between the two benchmarks.

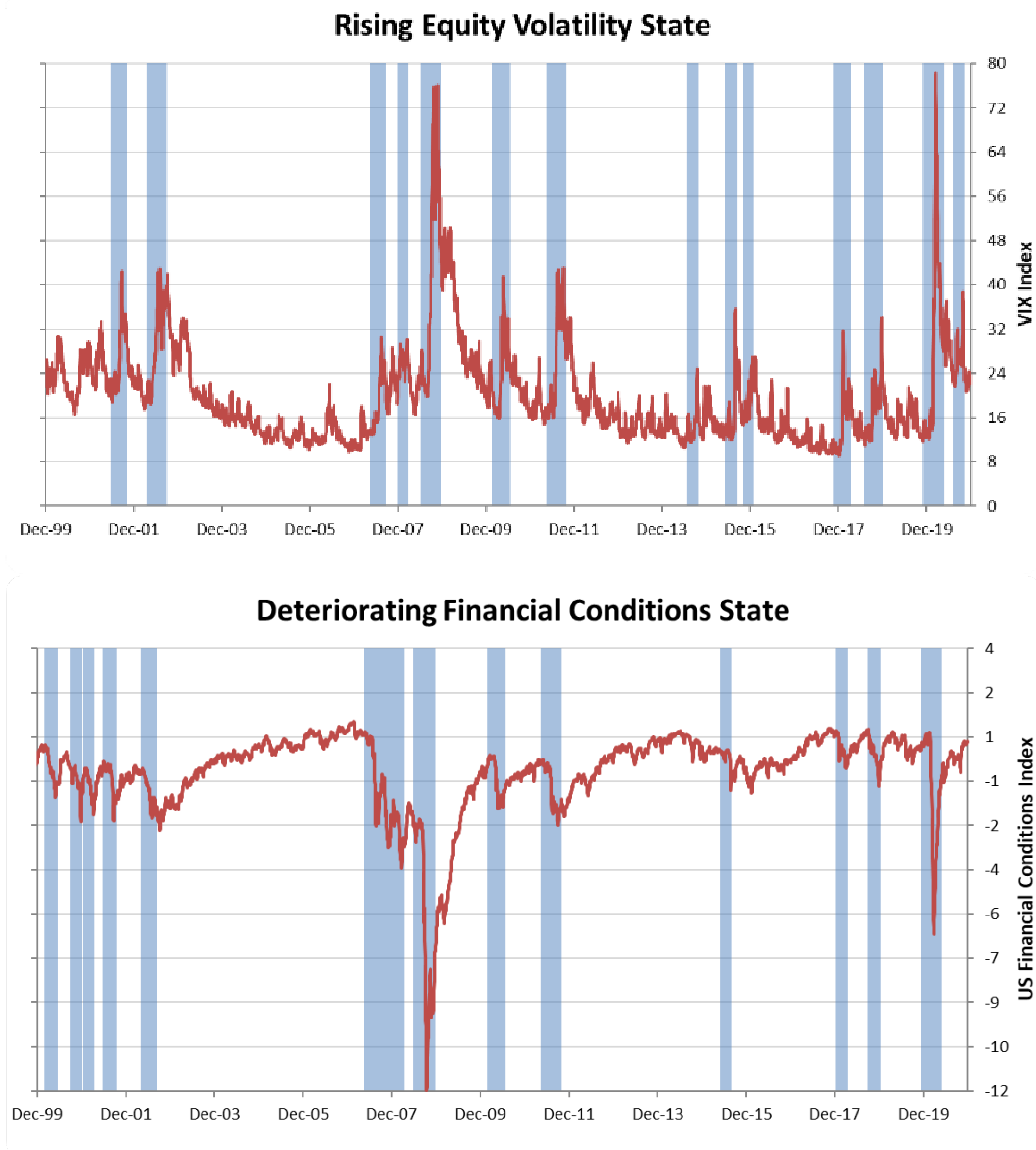


Figure 21 ***Market Environment State Indicators***

The figure contains two examples of market state representations, one for US equity volatility on the top and the other for US financial conditions on the bottom. The shaded areas indicate an n-day change of a signed minimum of x over a t-day period. An 11-point increase in the 3-day moving average for the CBOE VIX and a 1.25 point decrease in the 3-day moving average for the Bloomberg US Financials Conditions Index capture states representing 20-25% of the 2000-2020 history for each data set.

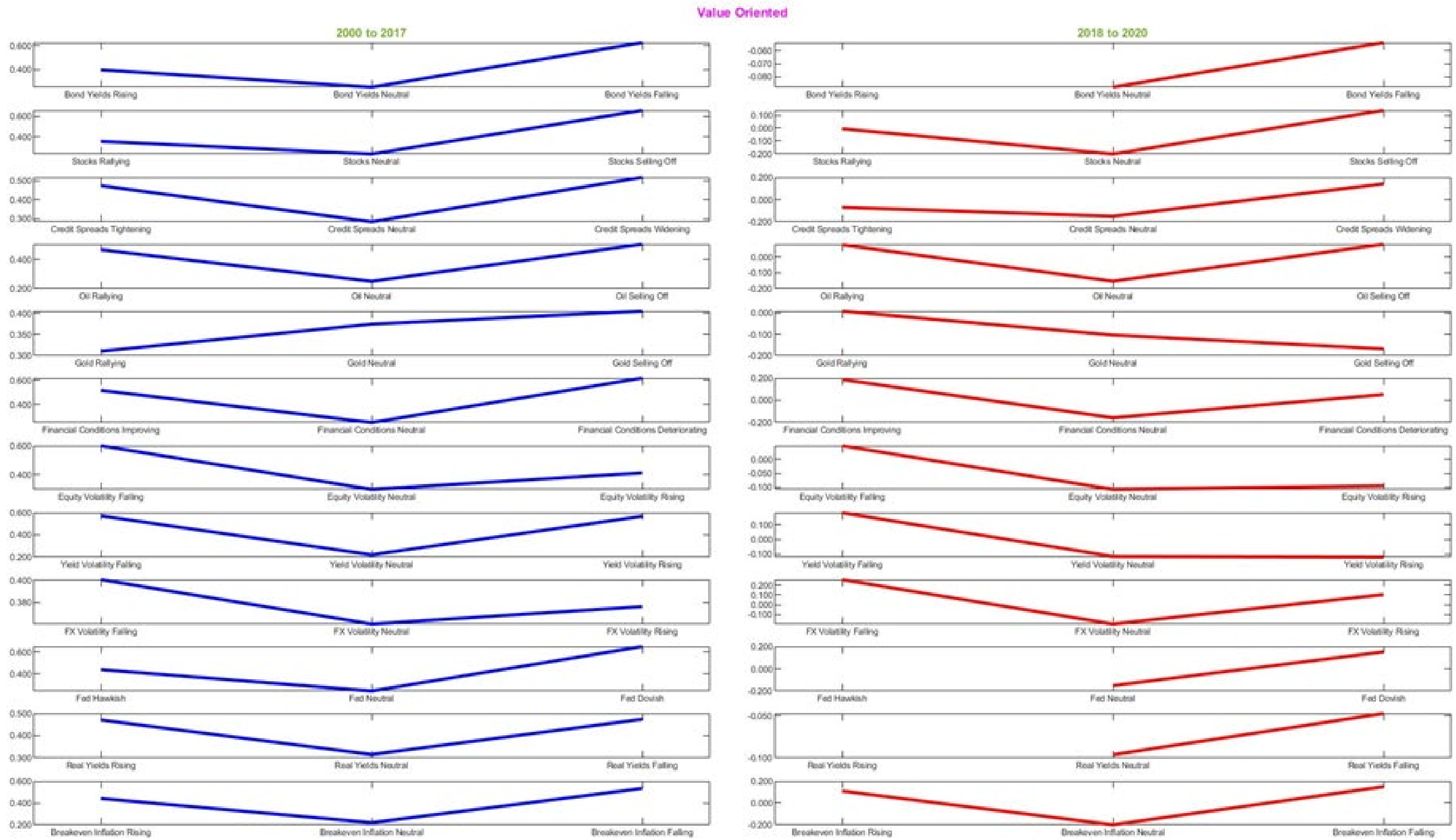
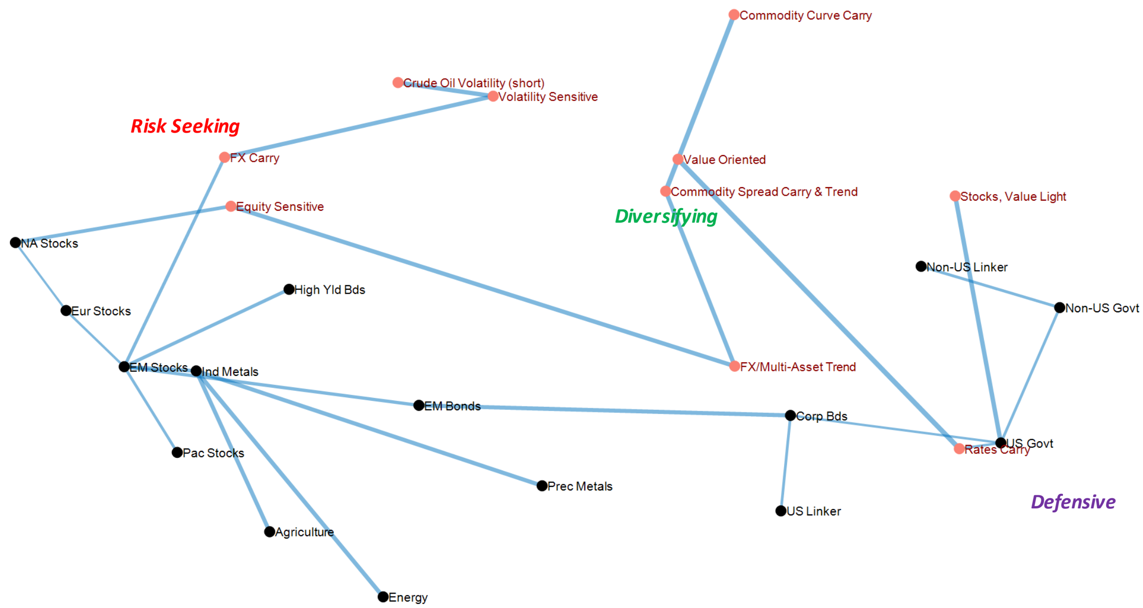


Figure 22 ***State-Based Conditional Means for Value-Oriented ARP Strategies***

Using weekly returns between January 2000 and December 2020 and 24 market environment indicators, the table provides the conditional weekly mean for the two opposing states and the neutral state for the value-oriented broad ARP benchmark. The left (right) column contains the 2000-2017 (2018-2020) results. The unconditional means are very different for the two time periods, so the shape of the lines, the neutral-relative difference, is most relevant. A missing line segment indicates no state representation during that time period.

Panel A: 2000 to 2017



Panel B: 2018 to 2020

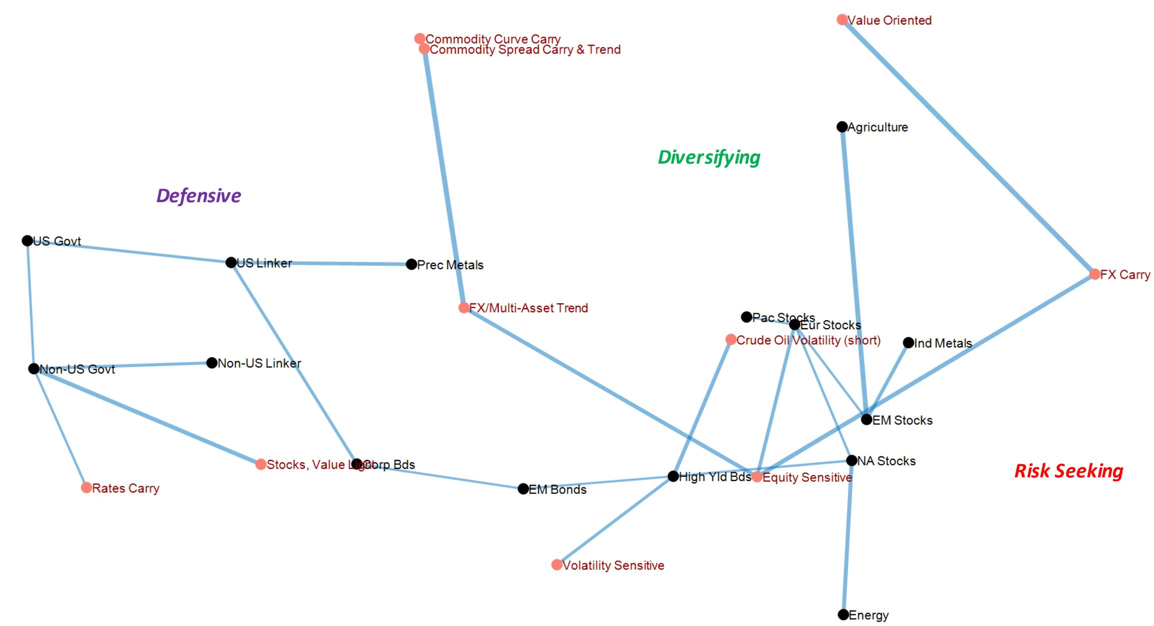


Figure 23 ***Minimum Spanning Trees for ARP Broad Benchmarks***

This exhibit provides an undirected graph summarizing the correlations among ARP and traditional benchmarks for the 2000-2017 (Panel A) and 2018-2020 (Panel B) periods. Vertices for ARP (traditional) benchmarks appear in salmon (black). Segment width indicates distance, with a thinner line indicating closer proximity. Large labels indicate three general neighborhoods in terms of portfolio construction role: risk seeking (red), diversifying (green) and defensive (purple). The orientation of Panel A and B is different, but the neighborhoods are essentially the same.

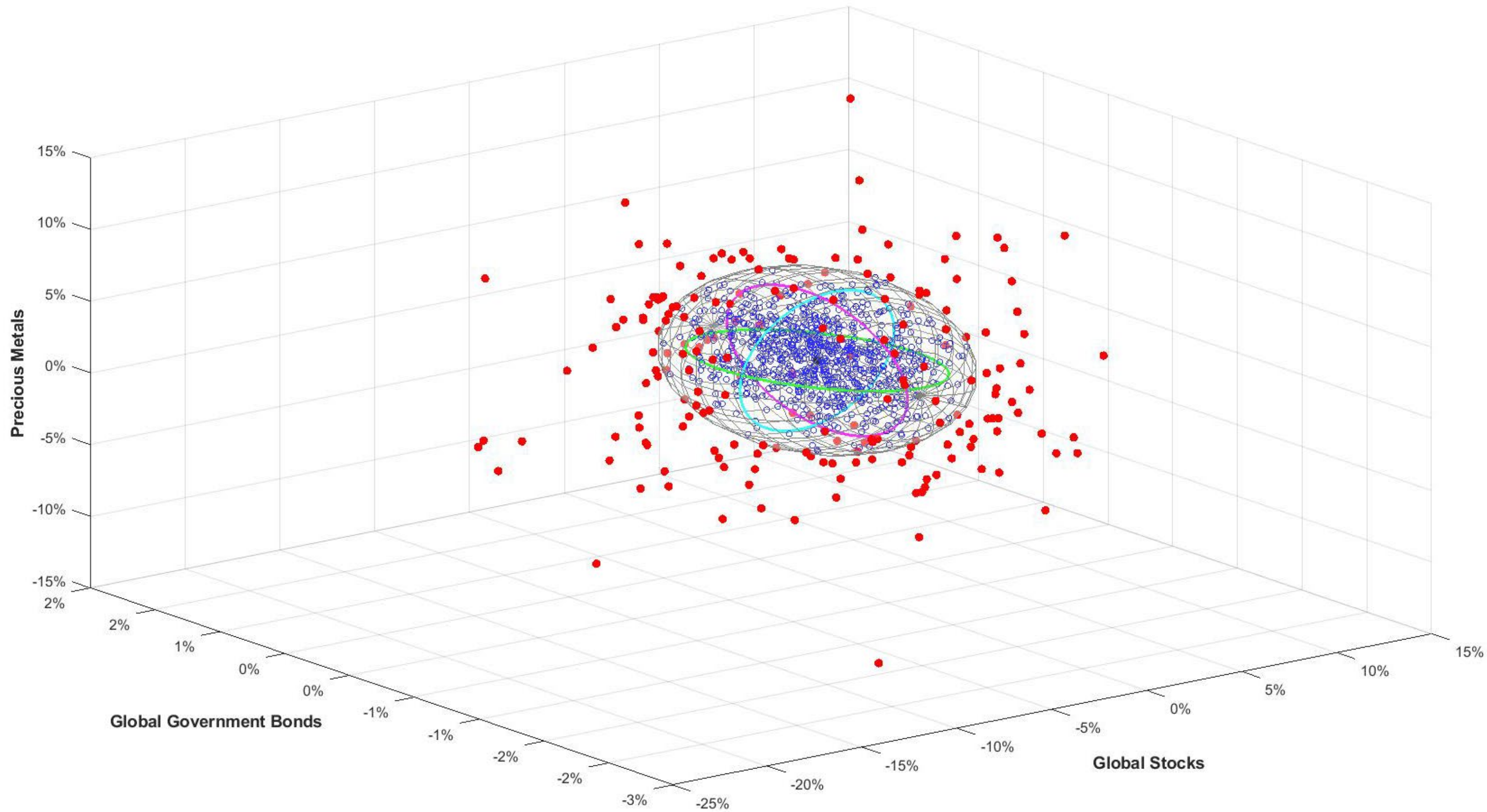


Figure 24 ***Turbulence Identification for ARP***
Based upon weekly excess returns for the MSCI ACWI, FTSE WGBI and Bloomberg Precious Metals indices between January 2000 and December 2020, this figure highlights turbulent data points with filled red circles – i.e. those observations falling outside the tolerance ellipsoid at a 0.75 probability. The unfilled blue circles indicate non-turbulent data points. The three underlying error ellipses appear in cyan, magenta and green.

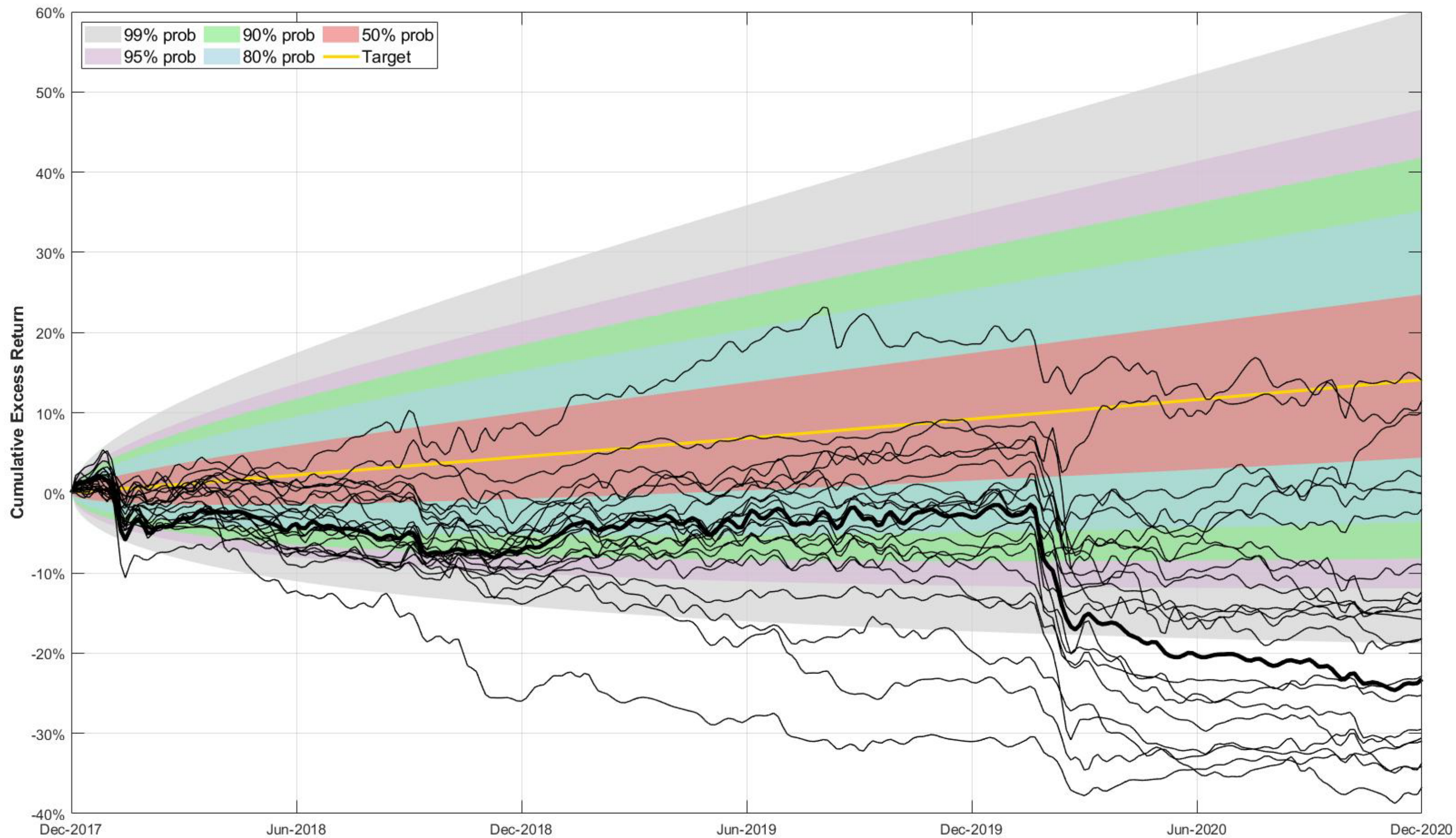
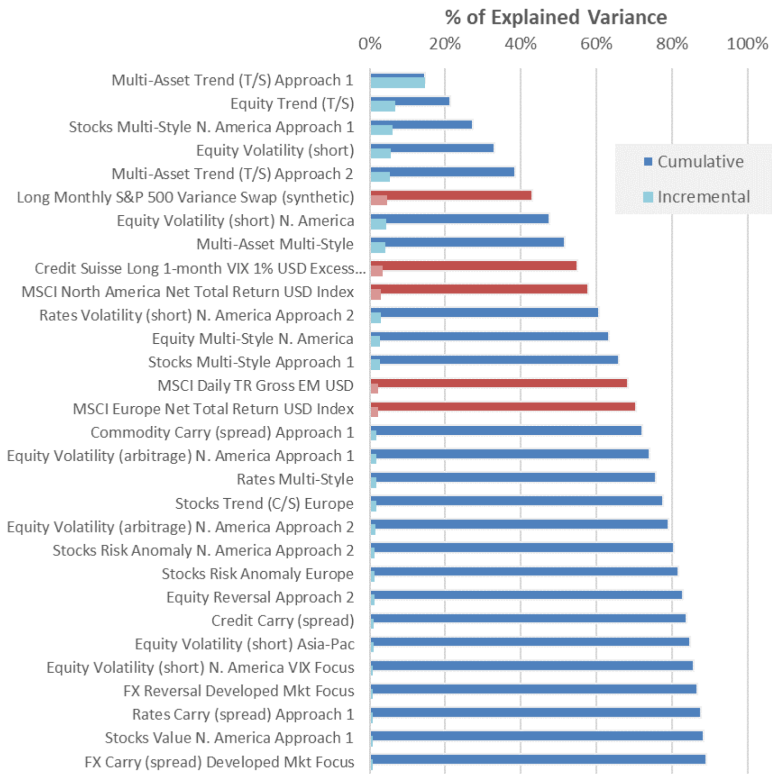


Figure 25 ***2018-2020 Performance by Diversified ARP Fund Managers***

This cone chart compares cumulative excess return across 22 diversified ARP managers (thin black lines) with weekly prices and distributions quoted in Bloomberg between December 2017 and December 2020. The exhibit scales all excess return histories to 8% annual volatility to facilitate comparison. The dark black line represents the SG Multi Alternative Risk Premia Index, an equally weighted blend of ARP managers with multi-asset and multi-style exposures. The target excess return line reflects an expected Sharpe ratio of 0.6 on 8% volatility. The multi-colored cones highlight the probability of various outcomes assuming normally distributed returns. Most funds deliver significantly negative deviations from expectation over the recent three-year period.

Panel A



Panel B

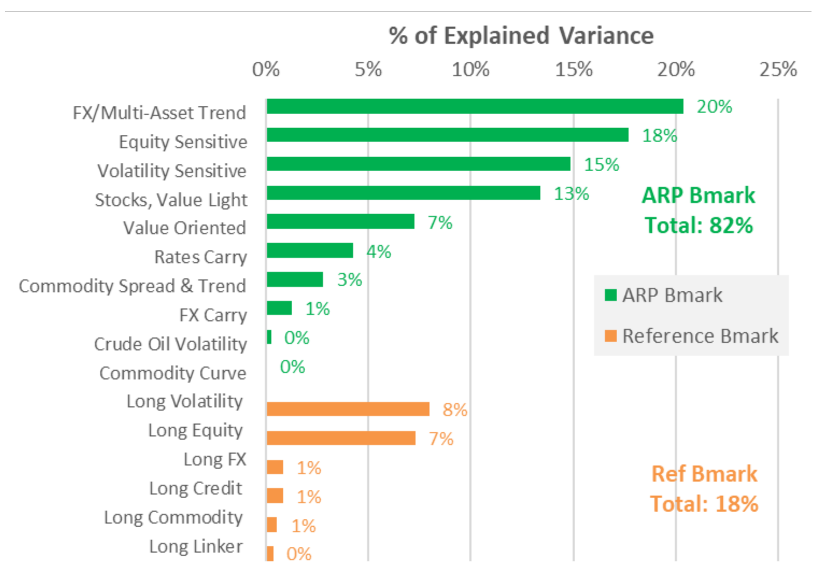
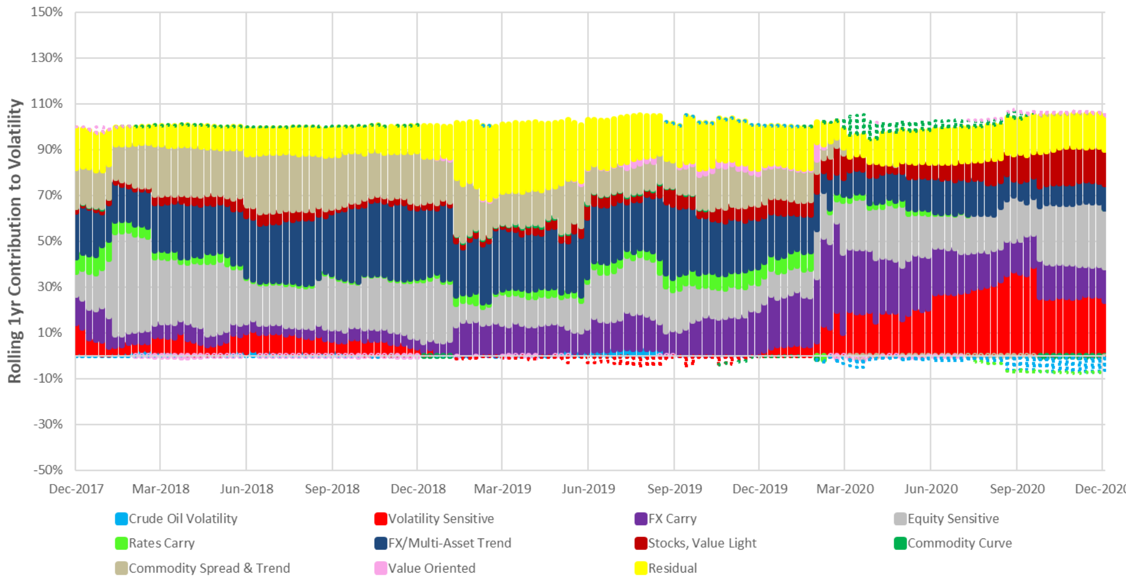


Figure 26 ***Diversified ARP Funds Risk Contribution Profile***

Using weekly returns between December 2017 and December 2020 in an EN regression, Panel A summarizes the contribution to explained variance from 85 base ARP benchmarks (blue) and 21 reference benchmarks (red) across 22 diversified ARP funds. The coefficient of determination is 58%, leaving 42% of total fund variance unexplained by these benchmarks. Panel B consolidates the variance explained in broad ARP benchmarks (green) and reference benchmark (orange) groups, highlighting that ARP benchmarks account for 82% of explained variance (over 90% excluding long volatility and FX reference benchmarks).

Panel A



Panel B

Figure 27 ***SG Multi Alternative Risk Premia Index Risk Contribution Profile***

Using weekly returns between December 2016 and December 2020 for the SG diversified ARP manager index, Panel A shows the explanatory power of the 10 broad statistical ARP benchmarks in a rolling 52-week EN regression. (The results for 2017 provide a baseline for the subsequent three years.) Dotted fill indicates a negative benchmark loading (potentially indicating a spread relationship with another benchmark) and a negative risk contribution indicates a diversifying role. The sum of all colors except the yellow residual is the R^2. Panel B displays the corresponding 52-week absorption ratio, a systemic risk measure indicating the fraction of variance across the 10 broad benchmarks explained by the first three principal components.

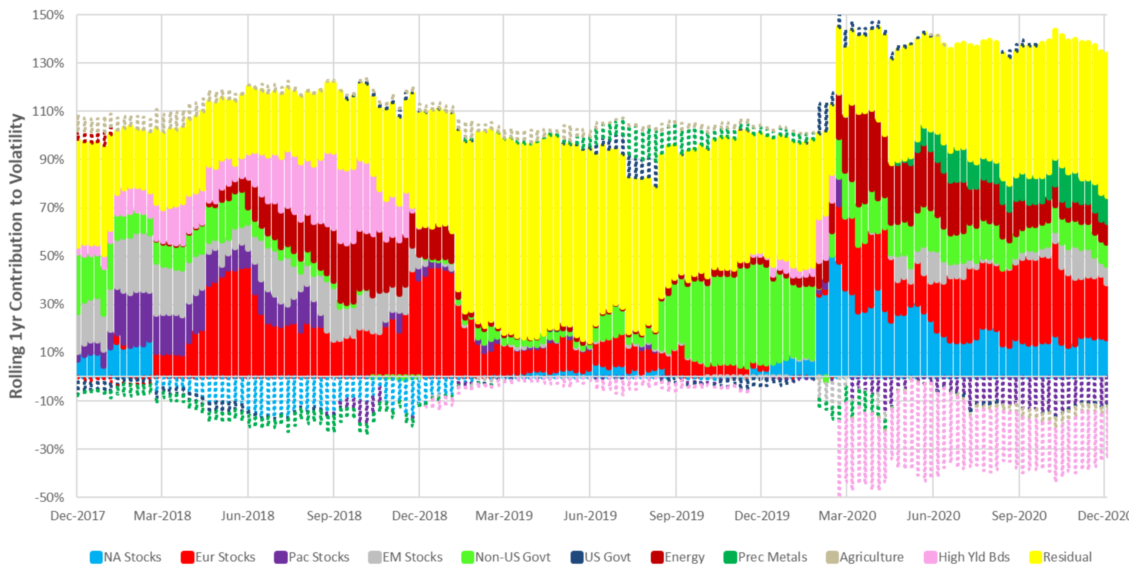


Figure 28 ***SG Multi Alternative Risk Premia Index Risk Contribution Profile***

This exhibit repeats the exercise from Panel A of Figure 27 using traditional, long-only benchmarks to much noisier effect. The amount of residual variance, instability of benchmark contributions, and dependence on negative loadings reinforce the distinct profile of ARP funds.

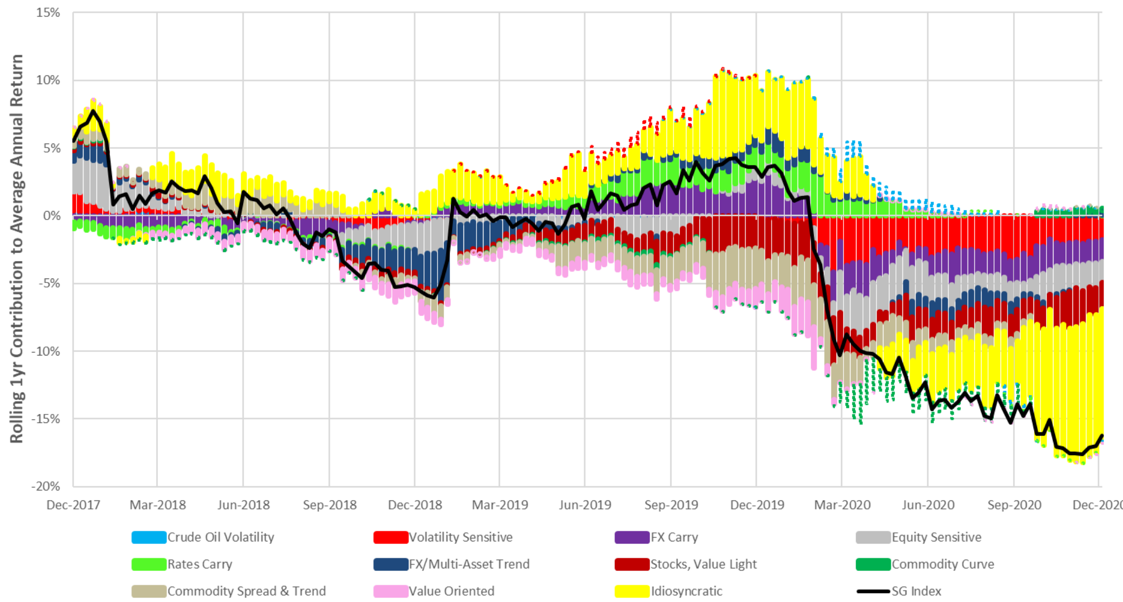
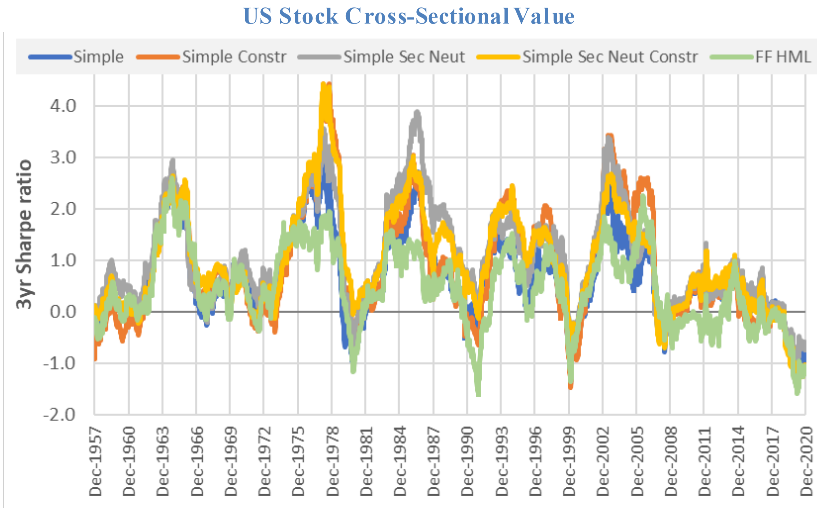


Figure 29 ***SG Multi Alternative Risk Premia Index Return Contribution Profile***

Using weekly returns between December 2016 and 2020 for the SG diversified ARP fund index, this chart provides the annual excess return contributions underlying the analysis in Figure 27. Dotted fill indicates a negative benchmark loading, which generally make a negligible return contribution -- commodity curve in 2020 is the exception, as the regression struggles to differentiate its footprint. The black line represents the average annualized 52-week excess return for the SG index.



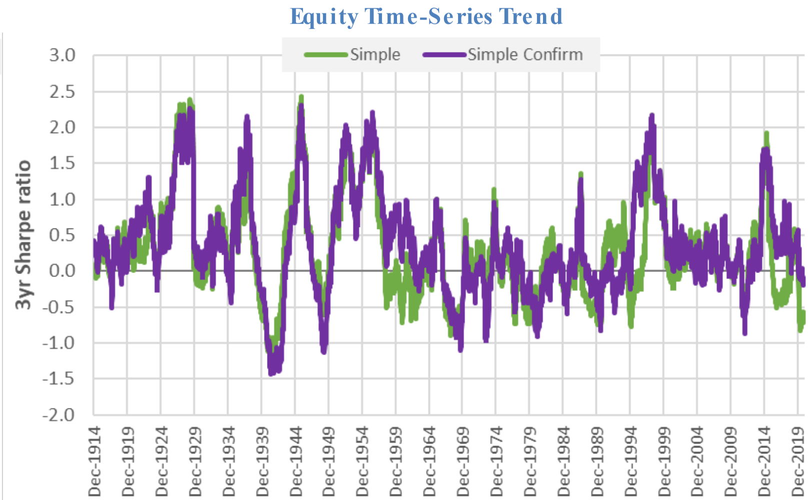


Figure 30 ***Primitive US Stock Value and Equity Trend Benchmarks***

Using gross weekly returns between December 1954 and December 2020 for the US stock (long-short) cross-sectional value benchmark and between December 1911 and December 2020 for the equity time-series trend benchmark, this exhibit provides historical context for the 2018-2020, three-year Sharpe ratio of two important contributors to the recent poor performance of ARP portfolios. The value chart on the left includes five primitive benchmarks — the first measuring the composite value score relative to the investable universe (Simple), the second adding size and market risk constraints to the first, the third utilizing the composite value score on a sector relative basis (Simple Sec Neut), the fourth adding size and market risk constraints to the third, and the fifth being the Fama-French high-minus-low factor (FF HML). The trend chart on the right includes two primitive benchmarks – the first taking a long or short position based upon the sign of the trailing 260-day S&P 500 return (Simple) and the second adding a basic confirmation to the first in the form of a z-score threshold to avoid taking positions based upon a weak trend signal (Simple Confirm).

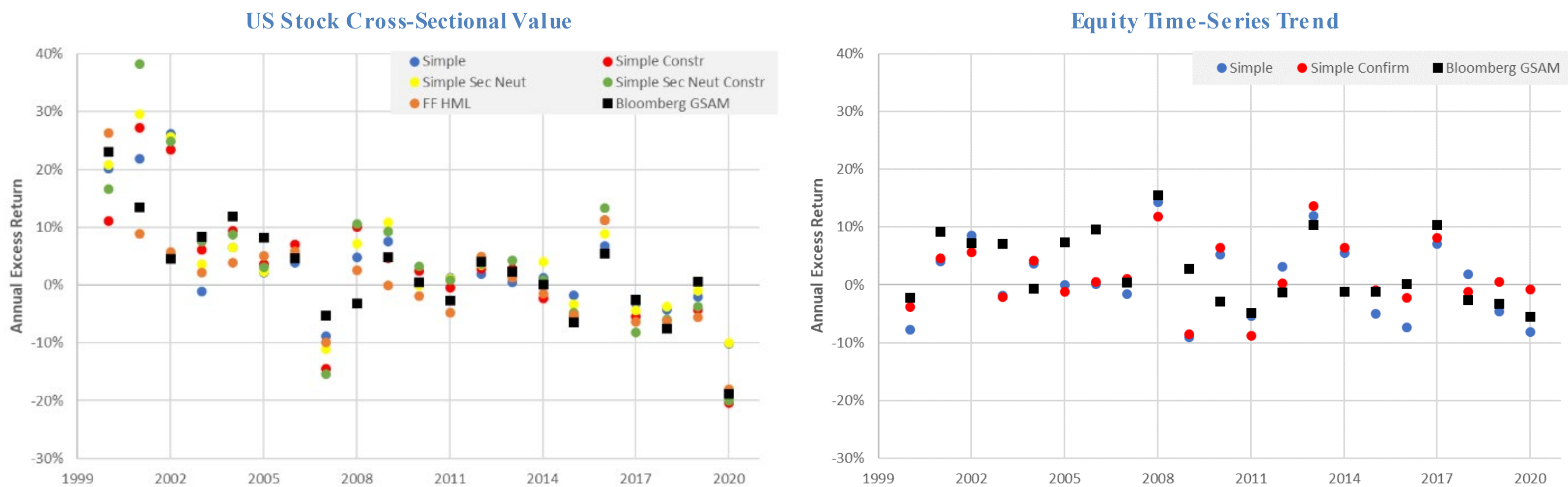


Figure 31 *Annual Returns to Primitive US Stock Cross-Sectional Value Benchmarks*

Using gross weekly returns from December 1999 to December 2020, this chart shows on the left the annual excess returns for the Bloomberg GSAM US Equity Value L/S Index (black box) with five alternative primitive benchmarks (multi-colored circles) — the first measuring the composite value score relative to the investable universe (Simple), the second adding size and market risk constraints to the first, the third utilizing the composite value score on a sector relative basis (Simple Sec Neut), and the fourth adding size and market risk constraints to the third, and the fifth being the Fama-French high-minus-low factor (FF HML). The Bloomberg GSAM Equity Trend Index (black box) appears on the right with two alternative primitive benchmarks (multi-colored circles) — the first taking a long or short position based upon the sign of the trailing 260-day S&P 500 return (Simple) and the second adding a basic confirmation to the first in the form of a z-score threshold to avoid taking positions based upon a weak trend signal (Simple Confirm). For comparability, returns for all benchmarks are scaled to a volatility of seven percent.

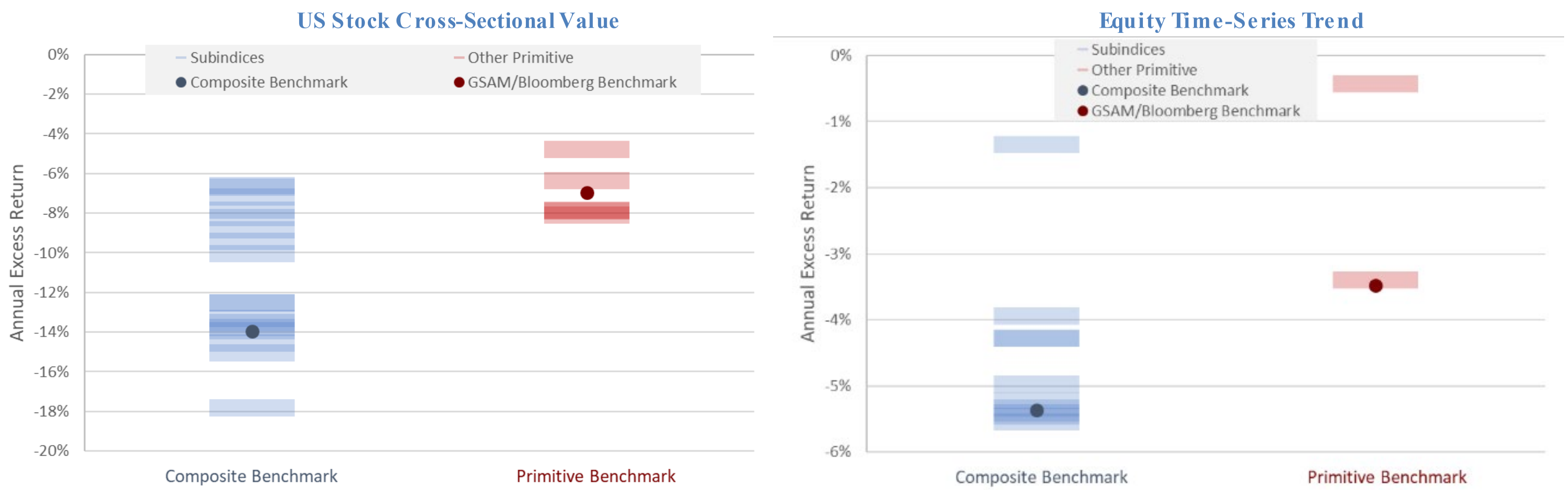


Figure 32 ***Comparison of Primitive and Composite Benchmarks in US Cross-Sectional Value and Equity Time-Series Trend***
The chart reports the three-year annual gross excess return for the 2018-2020 period for two important strategies within ARP portfolios. For US stock cross-sectional value on the left, the return for the base statistical composite benchmark appears as a dark blue dot, with blue bars representing the return of the 17 underlying tradable bank indices and dark shading indicating proximate results. A dark red dot shows the return for the GSAM/Bloomberg benchmark, with red bars indicating the return of five alternative primitive benchmark candidates. For equity time-series trend on the right, blue bars capture the 10 tradable bank indices comprising the composite and red bars show two primitive benchmark possibilities. For comparability, each return series is scaled to a volatility of seven percent.