PCA for Implied Volatility Surfaces

MARCO AVELLANEDA, BRIAN HEALY, ANDREW PAPANICOLAOU, AND GEORGE PAPANICOLAOU

MARCO AVELLANEDA
is a professor of mathematics at the Courant Institute of Mathematical Sciences at New York University in New York, NY.
avellaneda@courant.nyu.edu

BRIAN HEALY
is a researcher at the Department of Financial Computing and Analytics at University College London in London, UK.
brian@decisionsci.net

ANDREW PAPANICOLAOU
is an assistant professor at the Department of Finance and Risk Engineering at NYU Tandon School of Engineering in Brooklyn, NY.
ap1345@nyu.edu

GEORGE PAPANICOLAOU
is a professor at the Department of Mathematics at Stanford University in Stanford, CA.
papanicolaou@stanford.edu

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

KEY FINDINGS

- Principal component analysis of a comprehensive dataset of implied volatility surfaces from options on US equities shows that their collective behavior is captured by just nine factors, whereas the effective spatial dimension of the residuals is closer to 500 than to the nominal dimension of 28,000, revealing the large redundancy in the data.
- Portfolios of implied volatility surface returns, weighed suitably by open interest and Vega, track the principal eigenportfolio associated with a market portfolio of options, in analogy to equity portfolios.
- Retention of the tensor structure in the eigenportfolio analysis improves the tracking between the open interest–Vega weighted (tensor) implied volatility surface returns portfolio and the (tensor) eigenportfolio, indicating that data structure matters.

ABSTRACT: *Principal component analysis (PCA) is a useful tool when trying to construct factor models from historical asset returns. For the implied volatilities of US equities, there is a PCA-based model with a principal eigenportfolio whose return time series lies close to that of an overarching market factor. The authors show that this market factor is the index resulting from the daily compounding of a weighted average of implied-volatility returns, with weights based on the options' open interest and Vega. The authors also analyze the singular vectors derived from the tensor structure of the implied volatilities of S&P 500 constituents and find evidence indicating that some type of open interest- and Vega-weighted index should be one of at least two significant factors in this market.*

TOPICS: ***Statistical methods, simulations, big data/machine learning****

We show through principal component analysis (PCA) that a relatively small number of factors can account for most of the variation in the collective movements of the implied volatilities derived from US equity options. In fact, a matrix formed with normalized implied volatility returns over time has positive covariance with the first principal eigenvector, which is closely related to an open interest (OI)- and Vega-weighted basket of implied volatilities. This draws parallels with the first principal component obtained from the covariance of the matrix of normalized equity returns and its proximity to the capitalization-weighted portfolio (see Avellaneda and Lee 2010; Boyle 2014); OI is a measure of market size for options just as capitalization is a measure of market size for equities, and both implicitly

carry liquidity information. OI is the number of open contracts for a given option (name, strike, maturity) at a given time. Our findings highlight and give detailed insight into how PCA can be used to extract information from the covariance structure for a large dataset of implied volatilities and how new and improved implied volatility factors can be constructed using OI and Vega, both of which will be useful for portfolio and risk managers who have a need for better statistical prediction models to improve their estimated risk metrics (e.g., Value at Risk [VaR] and expected shortfall). To date, the preeminent volatility index is VIX, which is constructed from index options. The VIX has proven reliable but on occasion has shown susceptibility to outlier prices and manipulative trading (see Griffin and Shams 2017). The OI–Vega-based factors constructed in this article are more robust because they are based on hundreds of implied volatilities and place more emphasis on those contracts that have most trading interest.

The study described in this article builds on the work of Avellaneda and Dobi (2014), who considered a large dataset of implied volatility surfaces for a few thousand US equities and used PCA to find the smallest number of factors needed to explain the collective movements of these volatilities. They also constructed principal eigenportfolios and examined the qualitative structure of each from the first through fourth eigenvectors. We draw from the same data source as Avellaneda and Dobi (2014)—namely, the implied volatility surface (IVS) data available from OptionMetrics through Wharton Research Data Services. We hypothesize that the normalized covariance matrix of market-wide implied volatilities has a *low-rank plus random* structure (known as the *spike model*), and similar to Avellaneda and Dobi (2014), we find that removal of the low-rank components leaves a residual whose squared singular values are close in distribution to a Marchenko–Pastur (MP) law. Using random matrix theory (RMT), the presence of principal factors should make it possible to reject a model of purely random noise. RMT was used by Avellaneda and Dobi (2014), with the spectra-limiting MP distribution providing the basis for establishing cutoffs for the identification of the nonrandom structure. The analyses in this article consistently (for multiple years) show there to be at least two outliers in the singular value distribution, indicating that at least two factors drive the time series of IVS returns.

As noted, we focus on the first principal eigenportfolio and its closeness to various OI- and Vega-weighted portfolios. An eigenportfolio is a vector of portfolio weights that is derived from an eigenvector of the returns' covariance matrix. In equities, it is well known that the eigenportfolio constructed from the first principal eigenvector has explanatory power for the cross section of US equity returns (Avellaneda and Lee 2010; Boyle 2014) and that it tracks closely with a dominant factor such as a market portfolio. Perhaps the most significant finding in this article is a sizable body of evidence indicating that option OI and OI–Vega are key elements in the construction of factors to explain the collective cross-sectional changes of IVSs. Specifically, we show that various factors constructed from OI and OI–Vega weightings of implied volatility returns have significant explanatory power for interpreting the principal eigenportolio's returns. This finding can be considered the *implied volatility analogue* to the equity market's first principal factor, namely the capitalization-weighted returns portfolio. As noted, it appears that OI plays a role for implied volatility similar to that played by capitalization for equities. However, such a comparison is very informal because the capital asset pricing model (CAPM) and its related economic theory bind equities and capitalization closely together, whereas the implied volatility results shown in this article are, at present, statistical findings.

The time series of IVSs can be put into vector form, but its natural representation is a four-dimensional tensor, with the four dimensions being time, name, option maturity, and option delta (normalized strike). Our study of the first principal eigenportfolio can be extended to this tensor setting, which provides an example of how factor construction can, in fact, be improved by considering the natural representation offered by the tensor structure. The maturity and strike dimensions lend themselves to individual factors, for which we can construct individualized OI-weighted factors for each option maturity or for each maturity–delta pair. Individualized factors allow for a more nuanced weighing of changes in implied volatilities, and this leads to improved explanatory power for the covariance structure's, suitably defined, principal eigenportfolio.

REVIEW OF LITERATURE

The work of Avellaneda and Dobi (2014) and Dobi (2014) provides a cross-sectional classification of US

equity options based on implied volatility data for the period from August 2004 to August 2013, jointly with equity returns. The spectrum of the joint equity–IVS is used, in particular the leading eigenvalues, to classify options into those carrying mostly systemic risk and those carrying mostly idiosyncratic risk. Then, employing methods from PCA and results from random matrix theory, the significant eigenvalues are identified, and it is shown that approximately nine principal components suffice to reproduce the IVSs of all equities studied, with even fewer risk factors for so-called systemic names, such as SPY, QQQ, and AAPL. An explicit model is introduced that can be used to track the dynamics of the IVS, yet is compact and computationally tractable.

Focusing on the IVS for a single asset, Cont and Da Fonseca (2002) examined time series of option prices for options on the S&P 500 and FTSE 100 indexes. They showed how the IVS can be deformed and represented as a randomly fluctuating surface driven by a small number of orthogonal random factors. They found a simple factor model compatible with the empirical observations. Also of interest are methods for pricing baskets of many assets using option implied volatilities (see Avellaneda et al. 2002).

Before work on implied volatilities, there was a bounty of research on equities. The original work on portfolio composition dates back to Markowitz (1952); subsequent work on the CAPM model focused on the expected return of an asset relative to the risk-free instrument and derived a relationship between this and the excess return of a market or benchmark portfolio. Work in subsequent years suggested that factors other than the market excess return were being priced. In particular, Roll and Ross (1980) used data for individual equities during the 1962–1972 period and found that at least three are priced in the generating process of returns; Fama and French (1992) found the most significant factors to be market excess return, company size, and the ratio of the book value to the market value of the firm. Plerou et al. (2002) made early use of PCA and random matrix theory for analyzing equity returns.

The importance of eigenportfolios was highlighted by Boyle (2014), who examined the conditions under which frontier portfolios have positive weights on all assets. This is of interest because the market portfolio given by CAPM is mean–variance efficient and has positive weights on all assets. Before this work was that by Avellaneda and Lee (2010), who studied statistical arbitrage strategies in US equities with trading signals generated using PCA. Modeling the residuals of stock returns as a mean-reverting process, Avellaneda and Lee (2010) developed contrarian trading signals and then back tested these over the broad universe of US equities. The fact that these PCA-based strategies have an average annual Sharpe ratio that is statistically and economically significant provides empirical support for the PCA approach.

STRUCTURE AND RESULTS OF THE ARTICLE

This article has three main sections after this introduction. The first section addresses the estimation of the low-rank principal component structure from the standardized returns of options' implied volatility. The main result of this section is the introduction of an effective dimension approach for assessing the randomness of residuals, especially when there is only randomness in time because the vectorized IVS data produce residuals that do retain some of the structure of the data. The second section explores the role of the first principal component in constructing an eigenportfolio, with option OI and Vega as weights, as the primary factor in evaluating collective movements of IVSs. The final section makes use of the data's natural tensor structure for construction of improved principal eigenportfolios. The main contribution of this article is the presented evidence demonstrating the importance of OI when measuring changes in implied volatilities. Performing PCA to determine the number of relevant factors is a fairly standard procedure once we have standardized the data, but construction and analysis of eigenportfolios requires a deeper understanding of the data, including OI. The tensor analysis provides more depth of understanding because it shows us that construction of factors individualized to subcategories of options (e.g., separate OI-based factors for each of the options' maturities) leads to a clear improvement in the eigenportfolio's ability to account for implied volatilities' movements.

MATRIX OF IMPLIED VOLATILITY RETURNS

Let t be an index denoting calendar days. Let i be an index denoting an individual option contract, and denote the implied volatility for this particular option

contract as $\hat{\sigma}_i(t)$. We define at time t the vector of daily returns on the ith contract's implied volatility as

$$r_i(t) = \frac{d\hat{\sigma}_i(t)}{\hat{\sigma}_i(t)} \quad \text{for} \quad 1 \le i \le N \quad \text{and} \quad 1 \le t \le T, \tag{1}$$

where $d\hat{\sigma}_i(t) = \hat{\sigma}_i(t+dt) - \hat{\sigma}_i(t)$ with $dt = 1/252 = 1$ day. We assume throughout that the number of contracts far exceeds the number of days, $N \gg T$, which means that the covariance/correlation matrix has several eigenvalues equal to zero. We standardize these returns and then place them into a matrix $R \in \mathbb{R}^{N \times T}$, given by

$$R = \left[\frac{r_i(t) - \overline{r_i}}{h_i} \right]_{1 \le i \le N, 1 \le t \le T} \tag{2}$$

where $\overline{r_i} = \frac{1}{T}\Sigma_t r_i(t)$ and $h_i = \sqrt{\frac{1}{T-1}\Sigma_t (r_i(t) - \overline{r_i})^2}$. Our hypothesis is that R can be decomposed into a low-rank factor matrix F and a random matrix X

$$R = F + X \tag{3}$$

which can be tested using the spike-model approach (see Benaych-Georges and Nadakuditi 2011). The low-rank matrix F can be further decomposed into orthogonal components

$$F = \sum_{i=1}^{d} f_i \theta_i^* \tag{4}$$

where each vector f_i is a principal characteristic of R with orthogonality between f_i and f_j $\forall i \neq j$, each θ_i is its loading, and $*$ denotes matrix/vector adjoint. In particular, if $||f_i|| = 1$, then $||\theta_i||^2$ is an eigenvalue of FF^*.

There are two obvious issues to address: the value of d in Equation 4 and the vectors $(f_i)_{i=1,\ldots,d}$. In the RMT literature it is equally as important to determine the θ_i's because the criticality of a given θ_i will determine whether component f_i is distinguishable from R's bulk eigenvectors. In applications to financial data, however, the top components are typically much greater than the critical threshold, and attempts to include the middle ranks of F will usually lead to overfitting.

A brief description of the IVS data is given in the Appendix. For our analysis, we extracted 56 implied volatility values for each of the (roughly) 500 S&P 500 constituents for each of the approximately 252 business days in each year of data for 2012–2017. Specifically, we use options with time to maturity of 30, 60, 91, 122, 152, 182, 273, and 365 days and which have a delta (normalized strike; see Appendix) of −20, −30, −40, 50, 40, 30, and 20. Therefore, for each maturity we use three out-of-the-money put options (−20, −30, −40), one at-the-money option (50), and three out-of-the-money call options (40, 30, 20). Out-of-the-money options are used because these are more widely traded and hence are more liquid and have more reliable prices. Thus $N = 500 \times 8 \times 7 = 28{,}000$ and $T = 252$ (or 250 or 251 depending on when holidays fall in the year) if we use a one-year estimation window.

Singular Values of Nonprincipal Structure

A very basic estimator of the covariance matrix is $\hat{\rho} = RR^*/T$. Much of the literature has addressed methods of improving this estimator, including shrinkage of the eigenvalues by Ledoit and Wolf (2004) and asymptotic behavior of eigenvectors by Ledoit and Péché (2011) and Ledoit and Wolf (2012). Perhaps the most applicable reference for what we are seeking in this article is Benaych-Georges and Nadakuditi (2011), who gave a result stating that if the magnitude of the $||\theta||_i$'s is greater than some critical threshold with respect to the variance of X's entries in Equation 3, then the principal component vectors of $\hat{\rho}$ will be inside a cone centered around the principal eigenvectors of FF^*/T. For financial data, the first principal component often accounts for as much as 50% of the total variance and thus has an eigenvalue that is well over the threshold, but higher-order factors may be closer to the critical level. Detection of factors whose eigenvalues are near the critical threshold is interesting but not the main focus in this article. Instead, we focus on finding an estimate of the minimum number of factors needed to have a statistical nonrejection of the estimated low-rank model.

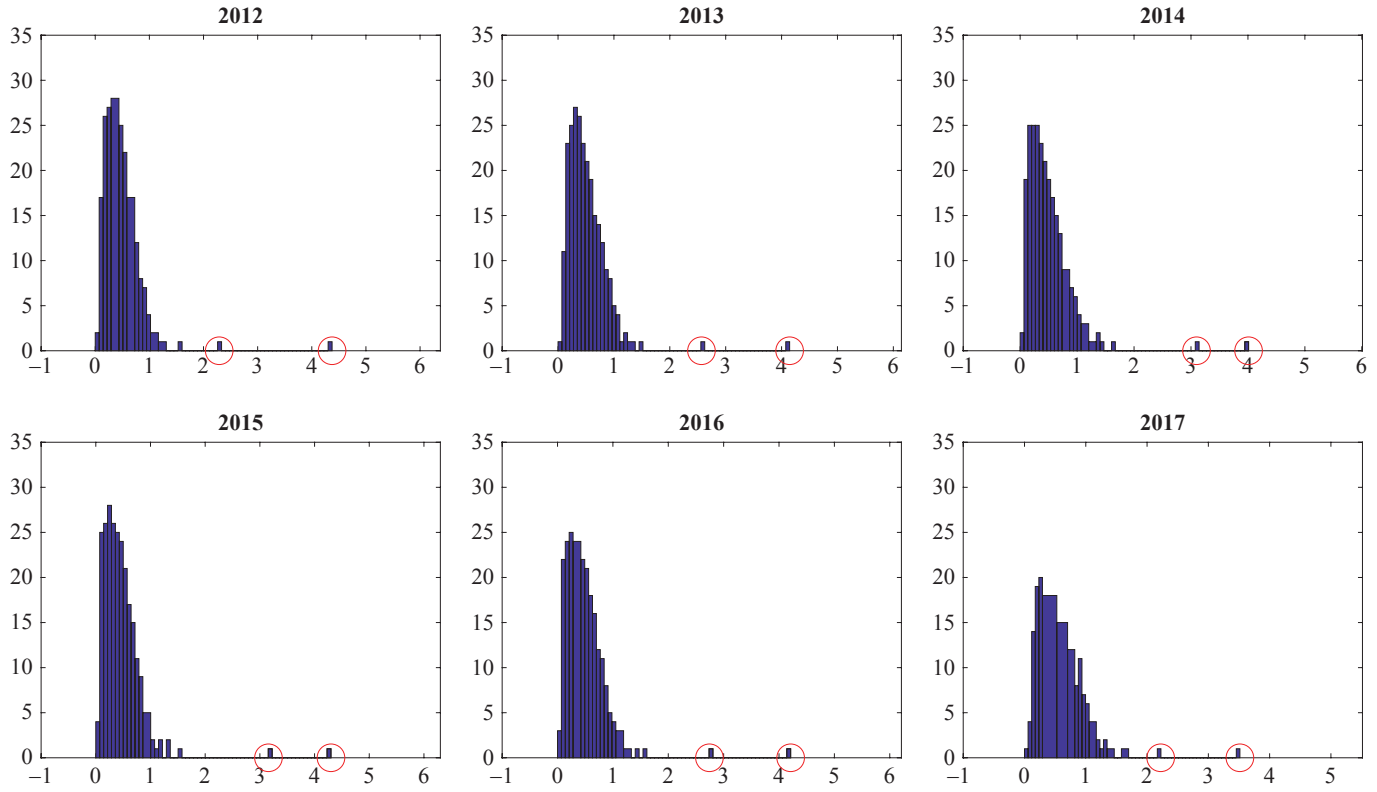
In practice, the estimator $\hat{\rho}$ is not calculated because usually $N \gg T$, making it more efficient to compute the singular value decomposition (SVD). The SVD represents R as

$$R = USV^*, \tag{5}$$

where $U = [U_1, U_2, \ldots, U_T]$ is an $N \times T$ matrix with orthonormal columns, S is a $T \times T$ diagonal matrix with entries $S_{11} \ge S_{22} \ge \ldots \ge S_{TT} \ge 0$, and $V = [V_1, V_2, \ldots V_T]$ is a $T \times T$ matrix with orthonormal columns.

EXHIBIT 1

Histograms of $\log(1 + S_{ii}^2/N)$ Where S_{ii} Are the Nonzero Singular Values of R, Estimated with One-Year Windows from 2012 to 2017



Notes: If $F = 0$, there will be low probability of outliers from the bulk, and the histogram can be fitted with a Marchenko–Pastur density. However, in each of these histograms we see at least two outliers (circled marks), which indicates the rank of F is at least two. Hence, principal components need to be removed.

The nonzero eigenvalues of the correlation structure are the S_{ii}^2 values, and if R were completely random (i.e., if $F = 0$ in Equation 3), then the histogram of these values would be close to an MP density when N and T are large. However, from Exhibit 1, it is clear that at least two values are separated visibly from the bulk of $(S_{ii}^2/N)_{1\le i\le T}$. Hence, the rank of F is at least two, which means at least two principal components need to be removed from the data for the remainder to be considered noise.

For some $d \le T$, the best rank-d estimator of F that minimizes the Frobenius norm of error is

$$\hat{F} = \sum_{i\le d} S_{ii} U_i V_i^*,$$

with residual $\tilde{R} = R - \hat{F}$. The true rank of F is greater than d if a statistical test of the residual rejects the hypothesis of purely random entries in $\tilde{R}$. A Kolmogorov–Smirnoff (KS) test is likely to reject this hypothesis if d is too small. We devise a simple KS test using the data's own estimated MP distribution; that is, we use the estimated asymptotic distribution parameters obtained from $(S_{ii}^2/N)_{d<i\le T}$ and then check for significance of the associated KS statistic, as we now describe.

The MP density is

$$\nu(x) = \frac{1}{2\pi\gamma^2}\frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{\lambda x} \quad \text{for } \lambda_- \le x \le \lambda_+ \qquad (6)$$

where

$$\lambda_\pm = \gamma^2(1 \pm \sqrt{\lambda})^2 \text{ and } \lambda > 0. \qquad (7)$$

If R were a random $N \times T$ matrix with independent identically distributed entries of mean zero and

variance γ^2, then for $N > T$ the RMT tells us that the empirical spectral distribution of the covariance $\frac{1}{N}R^*R$

$$\frac{1}{T}\sum_{i=1}^{T}\mathbf{1}_{\{S_{ii}^2/N\le x\}},$$

converges in probability (pointwise in x) to the distribution of the MP density in Equation 6 as N and T tend to infinity with $\lim\frac{T}{N}=\lambda\in(0,1)$ fixed. Conversely, the spectral distribution of $\frac{1}{T}RR^*$ is the case of $\lambda > 1$, wherein the limit law has an additional discrete mass at zero with weight $1 - 1/\lambda$, which appears because the $N\times N$ covariance has rank at most $T < N$, and so there are $N - T$ zero eigenvalues. Because we are interested only in eigenvalues through their empirical spectral density, we can consider the $T\times T$ covariance $\frac{1}{N}R^*R$ for which the dimension ratio $\lambda = T/N$ is less than one and there is no mass at zero in the asymptotic MP law.

The issue with the IVS data matrix R and its residual $\tilde{R} = R - \hat{F}$ is that we are not dealing with matrixes with independent identically distributed entries. We are, in fact, very far from it, and so it is not at all clear that the empirical spectral density of the residual matrixes will be close to the MP law. There are significant correlations among the entries of the residual matrixes, even without addressing the normalization issue. As noted, there is considerable theory on separating the bulk spectrum from the spike eigenvalues for idealized random matrix spike models; we may also cite the survey by Johnstone and Paul (2018) and, in dealing with normalization issues, El Karoui (2008). The theoretical criteria provided in the literature do not work with the IVS data, as expected. Writing a data matrix as a factor matrix plus a residual so that the residual is noise, or has no useful information, is a problem that arises often and in many different disciplines—not only with financial data but also, for example, in imaging in materials science (Berman 2019). With real data, this is almost always treated with a variety of empirical estimation methods whose validity is assessed on the basis of the results produced in specific applications.

For the IVS data, we will fit the empirical spectrum of the residuals to the MP law by matching supports, as we now describe. The quality of the fit is quantified by a KS test. The main result of this empirical fit, which works well for the IVS data, is to extract an estimated dimension ratio $\hat{\lambda}$ and standard deviation $\hat{\gamma}$ (Exhibit 2).

EXHIBIT 2

Estimated Parameters for the MP Distribution and Effective Dimension after Removal of Nine Principal Components for Each of the 251 or 250 IVS Daily Returns Observed in Each Year

Year	$\hat{\lambda}_+$	$\hat{\lambda}_-$	$\hat{\gamma}$	$\hat{\lambda}$	$\tilde{N}$
2012	1.77	0.07	0.80	0.44	573
2013	1.93	0.10	0.85	0.40	633
2014	2.15	0.07	0.86	0.49	519
2015	1.80	0.06	0.80	0.47	537
2016	1.98	0.06	0.83	0.50	506
2017	2.33	0.11	0.93	0.42	601

Notes: The data points are $X_i = S_{ii}^2/N$ for $i > 9$, the estimates are given by Equations 8 and 9, and the effective dimension is $\tilde{N} = T/\hat{\lambda}$. Note in particular that $\tilde{N} \ll N \sim 25{,}000$.

Let $X_i = S_{ii}^2/N$ be the eigenvalues (normalized) of the IVS data matrix R (normalized). Given the number of factors d that we want to retain, the estimators for the support $\lambda_\pm$ of the empirical spectral density, or histogram, of the residual that we use are

$$\hat{\lambda}_+ = X_{d+1} \quad \text{and} \quad \hat{\lambda}_- = X_T, \tag{8}$$

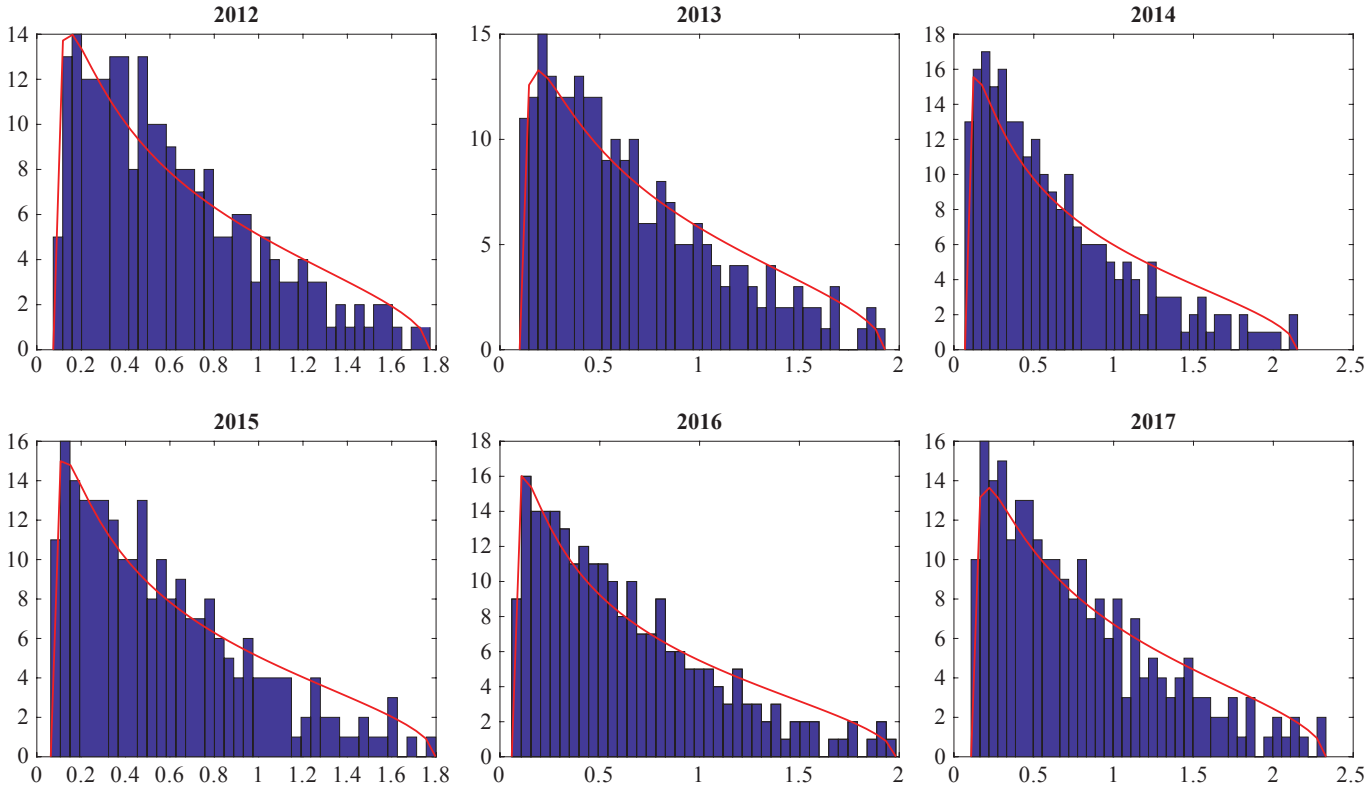
which, using Equation 7 with $\lambda \in (0,1)$, gives

$$\hat{\gamma} = \frac{\sqrt{\hat{\lambda}_+}+\sqrt{\hat{\lambda}_-}}{2} \quad \text{and} \quad \hat{\lambda} = \left(\frac{\sqrt{\hat{\lambda}_+}-\sqrt{\hat{\lambda}_-}}{2\hat{\gamma}}\right)^2. \tag{9}$$

The fitted MP densities to the empirical spectral densities of the residuals with nine components removed are shown in Exhibit 3 for each of the years from 2012 to 2017. A two-sample KS test does not reject any of these years, and increasing d to 10, 20, 30, and 50 continues to result in a visibly good fit and nonrejection by the KS test. Hence, we conclude that nine factors is typically sufficient to describe the daily systematic movements among all IVSs over a single year of daily data. In contrast, for equity returns for the S&P 500 constituents, (roughly) 20 factors are typically required to account for the majority of daily movement, with the number dropping below 10 during the 2008 financial crisis (see Avellaneda and Lee 2010).

EXHIBIT 3

MP Density Fitted to Histogram of Remaining $T - d$ Squared Singular Values over N, 2012–2017



Notes: For each of the years from 2012 to 2017, we removed the top $d = 9$ principal singular values and fitted an MP density to the histogram of the remaining $T - d$ squared singular values over N (i.e., S_{ii}^2/N for $i > d$). A two-sample Kolmogorov–Smirnoff test does not reject any of these years. Removal of more than nine components also leads to nonrejection of the hypothesis that the residual matrixes $\tilde{R}$ have purely random entries.

A direct comparison with work by Avellaneda and Dobi (2014), wherein a cutoff for purely random entries was determined to be around 108 factors (out of ~3,000 names), is not possible because the dataset used by Avellaneda and Dobi (2014) includes equity returns normalized using the strike of the at-the-money option. This is a more complex dataset because the vectorized time series of IVS plus equities (of size 28,000 + 500 in the context of this article) is quite heterogeneous, and therefore the "low-rank plus random" decomposition needs additional attention.

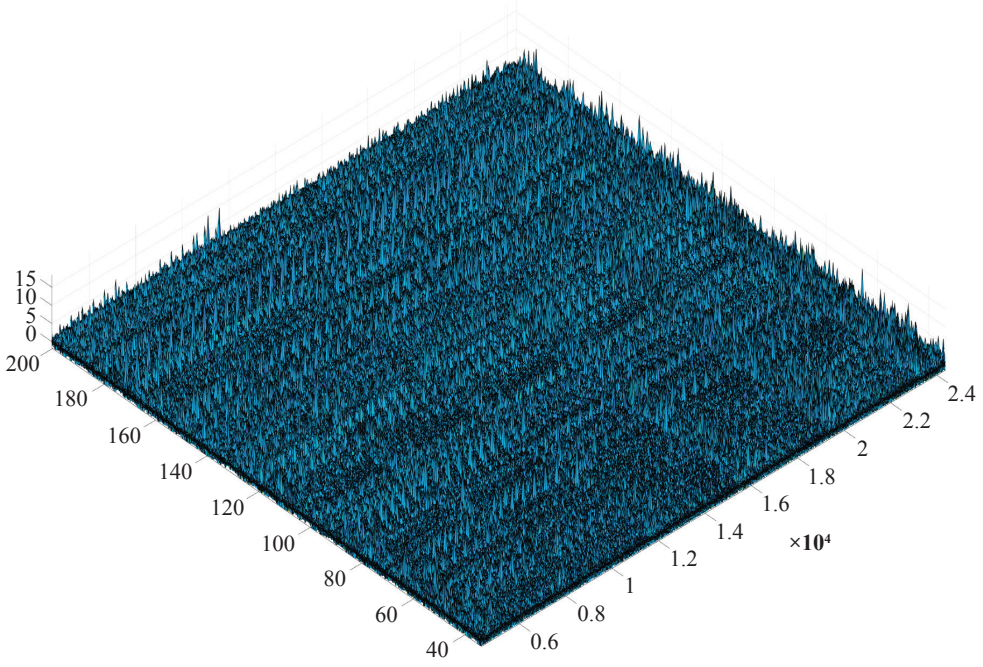
In this article, the nine factors (out of ~500 names for IVS) describe the systematic movements. Compared with equity returns by themselves, the number of factors is of comparable size for the different years between 2012 and 2017, and $d = 20$ factors are needed for the S&P 500 constituents to produce a random residual.

Effective Dimension in Residual Matrixes

We will now introduce the notion of effective dimension of the data as follows. The parameter estimates given in Equations 8 and 9 for the dimension ratio $\lambda = T/N$ and the data standard deviation γ imply that the empirical spectral density of the data fits an MP distribution associated with a random matrix whose dimensions are different from those of the actual data matrix R. This is because the IVS data has entries with significant correlations between them, and the effective dimension idea is a way to account for or quantify this feature. We could say roughly that the residual data matrix does not have independent entries, but it does have some kind of independence when the data are grouped in blocks and the number of such blocks plays the role of an effective dimension. This assessment is made using the empirical spectral density of the residual.

EXHIBIT 4

Nonrandom Structure Seen in the Residual Matrix $\tilde{R}$ Even after Sufficient Components Have Been Removed for Nonrejection by KS Test



Notes: The horizontal coordinate on the right is the IVS name, delta, and maturity (vectorized), and that on the left is time. This patterned structure suggests that the spatial modes are nonrandom and that the randomness we have concluded from the KS test is due to randomness in the temporal loadings. Random temporal loadings is precisely what is suggested in Equation 10.

Denoted by $\tilde{N} = T/\hat{\lambda}$, we call this the effective dimension associated with a residual that is purely random. In other words, the N-dimensional columns of U_{d+1}, U_{d+2}, … and U_T do not affect the spectrum, and we have

$$\frac{1}{N}\tilde{R}^*\tilde{R} = \frac{1}{N}\sum_{i=d+1}^{T} S_{ii}^2 V_i V_i^* \overset{\mathcal{D}}{\approx} \frac{1}{\tilde{N}} Y^* Y, \tag{10}$$

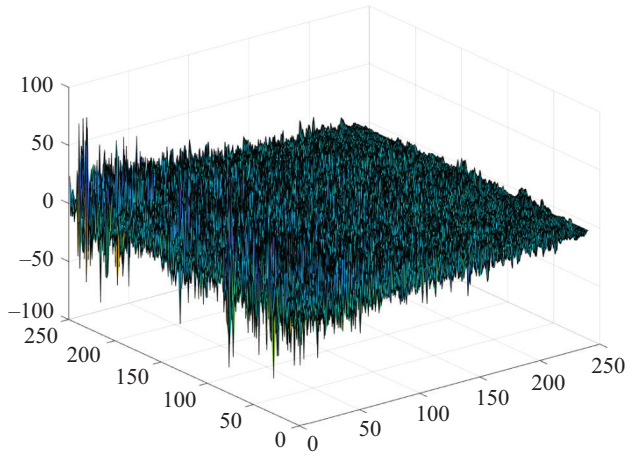
where Y is an $\tilde{N} \times T$ matrix with purely random entries and "$\overset{\mathcal{D}}{\approx}$" denotes approximate equality in distribution, in the sense of the fitting of the empirical spectral density to the MP law that was described in the previous section. Hence, we are looking for pure randomness in the temporal loadings, and we are not concerned if nonrandom structure remains in the higher-order spatial components. In fact, for the implied volatility data there are clear patterns of nonrandom structure in the residual $\tilde{R}$ even for d large enough for nonrejection of the randomness hypothesis, as is clearly seen in Exhibit 4 for the 2017 data. The reason for the nonrandom patterns is simple: The implied volatility's four-dimensional tensor structure was flattened into a two-dimensional $N \times T$ matrix using a lexicographical ordering for vectorizing the IVS data (name, strike, and maturity) that prevails even after removal of PCA factors. This patterned structure does not mean we have incorrectly concluded there is pure randomness in the residual as seen by the empirical spectral density; rather, it suggests that the spatial modes (name, strike, and maturity) retain some of their structure and that the randomness we have found from the KS test is due to randomness in the temporal loadings. Indeed, randomness of temporal loadings is precisely what is suggested by the approximate distributional equivalence expressed in Equation 10.

Denote the tth column of R as R_t, which we can write as

$$R_t = \sum_{i=1}^{d} f_i \theta_{it} + \sum_{i=d+1}^{T} S_{ii} U_i V_{it},$$

EXHIBIT 5

Matrix $S\tilde{V}^*$ of Temporal Loadings for the 2017 Data of Dimension $(T-d)\times T$ with $d=9$



Note: The distinctive tapering is not an indication of nonrandomness in the residual but instead says that $S\tilde{V}^*/N$ is close in distribution to ΣQ^*, where $Q\Sigma Q^*$ is the spectral form of another matrix $Y^*Y/\tilde{N}$ and Y is an $\tilde{N}\times T$ matrix with purely random entries of variance $\hat{\gamma}^2$.

where θ_{it} and V_{it} are the tth entry of θ_i and V_i, respectively; U_i is the ith column of U. Clearly θ_{it} are the temporal loadings on the ith principal factor, and for $i > d$ the temporal loadings are $S_{ii}V_{it}$. Denote the higher-order temporal modes as $\tilde{V} = [V_{d+1}, \ldots, V_T]$. The KS test has indicated randomness of these loadings for $i > d$, which means the $(T-d)\times T$ matrix $S\tilde{V}^*$ has a covariance spectrum close in distribution to that of a random matrix. However, this matrix has a distinctive tapering that would counter any claim of pure randomness. Indeed, Exhibit 5 shows this tapering for the 2017 data for S_{ii}^2/N with $i > 9$ (i.e., with $d = 9$). However, if one considers a purely random matrix Y of dimension $T\times\tilde{N}$ with entries having variance $\hat{\gamma}^2$ and expresses it in spectral form such that $Y^*Y/\tilde{N} = Q\Sigma Q^*$, then ΣQ^* is a tapered matrix that is derived from a purely random matrix. Moreover, the squared singular values of ΣQ^* fit the MP law and are equal in distribution to the squared singular values of $S\tilde{V}^*/N$.

The notion of effective dimension as we have introduced it here is robust and useful in understanding the real information content of the IVS data. Note in particular that in Exhibit 2, which has the parameter estimators for the MP fit as well as the effective dimension for each of the six years we have considered, a striking finding is that we have a very low effective dimension for each year (i.e., $500 \sim \tilde{N} \ll N \sim 25{,}000$). The effective dimension of the IVS residuals in a factor decomposition indicates that there is structure in these residuals; they are not purely random matrixes.

PRINCIPAL EIGENPORTFOLIOS AND OI-WEIGHTED INDEXES

The portfolio constructed from the first eigenvector of the normalized covariance, the eigenportfolio, has been analyzed by Avellaneda and Lee (2010) and Boyle (2014) and others for equities returns. In this section, we will apply a similar approach and terminology to construct a portfolio of implied volatilities and compare it with an analog of a market portfolio. Although such portfolios can exist in theory, they are not directly[1] tradeable. However, it is still of interest to introduce them and use them as a proxy for the market's collective volatility, similar to the VIX index.

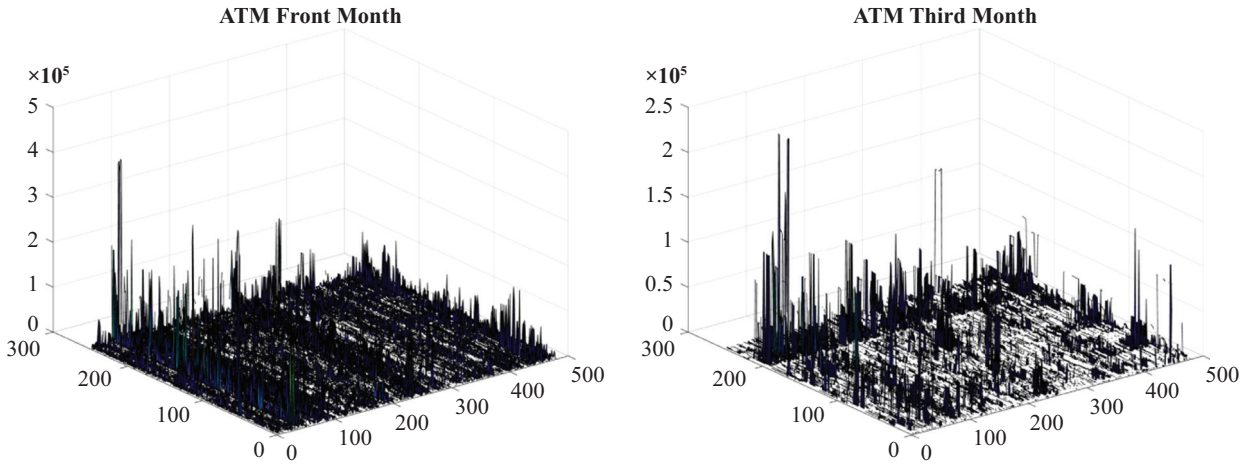
Central to our analysis of the eigenportfolio is its closeness to an OI-weighted factor of our construction. Analogous to the case in equities wherein the eigenportfolio's returns track the capitalization-weighted market portfolio (with varying degrees of closeness), we show that the implied-volatility eigenportfolio's returns track closely to various OI-weighted factors or indexes. It is essential to include the OI because it is the implied volatilities' analogue to capitalization, thereby capturing each option contract's importance.[2] In the analysis of both implied volatilities and equities, the underlying spike model (see Benaych-Georges and Nadakuditi 2011) provides a mathematical framework for the closeness of the eigenportfolio and factor returns. Exhibit 6 shows daily OI amounts for at-the-money options for 2017. It is important to note the spikes and the zeros in the OI. Most options have zero open trades at any point in time, which is an indication that a rather small percentage of options can explain changes in implied volatilities.

[1] A portfolio can be constructed with returns equal to a linear combination of implied volatilities returns (i.e., a linear combination of the $r_i(t)$ values). However, such a portfolio would probably incur a roll yield owing to the daily rebalancing required to maintain a constant maturity and constant delta position.

[2] In practice, the CBOE's VIX calculation uses OI as well. Specifically, the VIX construction includes only SPX options that are between the at-the-money mark and the last strike before the first two consecutive strikes that have zero OI.

EXHIBIT 6

Daily Values of OI for the At-the-Money Options with 30 Days to Maturity (front month) and 90 Days to Maturity (third month)



Notes: The horizontal coordinate on the left is time and that on the right is option name. Note the spikes and the zeros in the OI, which indicate that only a small percentage of options can explain changes in implied volatilities because most options do not have open trades any on given day.

Trading-Implied Volatility and Construction of OI-Based Factors

It is not only OI, however, that must be taken into consideration in constructing IVS returns portfolios. We must also account for the sensitivity of IVS to fluctuations as measured by the Vega, for example. We can motivate this sensitivity in portfolios we are about to create by first introducing a synthetic market of exchange-traded notes (ETNs). For each option we can think of an ETN whose prospectus states the daily returns to be

$$\frac{dE_i(t)}{E_i(t)} := r_i(t), \tag{11}$$

where $r_i(t)$'s are the components of the implied volatility return vector defined in Equation 1 and $dt = 1/252 = 1$ day. The ETN whose returns are given by Equation 11 is the stochastic component in the returns of a Δ-neutral options position. Indeed, letting $C_i(t)$ denote the option price with underlying price, time to maturity, and delta $(S_i(t), \tau_i, \Delta_i)$, from Itô's lemma we compute the (unitless) differential of a Δ-neutral position (up to a term of size Big-Oh),

$$\begin{aligned}\frac{dC_i(t) - \Delta_i dS_i(t)}{S_i(t)} &= \frac{\mathcal{V}_i(t)}{S_i(t)} d\hat{\sigma}_i(t) + O(dt) \\ &= \mathcal{V}_i^{untls}(t) r_i(t) + O(dt),\end{aligned} \tag{12}$$

where $\mathcal{V}_i(t)$ is the Vega for the ith option and

$$\mathcal{V}_i^{untls}(t) = \frac{\hat{\sigma}_i(t)\mathcal{V}_i(t)}{S_i(t)}, \tag{13}$$

is a unitless Vega, which is the dollar-Vega divided by the price of the underlying. We will use Equation 13 in constructing weighing factors, in addition to OI, after we first discuss them more generally.

We would like to construct a global factor that can describe upward of 50% of daily variance for all of the IVS ETNs. In equities, such a factor is the market portfolio, which suggests to us that the number of outstanding shares (contracts) should have some bearing on the relevance of an individual equity in factor construction. Indeed, in options this is precisely the OI, and a general form for a global factor with only OI weighing is

$$\frac{dQ(t)}{Q(t)} := \frac{\sum_i \omega(\mathcal{OI}_i(t)) r_i(t)}{\sum_i \omega(\mathcal{OI}_i(t))}, \tag{14}$$

where $\mathcal{OI}_i(t)$ is the OI for the ith option and $\omega(\cdot)$ is a weighting function of our choosing. In the simplest case we have $d = 1$ in Equation 4, and the ETN returns have a simple factor-based returns model

$$\frac{dE_i(t)}{E_i(t)} = \beta_i \frac{dQ(t)}{Q(t)} + \xi_i(t), \tag{15}$$

where ξ_i is an idiosyncratic noise component independent of $Q(t)$. Ordinary least squares regression shows us that the β_i's are given by the covariance with the factor

$$\beta_i = \text{cov}\left(\frac{dE_i}{E_i}, \frac{dQ}{Q}\right)/h_q^2, \tag{16}$$

where $h_q^2 = \text{var}(\frac{dQ}{Q})$

Remark 1. Boyle (2014) explained how the principal eigenportfolio is a frontier portfolio if the Perron–Frobenius theorem applies. However, for options, the frontier/CAPM theory does not apply because the lifetime of an option is too short. Therefore, comparisons with equities are merely an informal, statistical analogy.

The Spike Model and the Principal Eigenportfolio

Let us consider the $N \times N$ empirical covariance matrix for the returns of the synthetic ETNs

$$\hat{\Sigma}_{ij} := \frac{1}{T-1}\sum_{t=1}^{T}(r_i(t) - \overline{r_i})(r_j(t) - \overline{r_j}) \quad \text{for} \quad 1 \le i, j \le N.$$

Using the returns model in Equation 15, the population covariance matrix is

$$\Sigma = h_q^2 \beta\beta^* + \Omega, \tag{17}$$

where Ω is the covariance matrix with $\Omega_{ij} = cov(\xi_i(\cdot), \xi_j(\cdot))$, and $h_q^2 = \text{var}\left(\frac{dQ(\cdot)}{Q(\cdot)}\right)$. Equation 17 is a spike model, as referred to earlier, because the β's describe a substantial portion of variance and cause a single eigenvalue to stick out from the rest of the spectrum. The first principal component of the ETN empirical covariance matrix will be nearly proportional to the β's of the OI-weighted portfolio if the $\xi(t)$ covariance is not too large. For the model in Equation 3, the distribution of the empirical matrix principal component is shown by Benaych-Georges and Nadakuditi (2011) to be within a cone surrounding the spike model's low-rank component if the difference between $||\theta||^2$ and the variances of the noise is over a critical amount. For the spike model in Equation 17, the critical threshold is crossed if $h_q^2 ||\beta||^2$ exceeds a threshold determined by the covariances of the $\xi_i(t)$'s, which should happen as N grows.

In finance, there usually are differing sizes among the Σ_{ii}'s, which means better statistical estimation of principal eigenvectors results from consideration of correlations rather than covariances; this is the normalization issue we noted earlier. The ETNs' empirical correlation matrix is

$$\hat{\rho} = h^{-1}\hat{\Sigma}h^{-1},$$

where h is a diagonal matrix of standard deviations h_i defined in Equation 2. Letting u_1 denote the principal eigenvector of $\hat{\rho}$, the spike model suggests $u_1 \approx \frac{1}{c}h^{-1}\beta$ for c, a normalizing constant. Using another (orthogonal) eigenvector $\tilde{u}$ such that $\tilde{u} \perp u_1$, we can also construct portfolios as done by Avellaneda and Lee (2010):

$$\pi_1 = \frac{h^{-1}u_1}{\sum_i (h^{-1}u_1)_i}, \quad \tilde{\pi} = \frac{h^{-1}\tilde{u}}{\sum_i (h^{-1}\tilde{u})_i},$$

which are orthogonal in the sense that covariance of these portfolios' returns is zero

$$\pi_1^*\hat{\Sigma}\tilde{\pi} = u_1^*\hat{\rho}\tilde{u} = \lambda_1 u_1^*\tilde{u} = 0,$$

where $\lambda_1 > 0$ is the principal eigenvector such that $\hat{\rho}u_1 = \lambda_1 u_1$.

Proposition 0.1. Returns of the top eigenportfolio tend toward the factor returns plus some tracking error. Returns of the orthogonal portfolios tend toward factor neutrality.

Proof. Assuming the parameters are such that we are over the critical levels noted by Benaych-Georges and Nadakuditi (2011) and letting $EP_1(t)$ be the principal eigenportfolio, we have

$$\begin{aligned}
&\frac{dEP_1(t)}{EP_1(t)} - \frac{\sum_i (\beta_i/h_i)^2}{\sum_i \beta_i/h_i^2}\frac{dQ(t)}{Q(t)} \\
&= \left(\sum_i \left(\pi_{1i} - \frac{\beta_i/h_i^2}{\sum_j \beta_j/h_j^2}\right)\beta_i\right)\frac{dQ(t)}{Q(t)} + \sum_i \pi_{1i}\xi_i(t) \\
&\to \varepsilon_1(t),
\end{aligned}$$

where factor returns disappear because $\pi_{1i} - \frac{\beta_i/h_i^2}{\sum_j \beta_j/h_j^2} \to 0$ as the random matrix's dimensions grow, and where

$\varepsilon_1(t) = \lim_N \Sigma_i \pi_{1i} \xi_i(t)$ is the tracking error. All orthogonal portfolios $\tilde{\pi}$ are approximately factor neutral:

$$\frac{d\widetilde{EP}(t)}{\widetilde{EP}(t)} = \underbrace{\left(\sum_i \tilde{\pi}_i \beta_i\right)}_{\approx 0} \frac{dQ(t)}{Q(t)} + \sum_i \tilde{\pi}_i \xi_i(t) \to \tilde{\varepsilon}(t),$$

where $\tilde{\varepsilon}(t) = \lim_N \Sigma_i \tilde{\pi}_i \xi_i(t)$, and where the limit happens because $(h^{-1}\tilde{u})^* \beta \approx \tilde{u}^* h^{-1} h u_1 = 0$ becomes more accurate and tends toward an equality as the gap between $||h^{-1}\beta||$ and the critical level increases. □

Empirical Analysis

The first spatial[3] singular vector U_1 computed by the SVD in Equation 5 is the empirical estimator for u_1, and the estimator for each h_i is the empirical standard deviation of each r_i. Hence, we can compute the eigenportfolio from the data and then compare it with the empirically estimated β's. In our studies we will consider two weighting functions

$$\omega(\mathcal{OI}) = \mathcal{OI} \quad \text{or} \quad \omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}, \quad (18)$$

where $\mathcal{V}^{untls}$ denotes the unitless Vega of Equation 13. Generally speaking, $\omega(\mathcal{OI}) = \mathcal{OI}$ results in a factor with a less significant intercept in ex post regressions of eigenportfolio returns onto the factor returns, whereas $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}$ results in the same regression having a significant intercept but lower projection error. The plain OI weighting is a bit strange, however, because it counts contracts without taking into account the sensitivity of the contract to a change in the volatility of the underlying stock. It should also be noted that the log weighting has a factor loading that is closer to unity, whereas the plain OI weighting is significantly less than unity.

The OI and unitless Vega weighing does, in fact, matter, and it performs better when the tensor data structure is taken into consideration. We have chosen the combined OI and unitless Vega weighing in the form $\log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}$, which works well for the IVS data, although it is rather arbitrary at present.

In Exhibit 7, we see the comparison for each of the years, with each year's eigenportfolio computed using the 251 or 250 days of data from the year and with a $Q(t)$ factor computed using the weight function $\omega(\mathcal{OI}) = \mathcal{OI}$. If the eigenportfolio's weights are sorted in descending order (i.e., we sort π_1 in descending order) and the sorting index then is used to permute the vector $\text{diag}^{-2}(h)\beta / \Sigma_i (\text{diag}^{-2}(h)\beta)_i$, then the sorted vector and the permuted vector should line up. Indeed, the plots in Exhibit 7 show this lining up, with the sorted eigenportfolio in red and the permuted β's in blue.

We can also check the name, the tenor, the Δ, the β_i / h_i^2's, and the OI for the top-weighted options. These traits are listed in Exhibit 8 for the top 32 options in the 2012 sorting. It is interesting to note that most of the top options are out-of-the-money put options, all with 365 days to maturity (the longest-dated options in the dataset). Generally speaking, long-dated options have higher Vega[4] and therefore are most sensitive to changes in implied volatility. The interpretation of long-dated options dominating the first eigenportfolio is simple: The first eigenportfolio explains the most systematic movements among the options and should be the least sensitive to idiosyncratic noise; it therefore ignores short-dated options that may fluctuate idiosyncratically because of short-lived risk events. The years 2013–2017 had similar characteristics in the top 32 eigenportfolio options.

We also perform checks to make sure eigenportfolios can track in an online setting; that is, all quantities used to compute a portfolio in real time are adapted to the filtration generated by the options data (Exhibit 9).[5] We group all six years of data into one simulated run and then compute an eigenportfolio each month using the previous six months' daily returns. The returns are compounded daily, but we only update the portfolio weights monthly. Exhibit 9, Panels A–D, show the

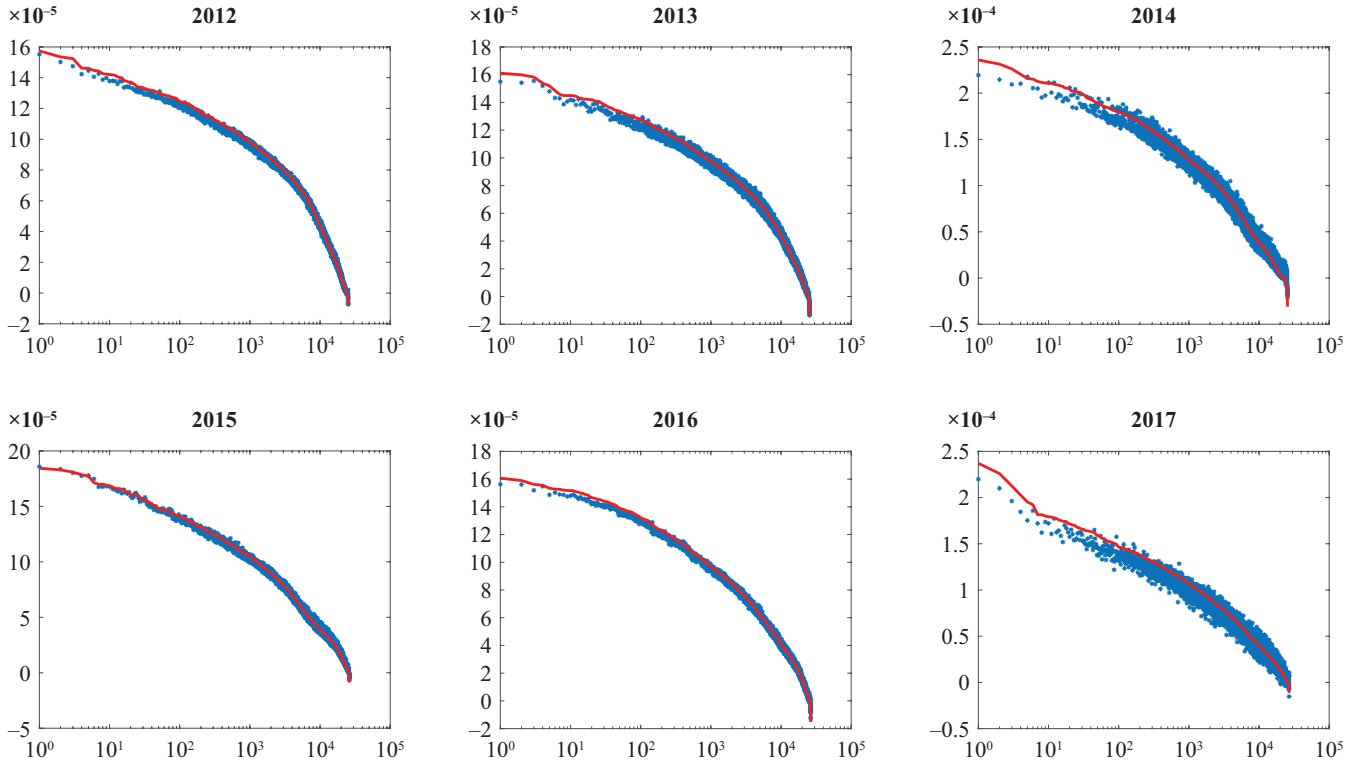
[3] That is, of the IVS vector of names, deltas, and maturities.

[4] The Black–Scholes call/put option Vega is $\mathcal{V}(t) = SN'(d_1)\sqrt{\tau}$, where τ is the time to maturity. Hence, all other things being equal, longer-dated options have higher Vega.

[5] The online eigenportfolio that we compute has one major anticipatory element: survivorship bias. We selected names of the S&P 500 constituents from 2017 and collected their options' data going back to 2012. Hence, there is some survivorship bias in favor of names that performed well enough to stay in the index for these six years. We acknowledge this bias and realize that it will persist over time, but for our analysis of daily statistics it does not make our findings any less valid. Moreover, the OI weighing favors liquid options for which survivorship bias is likely to be reduced.

EXHIBIT 7

Eigenportfolio Computed from the Data (red) Compared with the Theoretical Eigenportfolio That Is Close to β_i / h_i^2 (blue)



Notes: The horizontal axes are in log scale. Each $\beta_i = \mathrm{cov}(dE_i(\cdot)/E_i(\cdot), dQ(\cdot)/Q(\cdot))/h_q^2$*, where for these plots the factor Q(t) has been computed according to the formula in Equation 14 using the weighting function* $\omega(\mathcal{OI}) = \mathcal{OI}$*. To generate this plot, we first sort the eigenportfolio in descending order and then insert the sorting index into the vector* $h^{-2}\beta / \sum_i (h^{-2}\beta)_i$*. The lining up of the two vectors using a single sorting index is evidence that the factor computed using the weighting function* $\omega(\mathcal{OI}) = \mathcal{OI}$ *is close to the data's principal component.*

online six-month sliding window performance of the eigenportfolio alongside the factor returns and the VIX returns. Note in Exhibit 9, Panel B, that the eigenportfolio returns have significant explanatory power for the returns on VIX (i.e., $d\mathrm{VIX}(t)/\mathrm{VIX}(t)$) but also that returns for the OI factor and eigenportfolio have a much stronger linear dependence. Indeed, even if we use the weighting function $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}$, which allows the OI factor to have better tracking with the VIX (see Exhibit 9, Panel D), the stronger linear dependence in daily returns between factor and eigenportfolio still prevails.

Lastly, some discussion on performance with the two different weighting functions is in order. The visual evidence in Exhibits 7 and 9 should sufficiently convince the reader that OI needs to be included in factor construction. The discussion that remains is for us to decide how to determine the OI-weighted factor that is somehow best. Exhibit 10 provides evidence indicating that $\omega(\mathcal{OI}) = \mathcal{OI}$ is good because it leaves the least amount of unexplained systematic return after regression of the eigenportfolio onto the factor's returns (as well as two other factors). However, visually from Exhibit 9, Panel D, we see that $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}$ produces a factor that is better for tracking the VIX; the VIX is the US market's premier volatility index, but this choice of ω leaves a significant amount of unexplained systematic return after the regression. Bear in mind that leaving unexplained systematic return is not entirely bad because we have thus far only been able to construct a single factor for explained implied volatility movements, but we know that we need at least two (recall Exhibit 1,

EXHIBIT 8

Ordering of Names in Top 32 Slots in Eigenportfolio of S&P 500 Constituents' Implied Volatility for 2012

Ticker	Maturity (days)	Δ	β/h^2	OI (average)
ADSK	365	–20	3.5332×10^3	2,751
KLAC	365	–20	3.4206×10^3	3,962
DHR	365	–20	3.3575×10^3	18,017
KLAC	365	–30	3.2402×10^3	3,906
LRCX	365	–30	3.2904×10^3	4,063
KLAC	365	–40	3.1996×10^3	5,171
LRCX	365	–40	3.2450×10^3	4,002
INTU	365	–40	3.1577×10^3	4,229
INTU	365	–30	3.1832×10^3	5,457
ADI	365	–20	3.1372×10^3	15,907
XLNX	365	–20	3.1402×10^3	4,237
FLR	365	–20	3.1365×10^3	19,285
ADI	365	–30	3.0942×10^3	13,276
ADSK	365	–30	3.1348×10^3	4,018
TXN	365	–20	3.1324×10^3	19,669
DHR	365	–30	3.0415×10^3	23,395
CHRW	365	–40	3.0531×10^3	2,874
CHRW	365	–30	3.0543×10^3	3,044
FLR	365	–30	3.0444×10^3	22,425
ROST	365	–30	3.1108×10^3	5,809
RCL	365	–40	3.0327×10^3	7,943
FDX	365	–30	3.0153×10^3	34,131
RCL	365	–20	3.0019×10^3	24,656
EXPD	365	–20	2.9784×10^3	5,796
LRCX	365	40	3.0237×10^3	14,541
FDX	365	–20	2.9778×10^3	43,925
FLR	365	40	2.9695×10^3	18,146
LRCX	365	30	3.0293×10^3	12,629
ADSK	365	–40	2.9768×10^3	3,314
SBUX	365	–40	3.0220×10^3	39,345
IR	365	–20	2.9369×10^3	4,699
CF	365	–20	2.9476×10^3	72,143

Notes: The maturity for all these top-weighted options is 365 days, the longest-dated options in our dataset. Generally, long-dated options have higher Vega, which means that the first eigenportfolio finds the most systematic explanation for implied volatility surface movements by considering the options with the greatest sensitivity to implied volatility changes.

where it was clear that there are at least two principal components).

It turns out that there is a setting wherein $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}$ is an acceptable factor, but it will require us to move from ordinary flat linear algebra and use tensors in the next section.

FACTORS AND EIGENPORTFOLIOS USING TENSORS

The IVSs have a natural tensor structure with four dimensions: time, name, maturity, and delta (normalized strike). To work with this tensor, we first need to redefine our notation for implied volatility returns from the definition given in Equations 1 and 2. Before we had $1 \le i \le N$, where $N = 500 \cdot 8 \cdot 7 = 28{,}000$ (500 names, eight maturities, seven deltas). Now we have multiple indexes:

$$r_{ijk}(t) = \text{time-}t \text{ implied volatility return for } i\text{th name, } j\text{th maturity and } k\text{th } \Delta,$$

where $1 \le i \le N^{(1)}$, $1 \le j \le N^{(2)}$, $1 \le k \le N^{(3)}$, and $1 \le t \le T$ (i.e., $N^{(1)} = 500$, $N^{(2)} = 8$ and $N^{(3)} = 7$). We standardize the returns and place them in a four-dimensional tensor

$$R = \left[\frac{r_{ijk}(t) - \overline{r}_{ijk}}{h_{ijk}} \right]_{1 \le t \le T,\ 1 \le i \le N^{(1)},\ 1 \le j \le N^{(2)},\ 1 \le k \le N^{(3)}}, \qquad (19)$$

where $\overline{r}_{ijk} = \frac{1}{T}\Sigma_t r_{ijk}(t)$ and $h_{ijk} = \sqrt{\frac{1}{T-1}\Sigma_t (r_{ijk}(t) - \overline{r}_{ijk})^2}$. Similar to the sector-based hierarchical PCA done for equities by Avellaneda (2019), we can define individualized factors for each value of a certain tensor dimension. The maturity dimension and the delta dimension are the two candidates for individualized factors; the following subsection demonstrates the advantage of constructing the individualized factors.

Eigenportfolios via Multilinear SVD

A tensor analogue for the SVD is the multilinear singular value decomposition (MLSVD) (see De Lathauwer, De Moor, and Vandewalle 2000; Kolda and Bader 2009; Cichocki et al. 2015). The canonical polyadic decomposition (CPD) (see Tucker 1966; Kolda and Bader 2009) can also be used, but for computing principal components with the IVS data, the results produced by these two approaches are very similar. We use the MLSVD, which is briefly described in the Appendix, as follows.

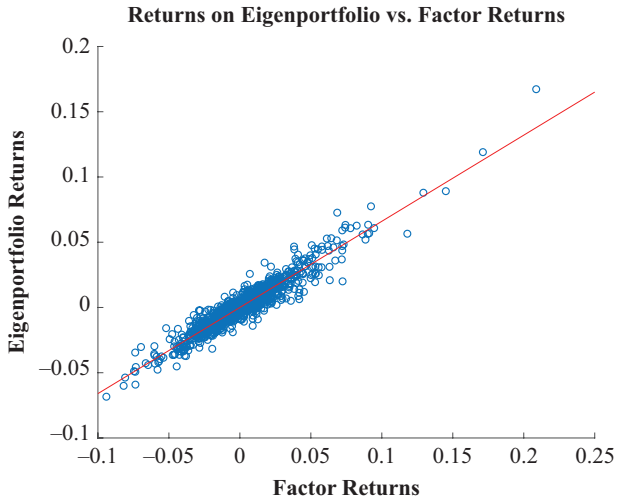
We can write the tensor R in the form

$$R = \sum_{t,i,j,k} S_{tijk}\ U_t^{(1)} \circ U_i^{(2)} \circ U_j^{(3)} \circ U_k^{(4)}, \qquad (20)$$

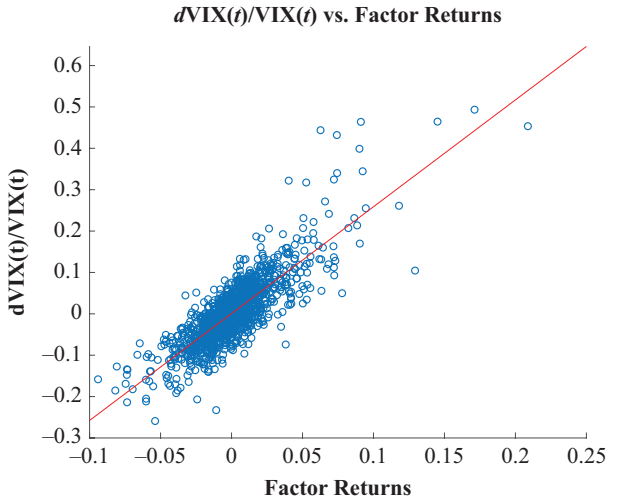
EXHIBIT 9

Output from Adapted Eigenportfolios Using a Six-Month Sliding Window

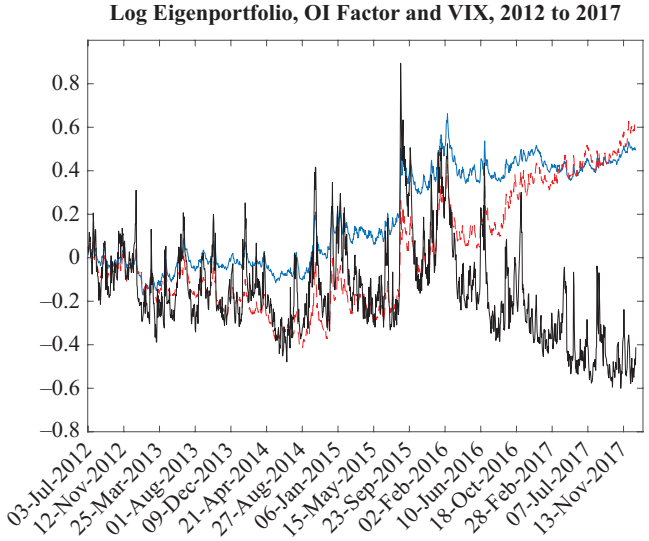
Panel A: Scatter Plot of Adapted Eigenportfolio Returns against Returns of OI-Weighted Factor Computed Using Weighting Function $\omega(\mathcal{OI}) = \mathcal{OI}$ (2012–2017)



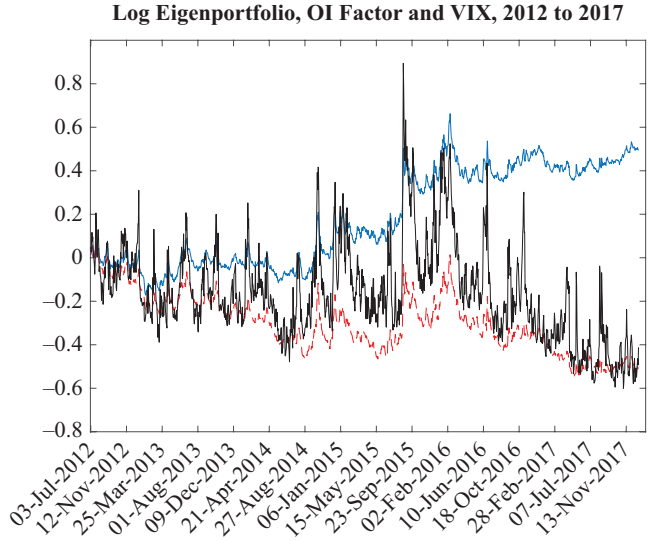
Panel B: Scatter Plot of VIX Returns against Returns of OI-Weighted Factor Computed Using Weighting Function $\omega(\mathcal{OI}) = \mathcal{OI}$ (2012–2017)



Panel C: Adapted Six-Month Sliding-Window Eigenportfolio and OI-Weighted Factor Computed Using Weighting Function $\omega(\mathcal{OI}) = \mathcal{OI}$



Panel D: Adapted Six-Month Sliding-Window Eigenportfolio and OI-Weighted Index Computed Using Weighting Function $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{\text{untls}}$ (logarithm of each index)



Notes: The eigenportfolio is computed using a six-month sliding window, which is why the first six months of 2012 are not included in the output. By taking the OI weights to be the logarithm of 1 plus OI, we see improved tracking of the VIX compared with Panel C. Regression of the eigenportfolio onto the factor has significant intercept coefficient but improved projection error compared with the same regression with $\omega(\mathcal{OI}) = \mathcal{OI}$ (see Exhibit 10).

where $U^{(1)}$ is a $T \times T$ orthonormal matrix, $U^{(2)}$ is an $N^{(1)} \times T$ orthonormal matrix, $U^{(3)}$ is an $N^{(2)} \times N^{(2)}$ orthonormal matrix, $U^{(4)}$ is an $N^{(3)} \times N^{(3)}$ orthonormal matrix, S is a $T \times N^{(1)} \times N^{(2)} \times N^{(3)}$ real-valued tensor, and $\circ$ denotes the vector outerproduct. The main disadvantage of the MLSVD is that the so-called *core tensor* S is, in general, not diagonal or even sparse, which results in difficulties when computing a rank-d decomposition that is best in

EXHIBIT 10

Factor Loadings and Unexplained Systematic Returns for Eigenportfolio from a Six-Month Sliding Window and Various OI Weightings in Factor Construction

$\omega(\mathcal{OI})$	α_3	β_3	α_2	β_2	α_1	β_1
2012–2014						
$\mathcal{OI}$	0.0629 (1.6307)	0.6348	0.0755 (1.8777)	0.7196	0.0908 (2.2614)	0.7187
	$R^2 = 0.9387$		$R^2 = 0.9332$		$R^2 = 0.9322$	
$\log(1 + \mathcal{OI}) \times \mathcal{V}^{\text{untls}}$	0.1224 (4.0669)	0.8947	0.1418 (4.4146)	1.0053	0.1745 (5.2169)	1.0000
	$R^2 = 0.9629$		$R^2 = 0.9574$		$R^2 = 0.9529$	
2013–2015						
$\mathcal{OI}$	0.0570 (1.1356)	0.5500	0.0728 (1.3413)	0.6794	0.0865 (1.5885)	0.6769
	$R^2 = 0.9303$		$R^2 = 0.9184$		$R^2 = 0.9172$	
$\log(1 + \mathcal{OI}) \times \mathcal{V}^{\text{untls}}$	0.1409 (4.1438)	0.8472	0.1664 (4.4377)	0.9842	0.1934 (4.8793)	0.9745
	$R^2 = 0.9681$		$R^2 = 0.9610$		$R^2 = 0.9561$	
2014–2016						
$\mathcal{OI}$	0.0190 (0.3091)	0.5259	0.0116 (0.1746)	0.6600	0.0165 (0.2491)	0.6594
	$R^2 = 0.9183$		$R^2 = 0.9049$		$R^2 = 0.9047$	
$\log(1 + \mathcal{OI}) \times \mathcal{V}^{\text{untls}}$	0.1841 (4.9527)	0.8731	0.2007 (5.1087)	0.9734	0.2184 (5.3412)	0.9687
	$R^2 = 0.9701$		$R^2 = 0.9664$		$R^2 = 0.9634$	
2015–2017						
$\mathcal{OI}$	–0.0505 (–0.8877)	0.5164	–0.0705 (–1.1757)	0.6168	–0.0803 (–1.3413)	0.6170
	$R^2 = 0.9102$		$R^2 = 0.8997$		$R^2 = 0.8992$	
$\log(1 + \mathcal{OI}) \times \mathcal{V}^{\text{untls}}$	0.1705 (6.5261)	0.8926	0.1769 (6.6106)	0.9389	0.1836 (6.8505)	0.9381
	$R^2 = 0.9810$		$R^2 = 0.9800$		$R^2 = 0.9797$	
2012–2017 (all years)						
$\mathcal{OI}$	–0.0010 (–0.0289)	0.5444	–0.0020 (–0.0546)	0.6604	0.0028 (0.0781)	0.6601
	$R^2 = 0.9156$		$R^2 = 0.9033$		$R^2 = 0.9032$	
$\log(1 + \mathcal{OI}) \times \mathcal{V}^{\text{untls}}$	0.1476 (7.2613)	0.8903	0.1619 (7.5857)	0.9760	0.1813 (8.2952)	0.9732
	$R^2 = 0.9697$		$R^2 = 0.9665$		$R^2 = 0.9645$	

Notes: For each sample period there are three regressions: a three-factor regression $\frac{dE^{\mathcal{P}}}{E^{\mathcal{P}}} = \alpha_3 + \beta_3 \frac{dQ}{Q} + b_{eq} \frac{dS\&P}{S\&P} + b_{vix} \frac{dVIX}{VIX} + \varepsilon$*, a two-factor regression* $\frac{dE^{\mathcal{P}}}{E^{\mathcal{P}}} = \alpha_2 + \beta_2 \frac{dQ}{Q} + b_{eq} \frac{dS\&P}{S\&P} + \varepsilon$*, and a one-factor regression* $\frac{dE^{\mathcal{P}}}{E^{\mathcal{P}}} = \alpha_1 + \beta_1 \frac{dQ}{Q} + \varepsilon$*. The factor obtained by weighting function* $\omega(\mathcal{OI}) = \mathcal{OI}$ *is perhaps desirable because it leaves no significant excess return in the residual, but there is no reason why we should reject a factor with nonzero intercept. In contrast, the weighting function* $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{\text{untls}}$ *has significant intercept and a higher* R^2 *and hence is perhaps a better factor.*

the sense of the Frobenius norm. In practice, the rank-1 decomposition used for computing the eigenportfolio can be estimated using the top MLSVD vectors, which we write as

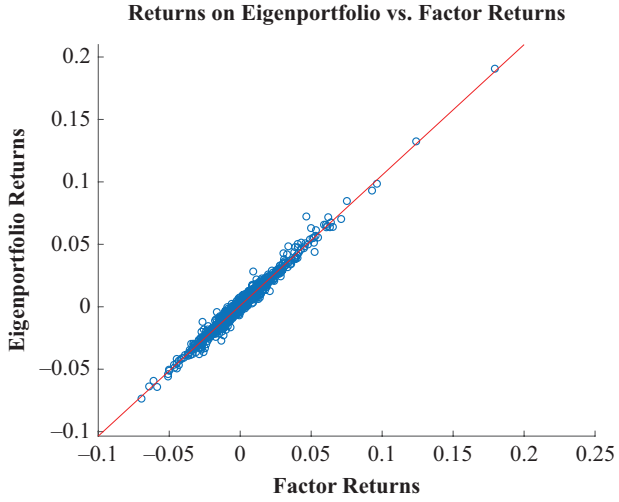
$$\tilde{U}^{(1)} = U_1^{(2)} \circ U_1^{(3)} \circ U_1^{(4)}, \tag{21}$$

which for the IVS data is very close to what one gets using CPD. To compute the tensor eigenportfolio, we

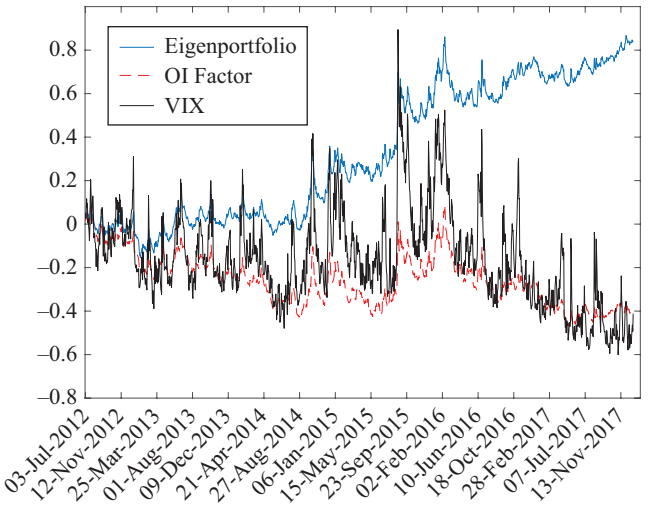
EXHIBIT 11

Output from the Adapted Tensor Eigenportfolio Using a Six-Month Sliding Window

Panel A: Scatter Plot of Adapted Tensor Eigenportfolio Returns against Returns of Tensor OI-Weighted $\bar{Q}$ Factor Given by Equation 24



Panel B: Adapted Tensor Eigenportfolio and Tensor *OI*-Weighted Index Computed Using Weighting Function $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{\text{untls}}$ (logarithm of each index) Log Eigenportfolio, OI Factor and VIX, 2012 to 2017



Notes: Panel A is computed using weighting function $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{\text{untls}}$, from 2012 through 2017. The eigenportfolio is computed using a six-month sliding window. The linear dependence seen here is an improvement from the linear dependence seen in Exhibit 9, Panel A, achieved using flat matrixes. Panel B does not suggest that the tensor approach to eigenportfolio and factor construction offers an improvement to simply using flat matrixes because it is roughly similar to its analogue in Exhibit 9, Panel D. Improvement as a result of tensors is more evident in Panel A and in Exhibits 12, 13, and 14.

do an elementwise division with the tensor of standard deviations:

$$\pi^{(1)}_{ijk} \propto \frac{\tilde{U}^{(1)}_{ijk}}{h_{ijk}}. \tag{22}$$

It remains to be decided how to normalize this $\pi^{(1)}$. If we normalize with one global summation over the multi-index (i, j, k), there would have been no need to use tensors. However, if there is a dimension such that each of the index values has a different corresponding OI-weighted factor, then the normalization needs to be done separately for the different values of this index. The logic for using different normalizations for different index values will become clearer in the following example.

Example: Different Factors for Each Maturity

Suppose that we construct a different OI-weighted factor for each maturity, for a total of eight factors. Moreover, suppose that our factor considers the β for each option to be its loading on the factor corresponding to the same maturity; that is, there are factors Q_j for $j = 1, 2, \ldots, 8$, with $dQ_j(t)/Q_j(t) \propto \sum_{i,k} \omega(\mathcal{OI}_{ijk}(t)) r_{ijk}(t)$, where $\mathcal{OI}$ is the tensor of open interests. The factor model for returns is then

$$r_{ijk}(t) = \beta_{ijk} \frac{dQ_j(t)}{Q_j(t)} + \xi_{ijk}(t),$$

where ξ is a tensor of idiosyncratic noise that is independent of the factor processes in Q_j. In this case, for the β_{ijk}/h^2_{ijk}'s to line up with the eigenportfolio in Equation 22, the normalization needs to be done as follows for each j:

$$\begin{aligned} \pi^{(1)}_{ijk} &= \frac{\tilde{U}^{(1)}_{ijk}/h_{ijk}}{\sum_{i,k} \tilde{U}^{(1)}_{ijk}/h_{ijk}} \\ \frac{dQ_j(t)}{Q_j(t)} &= \frac{\sum_{i,k} \omega(\mathcal{OI}_{ijk}(t)) r_{ijk}(t)}{\sum_{i,k} \omega(\mathcal{OI}_{ijk}(t))}. \end{aligned} \tag{23}$$

EXHIBIT 12

Factor Loadings and Systematic Unexplained Returns for Tensor Eigenportfolio from Six-Month Sliding Window and Various OI Weightings in Tensor-Factor Construction

$\omega(\mathcal{OI})$	α_3	β_3	α_2	β_2	α_1	β_1
2012–2017 (all years)						
$\mathcal{OI}$	0.0891 (3.0439)	0.7384	0.0923 (2.9875)	0.8432	0.0925 (3.0107)	0.8432
	$R^2 = 0.9466$		$R^2 = 0.9404$		$R^2 = 0.9404$	
$\log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}$	0.2092 (11.8532)	0.9673	0.2211 (11.8433)	1.0466	0.2339 (12.3862)	1.0450
	$R^2 = 0.9806$		$R^2 = 0.9782$		$R^2 = 0.9775$	

Notes: For each sample period there are three regressions: a three-factor regression $\frac{dE^{\varphi}}{E^{\varphi}} = \alpha_3 + \beta_3 \frac{d\bar{Q}}{\bar{Q}} + b_{eq} \frac{dS\&P}{S\&P} + b_{vix} \frac{dVIX}{VIX} + \varepsilon$*, a two-factor regression* $\frac{dE^{\varphi}}{E^{\varphi}} = \alpha_2 + \beta_2 \frac{d\bar{Q}}{\bar{Q}} + b_{eq} \frac{dS\&P}{S\&P} + \varepsilon$*, and a one-factor regression* $\frac{dE^{\varphi}}{E^{\varphi}} = \alpha_1 + \beta_1 \frac{d\bar{Q}}{\bar{Q}} + \varepsilon$*. The tensor regressions shown in this exhibit have higher* R^2 *than their counterparts in Exhibit 10, which is an indication that the tensor approach to factor construction is better.*

Finally, the global factor (to compare with the eigenportfolio) is simply the mean of the tenor factors:

$$\frac{d\bar{Q}(t)}{\bar{Q}(t)} = \frac{1}{N^{(2)}} \sum_j \frac{dQ_j(t)}{Q_j(t)}. \tag{24}$$

Exhibit 11 shows improved results if this maturity-wise approach is used with the normalizations in Equation 23 along with the weighting function $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}$. The linear dependence between the tensor eigenportfolio and the $\bar{Q}$ seen in Exhibit 11, Panel A, is a considerable improvement from the linear dependence shown in Exhibit 9, Panel A, which was obtained using flat matrixes. The ex post regression diagnostics of the tensor approach show an improvement over the flat matrix approach; Exhibit 12 lists regression outputs that have higher a R^2 (i.e., less projection error) than their counterparts in Exhibit 10.

Finally, we show the in-sample plots of cumulative returns for the tensor portfolios next to the cumulative returns of the portfolios from the flat matrixes; all portfolios are constructed with weighting function $\omega(\mathcal{OI}) = \log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}$. Exhibits 13 and 14 show these results, from which it is clear that the tensor factor constructed in this subsection allows for eigenportfolio tracking that is much closer to the factor. This is evidence that the family of OI-based tensor factors comprises a good factor, in the sense that it can track the eigenportfolio returns. We recall that we are concerned with the eigenportfolio when constructing factors because we know from theoretical analysis of the spike model that the first eigenportfolio will have weights close to the vector $h^{-2}\beta$, where β are loadings on a dominant factor. Hence, to determine whether we have a dominant factor, we should make comparisons with the eigenportfolio, and in the tensor IVS data context the results come out better, most likely because the data are heterogeneous. It is therefore beneficial to respect the tensor structure, which MLSVD does.

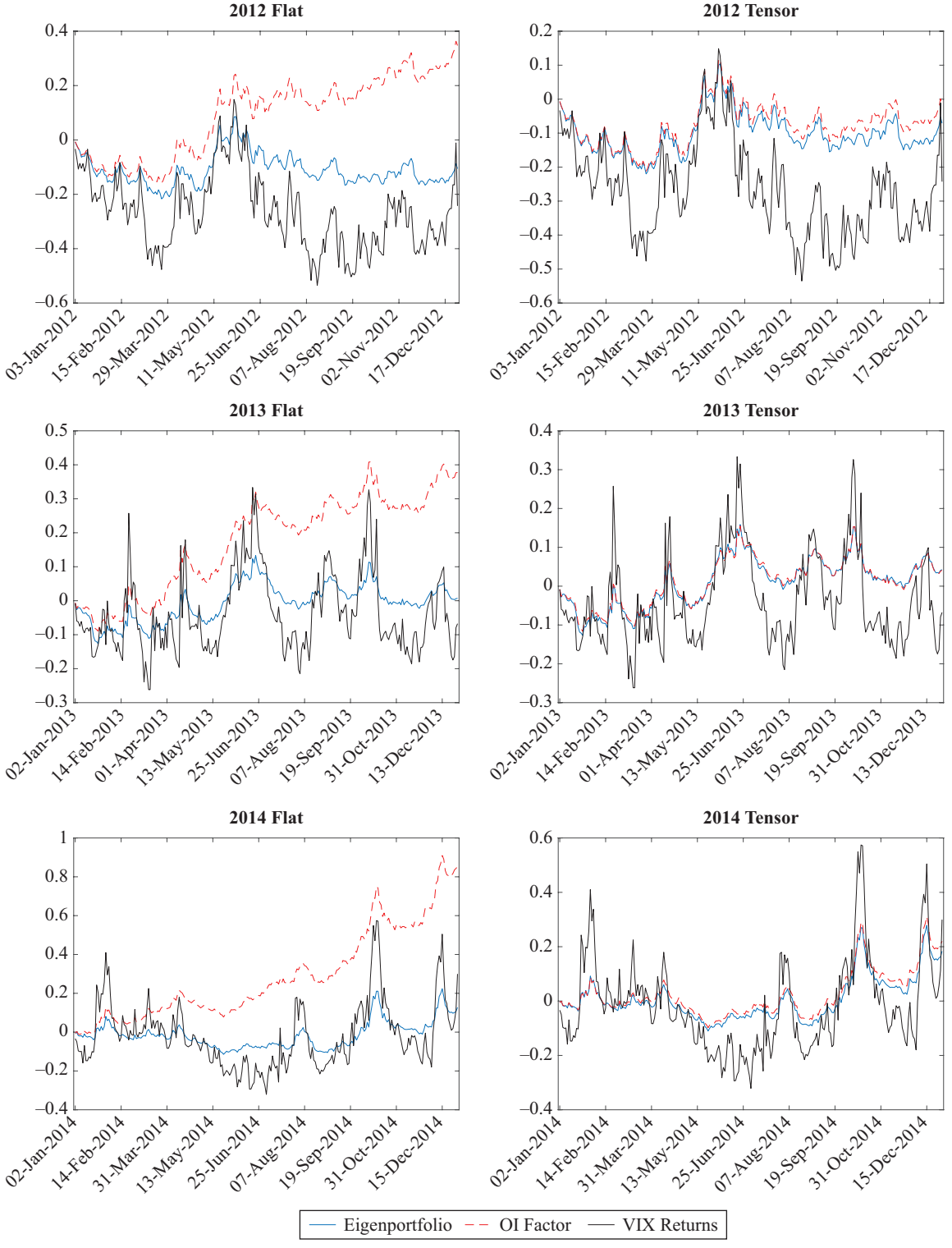
SUMMARY AND CONCLUSION

We have carried out PCA of a dataset of IVSs from options on US equities. The analysis contains daily implied volatilities for the six years 2012–2017 for 500 equities (names), eight maturities, and seven strikes (i.e., 28,000 data points daily for 251×6 days). We have posed and answered three questions in this data-driven study: (1) What is the essential information in this dataset, or how many factors are needed to represent the data with the residual being noise; (2) by analogy with equity-returns PCA analysis, can the principal eigenportfolio be associated with a market portfolio, and what should that market portfolio be; and (3) because the natural structure of the data is that of a four-dimensional tensor (time, name, maturity, strike), do we benefit from preserving this structure in the PCA analysis—that is, by doing a tensor PCA?

The analysis of question 1 is in the section Matrix of Implied Volatility Returns. Building on the work of Avellaneda and Dobi (2014), we perform matrix PCA on the flattened IVS data excluding stock returns and arrive at the conclusion that the number of significant factors is nine. For comparison, the number of significant

EXHIBIT 13

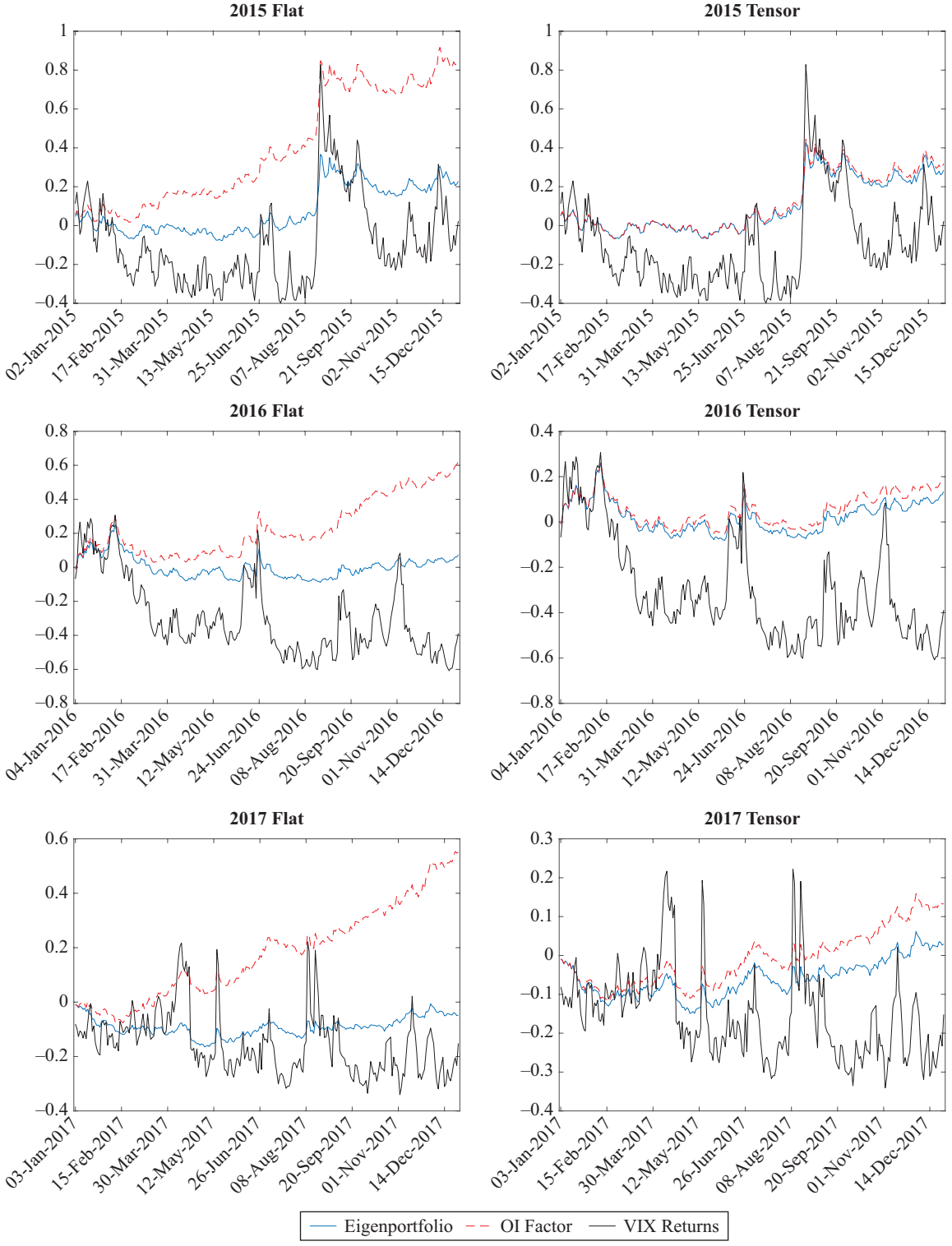
Comparison of the Eigenportfolio Tracking the OI Factor for Both Flat Matrixes (on the left) and Tensors (on the right)



Notes: All portfolios are constructed with weighting function $\omega(\mathcal{OI}) = log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}$. *The plots display the logarithm of each index. These are in-sample fits for each of the years 2012–2014.*

EXHIBIT 14

Comparison of Eigenportfolio Tracking OI Factor for Both Flat Matrixes (on the left) and Tensors (on the right)



Notes: All portfolios are constructed with weighting function $\omega(\mathcal{OI}) = log(1 + \mathcal{OI}) \times \mathcal{V}^{untls}$. *The plots display the logarithm of each index. These are in-sample fits for each of the years 2015–2017.*

factors for equity returns is normally in the range of 15 to 20. There is a very large dimension reduction to the IVS dataset because of the high degree of structure in option prices and hence in the IVS. In this section, we also introduced the concept of effective dimension for the residual because, contrary to theoretical factor analysis and equity returns analysis, the IVS residuals retain structural patterns or correlations. We find that the effective spatial dimension (name, maturity, strike) of the residuals is closer to 500 than to the nominal dimension of 28,000. This is the way spectral analysis (i.e., PCA) deals with the heterogeneity of the IVS data: The residuals have patterns even after the market information has been taken out.

The analysis of question 2 is in the section Principal Eigenportfolios and OI-Weighted Indexes. The main result here is the construction of the analog for IVS of the market portfolio for equity returns. This is based on using the open interest of the options and their (unitless) Vega. OI is the number of contracts on a given day for each name, maturity, and strike, and the Vega is an associated sensitivity to volatility. We find that portfolios of IVS returns weighed suitably by OI and Vega do track the IVS eigenportfolio. This provides an interpretation of the eigenportfolio that is analogous to the one for equity returns, is robust, and can be used in ways the VIX, the SPX volatility index, is used.

The analysis of question 3 can be found in the Factors and Eigenportfolios Using Tensors section. Yes, we find that retaining the tensor structure in the eigenportfolio analysis makes a difference in that the OI–Vega-weighted (tensor) IVS returns portfolio tracks the (tensor) eigenportfolio much better. This is a strong indication that data structure matters and data flattening should be avoided if possible.

There are obviously many, many more questions that can and should be asked about the IVS dataset, including theoretical ones about the methodology used. It is an evolving research enterprise.

APPENDIX A

DESCRIPTION OF DATA

Implied Volatility

Implied volatilities are not directly observable in the market and must be derived from the prices of traded options. At any point in time, for a given underlying stock, a variety of call and put options will be available to trade on the market, each of which will have a specific strike price and time to maturity. Given the observed prices of these options, it is then possible to infer an implied volatility using a numerical method.

Options on individual stocks have an American-style exercise feature and must be priced using a numerical algorithm because no closed-form solution is available. For this purpose, OptionMetrics uses the industry standard Cox–Ross–Rubinstein binomial tree model. This model can accommodate underlying securities with either discrete dividend payments or a continuous dividend yield. An option is priced by working backward through the tree from the maturity date when the payoff is known and incorporating any potential value arising from the possibility of early exercise at each node. The calculated price of the option at time $t = 0$ is the model price.

To compute the implied volatility of an option given its price, the model is run iteratively with different values of σ until the model price of the option converges to its market price, defined as the midpoint of the option's best closing bid and best closing offer prices. At this point, the final value of σ is the option's implied volatility.

This model can be adapted to account for the discrete dividends that stocks typically pay on a quarterly basis. The approach taken is to adjust the price of the underlying stock by subtracting the discounted value of all dividends to be paid between now and the expiry of the option.

Once the implied volatility has been calculated for an option, it is a simple matter to calculate its delta using the Black–Scholes model. We can then create a grid of time to maturity versus delta with a corresponding implied volatility where it is known. This will result in a grid of implied volatilities that have very different times to maturity and deltas from those needed to form a standardized grid, which is referred to as an IVS.

OptionMetrics calculates its standardized option-implied volatilities using a kernel-smoothing technique. The data are first organized by the log of days to expiration and by call-equivalent delta (delta for a call and one plus delta for a put). A kernel smoother is then used to generate a smoothed volatility value at each of the specified interpolation grid points. At each grid point j on the volatility surface, the smoothed volatility $\hat{\sigma}_j$ is calculated as a weighted sum of option-implied volatilities:

$$\hat{\sigma}_j = \frac{\sum_i \mathcal{V}_i \sigma_i \Phi(x_{i,j}, y_{i,j}, z_{i,j})}{\sum_i \mathcal{V}_i \Phi(x_{i,j}, y_{i,j}, z_{i,j})},$$

where i is indexed over all the options for that day, $\mathcal{V}_i$ is the Vega of the option, σ_i is the implied volatility, and Φ is the kernel function:

$$\Phi(x, y, z) = \frac{1}{\sqrt{2\pi}} e^{-\left(\left(\frac{x^2}{2h_1}\right)+\left(\frac{y^2}{2h_2}\right)+\left(\frac{z^2}{2h_3}\right)\right)}$$

EXHIBIT A1
Bucketing Scheme for Open Interest

τ Bucket (days)	Min Value (days)	Max Value (days)
30	1	45
60	46	75
91	76	106
122	107	137
152	138	167
182	168	227
273	228	319
365	320	

where

$$x_{i,j} = \log(T_i / T_j)$$
$$y_{i,j} = \Delta_i - \Delta_j$$
$$z_{i,j} = I_{CP_i = CP_j}.$$

Values $x_{i,j}$, $y_{i,j}$ and $z_{i,j}$ are measures of the distance between the option and the target grid point, $T_i(T_j)$ is the number of days to the expiration of the option (grid point), $\Delta_i(\Delta_j)$ is the call-equivalent delta of the option (grid point), $CP_i(CP_j)$ is the call/put identifier of the option (grid point), and I_0 is an indicator function. The kernel bandwidth parameters were chosen empirically and are set as $h_1 = 0.05$, $h_2 = 0.005$, and $h_3 = 0.001$.

Open Interest

Unlike implied volatility, open interest is directly observable in the market and represents the number of contracts in a particular option (underlying stock, strike, and expiry date) open at a point in time. For each open contract, there will be a party who is long the option and conversely one who is short the option. Open interest is therefore concentrated in the most popular contracts, which tend to be those closest to at-the-money and have less than one year but more than two weeks to expiry.

OptionMetrics provides an open interest for every available contract at the close of each trading day because this is the information available from the exchange. However, as described earlier, we are interested in the open interest at the grid points we have chosen for our IVSs, namely constant Δs and constant maturities. To estimate the open interest for these points, we used a bucketing approach.

First, we allocated each option to one of eight time buckets according to its days to expiry (τ), where the limits on these ranges are as given in Exhibit A1. Similarly, we allocated each option to one of eight Δ buckets according to where the limits on these ranges are as given in Exhibit A2. Once this was done, each option was allocated to one of our 56 grid points, and we then summed the open interest across all options at each grid point for each stock on each day in our dataset.

EXHIBIT A2
Bucketing Scheme for Delta

Δ Bucket	Min Value	Max Value
–20	–25	–1
–30	–35	–24
–40	–45	–34
50	45	54
40	35	44
30	25	34
20	1	24

APPENDIX B

THE MULTILINEAR SVD

If A is a matrix,[6] we can use the singular value decomposition (SVD) to expand

$$A = A(i,t), \quad i = 1, \ldots, N, \quad t = 1, \ldots, T$$

in the form

$$A = U \Sigma V^*.$$

Here U is an $N \times N$ unitary matrix, V is a unitary $T \times T$ matrix, and $\sum$ is a diagonal matrix with entries $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_R \geq 0$, where the index R is the rank of A and $R \leq \min\{N, T\}$. In terms of the components, we have

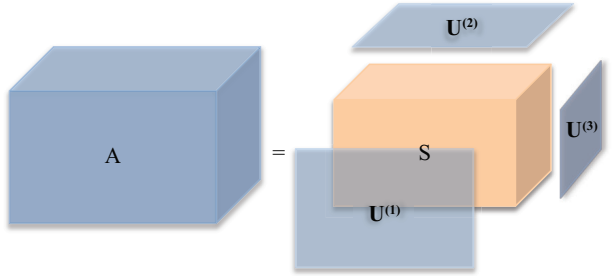
$$A(i,t) = \sum_{r=1}^{R} \sigma_r U(i,r) V(t,r), \; i = 1, \ldots, N, \; j = 1, \ldots, T.$$

For this expansion, only R columns of the unitary matrixes U and V are needed to represent the matrix A. When the sum stops at $\overline{R} < R$, the corresponding matrix $A_{\overline{R}}$ is the best rank $\overline{R}$ approximation of the matrix A in the Frobenius norm.

These properties do not generalize to tensors except in special and rather limited ways. A commonly used approach

[6] The notation in this Appendix is a bit different from that used earlier in the article and closer to that used in the linear algebra literature.

EXHIBIT B1
Core Decomposition of a Three-Dimensional Cube



for tensors is the MLSVD, or Tucker decomposition, which for a fourth-order tensor has the form:

$$A(\underline{i},t)=\sum_{\underline{1}\le\underline{i}'\le\underline{N},1\le t'\le T} S(\underline{i}',t')U(\underline{i},\underline{i}')V(t,t'). \tag{B1}$$

Here, $\underline{i}=(i_1,i_2,i_3)$, $\underline{N}=(N_1,N_2,N_3)$, V is a unitary $T\times T$ matrix, and U is a unitary tensor of the form

$$U(\underline{i},\underline{i}')=U^{(1)}(i_1,i_1')U^{(2)}(i_2,i_2')U^{(3)}(i_3,i_3') \tag{B2}$$

and the matrixes $U^{(1)}$, $U^{(2)}$, $U^{(3)}$ are unitary of size $N_1\times N_1$, $N_2\times N_2$, and $N_3\times N_3$, respectively. The fourth-order tensor S is not diagonal anymore, in general. However, it has the property of *all orthogonality*:

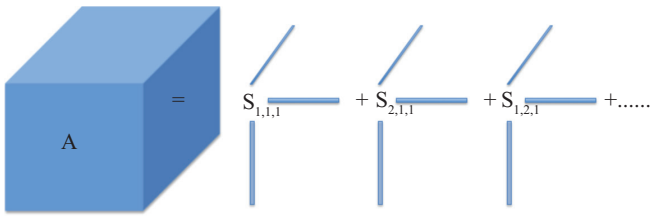
$$\sum S(\underline{i},t)S(\underline{i}',t')=0 \tag{B3}$$

where the indexes $(\underline{i},t)$ and $(\underline{i}',t')$ are equal except for one of the four components and the sum is over the three equal components. This means that distinct slices of each orientation (Horizontal (N_1), Vertical (N_2), Frontal (N_3), Time (T)) are orthogonal. Moreover, the indexes can be permuted so that the sums of squares over all except for one index are ordered by size

$$\begin{aligned}
\sigma_1^H&=\sqrt{\sum_{j,k,t}S^2(1,j,k,t)}\ge\sigma_2^H\ge\cdots\ge\sigma_{N_1}^H\\
\sigma_1^V&=\sqrt{\sum_{i,k,t}S^2(i,1,k,t)}\ge\sigma_2^V\ge\cdots\ge\sigma_{N_2}^V\\
\sigma_1^F&=\sqrt{\sum_{i,j,t}S^2(i,j,1,t)}\ge\sigma_2^F\ge\cdots\ge\sigma_{N_3}^F\\
\sigma_1^T&=\sqrt{\sum_{i,j,k}S^2(i,j,k,1)}\ge\sigma_2^T\ge\cdots\ge\sigma_T^T
\end{aligned} \tag{B4}$$

where we write here $\underline{i}=(i,j,k)$.

EXHIBIT B2
Tucker Decomposition



The MLSVD representation (Equation B1) is exact and is obtained from the application of SVDs to all the possible flattenings of the tensor and organizing the output suitably to get the result (see De Lathauwer, De Moor, and Vandewalle 2000; Kolda and Bader 2009; and Cichocki et al. 2015). A schematic of the MLSVD expansion is shown in Exhibit B1. Another expansion, not used here, is the canonical polyadic decomposition that is based on Frobenius norm minimization, a nonconvex problem; a schematic is shown in Exhibit B2.

In tensor PCA, we are interested in the covariance over time of the data, which we assume here is already normalized. In the notation of this Appendix, we define this covariance as

$$C(\underline{i},\underline{j})=\sum_t A(\underline{i},t)\overline{A(\underline{j},t)}. \tag{B5}$$

Using the MLSVD representation of A in Equation B1, we deduce the representation

$$C(\underline{i},\underline{j})=\sum_{\underline{i}',\underline{j}'} s(\underline{i}',\underline{j}')U(\underline{i},\underline{i}')\overline{U(\underline{j},\underline{j}')} \tag{B6}$$

where

$$s(\underline{i},\underline{j})=\sum_t S(\underline{i},t)S(\underline{j},t). \tag{B7}$$

Using the unitarity of the tensor U (Equation B2) in the covariance expansion (Equation B6) we obtain

$$\sum_{\underline{j}} C(\underline{i},\underline{j})U(\underline{j},\underline{k})=\sum_{\underline{i}'} s(\underline{i}',\underline{k})U(\underline{i},\underline{i}'). \tag{B8}$$

We see from this expression that the unitary tensor U is not an eigenvector of the covariance tensor C in Equation B8 because s is not diagonal in general. As noted earlier, U has the form

$$U(\underline{i},\underline{j})=U^{(1)}(i_1,j_1)U^{(2)}(i_2,j_2)U^{(3)}(i_3,j_3)$$

with $U^{(1)}(i_1, j_1)$, $U^{(2)}(i_2, j_2)$, $U^{(3)}(i_3, j_3)$ unitary matrixes of size $N_1 \times N_1$, $N_2 \times N_2$, $N_3 \times N_3$, respectively. Equation B8 captures rather clearly the scope the MLSVD representation of A in Equation B1 in that, although it is not a spectral form for the covariance, the tensor s on the right has positive diagonal elements that can be ordered and has off-diagonal elements that are often small because of the total orthogonality property (Equation B3); however, this has to be verified separately and it is not generally true. When, however, $s(\underline{1}, \underline{1})$ is large, then we can take

$$\tilde{U}^{(1)}(\underline{i}, \underline{1}) = U^{(1)}(i_1, 1), U^{(2)}(i_2, 1), U^{(3)}(i_3, 1) \tag{B9}$$

as the principal tensor eigenvector, as we did with a slightly different notation in Equation 21.

We close this Appendix with a few comments that complement the discussion up to now.

First, the truncation that produces the principal tensor eigenvector does not, in general, arise from a Frobenius norm minimization of the difference $||A - A_{PCA}||$. However, for the IVS data, the principal tensor eigenvector using MLSVD as described here and using CPD by Frobenius norm minimization produces essentially the same result. There will be a difference for multifactor tensor PCA, an issue that is not considered here.

Second, it is observed in practice that if there is a big gap in the size of the sigmas (see Equation B4), then this kind of truncation does behave like it does for matrixes. That is, the residual tends to behave as if it came from random entries.

Third, how does one test that a tensor has entries that behave as if they are random? We saw in the matrix case in the Matrix of Implied Volatility Returns section that one looks at the histogram of the singular values and compares this with a suitably adapted MP density—using a KS test, for example. For tensors, a test can be developed by using matrix flattenings of the tensor (e.g., horizontal, vertical, frontal). Because the MP law is an asymptotic one, care must be taken that the data structure is constructed so that it gives rise to flattenings for which this asymptotic law can actually be used. This is work in progress at present.

Fourth, there are other methods for constructing tensor principal components. Alternating least squares (or alternating SVDs) are often used, but little can be said theoretically, and direct (nonconvex in general) Frobenius norm optimizations using gradient descent or stochastic gradient descent avoid getting stuck in local minima early on.

REFERENCES

Avellaneda, M. "Hierarchical PCA and Applications to Portfolio Management." Working paper, New York University Courant, 2019.

Avellaneda, M., and D. Dobi. "Modeling Volatility Risk in Equity Options Market: A Statistical Approach." Working paper, New York University Courant, 2014.

Avellaneda, M., and J. H. Lee. 2010. "Statistical Arbitrage in the US Equities Market." *Quantitative Finance* 10 (7): 761–782.

Avellaneda, M., D. Boyer-Olson, J. Busca, and P. Friz. 2002. "Reconstruction of Volatility: Pricing Index Options by the Steepest Descent Approximation." *Risk*, October: 91–95.

Benaych-Georges, F., and R. R. Nadakuditi. 2011. "The Eigenvalues and Eigenvectors of Finite, Low Rank Perturbations of Large Random Matrices." *Advances in Mathematics* 227 (1): 494–521.

Berman, M. 2019. "Improved Estimation of the Intrinsic Dimension of a Hyperspectral Image Using Random Matrix Theory." *Remote Sensing* 11 (9): 1049.

Boyle, P. 2014. "Positive Weights on the Efficient Frontier." *North American Actuarial Journal* 18 (4): 462–477.

Cichocki, A., D. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, and H. A. Phan. 2015. "Tensor Decompositions for Signal Processing Applications: From Two-Way to Multiway Component Analysis." *IEEE Signal Processing Magazine* 32 (2): 145–163.

Cont, R., and J. Da Fonseca. 2002. "Dynamics of Implied Volatility Surfaces." *Quantitative Finance* 2 (1): 45–60.

De Lathauwer, B., L. De Moor, and J. Vandewalle. 2000. "A Multilinear Singular Value Decomposition." *SIAM Journal on Matrix Analysis and Applications* 21 (4): 1253–1278.

Dobi, D. "Modeling Volatility Risk in Equity Options: A Cross-Sectional Approach." Ph.D. dissertation, New York University, 2014.

El Karoui, N. 2008. "Spectrum Estimation for Large Dimensional Covariance Matrices Using Random Matrix Theory." *Annals of Statisics* 36 (6): 2757–2790.

Fama, E. F., and K. R. French. 1992. "The Cross-section of Expected Stock Returns." *The Journal of Finance*, 47 (2): 427–465.

Griffin, J. M., and A. Shams. 2017. "Manipulation in the VIX?" *The Review of Financial Studies* 31 (4): 1377–1417.

Johnstone, I. M., and D. Paul. 2018. "PCA in High Dimensions: An Orientation." *Proceedings of the IEEE* 106 (8): 1277–1292.

Kolda, T. G., and B. W. Bader. 2009. "Tensor Decompositions and Applications." *SIAM Review* 51 (3): 455–500.

Ledoit, O., and S. Péché. 2011. "Eigenvectors of Some Large Sample Covariance Matrix Ensembles." *Probability Theory and Related Fields* 151 (1–2): 233–264.

Ledoit, O., and M. Wolf. 2004. "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices." *Journal of Multivariate Analysis* 88 (2): 365–411.

——. 2012. "Nonlinear Shrinkage Estimation of Large-Dimensional Covariance Matrices." *The Annals of Statistics* 40 (2): 1024–1060.

Markowitz, H. 1952. "Portfolio Selection." *The Journal of Finance* 7 (1): 77–91.

Plerou, V., P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr, and H. E. Stanley. 2002. "Random Matrix Approach to Cross Correlations in Financial Data." *Physical Review E* 65 (6): 066126.

Roll, R., and S. A. Ross. 1980. "An Empirical Investigation of the Arbitrage Pricing Theory." *The Journal of Finance* 35 (5): 1073–1103.

Tucker, L. R. 1966. "Some Mathematical Notes on Three-Mode Factor Analysis." *Psychometrika* 31 (3): 279–311.

ADDITIONAL READING

Neural Networks in Finance: *Design and Performance*
Irene Aldridge and Marco Avellaneda
The Journal of Financial Data Science
https://jfds.pm-research.com/content/1/4/39

ABSTRACT: *Neural networks have piqued the interest of many financial modelers, but the concrete applications and implementation have remained elusive. This article discusses a step-by-step technique for building a potentially profitable financial neural network. The authors also demonstrate a successful application of the neural network to investing based on daily and monthly financial data. The article discusses various components of neural networks and compares popular neural network activation functions and their applicability to financial time series. Specifically, use of the tanh activation function is shown to closely mimic financial returns and produce the best results. Incorporating additional inputs, such as the S&P 500 prices, also helps improve neural networks' forecasting performance. Longer training periods deliver strategies that closely mimic common technical analysis strategies, such as moving-average crossovers, whereas shorter training periods deliver significant forecasting power. The resulting neural network–based daily trading strategies on major US stocks significantly and consistently outperform the buy-and-hold positions in the same stocks.*