CENTER FOR RISK MANAGEMENT RESEARCH

Working Paper #2015-09

Understanding Systematic Risk: A High-Frequency Approach

Markus Pelger, Stanford University

August 21, 2015

University of California
Berkeley

Understanding Systematic Risk: A High-Frequency Approach

Markus Pelger*

August 21, 2015

Abstract

Under a large dimensional approximate factor model for asset returns, I use high-frequency data for the S&P 500 firms to estimate the latent continuous and jump factors. I estimate four very persistent continuous systematic factors for 2007 to 2012 and three from 2003 to 2006. These four continuous factors can be approximated very well by a market, an oil, a finance and an electricity portfolio. The value, size and momentum factors play no significant role in explaining these factors. For the time period 2003 to 2006 the finance factor seems to disappear. There exists only one persistent jump factor, namely a market jump factor. Using implied volatilities from option price data, I analyze the systematic factor structure of the volatilities. There is only one persistent market volatility factor, while during the financial crisis an additional temporary banking volatility factor appears. Based on the estimated factors, I can decompose the leverage effect, i.e. the correlation of the asset return with its volatility, into a systematic and an idiosyncratic component. The negative leverage effect is mainly driven by the systematic component, while it can be non-existent for idiosyncratic risk.

Keywords: Systematic risk, High-dimensional data, High-frequency data, Latent factors, Approximate factor model, PCA, S&P 500 index constituents, Leverage effect, Volatility factors, Industry factors, Jumps

JEL classification: C38, C52, C55, C58, G12

*Department of Management Science & Engineering, Stanford University, Stanford, CA 94305, Email: mpelger@stanford.edu. This paper is part of my Ph.D. thesis in Economics at UC Berkeley. I thank Robert M. Anderson, Martin Lettau, Michael Jansson and Lisa Goldberg for extensive advice and support. I also thank Jason Zhu for excellent research assistance. I am very grateful for comments and helpful discussions with Richard Stanton, Ulrike Malmendier, Steve Evans, Nicolae Gârleanu, Johan Walden, Viktor Todorov and the participants at the Risk Management Seminar, Econometrics Seminar and Finance Pre-Seminar at UC Berkeley and the INFORMS Annual Meeting. This work was supported by the Center for Risk Management Research at UC Berkeley.

1 Introduction

One of the most popular methods for modeling and estimating systematic risk are factor models. This paper employs the new statistical methods developed in Pelger (2015) to estimate and analyze an unknown factor structure in a large cross-section of high-frequency equity data. Conventional factor analysis requires long time horizons, while this new methodology works with short time horizons, e.g. a month. The question of how to capture systematic risk is one of the most fundamental questions in asset pricing. This paper enhances our understanding about systematic risk by answering the following questions: (1) What is a good number of factors to explain the systematic movements and how does this number change over time? (2) What are the factors and how persistent is the factor structure over time? (3) Are continuous systematic risk factors, which capture the variation during "normal" times, different from jump factors, which can explain systematic tail events? (4) How does the leverage effect, i.e. the correlation of asset returns with its volatility, depend on systematic and nonsystematic risk.

The important contribution of this paper is that it does not use a pre-specified (and potentially miss-specified) set of factors. Instead I estimate the statistical factors, which can explain most of the common comovement in a large cross-section of high-frequency data. As the high-frequency data allows me to analyze different short time horizons independently, I do not impose restrictions on the potential time-variation in the factors. For a pre-specified set of factors studies have already shown that time-varying systematic risk factors capture the data better.[1] Empirical evidence also suggests that for a given factor structure systemic risk associated with discontinuous price movements is different from continuous systematic risk.[2] I confirm and extend these results to a latent factor structure.

The statistical theory underlying my estimations is very general and developed in Pelger (2015). It combines the two fields of high-frequency econometrics and large-dimensional factor analysis. Under the assumption of an approximate factor model it estimates an unknown factor structure for general continuous-time processes based on

[1]The idea of time-varying systematic risk factors contains the conditional version of the CAPM as a special case, which seems to explain systematic risk significantly better than its constant unconditional version. Contributions to this literature include for example Jagannathan and Wang (1996) and Lettau and Ludvigson (2001). Bali, Engle, and Tang (2014) have also shown that GARCH-based time-varying conditional betas help explain the cross-sectional variation in expected stock returns.

[2]Empirical studies supporting this hypothesis include Bollerslev, Li and Todorov (2015), Pan (2002), Eraker (2003), Eraker et al. (2003), Bollerslev and Todorov (2011) and Gabaix (2012).

high-frequency data. Using a truncation approach, I can separate the continuous and jump components of the price processes, which I use to construct a "jump covariance" and a "continuous risk covariance" matrix. The latent continuous and jump factors can be separately estimated by principal component analysis. The number of total, continuous and jump factors is estimated by analyzing the ratio of perturbed eigenvalues, which is a novel idea to the literature and shows an excellent performance in simulations. A new generalized correlation test allows me to compare the statistical factors with observed economic factors.

My empirical investigations are based on a novel high-frequency data set of 5-minutes prices for the S&P 500 firms from 2003 to 2012. My estimation approach indicates that the number and the factors do change over time. I estimate four very persistent continuous systematic factors for 2007 to 2012 and three from 2003 to 2006. These continuous factors can be approximated very well by an equally-weighted market portfolio and three industry factors, namely an oil, finance and electricity factor.[3] For the time period 2003 to 2006 the finance factor seems to disappear, while the remaining factor structure stays persistent. For the whole time period there seems to exist only one persistent jump factor, namely a market jump factor. My results are robust to the sampling frequency and microstructure noise.[4]

In order to measure how well my set of statistical factors can be explained by industry or the Fama-French Carhart factors I use the concept of generalized correlations. The generalized correlations are the highest correlations that can be achieved through linear orthogonal combinations of each of the two sets of factors.[5] The four continuous industry factors (market, oil, finance and electricity) yield the very high generalized correlations of 1.00, 098, 0.95 and 0.80 with the four statistical continuous factors. The value, size and momentum factors cannot explain the four continuous statistical factors as the generalized correlations with the Fama-French Carhart factors only take the lower values 0.95, 0.74, 0.60 and 0.00. The jump structure is different as the generalized correlations of my four industry jump factors with the first four statistical jump factors are significantly

[3]The industry factors are constructed as portfolios with equally-weighted returns for firms in the oil and gas industry, the banking and insurance industry and the electricity and electric utility industry.

[4]I can show that the estimated monthly and yearly factor structures are essentially identical based on 5 minutes data. Changes in the factor structure seem to occur only for different years. Within a year the estimated factor structure is basically the same if I use 5 minutes, 15 minutes or daily data. Microstructure noise becomes only relevant for high-frequencies. The fact that my results are robust to different time horizons indicates that they are robust to microstructure noise.

[5]For more details see Section 4.

lower with 0.99, 0.75, 0.29 and 0.05.[6]

In the second part of the paper I analyze the systematic factor structure of the volatilities based on a novel data set of daily short-maturity, at-the-money implied volatilities from option prices for the same firms and time period. There seems to be only one persistent market volatility factor, while during the financial crisis an additional temporary banking volatility factor appears.

This paper contributes to the understanding of the leverage effect by separating this effect into its systematic and idiosyncratic component based on the estimated latent factors. The leverage effect, which describes the generally negative correlation between an asset return and its volatility changes, is one of the most important empirical stylized facts about the volatility. High-frequency data is particularly suited for analyzing the leverage effect as it allows to estimate changes in the unobserved volatility. There is no consensus on the economic explanation for this statistical effect. The magnitude of the effect seems to be too large to be explained by financial leverage. Alternative economic interpretations use a risk-premium argument. An anticipated rise in volatility increases the risk premium and hence requires a higher rate of return from the asset. This leads to a fall in the asset price. The causality for these two interpretations is different. These different explanations have been tested by Bekaert and Wu (2000) who use a parametric conditional CAPM model under a GARCH specification to obtain results consistent with the risk-premium story. I estimate the leverage effect completely non-parametrically and decompose it into its systematic and nonsystematic part based on my general statistical factors. I show that the leverage effect appears predominantly for systematic risk, while it is smaller and can even be non-existent for idiosyncratic risk. These findings rule out the financial leverage story, as that explanation does not distinguish between different sources of risk.

As an illustration I plot the cross-sectional distribution for different components of the leverage effect in Figure 1. The continuous returns and the volatilties are first decomposed into a systematic and idiosyncratic component based on the 4 continuous return factors and the largest volatility factor. Then I calculate the correlations between the different components. The largest negative correlation with a mean of -0.3 is between the systematic return component and the volatility, while the idiosyncratic return component has only an average correlation of -0.13 with the volatility. Systematic returns and idiosyncratic volatility are almost orthogonal to each other with a mean correlation

[6]Section 5 explains in detail how these numbers are calculated.

of -0.03.[7]

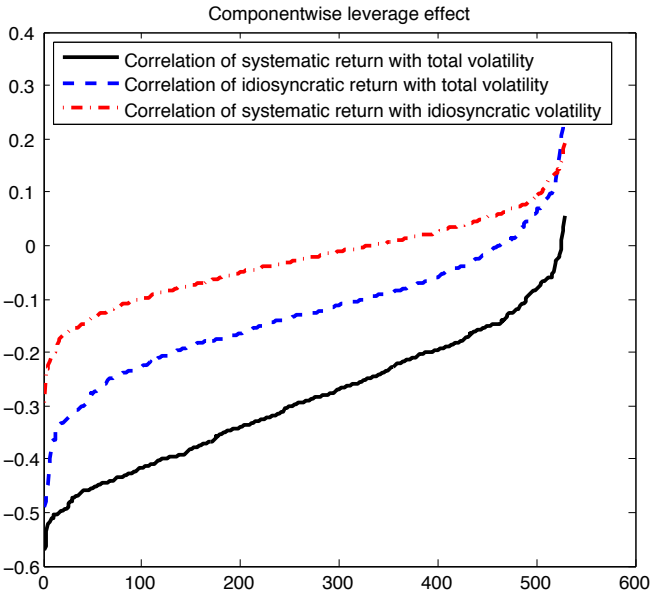

Figure 1: Cross-sectional distribution of componentwise leverage effect in 2012: Separation into systematic and idiosyncratic components uses 4 statistical return factors and 1 statistical volatility factor. The results are taken from Figure 6 in Section 6.

2 Related Literature

My paper contributes to the central question in empirical and theoretical asset pricing what constitutes systematic risk. There are essentially three common ways of selecting which factors and how many describe the systematic risk. The first approach is based on theory and economic intuition. The capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965) with the market as the only common factor falls into this category. The second approach bases factors on firm characteristics with the three-factor model of Fama and French (1993) as its most famous example. My approach falls into the third category where factor selection is statistical. This approach is motivated by the arbitrage pricing theory (APT) of Ross (1976). Factor analysis can be used to analyze

[7]As I show in Section 6 idiosyncratic returns and systematic volatility are also almost orthogonal with a mean correlation of 0.001 for the year 2012.

the covariance structure of returns. This approach yields estimates of factor exposures as well as returns to underlying factors, which are linear combinations of returns on underlying assets. The notion of an "approximate factor model" was introduced by Chamberlain and Rothschild (1983), which allowed for a non-diagonal covariance matrix of the idiosyncratic component. Connor and Korajczyk (1988, 1993) study the use of principal component analysis in the case of an unknown covariance matrix, which has to be estimated.[8] One distinctive feature of the factor literature described above is that it uses long-horizon data. The advantages of using high-frequency data are apparent as it provides more information for more precise estimation and there are sufficiently many data points for estimating a factor structure that varies over longer time-horizons. For example the factor analysis can be pursued on a monthly basis to test how the factor structure changes over time.[9]

So far most of the empirical literature that utilizes the tools of high-frequency econometrics to analyze a factor structure is limited to a pre-specified set of factors. For example, Bollerslev, Li and Todorov (2015) estimate the betas for a continuous and jump market factor. Fan, Furger and Xiu (2014) estimate a large-dimensional covariance matrix with high-frequency data for a given factor structure. My work goes further as I estimate the unknown continuous and jump factor structure in a large cross-section. An exception is Aït-Sahalia and Xiu (2015a) who apply nonparametric principal component analysis to a low-dimensional cross-section of high-frequency data.[10]

My results were derived simultaneously and independently to results by Aït-Sahalia and Xiu (2015b). Both of our papers consider the estimation of a large-dimensional factor model based on high-frequency observations. From the theoretical side their work

[8]The general case of a static large dimensional factor model is treated in Bai (2003) and Bai and Ng (2002). Forni, Hallin, Lippi and Reichlin (2000) introduced the dynamic principal component method. Fan, Liao and Mincheva (2013) study an approximate factor structure with sparsity.

[9]A disadvantage of working with high-frequency data is the relatively short time horizon for which appropriate data for a large cross-section is available. Arbitrage pricing theory links risk premiums to systematic risk. Factors that explain most of the comovements should also explain most of the risk premia. Unfortunately, the short-time horizon of 10 years puts restrictions on a reliable estimation of the risk premium and hence for testing this statement. Hence, this paper focusses on interpreting and understanding the properties of factors that explain most of the common comovements in the data without testing the asset pricing implications.

[10]My results were derived simultaneously and independently to results in Aït-Sahalia and Xiu (2015a). They find that the first three continuous principal components explain a large fraction of the variation in the S&P100 index. Their work is different from mine as they consider a low-dimensional regime for continuous processes, whereas I work in a large-dimensional regime and analyze both the continuous and jump structures.

is different from this paper and Pelger (2015) as I also include jumps and provide a distribution theory. My main identification condition is a bounded eigenvalue condition on the idiosyncratic covariance matrix, while their identification is based on a sparsity assumption on the idiosyncratic covariance matrix. In the empirical part they consider a similar data set with 15 minutes data and show that 4 statistical factors are sufficient to obtain a block-diagonal pattern in the idiosyncratic covariance matrix. Their study focuses on estimating the continuous covariance matrix, while my work tries to explain the factor structure itself and also considers the factor structures in jumps and volatilities.

The rest of the paper is organized as follows. Section 3 introduces the factor model. In Section 4 I explain the estimation method. Section 5 analyzes the systematic pattern in equity data based on high-frequency data. Section 6 is an empirical application to volatility data and includes the analysis of the leverage effect. Concluding remarks are provided in Section 7. All the mathematical statements and additional empirical results are deferred to the appendices.

3 Factor Model

The theoretical foundation for my empirical results assumes an asymptotic framework in which the number of cross-sectional and high-frequency observations both go to infinity. The high number of cross-sectional observations makes the large dimensional covariance analysis challenging, but under the assumption of a general approximate factor structure the "curse of dimensionality" turns into a "blessing" as it becomes necessary for estimating the systematic factors. I argue that my data set with around 20,000 yearly observations for each of the 500 cross-sectional assets is sufficiently large for invoking asymptotic theory.[11]

This paper assumes that log asset prices can be modeled by an approximate factor model. Hence most co-movements in asset prices are due to a systematic factor compo-

[11] I have run many robustness tests where I vary the number of cross-sectional and high-frequency observations and my general findings are not affected. The availability of reliable intra-day data for a large cross-section limits my study to the data that I am using in this paper. I cannot rule out the possibility that with more data I could find additional factors that are persistent. My estimations indicate that there are four respectively three strong factors in the equity data and they seem to follow a very strong pattern, which makes me believe that this is not a pure data-mining but real economic phenomena. It is possible that other factors, e.g. a value factor, do not explain much of the correlation for my cross-section and hence are not identified as a systematic factor.

nent. In more detail assume that I have N assets with log prices denoted by $X_i(t)$.[12] Assume the N-dimensional stochastic process $X(t)$ can be explained by a factor model, i.e.

$$X_i(t) = \Lambda_i^\top F(t) + e_i(t) \qquad i = 1, ..., N \text{ and } t \in [0, T]$$

where Λ_i is a $K \times 1$ dimensional vector and $F(t)$ is a K-dimensional stochastic process. The loadings Λ_i describe the exposure to the systematic factors F, while the residuals e_i are stochastic processes that describe the idiosyncratic component. However, I only observe the stochastic process X at M discrete time observations in the interval $[0, T]$. If I use an equidistant grid[13], I can define the time increments as $\Delta_M = t_{j+1} - t_j = \frac{T}{M}$ and observe

$$X(t_j) = \Lambda F(t_j) + e(t_j) \qquad j = 1, ..., M.$$

with $\Lambda = (\Lambda_1, ..., \Lambda_N)^\top$ and $X(t) = (X_1(t), ..., X_N(t))^\top$. In my setup the number of cross-sectional observations N and the number of high-frequency observations M is large, while the time horizon T and the number of systematic factors K is fixed. The loadings Λ, factors F, residuals e and number of factors K are unknown and have to be estimated.

I am also interested in estimating the continuous component, jump component and the volatility of the factors. Denoting by F^C the factors that have a continuous component and by F^D the factor processes that have a jump component, I can write

$$X(t) = \Lambda^C F^C(t) + \Lambda^D F^D(t) + e(t).$$

Note, that for factors that have both, a continuous and a jump component, the corresponding loadings have to coincide. In the following I assume a non-redundant representation of the K^C continuous and K^D jump factors. For example if we have K factors which have all exactly the same jump component but different continuous components, this results in K different total factors and $K^C = K$ different continuous factors, but in only $K^D = 1$ jump factor.

My approach requires only very weak assumptions which are summarized in Appendix B. First, the individual asset price dynamics are modeled as Itô-semimartingales, which

[12] Later in this paper I will also use volatilities for the process $X_i(t)$.

[13] My results would go through under a time grid that is not equidistant as long as the largest time increment goes to zero with speed $O\left(\frac{1}{M}\right)$.

is the most general class of stochastic processes, for which the general results of high-frequency econometrics are available. It includes many processes, for example stochastic volatility processes or jump-diffusion processes with stochastic intensity rate. Second, the dependence between the assets is modeled by an approximate factor structure similar to Chamberlain and Rothschild (1983). The idiosyncratic risk can be serially correlated and weakly cross-sectionally correlated and hence allows for a very general specification. The main identification criterion for the systematic risk is that the quadratic covariation matrix of the idiosyncratic risk has bounded eigenvalues, while the quadratic covariation matrix of the systematic factor part has unbounded eigenvalues. For this reason the principal component analysis can relate the eigenvectors of the exploding eigenvalues to the loadings of the factors. Third, in order to separate continuous systematic risk from jump risk, I allow only finite activity jumps, i.e. there are only finitely many jumps in the asset price processes. Many of my results work without this restriction and it is only needed for the separation of these two components. This still allows for a very rich class of models and for example general compound poisson processes with stochastic intensity rates can be accommodated. Last but not least, I work under the simultaneous limit of a growing number of high-frequency and cross-sectional observations. I do not restrict the path of how these two parameters go to infinity. However, my results break down if one of the two parameters stays finite. In this sense the "curse of dimensionality" becomes a "blessing".

4 Estimation

4.1 Estimating the Factors

I employ the estimation techniques developed in Pelger (2015). There are M observations of the N-dimensional stochastic process X in the time interval $[0,T]$. For the time increments $\Delta_M = \frac{T}{M} = t_{j+1} - t_j$ I denote the increments of the stochastic processes by

$$X_{j,i} = X_i(t_{j+1}) - X_i(t_j) \qquad F_j = F(t_{j+1}) - F(t_j) \qquad e_{j,i} = e_i(t_{j+1}) - e_i(t_j).$$

In matrix notation we have

$$\underset{(M\times N)}{X} = \underset{(M\times K)}{F}\underset{(K\times N)}{\Lambda^\top} + \underset{(M\times N)}{e}.$$

For a given K my goal is to estimate Λ and F. As in any factor model where only X is observed Λ and F are only identified up to invertible transformations. I impose the standard normalization that $\frac{\hat{\Lambda}^\top \hat{\Lambda}}{N} = I_K$ and that $\hat{F}^\top \hat{F}$ is a diagonal matrix.

The estimator for the loadings $\hat{\Lambda}$ is defined as the eigenvectors associated with the K largest eigenvalues of $\frac{1}{N} X^\top X$ multiplied by $\sqrt{N}$. The estimator for the factor increments is $\hat{F} = \frac{1}{N} X \hat{\Lambda}$. Note that $\frac{1}{N} X^\top X$ is an estimator for the quadratic covariation $\frac{1}{N}[X, X]$ for a finite N. The asymptotic theory is applied for $M, N \to \infty$. The systematic component of $X(t)$ is the part that is explained by the factors and defined as $C(t) = \Lambda F(t)$. The increments of the systematic component $C_{j,i} = F_j \Lambda_i^\top$ are estimated by $\hat{C}_{j,i} = \hat{F}_j \hat{\Lambda}_i^\top$.

Intuitively under some assumptions I can identify the jumps of the process $X_i(t)$ as the big movements that are larger than a specific threshold. I set the threshold identifier for jumps as $\alpha \Delta_M^{\bar{\omega}}$ for some $\alpha > 0$ and $\bar{\omega} \in \left(0, \frac{1}{2}\right)$ and define $\hat{X}^C_{j,i} = X_{j,i} \mathbb{1}_{\{|X_{j,i}| \leq \alpha \Delta_M^{\bar{\omega}}\}}$ and $\hat{X}^D_{j,i} = X_{j,i} \mathbb{1}_{\{|X_{j,i}| > \alpha \Delta_M^{\bar{\omega}}\}}$. The estimators $\hat{\Lambda}^C$, $\hat{\Lambda}^D$, $\hat{F}^C$ and $\hat{F}^D$ are defined analogously to $\hat{\Lambda}$ and $\hat{F}$, but using $\hat{X}^C$ and $\hat{X}^D$ instead of X. [14]

The quadratic covariation of the factors can be estimated by $\hat{F}^\top \hat{F}$ and the volatility component of the factors by $\hat{F}^{C\top} \hat{F}^C$. I show that the estimated increments of the factors $\hat{F}, \hat{F}^C$ and $\hat{F}^D$ can be used to estimate the quadratic covariation with any other process.

As I have already noted before, factor models are only identified up to invertible transformations. Two sets of factors represent the same factor model if the factors span the same vector space. When trying to interpret estimated factors by comparing them with economic factors, I need a measure to describe how close two vector spaces are to each other. As proposed by Bai and Ng (2006) the generalized correlation is a natural candidate measure. Let F be my K-dimensional set of factor processes and G be a K_G-dimensional set of economic candidate factor processes. I want to test if a linear combination of the candidate factors G can replicate some or all of the true factors F. The first generalized correlation is the highest correlation that can be achieved through a linear combination of the factors F and the candidate factors G. For the second generalized correlation I first project out the subspace that spans the linear combination for the first generalized correlation and then determine the highest possible correlation that can be achieved through linear combinations of the remaining $K-1$ respectively $K_G - 1$ dimensional subspaces. This procedure continues until I have calculated the $min(K, K_G)$ generalized correlation. Mathematically the generalized correlations are the square root

[14] For the jump threshold I use the *TOD* specification of Bollerslev, Li and Todorov (2013).

of the $min(K, K_G)$[15] largest eigenvalues of the matrix $[F,G]^{-1}[F,F][G,G]^{-1}[G,F]$. If $K = K_G = 1$ it is simply the correlation as measured by the quadratic covariation. Similarly the distance between two loading matrices Λ and $\tilde{\Lambda}$ with dimension $N \times K$ respectively $N \times \tilde{K}$ is measured as the square root of the $\min(K, \tilde{K})$ largest eigenvalues of $(\Lambda^\top\Lambda)^{-1}\Lambda^\top\tilde{\Lambda}(\tilde{\Lambda}^\top\tilde{\Lambda})^{-1}\tilde{\Lambda}^\top\Lambda$. If the two matrices span the same vector spaces, the generalized correlations are all equal to 1. Otherwise they denote the highest possible correlations that can be achieved through linear combinations of the subspaces. If for example for $K = K_G = 3$ the generalized correlations are $\{1, 1, 0\}$, it implies that there exists a linear combination of the three factors in G that can replicate two of the three factors in F.[16] I have shown that under general conditions the estimated factors $\hat{F}$, $\hat{F}^C$ and $\hat{F}^D$ can be used instead of the true unobserved factors for calculating the generalized correlations. Unfortunately, in this high-frequency setting there does not exist a theory for confidence intervals for the individual generalized correlations. However, I have developed an asymptotic distribution theory for the sum of squared generalized correlations, which I label as total generalized correlation. I use the total generalized correlation test described in Appendix B.6 to test if a set of economic factors represents the same factor model as the statistical factors.

The theorems and assumptions are collected in Appendix B.

4.2 Estimating the Number of Factors

In Pelger (2015) I also develop a new estimator for the number of factors, that can also distinguish between the number of continuous and jump factors. This estimator uses only the same weak assumptions that are needed for the consistency of my factor estimator. It can also easily be extended to long time horizon factor models and in simulations it outperforms the existing estimators while maintaining weaker assumptions. Intuitively the large eigenvalues are associated with the systematic factors and hence the problem of estimating the number of factors is roughly equivalent to deciding which eigenvalues are

[15] Using $min(K, K_G)$ instead of $max(K, K_G)$ is just a labeling convention. All the generalized correlations after $min(K, K_G)$ are zero and hence usually neglected.

[16] Although labeling the measure as a correlation, I do not demean the data. This is because the drift term essentially describes the mean of a semimartingale and when calculating or estimating the quadratic covariation it is asymptotically negligible. Hence, the generalized correlation measure is based only on inner products and the generalized correlations correspond to the singular values of the matrix $[F,G]$ if F and G are orthonormalized with respect to the inner product $[.,.]$. The generalized correlation between two sets of loadings is a measure of how well one set can be described as a linear combination of the other set.

considered to be large with respect to the rest of the spectrum. Under the assumptions that I need for consistency I can show that the first K "systematic" eigenvalues of $X^\top X$ are $O_p(N)$, while the nonsystematic eigenvalues are $O_p(1)$. A straightforward estimator for the number of factors considers the eigenvalue ratio of two successive eigenvalues and associates the number of factors with a large eigenvalue ratio. However, without very strong assumptions the small eigenvalues cannot be bounded from below, which could lead to exploding eigenvalue ratios in the nonsystematic spectrum. I propose a perturbation method to avoid this problem. As long as the eigenvalue ratios of the perturbed eigenvalues cluster, we are in the nonsystematic spectrum. As soon as we do not observe this clustering any more, but a large eigenvalue ratio of the perturbed eigenvalues, we are in the systematic spectrum.

The number of factors can be consistently estimated through the perturbed eigenvalue ratio statistic and hence, I can replace the unknown number K by its estimator $\hat{K}$. Denote the ordered eigenvalues of $X^\top X$ by $\lambda_1 \geq ... \geq \lambda_N$. I choose a slowly increasing sequence $g(N, M)$ such that $\frac{g(N,M)}{N} \to 0$ and $g(N, M) \to \infty$. Based on simulations a good choice for the perturbation term g is the median eigenvalue rescaled by $\sqrt{N}$, but the results are very robust to different choices of the perturbation.[17] Then, I define perturbed eigenvalues $\hat{\lambda}_k = \lambda_k + g(N, M)$ and the perturbed eigenvalue ratio statistic

$$ER_k = \frac{\hat{\lambda}_k}{\hat{\lambda}_{k+1}} \qquad \text{for } k = 1, ..., N-1.$$

The estimator for the number of factors is defined as the first time that the perturbed eigenvalue ratio statistic does not cluster around 1 any more:

$$\hat{K}(\gamma) = \max\{k \leq N-1 : ER_k > 1+\gamma\} \qquad \text{for } \gamma > 0.$$

The definitions of $\hat{K}^C(\gamma)$ and $\hat{K}^D(\gamma)$ are analogous but using λ_i^C respectively λ_i^D of the matrices $\hat{X}^{C\top}\hat{X}^C$ and $\hat{X}^{D\top}\hat{X}^D$. The results in my empirical analysis are robust to a wide range of values for the threshold γ.

[17] I estimate the number of factors using the perturbed eigenvalue ratio estimator with $g(N, M) = \sqrt{N} \cdot median\{\lambda_1, ..., \lambda_N\}$. For robustness I also use an unperturbed eigenvalue ratio test and $g(N, M) = log(N) \cdot median\{\lambda_1, ..., \lambda_N\}$. The results are the same.

5 High-Frequency Factors in Equity Data

5.1 Data

I use intraday log-prices from the Trade and Quote (TAQ) database for the time period from January 2003 to December 2012 for all the assets included in the S&P 500 index at any time between January 1993 and December 2012. In order to strike a balance between the competing interests of utilizing as much data as possible and minimizing the effect of microstructure noise and asynchronous returns, I choose to use 5-minute prices.[18] More details about the data selection and cleaning procedures are in Appendix A. For each of the 10 years I have on average 250 trading days with 77 log-price increments per day. Within each year I have a cross-section N between 500 and 600 firms.[19] The exact number for each year is in Table 1. After applying the cleaning procedure the intersection of the firms for the time period 2007 to 2012 is 498, while the intersection of all firms for the 10 years is only 304. The yearly results use all the available firms in that year, while the analysis over longer horizons uses the cross-sectional intersection.

Year	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Original	614	620	622	612	609	606	610	603	587	600
Cleaned	446	540	564	577	585	598	608	597	581	593
Dropped	27.36%	12.90%	9.32%	5.72%	3.94%	1.32%	0.33%	1.00%	1.02%	1.17%

Table 1: Observations after data cleaning

When identifying jumps, we face the tradeoff of finding all discontinuous movements against misclassifying high-volatility regimes as jumps. Therefore, the threshold should take into account changes in volatilities and intra-day volatility patterns. I use the TOD estimator of Bollerslev, Li and Todorov (2013) for separating the continuous from the jump movements. Hence the threshold is set as $a \cdot 77^{-0.49} \hat{\sigma}_{j,i}$, where $\hat{\sigma}_{j,i}$ estimates the daily volatility of asset i at time j by combining an estimated Time-of-Day volatility pattern with a jump robust bipower variation estimator for that day. Intuitively I classify all increments as jumps that are beyond a standard deviations of a local estimator of the stochastic volatility. For my analysis I use $a = 3$, $a = 4$ and $a = 4.5$.

[18] I have run robustness tests with 15min and daily data and the main results do not change.

[19] I do not extend my analysis to the time before 2003 as there are too many missing high-frequency observations for the large cross-section.

Table 2 lists the fraction of increments identified as jumps for different thresholds. Depending on the year for $a = 3$ more than 99% of the observations are classified as continuous, while less than 1% are jumps. In 2012, 99.2% of the movements are continuous and explain around 85% of the total quadratic variation, while the 0.8% jumps explain the remaining 15% of the total quadratic covariation. Changing the threshold either more or less movements are classified as jumps.[20] All the results for the continuous factors are extremely robust to this choice. However, the results for the jump factors are sensitive to the threshold. Therefore, I am very confident about the results for the continuous factors, while the jump factor results have to be interpreted with caution. If not noted otherwise, the threshold is set to $a = 3$ in the following.

As a first step Table 2 lists for each year the fraction of the total continuous variation explained by the first four continuous factors and the fraction of the jump variation explained by the first jump factor. As expected systematic risk varies over time and is larger during the financial crisis. The systematic continuous risk with 4 factors accounts for around 40-47% of the total correlation from 2008 to 2011, but explains only around 20-31% in the other years.[21] A similar pattern holds for the jumps where the first jump factor explains up to 10 times more of the correlation in 2010 than in the years before the financial crisis.

I have applied the factor estimation to the quadratic covariation and the quadratic correlation matrix, which corresponds to using the covariance or the correlation matrix in long-horizon factor modeling. For the second estimator I rescale each asset for the time period under consideration by the square-root of its quadratic covariation. Of course, the resulting eigenvectors need to be rescaled accordingly in order to obtain estimators for the loadings and factors. All my results are virtually identical for the covariation and the correlation approach, but the second approach seems to provide slightly more robust estimators for shorter time horizons. Hence, all results reported in this paper are based on the second approach.

[20] There is no consensus on the number of jumps in the literature. Christensen, Oomen and Podolskij (2014) use ultra high-frequency data and estimate that the jump variation accounts for about 1% of total variability. Most studies based on 5 minutes data find that the jump variation should be around 10 - 20% of the total variation. My analysis considers both cases.

[21] The percentage of correlation explained by the first four factors is calculated as the sum of the first four eigenvalues divided by the sum of all eigenvalues of the continuous quadratic correlation matrix.

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Percentage of increments identified as jumps										
a=3	0.011	0.011	0.011	0.010	0.010	0.009	0.008	0.008	0.007	0.008
a=4	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001	0.001	0.001
a=4.5	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.001
Variation explained by jumps										
a=3	0.19	0.19	0.19	0.16	0.21	0.16	0.16	0.15	0.12	0.15
a=4	0.07	0.07	0.07	0.05	0.10	0.06	0.06	0.06	0.03	0.05
a=4.5	0.05	0.04	0.05	0.04	0.08	0.04	0.05	0.05	0.02	0.04
Percentage of jump correlation explained by first 1 jump factor										
a=3	0.05	0.03	0.03	0.03	0.06	0.07	0.08	0.19	0.12	0.06
a=4	0.03	0.02	0.02	0.04	0.08	0.06	0.08	0.25	0.09	0.08
a=4.5	0.03	0.03	0.02	0.05	0.09	0.06	0.08	0.22	0.12	0.09
Percentage of continuous correlation explained by first 4 continuous factors										
	0.26	0.20	0.21	0.22	0.29	0.45	0.40	0.40	0.47	0.31

Table 2: (1) Fraction of increments identified as jumps for different thresholds. (2) Fraction of total quadratic variation explained by jumps for different thresholds. (3) Systematic jump correlation as measured by the fraction of the jump correlation explained by the first jump factor for different thresholds. (4) Systematic continuous correlation as measured by the fraction of the continuous correlation explained by the first four continuous factors.

5.2 Continuous Factors

5.2.1 Number of Factors

I estimate four continuous factors for each of the years from 2007 to 2012 and three continuous factors for the years 2003 to 2006. Figure 2 shows the estimation results for the numbers of continuous factors. Starting from the right I am looking for a visible strong increase in the perturbed eigenvalue ratio. Asymptotically any critical value larger than 1 should indicate the beginning of the systematic spectrum. However, for my finite sample I need to choose a critical value. In the plots I set the critical value equal to 1.08. Fortunately there are very visible humps at 4 for the years 2007 to 2012 and strong increases at 3 for the years 2003 to 2006, which can be detected for a wide range of critical values. Therefore, my estimator strongly indicates that there are 4 continuous factors from 2007 to 2012 and three continuous factors from 2003 to 2006. As a robustness test in Figure A.10 I also use an unperturbed eigenvalue ratio statistic. The results are the

same.[22]

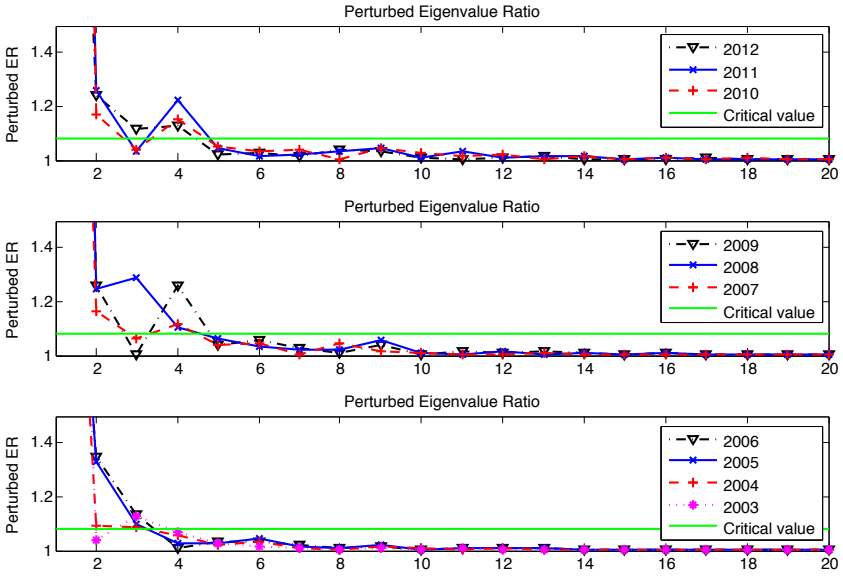

Figure 2: Number of continuous factors

In Figure 3 I apply the same analysis without separating the data into a continuous and jump component and obtain the same number of factors as in the continuous case. The perturbed eigenvalue ratios stop to cluster at the value 4 for 2007 to 2012 and at the value 3 for 2003 to 2006. This implies either that the continuous and jump factors are the same or that the continuous factors dominate the jump factors.

5.2.2 Persistence of Factors

The first four continuous factors are highly persistent for the time period 2007 to 2012, while there are three highly persistent factors for the time period 2003 to 2006. When comparing the systematic factor structures over time, I am interested if two sets of factors span the same vector space. I call a factor structure persistent if the vector spaces spanned by the factors stay constant. Persistence does not mean that the betas from a regression stay constant, which they usually do not do. If the factors estimated over a longer time horizon (e.g. 10 years) span the same vector space as factors estimated over all shorter horizons (e.g. 1 year) included in the longer period, persistence follows. The

[22]I have conducted the same analysis for more perturbation functions with the same findings. The results are available upon request.

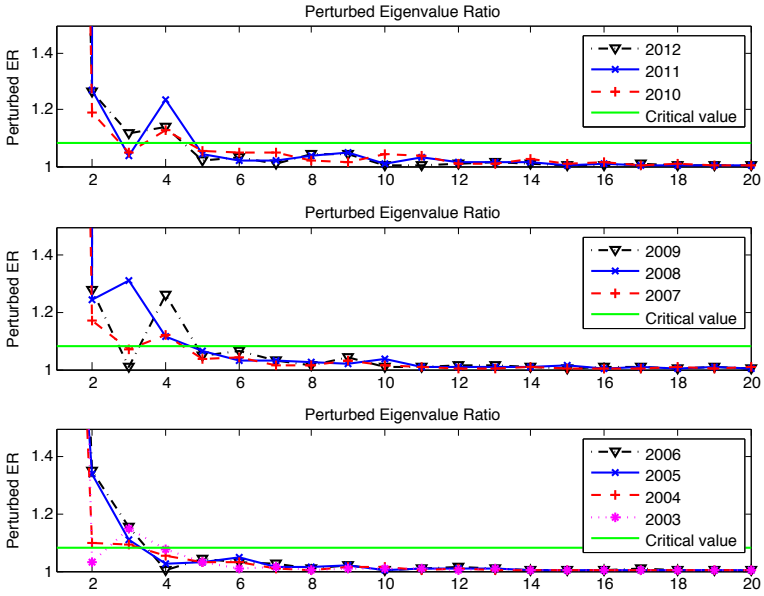

Figure 3: Number of total factors

difficulty in comparing the factor structure over time is that the same set of factors can lead to different principal components. An economic factor that explains a large fraction of the variation in one year and hence is associated with a large eigenvalue, might explain less variation and be linked to a smaller eigenvalue in another year. The generalized correlations allow us to compare the vector spaces that are spanned by different sets of factors. In Table 3 I calculate the generalized correlations for the first four largest statistical factors based on yearly quadratic correlations matrices and on a six-year quadratic correlation matrix. The number of generalized correlations that are close to one essentially suggests how many of the factors in the two sets are the same. The results indicate that it does not matter if I use a one year or six years horizon for the time period 2007 to 2012 for estimating the factors. In the same table I also compare the yearly loadings with the six-year loadings, which are essentially the same and hence represent the same portfolio space. As the loadings could be interpreted as portfolio weights, the same set of loadings implies the same factors for the same time period. Hence I have two ways to show the persistence in the factor structure.

In Table 4, I show that the first four yearly statistical factors and loadings are essentially identical to the first four monthly statistical factors and loadings in the year 2011. Identical results hold for the other years. This is another strong indication for the

persistence of the first four continuous factors.[23]

However, when doing the same analysis for the longer horizon 2003 to 2012 in Table 3, I observe that one factor disappears in 2003 to 2006. The first three generalized correlations are close to one, indicating that the two sets of factors share at least a three dimensional subspace, i.e. three of the factors coincide. The fourth generalized correlation for 2003 to 2006 however is significantly smaller implying that one of the four yearly factors cannot be written as a linear combination of the four factors estimated based on the 10 year horizon. This result is in line with my estimation results for the number of factors, where one factor seems to disappear before 2007. We observe exactly the same pattern for the loadings.

Factors, N=498					2007	2008	2009	2010	2011	2012
1. Generalized Correlation					1.00	1.00	1.00	1.00	1.00	1.00
2. Generalized Correlation					1.00	1.00	1.00	1.00	1.00	1.00
3. Generalized Correlation					0.99	0.99	1.00	1.00	1.00	0.98
4. Generalized Correlation					0.97	0.96	0.98	0.99	0.99	0.98
Loadings, N=498					2007	2008	2009	2010	2011	2012
1. Generalized Correlation					0.99	1.00	1.00	1.00	1.00	0.99
2. Generalized Correlation					0.97	0.99	0.99	0.99	0.99	0.97
3. Generalized Correlation					0.94	0.98	0.98	0.98	0.98	0.95
4. Generalized Correlation					0.93	0.97	0.96	0.97	0.95	0.93
Factors, N=302	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
1. Generalized Correlation	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2. Generalized Correlation	0.97	0.99	0.99	0.99	1.00	1.00	1.00	1.00	0.99	0.99
3. Generalized Correlation	0.95	0.97	0.98	0.99	0.99	0.99	0.97	0.98	0.99	0.98
4. Generalized Correlation	**0.47**	**0.63**	**0.17**	**0.67**	0.99	0.99	0.94	0.92	0.97	0.96
Loadings, N=302	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
1. Generalized Correlation	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	0.99
2. Generalized Correlation	0.91	0.96	0.97	0.97	0.98	0.99	0.99	0.99	0.98	0.96
3. Generalized Correlation	0.86	0.92	0.93	0.95	0.97	0.97	0.95	0.96	0.96	0.94
4. Generalized Correlation	**0.34**	**0.52**	**0.16**	**0.57**	0.95	0.96	0.90	0.88	0.93	0.91

Table 3: Persistence of continuous factors: Generalized correlations of the first four largest yearly continuous factors and their loadings with the first four statistical continuous factors and loadings for 2007-2012 respectively 2003-2012.

[23]I have done the same analysis for all the years and I will provide the results upon request.

1	2	3	4	5	6	7	8	9	10	11	12
Generalized correlations of monthly with yearly continuous factors											
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.99	0.99	0.99	0.99	0.99	1.00	1.00	0.99	0.99	1.00	0.99	0.99
0.98	0.93	0.99	0.97	0.98	0.98	0.98	0.99	0.99	0.96	0.90	0.96
Generalized correlations of monthly with yearly continuous loadings											
0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.96	0.96	0.97	0.97	0.97	0.97	0.98	0.98	0.98	0.99	0.98	0.97
0.95	0.95	0.96	0.96	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.96
0.94	0.86	0.94	0.90	0.93	0.94	0.94	0.95	0.96	0.90	0.84	0.91

Table 4: Persistence of continuous factors in 2011. Generalized correlation of monthly continuous factors and loadings with yearly continuous factors and loadings. The yearly number of factors is $K = 4$.

5.2.3 Interpretation of Factors

The four persistent continuous factors for 2007 to 2012 can be approximated very well by industry factors. The loading estimators can essentially be interpreted as portfolio weights for the factor construction. Simple eyeballing indicates that the first statistical factor seems to be an equally weighted market portfolio, a result which has already been confirmed in many studies. The loadings for the second to fourth statistical factors have a very particular pattern: Banks and insurance companies have very large loadings with the same sign, while firms related to oil and gas have large loadings with the opposite sign. Firms related to electricity seem to have their own pattern unrelated to the previous two. Motivated by these observations I construct four economic factors as

- Market (equally weighted)
- Oil and gas (40 equally weighted assets)
- Banking and Insurance (60 equally weighted assets)
- Electricity (24 equally weighted assets)

The details are in Appendix A.1.

Generalized correlations of 4 continuous factors with market, oil and finance factors										
N=498	2007-2012				2007	2008	2009	2010	2011	2012
1. Gen. Corr.	**1.00**				1.00	1.00	1.00	1.00	1.00	1.00
2. Gen. Corr.	**0.98**				0.98	0.97	0.98	0.97	0.98	0.93
3. Gen. Corr.	**0.95**				0.91	0.95	0.94	0.93	0.97	0.87
Generalized correlations of 4 continuous factors with market, oil, finance and electricity factors										
N=498	2007-2012				2007	2008	2009	2010	2011	2012
1. Gen. Corr.	**1.00**				1.00	1.00	1.00	1.00	1.00	1.00
2. Gen. Corr.	**0.98**				0.98	0.97	0.99	0.97	0.98	0.93
3. Gen. Corr.	**0.95**				0.91	0.95	0.95	0.93	0.94	0.90
4. Gen. Corr.	**0.80**				0.87	0.78	0.75	0.75	0.80	0.76
Generalized correlations of 4 continuous factors with market, oil, finance and electricity factors										
N=302	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
1. Gen. Corr.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2. Gen. Corr.	0.97	0.99	1.00	1.00	0.99	0.97	0.98	0.96	0.98	0.95
3. Gen. Corr.	0.57	0.75	0.77	0.89	0.85	0.92	0.95	0.92	0.93	0.83
4. Gen. Corr.	**0.10**	**0.23**	**0.16**	**0.35**	0.82	0.74	0.72	0.68	0.78	0.78
Generalized correlations of 4 continuous factors with market, oil and finance factors										
N=302	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
1. Gen. Corr.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2. Gen. Corr.	0.97	0.99	1.00	1.00	0.99	0.97	0.98	0.96	0.97	0.94
3. Gen. Corr.	**0.46**	**0.49**	**0.47**	**0.49**	0.84	0.92	0.94	0.89	0.93	0.83
Generalized correlations of 4 continuous factors with market, oil and electricity factors										
N=302	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
1. Gen. Corr.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2. Gen. Corr.	0.97	0.99	1.00	1.00	0.98	0.97	0.95	0.94	0.96	0.93
3. Gen. Corr.	**0.36**	**0.64**	**0.97**	**0.84**	0.83	0.76	0.73	0.69	0.78	0.78
Generalized correlations of 4 continuous factors with market, finance and electricity factors										
N=302	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
1. Gen. Corr.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2. Gen. Corr.	0.57	0.75	0.98	0.89	0.88	0.92	0.98	0.94	0.95	0.85
3. Gen. Corr.	**0.19**	**0.27**	**0.57**	**0.45**	0.83	0.74	0.73	0.72	0.78	0.78

Table 5: Interpretation of statistical continuous factors. Generalized correlation of economic factors (market, oil, finance and electricity factors) with first four largest statistical factors for different time periods.

The generalized correlations of the market, oil and finance factors with the first four largest statistical factors for 2007 to 2012 are very high as shown in the first analysis of Table 5. This indicates that three of the four statistical factors can almost perfectly be replicated by the three economic factors. This relationship is highly persistent over time.

In Table 5 the top of the first column uses the factors and generalized correlations based on a 6 year horizon, while in the last six columns I estimate the yearly statistical factors and calculate their generalized correlations with the yearly market, oil and finance factors. The generalized correlations close to one indicate that at least three of the statistical factors do not change over time and are persistent.

Identifying the fourth continuous factor is challenging and the closest approximation seems to be an electricity factor. The second analysis in Table 5 shows the generalized correlations of the four continuous statistical factors for 2007 to 2012 with the four economic factors. The fourth generalized correlation essentially measures how well the additional electricity factor can explain the remaining statistical factor. The fourth yearly generalized correlation takes values between 0.75 and 0.87, which means that the electricity factor can help substantially to explain the statistical factors, but it is not sufficient to perfectly replicate them. The first column shows the result for the total six year time horizon while the last six columns list the yearly results. In conclusion it seems that the relationship between the four economic and statistical factors is persistent over time.

The results in Subsection 5.2.2 indicate that the factor structure in 2003 to 2006 might be different compared to the later period. Based on the intersection of all the firms for 2003 to 2012 I analyze the generalized correlations of the first four yearly continuous statistical factors with the four yearly continuous industry factors. The third analysis in Table 5 shows that as expected one factor disappears in the early four years. A fourth generalized correlation between 0.16 and 0.35 for 2003 to 2006 suggests strongly that the statistical factors and industry factors have at most three factors in common. The fourth, fifth and sixth analyses in Table 5 try to identify the disappearing factor. Looking at the fifth analysis it seems that dropping the finance factor for the time period 2003 to 2006 leads to the smallest reduction in generalized correlations, i.e. the three statistical factors for 2003 to 2006 are not well-explained by a finance factor. On the other hand this finance factor is crucial for explaining the statistical factors for 2007 to 2012.

As a robustness test I extend the analysis to daily data and also include daily time-series of the Fama-French-Carhart factors. First I calculate daily continuous returns by adding up the continuous log-price increments for each day and creating this way the four continuous daily statistical factors. Then I calculate generalized correlations of the daily continuous factors with economic daily factors. I always include the daily continuous market, oil and finance factors in the analysis and in addition include either

1. Case 1: no additional factor
2. Case 2: a daily continuous electricity factor
3. Case 3: size, value and momentum factors
4. Case 4: daily continuous electricity, size, value and momentum factors

The results are summarized in Table 6. Obviously, the size, value and momentum factor do not explain much variation beyond the industry factors. The fourth generalized correlation in case 2 is with 0.81 almost the same as in case 4. In particular, the fourth factor seems to be much better explained by an electricity factor than by a size, value or momentum factor, which only account for a fourth generalized correlation of 0.43 in case 3.

	Case 1	Case 2	Case 3	Case 4
1. Generalized Correlation	1.00	1.00	1.00	1.00
2. Generalized Correlation	0.99	0.99	0.98	0.99
3. Generalized Correlation	0.95	0.95	0.95	0.95
4. Generalized Correlation		**0.81**	**0.43**	**0.83**

Table 6: Generalized correlations of the 4 daily continuous factors with daily economic factors for 2007-2012. N=498.

However, the Fama-French-Carhart factors are based on daily excess returns, which also include jumps and overnight returns. Additionally, daily returns are also mathematically different from daily increments in log asset prices. Hence the comparison with daily continuous returns might be misleading. Hence, I construct the 4 statistical and four economic industry factors using daily excess returns from CRSP. The estimated continuous loadings serve again as the portfolio weights for the statistical factors. Based on the daily excess returns from 2007 to 2012, I run simple OLS regressions in order to explain the four statistical factors. The short regression uses a market, oil, finance and electricity factor, while the long regression applies the same regressors and additionally the size, value and momentum factors. Table 7 shows that almost all the variation can be explained by the industry factors, while adding the Fama-French-Carhart factors does not change the explanatory power. In the second part of Table 7 I repeat a similar analysis as in Table 6 but using the daily excess returns. As before the generalized correlations with the 4 economic factors are very large indicating that a linear combination of daily excess returns of the industry portfolios can approximate the excess returns of the statistical factors very well. On the other hand the best linear combination of daily

excess returns of the Fama-French Carhart Factors provides only a poor approximation to the statistical factor returns.

R^2 for daily excess returns of portfolios based on continuous loadings				
N=498	1. Factor	2. Factor	3. Factor	4. Factor
Short	1.00	0.96	0.95	0.97
Long	1.00	0.97	0.95	0.97
Generalized correlations with 4 economic industry factors				
	1.00	**0.97**	**0.92**	**0.79**
Generalized correlations with 4 Fama-French Carhart Factors				
	0.95	0.74	0.60	0.00

Table 7: (1) R^2 for portfolios of daily CRSP excess returns based on continuous loadings $\hat{\Lambda}^C$. (2) Generalized correlations of daily CRSP excess returns based on continuous loadings with 4 economic factors. (3) Generalized correlations of daily CRSP excess returns based on continuous loadings with 4 Fama-French Carhart Factors. The time period is 2007-2012 and N=498.

	4 statistical and 3 economic factors			4 statistical and 4 economic factors		
	$\hat{\bar{\rho}}$	SD	95% CI	$\hat{\bar{\rho}}$	SD	95% CI
2007-2012	2.72	0.001	(2.71, 2.72)	3.31	0.003	(3.30, 3.31)
2007	2.55	0.06	(2.42, 2.67)	3.21	0.01	(3.19, 3.22)
2008	2.66	0.08	(2.51, 2.81)	3.18	0.29	(2.62, 3.75)
2009	2.86	0.10	(2.67, 3.05)	3.42	0.15	(3.14, 3.71)
2010	2.80	0.04	(2.72, 2.88)	3.38	0.01	(3.37, 3.39)
2011	2.82	0.00	(2.82, 2.82)	3.47	0.06	(3.35, 3.58)
2012	2.62	0.03	(2.56, 2.68)	3.25	0.01	(3.24, 3.26)

Table 8: Total generalized correlations (=sum of squared generalized correlations) with standard deviations and confidence intervals for the four statistical factors with three economic factors (market, oil and finance) and four economic factors (additional electricity factor). Number of assets $N = 498$.

As a last step I apply the statistical test of Appendix B.6 to test if the three respectively four continuous economic factors can perfectly replicate the statistical factors. So far I have not provided confidence intervals for the generalized correlations. Unfortunately, in this high-frequency setting there does not exist a theory for confidence intervals for the individual generalized correlations. However, I have developed an asymptotic distribution theory for the sum of squared generalized correlations, which I label as total

generalized correlation. The left part of Table 8 lists the total generalized correlation for different time periods for three economic factors. A total generalized correlation of three indicates that three of the four statistical factors can be perfectly replicated by the three economic factors. Only in 2009 I cannot reject the null hypothesis of a perfect factor replication. In the right half of Table 8 I apply the same test to four economic factors. Now a total generalized correlation of four implies that the four statistical factors are identical to the four economic factors. I reject the null hypothesis of a perfect linear combination. Hence, although my set of economic factors approximates the statistical factors very well, there seems to be a missing component.

5.3 Jump Factors

There seems to be a lower number of jump factors, which do not coincide with the continuous factors. Only the jump market factor seems to be persistent, while neither the number nor the structure of the other jump factors have the same persistence as for the continuous counterpart. Figures A.11, A.12 and A.13 estimate the number of jump factors for different thresholds. In most years the estimator indicates only one jump factor. Under almost all specifications there seems to be at most four jump factors and hence I will restrict the following analysis to the first four largest jump factors.

In Table 9 I analyze the persistence of the jump factors by comparing the estimation based on 6 years with the estimation based on yearly data. For the smallest threshold there seem to be two persistent factors as the first two generalized correlations are close to 1, but the structure is much less persistent than for the continuous data. For the largest threshold only the first generalized correlation is close to one suggesting only one persistent factor. Table A.1 shows that for the shorter time horizon of a month only one factor is persistent independently of the threshold. This could be explained by the fact that the systematic jumps do not necessarily happen during every month and hence the systematic structure measured on a monthly basis can be very different from longer horizons.

My estimator for identifying the jumps might erroneously classify high volatility time periods as jumps. Increasing the threshold in the estimator reduces this error, while I might misclassify small jumps as continuous movements. Increasing the threshold, reduces the persistence in the jump factors up to the point where only a market jump factors remains. It is unclear if the persistence for small jump thresholds is solely due to

Yearly vs. 6-year jump factors						Yearly vs. 6-year jump loadings					
2007	2008	2009	2010	2011	2012	2007	2008	2009	2010	2011	2012
						a=3					
1.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.95**	**0.99**	**0.98**	**0.98**	**0.97**	**0.95**
0.96	**1.00**	**0.95**	0.**98**	0.76	**0.85**	**0.84**	**0.97**	**0.84**	**0.85**	0.49	0.68
0.81	**0.88**	**0.84**	0.69	0.59	0.70	0.58	0.74	0.60	0.28	0.40	0.46
0.12	**0.81**	0.14	0.25	0.05	0.17	0.05	0.63	0.07	0.15	0.03	0.07
						a=4					
1.00	**1.00**	**0.99**	**1.00**	0.**99**	**0.99**	**0.88**	**0.97**	**0.81**	**0.95**	**0.85**	**0.85**
0.66	**0.99**	0.63	**1.00**	0.51	0.43	0.30	**0.87**	0.20	**0.93**	0.16	0.13
0.34	0.52	0.09	**0.97**	0.43	0.20	0.10	0.23	0.02	0.66	0.11	0.05
0.14	0.03	0.05	0.17	0.13	0.03	0.05	0.01	0.01	0.04	0.03	0.01
						a=4.5					
0.99	**0.99**	**0.98**	**1.00**	**0.99**	**0.99**	**0.85**	**0.95**	0.72	**0.95**	0.75	**0.82**
0.79	**0.97**	0.77	**1.00**	0.40	0.49	0.30	0.78	0.24	**0.93**	0.10	0.14
0.28	0.44	0.28	**0.96**	0.26	0.24	0.08	0.13	0.06	0.67	0.06	0.05
0.05	0.16	0.01	0.53	0.11	0.07	0.01	0.05	0.00	0.12	0.03	0.02

Table 9: Generalized correlations of 4 largest yearly jump factors with 4 jump factors for 2007-2012 and generalized correlations of 4 yearly jump loadings with 4 jump loadings for 2007-2012 for different thresholds. Here $K = 4$ and $N = 498$. Values larger than 0.8 are in bold.

misclassified high volatility movements.

Table 10 confirms that the jump factors are different from the continuous factors. Here I estimate the generalized correlations of the first four statistical jump factors with the market, oil, finance and electricity jump factors for 2007 to 2012. I can show that the first statistical jump factor is essentially the equally weighted market jump factor which is responsible for the first generalized correlation to be equal to 1. However, the correlations between the other statistical factors and the industry factors are significantly lower.

5.4 Comparison with Daily Data and Total Factors

The continuous factors dominate the jump factors and daily returns give similar but noisier estimators than the continuous high-frequency analysis. In this section I compare the estimators based on continuous, jump and total high-frequency data and daily CRSP returns. As I make the comparisons for each year separately, I can use my largest cross-sectional sample as listed in Table 1. I compare the loadings based on daily, total and

Generalized correlations of 4 economic jump factors with 4 statistical jump factors							
	2007-2012	2007	2008	2009	2010	2011	2012
a=3	**1.00**	**1.00**	**1.00**	**0.99**	**1.00**	**1.00**	**1.00**
	0.85	**0.95**	0.62	**0.86**	**0.81**	**0.86**	**0.83**
	0.61	0.77	0.40	0.76	0.31	0.61	0.59
	0.21	0.10	0.22	0.50	0.10	0.20	0.28
a=4	**0.99**	**0.99**	**0.95**	**0.94**	**1.00**	**0.99**	**0.99**
	0.74	0.53	0.41	0.59	**0.90**	0.53	0.57
	0.31	0.35	0.29	0.44	0.39	0.35	0.42
	0.03	0.19	0.20	0.09	0.05	0.14	0.16
a=4.5	**0.99**	**0.99**	**0.91**	**0.91**	**1.00**	**0.98**	**0.99**
	0.75	0.54	0.41	0.56	**0.93**	0.55	0.75
	0.29	0.35	0.30	0.40	0.68	0.38	0.29
	0.05	0.18	0.22	0.04	0.08	0.03	0.05

Table 10: Generalized correlations of market, oil, finance and electricity jump factors with first 4 jump factors from 2007-2012 for N=498 and for different thresholds. Values larger than 0.8 are in bold.

jump high-frequency loadings with the continuous loadings. As the loadings can be interpreted as portfolio weights, the same set of loadings also implies the same factors. However, some assets might be close substitutes in which case different portfolios might still describe the same factors. Thus, I use the loadings estimated from the different data sets to construct continuous factors and estimate the distance between the different sets of continuous factors.

Figure A.14 shows the yearly estimators for the number of factors based on approximately 250 daily observations. The pattern is similar to the continuous estimators, but much noisier. This can be either due to the fact that the number of observations is much smaller then the approximate 20,000 in the high-frequency case, but also because the daily returns include the jumps and overnight movements. Table A.2 indicates that the continuous factors and loadings are close to but different from those based on daily CRSP returns. This is a positive finding as it indicates that my results are in some sense robust to the frequency. On the other hand the high-frequency estimator seems to estimate the pattern in the data more precisely and there is a gain from moving from daily to intra-day data.

In Table A.2 I also show that the total factors and continuous high-frequency factors are essentially identical. This result has two consequences. First, it confirms that my

findings about the continuous factors are robust to the jump threshold. Even if all the movements are classified as continuous, I obtain essentially the same estimators for the loadings. Second, the systematic continuous pattern dominates the systematic jump pattern. The first jump factor is a market jump factor and hence is described by the same loadings as the first continuous factor. Even if there are systematic jump factors that are different from the second to fourth continuous factors their impact on the spectrum is so small, that it is not detected when considering only the first four total factors. This is also partly due to the fact that the jump quadratic covariation is only a small fraction of the total quadratic covariation.

Finally in Table A.3 I confirm that the systematic jump factors are different from the systematic continuous factors. The higher the jump threshold the less likely it is that the large increments are due to continuous movements with high volatility. Thus for a larger jump threshold the correlation between the jump factors and continuous factors decreases up to the point where only the market factor remains as having a common continuous and jump component.

5.5 Microstructure Noise

Non-synchronicity and microstructure noise are two distinguishing characteristics of high-frequency financial data. First, the time interval separating successive observations can be random, or at least time varying. Second, the observations are subject to market microstructure noise, especially as the sampling frequency increases. The fact that this form of noise interacts with the sampling frequency distinguishes this from the classical measurement error problem in statistics. Inference on the volatility of a continuous semimartingale under noise contamination can be pursued using smoothing techniques.[24] However, neither the microstructure robust estimators nor the non-synchronicity robust estimators can be easily extended to my large dimensional problem. The main results of my paper assume synchronous data with negligible microstructure noise. Using for example 5-minute sampling frequency as commonly advocated in the literature on realized volatility estimation, e.g. Andersen et al. (2001) and the survey by Hansen and Lunde (2006), seems to justify this assumption.

[24]Several approaches have been developed, prominent ones by Aït-Sahalia et al. (2005b), Barndorff-Nielsen et al. (2008) and Jacod et al. (2009) in the one-dimensional setting and generalizations for a noisy non-synchronous multi-dimensional setting by Aït-Sahalia et al. (2010), Podolskij and Vetter (2009), Barndorff-Nielsen et al. (2011) and Bibinger and Winkelmann (2014) among others.

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Median eigenvalue	0.048	0.036	0.033	0.034	0.040	0.119	0.070	0.030	0.032	0.025
Var of MN (T=12)	0.029	0.034	0.034	0.037	0.046	0.141	0.088	0.036	0.036	0.030
Var of MN (T=1)	0.001	0.001	0.001	0.001	0.001	0.003	0.002	0.001	0.001	0.001

Table 11: Estimation of the upper bound of the variance of microstructure noise for different years and different reference levels. (1) Median eigenvalue of yearly quadratic covariation matrix. (2) Upper bound on microstructure noise variance if the reference level for high and low frequency is a month (i.e. M is around 1,617). (3) Upper bound on microstructure noise variance if the reference level for high and low frequency is a year (i.e. M is around 19,250).

Volatility signature plots as used in Hansen and Lunde (2006) are very useful tools for identifying frequencies that are contaminated by noise. One common approach in the literature is to sample at lower frequencies to minimize the contamination by microstructure noise at the cost of using less data. Clearly when estimating the quadratic covariation without applying microstructure noise corrections there is a tradeoff between the higher noise variance and higher precision of using a finer frequency. An important question in this respect is the variance of the unobservable noise. For example Hansen and Lunde (2006) have estimated the microstructure noise variance for different assets. In Pelger (2015) I propose an estimator for the microstructure noise that utilizes the information in the cross-section. Under the assumptions outlined in Theorem 4, the increments of microstructure noise create a very specific spectral pattern. This allows us to derive upper bounds on the variance of the microstructure noise. These bounds are solely functions of the estimated eigenvalues and the ratio $\frac{M}{N}$. From a practical perspective it is ambiguous how to choose M for a given time horizon. For example Lee and Mykland (2009) use a year as the reference horizon for high to low frequencies, i.e. in my case M would be around $250 \cdot 77 = 19,250$. One could also argue that a month is a better cutoff between high and low frequencies which would set M to around $21 \cdot 77 = 1,617$. Obviously the results of my estimator for the microstructure noise variance depend on this choice.

Table 11 shows the estimation results for the different years. For a monthly reference level for high to low frequencies the upper bounds are very similar to the estimates in Hansen and Lunde (2006). For a yearly reference level they are significantly lower. In either case the microstructure noise contamination can be neglected when using 5 minute data. The fact that my results are robust to different time horizons, e.g. 5 minutes, 15

minutes and daily horizons, further confirms that the results are robust to microstructure noise.

6 Empirical Application to Volatility Data

Using implied volatilities from option price data, I analyze the systematic factor structure of the volatilities. There seems to be only one persistent market volatility factor, while during the financial crisis an additional temporary banking volatility factor appears. Based on the estimated factors, I can decompose the leverage effect, i.e. the correlation of the asset return with its volatility, into a systematic and an idiosyncratic component. The negative leverage effect is mainly driven by the systematic component, while it is substantially smaller for idiosyncratic risk. These findings are important as they can rule out popular explanations of the leverage effect, which do not distinguish between systematic and non-systematic risk.

6.1 Volatility Factors

As the volatility of asset price processes is not observed, I cannot directly apply my factor analysis approach to the data. There are essentially two ways to estimate the volatility. The first is based on high-frequency asset prices and estimates the spot volatility using the realized quadratic covariations for a short time window. The second approach infers the volatility under the risk-neutral measure using option price data. The VIX is the most prominent example for the second approach. I have pursued both approaches, but most of the results reported in this paper are based on the second one.

Using the realized quadratic variation for a short horizon, e.g. a day, I can obtain estimators for the spot volatilities, which can be used for my large-dimensional factor analysis. The details for the estimation of the spot volatilities and the construction of volatility of volatility estimators can be found in chapter 8 of Aït-Sahalia and Jacod (2014). The volatility of volatility estimator requires a bias correction. These estimators appear to be very noisy in practice and that is why I prefer an alternative approach.

Medvedev and Scaillet (2007) show that under a jump-diffusion stochastic volatility model the Black-Scholes implied volatility of an at-the-money option with a small time-to-maturity is close to the unobserved volatility. Using this insight I use implied volatilities

Year	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Original	408	479	507	525	543	557	565	561	549	565
Cleaned	399	465	479	495	508	528	536	530	523	529
Dropped	2.21%	2.92%	5.52%	5.71%	6.45%	5.21%	5.13%	5.53%	4.74%	6.37%

Table 12: Observations after data cleaning.

for my factor estimation approach.[25] Under relatively general conditions estimating the factor structure (and also later the leverage effect) under the risk-neutral measure yields the same results as under the physical measure.

In some sense I am trying to create a VIX-type times-series for all the assets in the cross-section. The older version of VIX, the VXO, was actually a measure of implied volatility calculated using 30-day S&P100 index at-the-money options. The VIX uses the concept of generalized implied volatility. I have also created a panel of generalized implied volatilities for my cross-section. However, the theoretical justification of this approach assumes an infinite number of strike prices for each asset and the quality of the estimator deteriorates for a small number of available strikes as it is the case for many assets in my sample. Therefore, the Black-Scholes implied volatility appears to be a much more robust estimator. For the assets in my sample, where I have a large number of strikes available, the generalized volatility and simple implied-volatility are very close, while for those with only few strike prices, the generalized volatility estimators seem to be unreliable.

Using daily implied volatilities from OptionMetrics for the same assets and time period as in the previous section, I apply my factor analysis approach. OptionMetrics provides implied volatilities for 30 days at-the-money standard call and put options using a linearly interpolated volatility surface. I average the implied call and put volatilities for each asset and each day. More details about the data and some results are in Appendix A.5. Unfortunately, the data is not available to construct intra-day implied volatilities for the whole cross-section. However, as the previous section has illustrated, the results with daily data seem to capture similar results as with higher frequency data. Table 12 reports the data after the data cleaning.

Figure 4 estimates the number of volatility factors. There seems to be only one

[25] A rigorous study would require me to take into account the estimation error for the implied volatility and to derive the simultaneous limit of M, N, the at-the-moneyness and the maturities of the options. As such an extension is beyond the scope of the paper, I treat the implied volatilities of short-maturity at-the-money options as the true observed volatilities under the risk-neutral measure.

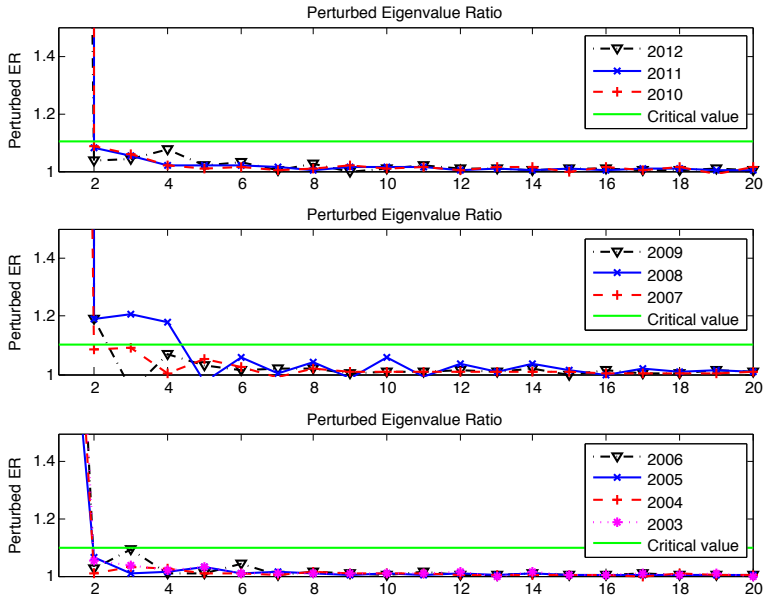

Figure 4: Number of volatility factors

strong persistent factor, which is essentially a market volatility factor and very highly correlated with the VIX. In 2008 there seems to be a second temporary volatility factor. The volatility factors seem to be different from the continuous, jump and daily factors. Table 13 shows the generalized correlations of the volatility loadings with the continuous, jump and daily loadings for each year. The volatility loadings cannot be interpreted as a portfolio of assets, but rather as a portfolio of volatilities. There does not seem to be a strong correlation with the other factors. Table 13 also calculates the generalized correlation of the first four volatility factors with market, oil, finance and electricity volatility factors constructed as equally weighted portfolios of volatilities for these industries. In 2008 and 2009 there seems to appear a temporary finance factor as the second generalized correlation jumps up.[26] This finding is in line with the results for the number of factors and not surprising given the financial crisis.

[26]I have calculated the generalized correlations between the first two statistical volatility factors and different combinations of the four economic volatility factors. It seems that the finance factor can explain most of the second statistical volatility factor.

Generalized correlations of 4 economic volatility factors with 4 statistical volatility factors										
					2007	2008	2009	2010	2011	2012
1. Generalized Correlation					1.00	1.00	1.00	1.00	1.00	1.00
2. Generalized Correlation					0.19	**0.90**	**0.92**	0.33	0.67	0.28
3. Generalized Correlation					0.07	0.34	0.13	0.06	0.11	0.05
4. Generalized Correlation					0.01	0.05	0.00	0.00	0.01	0.01
Generalized correlations between volatility and continuous loadings										
	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
1. Generalized Correlation	0.95	0.91	0.84	0.92	0.96	0.98	0.97	0.98	0.98	0.97
2. Generalized Correlation	0.59	0.68	0.84	0.34	0.69	0.79	0.57	0.37	0.31	0.21
3. Generalized Correlation	0.46	0.56	0.31	0.34	0.08	0.57	0.52	0.09	0.31	0.10
4. Generalized Correlation	0.14	0.11	0.14	0.03	0.08	0.12	0.02	0.08	0.02	0.05
Generalized correlations between volatility and jump loadings										
	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
1. Generalized Correlation	0.92	0.90	0.88	0.82	0.93	0.94	0.96	0.93	0.93	0.93
2. Generalized Correlation	0.47	0.61	0.72	0.33	0.21	0.20	0.46	0.15	0.38	0.09
3. Generalized Correlation	0.29	0.52	0.27	0.31	0.16	0.19	0.46	0.15	0.11	0.09
4. Generalized Correlation	0.29	0.10	0.04	0.07	0.16	0.19	0.03	0.10	0.03	0.01
Generalized correlations between volatility and daily loadings										
	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
1. Generalized Correlation	0.95	0.93	0.84	0.92	0.96	0.98	0.96	0.98	0.97	0.96
2. Generalized Correlation	0.44	0.64	0.84	0.34	0.61	0.64	0.46	0.29	0.49	0.23
3. Generalized Correlation	0.44	0.64	0.39	0.23	0.29	0.64	0.46	0.19	0.26	0.22
4. Generalized Correlation	0.22	0.06	0.21	0.18	0.02	0.38	0.02	0.07	0.03	0.09

Table 13: Generalized correlations between four economic (market, oil, finance and electricity) and four statistical volatility factors and between the loadings of the volatility factors and the loadings for continuous, jump and daily data.

6.2 Leverage Effect

One of the most important empirical stylized facts about the volatility is the leverage effect, which describes the generally negative correlation between an asset return and its volatility changes. The term "leverage" originates in one possible economic interpretation of this phenomenon, developed in Black (1976) and Christie (1982). When asset prices decline, companies become mechanically more leveraged as equity is a residual claim and the relative value of their debt rises relative to that of their equity. Therefore, their stock should become riskier, and as a consequence more volatile. Although this is only a hypothesis, this explanation has coined the term "leverage effect" to describe the

statistical regularity in the correlation between asset return and volatility.[27]

There is no consensus on the economic explanation for this statistical effect. The magnitude of the effect seems to be too large to be explained by financial leverage.[28] Alternative economic interpretations as suggested for example by French et al. (1987) and Campbell and Hentschel (1992) use a risk-premium argument. An anticipated rise in volatility increases the risk premium and hence requires a higher rate of return from the asset. This leads to a fall in the asset price. The causality for these two interpretations is different: The leverage hypothesis claims that return shocks lead to changes in volatility, while the risk premium story implies that return shocks are caused by changes in conditional volatility. Showing that the leverage effect appears only for systematic priced risk but not for unpriced nonsystematic risk could rule out the leverage story. These different explanations have been tested by Bekaert and Wu (2000) who use a parametric conditional CAPM model under a GARCH specification to obtain results consistent with the risk-premium story. I estimate the leverage effect completely non-parametrically and decompose it into its systematic and nonsystematic part based on my general statistical factors. I show that the leverage effect appears predominantly for systematic risk, while it can be non-existent for idiosyncratic risk. Although this does not prove that the risk-premium story is the only explanation, it provides a strong argument against the leverage story.

The estimation of the leverage effect is difficult because volatility is unobservable. As in the previous subsection there are essentially two non-parametric approaches to estimate the correlation between asset returns and the changes in their volatility. First, the common approach is to conduct preliminary estimation of the volatility over small windows, then to compute the correlation between returns and the increments of the estimated volatility. Wang and Mykland (2014), Aït-Sahalia, Fan and Li (2013) and Aït-Sahalia and Jacod (2014) are examples of this approach. Such estimators appear to be very noisy in practice. Second, Kalnina and Xiu (2014) use volatility instruments based on option data, such as the VIX or Black-Scholes implied volatilities. Their approach seems to provide better estimates than the first one.

[27]There are studies (e.g., Nelson (1991) and Engle and Ng (1993)) showing that the effect is generally asymmetric. Declines in stock prices are usually accompanied by larger increases in volatility than the declines in volatility with rising stock markets. Yu (2005) has estimated various discrete-time models with a leverage effect.

[28]Figlewski and Wang (2000) raise the question whether the effect is linked to financial leverage at all. They show that there is only an effect on volatility when leverage changes due to changes in stock prices but not when leverage changes because of a change in debt or number of shares.

Based on my factor estimation approach I decompose the leverage effect into a systematic and idiosyncratic component. The estimated high-frequency factors allow me to separate the returns and volatilities into a systematic and idiosyncratic component, which I can then use to calculate the different components of the leverage effect. I use two different methodologies. First, I employ only the high-frequency equity data and apply Aït-Sahalia and Jacod's (2014) approach as described in Theorem 6. As already noted in Kalnina and Xiu (2014), this estimator for correlation leads to a downward bias unless a long time horizon with a huge amount of high-frequency data is used. However, my main findings, namely that systematic risk drives the leverage effect is still apparent. Second, I estimate the correlation between daily continuous returns and daily implied volatilities. This approach follows the same reasoning as Kalnina and Xiu, except that they use intra-day data and and develop a bias reduction technique for volatility instruments that are unknown functions of the unobserved volatility. For a limited number of assets I have high-frequency prices of volatility instruments, for example the VIX. For these assets using simple daily increments in implied volatilities or the more sophisticated high-frequency bias-reduced estimators with volatility instruments yield very close results. Thus, I am confident that my main findings are robust to the estimation approach employed.

In this paper I use an average leverage effect, where I measure the leverage effect with the continuous quadratic covariation for the time horizon T[29]:

$$LEV = \frac{[\sigma_i^2, X_i]_T^C}{\sqrt{[X_i, X_i]_T^C}\sqrt{[\sigma_i^2, \sigma_i^2]_T^C}}.$$

Based on systematic factors, I can decompose this average leverage effect into a systematic and idiosyncratic part:

$$LEV_i^{syst} = \frac{[\sigma_i^2, X_i^{syst}]_T^C}{\sqrt{[X_i, X_i]_T^C}\sqrt{[\sigma_i^2, \sigma_i^2]_T^C}} \qquad LEV_i^{idio} = \frac{[\sigma_i^2, X_i^{idio}]_T^C}{\sqrt{[X_i, X_i]_T^C}\sqrt{[\sigma_i^2, \sigma_i^2]_T^C}}$$

where $X_i^{syst}(t) = \Lambda_i^\top F(t)$ and $X_i^{idio}(t) = e_i(t) = X_i(t) - \Lambda_i^\top F(t)$ for asset i. My estimator

[29] Aït-Sahalia and Jacod (2014) and Kalnina and Xiu (2014) also work with an aggregate leverage effect.

for this simple decomposition based on implied volatilities is therefore[30]

$$\widehat{LEV}_i = \frac{\hat{\sigma}_i^{2\top}\hat{X}_i^C}{\sqrt{\hat{X}_i^{C\top}\hat{X}_i^C}\sqrt{\hat{\sigma}_i^{2\top}\hat{\sigma}_i^2}},\ \widehat{LEV}_i^{syst} = \frac{\hat{\sigma}_i^{2\top}\hat{F}^C\hat{\Lambda}_i^C}{\sqrt{\hat{X}_i^{C\top}\hat{X}_i^C}\sqrt{\hat{\sigma}_i^{2\top}\hat{\sigma}_i^2}},\ \widehat{LEV}_i^{idio} = \frac{\hat{\sigma}_i^{2\top}\hat{e}_i^C}{\sqrt{\hat{X}_i^{C\top}\hat{X}_i^C}\sqrt{\hat{\sigma}_i^{2\top}\hat{\sigma}_i^2}}$$

where $\hat{\Lambda}^C$ is obtained from the high-frequency data, $\hat{X}^C$, $\hat{F}^C$ and $\hat{e}_i^C$ are based on the accumulated daily continuous increments and $\hat{\sigma}^2$ are the daily increments of an estimator of the implied volatility. In the following I use my four continuous statistical factors for estimating the systematic continuous part of $\hat{X}^C$. The decomposition of the leverage effect based on spot volatilities applies the systematic and non-systematic returns to Theorem 6.

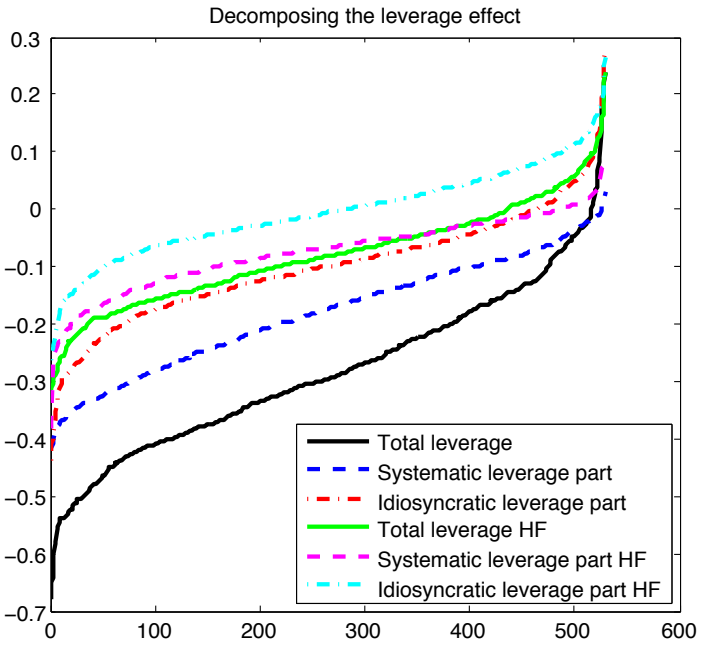

Figure 5: Decomposition of the leverage effect in 2012 using implied volatilities and high-frequency volatilities. I use 4 continuous asset factors.

Figure 5 plots the sorted decomposition of the leverage effect based on implied and realized volatility estimators. For each type of leverage I have sorted the values sepa-

[30]I have also done all the calculations in this subsection where I first demean the data before calculating the correlations. The results do not change and are available upon request.

rately. Hence the different curves should be interpreted as describing the cross-sectional distribution of the leverage effect for the different components.[31] There is clearly a difference between the systematic leverage and idiosyncratic leverage. The total leverage is close to the systematic leverage. Note however, that the absolute magnitude of high-frequency estimations of the leverage effects are significantly below the estimates based on the implied-volatility, which are more in line with the values usually assumed in the literature. In the online appendix the Figures C.35 to C.44 show the corresponding plots for all the years. We observe the same pattern: The systematic part of the leverage effect is larger than the idiosyncratic part and high-frequency based volatilities underestimate the leverage effect while implied volatility based estimates have the correct size.

The previous results could be driven by the fraction of total risk explained by the systematic part. Even if the idiosyncratic part of the return has the same correlation with the volatility, it can lead to a low covariance if it is only a small part of the total variation. For example from 2003 to 2006 the systematic factors explain only a small fraction of the total variation as can be seen in Table 2. This can explain the downward shift in the systematic leverage curves in Figures C.41 to C.44. A more meaningful measure is therefore the componentwise leverage effect. I decompose X_i and σ_i^2 into a systematic and idiosyncratic part: $X_i^{syst}, X_i^{idio}, {\sigma^2}_i^{syst}$ and ${\sigma^2}_i^{idio}$ based on the results in Section 5 and the previous subsection. X_i^{total} and ${\sigma_i^2}^{total}$ denote the total asset price respectively total volatility. Then for $y, z = syst$, $idio$ and $total$ I calculate the *componentwise leverage effect*

$$LEV_i^{y,z} = \frac{[X_i^y, {\sigma^2}_i^z]_T^C}{\sqrt{[X_i^y, X_i^y]_T^C}\sqrt{[{\sigma^2}_i^z, {\sigma^2}_i^z]_T^C}}$$

and obtain $LEV^{total,total}$, $LEV^{syst,total}$, $LEV^{idio,total}$, $LEV^{syst,syst}$, $LEV^{syst,idio}$, $LEV^{idio,syst}$ and $LEV^{idio,idio}$. For the componentwise leverage effect I use the implied volatility data, the four continuous statistical asset factors and the largest volatility factor.

Figure 6 depicts the sorted results for 2012. The other years are in the online appendix in Figures C.15 to C.24. There are three main findings:

1. $LEV^{total,total}$, $LEV^{syst,total}$ and $LEV^{syst,syst}$ yield the highest values and are similar to each other.
2. $LEV^{idio,total}$ and $LEV^{idio,idio}$ take intermediate values and are also very similar to

[31] For the same value on the x-axis different curves usually represent different firms.

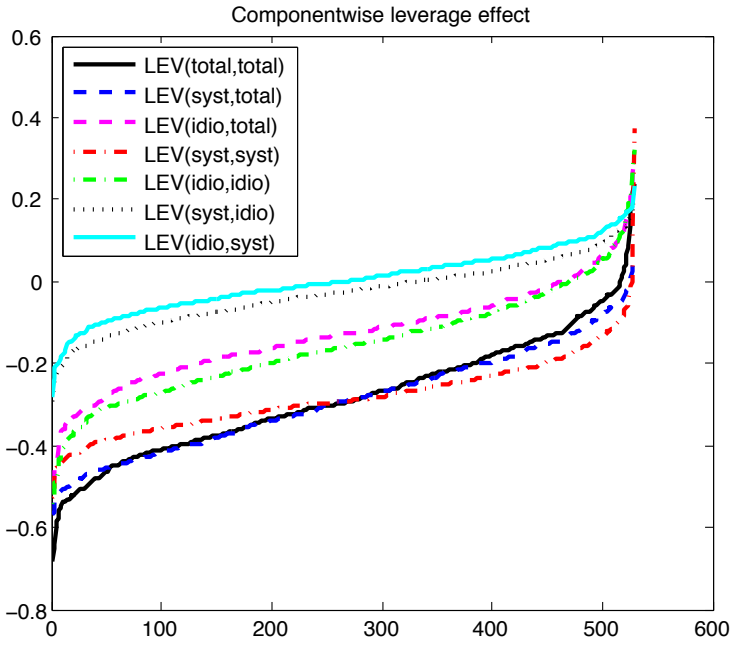

Figure 6: Componentwise leverage effect in 2012: Sorted correlations between total, systematic and idiosyncratic log-prices with total, systematic and idiosyncratic implied volatility. 4 asset factors and 1 volatility factor.

each other.

3. $LEV^{syst,idio}$ and $LEV^{idio,syst}$ are on average zero and very close to each other.

In conclusion it seems that the largest leverage effect is due to the systematic asset price and systematic volatility part. Without the systematic return component the leverage effect drops significantly. The systematic returns seem to be almost orthogonal to the idiosyncratic volatilitiy component and similarly the correlation between idiosyncratic returns and systematic volatility is on average zero. One interpretation of this finding is that the main contributor to the leverage effect is non-diversifiable risk. This finding lends support to the risk-premium explanation of the leverage effect and is a counterargument for the financial leverage story.

The strongest result is the very small correlation of idiosyncratic volatility with systematic returns and of systematic volatility with idiosyncratic returns. In the years from 2007 to 2012 the correlation of the systematic return with the total volatility is much larger than the correlation of the idiosyncratic return with the total volatility. This

particular pattern becomes weaker for 2003 to 2006. The results are robust to different variations of the leverage effect estimation. In Figure 7 and Figures C.25 to C.34 I calculate the componentwise leverage effect based on implied and high-frequency volatilities. As expected the high-frequency estimator underestimates the leverage effect, but the pattern is exactly the same and hence robust to the estimation methodology.

For the implied volatility based leverage estimator I calculate the correlation between the daily accumulated continuous log price increments and the increments of daily implied volatilities. Figure 8 and Figures C.45 to C.54 depict the componentwise leverage effect if I replace the accumulated continuous increments by daily CRSP returns. The results are essentially the same. In Figure 9 and Figures C.55 to C.64 I replace the four statistical factors based on continuous loadings applied to daily CRSP excess returns by the 4 Fama-French-Carhart factors. The pattern in the systematic and idiosyncratic leverage effect are not affected. This seems surprising at first as factors based on the continuous loadings are different from the Fama-French-Carhart factors except for the market factor. I can show that the leverage effect results are mainly driven by the market factor. If I replaced the four factors in our analysis by merely the market factor, the observed pattern would be very similar.

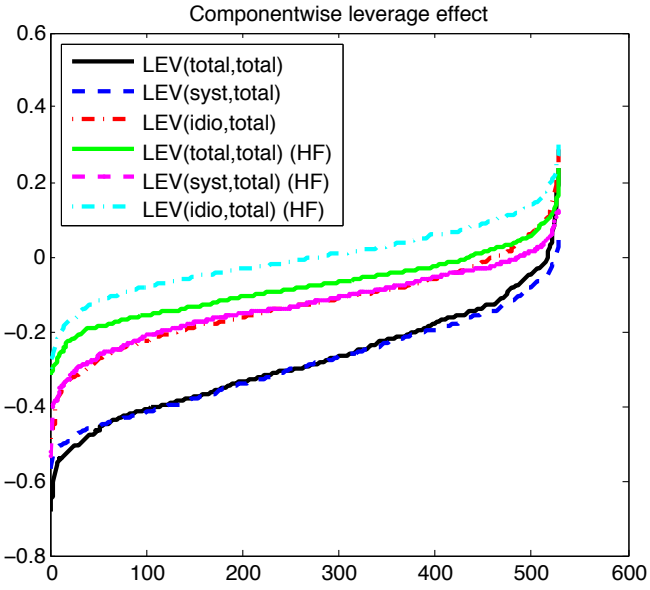

Figure 7: Componentwise leverage effect in 2012 based on implied and high-frequency volatilities. 4 continuous asset factors.

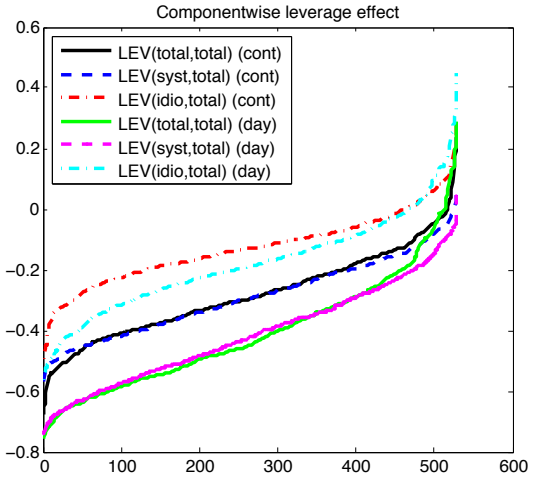

Figure 8: Componentwise leverage effect in 2012 with daily continuous log price increments $LEV(cont)$ and daily returns $LEV(day)$ and 4 asset factors.

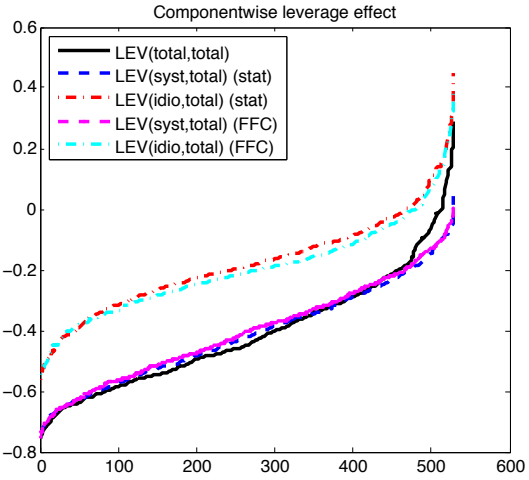

Figure 9: Componentwise leverage effect in 2012 with 4 continuous daily factors $LEV(stat)$ or 4 Fama-French-Carhart factors $LEV(FFC)$.

7 Conclusion

This paper studies factor models in the new setting of a large cross section and many high-frequency observations under a fixed time horizon. In an extensive empirical study I can show that the continuous factor structure is highly persistent. For the time period 2007 to 2012 I estimate four continuous factors which can be approximated very well by a market, oil, finance and electricity factor. The value, size and momentum factors play no significant role in explaining these factors. From 2003 to 2006 one continuous systematic factor disappears. There seems to exist only one persistent jump factor, namely a market jump factor. Using implied volatilities from option price data, I analyze the systematic factor structure of the volatilities. There seems to be only one persistent market volatility factor, while during the financial crisis an additional temporary banking volatility factor appears. Based on the estimated factors, I can decompose the leverage effect, i.e. the correlation of the asset return with its volatility, into a systematic and an idiosyncratic component. The negative leverage effect is mainly driven by the systematic component.

Arbitrage pricing theory links risk premiums to systematic risk. As a next step I want to analyze the ability of the high-frequency factors to price the cross-section of returns.

References

Aït-Sahalia, P. A. Mykland, Y., and L. Zhang, 2005a, How often to sample a continuous-time process in the presence of market microstructure noise, *Review of Financial Studies* 18, 351–416.

Aït-Sahalia, P. A. Mykland, Y., and L. Zhang, 2005b, A tale of two time scales: Determining integrated volatility with noisy high-frequency data, *Journal of the American Statistical Association* 100, 1394–1411.

Aït-Sahalia, Y., J. Fan, and Y. Li, 2013, The leverage effect puzzle: Disentangling sources of bias at high frequency, *Journal of Financial Economics* .

Aït-Sahalia, Y., J. Fan, and D. Xiu, 2010, High-frequency estimates with noisy and asynchronous financial data, *Journal of the American Statistical Association* 105, 1504–1516.

Aït-Sahalia, Y., and J. Jacod, 2014, *High-Frequency Financial Econometrics* (New Jersey: Princeton University Press).

Aït-Sahalia, Y., and D. Xiu, 2015a, Principal component analysis of high frequency data, *Working paper* .

Aït-Sahalia, Y., and D. Xiu, 2015b, Principal component estimation of a large covariance matrix with high-frequency data, *Working paper* .

Andersen, T. G., L. Benzoni, and J. Lund, 2002, An empirical investigation of continuous-time equity return models, *Journal of Finance* 57, 1239–1284.

Andersen, T.G., T. Bollerslev, F. X. Diebold, and P. Labys, 2001, The distribution of realized exchange rate volatility, *Journal of the American Statistical Association* 42, 42–55.

Bai, J., 2003, Inferential theory for factor models of large dimensions, *Econometrica* 71, 135–171.

Bai, J., and S. Ng, 2002, Determining the number of factors in approximate factor models, *Econometrica* 70, 191–221.

Bai, J., and S. Ng, 2006, Evaluating latent and observed factors in macroeconomics and finance, *Journal of Econometrics* 507–537.

Bali, T. T., R. F. Engle, and Y. Tang, 2014, Dynamic conditional beta is alive and well in the cross-section of daily stock returns., *Working paper* .

Bandi, F., and R. Reno, 2012, Time-varying leverage effects, *Journal of Econometrics* 12, 94–113.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard, 2008, Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise, *Econometrica* 76, 1481–1536.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard, 2011, Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading, *Journal of Econometrics* 162, 149–169.

Barndorff-Nielsen, O.E., and N. Shephard, 2002, Econometric analysis of realized volatility and its use in estimating stochastic volatility models, *Journal of the Royal Statistical Society* 253–280.

Bekaert, G., and G. Wu, 2000, Asymmetric volatility and risk in equity markets, *Review of Financial Studies* 13, 1–42.

Bibinger, M., and L. Winkelmann, 2014, Econometrics of co-jumps in high-frequency data with noise, *Journal of Econometrics* 184, 361–378.

Black, F., 1976, Studies of stock price volatility changes, *Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economic Statistics* 6, 177–181.

Bollerslev, T., S. Z. Li, and V. Todorov, 2015, Roughing up beta: Continuous vs. discontinuous betas, and the cross section of expected stock returns, *Working paper* .

Bollerslev, T., and V. Todorov, 2010, Jumps and betas: A new theoretical framework for disentangling and estimating systematic risks, *Journal of Econometrics* 157, 220–235.

Bollerslev, T., and V. Todorov, 2011, Tails, fears and risk premia, *Journal of Finance* 66, 2165–2211.

Campbell, J. Y., and L. Hentschel, 1992, No news is good news: An asymmetric model of changing volatility in stock returns, *Journal of Financial Economics* 31, 281–318.

Carhart, M. M., 1997, On persistence in mutual fund performance, *Journal of Finance* 1, 57–82.

Chamberlain, G., 1988, Asset pricing in multiperiod securities markets, *Econometrica* 56, 1283–1300.

Chamberlain, G., and M. Rothschild, 1983, Arbitrage, factor structure, and mean-variance analysis on large asset markets, *Econometrica* 51, 1281–1304.

Christensen, K., R. C. A. Oomen, and M. Podolskij, 2014, Fact or friction: Jumps at ultra high frequency, *Journal of Financial Economics* 114, 576–599.

Christie, A.A., 1982, The stochastic behavior of common stock variances: Value, leverage and interest rate effects, *Journal of Financial Economics* 10, 407–432.

Connor, G., and R. Korajczyk, 1988, Risk and return in an equilibrium APT: Application to a new test methodology, *Journal of Financial Economics* 21, 255–289.

Connor, G., and R. Korajczyk, 1993, A test for the number of factors in an approximate factor model,, *Journal of Finance* 58, 1263–1291.

Duffie, D., J. Pan, and K. J. Singleton, 2000, Transform analysis and asset pricing for affine jump-diffusions, *Econometrica* 68, 1343–1376.

Engle, R. F., and V. K. Ng, 1993, Measuring and testing the impact of news on volatility, *Journal of Finance* 48, 1749–1778.

Eraker, B., 2004, Do stock prices and volatility jump? Reconciling evidence from spot and option prices,, *Journal of Finance* 59, 1367–1404.

Eraker, B., M. Johannes, and N. Polson, 2003, The impact of jumps in volatility and returns, *Journal of Finance* 58.

Fama, E. F., and K. R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.

Fan, J., A. Furger, and D. Xiu, 2014, Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high frequency data, *Working paper* .

Fan, L., Y. Liao, and M. Mincheva, 2013, Large covariance estimation by thresholding principal orthogonal complements, *Journal of the Royal Statistical Society* 75, 603–680.

Figleskwi, S., and X. Wang, 2000, Is the "leverage effect" a leverage effect, *Working paper* .

Forni, M., M. Hallin, M. Lippi, and L. Reichlin, 2000, The generalized dynamic-factor model: Identification and estimation, *REVIEW* 82, 540–554.

French, R., K., G. W. Schwert, and R. F. Stambaugh, 1987, Expected stock returns and volatility, *Journal of Financial Economics* 19, 3–29.

Gabaix, X., 2012, Variable rare disasters: An exactly solved framework for ten puzzles in macrofinance, *Quarterly Journal of Economics* 645–700.

Hallin, M., and R. Liska, 2007, The generalized dynamic factor model: Determining the number of factors, *Journal of the American Statistical Association* 102, 603–617.

Hansen, P., and A. Lunde, 2006, Realized variance and market microstructure noise, *Journal of Business and Economic Statistics* 24, 127–161.

Jacod, J., Y. Li, P.A. Mykland, M. Podolskij, and M. Vetter, 2009, Microstructure noise in the continuous case: The pre-averaging approach, *Stochastic Processes and their Applications* 119, 2249–2276.

Johannes, M., 2004, The statistical and economic role of jumps in continuous-time interest rate models, *Journal of Finance* 59, 227–260.

Kalnina, I., and D. Xiu, 2014, Nonparametric estimation of the leverage effect using information from derivatives markets, *Working paper* .

Lee, S. S., and P. A. Mykland, 2008, Jumps in financial markets: A new nonparametric test and jump dynamics, *Review of Financial Studies* 21, 2535–2563.

Lettau, M., and S. Ludvigson, 2001, Resurrecting the (c)capm: a cross-sectional test when risk premia are time-varying, *Journal of Political Economy* 1238–1287.

Lintner, J., 1965, The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets, *Review of Economic and Statistics* 1, 13–37.

Lunde, A., N. Shephard, and K. Sheppard, 2014, Econometric analysis of vast covariance matrices using composite realized kernels, *Working paper* .

Mancini, C, 2009, Non parametric threshold estimation for models with stochastic diffusion coefficient and jumps, *Scandinavian Journal of Statistics* 42–52.

Medvedev, A., and O. Scaillet, 2007, Approximation and calibration of short-term implied volatilities under jump-diffusion stochastic volatility, *Review of Financial*

Studies 20, 427–459.

Nelson, D. B., 1991, Conditional heteroskedasticity in asset returns: A new approach, *Econometrica* 59, 347–370.

Pan, J., 2002, The jump risk premium implicit in options: Evidence from an integrated time-series study, *Journal of Financial Economics* 3–50.

Pelger, M., 2015, Large-dimensional factor modeling based on high-frequency observations, *Working paper* .

Podolskij, M., and M. Vetter, 2009, Bipower-type estimation in a noisy diffusion setting, *Stochastic Processes and their Applications* 11, 2803–2831.

Ross, S. A., 1976, The arbitrage theory of capital asset pricing, *Journal of Economic Theory* 13, 341–360.

Sharpe, W., 1964, Capital asset prices: a theory of market equilibrium under conditions of risk, *Journal of Finance* 3, 425–442.

Stock, J. H., and M. W. Watson, 2002, Forecasting using principal components from a large number of predictors, *Journal of American Statistical Association* 97, 1167–1179.

Todorov, V., 2009, Estimation of continuous-time stochastic volatility models with jumps using high-frequency data, *Journal of Econometrics* 148, 131–148.

Wang, D. C., and P. A. Mykland, 2014, The estimation of leverage effect with high frequency data, *Journal of the American Statistical Association* 109, 197–215.

Yu, J., 2005, On leverage in a stochastic volatility model, *Journal of Econometrics* 127, 165–178.

Zhang, L., 2011, Estimating covariation: Epps effect, microstructure noise, *Journal of Econometrics* 160, 33–47.

A Empirical Appendix

A.1 Equity Data

I collect the price data from the TAQ database for the time period 2003 to 2012. I construct the log-prices for 5 minutes sampling, which gives me on average 250 days per year with 77 daily increments. Overnight returns are removed so that there is no concern of price changes due to dividend distributions or stock splits. I use the price of the trade at or immediately proceeding each 5-min mark. For each year I take the intersection of stocks traded each day with the stocks that have been in the S&P500 index at any point during 1993-2012. This gives me a cross-section N of around 500 to 600 firms for each year. I apply standard data cleaning procedures:

- Delete all entries with a time stamp outside 9:30am-4pm
- Delete entries with a transaction price equal to zero
- Retain entries originating from a single exchange
- Delete entries with corrected trades and abnormal sale condition.
- Aggregate data with identical time stamp using volume-weighted average prices

In each year I eliminate stocks from my data set if any of the following conditions is true:

- All first 10 5-min observations are missing in any of the day of this year
- There are in total more than 50 missing values before the first trade of each day for this year
- There are in total more than 500 missing values in the year

Table 1 in the main text shows the number of observations after the data cleaning.

Missing observations are replaced by interpolated values. For each day if the first n observations are missing, I interpolate the first values with the $(n+1)$th observation. Otherwise I take the previous observation. As my estimators are based on increments, the interpolated values will result in increments of zeros, which do not contribute to the quadratic covariation.

Daily returns and industry classifications (SIC codes) for the above stocks are from CRSP. I rely on Kenneth R. French's website for daily returns on the Fama-French-Carhart four-factor portfolios. I define three different industry factors as equally weighted portfolios of assets with the following SIC codes

1. Oil and gas: 1200; 1221; 1311; 1381; 1382; 1389; 2870; 2911; 3533; 4922; 4923

2. Banking and finance: 6020; 6021; 6029; 6035; 6036; 6099; 6111; 6141; 6159; 6162; 6189; 6199; 6282; 6311; 6331; 6351; 6798
3. Energy: 4911; 4931; 4991

A.2 Factor Analysis

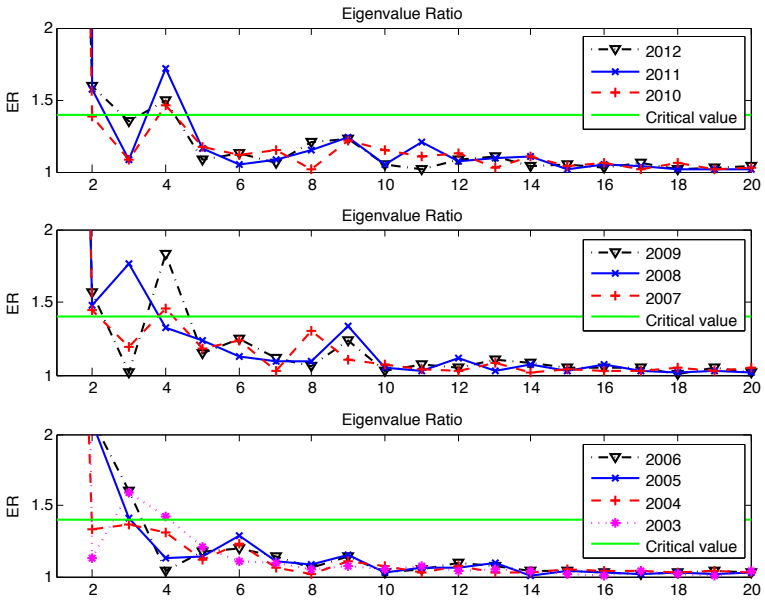

Figure A.10: Number of continuous factors using unperturbed eigenvalue ratios

A.3 Jump Factors

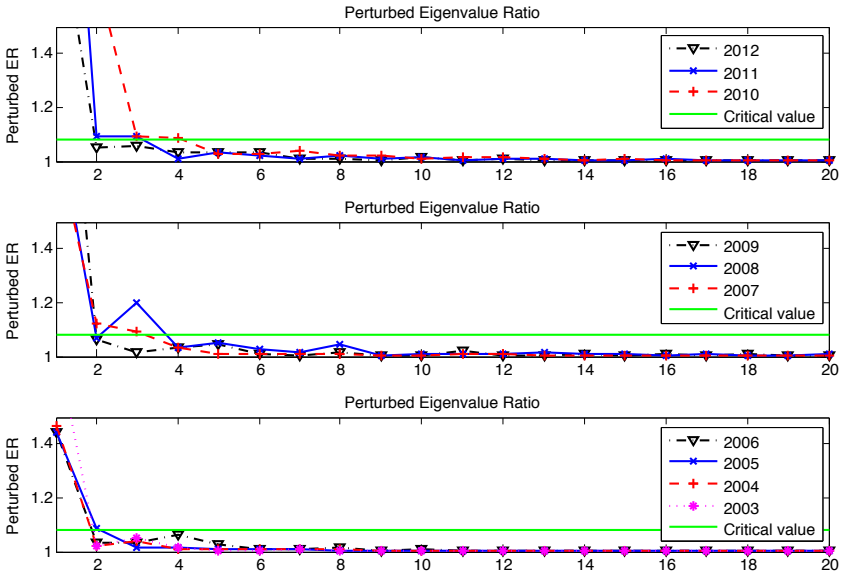

Figure A.11: Number of jump factors with truncation level $a = 3$.

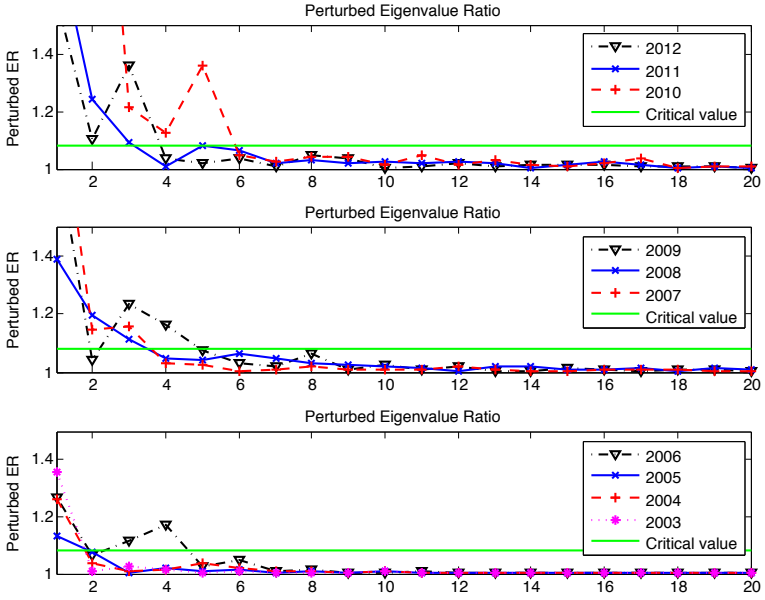

Figure A.12: Number of jump factors with truncation level $a = 4$.

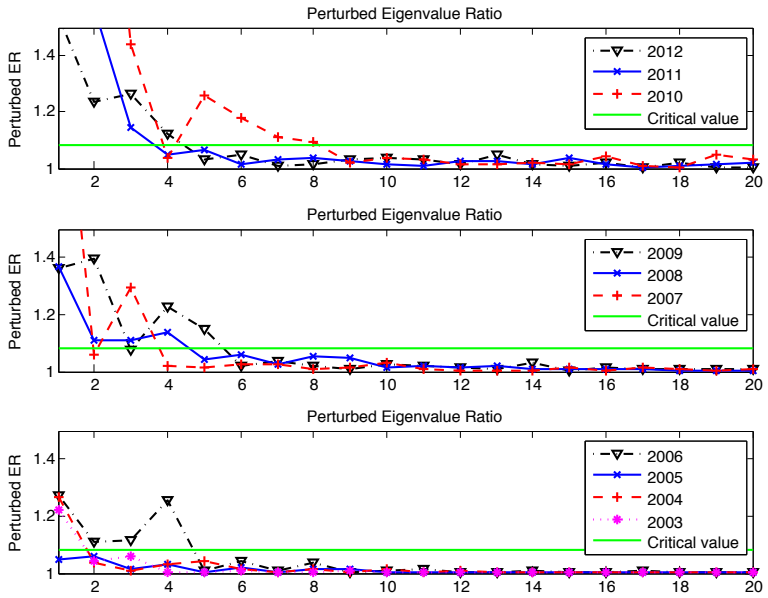

Figure A.13: Number of jump factors with truncation level $a = 4.5$.

1	2	3	4	5	6	7	8	9	10	11	12
Generalized correlations of monthly with yearly jump factors (a=3)											
0.98	**0.96**	**0.99**	**0.98**	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**	**1.00**	**1.00**
0.62	0.68	**0.87**	0.74	**0.88**	0.76	**0.95**	**0.95**	**0.96**	**0.87**	**0.95**	**0.80**
0.22	0.49	0.41	0.45	0.39	0.58	0.60	**0.93**	0.58	**0.81**	0.73	0.42
0.08	0.18	0.20	0.16	0.18	0.14	0.11	0.76	0.42	0.73	0.18	0.11
Generalized correlations of monthly with yearly jump loadings (a=3)											
0.85	**0.82**	**0.90**	**0.85**	**0.89**	**0.96**	**0.94**	**0.97**	**0.97**	**0.90**	**0.96**	**0.94**
0.29	0.32	0.42	0.38	0.48	0.43	0.66	0.77	0.56	0.52	0.64	0.44
0.11	0.22	0.16	0.26	0.17	0.30	0.22	0.71	0.30	0.42	0.33	0.19
0.03	0.08	0.09	0.06	0.07	0.05	0.05	0.40	0.19	0.32	0.08	0.04
Generalized correlations of monthly with yearly jump factors (a=4)											
0.73	0.75	**0.80**	0.77	**0.90**	**1.00**	**0.99**	**0.88**	**1.00**	**0.89**	**1.00**	**0.97**
0.35	0.20	0.63	0.44	**0.82**	**0.89**	**0.93**	0.73	**0.97**	0.71	**1.00**	**0.80**
0.06	0.11	0.56	0.21	0.37	0.21	0.76	0.42	0.50	0.47	**0.98**	0.45
0.02	0.01	0.28	0.03	0.03	0.08	0.32	0.11	0.14	0.08	0.76	0.30
Generalized correlations of monthly with yearly jump loadings (a=4)											
0.35	0.29	0.31	0.32	0.42	**0.95**	0.60	0.24	**0.96**	0.30	**0.95**	0.53
0.10	0.06	0.23	0.12	0.19	0.25	0.29	0.17	0.41	0.11	**0.89**	0.15
0.02	0.03	0.15	0.05	0.06	0.03	0.17	0.08	0.06	0.08	0.69	0.09
0.01	0.00	0.07	0.01	0.00	0.02	0.05	0.01	0.02	0.01	0.11	0.05
Generalized correlations of monthly with yearly jump factors (a=4.5)											
0.67	0.72	0.69	0.66	**0.91**	**1.00**	**0.97**	0.72	**0.99**	0.53	**1.00**	**0.95**
0.31	0.36	0.63	0.31	0.66	0.64	0.73	0.69	**0.90**	0.32	**1.00**	0.66
0.28	0.30	0.32	0.11	0.45	0.26	0.51	0.29	0.25	0.14	**0.85**	0.44
0.05	0.05	0.20	0.04	0.18	0.04	0.13	0.21	0.02	0.03	0.04	0.13
Generalized correlations of monthly with yearly jump loadings (a=4.5)											
0.22	0.19	0.20	0.18	0.31	**0.93**	0.40	0.11	0.31	0.09	**0.96**	0.32
0.09	0.11	0.15	0.08	0.12	0.10	0.11	0.09	0.12	0.05	**0.94**	0.09
0.08	0.08	0.06	0.03	0.08	0.04	0.06	0.04	0.04	0.02	0.77	0.07
0.01	0.01	0.04	0.01	0.03	0.01	0.01	0.03	0.00	0.01	0.01	0.02

Table A.1: Persistence of jump factors in 2011. Generalized correlation of monthly jump factors and loadings with yearly jump factors and loadings. The yearly number of factors is $K = 4$. Values larger than 0.8 are in bold.

A.4 Comparison with Daily Data and Total Factors

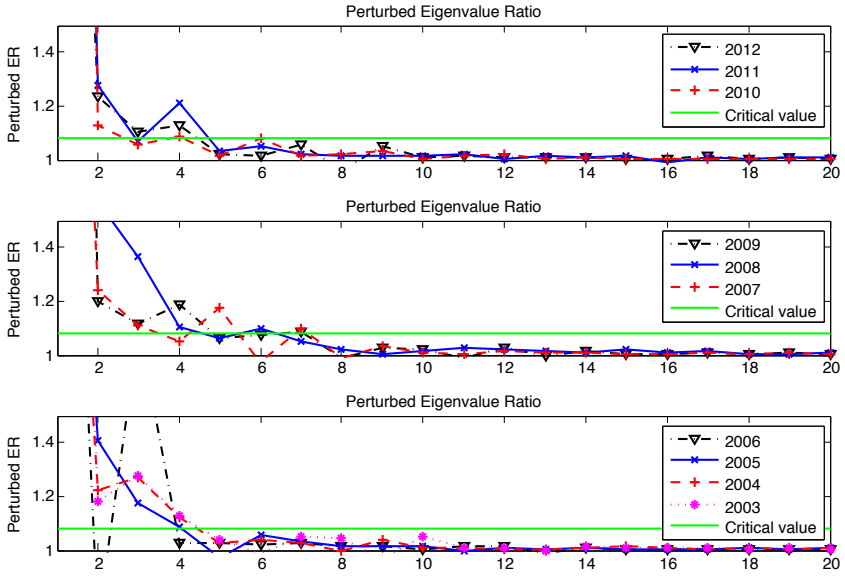

Figure A.14: Number of daily factors

2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Generalized correlations between continuous and total factors									
1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.9999	0.9999	1.0000	1.0000	0.9999	1.0000	0.9999	0.9997	1.0000	0.9999
0.9997	0.9996	0.9998	0.9998	0.9997	0.9999	0.9999	0.9993	0.9999	0.9999
0.9979	0.9982	0.9983	0.9748	0.9995	0.9989	0.9998	0.9855	0.9997	0.9997
Generalized correlations between continuous and total loadings									
0.9992	0.9982	0.9998	0.9997	0.9988	0.9994	0.9991	0.9993	0.9997	0.9992
0.9977	0.9982	0.9995	0.9994	0.9976	0.9988	0.9991	0.9720	0.9993	0.9991
0.9967	0.9974	0.9972	0.9982	0.9965	0.9988	0.9988	0.9720	0.9993	0.9991
0.9967	0.9974	0.9972	0.9708	0.9965	0.9950	0.9988	0.9698	0.9987	0.9988
Generalized correlations between continuous and daily factors									
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.99	0.97	0.99	0.99	0.99	1.00	0.99	0.98	0.99	0.99
0.98	0.94	0.94	0.97	0.95	0.98	0.98	0.98	0.98	0.97
0.55	**0.65**	0.83	**0.17**	**0.47**	**0.76**	0.98	0.96	0.93	0.93
Generalized correlations between continuous and daily loadings									
0.99	0.96	0.98	0.98	0.99	0.98	0.99	0.99	0.96	0.98
0.82	0.86	0.83	0.89	0.76	0.95	0.91	0.89	0.96	0.92
0.82	0.86	0.83	0.86	0.56	0.83	0.91	0.86	0.83	0.86
0.51	**0.48**	**0.73**	**0.13**	**0.41**	**0.61**	0.91	0.86	0.83	0.82

Table A.2: Generalized correlations between continuous factors and loadings based on continuous data and on total HF and daily data for $K = 4$ factors and for each year. I use the loadings estimated from the different data sets to construct continuous factors and estimate the distance between the different sets of continuous factors. Values smaller than 0.8 are in bold.

2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Generalized correlations between continuous and jump factors (a=3)									
1.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
0.99	**0.99**	**1.00**	**1.00**	**0.98**	**0.98**	**0.99**	**0.96**	**0.98**	**0.99**
0.98	**0.97**	**0.98**	0.64	**0.98**	0.77	**0.97**	0.56	**0.93**	**0.87**
0.82	0.58	**0.87**	0.29	0.18	0.38	**0.93**	0.39	0.67	0.38
Generalized correlations between continuous and jump loadings (a=3)									
0.94	**0.98**	**0.97**	**0.95**	**0.96**	**0.93**	**0.97**	**0.95**	**0.92**	**0.96**
0.94	**0.90**	**0.86**	0.72	0.50	0.32	**0.83**	0.32	0.77	0.79
0.84	**0.90**	**0.84**	0.34	0.30	0.31	**0.80**	0.32	0.63	0.44
0.68	0.31	**0.84**	0.14	0.30	0.31	**0.80**	0.26	0.48	0.15
Generalized correlations between continuous and jump factors (a=4)									
1.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
0.97	0.66	**0.85**	**0.81**	**0.92**	**0.95**	0.88	**0.89**	**0.89**	**0.88**
0.88	0.17	0.29	0.58	0.65	0.65	0.45	0.41	0.60	0.69
0.10	0.11	0.21	0.21	0.08	0.07	0.22	0.30	0.25	0.54
Generalized correlations between continuous and jump loadings (a=4)									
0.86	0.73	**0.82**	**0.83**	**0.88**	0.79	0.78	0.76	**0.87**	**0.81**
0.31	0.13	0.34	0.20	0.50	0.26	0.28	0.33	0.31	0.32
0.25	0.08	0.09	0.17	0.08	0.26	0.14	0.24	0.19	0.32
0.25	0.08	0.09	0.17	0.08	0.03	0.08	0.10	0.09	0.22
Generalized correlations between continuous and jump factors (a=4.5)									
1.00	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
0.84	0.66	0.79	**0.81**	**0.92**	**0.94**	**0.83**	**0.87**	0.79	**0.81**
0.55	0.23	0.37	0.52	0.49	0.60	0.47	0.59	0.52	0.66
0.04	0.10	0.27	0.16	0.03	0.03	0.13	0.25	0.18	0.43
Generalized correlations between continuous and jump loadings (a=4.5)									
0.73	0.60	0.75	0.79	**0.82**	0.58	0.54	0.54	0.75	0.72
0.27	0.14	0.25	0.19	0.38	0.45	0.25	0.43	0.21	0.26
0.08	0.14	0.12	0.14	0.08	0.32	0.15	0.24	0.10	0.22
0.08	0.04	0.12	0.14	0.02	0.01	0.04	0.11	0.10	0.22

Table A.3: Generalized correlations between continuous factors and loadings based on continuous data and on jump data for $K = 4$ factors and for each year. I use the loadings estimated from the different data sets to construct continuous factors and estimate the distance between the different sets of continuous factors. Values larger than 0.8 are in bold.

A.5 Implied Volatility Data

I use daily prices for standard call and put options from OptionMetrics for the same firms and time periods as for the high-frequency data. OptionMetrics provides implied volatilities for 30 days at the money options using a linearly interpolated volatility surface. I average the implied call and put volatilities for each asset and each day. Then I apply the following data cleaning procedure in order to identify outliers. For each year I remove a stock if

- for days 1-15 any of the volatilities is greater than 200% of the average volatility of the first 31 days
- for the last 15 days any of the volatilities is greater than 200% of the average volatility of the last 31 days
- for all the other days any of the volatilities is greater than 200% of the average of a 31 days moving window centered at that day.

The observations after the data cleaning are reported in Table 12 in the main text.

B Theoretical Appendix

All the theoretical results are taken from Pelger (2015), where I also provide a more in-depth discussion of the underlying arguments.

B.1 Assumptions

All the stochastic processes considered in this paper are locally bounded special Itô semimartingales as defined in Definition 1 in Pelger (2015). These particular semimartingales are the most general stochastic processes for which we can develop an asymptotic theory for the estimator of the quadratic covariation. A d-dimensional locally bounded special Itô semimartingale Y can be represented as

$$Y_t = Y_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s + \int_0^t \int_E \delta(s,x)(\mu - \nu)(ds, dx)$$

where b_s is a locally bounded predictable drift term, σ_s is an adapted cádlág volatility process, W is a d-dimensional Brownian motion and $\int_0^t \int_E \delta(s,x)(\mu-\nu)(ds,dx)$ describes

a jump martingale. μ is a Poisson random measure on $\mathbb{R}_+ \times E$ with $(E, \mathbb{E})$ an auxiliary measurable space on the space $(\Omega, \mathfrak{F}, (\mathfrak{F}_t)_{t\geq 0}, \mathbb{P})$. The predictable compensator (or intensity measure) of μ is $\nu(ds, dx) = ds \times v(dx)$ for some given finite or sigma-finite measure on $(E, \mathbb{E})$. These dynamics are very general and completely non-parametric. They allow for correlation between the volatility and asset price processes. I only impose some week regularity conditions in Definition 1. The model includes many well-known continuous-time models as special cases: for example stochastic volatility models like the CIR or Heston model, the affine class of models in Duffie, Pan and Singleton (2000), Barndorff-Nielsen and Shephard's (2002) Ornstein-Uhlenbeck stochastic volatility model with jumps or Andersen, Benzoni, and Lund's (2002) stochastic volatility model with log-normal jumps generated by a non-homogenous Poisson process.

The key assumption for obtaining a consistent estimator for the loadings and factors is an approximate factor structure. It requires that the factors are systematic in the sense that they cannot be diversified away, while the idiosyncratic residuals are nonsystematic and can be diversified away. The approximate factor structure assumption uses the idea of appropriately bounded eigenvalues of the residual quadratic covariation matrix, which is analogous to Chamberlain and Rothschild (1983) and Chamberlain (1988). Let $\|A\| = (tr(A\top A))^{1/2}$ denote the norm of a matrix A and $\lambda_i(A)$ the i's singular value of the matrix A, i.e. the square-root of the i's eigenvalue of $A^\top A$. If A is a symmetric matrix then λ_i is simply the i's eigenvalue of A.

Assumption 1. Factor structure assumptions

1. **Underlying stochastic processes**
   F and e_i are Itô-semimartingales as defined in Definition 1

$$F(t) = F(0) + \int_0^t b_F(s)ds + \int_0^t \sigma_F(s)dW_s + \sum_{s\leq t} \Delta F(s)$$
$$e_i(t) = e(0) + \int_0^t b_{e_i}(s)ds + \int_0^t \sigma_{e_i}(s)dW_s + \sum_{s\leq t} \Delta e_i(s)$$

   In addition each e_i is a square integrable martingale.
2. **Factors and factor loadings**

The quadratic covariation matrix of the factors Σ_F is positive definite a.s.

$$\sum_{j=1}^{M} F_j F_j^\top \xrightarrow{p} [F,F]_T =: \Sigma_F$$

and

$$\left\| \frac{\Lambda^\top \Lambda}{N} - \Sigma_\Lambda \right\| \to 0.$$

where the matrix Σ_Λ is also positive definite. The loadings are bounded, i.e. $\|\Lambda_i\| < \infty$ for all $i = 1, ..., N$.

3. **Independence of F and e**
   The factor process F and the residual processes e are independent.
4. **Approximate factor structure**
   The largest eigenvalue of the residual quadratic covariation matrix is bounded in probability, i.e.

$$\lambda_1([e,e]) = O_p(1).$$

As the predictable quadratic covariation is absolutely continuous, I can define the instantaneous predictable quadratic covariation as

$$\frac{d\langle e_i, e_k \rangle_t}{dt} = \sigma_{e_{i,k}}(t) + \int \delta_{i,k}(z) v_t(z) =: G_{i,k}(t)$$

I assume that the largest eigenvalue of the matrix $G(t)$ is almost surely bounded for all t:

$$\lambda_1(G(t)) < C \qquad \text{a.s. for all } t \text{ for some constant } C.$$

5. **Identification condition** *All Eigenvalues of $\Sigma_\Lambda \Sigma_F$ are distinct a.s..*

Assumption 2. Weak dependence of error terms
The row sum of the quadratic covariation of the residuals is bounded in probability:

$$\sum_{i=1}^{N} \|[e_k, e_i]\| = O_p(1) \qquad \forall k = 1, ..., N$$

B.2 Consistency

As pointed out before, the factors F and loadings Λ are not separately identifiable. However, they can be estimated up to an invertible $K \times K$ matrix H. Hence, my estimator $\hat{\Lambda}$ will estimate ΛH and $\hat{F}$ will estimate $FH^{\top -1}$. Note, that the common component is well-identified and $\hat{F}\hat{\Lambda}^\top = \hat{F}H^{\top -1}H^\top\Lambda^\top$. For almost all purposes knowing ΛH or $FH^{\top -1}$ is as good as knowing Λ or F as what is usually of interest is the vector space spanned by the factors. In my general approximate factor models I require N and M to go to infinity. The rates of convergence will usually depend on the smaller of these two values denoted by $\delta = \min(N, M)$.[32]

Theorem 1. Consistency of estimators
Define the rate $\delta = \min(N, M)$. Then the following consistency results hold:

1. *Consistency of loadings estimator: Under Assumption 1 it follows that*

$$\hat{\Lambda}_i - H^\top \Lambda_i = O_p\left(\frac{1}{\sqrt{\delta}}\right).$$

2. *Consistency of factor estimator and common component: Under Assumptions 1 and 2 we have*

$$\hat{F}_T - H^{-1}F_T = O_p\left(\frac{1}{\sqrt{\delta}}\right), \qquad \hat{C}_{T,i} - C_{T,i} = O_p\left(\frac{1}{\sqrt{\delta}}\right).$$

3. *Consistency of quadratic variation: Under Assumptions 1 and 2 and for any stochas-*

[32]Note that F_j is the increment $\Delta_j F$ and goes to zero for $M \to \infty$ for almost all increments. It can be shown that in a specific sense I can also consistently estimate the factor increments, but the asymptotic statements will be formulated in terms of the stochastic process F evaluated at a discrete time point t_j. For example $F_T = \sum_{j=1}^{M} F_j$ denotes the factor process evaluated at time T. Similarly I can evaluate the process at any other discrete time point $T_m = m \cdot \Delta_M$ as long as $m \cdot \Delta_M$ does not go to zero. Essentially m has to be proportional to M. For example, I could chose T_m equal to $\frac{1}{2}T$ or $\frac{1}{4}T$. The terminal time T can always be replaced by the time T_m in all the theorems. The same holds for the common component.

tic process $Y(t)$ satisfying Definition 1 it holds for $\frac{\sqrt{M}}{N} \to 0$ and $\delta \to \infty$ that

$$\sum_{j=1}^{M} \hat{F}_j \hat{F}_j^\top = H^{-1}[F,F]_T H^{-1^\top} + o_p(1) \;, \qquad \sum_{j=1}^{M} \hat{F}_j Y_j = H^{-1}[F,Y]_T + o_p(1)$$
$$\sum_{j=1}^{M} \hat{e}_{j,i}\hat{e}_{j,k} = [e_i,e_k]_T + o_p(1) \;, \qquad \sum_{j=1}^{M} \hat{e}_{j,i} Y_j = [e_i,Y]_T + o_p(1)$$
$$\sum_{j=1}^{M} \hat{C}_{j,i}\hat{C}_{j,k} = [C_i,C_k]_T + o_p(1) \;, \qquad \sum_{j=1}^{M} \hat{C}_{j,i} Y_j = [C_i,Y]_T + o_p(1).$$

for $i,k = 1,...,N$.

B.3 Separating Continuous and Jump Factors

Assumption 3. Truncation identification
F and e_i have only finite activity jumps and factor jumps are not "hidden" by idiosyncratic jumps:

$$\mathbb{P}\left(\Delta X_i(t) = 0 \text{ if } \Delta(\Lambda_i^\top F(t)) \neq 0 \text{ and } \Delta e_i(t) \neq 0\right) = 0.$$

The quadratic covariation matrix of the continuous factors $[F^C, F^C]$ and of the jump factors $[F^D, F^D]$ are each positive definite a.s. and the matrices $\frac{\Lambda^{C\top}\Lambda^C}{N}$ and $\frac{\Lambda^{D\top}\Lambda^D}{N}$ each converge in probability to positive definite matrices.

Assumption 3 has three important parts. First, I require the processes to have only finite jump activity. Second, I assume that a jump in the factors or the idiosyncratic part implies a jump in the process X_i. This second part is always satisfied as soon as the Lévy measure of F_i and e_i have a density, which holds in most popular models in the literature. The third statement is a non-redundancy condition and requires each systematic jump factor to jump at least once in the data. This is a straightforward and necessary condition to identify any jump factor. Hence, the main restriction in Assumption 3 is the finite jump activity. For example compound poisson processes with stochastic intensity rate fall into this category.

Theorem 2. Separating continuous and jump factors:
Assume Assumptions 1 and 3 hold. The estimators $\hat{\Lambda}^C$, $\hat{\Lambda}^D$, $\hat{F}^C$ and $\hat{F}^D$ are defined

analogously to $\hat{\Lambda}$ and $\hat{F}$, but using $\hat{X}^C$ and $\hat{X}^D$ instead of X.[33]

1. *The continuous and jump loadings can be estimated consistently:*

$$\hat{\Lambda}_i^C = {H^C}^\top \Lambda_i^C + o_p(1) \qquad , \qquad \hat{\Lambda}_i^D = {H^D}^\top \Lambda_i^D + o_p(1).$$

2. *Assume that additionally Assumption 2 holds. The continuous and jump factors can only be estimated up to a finite variation bias term*

$$\hat{F}_T^C = {H^C}^{-1} F_T^C + o_p(1) + \textit{finite variation term}$$
$$\hat{F}_T^D = {H^D}^{-1} F_T^D + o_p(1) + \textit{finite variation term.}$$

3. *Under the additional Assumption 2 I can estimate consistently the covariation of the continuous and jump factors with other processes. Let $Y(t)$ be an Itô-semimartingale satisfying Definition 1. Then we have for $\frac{\sqrt{M}}{N} \to 0$ and $\delta \to \infty$:*

$$\sum_{j=1}^{M} \hat{F}_j^C Y_j = {H^C}^{-1} [F^C, Y]_T + o_p(1) \qquad , \qquad \sum_{j=1}^{M} \hat{F}_j^D Y_j = {H^D}^{-1} [F^D, Y]_T + o_p(1).$$

The theorem states that I can estimate the factors only up to a finite variation term, i.e. I can only estimate the martingale part of the process correctly. The intuition behind this problem is simple. The truncation estimator can correctly separate the jumps from the continuous martingale part. However, all the drift terms will be assigned to the continuous component. If a jump factor also has a drift term, this will now appear in the continuous part and as this drift term affects infinitely many cross-sectional X_i, it cannot be diversified away.

B.4 Estimating the Number of Factors

Theorem 3. Estimator for number of factors
Assume Assumption 1 holds and $O\left(\frac{N}{M}\right) \leq O(1)$. Then for any $\gamma > 0$

$$\hat{K}(\gamma) \xrightarrow{p} K.$$

[33] Define $H^C = \frac{1}{N}\left({F^C}^\top F^C\right)\left({\Lambda^C}^\top \hat{\Lambda}^C\right) {V_{MN}^C}^{-1}$ and $H^D = \frac{1}{N}\left({F^D}^\top F^D\right)\left({\Lambda^D}^\top \hat{\Lambda}^D\right) {V_{MN}^D}^{-1}$.

Assume in addition that Assumption 3 holds. Then for any $\gamma > 0$

$$\hat{K}^C(\gamma) \xrightarrow{p} K^C \qquad \hat{K}^D(\gamma) \xrightarrow{p} K^D$$

where K^C *is the number of continuous factors and* K^D *is the number of jump factors.*

B.5 Microstructure noise

Theorem 4. Upper bound on impact of noise
Assume we observe the true asset price with noise:

$$Y_i(t_j) = X_i(t_j) + \tilde{\epsilon}_{j,i}$$

where the noise $\tilde{\epsilon}_{j,i}$ *is i.i.d.* $(0, \sigma_\epsilon^2)$ *and independent of* X *and has finite fourth moments. Furthermore assume that Assumption 1 holds and that* $\frac{N}{M} \to c < 1$*. Denote increments of the noise by* $\epsilon_{j,i} = \tilde{\epsilon}_{j+1,i} - \tilde{\epsilon}_{j,i}$*. Then the variance of the microstructure noise is bounded by*

$$\sigma_\epsilon^2 \leq \frac{c}{2(1-\sqrt{c})^2} \min_{s \in [K+1, N-K]} \left(\lambda_s \left(\frac{Y^\top Y}{N} \right) \frac{1}{1 + cos\left(\frac{s+r+1}{N}\pi\right)} \right) + o_p(1)$$

where $\lambda_s \left(\frac{Y^\top Y}{N} \right)$ *denotes the sth largest eigenvalue of a symmetric matrix* $\frac{Y^\top Y}{N}$*.*

Remark 1. *For* $s = \frac{1}{2}N - K - 1$ *the inequality simplifies to*

$$\sigma_\epsilon^2 \leq \frac{c}{2(1-\sqrt{c})^2} \cdot \lambda_{1/2N-K-1} \left(\frac{Y^\top Y}{N} \right) + o_p(1).$$

Hence the microstructure noise variance can be bounded by approximately the median eigenvalue of the observed quadratic covariation matrix multiplied by a constant that depends only on the ratio of M *and* N*.*

B.6 Identifying the Factors

Theorem 5. A central limit theorem for the generalized continuous correlation
Assume Assumptions 1 to 3 hold. The process G *is either (i) a well-diversified portfolio of* X*, i.e. it can be written as* $G(t) = \frac{1}{N}\sum_{i=1}^N w_i X_i(t)$ *with* $\|w_i\|$ *bounded for all* i *or (ii)*

G is independent of the residuals $e(t)$. Furthermore assume that $\frac{\sqrt{M}}{N} \to 0$. Denote the threshold estimators for the continuous factors as $\hat{F}^C$ and for the continuous component of G as $\hat{G}^C$. The total generalized continuous correlation is

$$\bar{\rho}^C = trace\left([F^C, F^C]^{-1}[F^C, G^C][G^C, G^C]^{-1}[G^C, F^C]\right)$$

and its estimator is

$$\hat{\bar{\rho}}^C = trace\left((\hat{F^C}^\top \hat{F}^C)^{-1}(\hat{F^C}^\top \hat{G}^C)(\hat{G^C}^\top \hat{G}^C)^{-1}(\hat{G^C}^\top \hat{F}^C)\right).$$

Then

$$\frac{\sqrt{M}}{\sqrt{\hat{\Xi}^C}}\left(\hat{\bar{\rho}}^C - \bar{\rho}^C\right) \xrightarrow{D} N(0,1)$$

Theorem 12 in Pelger (2015) provides the explicit estimator for $\hat{\Xi}^C$.

B.7 Leverage Effect

Theorem 6. Estimation of the leverage effect with high-frequency data
Assume Y is a 1-dimensional Itô-semimartingale as in Definition 1 and in addition has only finite jump activity and its volatility process σ_Y^2 is continuous. I want to estimate the leverage effect defined as

$$LEV = \frac{[\sigma_Y^2, Y]_T^C}{\sqrt{[Y,Y]_T^C}\sqrt{[\sigma_Y^2, \sigma_Y^2]_T^C}}.$$

Denote the M increments of Y as $Y_j = Y_{t_{j+1}} - Y_{t_j}$. A consistent estimator of the leverage effect is

$$\widehat{LEV} = \frac{\widehat{[\sigma_Y^2, Y]_T^C}}{\sqrt{\widehat{[Y,Y]_T^C}}\sqrt{\widehat{[\sigma_Y^2, \sigma_Y^2]_T^C}}}.$$

with

$$\begin{aligned}
\hat{Y}_j^C &= Y_j \mathbb{1}_{\{|Y_j| \leq \alpha \Delta_M^{\varpi}\}} \\
\hat{\sigma}_l^2 &= \frac{1}{k\Delta_M} \sum_{j=1}^{k} \hat{Y}_{l+j}^{C\,2} \\
\widehat{[\sigma_Y^2, Y]}_T^C &= \frac{2}{k} \sum_{l=0}^{M-2k} \left(\hat{\sigma}_{(l+k)\Delta_M}^2 - \hat{\sigma}_{l\Delta_M}^2\right) \left(\hat{Y}_{(l+k)\Delta_M}^C - \hat{Y}_{l\Delta_M}^C\right) \\
\widehat{[Y, Y]}_T^C &= \sum_{j=1}^{M} \hat{Y}_j^{C2} \\
\widehat{[\sigma_Y^2, \sigma_Y^2]}_T &= \frac{3}{2k} \sum_{l=1}^{M-2k} \left(\hat{\sigma}_{(l+k)\Delta_M}^2 - \hat{\sigma}_{l\Delta_M}^2\right)^2 - \sum_{j=0}^{M-2k} \frac{6}{k^2} \left(1 - \frac{2}{k}\right) \hat{\sigma}_{l\Delta_M}^4.
\end{aligned}$$

Let $\Delta_M = \frac{T}{M}$, $k \sim \Delta_M^{-1/2}$, $\alpha > 0$ *and* $\bar{\omega} \in \left(0, \frac{1}{2}\right)$. *Then for any fixed* T *and as* $M \to \infty$

$$\widehat{LEV} \xrightarrow{p} LEV$$

Proof. See Theorem 8.14 in Aït-Sahalia and Jacod (2014) and Theorem 3 in Kalnina and Xiu (2014). □

Appendix C is available online.[34]

[34]The online appendix is available on my website: https://people.stanford.edu/mpelger/research