Understanding Systematic Risk - A High-Frequency Approach

MARKUS PELGER*

March 21, 2020

ABSTRACT

Based on a novel high-frequency data set for a large number of firms, I estimate the time-varying latent continuous and jump factors that explain individual stock returns. The factors are estimated using principal component analysis applied to a local volatility and jump covariance matrix. I find four stable continuous systematic factors, which can be well-approximated by a market, oil, finance, and electricity portfolio, while there is only one stable jump market factor. The exposure of stocks to these risk factors and their explained variation is time-varying. The four continuous factors carry an intraday risk premium that reverses overnight.

*Department of Management Science & Engineering, Stanford University, Stanford, CA 94305, Email: mpelger@stanford.edu. I thank Jason Zhu for excellent research assistance. I thank Yacine Aït-Sahalia, Torben Andersen, Robert M. Anderson, Svetlana Bryzgalova, Mikhail Chernov, John Cochrane, Frank Diebold, Darrell Duffie, Noureddine El Karoui, Steve Evans, Jianqing Fan, Kay Giesecke, Lisa Goldberg, Valentin Haddad, Michael Jansson, Martin Lettau, Ulrike Malmendier, Stefan Nagel (Editor), Olivier Scaillet, Ken Singleton, George Tauchen, Viktor Todorov, Neil Shephard, Dacheng Xiu, two anonymous referees, and audience participants at UC Berkeley, Stanford, University of Pennsylvania, University of Bonn and SoFiE, INFORMS, FERM, Econometric society, and NBER Time-Series meetings. This work was supported by the Center for Risk Management Research at UC Berkeley. I have read the *Journal of Finance* disclosure policy and have no conflict of interest to disclose.

One of the most popular methods for modeling systematic risk is estimation of factor models - finding the "right" systematic factors has become the central question of asset pricing. I contribute to our understanding about systematic risk by shedding light on the following three questions: (1) what are the factors that explain the systematic comomvent in individual stocks, (2) how does the systematic factor structure for stocks change over time, (3) what are the asset pricing implications of the systematic factors?

To do so, the approach that I follow has three key elements: First, rather than using a pre-specified (and potentially mis-specified) set of factors, I estimate the statistical factors, which can explain most of the common comovement in a large cross-section of stock returns. Second, I use high-frequency data which allows me to study the time-variation in the factor structure under minimal assumptions as I can analyze very short time horizons independently. Allowing for time-variation in the factor structure is crucial as individual stocks do not have constant risk exposure in contrast to some characteristic-sorted portfolios.[1] And third, I separate high frequency returns into continuous intraday returns and intraday and overnight jumps which allows me to decompose the systematic risk structure into its smooth and rough components, which have different asset pricing implications.

The statistical theory underlying my estimations, developed by Pelger (2019), combines high-frequency econometrics and large-dimensional factor analysis and is very general. Specifically, under the assumption of an approximate factor model, it estimates an unknown factor structure for general continuous-time processes based on 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" matrix and a "continuous risk covariance" matrix. The latent continuous and jump factors can be separately estimated by principal component analysis (PCA).

My empirical investigation is based on a novel high-frequency data set of five-minute returns of the stocks in the WRDS TAQ millisecond trades database for the period 2004 to 2016. My main findings are as follows. First, I find four high-frequency factors, with time-variation in the exposure of stocks to these risk factors and the amount of variation that they explain (e.g., one factor is only systematic during the financial crisis). Surprisingly, the portfolio weights used to construct the statistical factors are stable over time. Second, these four factors have an economically meaningful interpretation and can be closely approximated by market, oil, finance and electricity factors, whereas the size, value, and momentum factors of the Fama-French-Carhart model cannot span the statistical factors. Third, the factor structure for smooth continuous movements is different from that for rough jump movements.

[1] See Lettau and Pelger (2020).

In particular, there seems to be only one intraday jump market factor, while the continuous factors have the same composition as the high-frequency factors including the jumps. Fourth, PCA factors estimated at lower frequencies (e.g. weekly or monthly returns) are different from high-frequency PCA factors. The lower frequencies result in a loss in information and the resulting PCA factors have a less interpretable structure and explain less of the time-series variation in stock returns than the high-frequency PCA factors. Fifth, the high-frequency factors carry an intraday risk premium that reverses overnight. Decomposing returns into their intraday and overnight components, I document a strong reversal pattern in individual stock returns, which is captured by the high-frequency factors. Finally, the high-frequency factors explain the expected return structure of industry portfolios, while the Fama-French-Carhart factors better explain size- and value-sorted portfolios. This result suggests that time-varying factors that explain the comovement in individual stocks are indeed informative about cross-sectional pricing, but are not necessarily related to characteristics.

This paper addresses the central question in empirical and theoretical asset pricing of what constitutes systematic risk. There are essentially three common approaches to identify the factors that describe the systematic risk. Under the first approach, factor selection is based on theory and economic intuition. The capital asset pricing model (CAPM) of Sharpe (1964), in which the market is the only common factor, belongs to this category. Under the second approach, factors are based on firm characteristics. The three-factor model of Fama and French (1993) is the most famous example of this approach. Under the third approach - the one to which this paper belongs - factor selection is statistical. This approach is motivated by the arbitrage pricing theory (APT) of Ross (1976). Factor analysis can be used to analyze 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" is introduced by Chamberlain and Rothschild (1983), which allows for a nondiagonal covariance matrix of the idiosyncratic component. Connor and Korajczyk (1988, 1993) study the use of PCA in the case of an unknown covariance matrix, that has to be estimated.[2]

One distinctive feature of the factor studies above is that they employ a constant factor model that does not allow for time-variation. The objects of interest are sorted portfolios based on previously established knowledge about the empirical behavior of average returns. Lettau and Pelger (2020) show that characteristic-sorted portfolios are well described by a constant factor model. However, a shortcoming of the characteristic-sorted approach is that

[2]The general case of a static large-dimensional factor model is treated in Bai (2003) and Bai and Ng (2002). Fan, Liao, and Mincheva (2013) study an approximate factor structure with sparsity.

the results depend on the choice of conditioning variables to generate the portfolios (Nagel (2013)).

In this paper I work directly on a large cross-section of individual stocks. However, Lettau and Pelger (2020) show that a constant loading model is not appropriate to model individual stock returns over longer time horizons. The dominant approach is to model time-variation in risk exposure through time-varying characteristics. Kelly, Pruitt, and Su (2017) and Fan, Liao, and Wang (2016b) follow this approach.[3] In effect, these studies apply a version of PCA to characteristic-managed portfolios. Their results therefore depend on the choice of characteristics and the basis functions used to model the functional relationship between characteristics and loadings. I propose an alternative approach that accounts for time-variation and is completely general in that it does not depend on the choice of characteristics or basis functions. As a result, I identify the factors that explain most of the variation in individual stocks without using any (potentially incorrect) prior about the characteristics that drive the variation. Interestingly, I find that the factors that explain the variation in individual stocks are not related to popular characteristic-based factors such as the Fama-French-Carhart factors.[4]

I combine tools from high-frequency econometrics with a large-dimensional panel data set. The advantage of high-frequency observations is to estimate a time-varying factor model without any prior assumption about the time-variation. Furthermore, the significantly larger amount of time-series observations allows for a more precise estimation of the risk structure. To date most of the empirical literature that uses high-frequency econometrics to analyze a factor structure is limited to a pre-specified set of factors. For example, Bollerslev, Li, and Todorov (2016) and Alexeev, Dungey, and Yao (2017) estimate betas for a continuous and jump market factor, and Fan, Furger, and Xiu (2016a) estimate a large-dimensional covariance matrix with high-frequency data for a given factor structure. In this paper I go further by estimating the unknown continuous and jump factor structure in a large cross-section. My method employs a purely statistical criterion to derive factors and has the advantage of requiring no ex ante knowledge of the structure of returns.[5] In addition, the high-frequency data allow me to separately study smooth continuous factor risk and

[3] In a related approach, Fama and French (2020) extract factors by cross-sectional regressions on pre-specified characteristics and model time-varying loadings through time-varying characteristics.

[4] For a pre-specified set of factors, prior studies show that time-varying systematic risk factors capture the data better. Time-varying systematic risk factors contain the conditional version of the CAPM as a special case, which appears 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 (2017) also show that GARCH-based time-varying conditional betas help explain the cross-sectional variation in expected stock returns.

[5] Aït-Sahalia and Xiu (2019) apply nonparametric principal component analysis to a low-dimensional

rough intraday and overnight jump risk, which cannot be estimated from observing only daily returns. Empirical evidence suggests that for a given factor structure, systemic risk associated with discontinuous price movements is different from continuous systematic risk.[6] I confirm and extend these results to a latent factor structure.

I find that risk-return patterns differ between trading and nontrading hours. While Branch and Ma (2012), Cliff, Cooper, and Gulen (2008), Berkman, Koch, Tuttle, and Zhang (2012), and Lou, Polk, and Skouras (2018) also find distinct return patterns during trading and nontrading hours, I link these patterns to the underlying factor structure. The risk premium for characteristic-based factors such as size and value is earned mainly overnight. My high-frequency factors are based only on intraday information and hence are expected to capture the intraday risk premium. In line with this prediction, I find that these factors have a positive intraday risk premium that reverses overnight. Thus, the daily risk premium of the high-frequency factors is lower in absolute value than its intraday and overnight components. Lou, Polk, and Skouras (2018) show that institutional investors tend to initiate trades at the close, while individual investors are more likely to initiate trades near the open. This suggests that the reversal pattern in the factor risk premia might be driven by the demand of different clienteles.

My paper contributes to an emerging literature that uses new econometric techniques in asset pricing for high-dimensional data. Lettau and Pelger (2020) generalize PCA by including a penalty on the pricing error in expected returns. Kozak, Nagel, and Santosh (2020) apply mean-variance optimization with an elastic net to PCA-based factors. The three-pass model in Giglio and Xiu (2017) corrects for missing factors by including PCA-based factors. These approaches use a constant loading model. Kelly, Pruitt, and Su (2017) and Fan, Liao, and Wang (2016b) apply a version of PCA to projected portfolio data. All of these papers argue that the stochastic discount factor based on some version of dominant principal components can well explain the cross-section of expected returns.

The rest of the paper is organized as follows. Section I introduces the factor model setup. Section II explains the estimation method, and Section III reports the empirical findings. Section IV concludes. The Internet Appendix contains additional empirical results.[7]

cross-section of high-frequency data. Aït-Sahalia and Xiu (2017) also estimate a large-dimensional factor model based on continuous high-frequency observations. Their studies focus on estimating the continuous covariance matrix, while my work tries to explain the factor structure itself.

[6]Empirical studies supporting this hypothesis include Bollerslev, Li, and Todorov (2016), Alexeev, Dungey, and Yao (2017), Pan (2002), Eraker, Johannes, and Polson (2003), Eraker (2004), Bollerslev and Todorov (2011) and Gabaix (2012).

[7]The Internet Appendix is available in the online version of the article on the Journal of Finance website.

I. Methodology

A. *Factor Model*

Assume that log asset prices can be modeled by a local approximate factor model. Most comovements in asset prices are due to a systematic factor component. In more detail, consider N assets with log prices denoted by $P_i(t)$. Assume that the N-dimensional log price process $P(t)$ can be explained by a local factor model, that is,

$$P_i(t) = \Lambda_{i,s}^\top F(t) + e_i(t), \qquad i = 1, ..., N,\ t \in [T_s, T_{s+1}], \text{ and } s = 1, ..., S, \tag{1}$$

where $\Lambda_{i,s}$ is a $K \times 1$ dimensional vector and $F(t)$ is a K-dimensional stochastic process. The loadings $\Lambda_{i,s}$ describe the exposure to the systematic factors F, while the residuals e_i are stochastic processes that describe the idiosyncratic component. The assumption of constant loadings Λ_s only needs to hold locally, that is, for the time interval $[T_s, T_{s+1}]$. In my empirical analysis I study local time windows of one week (five trading days), one month, three months, one year and 13 years. The loadings, factors and residual structure is allowed to differ completely across different local time windows.[8] This model is very general and includes the most popular factor models as a special case, for example, time-varying factor models in which the time-variation in loadings is due to characteristic variables that vary monthly as in Kelly, Pruitt, and Su (2017).

As my analysis is applied only locally, to simplify notation I drop the subscript s and formulate the model and results in terms of the locally constant model

$$P_i(t) = \Lambda_i^\top F(t) + e_i(t), \qquad i = 1, ..., N \text{ and } t \in [0, T]. \tag{2}$$

For the fixed (and potentially short) time interval $[0, T]$, I only observe the stochastic process P at discrete time observations $t_0 = 0, t_1 = \Delta_M, t_2 = 2\Delta_M, ..., t_M = M\Delta_M$, where the time increment is defined as $\Delta_M = t_{j+1} - t_j = \frac{T}{M}$. Hence, I observe the log prices

$$P(t_j) = \Lambda F(t_j) + e(t_j), \qquad j = 0, ..., M. \tag{3}$$

where $\Lambda = (\Lambda_1, ..., \Lambda_N)^\top$ and $P(t) = (P_1(t), ..., P_N(t))^\top$. In my setup the number of cross-sectional observations N and the number of high-frequency observations M is large, while

[8] Note that different factors for different time intervals can be captured by setting the corresponding time-varying loadings to zero.

the time horizon T is short 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.

The asset price dynamics are completely nonparametric and very general. The log prices are modeled as Itô-semimartingales, which is the most general class of stochastic processes for which the general results of high-frequency econometrics are available[9]:

$$P(t) = P(0) + \int_0^t \mu_s ds + \int_0^t \sigma_s dW_s + \int_{\mathbb{R}} x\nu(ds, dx). \tag{4}$$

The above process consists of a predictable drift term, a continuous martingale with N-dimensional Brownian motion W_t and volatility process σ_t and a jump martingale described by a compensated jump counting measure that accounts for discontinuous price movements. These particular semimartingales are standard in high-frequency econometrics (see for example Aït-Sahalia and Jacod (2014)) and allow for correlation between the volatility and asset price processes. The model includes many well-known continuous-time models as special cases, for example, stochastic volatility models such as the CIR or Heston model, the affine class of models in Duffie, Pan, and Singleton (2000), the Ornstein-Uhlenbeck stochastic volatility model with jumps of Barndorff-Nielsen and Shephard (2002) and the stochastic volatility model with log-normal jumps generated by a nonhomogeneous Poisson process as in Andersen, Benzoni, and Lund (2002).

The factors in an approximate factor model are systematic in the sense that they explain most of the comovement in the data. A large fraction of the correlation between stock returns is explained by exposure to a small number of factors. As I will discuss in Section I.B this correlation structure has direct implications for asset pricing. Under essentially the same model assumptions, the pricing kernel has to be spanned by the systematic factors and the differences in portfolio returns' risk premia should be explained by differences in factor exposure.

My approach requires only very weak assumptions, which are stated formally in Pelger (2019). First, i assume that the dependence between assets can be captured by an approximate factor structure similar to that in Chamberlain and Rothschild (1983). Idiosyncratic returns can be serially and weakly cross-sectionally correlated, which allows for a very general specification. The main identification criterion for systematic risk is that the quadratic covariation matrix of the idiosyncratic component has bounded eigenvalues, while the quadratic covariation matrix of the systematic factor component has unbounded eigen-

[9]Locally bounded special Itô semimartingales are specified in Pelger (2019). I assume that the factors and idiosyncratic component are also of this form.

values. This allows PCA to relate the eigenvectors of the exploding eigenvalues to the factor loadings. Second, to separate continuous systematic risk from jump risk, I assume only finite activity jumps, that is, that there are only finitely many jumps in the asset price process of each stock.[10] 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 these two parameters' paths to infinity. The large 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 to estimate the systematic factors.

I am also interested in estimating the continuous, jump, and overnight components of the factor return. I can separate factors into continuous factors which consists of a continuous martingale and a predictable finite variation term and jump factors which consist of a jump martingale and a predicable finite variation term but no continuous martingale. The continuous and jump components constitute the intraday return. The overnight component is the difference between the logarithm of the closing price and the opening price which I model following Bollerslev, Li, and Todorov (2016) as an overnight jump.[11] Without loss of generality, I can decompose the instantaneous returns according to

$$dP(t) = \Lambda^C dF^C(t) + \Lambda^D dF^D(t) + \Lambda^N dF^N(t) + de(t), \tag{5}$$

where Λ^C, Λ^D, and Λ^N denote the continuous, jump and overnight loadings, and $F^C(t)$, $F^D(t)$, and $F^N(t)$ are the corresponding factor components. The model allows the loadings on jumps and overnight movements to differ from the continuous loadings. The model also allows the factors themselves to differ. For example, if a certain risk factor only contributes to continuous comovements and not to jumps, I set the jump loadings for this factor to zero.

A special case is the CAPM model of Bollerslev, Li, and Todorov (2016), which allows for different continuous, jump, and overnight betas for the market factor. Under the assumption of this one-factor model, I would estimate one factor when separately considering continuous returns, intraday jumps, or overnight jumps. If the market betas were different for continuous, jump, and overnight returns, I would then estimate a three-factor model for the combined returns with a pure continuous, an intraday jump, and an overnight jump market

[10] Many of my results go through without this restriction; it is only needed for the separation of these two components.

[11] Overnight returns are adjusted for dividends and stock splits.

factor. If the market betas were the same, then the combined returns would be described by one market factor with continuous and jump movements.

B. Asset Pricing Implications

Under minimal no-arbitrage assumptions on the approximate factor model (Chamberlain (1988) and Reisman (1992)), there exists an economy-wide pricing kernel that is spanned by the risk factors. Specifically, for any well-diversified portfolio the drift term of the excess return, which corresponds to the risk premium, equals the exposure to risk times the price of risk. Note that the approximate no-arbitrage condition in Chamberlain (1988) and Reisman (1992) results in only an approximate pricing statement, that is, only the risk premia of well-diversified portfolios are explained by the systematic factor risk. The pricing statement can be extended to individual stocks with idiosyncratic risk under either additional preference restrictions as in Connor (1984) or "no good deal" bounds on the risk-reward ratio and takes the form[12]

$$E_t[dP(t) - r_t dt] = \mu_t dt - r_t dt = \Lambda E_t[dF(t) - r_t dt] = \Lambda \left(\mu_t^F dt - r_t dt\right), \tag{6}$$

where r_t denotes the instantaneous risk-free interest rate and μ_t^F the drift term of the factors.[13] The stochastic discount factor (SDF) is spanned by the systematic risk factors. The potentially different exposures to continuous, jump, and overnight risk is reflected by different weights on these components in the SDF (Bollerslev, Li, and Todorov (2016) and Duffie, Pan, and Singleton (2000)).

As shown in Back (1991), the risk premium can be decomposed into its continuous, jump and overnight components.[14] However, in general using only return data I cannot separately estimate the continuous risk premium, the jump risk premium and the overnight risk premium. If the span of the loadings Λ^C, Λ^D, and Λ^N differed, I could separate the different risk premia by a cross-sectional regression on expected excess returns. However, below I show empirically that Λ^D is spanned by Λ^C. This finding rules out estimating the

[12] My asset pricing tests are based on well-diversified portfolios and hence only require an approximate no-arbitrage condition.

[13] The factors can be represented by $dF_t = d\mu_t^F + dM_t^F$, where $dM_t = \sigma^F dW_t^F + \int_{\mathbb{R}} x\nu^F(dt, dx)$ is a local martingale consisting of a Brownian motion W_t^F and a compensated jump martingale with compensated jump measure ν^F (see Pelger (2019) for technical details). In this formulation the drift term minus the risk-free rate corresponds to the risk premium for continuous and jump risk. Note that all factors under consideration are traded portfolios of the underlying assets.

[14] Back (1991) proves the decomposition into a continuous and a jump component but the same arguments can be applied to obtain the overnight component as well.

jump risk premium with a two-stage Fama and MacBeth (1973) type regression as proposed by Bollerslev, Li, and Todorov (2016) and Alexeev, Dungey, and Yao (2017). I therefore focus instead on the intraday and overnight components of returns.

Separating returns into their intraday and overnight components, I examine how systematic risk affects the expected intraday and overnight excess returns. The daily returns consist of intraday and overnight price movements,

$$dP(t) = dP^{intra}(t) + dP^{night}(t) = \Lambda(dF^{intra}(t) + dF^{night}(t)) + de^{intra}(t) + de^{night}(t). \quad (7)$$

By the no-arbitrage condition the risk premium can also be separated into intraday and overnight components,

$$\mu_t^{intra}dt - r_t^{intra}dt = \Lambda\left(\mu_t^{F,intra}dt - r_t^{intra}dt\right), \quad (8)$$

$$\mu_t^{night}dt - r_t^{night}dt = \Lambda\left(\mu_t^{F,night}dt - r_t^{night}dt\right), \quad (9)$$

where $\mu_t^{F,intra}$ and $\mu_t^{F,intra}$ are the intraday and overnight drift terms of the factors and r_t^{intra} and r_t^{night} are the intraday and overnight risk-free interest rates respectively. The time-series average of dP^{intra} estimates the intraday expected return $\frac{1}{T}\int_0^T \mu_t^{intra}dt$, and the analogue applies for the overnight part. Since factors estimated from high-frequency data only incorporate intraday price information, I examine whether the average intraday return differs from the average overnight return.

The SDF implies a tangency portfolio based only on the factors that lead to the maximal conditional Sharpe ratio. Given the different set of factors in my analysis, the Sharpe ratio of the corresponding tangency portfolio serves as a measure of the factors' pricing performance. I also decompose the returns of the tangency portfolio into an intraday component and an overnight component and examine when the risk premium for different factors is earned. Since I cannot estimate the time-varying drift term without further model assumptions, I can only use factor means $\bar{\mu}^F - \bar{r} = \int_0^T(\mu_t^F - r_t)dt$ to construct the tangency portfolio weights $w^{SDF} = [F,F]^{-1}(\bar{\mu}^F - \bar{r})$, and study the overall return $dF(t)w^{SDF}$, intraday return $dF^{intra}(t)w^{SDF}$ and overnight return $dF^{night}(t)w^{SDF}$ of the tangency portfolio.[15] I also analyze the weights of the tangency portfolio based only on a specific component of the factors returns, for example, optimal overnight weights $w^{SDF,night} = [F^{night},F^{night}]^{-1}(\bar{\mu}^{F,night} - \bar{r}^{night})$.

[15]The quadratic covariation $[F,F]$ is defined as $\lim_{M\to\infty}\sum_{j=1}^{M}(F(t_j) - F(t_{j-1}))^2$ for $t_j - t_{j-1} = \frac{1}{M}$ and can be interpreted as a high-frequency covariance matrix of the changes in the stochastic process.

II. Estimation

A. *Factor Estimation*

I employ the estimation technique developed in Pelger (2019), which is essentially PCA applied to a volatility and jump covariance matrix. 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

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

The term R denotes the panel matrix of the high-frequency log-returns. In matrix notation, I have

$$\underset{(M\times N)}{R} = \underset{(M\times K)}{\Delta F}\,\underset{(K\times N)}{\Lambda^\top} + \underset{(M\times N)}{\Delta e}. \tag{10}$$

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

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

Intuitively, under some assumptions I can identify the jumps of the process $P_i(t)$ as the large movements that are greater 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 $\hat R^C_{j,i} = R_{j,i}\mathbb{1}_{\{|R_{j,i}|\le\alpha\Delta_M^{\bar\omega}\}}$ and $\hat R^D_{j,i} = R_{j,i}\mathbb{1}_{\{|R_{j,i}|>\alpha\Delta_M^{\bar\omega}\}}$.[16] The estimators $\hat\Lambda^C$, $\hat\Lambda^D$, $\Delta\hat F^C$, and $\Delta\hat F^D$ are defined analogously as $\hat\Lambda$ and $\Delta\hat F$, but using $\hat R^C$ and $\hat R^D$ instead of R. Overnight returns are modeled as separate jumps.

The quadratic covariation of the factors can be estimated by $\Delta\hat F^\top\Delta\hat F$ and the volatility

[16] I set $\bar\omega = 0.49$ (see Pelger (2019), Aït-Sahalia and Xiu (2017) and Bollerslev, Li, and Todorov (2013).) The threshold rate $\bar\omega$ is typically set between 0.47 and 0.49; the results are insensitive to this choice. Intuitively, I classify all increments as jumps that are beyond α standard deviations of a local estimator of stochastic volatility with $\alpha = 3, 4, 4.5$, or 5. For the jump threshold I use the *TOD* specification of Bollerslev, Li, and Todorov (2013), which takes into account the time-of-day pattern in the spot volatility estimation.

component of the factors by $\Delta\hat{F}^{C^\top}\Delta\hat{F}^{C}$. The estimated increments of the factors $\Delta\hat{F}, \Delta\hat{F}^{C}$, and $\Delta\hat{F}^{D}$ can be used to estimate the quadratic covariation with any other process, that is, I can use them in a high-frequency regression to consistently estimate the loadings for the different components of the different high-frequency factors.

The estimated loadings $\hat{\Lambda}, \hat{\Lambda}^{C}$, and $\hat{\Lambda}^{D}$ measure the risk exposure to the factors as well as serve as portfolio weights to construct the factors. For example, the portfolio weights for the continuous factors are $w^{C} = \frac{1}{\sqrt{N}}\hat{\Lambda}^{C}$. Based on the portfolio weights I do not only study the continuous returns of the continuous factors $R^{C}w^{C}$ but also their overall daily $R^{day}w^{C}$, intraday $R^{intra}w^{C}$, and overnight returns $R^{night}w^{C}$. In the analysis below I refer to a continuous factor if it is constructed with the continuous portfolio weights and specify the return component under consideration. The loadings that I estimate on a local time window coincide with the regression coefficients on the same local time window, that is, the rescaled eigenvectors $\hat{\Lambda}^{C}$ are equal to $R^{C^\top}\Delta\hat{F}^{C}(\Delta\hat{F}^{C^\top}\Delta\hat{F}^{C})^{-1}$. Since I also study the regression coefficients of different return components of the continuous factors, I label as loadings the regression coefficients for these factors and I specify both, the return component and time window under consideration.

B. Number of Factors

In Pelger (2019) I develop a new diagnostic criterion for the number of factors that can also distinguish between the number of continuous and jump factors.[17] Intuitively, the large eigenvalues are associated with the systematic factors and hence the problem of estimating the number of factors is roughly equivalent to determining which eigenvalues are considered large with respect to the rest of the spectrum. Under the approximate factor model assumptions, the first K "systematic" eigenvalues of $R^\top R$ 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, 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 around one, we are in the nonsystematic spectrum. As soon as we do not observe this clustering, but a rather large eigenvalue ratio of the perturbed eigenvalues, we are in the systematic spectrum.

[17] This estimator uses only the same weak assumptions that are needed for the consistency of my factor estimator. In simulations it outperforms the existing estimators while maintaining weaker assumptions.

The number of factors can be consistently estimated using the perturbed eigenvalue ratio statistic and hence I can replace the unknown number K by its estimator $\hat{K}$. I denote the ordered eigenvalues of $R^\top R$ by $\lambda_1 \geq ... \geq \lambda_N$ and define the perturbed eigenvalues $\hat{\lambda}_k = \lambda_k + g(N, M)$. Here, $g(N, M)$ is any slowly increasing sequence 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 strongly robust to different choices of the perturbation.[18] The perturbed eigenvalue ratio statistic equals

$$ER_k = \frac{\hat{\lambda}_k}{\hat{\lambda}_{k+1}} \qquad \text{for } k = 1, ..., N-1. \tag{11}$$

The estimator for the number of factors is defined as the first time (for descending k) for which 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 $\hat{R}^{C\top}\hat{R}^C$ and $\hat{R}^{D\top}\hat{R}^D$. The results in my empirical analysis are robust to a wide range of values for the threshold γ as illustrated by the eigenvalue ratio plots in Section III.B.

C. *Comparison between Factors*

One of the major challenges that arise when comparing two different sets of factors is that a factor model is identified only up to invertible linear transformations. Two sets of factors represent the same factor model if the factors span the same vector space. Thus, interpreting estimated factors by comparing them with economic factors, I need a measure that describes how close two vector spaces are to each other. The generalized correlation as proposed by Bai and Ng (2006) is a natural candidate.[19] Intuitively, I calculate the correlation between the latent and candidate factors after rotating them appropriately. Generalized correlations close to one indicate how many factors two sets have in common.

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

[18] 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 $g(N, M) = log(N) \cdot median\{\lambda_1, ..., \lambda_N\}$ and the Onatski (2010) eigenvalue difference estimator. The results are the same. See Pelger (2019).

[19] The generalized correlation is also called the canonical correlation.

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 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 $K-1$ and K_G-1 dimensional subspaces. This procedure continues until I have calculated the $min(K, K_G)$ generalized correlation. If $K = K_G = 1$, it is simply the correlation as measured by the quadratic covariation. [20] If two matrices span the same vector spaces, the generalized correlations are all equal to one. Otherwise they represent 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 $\{1, 1, 0\}$, then there exists a linear combination of the three factors in G that can replicate two of the three factors in F. Note that the number of candidate factors K_G can be different than the number of factors K that I want to explain.

To interpret latent factor models, Pelger and Xiong (2019) propose the use of proxy factors. The proxy factors use only the largest portfolio weights of the latent factors and set the smaller portfolio weights to zero. Pelger and Xiong (2019) show that the largest factor portfolio weights already contain most of the information signal to construct the latent factor even if the true factor itself is not sparse.

III. Empirical Results

A. Data

I combine data from the WRDS TAQ millisecond trades database, WRDS CRSP daily security database, and WRDS Compustat from January 2004 to December 2016. This is the earliest time period for which the TAQ Millisecond data are available.[21] I calculate high-frequency, daily and overnight returns for all assets included in the S&P 500 index.[22]

[20] Mathematically, the generalized correlations are the square root of the $min(K, K_G)$ largest eigenvalues of the matrix $[G,F]^{-1}[F,F][F,G][G,G]^{-1}$. Similarly, the distance between two loading matrices Λ and $\tilde{\Lambda}$ with dimensions $N \times K$ and $N \times \tilde{K}$ is the square root of the $\min(K, \tilde{K})$ largest eigenvalues of $(\Lambda^\top\Lambda)^{-1}\Lambda^\top\tilde{\Lambda}(\tilde{\Lambda}^\top\tilde{\Lambda})^{-1}\tilde{\Lambda}^\top\Lambda$.

[21] In a previous version of this paper I used data from the WRDS TAQ second trades database. The second trades database goes back further, but includes a smaller number of trades for fewer stocks.

[22] I include only those stocks that have been in the S&P 500 index at some point between 1993 to 2012 to ensure that my results are not driven by small, illiquid stocks. In addition, large stocks typically have a longer time series. When creating a balanced panel based on either all available stocks or the stocks in the S&P 500, I obtain roughly the same data set after the data cleaning procedure.

To strike a balance between using as much data as possible and minimizing the effect of microstructure noise and asynchronous returns, I choose to use five-minute prices. [23] More details about the data selection and cleaning procedures are in the Appendix. Each trading day contains 79 price observations. For a significant number of stocks I do not observe trading at the opening at 9:30am but only some minutes later. Therefore, I start the intraday sample at 9:35am, which results in 77 log returns for each asset and day.[24] For each of the 13 years I have on average 250 trading days with a cross-section between 555 to 667 firms. The intersection of all firms over the full time horizon contains 332 firms that form my balanced panel.

Daily returns are downloaded from the WRDS CRSP daily security database and adjusted for dividends and stock splits. Overnight log returns are calculated as the difference between intraday log returns and daily adjusted log returns. I extend my data set to all stocks available on any day from January 2004 to December 2016 in the WRDS TAQ millisecond trades database and I calculate their high-frequency, daily, and overnight returns as well as the size, book-to-market, and momentum characteristics. Following the standard procedure of Fama and French (1992) and using the breakpoints from Kenneth French's website, I create six portfolios formed on size, book-to-market, and a value-weighted market, size, value, and momentum factor for a high-frequency version of the Fama-French-Carhart model. Using the daily interest rates from Kenneth French's website, I calculate the high-frequency and overnight interest rate under the assumption that the interest rate is constant over the day.[25]

When identifying jumps, I face the trade-off between finding all discontinuous movements and misclassifying high-volatility regimes as jumps. The threshold should therefore account for changes in volatilities and intraday volatility patterns. I use the TOD estimator of Bollerslev, Li, and Todorov (2013) to separate the continuous from the jump movements. Hence, the threshold is set to $a \cdot 77^{-0.49}\hat{\sigma}_{j,i}$, 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 bi-power variation estimator for that day. Intuitively, I classify as jumps all increments that

[23] The five-minute sampling frequency is commonly advocated in the literature on realized volatility estimation, see, for example, Aït-Sahalia and Jacod (2014). In Section IA.B of the Internet Appendix I show that my results are robust to the sampling frequency. As lower sampling frequencies are less affected by microstructure noise, this suggests that microstructure noise does not affect my findings. My returns are also not affected by bid-ask bounces, as I use the volume-weighted transaction price in the last second of each five-minute interval.

[24] The main results are not affected by this choice. Section IX of the Internet Appendix I rerun the main analysis with 78 intraday log returns and find that the results are essentially identical to those using the 77 intraday log returns.

[25] As interest rates are significantly smaller than stock returns over the time horizon under consideration, the effect of this assumption is negligible.

are beyond a standard deviations of a local estimator of the stochastic volatility. For my analysis I use $a = 3$, $a = 4$, $a = 4.5$, and $a = 5$.

I apply 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 to obtain estimators for the loadings and factors. All of my results are virtually identical across the covariation and correlation approaches, 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.

Table I reports the fraction of increments identified as jumps for different thresholds for the balanced and unbalanced data, where I use the full cross-section available for each year. Depending on the year, for $a = 3$ more than 99% of observations are classified as continuous, while less than 1% are jumps. In 2012, 99.4% of the movements are continuous and explain around 87% to 88% of the total quadratic variation, while the 0.6% of the movements that are jumps explain the remaining 12% to 13% of the total quadratic covariation. Increasing the threshold leads to fewer movements being classified as jumps.[26] All of the results for the continuous factors are highly robust to this choice. However, results for the jump factors are sensitive to the threshold. If not noted otherwise, the threshold is set to $a = 3$ in the analysis below.

[Table 1 about here.]

One of the main contributions of the paper is to shed light on the the time-variation in the factor structure. I start by estimating the factor structure within each year, that is, I apply PCA-based estimators to each year independently; in Section III.E I study the time-varying structure at the monthly level. First, I determine the number of high-frequency, continuous, and jump factors (Section III.B). My main argument is based on the perturbed eigenvalue ratio diagnostic criterion, but I complement it by showing that including more factors than indicated by my estimator creates an unstable pattern when using either different samples of stocks or different time periods. Second, I show that the factor structure is essentially identical for the larger unbalanced panel and the balanced panel. This result implies that

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

without loss of generality I can study the factor structure for the representative balanced panel, which is important for considering the asset pricing applications that require long-term means. Third, I show that the high-frequency factors constructed using the portfolio weights estimated over the full horizon are essentially identical to those constructed using locally estimated portfolio weights. This result implies that I can use the portfolio weights estimated over the full horizon to interpret the factors. It gives me a benchmark "rotation" of the factors to study the time-varying loadings and long-term means of the factor returns.

As a first step Table I reports 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.[27] The choice of four continuous factors and one jump factor is motivated by the results in the subsequent sections. As expected systematic risk varies over time and is larger during the financial crisis. The systematic risk for four continuous factors accounts for around 40% to 47% of the total correlation over the 2008 to 2011 period, but explains only around 20% to 36% of the total correlation in the other years.[28] A similar pattern holds for 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.

B. Number of Factors

I estimate four high-frequency factors for the years from 2007 to 2012 and 2016 and three factors for the years 2004 to 2006 and 2013 to 2015. Figures 1 and 2 show the estimation results for the perturbed eigenvalue ratio diagnostic criterion for the balanced and unbalanced panels.[29] Starting from the right, I am looking for a strong increase in the perturbed eigenvalue ratio. While asymptotically any critical value larger than one should indicate the beginning of the systematic spectrum, for my finite sample I need to choose a critical value. Based on the simulation studies in Pelger (2019) I set the critical value to 1.08 in these figures. As can be seen, clearly visible increases at four for the years 2007 to 2012 and 2016, and at three for the years 2004 to 2006 and 2013 to 2015, can be detected for a wide range of critical values. My estimator therefore strongly indicates that there are four high-frequency factors from 2007 to 2012 and 2016 and three high-frequency factors for the other years. The number of factors is the same for the balanced and unbalanced panels except for the year

[27] Each year I apply PCA to the yearly continuous and jump quadratic correlation matrices to estimate the underlying factor structure.

[28] 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.

[29] I obtain the same results when I conduct this analysis using alternative perturbation functions or with the Onatski (2010) eigenvalue difference estimator.

2015 where the balanced panel seems to have four factors. This finding is in line with the results in Table II below.

[Figure 1 about here.]

[Figure 2 about here.]

The number of continuous and jump factors should be bounded by the total number of total high-frequency factors.[30] When applying the diagnostic criterion to the continuous movements, the number of continuous factors appears to be the same, that is, four continuous factors from 2007 to 2012 and three continuous factors for the other years. The only outlier is the year 2015, for which the number of continuous factors is estimated to be four and five in the balanced and unbalanced panels, respectively. As this outlier result is sensitive to the cutoff threshold, it is likely due to estimation noise. The number of jump factors appears to be lower. Indeed, in most years there is only one jump factor, the diagnostic criterion applied to jumps identified by three standard deviations ($a = 3$) suggests one factor in six (seven) of the 13 years for the balanced (unbalanced) panel. Overall, the estimation is much noisier. When classifying jumps based on more standard deviations ($a > 3$), there are too few jump observations to make a reliable prediction. In particular, only a small number of co-jumps between different assets remain. Thus, while a relatively small number of assets jumping together may not be systematic from an economic point of view, it can lead to a large eigenvalue in the jump covariance matrix. Due to the instability in the estimation for jumps factors for $a > 3$, I focus on $a = 3$ in the following.

C. Balanced Panel

I find that the factor structure of the balanced subsample is representative of the full unbalanced data set, and thus I focus on the balanced panel in the following. Table II reports the generalized correlations between the first three and four latent continuous factors estimated on the full data and their intersections for each year. Generalized correlations equal to one indicate that the factors are the same. It is apparent that the first three continuous PCA factors are essentially identical using both data sets. The first four factors coincide for the years in which I estimate a four-factor structure. In the years in which I estimate only three factors, the fourth PCA factor can differ across the two data sets. This makes

[30] Section III in the Internet Appendix presents the results for the continuous and jump factors.

sense if there are only three factors in a given year, the fourth PCA factor is expected to fit noise and hence should not be the same in the two panels. I therefore view this evidence as confirmation of the number of estimated factors. The jump structure appears to share only one or two common jump factors. This result is again in line with the results on the number of factors in the previous subsection.

[Table 2 about here.]

D. High-Frequency Factors

The four continuous, high-frequency factors for the period 2004 to 2016 can be well approximated by industry factors which allows me to put an economic label on the statistical factors. As I can show that the factor portfolio weights for the continuous and high-frequency (continuous plus jumps) factors are identical, most of the analysis below is based on the continuous portfolio weights. In this section I study factor portfolio weights estimated over the full time horizon of 13 years. In Section E I extend this analysis to shorter local windows.

First, I examine the composition of the continuous factors. Pelger and Xiong (2019) suggest the use of proxy factors to interpret latent PCA factors. My first proxy factor is an equally weighted market portfolio. The second proxy factor has the 15%, and the third and fourth proxy factors have the 11%, largest portfolios weights of the corresponding statistical factors.[31] Figure 3 plots the portfolio weights of the proxy factors sorted according to industries. The Appendix IV provides details on the industry classifications. The second proxy factor is a long-short factor in the oil and finance industry. The third proxy factor is a finance industry factor. The fourth proxy factor appears to be an electricity industry factor.

[Figure 3 about here.]

Figure 4 shows the factor portfolio weights of the continuous PCA factors without the proxy shrinking. The stocks are again sorted according to industries. The first factor has only long positions of similar magnitude, which justifies the interpretation of an equally weighted market portfolio. The second continuous PCA factor has exactly the same interpretation as a long-short factor in the oil and finance industries with mostly negligible weights in the other industries. The third factor has the largest weights in the finance industry, but also nonnegligible weights in the oil and technology industries. The fourth factor has the largest weights in the electricity industry with some minor outliers in other industries.

[31] The fraction of largest portfolio weights is chosen to obtain a large generalized correlation with the high-frequency PCA factors.

[Figure 4 about here.]

Based on these insights I construct four industry factors: (1) an equally weighted market portfolio, (2) an equally weighted oil industry factor, (3) an equally weighted finance industry factor, and (4) an equally weighted electricity factor. In addition, I compare the latent factors to the four Fama-French-Carhart factors. Table III reports the generalized correlations for different continuous factors with the four continuous PCA factors estimated over the full time horizon. The proxy factors have generalized correlations of $\{1, 0.99, 0.95, 0.91\}$, which confirms that they approximate the latent PCA factors very well. Moreover, the generalized correlations between the PCA factors and the market, oil, and finance factors are $\{1, 0.98, 0.95\}$ indicating that these three factors capture three of the PCA factors very well, while adding the electricity factor provides a good but not perfect approximation of the fourth PCA factor. The results are robust to constructing the industry portfolios using all stocks in the unbalanced panel. Using the factor portfolio weights based on all high-frequency returns (continuous plus jumps), I construct high-frequency PCA factors with continuous returns, which are perfectly correlated with the continuous PCA factors.

[Table 3 about here.]

Next in Table III, I compare the portfolio weights based on jump, overnight, and daily log returns. As already noted, the high-frequency (continuous + jumps) factor portfolio weights have the same span as the continuous weights. As expected, the first four PCA jump portfolio weights are different, and share at most two weights in common factors with the continuous PCA. The portfolio weights suggest that the jumps pick up market and oil factors. However, when increasing the jump threshold (i.e. $a > 3$), the commonality between continuous and jump factors shrinks until they have only a market factor in common. Overnight and daily data seem to yield a similar but noisier pattern than the high-frequency returns. When moving to lower frequencies it seems that information is lost. The portfolios weights based on PCA applied to weekly returns has one less factor in common with the continuous returns, while the portfolio weights of monthly PCA factors have two fewer factors in common. The Internet Appendix reports the portfolio weights for overnight, daily, and weekly PCA factors. The results suggest that the lower frequencies have a less clear pattern and a weaker link to the industry classification. In particular, the daily and overnight portfolio weights show a similar pattern in the three industries and the market, but with more outliers, while the monthly portfolio weights seem to be linked only to a market portfolio and not to the industries. Figure 5 depicts these results.

[Figure 5 about here.]

The Fama-French-Carhart factors and the continuous PCA factors, only have the market factor in common. Indeed, while Table III shows that the market factor has the almost perfect correlation of 0.98 with the continuous PCA factors, the size and value factors only have a correlation of 0.65 and 0.43, respectively, and the momentum factor is essentially orthogonal to the PCA factors.[32]

E. Time-Variation

Using a short time horizon of one month (21 trading days), I study the time-variation in the portfolio weights and loadings of different factors; the Internet Appendix reports the results for a one-week horizon (5 trading days) and a three-month window (63 trading days) with essentially identical results. The four continuous PCA factors are very stable over time, while the Fama-French-Carhart factors have a time-varying factor structure. Given the four continuous PCA factors estimated over the full time horizon, the industry factors and the Fama-French-Carhart factors, I estimate the loadings for each on a rolling window of one month with continuous log returns.

First, I calculate the generalized correlations of the loadings estimated on the full time horizon with those estimated over the moving window as depicted in Figure 6. Here, I keep the factor portfolio weights constant but allow for arbitrary monthly time-variation in the loadings. Surprisingly, the loadings on the PCA and industry factors are stable over time. However, the loadings on the size, value and momentum factors do not have the same span for different time periods. This finding does not imply that the loadings for the PCA factors are constant over time. The finding that the generalized correlation of the PCA loadings is the same for different time periods corresponds to a model of the form

$$\Lambda(t) = \underbrace{\bar{\Lambda}}_{N\times K}\underbrace{H(t)}_{K\times K}. \tag{12}$$

This implies that the projection of the return space on the common component or residual space is the same for each $\Lambda(t)$, that is, the cross-sectional relationship has a stable structure.

[32] In the Internet Appendix I show that the above findings are robust to the type of returns that I study. Specifically, using the same portfolio weights I calculate the intraday, overnight, and daily returns for the continuous PCA, high-frequency PCA, proxy PCA, industry and Fama-French-Carhart factor portfolio weights. The generalized correlations are identical to the continuous returns, that is, the proxy and industry factors provide a good approximation while the size, value, and momentum factors are only weakly correlated with the PCA factors.

This justifies the estimation of the portfolio weights on PCA factors over the full time horizon. However, the loadings on an individual asset can have very different values for different time periods. Importantly, this stable structure does not hold for the Fama-French-Carhart factors, implying that regressions of stocks on these factors are biased even for very short time horizons.

[Figure 6 about here.]

My previous analysis on the number of factors indicates that some time periods have more systematic factors than others. This does not contradict the stable span of the loadings. It is possible that the loadings on a factor take very small values in a specific time period or that the volatility of a factor is locally very small and thus this factor does not explain a systematic portion of the correlation in the data for this period. In this case PCA would not detect this factor. Figure 7 estimates the PCA factor portfolio weights on the local one-month window and the total time horizon and depicts the generalized correlations. Note that in contrast to the previous analysis, I apply a separate PCA to each of the local one-month windows to obtain the local factor portfolio weights. When estimating only the first four PCA factors, the highest generalized correlations correspond to the 2007 to 2012 periods, with a sharp drop for the other years. When adding the first seven PCA factors, the fifth to seven generalized correlations are close to zero, which is a clear indication that they are fitting noise. Overall, these findings confirm the estimation for the number of factors.

[Figure 7 about here.]

Figure 8 takes a closer look at the time-variation in the locally estimated continuous PCA factors by plotting the factor portfolio weights for November 2006 and April 2008. In 2006 there is no finance factor present, resulting in the three-factor structure. In 2008 the fourth PCA factor is clearly loading heavily on the finance industry. This suggests that the fourth factor over the 2007 to 2012 period is the finance factor. This hypothesis is supported by the generalized correlation between the locally estimated PCA factors and different combinations of the industry factors. The strong increase in the generalized correlation from 2007 to 2012 appears only when including the finance factor.

[Figure 8 about here.]

The amount of variation explained by the factors changes over time. Table I above indicates that the proportion of the variation explained by factors increased during the

financial crisis. Figure 9, which provides a more refined analysis using the monthly window with local regressions, confirms this pattern. The proportion of systematic risk rises from 2009 to 2012 and again at the end of 2015. The four continuous PCA and industry factors explain roughly the same proportion of risk, while the Fama-French-Carhart factors capture a smaller portion. The market factor alone already explains a large part of the variation. The amount of explained variation is calculated with time-varying continuous loadings. In Table IV I compare the overall variation explained with time-varying continuous loadings to that explained with constant continuous loadings. I report the increment in explained variation relative to the market factor. Because all of the four-factor models span approximately the market factor, the increment can be interpreted as the variation explained by the three remaining factors orthogonal to the market. As expected, a time-varying loading model can capture more variation than the constant loading model. The gain for the continuous factors is modest, as expected given the stability results about the span in the loadings. The increase for the Fama-French-Carhart factors almost doubles, which is in line with the more pronounced time-variation in the loadings for this model. Recall that Table III above shows that the PCA factors estimated at lower frequencies differ from the high-frequency PCA factors. Table IV indicates that the lower frequencies result in a loss in information as the lower-frequency PCA factors explain less of the variation in the data.

[Figure 9 about here.]

[Table 4 about here.]

In Figures 10 and 11 I take a closer look at the source of the time-variation in the factor structure of the continuous PCA and Fama-French-Carhart factors. Specifically, I keep the factor portfolio weights constant and estimate the regression loadings and volatilities locally at the one-month horizon. First, I study the time-variation in the systematic part, which is the product of the loadings and the corresponding factor volatility $\frac{\Lambda_k(t)^\top \Lambda_k(t)}{N}\sigma_k^2(t)$. This product corresponds roughly to the eigenvalues that are due to the individual factors. The left plots show the absolute values while the right plots are a normalization with the time average of the quantity. As can be seen, all factors have a larger systematic portion from 2008 to 2012. When the eigenvalues due to the second and third finance-related factors are small, I estimate only three factors. The middle plots show the average loadings $\frac{\Lambda_k(t)^\top \Lambda_k(t)}{N}$ over time. Here is the biggest difference between the PCA and Fama-French-Carhart factors. The average loadings on the PCA factors are almost constant, while the average loadings on the Fama-French-Carhart factors fluctuate wildly, which supports the finding that this

factor structure is not stable. Finally, as expected, the volatility of the different factors varies substantially over time, which is why the number of PCA factors is time-varying.

[Figure 10 about here.]

Overall, the findings of this section suggest that PCA analysis applied to a 13-year horizon on intraday data provides a valid set of factors, and hence it is not necessary to estimate the factors locally. In contrast, even for very short time horizons Fama-French-Carhart factors require an adjustment for time-variation. The high-frequency PCA results do not extend to lower frequencies: PCA factors extracted from monthly returns are different, and Lettau and Pelger (2020) show that the monthly PCA portfolio weights of individual monthly stock returns are not stable over the longer time horizon of 40 years.

[Figure 11 about here.]

F. Asset Pricing

APT postulates a connection between factors that explain the comovement and the cross-section of expected returns. Under the assumption of APT the SDF is spanned by the factors that explain the systematic comovement. Hence, the tangency portfolio with the optimal Sharpe ratio has to consist only of these factors. Comparing the Sharpe ratios of mean-variance efficient portfolios based on factors therefore tests the pricing performance of the factors.[33]

Table V reports the maximum Sharpe ratios for tangency portfolios based on different factors. The factors are excess returns of traded assets and hence their risk premium can be calculated using their time-series mean.[34] The four daily PCA factors have a combined annual Sharpe ratio of 0.4, which is the same as for a market factor. In contrast, the tangency portfolio based on the Fama-French-Carhart daily factor returns strongly outperforms the market portfolio for this time period. As before, continuous PCA factors are constructed using the portfolio weights estimated from the continuous loadings over the full time horizon. The corresponding tangency portfolio significantly outperforms the characteristic-based factors at the daily horizon.

[33] My focus on Sharpe ratios is motivated by Barillas and Shanken (2018), who show that a Sharpe ratio criterion can compare asset pricing performance of factors regardless of the test assets.

[34] The Sharpe ratio is my measure of factors' risk premium. As my latent factors are excess returns and hence zero-cost portfolios their scaling is not identified. I normalize all factors by their standard deviation. As a result, their time-series mean equals their Sharpe ratio.

[Table 5 about here.]

Interestingly, the risk premium earned intraday and overnight differs significantly for characteristic-based and statistical factor portfolios. The four statistical factors earn the largest part of the risk premium intraday, which is not surprising as they are estimated to explain the intraday variation. In contrast, the size, value, and momentum factors earn their positive risk compensation overnight while the intraday compensation is minor and negative. The large intraday risk premium of the PCA factors reverses overnight. In particular, the second and fourth factors show a strong reversal pattern. This raises the question of what type of risk leads to this result. Lou, Polk, and Skouras (2018) show that institutional investors tend to initiate trades at the close while individuals are more likely to initiate trades near the open. Specifically, they show that small trades linked to individuals are more likely to occur near the open while large trades linked to institutions are more likely to occur near the close. Trading near the close may also not be purely information based but also due to rebalancing requirements, which could again be linked to institutional capital flows. A positive intraday return is associated with higher demand near the close, while a negative overnight return should follow from lower demand near the open. Hence, the reversal pattern in the factor risk premia might be driven by the demands of different clienteles.

Based on the intraday and the overnight means and covariances, I construct the tangency portfolios for each time segment. The first two columns of Table V show that the intraday tangency portfolio of PCA factors has almost twice the Sharpe ratio as the daily return portfolio. In contrast, the overnight tangency portfolio of the Fama-French-Carhart factors has twice the Sharpe ratio compared to the daily returns. These results confirm that the PCA factors are compensated for intraday risk while the characteristic-based factors earn risk compensation mainly overnight. These observations suggest that a long-short strategy with intraday long position and overnight short position in the tangency portfolios (respectively overnight long position and intraday short position for the characteristic-based factors) should perform better than a simple long position in the optimal portfolios. Indeed, the long-short strategy of optimal intraday and overnight portfolios yields an annual Sharpe ratio of 2.5 for the PCA factors and 1.7 for the Fama-French-Carhart factors.

To guard against survivorship bias, I construct PCA factors based on the full panel of unbalanced returns. I estimate four continuous PCA factors in each year and rotate them to be as close as possible to the continuous PCA factors estimated on the balanced panel over the full time horizon. This rotation is necessary to connect the yearly time series of the locally estimated factors over the full time horizon. Because in most years the span of the PCA factors based on the balanced panel is close to that of locally estimated factors based

on all stocks, I expect to obtain a close approximation of the PCA factors when conditioning on survival only for one year instead over the full time horizon. In addition, I include the industry factors based on the unbalanced panel of all stocks. The results for the tangency portfolio are robust to the survivorship bias that the balanced panel is exposed to. The Sharpe ratio for intraday returns dominates the daily Sharpe ratio. The reversal pattern in the second and fourth PCA factors are not affected by selection bias. Similar results hold using the industry factors.

[Figure 12 about here.]

In Figure 12 I take a closer look at the reversal pattern in stock returns and portfolios. The expected intraday excess returns are clearly negatively correlated with expected overnight returns. The same pattern holds for the expected returns of the 14 industry portfolios based on the unbalanced panel of all stocks. The six size- and value-sorted portfolios show only a very weak reversal pattern. The overnight reversal pattern is captured by the continuous PCA factors. Figure 13 plots cumulative returns for the four statistical and characteristic-based factors for daily returns and their intraday and overnight components. As expected, the PCA factors show a strong reversal pattern, that is, they earn higher average returns during the day but the returns reverse overnight, resulting in lower daily returns. The same result holds for the approximation of the PCA factors based on the unbalanced panel. In contrast, the Fama-French-Carhart factors show a strongly increasing return pattern for size and value overnight that is lower for daily returns and close to zero during the day.

[Figure 13 about here.]

How well can the statistical factors explain asset returns? I run time-series regressions with intercept on different test assets for their daily, intraday, and overnight returns for two sets of factors. As the estimated means of individual stocks are noisy, I test 14 industry portfolios based on the unbalanced panel and the six size- and value-sorted portfolios. Because the industry portfolios exhibit strong overnight reversal, I expect the PCA factors to better price these assets. Indeed, Figure 14 shows that the PCA factors do indeed explain the industry portfolios better than the Fama-French-Carhart factors. Separating the returns into their intraday and overnight components, I show that the PCA factors succeed in explaining the intraday and overnight patterns. In contrast, the Fama-French-Carhart factors essentially miss the intraday risk premium and assign too large a risk premium to the overnight returns. As this over- and underestimation partially cancels out, the daily pricing errors are smaller than for the intraday and overnight components. I next repeat this analysis with

time-varying loadings and pricing errors, using a 120 day moving window to estimate the intercepts and predicted returns, which are then averaged over time. The results are virtually identical, suggesting that time-variation in the loadings does not improve the weaker performance of the Fama-French-Carhart factors.

[Figure 14 about here.]

In Figure 15 I test the six size- and value-sorted portfolios.[35] As expected, these portfolios are better explained by the Fama-French-Carhart factors. In particular, the overnight returns are better explained by the anomaly factors compared to PCA. This finding suggests that intraday comovement does not capture size and value information. Estimating predicted returns and pricing errors with a 120 day moving window yields similar results. The performance of the PCA factors does not improve with the time-varying loadings and pricing errors, while the Fama-French-Carhart factors have slightly better pricing performance.[36] These results are in line with the results in Section III.E showing that the characteristic-based factors are less stable over time.

[Figure 15 about here.]

IV. Conclusion

This paper makes four main contributions to the literature. First, by estimating latent factors as opposed to relying on pre-specified observable factors, I show that a low-dimensional model can explain the comovement in intraday stock returns. Time-variation in the factor structure matters, but is subtle. The span of the regression loadings for my high-frequency PCA factors is stable over time. This means that projections on the systematic or idiosyncratic return component are (approximately) the same if they are done locally or with loadings estimated over the full time horizon. However, due to the time-varying volatility matrix of the factors, the contribution of a factor to the systematic component and the regression loadings on individual stocks vary over time. In contrast, the Fama-French-Carhart factors have a time-varying span over the regression loadings and lead to biased pricing errors for stocks if they are not estimated locally. Second, I show that the type of factors that are

[35] The characteristic-sorted portfolios are labeled as follows: 1=small growth, 2= small neutral, 3= small value, 4= big growth, 5= big neutral, 6=big value.

[36] The additional results are presented in the Internet Appendix.

obtained by PCA-based methods depends crucially on the underlying assets. Characteristic-based factors are obtained by applying PCA to characteristic-managed portfolios as in Kelly, Pruitt, and Su (2017). However, most of the time-variation in individual stocks returns is explained by industry factors. This raises the question of how characteristics are related to covariances of individual stocks. Third, I show that the systematic structure of smooth continuous intraday movements is different than that of rough intraday and overnight jumps. Finally, I show that the factors that explain the intraday comovements in stocks earn a significant intraday risk premium. However, as this risk premium reverses overnight, the daily risk premium is comparatively low for these factors. This finding suggests that studying only the cross-section of daily returns might neglect risk-return trade-offs that have an overnight reversal. The findings of this paper have direct implications for investment strategies as there is sizeable risk compensation from exploiting the reversal pattern.

REFERENCES

Aït-Sahalia, Yacine, and Jean Jacod, 2014, *High-Frequency Financial Econometrics* (Princeton University Press, NY).

Aït-Sahalia, Yacine, and Dacheng Xiu, 2017, Principal component estimation of a large covariance matrix with high-frequency data, *Journal of Econometrics* 201, 384–399.

Aït-Sahalia, Yacine, and Dacheng Xiu, 2019, Principal component analysis of high frequency data, *Journal of American Statistical Association* 114, 287–303.

Alexeev, Vitali, Mardi Dungey, and Wenying Yao, 2017, Time-varying continuous and jump betas: The role of firm characteristics and periods of stress, *Journal of Empirical Finance* 40, 1–19.

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

Back, Kerry, 1991, Asset prices for general processes, *Journal of Mathematical Economics* 20, 371–395.

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

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

Bai, Jushan, and Serena Ng, 2006, Evaluating latent and observed factors in macroeconomics and finance, *Journal of Econometrics* 1-2, 507–537.

Bali, Turan T., Robert F. Engle, and Yi Tang, 2017, Dynamic conditional beta is alive and well in the cross-section of daily stock returns., *Management Science* 63, 3760–3779.

Barillas, Francisco, and Jay Shanken, 2018, Comparing asset pricing models, *Journal of Finance* 73, 715–754.

Barndorff-Nielsen, Ole E., and Neil Shephard, 2002, Econometric analysis of realized volatility and its use in estimating stochastic volatility models, *Journal of the Royal Statistical Society* 64, 253–280.

Berkman, Henk, Paul Koch, Laura Tuttle, and Ying Zhang, 2012, Paying attention: Overnight returns and the hidden cost of buying at the open, *Journal of Financial and Quantitative Analysis* 47, 715–741.

Bollerslev, Tim, Sophia Z. Li, and Viktor Todorov, 2013, Jump tails, extreme dependencies and the distribution of stock returns, *Journal of Econometrics* 172, 307–324.

Bollerslev, Tim, Sophia Z. Li, and Viktor Todorov, 2016, Roughing up beta: Continuous vs. discontinuous betas, and the cross section of expected stock returns, *Journal of Financial Economics* 120, 464–490.

Bollerslev, Tim, and Viktor Todorov, 2011, Estimation of jump tails, *Econometrica* 79, 1727–1783.

Branch, Ben, and Aixin Ma, 2012, Overnight return, the invisible hand behind intraday returns, *Journal of Applied Finance* 22, 90–100.

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

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

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

Cliff, Michael, Michael Cooper, and Huseyin Gulen, 2008, Return differences between trading and non-trading hours: Like night and day, Technical report, Virginia Tech.

Connor, Gregory, 1984, A unified beta pricing theory, *Journal of Economic Theory* 34, 13–31.

Connor, Gregory, and Robert A. Korajczyk, 1988, Risk and return in an equilibrium apt: Application to a new test methodology, *Journal of Financial Economics* 21, 255–289.

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

Duffie, Darrell, Jun Pan, and Kenneth Singleton, 2000, Transform analysis and asset pricing for affine jump-diffusions, *Econometrica* 68, 1343–1376.

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

Eraker, Bjorn, Michael Johannes, and Nick Polson, 2003, The impact of jumps in volatility and returns, *Journal of Finance* 58.

Fama, Eugene F., and Kenneth R. French, 1992, The cross section of expected stock returns, *Journal of Finance* 47, 427–465.

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

Fama, Eugene F., and Kenneth R. French, 2020, Comparing cross-section and time-series factor models, *Review of Financial Studies, forthcoming* .

Fama, Eugene F., and James D. MacBeth, 1973, Risk, return and equilibrium: Empirical tests, *Journal of Political Economy* 81, 607–636.

Fan, Jianqing, Alex Furger, and Dacheng Xiu, 2016a, Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high frequency data, *Journal of Business and Economic Statistics* 34, 489–503.

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

Fan, Jianqing, Yuan Liao, and Weichen Wang, 2016b, Projected principal component analysis in factor models, *Annals of Statistics* 44, 219–254.

Gabaix, Xavier, 2012, Variable rare disasters: An exactly solved framework for ten puzzles in macrofinance, *Quarterly Journal of Economics* 127, 645–700.

Giglio, Stefano, and Dacheng Xiu, 2017, Asset pricing with omitted factors, Technical report, Yale.

Jagannathan, Ravi, and Zhenyu Wang, 1996, The conditional CAPM and the cross-section of expected stock returns, *Journal of Finance* 51, 3–53.

Kelly, Bryan, Seth Pruitt, and Yinan Su, 2017, Instrumented principal component analysis, Technical report, Yale.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2020, Shrinking the cross section, *Journal of Financial Economics, forthcoming* 135, 271–292.

Lettau, Martin, and Sydney Ludvigson, 2001, Resurrecting the (C)CAPM: A cross-sectional test when risk premia are time-varying, *Journal of Political Economy* 1238–1287.

Lettau, Martin, and Markus Pelger, 2020, Factors that fit the time-series and cross-section of stock returns, *Review of Financial Studies, forthcoming* .

Lou, Dong, Christopher Polk, and Spyros Skouras, 2018, A tug of war: Overnight versus intraday expected returns, Technical report, London School of Economics.

Nagel, Stefan, 2013, Empirical cross-sectional asset pricing, *Annual Review of Financial Economics,* 5, 167–199.

Onatski, Alexei, 2010, Determining the number of factors from empirical distribution of eigenvalues, *Review of Economic and Statistics* 92, 1004–1016.

Pan, Jun, 2002, The jump risk premium implicit in options: Evidence from an integrated time-series study, *Journal of Financial Economics* 63, 3–50.

Pelger, Markus, 2019, Large-dimensional factor modeling based on high-frequency observations, *Journal of Econometrics* 208, 23–42.

Pelger, Markus, and Ruoxuan Xiong, 2019, Interpretable proximate factors for large dimensions, Technical report, Stanford University.

Reisman, Haim, 1992, Intertemporal arbitrage pricing theory, *Review of Financial Studies* 5, 105–122.

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

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

Table I
Summary Statistics for Continuous and Jump Returns

This table presents the fraction of jump increments and the explained variation for the balanced and unbalanced panel.

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
	Intersection of balanced panel ($N = 332$)												
	Fraction of increments identified as jump for different thresholds (in %)												
$a = 3$	0.89	0.87	0.79	0.84	0.71	0.60	0.60	0.56	0.63	0.60	0.56	0.53	0.53
$a = 4$	0.20	0.19	0.17	0.19	0.14	0.11	0.12	0.09	0.14	0.13	0.11	0.10	0.10
$a = 4.5$	0.10	0.10	0.09	0.11	0.07	0.06	0.06	0.04	0.07	0.07	0.06	0.05	0.05
$a = 5$	0.05	0.05	0.05	0.06	0.04	0.03	0.03	0.02	0.04	0.04	0.03	0.03	0.02
	Fraction of total quadratic variation explained by jumps												
$a = 3$	0.16	0.16	0.14	0.19	0.15	0.14	0.13	0.09	0.13	0.13	0.12	0.25	0.11
$a = 4$	0.06	0.06	0.06	0.09	0.06	0.06	0.06	0.03	0.05	0.05	0.05	0.19	0.04
$a = 4.5$	0.04	0.04	0.04	0.07	0.04	0.05	0.04	0.02	0.04	0.03	0.03	0.17	0.03
$a = 5$	0.03	0.03	0.03	0.06	0.03	0.04	0.03	0.01	0.03	0.03	0.03	0.17	0.02
	Fraction of jump correlation explained by first jump factor												
$a = 3$	0.03	0.03	0.03	0.05	0.07	0.08	0.20	0.11	0.06	0.10	0.05	0.08	0.05
$a = 4$	0.02	0.02	0.04	0.07	0.05	0.07	0.28	0.07	0.07	0.18	0.07	0.09	0.06
$a = 4.5$	0.02	0.01	0.05	0.08	0.05	0.07	0.29	0.07	0.09	0.22	0.10	0.09	0.07
$a = 5$	0.02	0.01	0.05	0.09	0.06	0.08	0.30	0.07	0.11	0.22	0.12	0.09	0.07
	Fraction of continuous correlation explained by first four continuous factors ($a = 3$)												
	0.20	0.21	0.22	0.29	0.44	0.41	0.40	0.47	0.30	0.27	0.32	0.36	0.33
	All observations (unbalanced panel)												
N	555	607	649	649	674	674	667	659	609	656	668	650	643
	Fraction of increments identified as jump for different thresholds (in %)												
$a = 3$	0.87	0.82	0.75	0.80	0.67	0.58	0.57	0.53	0.59	0.57	0.53	0.52	0.52
$a = 4$	0.19	0.17	0.16	0.17	0.13	0.10	0.11	0.09	0.13	0.12	0.10	0.10	0.10
$a = 4.5$	0.10	0.09	0.08	0.10	0.06	0.05	0.06	0.04	0.06	0.06	0.05	0.05	0.05
$a = 5$	0.05	0.05	0.05	0.06	0.03	0.02	0.03	0.02	0.04	0.04	0.03	0.03	0.02
	Fraction of total quadratic variation explained by jumps												
$a = 3$	0.17	0.15	0.14	0.21	0.13	0.11	0.12	0.09	0.12	0.12	0.11	0.36	0.11
$a = 4$	0.06	0.06	0.05	0.11	0.05	0.03	0.05	0.03	0.05	0.04	0.04	0.31	0.04
$a = 4.5$	0.04	0.04	0.04	0.09	0.03	0.02	0.04	0.02	0.03	0.03	0.03	0.30	0.03
$a = 5$	0.03	0.03	0.03	0.08	0.02	0.01	0.03	0.01	0.03	0.02	0.03	0.30	0.03
	Fraction of jump correlation explained by first jump factor												
$a = 3$	0.03	0.03	0.03	0.24	0.07	0.08	0.24	0.10	0.06	0.11	0.06	0.82	0.05
$a = 4$	0.02	0.03	0.04	0.36	0.05	0.07	0.36	0.08	0.07	0.19	0.10	0.92	0.09
$a = 4.5$	0.03	0.04	0.05	0.45	0.05	0.07	0.41	0.12	0.08	0.23	0.13	0.94	0.11
$a = 5$	0.04	0.05	0.06	0.51	0.05	0.06	0.46	0.17	0.09	0.23	0.15	0.95	0.14
	Fraction of continuous correlation explained by first four continuous factors ($a = 3$)												
	0.22	0.23	0.24	0.31	0.47	0.42	0.42	0.49	0.33	0.29	0.34	0.39	0.36

Table II
Generalized Correlation of Factors from Unbalanced and Balanced Panel.

Note: This Table presents the generalized correlations between three or four PCA factors estimated on continuous and jump ($a = 3$) returns using the balanced and unbalanced data ($N = 332$ for balanced data). The number of high-frequency factors estimated and the number of stocks in the unbalanced panel are indicated in the top row.

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
$\hat{K}$	3	3	3	4	4	4	4	4	4	3	3	3	4
N	555	607	649	649	674	674	667	659	609	656	668	650	643
	First three continuous PCA factors												
1. GC	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2. GC	0.99	0.99	0.99	0.98	0.98	0.97	0.96	0.97	0.97	0.97	0.98	0.99	0.98
3. GC	0.96	0.87	0.92	0.87	0.97	0.90	0.90	0.87	0.97	0.94	0.98	0.98	0.98
	First four continuous PCA factors												
1. GC	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2. GC	0.99	0.99	0.99	0.98	0.99	0.98	0.96	0.97	0.98	0.98	0.98	0.99	0.98
3. GC	0.96	0.96	0.93	0.96	0.97	0.97	0.96	0.96	0.97	0.96	0.98	0.98	0.98
4. GC	0.92	0.64	0.71	0.94	0.91	0.95	0.90	0.92	0.91	0.14	0.57	0.88	0.93
	First three jump PCA factors												
1. GC	0.97	0.97	0.97	0.99	0.98	0.99	1.00	0.99	0.99	0.99	0.99	1.00	0.98
2. GC	0.37	0.48	0.90	0.08	0.97	0.88	0.99	0.93	0.90	0.71	0.97	0.99	0.42
3. GC	0.07	0.05	0.16	0.00	0.93	0.51	0.94	0.03	0.45	0.02	0.07	0.96	0.01
	First four jump PCA factors												
1. GC	0.97	0.97	0.98	0.99	0.98	0.99	1.00	0.99	0.99	0.99	0.99	1.00	0.98
2. GC	0.52	0.53	0.92	0.94	0.97	0.89	0.99	0.94	0.94	0.91	0.97	0.99	0.96
3. GC	0.17	0.14	0.88	0.08	0.93	0.57	0.96	0.76	0.89	0.06	0.74	0.98	0.08
4. GC	0.06	0.06	0.00	0.00	0.68	0.02	0.83	0.04	0.02	0.02	0.08	0.03	0.01

Table III
Generalized Correlations between Continuous PCA Factors with Other Factors

The top panel of this table presents the continuous generalized correlations of the first four statistical continuous PCA factors with the four PCA factors based on all high-frequency data (continuous+jumps), the four continuous PCA proxy factors, different combinations of industry factors (market, oil, finance and electricity) based on the balanced and unbalanced panel, the four Fama-French-Carhart continuous factors, the three Frama-French factors, and the continuous market factor. The bottom panel of this table presents the generalized correlations of factor portfolio weights of the four continuous PCA factors with portfolio weights of 4 PCA factors based on all high-frequency data (continuous+jumps), jump, overnight, daily, weekly or monthly returns. The results correspond to the balanced panel from 2004 to 2016 ($M = 252,021$ high-frequency increments, $T = 3,273$ days, and $N = 332$).

	1. GC	2. GC	3. GC	4. GC
Continuous generalized correlations with four continuous PCA factors				
HF PCA	1.00	1.00	1.00	1.00
PCA Proxy	1.00	0.99	0.95	0.91
Industry (M,O,F,E)	1.00	0.98	0.95	0.78
Industry (M,O,F)	1.00	0.98	0.95	0.00
Industry (M,F,E)	1.00	0.98	0.78	0.00
Industry (M,O,E)	1.00	0.96	0.78	0.00
Industry (M,O,F,E, unbalanced)	1.00	0.96	0.90	0.77
Fama-French-Carhart	0.98	0.66	0.45	0.05
Fama-French 3	0.98	0.65	0.43	0.00
Market	0.98	0.00	0.00	0.00
Generalized correlations of factor portfolio weights with 4 continuous PCA factors				
ω HF	1.00	1.00	1.00	1.00
ω Jump	0.98	0.96	0.65	0.39
ω Overnight	0.99	0.98	0.93	0.82
ω Daily	1.00	0.98	0.97	0.94
ω Week	0.98	0.94	0.79	0.20
ω Month	0.89	0.54	0.31	0.10

Table IV
Explained Variation for Constant and Time-Varying Loadings

This table presents the increments in explained variation R^2 relative to the market factor. Continuous factor loadings are estimated locally on a moving window of one month (21 trading days) or over the whole time-horizon, where the increments in explained variation is the difference between the explained variation of a four factor model and the explained variation of the market factor.

	Continous PCA	Weekly PCA	Monthly PCA	FFC	Market
	Time-varying continuous loadings				
HF R^2	0.09	0.08	0.06	0.04	0.29
Daily R^2	0.09	0.08	0.06	0.06	0.40
Intraday R^2	0.11	0.10	0.08	0.06	0.38
Overnight R^2	0.08	0.07	0.05	0.03	0.38
	Constant continuous loadings				
HF R^2	0.07	0.06	0.05	0.02	0.27
Daily R^2	0.10	0.08	0.06	0.03	0.37
Intraday R^2	0.10	0.08	0.06	0.03	0.35
Overnight R^2	0.07	0.06	0.04	0.02	0.34

Table V
Sharpe Ratios of Factors Intraday and Overnight

The top panel of the table presents the maximum Sharpe ratio of the optimal factor portfolio for intraday, overnight and daily (intraday + overnight) returns. The bottom panel presents the Sharpe ratios of individual factors for intraday, overnight, and daily returns.

	Intraday	Overnight	Daily
Sharpe ratio of tangency portfolio			
Continuous PCA	1.45	1.02	0.89
Continuous PCA (unbalanced)	1.46	1.24	0.95
Proxy PCA	1.09	1.09	0.65
Industry	0.84	0.71	0.61
Industry (unbalanced)	0.73	0.79	0.51
Fama-French-Carhart	0.41	1.32	0.60
PCA Overnight	0.43	1.01	0.42
PCA Daily	0.16	0.47	0.40
Sharpe ratios of individual factors			
1. Continuous PCA Factor	0.33	-0.03	0.25
2. Continuous PCA Factor	0.43	-0.61	0.04
3. Continuous PCA Factor	0.62	0.09	0.57
4. Continuous PCA Factor	1.05	-0.79	0.52
Market	0.16	0.47	0.41
Size	-0.07	0.71	0.27
Value	-0.17	0.48	0.17
Momentum	-0.14	0.69	0.22
1. Cont. PCA Factor (unbalanced)	0.47	0.20	0.49
2. Cont. PCA Factor (unbalanced)	0.67	-0.56	0.26
3. Cont. PCA Factor (unbalanced)	0.49	0.47	0.65
4. Cont. PCA Factor (unbalanced)	1.21	-0.90	0.59

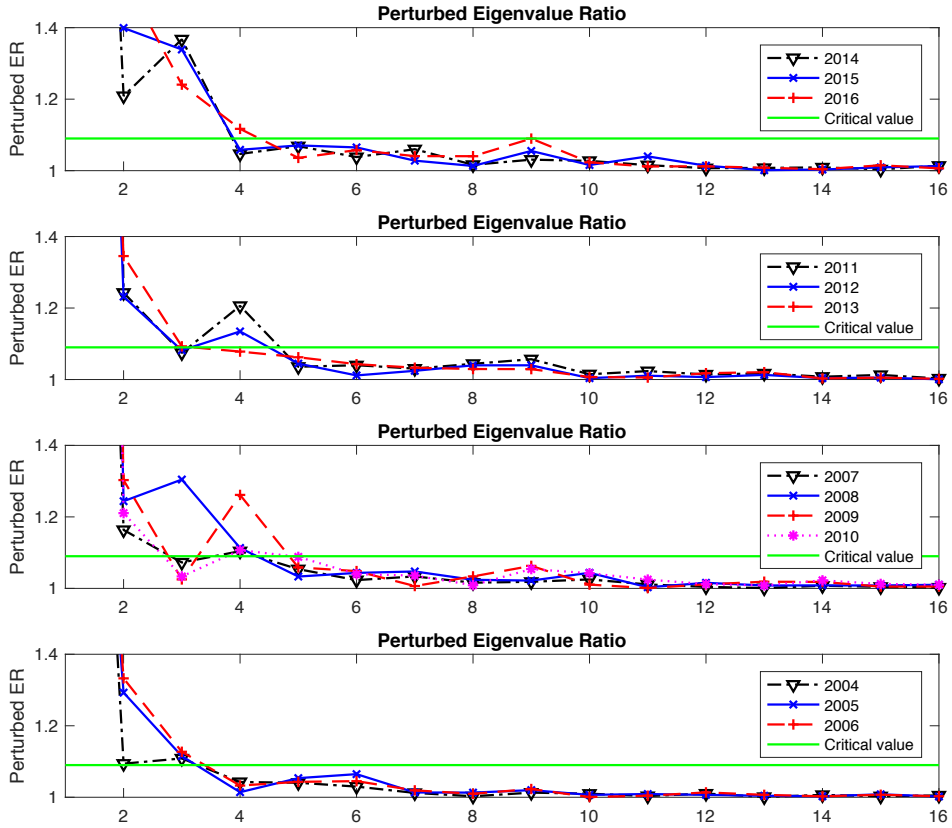

Figure 1. Number of HF factors. This figure plots the perturbed eigenvalue ratio statistics for the unbalanced panel of all high-frequency (continuous + jump) returns.

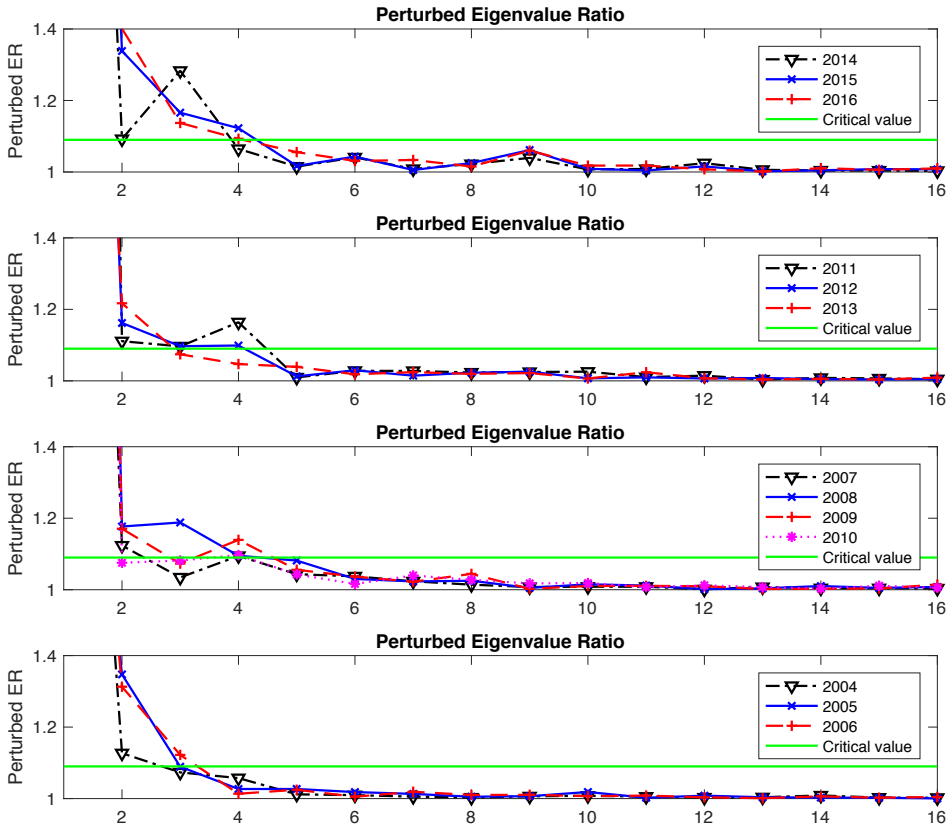

Figure 2. Number of HF factors. This figure plots the perturbed eigenvalue ratio statistics for the balanced panel of high-frequency (continuous + jump) returns.

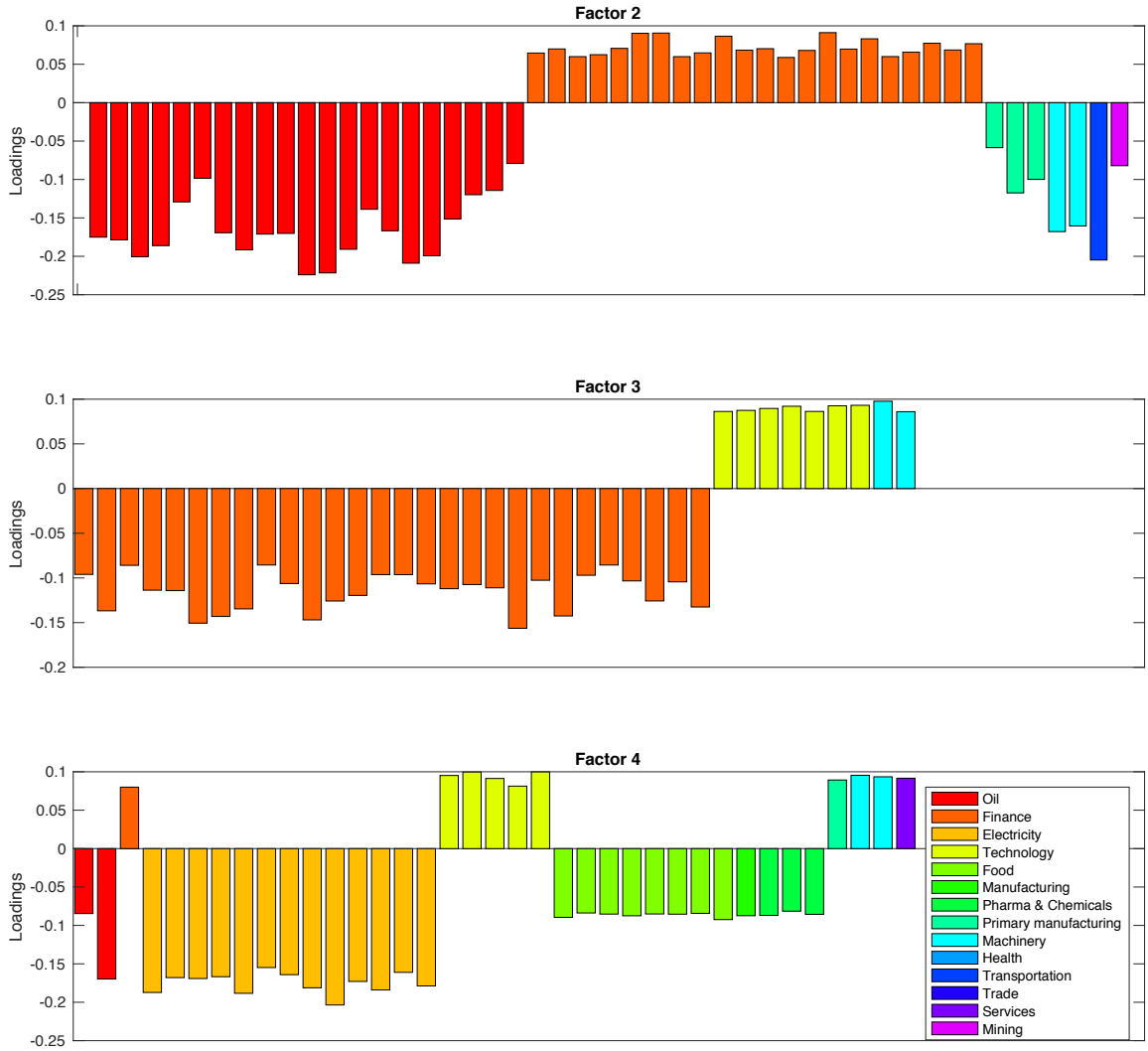

Figure 3. Portfolio weights of proxy factors. This figure shows the portfolio weights of the proxy factors for the four continuous PCA factors. The first proxy factor is an equally weighted market portfolio. The second proxy factors has 15%, and the third and fourth proxy factors have the 11%, largest portfolios weights of the corresponding statistical factors.

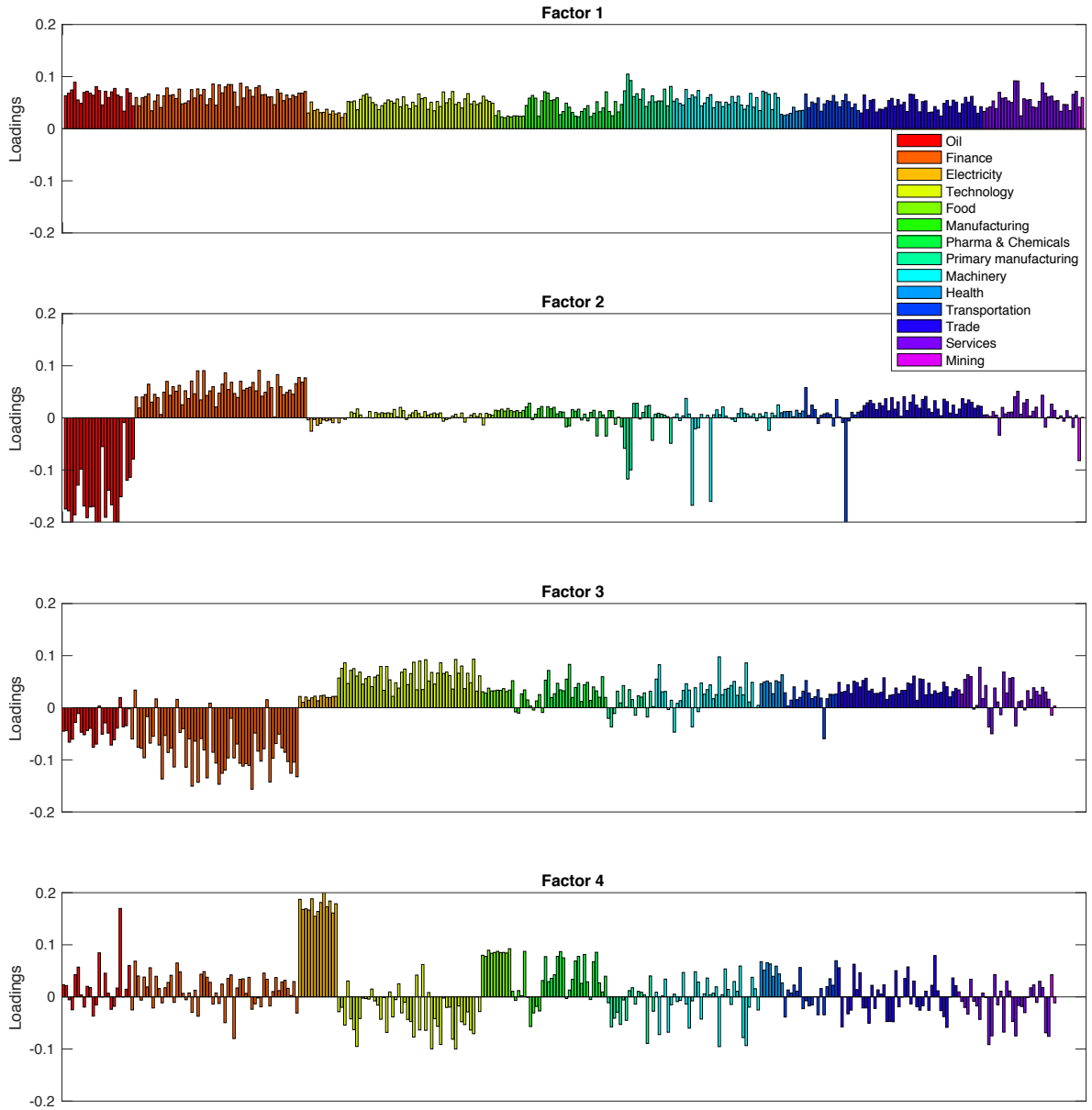

Figure 4. Portfolio weights of four continuous PCA factors. This figure depicts the portfolio weights of four continuous PCA factors over the full time horizon from 2004 to 2016. Stocks are sorted according to industries.

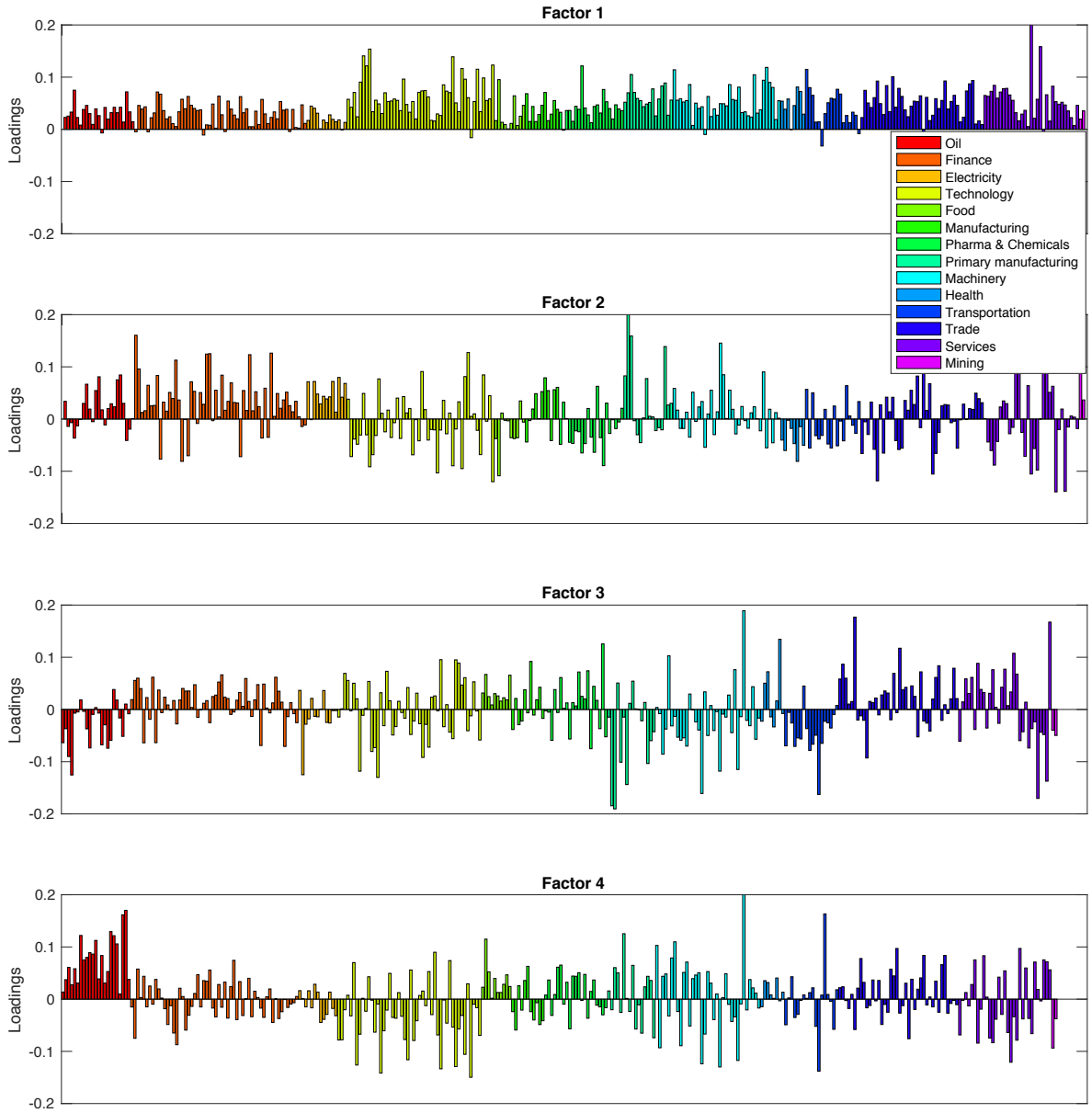

Figure 5. Portfolio weights of four monthly PCA factors. This figure shows the portfolio weights of four monthly PCA factors over the full time horizon from 2004 to 2016. Stocks are sorted according to industries.

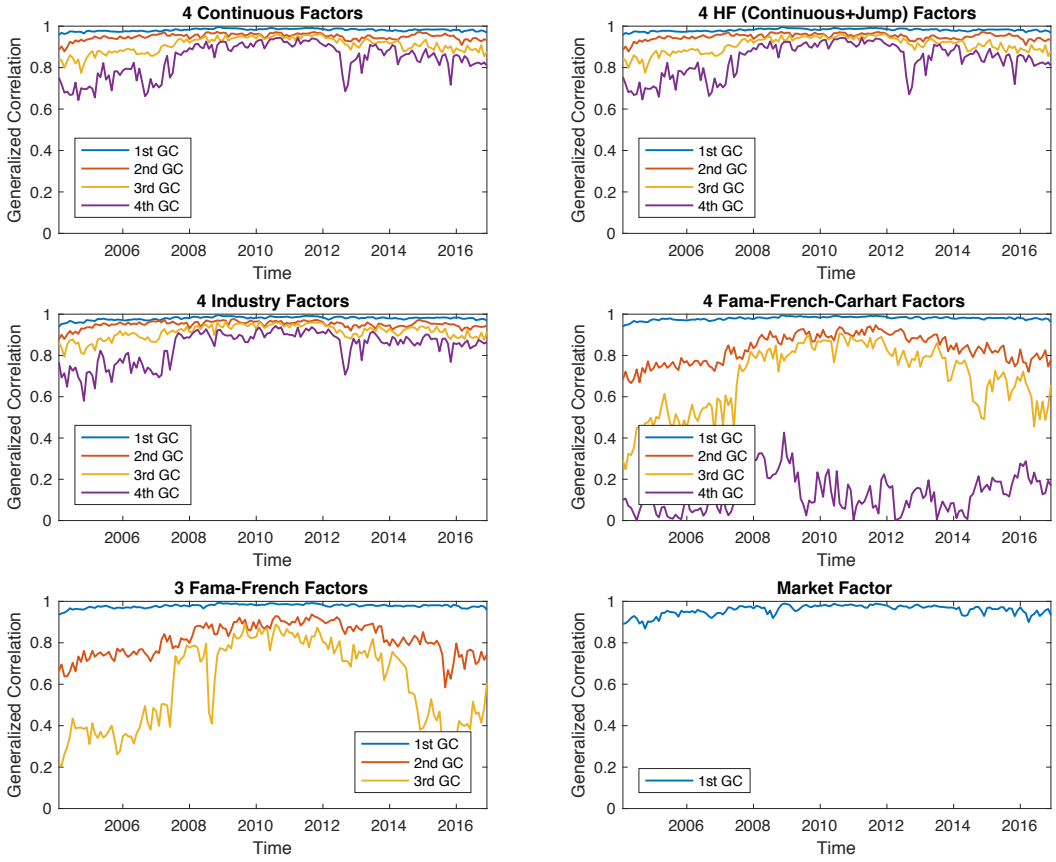

Figure 6. Time-variation in loadings. This figure plots the generalized correlations of continuous loadings estimated over the full time horizon and on a moving window of one month (21 trading days).

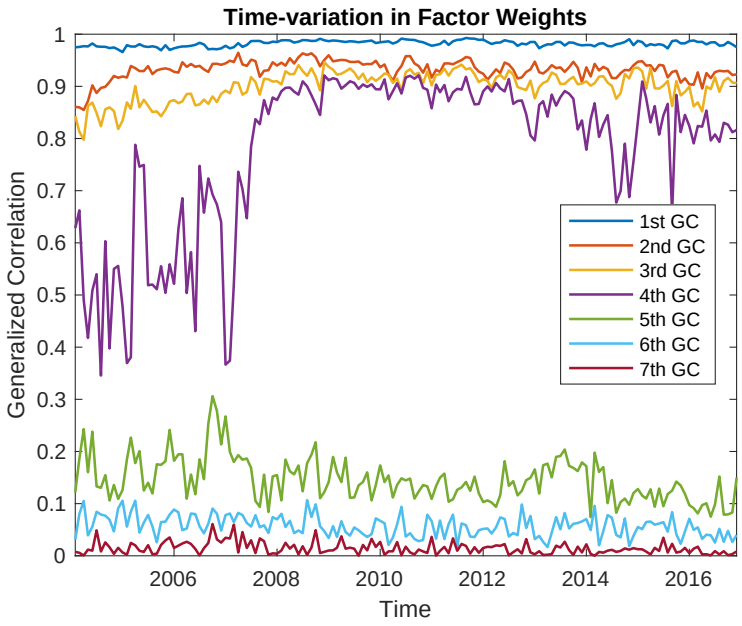

Figure 7. Time-variation in locally estimated continuous factors. This figure plots the generalized correlations of factor portfolio weights for the first seven continuous PCA factors estimated over the full time horizon and on a moving window of one month (21 trading days).

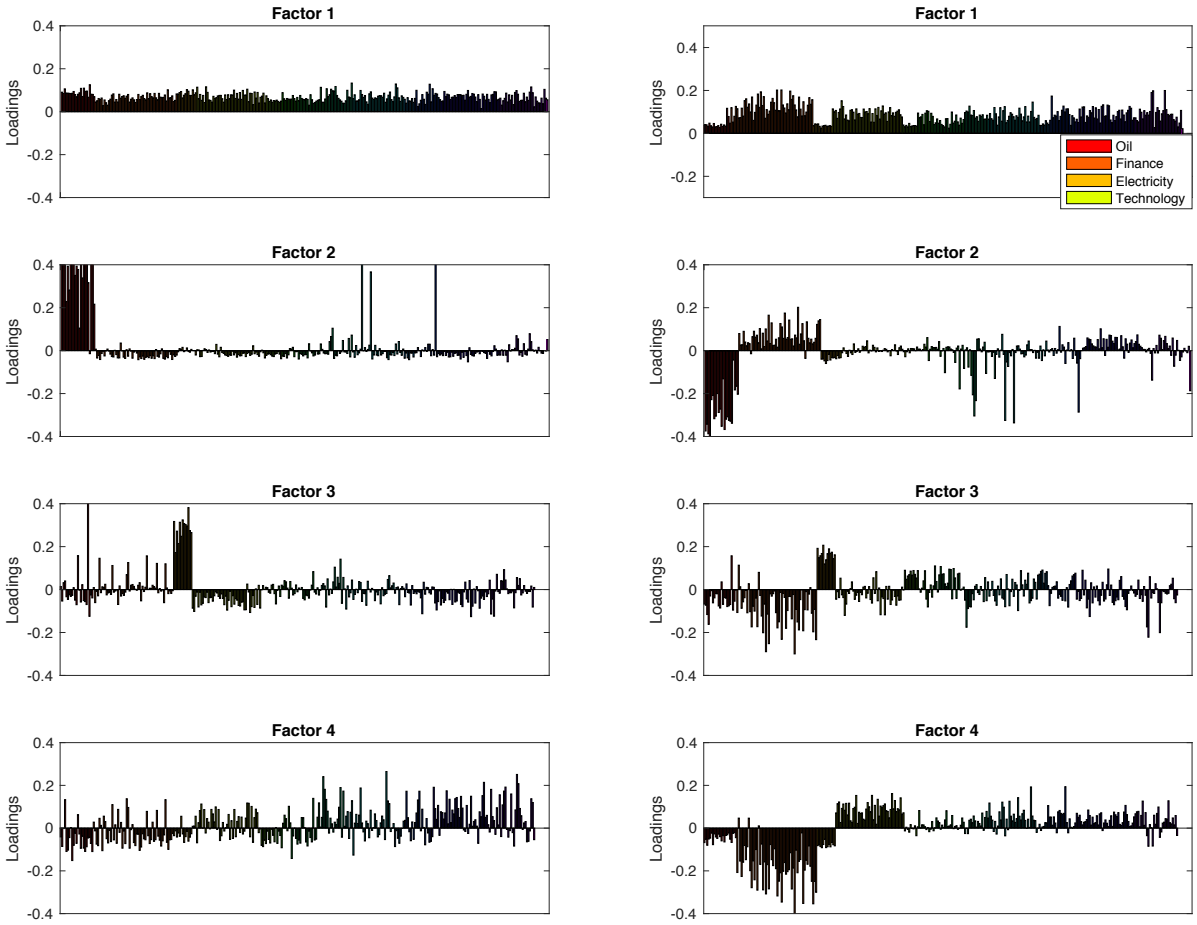

Figure 8. Time-varying portfolio weights of four continuous PCA factors. This figure plots the portfolio weights for the four continuous PCA factors in November 2006 (left) and April 2008 (right).

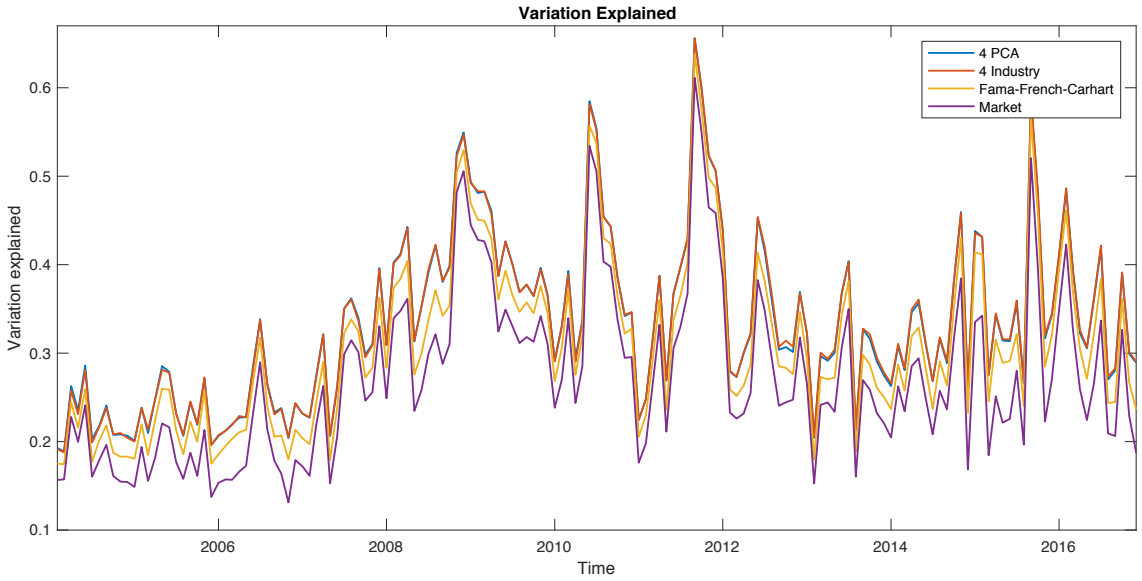

Figure 9. Time-variation in the percentage of explained variation for different factors. This figure plots the continuous variation calculated on a moving window of one month (21 trading days).

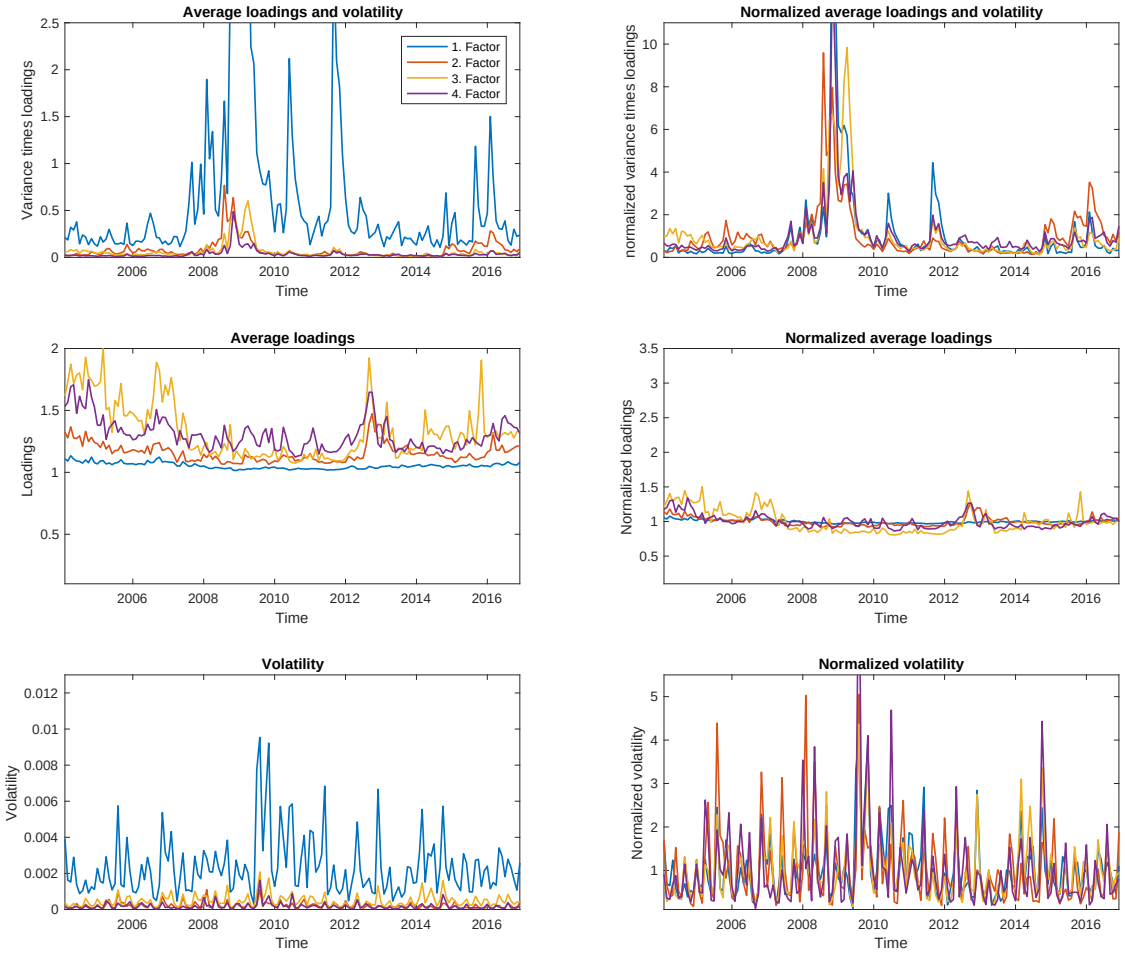

Figure 10. Decomposition of time-variation in factor structure for four continuous PCA factors. This figure plots the time-varying loadings and volatilities estimated on a moving window of one month (21 trading days) based on continuous returns. Left: Systematic impact of factors $\frac{\Lambda_k(t)^\top \Lambda_k(t)}{N}\sigma_k^2(t)$, where $\Lambda_k(t)$ is the continuous loading of factor k in month t and σ_k^2 is the continuous quadratic variation of factor k in month t, average loadings $\frac{\Lambda_k(t)^\top \Lambda_k(t)}{N}$, and volatility $\sigma_k^2(t)$. Right: The same three quantities but normalized by the time average of the quantity of interest.

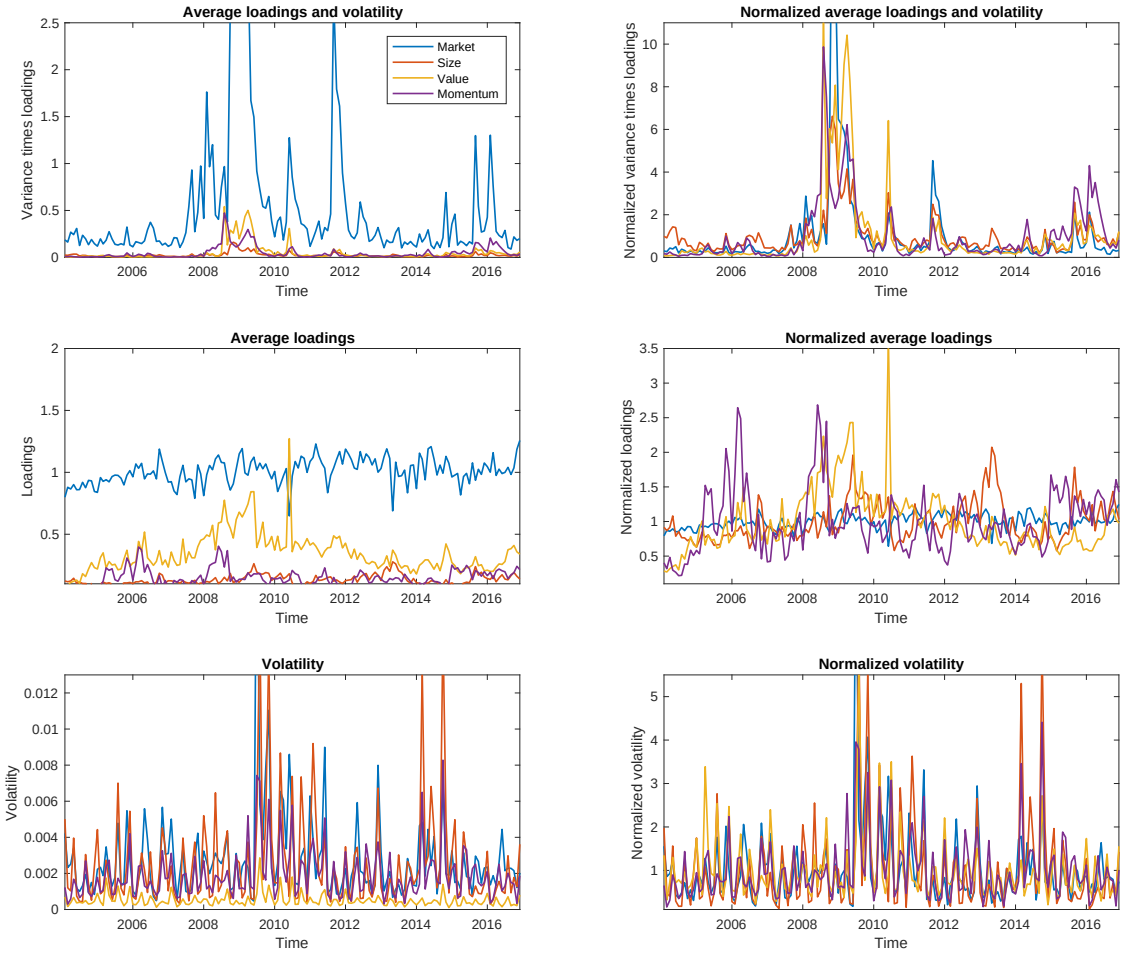

Figure 11. Decomposition of time-variation in factor structure for four continuous Fama-French Carhart factors. This figure plots the time-varying loadings and volatilities estimated on a moving window of one month (21 trading days) based on continuous returns. Left: Systematic impact of factors $\frac{\Lambda_k(t)^\top \Lambda_k(t)}{N}\sigma_k^2(t)$, where $\Lambda_k(t)$ is the continuous loading of factor k in month t and σ_k^2 is the continuous quadratic variation of factor k in month t, average loadings $\frac{\Lambda_k(t)^\top \Lambda_k(t)}{N}$, and volatility $\sigma_k^2(t)$. Right: The same three quantities but normalized by the time average of the quantity of interest.

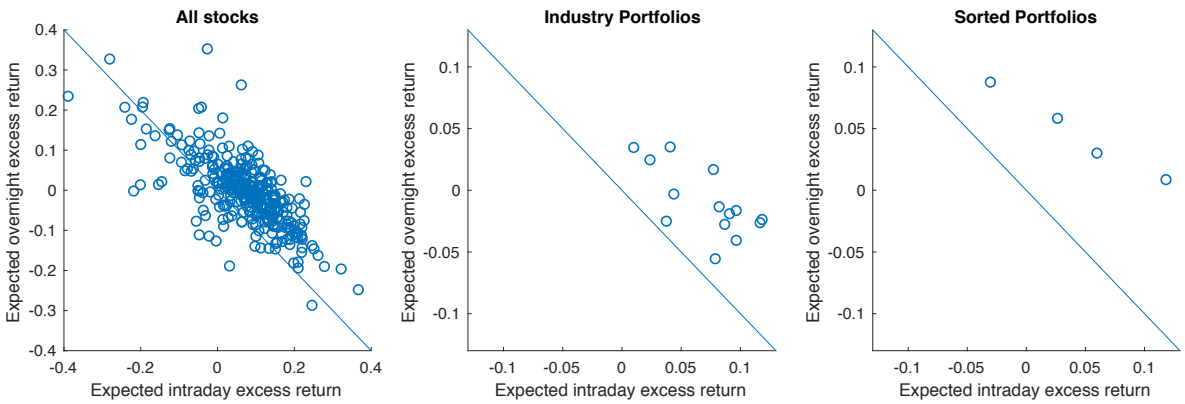

Figure 12. Expected intraday and overnight returns. This figure plots the expected intraday and overnight excess returns from 2004 to 2016 ($T = 3273$) for the balanced panel of all stocks ($N = 332$), 14 industry portfolios based on the unbalanced panel and six Fama-French size- and value-sorted portfolios (two size and three book-to-market quantiles).

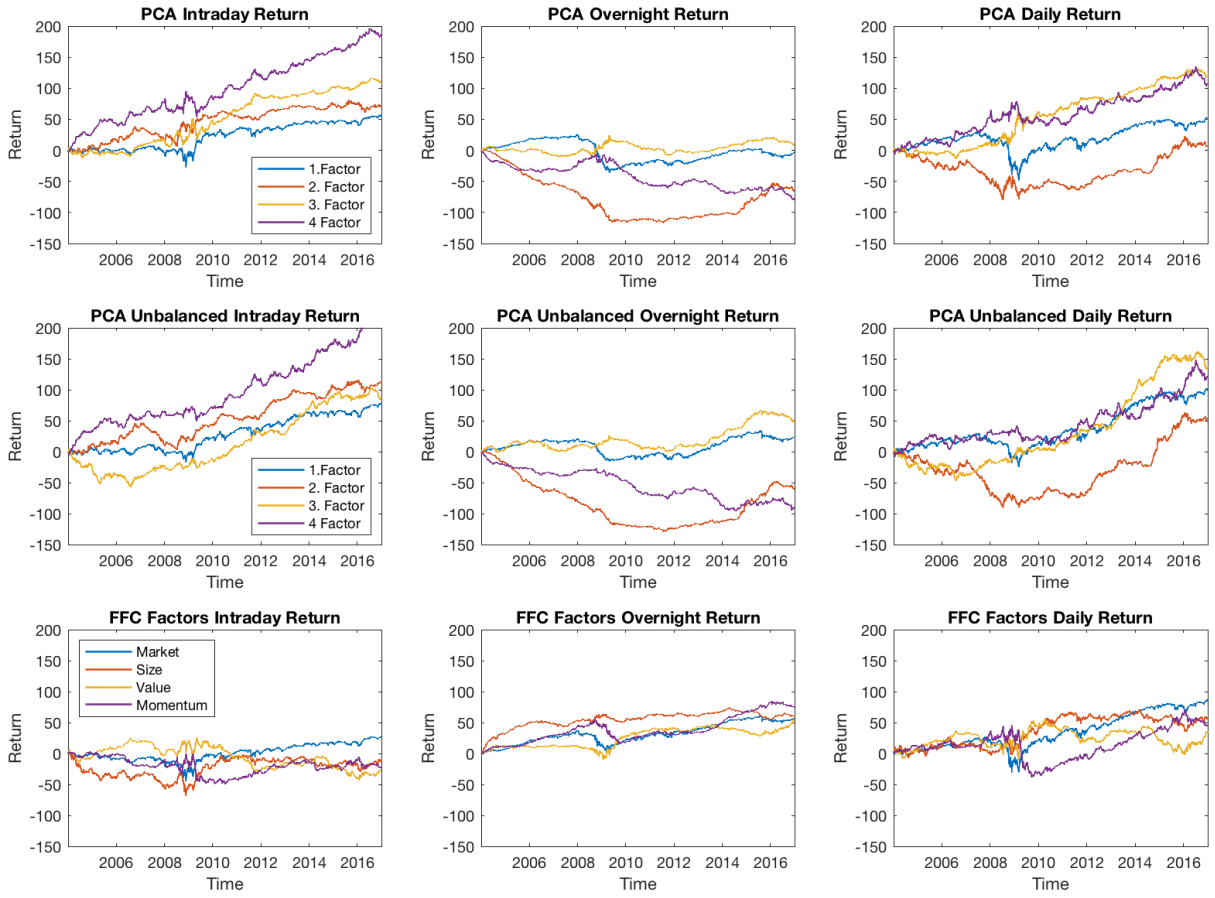

Figure 13. Normalized cumulative factor returns for intraday, overnight and daily returns. This figure plots the continuous PCA and Fama-French-Carhart factors normalized by their daily standard deviation.

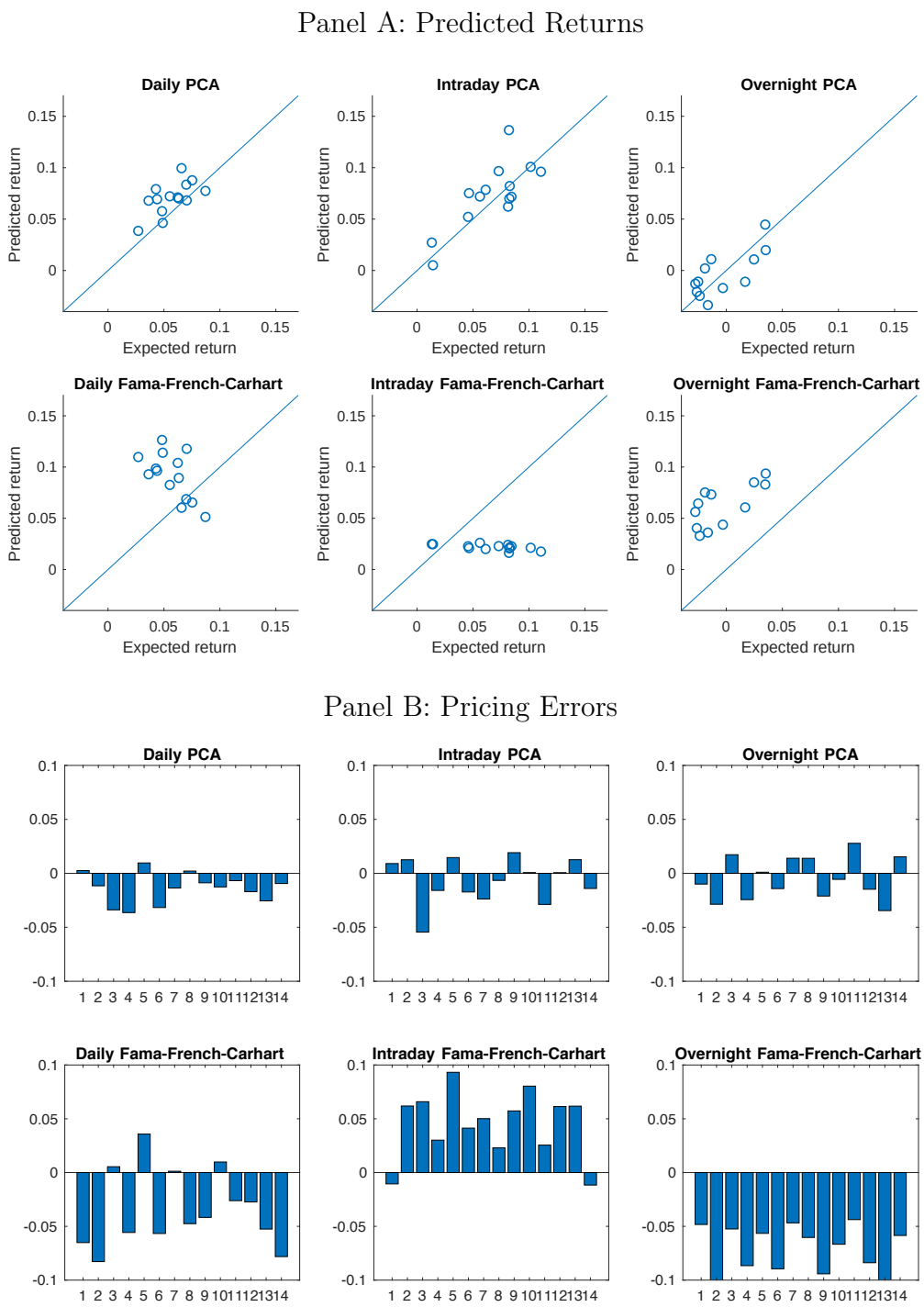

Figure 14. Asset pricing of industry portfolios. This figure depicts return predictions for 14 industry portfolios (unbalanced) based on four continuous PCA factors and the Fama-French-Carhart factors. Regressions are based on daily, intraday, or overnight data. Panel A shows the predicted and expected returns. Panel B plots the time-series pricing errors.

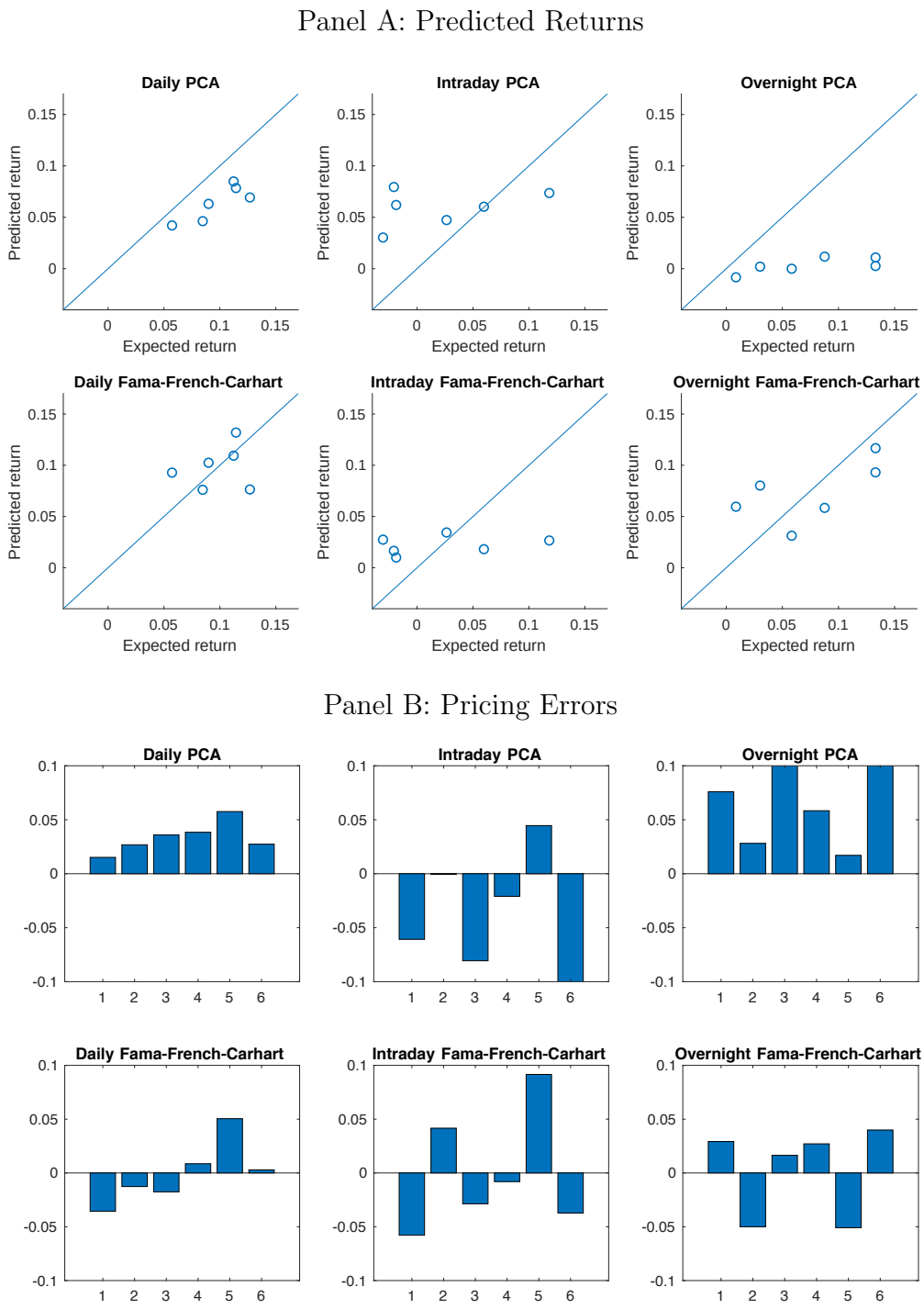

Figure 15. Asset pricing of size- and value-sorted portfolios. This figure depicts return predictions for six Fama-French size- and value-sorted portfolios (two size and three book-to-market quantiles) based on four continuous PCA factors and the Fama-French-Carhart factors. Regressions are based on daily, intraday, or overnight data. Panel A shows the predicted and expected returns. Panel B plots the time-series pricing errors.

Appendix. High-Frequency Data

The estimated high-frequency and overnight factors and stock returns are available on my website mpelger.people.stanford.edu.[37]

I collect the price data from the WRDS TAQ millisecond trades database for the period 2004 to 2016. I construct the log-prices for five-minute sampling, which gives me on average 250 days per year with 77 daily increments.

The first observation is the volume-weighted trading price in the exact second of 9:30:00. For the remaining 78 observations, the volume-weighted trading prices are calculated for each second, and the last observations in each 300 second interval are taken. For a significant portion of the stocks, there is no trade in the first seconds of the day, and thus I start my sample at 9:35am. The Internet Appendix presents all results for the sample starting at 9:30am with essentially identical findings.

I use the price of the trade at or immediately preceding each five-minute mark. For each year I take the intersection of the stocks traded each day and the stocks in the S&P 500 index over the period 1993 to 2012 based on the Bloomberg terminal. This gives me a cross-section N of 555 to 667 firms for each year with an intersection of $N = 332$ for all 13 years. 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 stamps using volume-weighted average prices.

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

- All 10 first five-minute observations are missing in any day of the year.
- There are more than 50 missing values before the first trade for each day on the year.
- There are more than 500 missing values in the year.

Missing observations are replaced by interpolated values. If no available five-minute price is available for the same day, the next available five-minute price is used. Otherwise, the last available five-minute price is used. Because my estimators are based on increments,

[37]The NYSE Daily TAQ data is property of NYSE Daily TAQ ©2019 New York Stock Exchange, LLC. Calculated based on data from Stock/Security Files©2019 Center for Research in Security Prices (CRSP), The University of Chicago Booth School of Business.

the interpolated values will result in increments of zeros, which do not contribute to the quadratic covariation.

I classify stocks into 14 different industries based on SIC codes. Table A.I lists the classifications and number of stocks based on the unbalanced and balanced panels. I construct the Fama-French-Carhart factors and characteristic-sorted portfolios based on all stocks in the WRDS TAQ millisecond trades database that satisfy the above requirements, that is, they include also the stocks that have not been part of the S&P 500 index.

Table A.I
Industry Portfolios

This table summarizes the composition of 14 industry portfolios for the unbalanced panel with an average of $N = 643$ firms per year and the balanced panel of $N = 332$ stocks (unbalanced/balanced number).

Industry	Number of firms	SIC Codes
Oil and gas	47/23	1200,1221,1222,1311,1381,1382,1389, 2870, 2911, 3533, 4922, 4923, 4932, 4924, 4922, 4923
Finance	108/56	6020, 6021, 6029, 6035, 6036, 6099, 6111, 6141, 6159, 6162, 6189, 6199, 6282, 6311, 6331, 6351, 6798, 6022, 6330, 6711, 6211, 6321, 6324, 6411, 6722, 6200, 6726, 6030, 6035, 6036, 6099, 6111, 6150, 6153, 6162, 6231, 6712, 6799
Electricity	30/13	4911, 4931, 4991
Technology	83/48	3571, 3577, 3674, 3572, 2559, 3825, 3578, 3651, 3663, 3670, 3672, 3674, 3678, 3679, 3761, 3827, 3570, 3573, 3625, 3660, 3661, 3662, 3764, 7370, 7372, 7373, 7374, 7375, 7379
Food	21/10	2000, 2015, 2026, 2033, 2067, 2033, 2075, 2082, 2084, 2095, 2096, 2032, 2041, 2043, 2066, 2099, 2111, 2086, 5461
Manufacturing	28/11	2273, 2320, 2321, 2325, 2341, 2385, 2439, 2421, 2431, 2435, 2491, 2493, 2499, 2515, 2531, 2621, 2679, 2711, 2329, 2672, 2674, 2676, 2721, 2731, 2750, 2761, 2771, 2782, 3149, 3171, 3842
Pharmaceuticals and chemicals	37/21	2834, 2841, 2842, 2844, 2851, 2879, 2830, 2813, 2869, 2812
Primary manufacturing	24/16	3011, 3089, 3199, 3221, 3275, 3312, 3317, 3324, 3334, 3357, 3411, 3423, 3429, 3334, 3021, 3316, 3331, 3341, 3350, 3351, 3432, 3491, 3492
Machinery	57/35	3511, 3519, 3519, 3519, 3533, 3533, 3550, 3559, 3561, 3562, 3579, 3621, 3633, 3711, 3714, 3714, 3714, 3721, 3721, 3721, 3728, 3810, 3822, 3825, 3826, 3829, 3873, 3942, 3944, 3812, 3523, 3523, 3531, 3532, 3534, 3546, 3563, 3565, 3569, 3594, 3641, 3713, 3724, 3731, 3751
Health	8/8	3840, 3841, 3845
Transportation	35/18	4011, 4213, 4481, 4492, 4512, 4512, 4513, 4730, 4813, 4813, 4813, 4841, 4841, 4841, 4899, 4225, 4510
Trade	68/40	5013, 5047, 5063, 5122, 5149, 5200, 5211, 5231, 5311, 5330, 5331, 5411, 5521, 5531, 5650, 5651, 5700, 5731, 5810, 5812, 5812, 5912, 5912, 5944, 5990, 5999, 5661, 5940, 5015, 5045, 5049, 5051, 5065, 5075, 5111, 5141, 5169, 5199, 5511, 5530, 5541, 5611, 5943, 5945
Services	49/31	1531,1731,4950, 7011, 7311, 7322, 7323, 7359, 7363, 7389, 7513, 7812, 7841, 7990, 7999, 8062, 8062, 8071, 8093, 8200, 8700, 8711, 8731, 8741, 3840, 4941, 4953, 4955, 7291, 7310, 7353, 7380, 7381, 7382, 7369, 7993, 7996, 8051, 8082, 8092, 8099, 8244, 8742
Mining	3/2	1422, 1423, 1442, 1011