Latent Common Return Volatility Factors: Capturing Elusive Predictive Accuracy Gains When Forecasting Volatility*

Mingmian Cheng, Norman R. Swanson and Xiye Yang

June 2017

Abstract

In this paper, we use factor-augmented HAR-type models to predict the daily integrated volatility of asset returns. Our approach is based on a proposed two-step dimension reduction procedure designed to extract latent common volatility factors from a large dimensional and high-frequency returns dataset with 267 constituents of the S&P 500 index. In the first step, we apply either LASSO or elastic net shrinkage on estimates of integrated volatility of all constituents in the dataset, in order to select a subset of asset return series for further processing. In the second step, we utilize (sparse) principal component analysis to estimate latent common asset return factors, from which latent integrated volatility factors are extracted. Although we find limited in-sample fit improvement, relative to a benchmark HAR model, all of our proposed factor-augmented models result in substantial out-of-sample predictive accuracy improvement. In particular, forecasting gains are observed at market, sector, and individual-stock levels, with the exception of the financial sector. Further investigation of the factor structures for non-financial assets shows that industrial and technology stocks are characterized by minimal exposure to financial assets, inasmuch as forecasting gains associated with factor-augmented models for these types of assets are largely attributable to the inclusion of non-financial stock price return volatility in our latent factors.

Keywords: Forecasting, Latent common volatility factor, Dimension reduction, Factor-augmented regression, High-frequency data, High-dimensional data

JEL Classification: C22, C52, C53, C58.

---

*Mingmian Cheng, Department of Economics, Rutgers University, 75 Hamilton Street, New Brunswick, NJ 08901, USA; mcheng@econ.rutgers.edu. Norman R. Swanson, Department of Economics, Rutgers University, 75 Hamilton Street, New Brunswick, NJ 08901, USA; nswanson@economics.rutgers.edu. Xiye Yang, Department of Economics, Rutgers University, 75 Hamilton Street, New Brunswick, NJ 08901, USA; xiyeyang@econ.rutgers.edu. The authors are grateful to Yacine Aït-Sahalia, Jianqing Fan, Yuan Liao, Bruce Mizrach, Markus Pelger and seminar participants at the 2017 China Meeting of the Econometric Society, the 2017 Canadian Economic Association meetings and the 23rd International Conference on Computing in Economics and Finance for comments that have been utilized in the preparation of this paper.

1 Introduction

Accurate price volatility estimation and prediction is crucial to successful risk management and asset allocation. In light of this fact, many new estimators relevant for volatility analysis have recently been introduced, including but not limited to, realized variance (RV) (Andersen et al. (2001)), jump robust RV based on multi-power variation and truncation (Barndorff-Nielsen and Shephard (2004), Mancini (2009), Corsi et al. (2010), Podolskij and Ziggel (2010)), and multi-scale (Aït-Sahalia et al. (2011)) and pre-averaging (Jacod et al. (2009)) estimators, which are designed to eliminate microstructure effects. Making use of these sorts of integrated volatility estimators, heterogeneous autoregressive (HAR) type forecasting models have been studied extensively in the financial econometrics literature. For example, Corsi (2009) introduces a basic HAR-RV model, and Andersen et al. (2007) and Corsi et al. (2010) analyze jump variation augmented HAR-RV models. Additionally, Duong and Swanson (2015) examine HAR model performance when so-called upside and downside jump variations are included. The authors also utilize q-th order variations of jump components, with $0.1 \leq q \leq 6$, and consider the usefulness of large jumps (i.e., jump size exceeds a given threshold) in HAR-RV type regressions. Patton and Sheppard (2015) study how the positive and negative price jumps affect the future volatility, respectively. Audrino and Hu (2016) exploit the prevalence of the leverage effect and investigate the characteristics of different components of continuous risks and jump risks on volatility persistence. Bollerslev et al. (2016) further improve volatility forecasting by allowing for the change of model coefficients according to the degree of measurement error. Other types of volatility forecasting model are also widely used, such as stochastic volatility (SV) models (Meddahi et al. (2001), Andersen et al. (2004), Andersen et al. (2011)), (G)ARCH-type models (Andersen et al. (2003), Hansen and Lunde (2005), Brandt and Jones (2006)), and Mixed Data Sampling (MIDAS) models (Ghysels et al. (2006), Ghysels and Sinko (2011)).

Although very parsimonious, the HAR-type models discussed in the above papers only utilize information on the target asset that is being predicted. A little explored question is whether there are sources of information other than the target asset itself can help improve the predictive accuracy in HAR regressions. In this paper, we attempt to answer this question by augmenting benchmark HAR models with estimates of latent integrated volatility (IV) factors extracted from latent common asset return factors, which are themselves extracted from a large dimensional and high-frequency asset returns dataset, and investigating whether these latent IV estimates are informative about the future volatility of selected target assets. As shall be discussed below, inclusion of latent IV factors substantially improve volatility forecasting performance for various assets at market, sector and individual-stock levels, with the notable exception of the financial sector.

The dimension reduction approach that we use in order to estimate factors combines several cutting-edge methods widely used in the literature. In particular, the motivation for our two-step dimension reduction procedure is based on new results on the use of principal component analysis (PCA) in the construction of latent factors using large dimensional and high-frequency asset return datasets that are developed in Aït-Sahalia and Xiu (2016a) and Aït-Sahalia and Xiu (2016b). However, in addition to focusing on PCA, our procedure attempts to take account of the fact that we are interested in targeted or individual market, sector or stock return prediction. Such targeted prediction is potentially inconsistent with the direct use of principal component analysis (PCA) for the extraction of common factors, since the common factors estimated using PCA that are used in prediction models are usually those associated with the largest eigenvalues in an eigenvalue-eigenvector decomposition of the correlation matrix of the dataset being examined (e.g., see Stock and Watson (2002a,b, 2006), Bai and Ng (2006a,b, 2008), and the references cited therein). Namely, the factors that account for the largest share of the variability of the covariance (correlation) matrix are assumed to be the best candidate predictors for a given target variable. Clearly, this may not always be the case, as discussed in Bai and Ng (2008), Carrasco and Rossi (2016), and Swanson and Xiong (2017). To address this problem, we begin, in a first step, by selecting a subset of assets from the total asset pool. This is done by carrying out shrinkage of the set of all integrated volatility estimates for the asset return variables in our dataset. Shrinkage is done using the least absolute shrinkage operator (LASSO) or the elastic net. Then, in a second step, we estimate latent asset return factors by applying either PCA or sparse PCA (SPCA) to the selected subset of asset return variables corresponding to the integrated volatility variables selected in our first step. Finally, these latent asset return factors are used to construct latent integrated volatility factors, which are in turn used as explanatory variables in our HAR-type regression model prediction experiments.

One important aspect of our investigation is our novel use of SPCA. While PCA is well known, sparse principal component analysis (SPCA) is relatively new to the field, as discussed in Kim and Swanson (2017). Intuitively, SPCA can be viewed as a form of "double" shrinkage (see Zou et al. (2006) and Qi et al. (2013)). More specifically, while PCA can be interpreted as penalized regression with an L-2 penalty (akin to the penalty used in ridge regression), SPCA can be interpreted as penalized regression with either an L-1 norm penalty (i.e., a LASSO variant of PCA), or a combined L-1 and L-2 norm penalty (i.e., an elastic net variant of PCA). In both cases, sparseness is imposed on the factor loadings, with a regularization parameter controlling the degree of sparseness. In our setup, thus, sparseness is first imposed in our variable selection step (i.e., in our first step, where the lasso and elastic net are used to analyze integrated volatility variables), and then again imposed in our latent factor construction step (i.e., in our second step, where PCA and SPCA are used to analyze high frequency asset returns). Broadly speaking, the first step of our approach follows, and builds on, methods developed in in Bai and Ng (2008) in which "targeted predictors" are selected before the estimation of common factors. Again broadly speaking, our second step follows, and builds on, methods developed in Aït-Sahalia and Xiu (2016a) and Aït-Sahalia and Xiu (2016b), in which latent integrated volatility variables are constructed using PCA.

Our dataset consists of intra-day observations on 267 constituents of the S&P 500 index, 9 sector ETFs, and one market EFT (i.e., SPY, which is the SPDR S&P 500 ETF). Data were analyzed for the sample period from January 3, 2006 to December 31, 2010, and were collected from the TAQ database. We report the results based on prediction of SPY, 9 sector ETFs, and 11 individual stocks, for the period of July 1, 2009 to December 31, 2010. We also report the in-sample fit of various forecasting models, common factor estimators, and data aggregation permutations. Our key findings are summarized below, and explained in detail in a later section of the paper.

First, *in-sample* fit is surprisingly stable across different models, including our benchmark HAR model and our volatility-factor augmented models, across three different data frequencies, including 1-minute, 5-minute, and 10-minute frequencies. Thus, there is little to choose between data frequencies when comparing in-sample model fit. Moreover, in-sample model fit is surprisingly similar across different asset classes (i.e., market index, sector ETFs, and individual stocks), with most R^2 values ranging rather tightly between 0.35 and 0.55.

Second, our in-sample findings are highly mis-leading, when the objective of interest is *out-*

of-sample volatility prediction. Namely, all of the above findings become irrelevant when ex ante prediction experiments are carried out. In particular, for forecasting, data frequency is crucial, and the “best” frequency varies across different assets and asset classes. However, we still recommend using the 5-minute frequency, as a general rule-of-thumb. This is because our factor augmented HAR models generally yield the “best” predictions (see below for further discussion) using 5-minute frequency data, when comparing results factor augmented model predictive accuracy across different frequencies. Intuitively, note that on one hand, using higher frequency data may result in a substantial amount of microstructure noise being absorbed by extracted factors, hence potentially deteriorating predictive performance. On the other hand, if the sampling frequency is relatively low, it is more difficult to eliminate individual jumps when estimating latent factors, leading to forecast deterioration.

The above argument is buttressed by our finding that models utilizing SPCA in factor construction generally forecast “better” than those utilizing PCA. Moreover, the performance of SPCA, relative to PCA, is greatest when one moves from using 10-minute to 5-minute frequency data, as well as when one moves from using 1-minute to 5-minute frequency data.

Third, and perhaps most importantly, predictive accuracy improves appreciably when latent common volatility factors are included in benchmark HAR-type models. For example, for Johnson & Johnson (see Table 15), the benchmark model using 5-minute frequency data achieves an out-of-sample R^2 value of only 0.14. This is approximately one-third of the out-of-sample R^2 value associated with our “best” factor-augmented model. This pattern occurs for many firms and sectors; as well as for the market ETF. Interestingly, if only in-sample R^2 values were examined in order to assess the usefulness of common factors, then the story would change markedly. For example, again using Johnson & Johnson to illustrate our findings, the benchmark model using 5-minute frequency data (without a common factor) achieves an in-sample R^2 value of 0.39, while in-sample R^2 values for our factor-augmented models are all between 0.43 and 0.48. This small increase associated with utilizing common factors in an in-sample context characterizes all of our experiments. Indeed, substantial increases in performance only arise when using latent factors for ex ante prediction. This finding constitutes strong evidence of an important difference between findings based on in- and out-of-sample experiments.

A different way to interpret the above key finding is as follows. In-sample R^2 values are widely known to be substantively greater than out of sample R^2 values in financial forecasting applications. This feature has been extensively discussed in the literature, and reasons for it range from the presence of (smooth) structural breaks and state transitions, to the general inability of linear models to capture inherently nonlinear interactions among financial variables and markets (e.g., see Paye and Timmermann (2006), Aiolfi et al. (2009), and Ang and Timmermann (2012)). In our experiments, when comparing benchmark HAR models, in-sample R^2 values are indeed much greater than their out-of-sample benchmark HAR counterparts, as might be expected. For example, using IBM (see the 5-minute panel in Table 14) to illustrate our findings, the benchmark model (without a common factor) achieves an in-sample R^2 value of 0.61, as opposed to an out-of-sample R^2 value of 0.24. However, when the "best" factor augmented in-sample and out-of sample performances are compared in this example, the R^2 values are 0.65 and 0.38, respectively. Thus, the relative out-of-sample gains associated with utilizing latent volatility factors are greater than the in-sample gains. This feature characterizes our results at all market, sector, and individual-stock levels, although it is more starkly apparent at the individual stock level.

Fourth, there is an important wrinkle to the above story. Namely, for financial assets, out-of sample R^2 values are approximately 0 in some cases. A particularly interesting example of this is the financial sector ETF. For this ETF, in-sample R^2 values range from around 0.53 to 0.64, while out-of-sample R^2 range from around 0.08 to 0.30. At the individual stock level, the picture is even more stark. Consider Goldman Sachs (see Table 13). In-sample R^2 values are always around 0.40, while out-of-sample R^2 values are always less than 0. However, all is not lost. As discussed above, for many of our target variables, there is substantial predictable content. For example, out-of-sample R^2 values for Coca-Cola (see Table 17), Exxon Mobil (see Table 22) and IBM (see Table 14) range from 0.35 to 0.41, from 0.30 to 0.37, and from 0.23 to 0.38, respectively, when using common factors constructed via our two-step procedure, and based on IV estimators constructed using 5-minute frequency data.

Fifth, financial stocks are frequently selected in our first variable selection (or shrinkage) step. However, they are often assigned small weights in the second step (i.e., the latent factor estimation step), particularly when SPCA is used in this step. For instance, when we forecast the volatility of our energy sector ETF using 1-minute frequency data, over 33% of the most frequently selected stocks in the first step are in financial sector. However, the average weight assigned by PCA to, for instance, Goldman Sachs is only around 0.09, while the corresponding weight assigned to Texas Instruments is around double that (see Table 24). Even more starkly, the average weight assigned by SPCA to Goldman Sachs drops is only around 0.02. This is in part due to the fact that over 50% of weights assigned by SPCA are identically zero. On the contrary, the average weight on Texas Instruments Incorporated rises to 0.19. Therefore, we conjecture that the contribution of financial stocks to common volatility factors may be less than that of stocks in other sectors, based on these rather surprising findings. Moreover, and as a result of the above findings, it is very likely that the marginal predictive content of common volatility factors is largely accounted for by information in sectors other than the financial sector, such as the industrial and technology sectors.

The rest of the paper is organized as follows. Section 2 outlines our setup and modeling assumptions, and includes a brief discussion of some of the realized measures that we construct. Section 3 discusses the forecasting framework used, and briefly introduces PCA, SPCA, LASSO and elastic net methods. Section 4 includes a discussion of the data used in our forecasting experiments, and summarizes our key empirical findings. Finally, Section 5 contains concluding remarks.

2 Setup

Denote by X the d-dimensional log-price process of d assets. Following the high-frequency literature, we assume that X follows an Itô-semimartingale defined on some filtered probability space $(\Omega, \mathbb{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, and has the following representation:

$$\begin{aligned} X_t = X_0 &+ \int_0^t b_s ds + \int_0^t \sigma_s dW_s \\ &+ \int_0^t \int_{\{|x|\leq\epsilon\}} x(\mu-\nu)(ds,dx) + \int_0^t \int_{\{|x|\geq\epsilon\}} x\mu(ds,dx), \end{aligned} \tag{1}$$

where b_t is the instantaneous drift term, σ_t is the spot volatility, and both are adapted and càdlàg. Additionally, W_t is a multidimensional standard Brownian motion, μ is a Poisson random measure with compensator ν, and $\epsilon > 0$ is an arbitrary number. For more details on Itô-semimartingale and continuous-time asset price modeling, see Aït-Sahalia and Jacod (2014) and the references therein.

Since volatility is unobservable, realized measures are often employed to consistently estimate it on a fixed interval $[0, T]$, using high-frequency intraday data. For instance, one of the most widely known measures, realized volatility, is defined as follows:

$$\mathrm{RV}_t = \sum_{i=1}^{\lfloor t/\Delta_n \rfloor} (\Delta_i^n X)^2, \quad \forall t \in [0, T], \tag{2}$$

where $\lfloor m \rfloor$ is the integer part of m and $\Delta_i^n X = X_{i\Delta_n} - X_{(i-1)\Delta_n}$, where Δ_n is the equally-spaced sampling interval that shrinks to zero. It is well-known that when asset prices are continuous on a fixed interval $[0, T]$, we have that:

$$\sum_{i=1}^{\lfloor t/\Delta_n \rfloor} (\Delta_i^n X)^2 \xrightarrow{\mathbb{P}} \int_0^t \sigma_s^2 ds, , \quad \forall t \in [0, T]. \tag{3}$$

However, when asset prices are discontinuous on $[0, T]$:

$$\sum_{i=1}^{\lfloor t/\Delta_n \rfloor} (\Delta_i^n X)^2 \xrightarrow{\mathbb{P}} \int_0^t \sigma_s^2 ds + \sum_{0 \le s \le t} (\Delta X_s)^2, \quad \forall t \in [0, T]. \tag{4}$$

where $\Delta X_s := X_s - X_{s-} \neq 0$, if and only if X jumps at time s.

To separate the integrated volatility from jump variation, one can use the threshold technique developed in Mancini (2001, 2009):

$$\sum_{i=1}^{\lfloor t/\Delta_n \rfloor} (\Delta_i^n X)^2 \mathbf{1}_{\{|\Delta_i^n X| \le \alpha \Delta_n^{\varpi}\}} \xrightarrow{\mathbb{P}} \int_0^t \sigma_s^2 ds, \tag{5}$$

or use the multipower variation (MPV) estimator developed in Barndorff-Nielsen and Shephard (2004) and Barndorff-Nielsen et al. (2006):

$$\Delta_n^{1-p^+/2} \sum_{i=1}^{\lfloor t/\Delta_n \rfloor - k+1} |\Delta_i^n X|^{p_1} ... |\Delta_{i+k-1}^n X|^{p_k} \xrightarrow{\mathbb{P}} m_{p_1} ... m_{p_k} \int_0^t |\sigma_s|^{p^+} ds \tag{6}$$

where $p_j \geq 0$, $p^+ = p_1 + \cdots + p_k$ and $m_p = \mathbb{E}[|\mathcal{N}(0, 1)|^p]$. One can also combine these two methods and use a truncated multipower variation estimator. Apparently, different components of the quadratic variation can be analyzed or used separately in econometric analysis.

We also assume that the continuous part of asset log-prices follows an underlying continuous-time factor model on $[0, T]$. Namely, define:

$$Y_t = \Lambda_t F_t + Z_t \tag{7}$$

where $Y_t := X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s$ is the continuous part of X, F_t is an r-dimensional continuous factor ($r < d$), Z_t is an idiosyncratic component, and Λ_t is a d-by-r factor loading matrix, each element of which is adapted and has càdlàg paths almost surely. Here, we specifically call F_t the common price factor in order to distinguish it from the common volatility factor defined later. The common price factor F's and the idiosyncratic component Z's are continuous Itô semimartingales as well, with:

$$F_t = F_0 + \int_0^t h_s ds + \int_0^t \eta_s dB_s \tag{8}$$

and

$$Z_t = Z_0 + \int_0^t g_s ds + \int_0^t \gamma_s d\tilde{B}_s, \tag{9}$$

where B_s and $\tilde{B}_s$ are independent Brownian motions. All of the coefficient processes, h, η, g and γ are adapted to $(\mathcal{F}_t)_{t\geq 0}$ and have càdlàg paths, almost surely. The above factor models and general settings follow Aït-Sahalia and Xiu (2016b).

3 Dimension Reduction and Forecasting Methods

The original HAR model is given below.

$$\mathrm{RM}_{t+h} = \beta_0 + \beta_1 \mathrm{RM}_t + \beta_2 \mathrm{RM}_{[t,t-4]} + \beta_3 \mathrm{RM}_{[t,t-21]} + \epsilon_t, \tag{10}$$

where RM's are realized measures of integrated volatility, and $\mathrm{RM}_{[t,t-p]}$ is the average of RM's over the most recent $p+1$ days. For instance, if realized volatility is used in the model, then:

$$\mathrm{RV}_{[t,t-p]} = \frac{1}{p+1} \sum_{i=0}^{p} \mathrm{RV}_{t-i}. \tag{11}$$

To eliminate the jump variation from the total quadratic variation, we employ the truncated realized volatility in (5) to consistently estimate the integrated volatility[1]. Therefore, the benchmark model that we consider in this paper is as follows:

$$\mathrm{TRV}_{t+h} = \beta_0 + \beta_1 \mathrm{TRV}_t + \beta_2 \mathrm{TRV}_{[t,t-4]} + \beta_3 \mathrm{TRV}_{[t,t-21]} + \epsilon_t, \tag{12}$$

where TRV stands for truncated realized volatility.

We propose using the following factor-augmented model in our forecasting experiments,

$$y_{t+h} = \beta_0 + \beta_\Psi^\intercal \Psi_t + \beta_w^\intercal w_t + \varepsilon_t, \tag{13}$$

where y_{t+h} is the h-step-ahead forecast of daily integrated volatility. We focus on one-day-ahead forecasts (i.e., $h = 1$). Here, w_t is a vector consisting of truncated realized volatility on day t, the weekly average of truncated realized volatility from days $t-4$ to t, and the monthly average of truncated realized volatility from days $t-21$ to t (i.e., w_t contains all predictors in the benchmark model). Furthermore, Ψ_t consists of r-dimensional unobservable predictors. Based on the structure of factors assumed in (7), we define

$$\Psi_t := \int_0^t \mathrm{diag}(\Lambda_s \eta_s \eta_s^\intercal \Lambda_s^\intercal) ds$$

and name it the common volatility factor. Note that we can not disentangle Λ from η unless imposing certain identification condition such as $\eta\eta^\intercal = I_r$. So we don't distinguish them and treat Ψ_t as the integrated volatility matrix of the r uncorrelated common factors.

Here, common price factors are extracted using PCA or SPCA applied to a high-frequency dataset, the constituent members of which are specified using LASSO or elastic net shrinkage on our 274 variable original dataset. Intuitively, common price factors in (7) can be interpreted as "composite stocks" (the name comes from the fact that they are linear combinations of all individual stocks in the data set) that in general affect a majority of stocks in the market. Therefore,

[1]We actually combine the two methods, i.e. (5) and (6), in the following way: we first use bipower variation to get an initial consistent estimate of the integrated volatility, and then use this to determine an initial choice for α. Then, we obtain a second estimate of the integrated volatility using the truncation method, and a second choice of α. We iterate this procedure until the estimated integrated volatility converges.

we first construct those "composite stocks", next estimate the integrated volatility for each, and finally use the estimated integrated volatilities as predictors in (13) to forecast the integrated volatility of the target asset. Of note is that unlike many other applications of factor-augmented regressions, we do not directly use common factors $\Lambda_t F_t$ extracted from a large number of assets. Instead, what we actually use as predictors in forecasting models are the estimated integrated volatilities of these common factors, i.e. Ψ_t. As discussed above, we use PCA and SPCA when constructing "composite stocks" in this paper. These dimension reduction methods will be briefly discussed after we summarize the shrinkage methods utilized in the first step of our two step volatility factor extraction procedure.

3.1 LASSO and Elastic Net

Prior to construction of latent factors using PCA and SPCA, we first select targeted predictor assets. For this, we use two shrinkage or variable selection methods, including the LASSO (see Tibshirani (1996)) and the elastic net (see Zou and Hastie (2005)). Both techniques can be interpreted as regularized or penalized regression methods. Briefly, let RSS be the sum of squared residuals from a regression of y_{t+h} on w_t and χ_t, where χ_t is a vector of estimates of integrated volatility on day t for all assets in X_t. The LASSO estimator is the solution to:

$$\min_{\phi} \ \text{RSS} + \lambda \sum_j |\phi_j|, \tag{14}$$

where the ϕ's are coefficients in the regression. Only assets with nonzero ϕ's are retained in our final set of selected target predictor assets, say $\tilde{X}_t$, and the sparsity (number of variables) in $\tilde{X}_t$ only depends on λ. Therefore, instead of X_t, we actually apply PCA or SPCA to the variance-covariance matrix of $\tilde{X}_t$ when constructing estimates of latent asset return factors that are in turn used to construct latent volatility factors.

Similarly, the elastic net estimator is the solution to:

$$\min_{\phi} \ \text{RSS} + \lambda \sum_j \Big(\frac{(1-\alpha)}{2} \phi_j^2 + \alpha |\phi_j| \Big), \tag{15}$$

with $\alpha \in [0, 1]$. Of note is that when $\alpha = 1$, the elastic net is equivalent to LASSO. As α shrinks toward 0, the elastic net approaches ridge regression. In our experiments, we set α = 0.2 and 0.6 for the elastic net. For both the LASSO and the elastic net, λ is selected using 10-fold cross validation in the training dataset used to calibrate prediction models.

Note that for any two different target assets, the information pool ($\tilde{X}_t$) from which we construct the $\hat{F}_t$'s and subsequently the $\hat{\Psi}_t$'s can be quite different (though the probability of them being equivalent is still positive). Additionally, note that after selecting $\tilde{X}_t$ via LASSO or elastic net shrinkage targeted to a specific asset, we construct (sparsely loaded) latent factors that are specifically related to the asset of interest. Therefore, it is reasonable to assume that their integrated volatilities (i.e., the $\hat{\Psi}_t$'s) will potentially have better predictive power for the volatility of the target asset, than were the entire dataset, X_t used to construct latent factors.

3.2 Principal Component Analysis

On any fixed interval $[0, T]$, define the following covariance matrix estimator:

$$\hat{\Sigma} = \frac{1}{t} \sum_{i=1}^{\lfloor t/\Delta_n \rfloor} \{(\Delta_i^n X)(\Delta_i^n X)^\intercal\} \mathbf{1}_{\{|\Delta_i^n X| \leq \alpha \Delta_n^{\varpi}\}}, \quad \forall t \in [0, T]. \tag{16}$$

Applying an eigenvalue-eigenvector decomposition to $\hat{\Sigma}$ yields estimates of eigenvalues in descending order, $\hat{\lambda}_1 > \hat{\lambda}_2 > \cdots > \hat{\lambda}_r$, and estimates of corresponding eigenvectors, $\hat{\xi}_1$, $\hat{\xi}_2, \cdots, \hat{\xi}_r$. Therefore, the first r principal components on day t can be estimated as follows:

$$\begin{aligned}
\Delta_i^n \hat{F}_{1,t} &= (\Delta_i^n X_t) \mathbf{1}_{\{|\Delta_i^n X_t| \leq \alpha \Delta_n^{\varpi}\}} \, \hat{\xi}_1 \\
\Delta_i^n \hat{F}_{2,t} &= (\Delta_i^n X_t) \mathbf{1}_{\{|\Delta_i^n X_t| \leq \alpha \Delta_n^{\varpi}\}} \, \hat{\xi}_2 \\
&\cdots\cdots\cdots\cdots\cdots\cdots \\
\Delta_i^n \hat{F}_{r,t} &= (\Delta_i^n X_t) \mathbf{1}_{\{|\Delta_i^n X_t| \leq \alpha \Delta_n^{\varpi}\}} \, \hat{\xi}_r
\end{aligned} \tag{17}$$

With these estimated principal components, latent common volatility factors on day t can be subsequently estimated as follows:

$$
\begin{aligned}
\hat{\Psi}_{1,t} &= \sum_{i=1}^{\lfloor t/\Delta_n \rfloor} (\Delta_i^n \hat{F}_{1,t})^2 \\
\hat{\Psi}_{2,t} &= \sum_{i=1}^{\lfloor t/\Delta_n \rfloor} (\Delta_i^n \hat{F}_{2,t})^2 \\
&\cdots\cdots\cdots\cdots\cdots\cdots \\
\hat{\Psi}_{r,t} &= \sum_{i=1}^{\lfloor t/\Delta_n \rfloor} (\Delta_i^n \hat{F}_{r,t})^2.
\end{aligned}
\tag{18}
$$

Aït-Sahalia and Xiu (2016b) show that the number of common factors can be consistently estimated, and that $\sum_{j=1}^{\hat{r}} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^{\intercal}$, where $\hat{r}$ is the estimate of the number of common factors, converges to $\Lambda[\frac{1}{t}\int_0^t (\eta_s \eta_s^{\intercal}) ds]\Lambda^{\intercal}$, with dimension diverging to infinity. As a result, Ψ_t can be consistently estimated by the diagonal elements of $\sum_{j=1}^{\hat{r}} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^{\intercal}$. Once we have the estimates of Ψ_t (i.e., $\hat{\Psi}_t$), we can plug them into model (13) to forecast y. In both the case of PCA and SPCA, the latent integrated volatility variables used in the HAR regressions discussed below are estimated using high frequency uncorrelated latent asset return factor data.

We conclude this subsection with two remarks.

First, the above PCA procedure delivers the eigens (eigenvalues and eigenvectors) of the integrated volatility matrix. According to Aït-Sahalia and Xiu (2016a), these eigens are different from the integrated eigens of the spot volatility matrix, when t does not shrink to zero. However, in finite samples (e.g., in our empirical application), the time horizon t (one day) is small relative to Δ_n (1-minute, 5-minute, and 10-minute), and this difference is small compared with other sources of estimation error. Therefore, we do not address eigens of integrated volatility versus integrated eigens of spot volatility differences in our empirical application, following the approach taken by Aït-Sahalia and Xiu (2016b).

Second, it is well-known that eigens are nonlinear functions of the corresponding data matrix. Jacod and Rosenbaum (2013) show that various bias terms arise when estimating integrals of nonlinear functions of the spot volatility matrix, although only one bias term remains when local window sizes that are used are chosen to be relatively small. They further demonstrate that this remaining bias can be consistently estimated. Hence, it is possible to construct bias-corrected estimators. Moreover, according to Aït-Sahalia and Xiu (2016a), these bias terms are proportional to their associated eigens. Consequently, they share the same source of predictive power as eigens. In addition, analogous to our earlier arguments, the ratio t/Δ_n is small in our empirical application, making the bias term that can be treated using the methods of Jacod and Rosenbaum (2013), which is an integral over $[0, t]$, relatively small compared with other estimation errors. In view of these observations, we don't remove this bias term in our empirical application.

3.3 Sparse Principal Component Analysis

In general, PCA yields nonzero factor loadings for (almost) all variables, which exacerbates difficulty in interpretation. To avoid this drawback of PCA, and to generally induce parsimony, we also consider estimating factors using SPCA, as developed by Zou et al. (2006) and Qi et al. (2013).

Let $\hat{\Sigma}$ be the same covariance matrix estimator defined in (16). The eigenvector $\hat{\xi}_1$ of the first sparse principal component is the solution to:

$$\max_{||\xi_1||_2=1} \frac{\xi_1^\intercal \hat{\Sigma} \xi_1}{||\xi_1||_{\lambda_1}^2}, \tag{19}$$

where $||\cdot||_{\lambda_1}$ is a mixed norm defined as $\sqrt{(1-\lambda_1)||\cdot||_2^2 + \lambda_1||\cdot||_1^2}$, with $\lambda_1 \in [0, 1]$. Note that if $\lambda_1 = 0$, this mixed norm is equivalent to the L-2 norm, while it is equivalent to the L-1 norm if $\lambda_1 = 1$. With $\hat{\xi}_1$, one can sequentially obtain subsequent eigenvectors by solving the following optimization problems for j = 2,3,...,r:

$$\max_{||\xi_j||_2=1, \xi_{j-1}\perp\xi_j} \frac{\xi_j^\intercal \hat{\Sigma} \xi_j}{||\xi_j||_{\lambda_j}^2}, \tag{20}$$

where λ_k is the tuning parameter for ξ_k (which might be different for each k). In short, SPCA produces "sparse" factor loadings in the sense that many of them are identically zero, while factors are still constructed in the spirit of PCA, since explained data variances are maximized under constraints. Qi et al. (2013) show that their proposed algorithm for optimizing objective functions yields a stable limit which consistently estimates the eigenvectors under certain conditions. Our apporach, as discussed above, is to first estimate high-frequency sparse principal components, and then construct and utilize realized volatilities from these estimated factors in (13) to forecast any given target of interest.

3.4 Forecasting Methods

The proposed one-step forecasting model is:

$$\widehat{\text{TRV}}_{t+1} = \beta_0 + \beta_1 \widehat{\text{TRV}}_t + \beta_2 \widehat{\text{TRV}}_{[t,t-4]} + \beta_3 \widehat{\text{TRV}}_{[t,t-21]} + \beta_\Psi^\intercal \hat{\Psi}_t + \epsilon_t. \tag{21}$$

The estimated factors' volatilities, $\hat{\Psi}_t$, are constructed by implementing the above mentioned two-step procedure. Recall that the first step involves using LASSO or elastic net shrinkage to select a subset of the asset dataset, as outlined in Section 3.1. The second step involves latent volatility factor construction, as discussed in Sections 3.2 and 3.3.

We choose the number of latent factors in our experiments by following an easy-to-implement, albeit ad-hoc rule. First, we sort all eigenvalues in descending order and select (additional) principal components based on their corresponding eigenvalues until their cumulative contribution exceeds (or is equal to) 90% of the total variation of the dataset. Next, we discard principal components with individual contributions that are less than 5% of total variation. For instance, if the first 5 principal components contribute 60%, 10%, 10%, 6%, 4%, respectively, we keep the first 4 principal components. The idea is very simple and natural: there is a trade-off between a more parsimonious model and a less informative one. Although the choice of cutoffs is somewhat arbitrary, our experiments suggest that the findings are robust to other cutoffs within a reasonable range of the above ones. Finally, we estimate daily integrated volatility of selected latent factors and use them as predictors in (21).

In summary, we consider six "permutations" of our two-step procedure in forecasting experiments, as follows:

I. EN1-PCA: First step - assets selected using elastic net (EN) shrinkage, with parameter α = 0.2. Second step - latent integrated volatility factors constructed using PCA.

II. EN2-PCA: First step - assets selected using elastic net (EN) shrinkage, with parameter α = 0.6. Second step - latent integrated volatility factors constructed using PCA.

III. LASSO-PCA: First step - assets selected using LASSO shrinkage, with parameter α = 0.2. Second step - latent integrated volatility factors constructed using PCA.

IV. EN1-SPCA: First step - assets selected using elastic net (EN) shrinkage, with parameter α = 0.2. Second step - latent integrated volatility factors constructed using SPCA.

V. EN2-SPCA: First step - assets selected using elastic net (EN) shrinkage, with parameter α = 0.6. Second step - latent integrated volatility factors constructed using SPCA.

VI. LASSO-SPCA: First step - assets selected using LASSO shrinkage, with parameter α = 0.2. Second step - latent integrated volatility factors constructed using SPCA.

Model estimation and volatility prediction are carried out anew, each day, using a rolling-window estimation scheme. The length of rolling window (i.e. the in-sample period), is 630 days. For example, we first estimate models using data from December 28, 2006 to June 30, 2009 (630 trading days), and then construct one-day-ahead forecasts for July 1, 2009. Then, in order to forecast the volatility on July 2, 2009, we first estimate our models using data from December 29, 2006 to July 1, 2009 (630 trading days). We continue this procedure until we reach the end of our dataset. Finally, we obtain sequences of daily out-of-sample volatility forecasts for the sample period from July 1, 2009 to December 31, 2010, which constitutes 380 trading days.

Our benchmark HAR model is estimated using ordinary least squares. All factor-augmented regressions are estimated using constrained least squares, in order to guarantee that all parameters are nonnegative. By doing so, we avoid any potential negative forecasts of volatility.

To evaluate the forecasting performance of our factor-augmented models and compare them with the benchmark model, we consider three different criteria:

(a) In-sample R^2.

(b) Out-of-sample R^2 (Campbell and Thompson (2008)), defined as:

$$R^2_{\mathrm{OOS}} = 1 - \frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{T}(y_t - \bar{y}_t)^2}, \tag{22}$$

where y_t is the ex-post value of volatility, $\bar{y}_t$ is the historical average of volatility, and $\hat{y}_t$ is our forecast.

(c) Heteroskedasticity adjusted root mean square error (HARMSE) (Corsi et al. (2010)), defined as:

$$\text{HARMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\frac{y_t - \hat{y}_t}{y_t})^2} \tag{23}$$

The experimental setup discussed in this section is summarized in Table 1.

4 Empirical Results

4.1 Data

We collect intraday observations on 267 constituents of the S&P 500 index[2]; 9 sector ETFs, including; Materials (XLB), Energy (XLE), Financial (XLF), Industrial (XLI), Technology (XLK), Consumer Staples (XLP), Utilities (XLU), Health Care (XLV), and Consumer Discretionary (XLY); and the SPDR S&P 500 ETF (SPY). Our sample period is January 3, 2006 to December 31, 2010, and data are collected from the TAQ database.

In our forecasting experiments, target assets include SPY; the 9 sector ETFs listed above; and 11 individual stocks, including: Coca-Cola Company (KO), Exxon Mobil Corporation (XOM), General Electric Company (GE), Goldman Sachs (GS), International Business Machines (IBM), Johnson & Johnson (JNJ), JPMorgan Chase (JPM), McDonald's (MCD), Merck (MRK), Microsoft (MSFT) and Wal-Mart (WMT).

It is worth mentioning that the original dataset we collected consists of 274 constituents of S&P 500 index. Of these, seven stocks, including AIG, C, F, GNW, HIG, LVLT and STT, are deleted, leaving 267 stocks. The reason for this is that these stocks generate a small number of extreme integrated volatility values, even when data are filtered with using a judiciously chosen jump threshold. These stocks are thus viewed as "outliers" that contains strong microstructure noises and/or recording error, which are not informative about future volatilities, hence may consequently deteriorate forecasting performance of our models. As a robustness check, however, we did compare empirical results based on 267 constituents with those based on 274 constituents, although comparable results are only shown from the SPY case. Complete results based on the original data set of 274 constituents are available upon request, although it is clear, upon

[2]Since the constituents of S&P 500 index change over time, we only collect those that are always present in the index between 2006 to 2010.

comparison of our results in these two cases, that utilizing the 7 additional stocks result in a deterioration of the predictive performance of out latent volatility factors.

Finally, data cleaning, subsampling, etc., all follow standard procedures described in Aït-Sahalia and Jacod (2012). Overnight returns are excluded. Less frequently traded stocks are also excluded from the dataset since they do not generate high-frequency data.

4.2 Empirical Findings: Forecasting Performance

Tables 2–22 show the one-day ahead forecast performance of the benchmark HAR model and various factor-augmented HAR models, for the forecasting sample period from July 1, 2009 to December 31, 2010. All tables report in-sample and out-of-sample R^2 values, as well as HARMSE values. Table 2 (SPY) also compares the results with and without the aforementioned seven "outlier" stocks (first and second columns under each criterion). Moreover, to compare the performance across different sampling frequencies, we construct factors using 1-minute, 5-minute, and 10-minute frequency data, respectively. Finally, as discussed above, forecasting experiments are carried out using rolling windows to estimate all models, prior to ex ante forecast construction at each point in time. A number of clear-cut conclusions emerge upon inspection of the results contained in these tables.

First, *in-sample fit* is surprisingly stable across different models, including our benchmark HAR model and our volatility-factor augmented models, across three different data frequencies, including 1-minute, 5-minute, and 10-minute frequencies. Thus, there is little to choose between data frequencies when comparing in-sample model fit. Moreover, in-sample model fit is surprisingly similar across different asset classes (i.e., market index, sector ETFs, and individual stocks), with most R^2 values ranging rather tightly between 0.35 and 0.55. More specifically, most in-sample R^2 values for sector ETFs range rather tightly between approximately 0.50 and 0.65, regardless of whether our HAR specifications include a latent volatility factor or not. The exception to this appears to be XLP (Consumer Staples, Table 8), for which values range from 0.38 to 0.50. The market ETF (SPY, Table 2) delivers in-sample R^2 values between approximately 0.55 and 0.65. Finally, for individual stocks, the range is somewhat wider, including values from 0.35 to 0.65. Finally, in-sample fit changes little when volatility factors are added to benchmark HAR models, regardless of asset class. Thus, based solely on in-sample diagnostics, there appears to be little gain to deploying volatility factors in HAR analysis. However, we shall see that this finding changes dramatically when out-of-sample, or true ex ante forecasting, is carried out.

Second, our in-sample findings are highly mis-leading, when the objective of interest is *out-of-sample* volatility prediction. Namely, all of the above findings become irrelevant when ex ante prediction experiments are carried out. For example, for forecasting, data frequency is crucial, and the "best" frequency varies across different assets and asset classes. However, we still recommend using the 5-minute frequency, as a general rule-of-thumb. This is because our factor augmented HAR models generally yield the "best" predictions (see below for further discussion) using 5-minute frequency data, when comparing results factor augmented model predictive accuracy across different frequencies. Intuitively, note that on one hand, using higher frequency data may result in a substantial amount of microstructure noise being absorbed by extracted factors, hence potentially deteriorating predictive performance. On the other hand, if the sampling frequency is relatively low, it is more difficult to eliminate individual jumps when estimating latent factors, leading to forecast deterioration.

Third, note that the the above findings are based on a comparison of predictions made using factor augmented HAR models. This is the correct comparison to make because predictive accuracy improves appreciably when latent common volatility factors are included in our benchmark HAR-type model. For example, for Johnson & Johnson (see Table 15), the benchmark model using 5-minute frequency data achieves an out-of-sample R^2 value of only 0.14. This is approximately one-third of the out-of-sample R^2 value associated with our "best" factor-augmented model. This pattern occurs for many firms and sectors; as well as for the market ETF. Interestingly, if only in-sample R^2 values were examined in order to assess the usefulness of common factors, then the story would change markedly. For example, again using Johnson & Johnson to illustrate our findings, the benchmark model using 5-minute frequency data (without a common factor) achieves an in-sample R^2 value of 0.39, while in-sample R^2 values for our factor-augmented models are all between 0.43 and 0.48. This small increase associated with utilizing common factors in an in-sample context characterizes all of our experiments. Indeed, substantial increases in performance only arise when using latent factors for ex ante prediction. As discussed in the introduction to this paper, this finding constitutes strong evidence of an important difference between findings based on in- and out-of-sample experiments.

The above conclusion can perhaps best be understood by noting that in-sample R^2 values are widely known to be substantively greater than out of sample R^2 values in financial forecasting applications. This feature has been extensively discussed in the literature, and reasons for it range from the presence of (smooth) structural breaks and state transitions, to the general inability of linear models to capture inherently nonlinear interactions among financial variables and markets (e.g., see Ang and Timmermann (2012), Aiolfi et al. (2009), and Paye and Timmermann (2006)). Naturally, arguments centering around market efficiency may also play a role in explaining this phenomenon. Not surprisingly, then, when comparing benchmark HAR models, we find that in-sample R^2 values are indeed much greater than their out-of-sample benchmark HAR counterparts. For example, using IBM (see the 5-minute panel in Table 14) to illustrate our findings, the benchmark model (without a common factor) achieves an in-sample R^2 value of 0.61, as opposed to an out-of-sample R^2 value of 0.24. However, when the "best" factor augmented in-sample and out-of sample performances are compared in this example, the R^2 values are 0.65 and 0.38, respectively. Thus, the relative out-of-sample gains associated with utilizing latent volatility factors are greater than the in-sample gains, as the out-of-sample R^2 value increases from 0.24 to 0.38, which is more than a 50% gain. Indeed, analogous predictive accuracy gains exceed 50% for GE, JNJ, JPM, KO, MCD, MRK, WMT, and XOM (see Tables 12, 15, 16, 17, 18, 19, 21 and 22, respectively), with 5-minute frequency data. Lesser gains arise for only 2 of 11 stocks that we analyze. Broadly speaking, this feature also characterizes our results at all market and sector levels, although it is more starkly apparent at the individual stock level.

Fourth, models utilizing SPCA in factor construction generally forecast "better" than those utilizing PCA. Moreover, the gains to using SPCA, relative to PCA, are greatest when one moves from using 10-minute to 5-minute frequency data, as well as when one moves from using 1-minute to 5-minute frequency data. This two-pronged finding is as expected, given that using high frequency data across many stocks, when constructing latent volatility factors, involves accounting for noisiness due not only to sampling frequency (i.e., microstructure noise), but also due to the large number of assets, a increasing number of which are transmitting noisy signals, as the cross sectional dimension of our dataset increases. This argument, parallels the argument outlined above, whereby using higher frequency data may result in more microstructure noise being absorbed by extracted factors, while when the sampling frequency is relatively low (or when the number of assets is relatively high), it may be more difficult to eliminate individual jumps when estimating latent factors.

Drilling down a bit further, the results in Table 12 indicate that at 1- and 5-minute frequencies, factor-augmented models with SPCA have a 25%–35% larger out-of-sample R^2 than those with PCA. Similar results can also be found in Tables 13, 14, 18, 20, 21 and 22. This pattern, however, becomes insignificant or even reversed at our lowest sampling frequency (i.e., the 10-minute frequency). Moreover, when forecasting individual stocks, as well as some ETFs, such as SPY, XLB, XLE, XLI, XLK and XLY (see Tables 2, 3, 4, 6, 7 and 11, respectively), factor-augmented models with SPCA yield much lower HARMSE, especially at when using higher frequency data. Again, this pattern becomes less significant at lower frequency. As discussed above, this finding likely due to the presence of microstructure noise in our data, given that SPCA assigns many identically zero weights on stocks, and consequently alleviates some of the effect of microstructure noise; particularly from stocks, which are non-informative about the volatility of the target asset. Therefore, we are not surprised that factor-augmented models using SPCA are more likely to perform better than those using PCA at higher frequencies. Of course, it is perhaps worth noting that due to aggregation, the impact of microstructure noise on our market index ETF and sector ETFs is much weaker. As a consequence, the difference among models utilizing SPCA and PCA when forecasting our ETFs is less pronounced, as mentioned above.

Fifth, there is an important wrinkle to the above story. Namely, for financial assets, out-of sample R^2 values are approximately 0 in some cases. A particularly interesting example of this is the financial sector ETF. For this ETF, in-sample R^2 values range from around 0.53 to 0.64, while out-of-sample R^2 range from around 0.08 to 0.30 (see Table 5). At the individual stock level, the picture is even more stark. Consider Goldman Sachs (see Table 13). In-sample R^2 values are always around 0.40, while out-of-sample R^2 values are always less than 0. Evidently, integrated volatility of individual financial stocks is the most difficult to forecast. Unlike forecasting the financial sector as a whole, when it comes to individual financial stocks, HAR-type models performs very poorly. In Tables 13 and 16, entries in the column of out-of-sample R^2 for the benchmark model are almost all negative, HARMSE are in general much larger than those for other assets, and even in-sample R^2 values are much lower compared to other assets.

However, all is not lost. Incorporating common volatility factors extracted from a broad range of stocks into benchmark models sometimes helps in obtaining more precise forecasts for financial stocks, but only to a very limited extent. As discussed above, for many of our target variables, there is substantial predictable content. For example, out-of-sample R^2 values for Coca-Cola (see Table 17), Exxon Mobil (see Table 22), and IBM (see Table 14) range from 0.35 to 0.41, from 0.30 to 0.37, and from 0.23 to 0.38, respectively, when using common volatility factors constructed via our two-step procedure, and based on IV estimators constructed using 5-minute frequency data.

Sixth, financial stocks are frequently selected in our first variable selection (or shrinkage) step. However, they are often assigned small weights in the second step (i.e., the latent factor estimation step), particularly when SPCA is used in this step. For instance, when we forecast the volatility of our energy sector ETF using 1-minute frequency data, over 33% of the most frequently selected stocks in the first step are in financial sector. However, the average weight assigned by PCA to, for instance, Goldman Sachs is only around 0.09, while the corresponding weight assigned to Texas Instruments is around double that (see Table 24). Even more starkly, the average weight assigned by SPCA to Goldman Sachs drops is only around 0.02. This is in part due to the fact that over 50% of weights assigned by SPCA are identically zero. On the contrary, the average weight on Texas Instruments Incorporated rises to 0.19. Therefore, we conjecture that the contribution of financial stocks to common volatility factors may be less than that of stocks in other sectors, based on these rather surprising findings. Moreover, and as a result of the above findings, it is very likely that the marginal predictive content of common volatility factors is largely accounted for by information in sectors other than the financial sector, such as the industrial and technology sectors.

4.3 Empirical Findings: Latent Factor Structures

Tables 23–25 contain factor structure details, for the case where we are interested in forecasting non-financial sector ETFs and individual stocks. A number of conclusions emerge when examining these results.

First, note that different shrinkage methods in the first step of our procedure select almost the same pool of stocks, for each sampling frequency. Thus, there appears to be little to choose between the LASSO and elastic net shrinkage. However, the pool of selected stocks changes with data frequency. For instance, consider the SPY ETF. Table 23 shows that at the 1-minute frequency, almost 32% of selected stocks belong to the financial sector. In contrast, at 5-minute and 10-minute frequencies, only around 15% to 20% of selected stocks are financials. Similar results can be seen upon inspection of Table 24 (sector ETF) and 25 (individual stock).

Second, an important feature of our volatility factors is that financial stocks tend to be selected frequently in the first step of our procedure, particularly when using higher frequency data. However, relatively little weight is placed on such stocks in the second step of our procedure, when utilizing PCA and SPCA to estimate asset return factors. For instance, in columns denoted "PCA" in these three tables, the average weight on HBAN (Huntington Bancshares) is only between 0.06 and 0.07, when using 1-minute frequency data. Similarly, BK (Bank of New York Mellon) has average weight around 0.06–0.09, when using 5-minute frequency data, and MMC (Marsh & McLennan Companies) in Table 23, GS (Goldman Sachs) in Table 24 and LM (Legg Mason) in Table 25 have average weights of around 0.1 or less, when using 10-minute frequency data. Furthermore, under "SPCA", the average weights on financial stocks are even smaller, and many are identically zero. For instance, Table 23 shows that at the 1-minute frequency, the average weight on PRU (Prudential Financial) decreases dramatically from 0.104 to 0.047 when factor estimation utilizes SPCA instead of PCA (under SPCA almost 28% of daily weights are zero). This finding is consistent with our above microstructure noise explanation of the superior performance of models that utilize SPCA, in conjunction with the use of higher frequency data.

Third, notice that stocks in the industrial and technology sectors usually have larger factor loadings (weights) under both PCA and SPCA. For instance, in Table 25, CSCO (Cisco), LLTC (Linear Technology) and SWKS (Skyworks Solutions) - in the technology sector, and MAS (Masco), UPS and UTX (United Technologies) - in the industrial sector, all have average weights greater than 0.15. Similarly, in Table 24, CERN (Cerner), NFLX (Netflix) and TXN (Texas Instruments) - in technology sector, and CSX, FAST (Fastenal) and HON (Honeywell) - in the industrial sector - have average weights larger than 0.15. Putting all of the above evidence together, we conclude that although financial stocks are frequently chosen in our first step shrinkage procedure, their contributions to common volatility factors appears to be less than that of industrial and technology stocks.

5 Concluding Remarks

This paper investigates whether latent common volatility factors extracted from a large-dimensional high-frequency intraday stock returns improve volatility forecasting. We propose a factor-augmented version of the widely studied HAR model. In our new model, factors are estimated using a two-step procedure involving variable selection using least absolute selection operator (LASSO) and elastic net shrinkage, followed by factor estimation using (sparse) principal components analysis (SPCA).

Our key findings are summarized as follows. First and foremost, we uncover substantial empirical evidence indicating that latent common volatility factors greatly improve the out-of-sample predictive accuracy of HAR models, as measured by both HARMSE and out-of-sample R^2. This improvement is seen across markets, sectors, and individual companies, with the greatest improvements noted at the individual company level. Second, in-sample performance is often irrelevant to out-of-sample performance. Indeed, if volatility modeling is viewed solely through the lens of in-sample fit, then little is gained by generalizing the HAR model using our procedure. Almost all gains are seen only when true ex ante prediction is carried out. Third, we recommend using high frequency datasets consisting of data sampled a a 5-minute frequency, when constructing predictions of volatility using factor augmented regressions. This recommendation arises because of microstructure noise considerations, as well as because of the incidence of heterogeneous jumps associated with the large cross sectional dimension of our dataset. We also find that models utilizing SPCA perform better than those with PCA, when these methods are used to extract common volatility factors.

This paper is meant as a starting point, as much remains to be done. For example, although substantial theoretical advances in the application of principal component analysis to high dimensional asset return datasets are made in Aït-Sahalia and Xiu (2016a) and Aït-Sahalia and Xiu (2016b), it remains to ascertain whether the results carry over to the use of SPCA. It also remains to theoretically analyze higher order latent (e.g., volatility) factors that are estimated based using first order latent factors constructed using observed (asset) data. From an empirical perspective, it will be of interest to further examine the robustness of the findings in this paper to the use of alternative sample periods for both in-sample estimation and out-of sample prediction. It will also be of interest to assess whether the findings in this paper can be translated into profitable investment strategies, in real-time trading contexts.

Table 1: Experimental Setup

Benchmark Model:	
$\widehat{\mathrm{TRV}}_{t+1} = \beta_0 + \beta_1 \widehat{\mathrm{TRV}}_t + \beta_2 \widehat{\mathrm{TRV}}_{[t,t-4]} + \beta_3 \widehat{\mathrm{TRV}}_{[t,t-21]} + \epsilon_t$	
Two-Step Procedure:	
Step 1: Shrinkage Methods (Variable Selection)	Step 2: Factor Estimation Methods
1. LASSO (α = 0)	1. PCA
2. EN1 (α = 0.2)	
3. EN2 (α = 0.6)	2. SPCA
Sample Periods:	
In-sample period: January 3, 2006 – June 30, 2009	
Out-of-sample period: July 1, 2009 – December 31, 2010	
Regression Estimation Scheme:	
Rolling-window estimation.	
Window length: 630 days.	
Sampling Frequencies:	
1, 5, and 10 minutes.	
Factor Selection Rules:	
Contribution of all selected factors exceeds 90% of total variation.	
Contribution of every selected factor exceeds 5% of total variation.	
Evaluation Criteria:	
1. In-sample R^2	
2. Out-of-sample R^2	
3. Heteroskedasticity adjusted root mean square error (HARMSE)	

Table 2: SPDR S&P 500 ETF (SPY)

Frequency	Model	In-Sample R^2		Out-of-Sample R^2		HARMSE	
1-minute	Benchmark	0.5218	0.5218	0.2737	0.2737	1.2493	1.2493
	EN1-PCA	0.5302	0.5279	0.3030	0.3004	0.8443	1.0169
	EN2-PCA	0.5304	0.5279	0.3181	0.2823	0.8276	0.9794
	Lasso-PCA	0.5304	0.5280	0.3164	0.2985	0.8347	0.9981
	EN1-SPCA	0.5458	0.5408	0.3312	0.1822	0.6245	0.9497
	EN2-SPCA	0.5461	0.5408	0.3421	0.1626	0.6313	0.9911
	Lasso-SPCA	0.5461	0.5413	0.3197	0.1601	0.6350	0.9959
5-minute	Benchmark	0.6006	0.6006	0.3605	0.3605	1.2629	1.2629
	EN1-PCA	0.6071	0.6029	0.3897	0.3801	0.9828	1.1222
	EN2-PCA	0.6047	0.6031	0.3931	0.3780	1.0646	1.0984
	Lasso-PCA	0.6039	0.6030	0.3774	0.3759	1.0240	1.1122
	EN1-SPCA	0.6204	0.6088	0.4313	0.3995	0.7066	0.9393
	EN2-SPCA	0.6202	0.6088	0.4381	0.4000	0.7141	0.9156
	Lasso-SPCA	0.6193	0.6086	0.4233	0.4071	0.7012	0.9497
10-minute	Benchmark	0.5039	0.5039	0.2609	0.2609	1.6082	1.6082
	EN1-PCA	0.5445	0.5461	0.3829	0.3342	1.0496	1.0796
	EN2-PCA	0.5440	0.5373	0.3705	0.2729	1.0213	1.1176
	Lasso-PCA	0.5453	0.5363	0.3725	0.2810	1.0323	1.1523
	EN1-SPCA	0.5457	0.5428	0.3960	0.3239	1.0672	1.0790
	EN2-SPCA	0.5434	0.5362	0.3800	0.2816	1.0670	1.0963
	Lasso-SPCA	0.5449	0.5361	0.3833	0.2992	1.1066	1.1081

Note: See Table 1. Entries are statistics that measure in-sample and out-of-sample volatility forecasting performance of the HAR model given in equation (21) of Section 3.4, for the target variable given in the title of the table (i.e., the SPY ETF). All models other than the benchamrk (HAR) model, denoted as "Benchmark", include latent volatility factors. EN1 and EN2 denote models for which elastic net shrinkage is used in initial variable selection, with α = 0.2 and 0.6, respectively. Lasso denotes use of the least absolute shrinkage operator in initial variable selection. After initial variable selection, either PCA or sparse PCA (i.e., SPACA) are utilized to obtain the laten volatility factor used in all models denoted as such. In-Sample R^2, Out-of-Sample R^2 and HARMSE entries in this table consist of 2 columns each, the first of which corresponds to predictions made using 267 stocks in factor construction, and the second of which utilizes 274 stocks in the step of our analysis (see Section 4.1 for further details). All other tables report results based only on the analysis of 267 stocks. Complete details are given in Sections 3 and 4.

Table 3: Materials Sector ETF (XLB)

Frequency	Model	In-Sample R^2	Out-of-Sample R^2	HARMSE
1-minute	Benchmark	0.5598	0.3204	0.8258
	EN1-PCA	0.5598	0.3204	0.8258
	EN2-PCA	0.5598	0.3204	0.8258
	Lasso-PCA	0.5598	0.3204	0.8258
	EN1-SPCA	0.5678	0.3616	0.6902
	EN2-SPCA	0.5673	0.3647	0.6904
	Lasso-SPCA	0.5673	0.3686	0.6910
5-minute	Benchmark	0.6234	0.2853	1.0050
	EN1-PCA	0.6274	0.3107	0.9057
	EN2-PCA	0.6271	0.3047	0.9316
	Lasso-PCA	0.6269	0.3053	0.9303
	EN1-SPCA	0.6341	0.3322	0.7841
	EN2-SPCA	0.6345	0.3351	0.7887
	Lasso-SPCA	0.6348	0.3445	0.7970
10-minute	Benchmark	0.5497	0.1131	1.2993
	EN1-PCA	0.5712	0.1699	1.0226
	EN2-PCA	0.5717	0.1684	1.0258
	Lasso-PCA	0.5709	0.1682	1.0329
	EN1-SPCA	0.5702	0.1833	1.0078
	EN2-SPCA	0.5694	0.1815	1.0187
	Lasso-SPCA	0.5690	0.1735	1.0100

Notes: See notes to Table 2.

Table 4: Energy Sector ETF (XLE)

Frequency	Model	In-Sample R^2	Out-of-Sample R^2	HARMSE
1-minute	Benchmark	0.5221	0.1932	1.1601
	EN1-PCA	0.5239	0.2910	0.9712
	EN2-PCA	0.5236	0.2925	0.9773
	Lasso-PCA	0.5236	0.2910	0.9764
	EN1-SPCA	0.5445	0.3592	0.6126
	EN2-SPCA	0.5451	0.3664	0.6065
	Lasso-SPCA	0.5462	0.3637	0.6018
5-minute	Benchmark	0.6203	0.3153	1.1597
	EN1-PCA	0.6240	0.3750	1.0010
	EN2-PCA	0.6242	0.3577	0.9668
	Lasso-PCA	0.6231	0.3608	0.9830
	EN1-SPCA	0.6286	0.4192	0.7402
	EN2-SPCA	0.6308	0.4277	0.7335
	Lasso-SPCA	0.6298	0.3993	0.7473
10-minute	Benchmark	0.5374	0.1878	1.4904
	EN1-PCA	0.5667	0.3442	0.9174
	EN2-PCA	0.5656	0.3575	0.8816
	Lasso-PCA	0.5652	0.3547	0.9139
	EN1-SPCA	0.5601	0.3315	0.8931
	EN2-SPCA	0.5595	0.3793	0.8741
	Lasso-SPCA	0.5591	0.3820	0.8863

Notes: See notes to Table 2.

Table 5: Financial Sector ETF (XLF)

Frequency	Model	In-Sample R^2	Out-of-Sample R^2	HARMSE
1-minute	Benchmark	0.5423	0.3026	0.7466
	EN1-PCA	0.5441	0.2695	0.7847
	EN2-PCA	0.5445	0.2433	0.7978
	Lasso-PCA	0.5450	0.2106	0.8075
	EN1-SPCA	0.6258	0.3585	0.6668
	EN2-SPCA	0.6299	0.3085	0.7173
	Lasso-SPCA	0.6335	0.1985	0.7334
5-minute	Benchmark	0.5823	0.2853	1.3230
	EN1-PCA	0.6085	0.2565	1.2094
	EN2-PCA	0.6028	0.2896	1.2651
	Lasso-PCA	0.5972	0.2508	1.3114
	EN1-SPCA	0.6145	0.2612	1.2482
	EN2-SPCA	0.6150	0.2648	1.3372
	Lasso-SPCA	0.6149	0.2652	1.3766
10-minute	Benchmark	0.4950	0.1276	1.7122
	EN1-PCA	0.5386	0.0960	1.8255
	EN2-PCA	0.5393	0.1032	1.8221
	Lasso-PCA	0.5391	0.1053	1.8167
	EN1-SPCA	0.5427	0.0852	1.8423
	EN2-SPCA	0.5400	0.0881	1.8621
	Lasso-SPCA	0.5402	0.0984	1.8578

Notes: See notes to Table 2.

Table 6: Industrial Sector ETF (XLI)

Frequency	Model	In-Sample R^2	Out-of-Sample R^2	HARMSE
1-minute	Benchmark	0.5573	0.3389	0.8208
	EN1-PCA	0.5668	0.3589	0.6438
	EN2-PCA	0.5659	0.3607	0.6585
	Lasso-PCA	0.5658	0.3491	0.6513
	EN1-SPCA	0.5848	0.3724	0.5040
	EN2-SPCA	0.5887	0.3681	0.4973
	Lasso-SPCA	0.5890	0.3771	0.4896
5-minute	Benchmark	0.6219	0.3217	1.6667
	EN1-PCA	0.6398	0.3211	1.0840
	EN2-PCA	0.6389	0.3155	1.3193
	Lasso-PCA	0.6380	0.3094	1.2992
	EN1-SPCA	0.6547	0.3363	0.9299
	EN2-SPCA	0.6534	0.3324	1.0008
	Lasso-SPCA	0.6528	0.3364	0.9307
10-minute	Benchmark	0.5309	0.0715	1.5196
	EN1-PCA	0.5566	0.0982	1.0240
	EN2-PCA	0.5571	0.1125	1.0148
	Lasso-PCA	0.5575	0.1172	1.0048
	EN1-SPCA	0.5529	0.1136	1.0563
	EN2-SPCA	0.5540	0.1211	1.0552
	Lasso-SPCA	0.5538	0.1244	1.0401

Notes: See notes to Table 2.

Table 7: Technology Sector ETF (XLK)

Frequency	Model	In-Sample R^2	Out-of-Sample R^2	HARMSE
1-minute	Benchmark	0.5311	0.2340	0.6839
	EN1-PCA	0.5370	0.2848	0.5778
	EN2-PCA	0.5372	0.2800	0.5765
	Lasso-PCA	0.5372	0.2801	0.5757
	EN1-SPCA	0.5458	0.3103	0.4815
	EN2-SPCA	0.5458	0.3046	0.4806
	Lasso-SPCA	0.5456	0.2981	0.4914
5-minute	Benchmark	0.6171	0.2849	0.9884
	EN1-PCA	0.6207	0.3009	0.9021
	EN2-PCA	0.6192	0.2892	0.9350
	Lasso-PCA	0.6198	0.3031	0.9043
	EN1-SPCA	0.6302	0.3183	0.7123
	EN2-SPCA	0.6309	0.3077	0.7020
	Lasso-SPCA	0.6311	0.3093	0.7011
10-minute	Benchmark	0.5118	0.0451	1.3730
	EN1-PCA	0.5363	0.0929	0.9936
	EN2-PCA	0.5362	0.1002	0.9901
	Lasso-PCA	0.5364	0.0937	0.9833
	EN1-SPCA	0.5342	0.1050	0.9818
	EN2-SPCA	0.5341	0.1051	0.9791
	Lasso-SPCA	0.5344	0.1028	0.9476

Notes: See notes to Table 2.

Table 8: Consumer Staples Sector ETF (XLP)

Frequency	Model	In-Sample R^2	Out-of-Sample R^2	HARMSE
1-minute	Benchmark	0.3840	0.1047	0.7164
	EN1-PCA	0.4066	0.1905	0.4557
	EN2-PCA	0.4066	0.1931	0.4510
	Lasso-PCA	0.4067	0.1988	0.4554
	EN1-SPCA	0.4405	0.1830	0.3920
	EN2-SPCA	0.4353	0.1194	0.4026
	Lasso-SPCA	0.4342	0.1703	0.3983
5-minute	Benchmark	0.4578	0.2790	0.9885
	EN1-PCA	0.4929	0.4753	0.5346
	EN2-PCA	0.4908	0.4162	0.5487
	Lasso-PCA	0.4910	0.4236	0.5675
	EN1-SPCA	0.5165	0.4089	0.5666
	EN2-SPCA	0.5188	0.3479	0.5794
	Lasso-SPCA	0.5180	0.4234	0.5744
10-minute	Benchmark	0.4001	0.1796	1.3737
	EN1-PCA	0.4870	0.2990	1.0413
	EN2-PCA	0.4887	0.2953	1.0044
	Lasso-PCA	0.4893	0.2789	1.0221
	EN1-SPCA	0.4840	0.2621	1.0628
	EN2-SPCA	0.4930	0.2432	1.0490
	Lasso-SPCA	0.4942	0.2278	1.0707

Notes: See notes to Table 2.

Table 9: Utilities Sector ETF (XLU)

Frequency	Model	In-Sample R^2	Out-of-Sample R^2	HARMSE
1-minute	Benchmark	0.5242	0.1309	0.8652
	EN1-PCA	0.5331	0.1515	0.5590
	EN2-PCA	0.5333	0.1359	0.5620
	Lasso-PCA	0.5332	0.1549	0.5587
	EN1-SPCA	0.5380	0.1869	0.5176
	EN2-SPCA	0.5374	0.1747	0.5161
	Lasso-SPCA	0.5376	0.1806	0.5265
5-minute	Benchmark	0.5683	0.1887	1.1073
	EN1-PCA	0.5906	0.2594	0.6719
	EN2-PCA	0.5870	0.2559	0.6629
	Lasso-PCA	0.5845	0.2746	0.6580
	EN1-SPCA	0.6160	0.2751	0.8010
	EN2-SPCA	0.6118	0.2618	0.7718
	Lasso-SPCA	0.6108	0.2655	0.7661
10-minute	Benchmark	0.4960	0.1882	1.3382
	EN1-PCA	0.5320	0.3671	0.8231
	EN2-PCA	0.5307	0.3879	0.8198
	Lasso-PCA	0.5304	0.3563	0.8261
	EN1-SPCA	0.5307	0.3662	0.8257
	EN2-SPCA	0.5292	0.3896	0.8234
	Lasso-SPCA	0.5292	0.3577	0.8382

Notes: See notes to Table 2.

Table 10: Health Care Sector ETF (XLV)

Frequency	Model	In-Sample R^2	Out-of-Sample R^2	HARMSE
1-minute	Benchmark	0.5053	0.2481	0.6576
	EN1-PCA	0.5185	0.2678	0.4433
	EN2-PCA	0.5186	0.2706	0.4399
	Lasso-PCA	0.5186	0.2610	0.4389
	EN1-SPCA	0.5257	0.2629	0.4309
	EN2-SPCA	0.5269	0.2796	0.4163
	Lasso-SPCA	0.5276	0.2395	0.4230
5-minute	Benchmark	0.4735	0.2067	1.0695
	EN1-PCA	0.5027	0.3332	0.6355
	EN2-PCA	0.5019	0.3218	0.6613
	Lasso-PCA	0.5014	0.3263	0.6656
	EN1-SPCA	0.5400	0.3070	0.6025
	EN2-SPCA	0.5404	0.3218	0.6022
	Lasso-SPCA	0.5423	0.2934	0.6048
10-minute	Benchmark	0.4566	0.2016	1.2785
	EN1-PCA	0.5087	0.3486	0.7555
	EN2-PCA	0.5102	0.3498	0.7428
	Lasso-PCA	0.5098	0.3648	0.7550
	EN1-SPCA	0.5056	0.3654	0.7431
	EN2-SPCA	0.5077	0.3533	0.7678
	Lasso-SPCA	0.5071	0.3733	0.7317

Notes: See notes to Table 2.

Table 11: Consumer Discretionary Sector ETF (XLY)

Frequency	Model	In-Sample R^2	Out-of-Sample R^2	HARMSE
1-minute	Benchmark	0.5255	0.3513	0.8867
	EN1-PCA	0.5370	0.4082	0.7845
	EN2-PCA	0.5367	0.3972	0.7855
	Lasso-PCA	0.5366	0.4126	0.7861
	EN1-SPCA	0.5557	0.4236	0.5854
	EN2-SPCA	0.5550	0.4018	0.5686
	Lasso-SPCA	0.5544	0.4171	0.5993
5-minute	Benchmark	0.5724	0.3408	1.3435
	EN1-PCA	0.5832	0.3581	1.2122
	EN2-PCA	0.5841	0.3691	1.2269
	Lasso-PCA	0.5849	0.3724	1.2084
	EN1-SPCA	0.6120	0.4058	0.9128
	EN2-SPCA	0.6119	0.4056	0.9064
	Lasso-SPCA	0.6117	0.4103	0.9152
10-minute	Benchmark	0.4784	0.1197	1.5265
	EN1-PCA	0.5177	0.1966	1.0291
	EN2-PCA	0.5195	0.1848	1.0086
	Lasso-PCA	0.5196	0.1853	1.0053
	EN1-SPCA	0.5133	0.1880	1.0273
	EN2-SPCA	0.5157	0.1854	0.9921
	Lasso-SPCA	0.5161	0.1904	0.9789

Notes: See notes to Table 2.

Table 12: General Electric Company (GE)

Frequency	Model	In-Sample R^2	Out-of-Sample R^2	HARMSE
1-minute	Benchmark	0.5211	0.2898	0.8151
	EN1-PCA	0.5243	0.3123	0.7342
	EN2-PCA	0.5240	0.3132	0.7350
	Lasso-PCA	0.5240	0.3130	0.7367
	EN1-SPCA	0.5792	0.3823	0.5655
	EN2-SPCA	0.5816	0.4005	0.5602
	Lasso-SPCA	0.5800	0.3954	0.5665
5-minute	Benchmark	0.5189	0.1576	1.2367
	EN1-PCA	0.5554	0.1710	0.9424
	EN2-PCA	0.5435	0.2130	1.0303
	Lasso-PCA	0.5522	0.1848	0.9533
	EN1-SPCA	0.5821	0.2586	0.8025
	EN2-SPCA	0.5823	0.2576	0.8256
	Lasso-SPCA	0.5830	0.2665	0.8301
10-minute	Benchmark	0.4825	0.0555	1.5839
	EN1-PCA	0.4893	0.0943	1.3869
	EN2-PCA	0.4900	0.1012	1.3659
	Lasso-PCA	0.4908	0.1041	1.3368
	EN1-SPCA	0.4901	0.0997	1.3638
	EN2-SPCA	0.4905	0.0980	1.3590
	Lasso-SPCA	0.4912	0.1020	1.3359

Notes: See notes to Table 2.

Table 13: The Goldman Sachs Group, Inc. (GS)

Frequency	Model	In-Sample R^2	Out-of-Sample R^2	HARMSE
1-minute	Benchmark	0.3939	-0.2130	1.6083
	EN1-PCA	0.3920	-0.2120	1.6196
	EN2-PCA	0.3920	-0.2120	1.6196
	Lasso-PCA	0.3920	-0.2120	1.6196
	EN1-SPCA	0.4078	0.0106	0.9982
	EN2-SPCA	0.4180	-0.0706	1.0676
	Lasso-SPCA	0.4317	-0.0856	1.0974
5-minute	Benchmark	0.4206	-0.2341	2.0352
	EN1-PCA	0.4187	-0.2100	2.0106
	EN2-PCA	0.4187	-0.2228	2.0288
	Lasso-PCA	0.4187	-0.2213	2.0235
	EN1-SPCA	0.4258	-0.1169	1.7640
	EN2-SPCA	0.4226	-0.1392	1.7918
	Lasso-SPCA	0.4402	-0.0366	1.3739
10-minute	Benchmark	0.3758	-0.2761	2.5707
	EN1-PCA	0.3957	-0.0406	1.5839
	EN2-PCA	0.4006	-0.0435	1.6176
	Lasso-PCA	0.3990	-0.0497	1.6556
	EN1-SPCA	0.3971	-0.0670	1.7213
	EN2-SPCA	0.4016	-0.0807	1.7189
	Lasso-SPCA	0.4007	-0.0789	1.7528

Notes: See notes to Table 2.

Table 14: International Business Machines Corporation (IBM)

Frequency	Model	In-Sample R^2	Out-of-Sample R^2	HARMSE
1-minute	Benchmark	0.5380	0.1149	1.1117
	EN1-PCA	0.5423	0.1855	0.9083
	EN2-PCA	0.5424	0.1707	0.9089
	Lasso-PCA	0.5424	0.1693	0.9146
	EN1-SPCA	0.5623	0.1478	0.6466
	EN2-SPCA	0.5621	0.2774	0.6214
	Lasso-SPCA	0.5632	0.2785	0.6124
5-minute	Benchmark	0.6140	0.2374	1.0384
	EN1-PCA	0.6194	0.3004	0.9106
	EN2-PCA	0.6281	0.3167	0.8908
	Lasso-PCA	0.6319	0.3215	0.8284
	EN1-SPCA	0.6463	0.3826	0.7098
	EN2-SPCA	0.6518	0.3709	0.7436
	Lasso-SPCA	0.6538	0.3578	0.7521
10-minute	Benchmark	0.5936	0.1696	1.1609
	EN1-PCA	0.5993	0.2111	1.0156
	EN2-PCA	0.5993	0.2138	1.0000
	Lasso-PCA	0.5995	0.2177	0.9866
	EN1-SPCA	0.5989	0.2306	0.9837
	EN2-SPCA	0.5987	0.2295	0.9792
	Lasso-SPCA	0.5986	0.2283	0.9772

Notes: See notes to Table 2.

Table 15: Johnson & Johnson (JNJ)

Frequency	Model	In-Sample R^2	Out-of-Sample R^2	HARMSE
1-minute	Benchmark	0.3611	0.1158	0.8426
	EN1-PCA	0.4040	0.2748	0.4300
	EN2-PCA	0.4039	0.2580	0.4331
	Lasso-PCA	0.4041	0.2588	0.4411
	EN1-SPCA	0.4415	0.2210	0.4721
	EN2-SPCA	0.4407	0.2319	0.4500
	Lasso-SPCA	0.4380	0.2300	0.4568
5-minute	Benchmark	0.3882	0.1398	1.0297
	EN1-PCA	0.4356	0.3251	0.5415
	EN2-PCA	0.4316	0.3355	0.5533
	Lasso-PCA	0.4310	0.3738	0.5387
	EN1-SPCA	0.4814	0.2968	0.5557
	EN2-SPCA	0.4816	0.3496	0.5746
	Lasso-SPCA	0.4815	0.3118	0.5709
10-minute	Benchmark	0.3740	0.1091	1.3244
	EN1-PCA	0.4395	0.2914	0.8430
	EN2-PCA	0.4432	0.3202	0.8348
	Lasso-PCA	0.4391	0.3136	0.8480
	EN1-SPCA	0.4450	0.3015	0.8406
	EN2-SPCA	0.4489	0.2860	0.8483
	Lasso-SPCA	0.4444	0.3513	0.8371

Notes: See notes to Table 2.

Table 16: JPMorgan Chase & Co. (JPM)

Frequency	Model	In-Sample R^2	Out-of-Sample R^2	HARMSE
1-minute	Benchmark	0.5122	-0.0657	1.1058
	EN1-PCA	0.5122	-0.0659	1.1060
	EN2-PCA	0.5122	-0.0657	1.1058
	Lasso-PCA	0.5122	-0.0657	1.1058
	EN1-SPCA	0.5730	-0.0419	0.9726
	EN2-SPCA	0.5742	-0.0790	1.0505
	Lasso-SPCA	0.5711	-0.0569	1.0658
5-minute	Benchmark	0.5543	0.0369	1.3160
	EN1-PCA	0.5640	0.0699	1.3026
	EN2-PCA	0.5592	0.0438	1.2999
	Lasso-PCA	0.5574	0.0552	1.3058
	EN1-SPCA	0.5745	0.0542	1.2892
	EN2-SPCA	0.5703	0.0713	1.2437
	Lasso-SPCA	0.5709	0.0877	1.2393
10-minute	Benchmark	0.4516	-0.2183	1.8278
	EN1-PCA	0.4682	-0.2899	1.8026
	EN2-PCA	0.4763	-0.2392	1.7838
	Lasso-PCA	0.4787	-0.2438	1.7681
	EN1-SPCA	0.4810	-0.2596	1.7216
	EN2-SPCA	0.4899	-0.2128	1.7214
	Lasso-SPCA	0.4911	-0.2275	1.7314

Notes: See notes to Table 2.

Table 17: The Coca-Cola Company (KO)

Frequency	Model	In-Sample R^2	Out-of-Sample R^2	HARMSE
1-minute	Benchmark	0.4082	0.1379	0.9190
	EN1-PCA	0.4384	0.2487	0.5931
	EN2-PCA	0.4386	0.2568	0.5920
	Lasso-PCA	0.4385	0.2417	0.5880
	EN1-SPCA	0.4504	0.2621	0.4465
	EN2-SPCA	0.4500	0.2681	0.4571
	Lasso-SPCA	0.4501	0.2478	0.4592
5-minute	Benchmark	0.5598	0.2292	1.1106
	EN1-PCA	0.6039	0.3952	0.7784
	EN2-PCA	0.5996	0.3626	0.7949
	Lasso-PCA	0.5998	0.3954	0.7813
	EN1-SPCA	0.6194	0.3500	0.7107
	EN2-SPCA	0.6166	0.4174	0.6635
	Lasso-SPCA	0.6170	0.3807	0.6651
10-minute	Benchmark	0.5006	0.1572	1.4081
	EN1-PCA	0.5687	0.2966	1.0139
	EN2-PCA	0.5702	0.2943	0.9600
	Lasso-PCA	0.5679	0.2459	1.0471
	EN1-SPCA	0.5707	0.2637	0.9683
	EN2-SPCA	0.5699	0.2831	0.9421
	Lasso-SPCA	0.5671	0.2428	0.9493

Notes: See notes to Table 2.

Table 18: McDonald's Corporation (MCD)

Frequency	Model	In-Sample R^2	Out-of-Sample R^2	HARMSE
1-minute	Benchmark	0.3738	-0.1118	0.9516
	EN1-PCA	0.3921	0.1359	0.6769
	EN2-PCA	0.3920	0.1485	0.6735
	Lasso-PCA	0.3920	0.1700	0.6680
	EN1-SPCA	0.4606	0.2318	0.5474
	EN2-SPCA	0.4576	0.1939	0.5311
	Lasso-SPCA	0.4604	0.0285	0.5411
5-minute	Benchmark	0.3785	-0.1553	1.3427
	EN1-PCA	0.4219	0.2491	0.8218
	EN2-PCA	0.4129	0.2093	0.8621
	Lasso-PCA	0.4152	0.2078	0.8535
	EN1-SPCA	0.4709	0.2464	0.7252
	EN2-SPCA	0.4711	0.2126	0.7607
	Lasso-SPCA	0.4721	0.1929	0.7566
10-minute	Benchmark	0.3299	-0.1506	1.9629
	EN1-PCA	0.4212	-0.0134	1.5636
	EN2-PCA	0.4174	0.0291	1.5366
	Lasso-PCA	0.4192	0.0759	1.5963
	EN1-SPCA	0.4241	0.0333	1.2639
	EN2-SPCA	0.4226	0.1089	1.3563
	Lasso-SPCA	0.4240	0.1185	1.3150

Notes: See notes to Table 2.

Table 19: Merck & Co., Inc. (MRK)

Frequency	Model	In-Sample R^2	Out-of-Sample R^2	HARMSE
1-minute	Benchmark	0.3563	0.1561	0.7700
	EN1-PCA	0.3910	0.2377	0.5873
	EN2-PCA	0.3916	0.2401	0.5876
	Lasso-PCA	0.3917	0.2400	0.5863
	EN1-SPCA	0.4508	0.2370	0.3857
	EN2-SPCA	0.4525	0.2424	0.3835
	Lasso-SPCA	0.4523	0.2381	0.3654
5-minute	Benchmark	0.4129	0.1914	0.9668
	EN1-PCA	0.4721	0.2825	0.6938
	EN2-PCA	0.4686	0.2804	0.7012
	Lasso-PCA	0.4687	0.2847	0.7420
	EN1-SPCA	0.5452	0.2966	0.4908
	EN2-SPCA	0.5444	0.2929	0.4956
	Lasso-SPCA	0.5454	0.2965	0.4892
10-minute	Benchmark	0.4723	0.2843	1.1830
	EN1-PCA	0.5028	0.3981	1.0175
	EN2-PCA	0.5020	0.3873	1.0178
	Lasso-PCA	0.5020	0.4072	1.0136
	EN1-SPCA	0.5010	0.4063	0.9486
	EN2-SPCA	0.5005	0.4143	0.9526
	Lasso-SPCA	0.5006	0.4117	0.9304

Notes: See notes to Table 2.

Table 20: Microsoft Corporation (MSFT)

Frequency	Model	In-Sample R^2	Out-of-Sample R^2	HARMSE
1-minute	Benchmark	0.5622	0.2316	0.7706
	EN1-PCA	0.5699	0.2751	0.7567
	EN2-PCA	0.5701	0.2639	0.7563
	Lasso-PCA	0.5702	0.2681	0.7555
	EN1-SPCA	0.5944	0.3269	0.5756
	EN2-SPCA	0.5955	0.3086	0.5804
	Lasso-SPCA	0.5952	0.3499	0.5781
5-minute	Benchmark	0.6116	0.2394	1.0603
	EN1-PCA	0.6173	0.2816	1.0187
	EN2-PCA	0.6172	0.2784	1.0136
	Lasso-PCA	0.6171	0.2777	1.0110
	EN1-SPCA	0.6329	0.3305	0.8306
	EN2-SPCA	0.6324	0.3169	0.7993
	Lasso-SPCA	0.6327	0.3141	0.8153
10-minute	Benchmark	0.4991	0.1190	2.2867
	EN1-PCA	0.5126	0.1930	2.0942
	EN2-PCA	0.5120	0.1798	2.1108
	Lasso-PCA	0.5114	0.1817	2.1030
	EN1-SPCA	0.5158	0.2166	2.0068
	EN2-SPCA	0.5146	0.1975	1.8044
	Lasso-SPCA	0.5136	0.2134	1.9075

Notes: See notes to Table 2.

Table 21: Wal-Mart Stores, Inc. (WMT)

Frequency	Model	In-Sample R^2	Out-of-Sample R^2	HARMSE
1-minute	Benchmark	0.3552	-0.3547	1.0014
	EN1-PCA	0.3693	-0.2423	0.8893
	EN2-PCA	0.3694	-0.2523	0.8926
	Lasso-PCA	0.3695	-0.2480	0.8926
	EN1-SPCA	0.3991	0.0710	0.6201
	EN2-SPCA	0.3977	0.1636	0.6056
	Lasso-SPCA	0.3976	0.1615	0.5976
5-minute	Benchmark	0.4571	-0.2414	1.1750
	EN1-PCA	0.4822	0.0186	0.8979
	EN2-PCA	0.4822	0.0918	0.9150
	Lasso-PCA	0.4849	0.1143	0.8939
	EN1-SPCA	0.5376	0.1146	0.8349
	EN2-SPCA	0.5274	0.0704	0.8387
	Lasso-SPCA	0.5263	0.0663	0.8486
10-minute	Benchmark	0.4012	-0.2658	1.5235
	EN1-PCA	0.4723	0.0772	1.0864
	EN2-PCA	0.4758	0.0955	1.0586
	Lasso-PCA	0.4759	0.0843	1.0716
	EN1-SPCA	0.4659	0.0665	1.1765
	EN2-SPCA	0.4684	0.0546	1.1446
	Lasso-SPCA	0.4690	0.0501	1.1425

Notes: See notes to Table 2.

Table 22: Exxon Mobil Corporation (XOM)

Frequency	Model	In-Sample R^2	Out-of-Sample R^2	HARMSE
1-minute	Benchmark	0.3620	-0.0409	1.2201
	EN1-PCA	0.3676	0.2644	0.7383
	EN2-PCA	0.3668	0.2590	0.7495
	Lasso-PCA	0.3667	0.2555	0.7482
	EN1-SPCA	0.3998	0.3272	0.4526
	EN2-SPCA	0.3995	0.3065	0.4518
	Lasso-SPCA	0.4001	0.3000	0.4684
5-minute	Benchmark	0.4823	0.0819	1.2181
	EN1-PCA	0.5032	0.3501	0.7590
	EN2-PCA	0.5011	0.3082	0.7848
	Lasso-PCA	0.4989	0.3053	0.7824
	EN1-SPCA	0.5444	0.3630	0.7415
	EN2-SPCA	0.5408	0.3450	0.7232
	Lasso-SPCA	0.5387	0.3744	0.7138
10-minute	Benchmark	0.4102	0.0355	1.5327
	EN1-PCA	0.4801	0.2805	1.0449
	EN2-PCA	0.4791	0.2996	1.0267
	Lasso-PCA	0.4812	0.2546	1.0546
	EN1-SPCA	0.4786	0.2386	1.0775
	EN2-SPCA	0.4781	0.2673	1.0312
	Lasso-SPCA	0.4809	0.1981	1.0852

Notes: See notes to Table 2.

Table 23: Factor Structure (SPY)

Stock			EN-1			Stock			EN-2			Stock			Lasso		
Sampling Frequency: 1 Minute																	
Ticker	Sector	Freq.	PCA	SPCA		Ticker	Sector	Freq.	PCA	SPCA		Ticker	Sector	Freq.	PCA	SPCA	
AFL	F	1.000	0.129	0.105	0.955	AFL	F	1.000	0.131	0.106	0.961	AFL	F	1.000	0.131	0.105	0.958
MO	CS	1.000	0.083	0.024	0.397	MO	CS	1.000	0.085	0.025	0.392	MO	CS	1.000	0.086	0.025	0.389
AMT	T	1.000	0.107	0.057	0.808	AMT	T	1.000	0.109	0.057	0.808	AMT	T	1.000	0.109	0.057	0.808
ABC	H	1.000	0.126	0.112	0.787	ABC	H	1.000	0.128	0.114	0.779	ABC	H	1.000	0.129	0.112	0.789
AMGN	H	1.000	0.126	0.101	0.832	AMGN	H	1.000	0.129	0.106	0.829	AMGN	H	1.000	0.129	0.105	0.832
BK	F	1.000	0.087	0.019	0.463	BK	F	1.000	0.088	0.019	0.463	BK	F	1.000	0.088	0.019	0.468
BBY	CD	1.000	0.137	0.126	0.963	BBY	CD	1.000	0.139	0.128	0.963	BBY	CD	1.000	0.140	0.128	0.963
CPB	CS	1.000	0.103	0.055	0.668	CPB	CS	1.000	0.105	0.055	0.661	CPB	CS	1.000	0.105	0.056	0.639
CAH	H	1.000	0.135	0.126	0.892	CAH	H	1.000	0.138	0.130	0.882	CAH	H	1.000	0.138	0.128	0.884
CI	H	1.000	0.109	0.059	0.805	CI	H	1.000	0.111	0.060	0.813	CI	H	1.000	0.112	0.060	0.797
CAG	CS	1.000	0.110	0.071	0.629	CAG	CS	1.000	0.111	0.073	0.645	CAG	CS	1.000	0.112	0.073	0.639
HSY	CS	1.000	0.105	0.054	0.632	HSY	CS	1.000	0.107	0.057	0.653	HSY	CS	1.000	0.107	0.057	0.642
HBAN	F	1.000	0.064	0.011	0.168	HBAN	F	1.000	0.066	0.011	0.184	HBAN	F	1.000	0.065	0.011	0.168
ILMN	H	1.000	0.105	0.055	0.629	ILMN	H	1.000	0.106	0.055	0.642	ILMN	H	1.000	0.108	0.056	0.634
KR	CS	1.000	0.090	0.034	0.487	KR	CS	1.000	0.092	0.033	0.487	KR	CS	1.000	0.092	0.033	0.484
LM	F	1.000	0.095	0.034	0.624	LM	F	1.000	0.096	0.034	0.618	LM	F	1.000	0.096	0.033	0.618
LNC	F	1.000	0.117	0.074	0.861	LNC	F	1.000	0.118	0.074	0.861	LNC	F	1.000	0.119	0.074	0.863
MMC	F	1.000	0.092	0.033	0.600	MMC	F	1.000	0.094	0.034	0.595	MMC	F	1.000	0.094	0.035	0.600
PEP	CS	1.000	0.114	0.072	0.771	PEP	CS	1.000	0.116	0.076	0.768	PEP	CS	1.000	0.116	0.074	0.766
PRU	F	1.000	0.104	0.047	0.718	PRU	F	1.000	0.106	0.048	0.721	PRU	F	1.000	0.106	0.047	0.734
SWN	E	1.000	0.118	0.082	0.837	SWN	E	1.000	0.119	0.083	0.847	SWN	E	1.000	0.119	0.081	0.834
SBUX	CD	1.000	0.151	0.158	0.987	SBUX	CD	1.000	0.154	0.160	0.987	SBUX	CD	1.000	0.154	0.158	0.987
SYY	CS	1.000	0.108	0.063	0.729	TROW	F	1.000	0.115	0.068	0.889	TROW	F	1.000	0.116	0.068	0.895
TROW	F	1.000	0.114	0.067	0.887	TIF	CD	1.000	0.133	0.112	0.968	TIF	CD	1.000	0.133	0.110	0.968
TIF	CD	1.000	0.130	0.110	0.971	UNP	I	1.000	0.142	0.132	0.966	UNP	I	1.000	0.142	0.132	0.963
UNP	I	1.000	0.139	0.129	0.966	UNM	F	1.000	0.125	0.092	0.887	UNM	F	1.000	0.125	0.092	0.884
UNM	F	1.000	0.123	0.092	0.882	WYNN	CD	1.000	0.143	0.133	0.987	WYNN	CD	1.000	0.143	0.132	0.982
CNX	U	0.997	0.133	0.114	0.955	MET	F	1.000	0.114	0.068	0.853	MET	F	1.000	0.114	0.068	0.853
WYNN	CD	0.997	0.141	0.131	0.984	CNX	U	0.992	0.135	0.115	0.960	SYK	H	0.992	0.144	0.140	0.947
CERN	T	0.992	0.147	0.150	0.960	CERN	T	0.992	0.150	0.154	0.960	CNX	U	0.987	0.136	0.114	0.960
SYK	H	0.989	0.142	0.138	0.939	SYK	H	0.982	0.144	0.142	0.944	AXP	F	0.979	0.123	0.088	0.949
MRK	H	0.971	0.123	0.094	0.894	SYY	CS	0.976	0.109	0.062	0.733	CERN	T	0.976	0.151	0.154	0.962
GE	I	0.968	0.134	0.120	0.927	AXP	F	0.961	0.124	0.090	0.945	SYY	CS	0.974	0.110	0.064	0.743
Sampling Frequency: 5 Minute																	
Stock			EN-1			Stock			EN-2			Stock			Lasso		
Ticker	Sector	Freq.	PCA	SPCA		Ticker	Sector	Freq.	PCA	SPCA		Ticker	Sector	Freq.	PCA	SPCA	
BK	F	1.000	0.077	0.020	0.418	BK	F	1.000	0.075	0.018	0.439	BK	F	1.000	0.075	0.018	0.434
BBY	CD	1.000	0.125	0.108	0.924	BBY	CD	1.000	0.123	0.105	0.913	BBY	CD	1.000	0.121	0.103	0.929
BSX	H	1.000	0.097	0.063	0.542	BSX	H	1.000	0.093	0.058	0.539	BSX	H	1.000	0.092	0.056	0.532
CNX	U	1.000	0.125	0.103	0.924	CNX	U	1.000	0.122	0.098	0.929	CNX	U	1.000	0.121	0.098	0.937
CSX	I	1.000	0.130	0.115	0.932	CSX	I	1.000	0.127	0.111	0.934	CSX	I	1.000	0.127	0.112	0.926
CAH	H	1.000	0.128	0.117	0.861	CAH	H	1.000	0.124	0.112	0.845	CAH	H	1.000	0.124	0.114	0.826
CERN	T	1.000	0.146	0.149	0.916	CERN	T	1.000	0.142	0.146	0.924	CERN	T	1.000	0.141	0.145	0.916
CI	H	1.000	0.098	0.053	0.689	CI	H	1.000	0.095	0.050	0.676	CI	H	1.000	0.094	0.050	0.668
CCI	T	1.000	0.085	0.040	0.584	CCI	T	1.000	0.084	0.039	0.584	CCI	T	1.000	0.083	0.039	0.587
DHI	CD	1.000	0.099	0.056	0.763	DHI	CD	1.000	0.098	0.056	0.750	DHI	CD	1.000	0.098	0.057	0.758
FITB	F	1.000	0.080	0.028	0.497	FITB	F	1.000	0.079	0.026	0.492	FITB	F	1.000	0.079	0.026	0.492
HSY	CS	1.000	0.094	0.051	0.574	HSY	CS	1.000	0.091	0.048	0.558	HSY	CS	1.000	0.090	0.048	0.582
ILMN	H	1.000	0.096	0.059	0.632	ILMN	H	1.000	0.097	0.061	0.629	ILMN	H	1.000	0.092	0.056	0.629
LNC	F	1.000	0.103	0.063	0.805	LNC	F	1.000	0.100	0.059	0.784	LNC	F	1.000	0.100	0.060	0.784
MUR	E	1.000	0.130	0.114	0.934	MUR	E	1.000	0.126	0.109	0.934	MUR	E	1.000	0.125	0.108	0.947
NWSA	CD	1.000	0.143	0.148	0.947	NWSA	CD	1.000	0.141	0.144	0.950	NWSA	CD	1.000	0.141	0.147	0.942
PRU	F	1.000	0.092	0.041	0.716	PRU	F	1.000	0.090	0.040	0.713	PRU	F	1.000	0.090	0.039	0.711
DGX	H	1.000	0.114	0.084	0.771	DGX	H	1.000	0.110	0.083	0.742	DGX	H	1.000	0.107	0.076	0.758
HOT	CD	1.000	0.133	0.125	0.911	HOT	CD	1.000	0.130	0.122	0.913	HOT	CD	1.000	0.131	0.125	0.918
TROW	F	1.000	0.104	0.061	0.826	TROW	F	1.000	0.101	0.057	0.818	TROW	F	1.000	0.102	0.059	0.832
THC	H	1.000	0.076	0.034	0.487	THC	H	1.000	0.076	0.034	0.497	THC	H	1.000	0.073	0.030	0.476
TIF	CD	1.000	0.122	0.097	0.924	TIF	CD	1.000	0.119	0.095	0.926	TIF	CD	1.000	0.120	0.098	0.937
UPS	I	1.000	0.143	0.148	0.966	UPS	I	1.000	0.138	0.141	0.942	UPS	I	1.000	0.140	0.145	0.963
UTX	I	1.000	0.136	0.132	0.966	UTX	I	1.000	0.133	0.128	0.968	UTX	I	1.000	0.133	0.130	0.955
WYNN	CD	1.000	0.129	0.114	0.934	WYNN	CD	1.000	0.127	0.111	0.926	WYNN	CD	1.000	0.127	0.113	0.929
PEP	CS	0.992	0.100	0.062	0.687	TYC	I	1.000	0.103	0.069	0.784	PEP	CS	1.000	0.096	0.058	0.689
WMB	E	0.989	0.126	0.106	0.904	PEP	CS	0.997	0.098	0.061	0.691	MCD	CD	0.997	0.091	0.045	0.691
MCD	CD	0.987	0.093	0.047	0.699	MCD	CD	0.992	0.091	0.046	0.706	TYC	I	0.997	0.103	0.070	0.773
TYC	I	0.984	0.106	0.072	0.794	WMB	E	0.966	0.127	0.107	0.916	TGT	CD	0.984	0.101	0.058	0.797
SYK	H	0.979	0.129	0.117	0.874	FDO	CD	0.955	0.073	0.021	0.408	WMB	E	0.958	0.123	0.104	0.915
SWN	E	0.971	0.112	0.078	0.818	SYK	H	0.939	0.124	0.110	0.866	FDO	CD	0.939	0.073	0.021	0.434
FDO	CD	0.966	0.075	0.024	0.420	TGT	CD	0.937	0.098	0.055	0.795	MRK	H	0.924	0.101	0.070	0.758
TGT	CD	0.963	0.101	0.059	0.803	SWN	E	0.926	0.110	0.074	0.807	LLTC	T	0.921	0.161	0.193	0.989
Sampling Frequency: 10 Minute																	
Stock			EN-1			Stock			EN-2			Stock			Lasso		
Ticker	Sector	Freq.	PCA	SPCA		Ticker	Sector	Freq.	PCA	SPCA		Ticker	Sector	Freq.	PCA	SPCA	
AFL	F	1.000	0.115	0.083	0.837	AFL	F	1.000	0.115	0.083	0.834	AFL	F	1.000	0.117	0.084	0.834
BBBY	CD	1.000	0.153	0.145	0.929	BBBY	CD	1.000	0.151	0.142	0.924	BBBY	CD	1.000	0.153	0.143	0.924
BBY	CD	1.000	0.138	0.120	0.876	BBY	CD	1.000	0.136	0.119	0.900	BBY	CD	1.000	0.139	0.121	0.892
CERN	T	1.000	0.157	0.153	0.871	CERN	T	1.000	0.153	0.155	0.887	CERN	T	1.000	0.160	0.158	0.895
CI	H	1.000	0.104	0.065	0.747	CI	H	1.000	0.100	0.060	0.750	CI	H	1.000	0.105	0.064	0.745
CCI	T	1.000	0.091	0.050	0.595	CCI	T	1.000	0.089	0.049	0.605	CCI	T	1.000	0.091	0.050	0.608
DHI	CD	1.000	0.108	0.068	0.747	DHI	CD	1.000	0.106	0.066	0.768	DHI	CD	1.000	0.109	0.068	0.750
DVN	E	1.000	0.130	0.108	0.863	FCX	M	1.000	0.139	0.119	0.908	FCX	M	1.000	0.142	0.122	0.897
FCX	M	1.000	0.141	0.122	0.918	FITB	F	1.000	0.086	0.035	0.529	FITB	F	1.000	0.087	0.037	0.547
FITB	F	1.000	0.087	0.037	0.542	ILMN	H	1.000	0.109	0.076	0.658	ILMN	H	1.000	0.108	0.075	0.676
ILMN	H	1.000	0.110	0.078	0.666	JNJ	H	1.000	0.114	0.084	0.771	JNJ	H	1.000	0.117	0.087	0.774
JNJ	H	1.000	0.117	0.089	0.768	LNC	F	1.000	0.110	0.071	0.808	LNC	F	1.000	0.111	0.074	0.805
LNC	F	1.000	0.111	0.074	0.813	MMC	F	1.000	0.098	0.061	0.671	MMC	F	1.000	0.100	0.061	0.671
MMC	F	1.000	0.099	0.061	0.682	OXY	E	1.000	0.125	0.096	0.892	OXY	E	1.000	0.128	0.098	0.897
OXY	E	1.000	0.127	0.098	0.900	PNC	F	1.000	0.105	0.065	0.726	PNC	F	1.000	0.106	0.067	0.734
PNC	F	1.000	0.105	0.067	0.734	PRU	F	1.000	0.098	0.052	0.745	PRU	F	1.000	0.100	0.054	0.724
PRU	F	1.000	0.099	0.054	0.724	DGX	H	1.000	0.113	0.089	0.745	DGX	H	1.000	0.116	0.093	0.729
DGX	H	1.000	0.117	0.094	0.742	SYK	H	1.000	0.131	0.114	0.818	SYK	H	1.000	0.136	0.120	0.826
SYK	H	1.000	0.134	0.118	0.842	SYMC	T	1.000	0.137	0.121	0.876	SYMC	T	1.000	0.139	0.123	0.858
SYMC	T	1.000	0.139	0.124	0.866	THC	H	1.000	0.086	0.048	0.563	THC	H	1.000	0.091	0.053	0.571
THC	H	1.000	0.089	0.052	0.555	TIF	CD	1.000	0.132	0.110	0.905	TIF	CD	1.000	0.134	0.113	0.900
TIF	CD	1.000	0.134	0.112	0.903	UPS	I	1.000	0.149	0.148	0.937	UPS	I	1.000	0.152	0.147	0.929
UPS	I	1.000	0.151	0.146	0.934	UTX	I	1.000	0.141	0.130	0.918	UTX	I	1.000	0.144	0.131	0.908
UTX	I	1.000	0.143	0.131	0.913	WYNN	CD	1.000	0.134	0.114	0.900	WYNN	CD	1.000	0.137	0.118	0.908
WYNN	CD	1.000	0.136	0.118	0.905	CNX	U	1.000	0.131	0.106	0.858	CNX	U	1.000	0.133	0.110	0.850
CNX	U	0.997	0.133	0.110	0.865	FAST	I	1.000	0.147	0.139	0.929	WHR	CD	1.000	0.157	0.156	0.929
FAST	I	0.997	0.151	0.145	0.934	NDAQ	F	1.000	0.094	0.047	0.705	FAST	I	0.997	0.152	0.146	0.945
NDAQ	F	0.997	0.096	0.049	0.715	WHR	CD	1.000	0.153	0.151	0.918	NDAQ	F	0.997	0.096	0.049	0.715
SWKS	T	0.997	0.156	0.151	0.881	SWKS	T	0.997	0.152	0.144	0.873	SWKS	T	0.992	0.154	0.147	0.881
WHR	CD	0.992	0.156	0.158	0.937	EMR	I	0.997	0.145	0.136	0.953	EMR	I	0.992	0.149	0.140	0.952
MO	CS	0.987	0.084	0.045	0.552	DVN	E	0.979	0.131	0.108	0.866	MCK	H	0.984	0.121	0.095	0.759
MCK	H	0.979	0.118	0.093	0.774	MCK	H	0.961	0.117	0.089	0.767	MO	CS	0.976	0.084	0.045	0.555
EMR	I	0.976	0.147	0.139	0.938	HRB	F	0.953	0.116	0.093	0.677	DVN	E	0.966	0.132	0.108	0.856

*Notes: Factor loadings (weights) associated with the asset return volatilities contained in our latent volatility factors are reported in this table. Recalling that factors are estimated anew, prior to the construction of each daily volatility forecast, entries in the column entitled "Freq." indicate the frequency with which a particular variable appears in volatility factors for models of the target variable for which volatility is being predicted (i.e., the SPY ETF in this table). Only stocks that are selected for use in the construction of almost all latent factors are listed. Stock tickers of selected stocks appearing in the latent factors are given in the first column of entries in the table. In the second column, these stocks are roughly categorized as belonging to one of three sectors, including Financials (F), Consumer Discretionary (CD), and Consumer Staples (CS). Entries in the columns denoted by "PCA" and "SPCA" indicate the sample averages of the weight assigned to each stock in the construction of the first principal component (i.e., volatility factor) in our prediction experiments, based on these two alternative factor estimation methods. All results are based on experiments carried out using our dataset of 267 stocks. See Section 4 for further details.

Table 24: Factor Structure (XLE)

Sampling Frequency: 1 Minute

Stock			EN-1			Stock			EN-2			Stock			Lasso		
Ticker	Sector	Freq.	PCA	SPCA		Ticker	Sector	Freq.	PCA	SPCA		Ticker	Sector	Freq.	PCA	SPCA	
AFL	F	1.000	0.133	0.109	0.961	AFL	F	1.000	0.130	0.105	0.945	AFL	F	1.000	0.130	0.104	0.953
ALL	F	1.000	0.105	0.048	0.682	ALL	F	1.000	0.102	0.045	0.684	ALL	F	1.000	0.102	0.045	0.695
ABC	H	1.000	0.130	0.113	0.771	ABC	H	1.000	0.127	0.109	0.763	ABC	H	1.000	0.129	0.109	0.776
BBY	CD	1.000	0.140	0.128	0.950	BBY	CD	1.000	0.137	0.127	0.961	BBY	CD	1.000	0.139	0.127	0.955
CNX	U	1.000	0.139	0.120	0.958	CAH	H	1.000	0.136	0.123)	0.876	CAH	H	1.000	0.137	0.123	0.863
CAH	H	1.000	0.140	0.130	0.879	CI	H	1.000	0.110	0.057	0.779	CI	H	1.000	0.110	0.057	0.795
CI	H	1.000	0.113	0.061	0.797	CAG	CS	1.000	0.111	0.071	0.618	CAG	CS	1.000	0.108	0.066	0.600
CAG	CS	1.000	0.111	0.069	0.611	FITB	F	1.000	0.088	0.023	0.471	FITB	F	1.000	0.088	0.024	0.471
FITB	F	1.000	0.090	0.024	0.474	GS	F	1.000	0.090	0.022	0.479	GS	F	1.000	0.090	0.022	0.476
GE	I	1.000	0.138	0.123	0.926	HSY	CS	1.000	0.104	0.051	0.629	HSY	CS	1.000	0.103	0.049	0.639
GS	F	1.000	0.092	0.022	0.487	ILMN	H	1.000	0.104	0.052	0.637	ILMN	H	1.000	0.105	0.056	0.642
HSY	CS	1.000	0.106	0.052	0.629	KR	CS	1.000	0.092	0.034	0.474	KR	CS	1.000	0.091	0.036	0.479
HBAN	F	1.000	0.067	0.012	0.161	LM	I	1.000	0.095	0.032	0.637	LM	I	1.000	0.095	0.032	0.621
ILMN	H	1.000	0.104	0.053	0.611	LNC	F	1.000	0.117	0.073	0.874	LNC	F	1.000	0.117	0.073	0.871
KR	CS	1.000	0.092	0.032	0.476	LLTC	T	1.000	0.178	0.222	0.997	LLTC	T	1.000	0.179	0.222	0.995
LM	I	1.000	0.097	0.033	0.621	MMC	F	1.000	0.092	0.032	0.584	MMC	F	1.000	0.092	0.033	0.592
LNC	F	1.000	0.121	0.077	0.874	MDT	H	1.000	0.140	0.129	0.918	MDT	H	1.000	0.139	0.127	0.908
LLTC	T	1.000	0.182	0.228	1.000	MET	F	1.000	0.113	0.066	0.853	MET	F	1.000	0.113	0.065	0.858
MMC	F	1.000	0.095	0.034	0.600	PRU	F	1.000	0.105	0.047	0.726	PRU	F	1.000	0.105	0.047	0.732
MDT	H	1.000	0.143	0.133	0.918	SWN	E	1.000	0.118	0.080	0.834	SWN	E	1.000	0.118	0.079	0.826
MET	F	1.000	0.116	0.070	0.837	SYK	H	1.000	0.142	0.136	0.947	TROW	F	1.000	0.114	0.066	0.897
PEP	CS	1.000	0.116	0.071	0.750	TROW	F	1.000	0.114	0.066	0.892	TIF	CD	1.000	0.130	0.107	0.968
PRU	F	1.000	0.108	0.048	0.711	TIF	CD	1.000	0.130	0.106	0.961	UNM	F	1.000	0.123	0.090	0.889
SWN	E	1.000	0.123	0.086	0.847	UNM	F	1.000	0.123	0.090	0.889	AMT	T	1.000	0.107	0.054	0.792
SYK	H	1.000	0.146	0.142	0.947	AMT	T	1.000	0.108	0.056	0.803	GRMN	CD	1.000	0.114	0.073	0.718
SYY	CS	1.000	0.110	0.061	0.711	CNX	U	0.989	0.135	0.114	0.963	SYK	H	0.995	0.142	0.134	0.939
TROW	F	1.000	0.117	0.069	0.903	PEP	CS	0.987	0.113	0.068	0.755	CA	T	0.995	0.138	0.127	0.944
TIF	CD	1.000	0.133	0.109	0.958	GRMN	CD	0.982	0.115	0.077	0.724	CNX	U	0.992	0.134	0.112	0.958
UNM	F	1.000	0.126	0.094	0.897	HBAN	F	0.979	0.065	0.010	0.164	PEP	CS	0.971	0.113	0.068	0.772
VLO	E	1.000	0.120	0.081	0.847	MON	M	0.971	0.099	0.041	0.604	MON	M	0.963	0.099	0.040	0.590
CSX	I	0.997	0.150	0.150	0.976	CA	T	0.966	0.138	0.129	0.948	AXP	F	0.961	0.124	0.091	0.953
MON	M	0.997	0.102	0.042	0.612	AXP	F	0.961	0.124	0.090	0.942	HBAN	F	0.958	0.066	0.011	0.162
AMT	T	0.995	0.110	0.057	0.794	TXN	T	0.939	0.166	0.193	0.997	TXN	T	0.945	0.167	0.194	0.992
SBUX	CD	0.958	0.156	0.159	0.978	GE	I	0.929	0.134	0.118	0.912	UNP	I	0.929	0.136	0.122	0.963
CERN	T	0.950	0.150	0.152	0.972	UTX	I	0.916	0.142	0.139	0.994	ENDP	H	0.929	0.114	0.082	0.717

Sampling Frequency: 5 Minute

Stock			EN-1			Stock			EN-2			Stock			Lasso		
Ticker	Sector	Freq.	PCA	SPCA		Ticker	Sector	Freq.	PCA	SPCA		Ticker	Sector	Freq.	PCA	SPCA	
BK	F	1.000	0.069	0.017	0.437	BK	F	1.000	0.065	0.015	0.432	BK	F	1.000	0.066	0.016	0.429
BSX	H	1.000	0.080	0.048	0.511	BSX	H	1.000	0.077	0.046	0.521	BSX	H	1.000	0.078	0.048	0.518
CNX	U	1.000	0.111	0.089	0.908	CNX	U	1.000	0.107	0.085	0.911	CNX	U	1.000	0.107	0.085	0.918
CSX	I	1.000	0.116	0.100	0.929	CSX	I	1.000	0.112	0.097	0.926	CSX	I	1.000	0.112	0.098	0.918
CAH	H	1.000	0.108	0.095	0.829	CAH	H	1.000	0.105	0.092	0.829	CAH	H	1.000	0.106	0.093	0.826
CERN	T	1.000	0.129	0.130	0.908	CERN	T	1.000	0.123	0.122	0.921	CERN	T	1.000	0.121	0.119	0.921
CHK	E	1.000	0.107	0.081	0.868	CHK	E	1.000	0.101	0.076	0.874	CI	H	1.000	0.081	0.039	0.661
CI	H	1.000	0.083	0.040	0.658	CI	H	1.000	0.080	0.039	0.661	CCI	T	1.000	0.074	0.035	0.550
CCI	T	1.000	0.078	0.037	0.579	CCI	T	1.000	0.074	0.034	0.571	DHI	CD	1.000	0.087	0.049	0.734
DHI	CD	1.000	0.090	0.052	0.771	DHI	CD	1.000	0.088	0.051	0.766	FDO	CD	1.000	0.065	0.020	0.397
FDO	CD	1.000	0.068	0.021	0.418	FDO	CD	1.000	0.065	0.019	0.424	FITB	F	1.000	0.070	0.023	0.461
FITB	F	1.000	0.073	0.024	0.474	FITB	F	1.000	0.070	0.023	0.471	HD	CD	1.000	0.100	0.074	0.868
HD	CD	1.000	0.104	0.078	0.897	HD	CD	1.000	0.099	0.073	0.879	ILMN	H	1.000	0.080	0.046	0.605
ILMN	H	1.000	0.086	0.050	0.582	ILMN	H	1.000	0.085	0.053	0.603	LLTC	T	1.000	0.148	0.176	0.992
LNC	F	1.000	0.092	0.055	0.787	LNC	F	1.000	0.089	0.053	0.784	MUR	E	1.000	0.109	0.090	0.929
LLTC	T	1.000	0.152	0.179	0.989	LLTC	T	1.000	0.149	0.178	0.992	PRU	F	1.000	0.080	0.035	0.679
MUR	E	1.000	0.115	0.098	0.921	MUR	E	1.000	0.110	0.093	0.926	SWN	E	1.000	0.096	0.066	0.826
PRU	F	1.000	0.082	0.036	0.703	PRU	F	1.000	0.079	0.034	0.700	SWKS	T	1.000	0.128	0.133	0.921
SWN	E	1.000	0.100	0.070	0.813	SWN	E	1.000	0.096	0.065	0.821	SYK	H	1.000	0.108	0.094	0.850
SWKS	T	1.000	0.131	0.134	0.932	SWKS	T	1.000	0.128	0.133	0.942	TROW	F	1.000	0.090	0.051	0.816
HOT	CD	1.000	0.121	0.113	0.929	SYK	H	1.000	0.107	0.092	0.858	TIF	CD	1.000	0.106	0.085	0.942
SYK	H	1.000	0.110	0.094	0.850	TROW	F	1.000	0.090	0.052	0.826	UTX	I	1.000	0.115	0.108	0.961
TROW	F	1.000	0.093	0.053	0.826	TIF	CD	1.000	0.106	0.086	0.937	WYNN	CD	1.000	0.113	0.099	0.921
TIF	CD	1.000	0.110	0.089	0.932	UTX	I	1.000	0.116	0.109	0.950	KEY	F	1.000	0.072	0.032	0.458
UTX	I	1.000	0.120	0.113	0.958	WYNN	CD	1.000	0.113	0.100	0.932	BBY	CD	1.000	0.107	0.091	0.911
WYNN	CD	1.000	0.118	0.103	0.929	KEY	F	1.000	0.072	0.032	0.476	CHK	E	0.997	0.101	0.076	0.868
CSCO	T	0.992	0.122	0.118	0.944	CSCO	T	0.992	0.119	0.116	0.944	LNC	F	0.997	0.090	0.053	0.802
HSY	CS	0.989	0.080	0.039	0.527	HOT	CD	0.989	0.116	0.109	0.928	CSCO	T	0.995	0.118	0.115	0.944
MCD	CD	0.989	0.083	0.040	0.670	HSY	CS	0.989	0.076	0.037	0.513	TXN	T	0.992	0.134	0.146	0.976
LMT	I	0.987	0.093	0.061	0.744	BBY	CD	0.989	0.106	0.091	0.904	HOT	CD	0.989	0.117	0.110	0.926
NFLX	T	0.979	0.120	0.116	0.820	TXN	T	0.976	0.133	0.145	0.976	HSY	CS	0.987	0.076	0.036	0.504
KEY	F	0.966	0.074	0.033	0.488	NFLX	T	0.968	0.115	0.112	0.821	NFLX	T	0.976	0.113	0.108	0.814
NWSA	CD	0.947	0.126	0.128	0.933	CTSH	T	0.963	0.097	0.065	0.866	CTSH	T	0.953	0.097	0.066	0.867
MAS	I	0.945	0.132	0.132	0.916	LMT	I	0.921	0.090	0.060	0.749	TYC	I	0.937	0.088	0.057	0.784
BHI	E	0.921	0.131	0.131	0.934	MCD	CD	0.908	0.077	0.037	0.690	ALL	F	0.903	0.075	0.034	0.653

Sampling Frequency: 10 Minute

Stock			EN-1			Stock			EN-2			Stock			Lasso		
Ticker	Sector	Freq.	PCA	SPCA		Ticker	Sector	Freq.	PCA	SPCA		Ticker	Sector	Freq.	PCA	SPCA	
AEP	U	1.000	0.105	0.068	0.639	AEP	U	1.000	0.108	0.071	0.674	AEP	U	1.000	0.110	0.075	0.679
BBBY	CD	1.000	0.166	0.160	0.937	BBBY	CD	1.000	0.170	0.165	0.937	BBBY	CD	1.000	0.167	0.162	0.934
CBS	CD	1.000	0.162	0.156	0.916	CBS	CD	1.000	0.165	0.158	0.913	CBS	CD	1.000	0.164	0.157	0.908
CNX	U	1.000	0.140	0.114	0.871	CNX	U	1.000	0.142	0.116	0.871	CNX	U	1.000	0.142	0.117	0.871
CERN	T	1.000	0.166	0.160	0.882	CERN	T	1.000	0.167	0.161	0.876	CERN	T	1.000	0.165	0.158	0.876
CI	H	1.000	0.106	0.062	0.708	CI	H	1.000	0.108	0.062	0.716	CI	H	1.000	0.108	0.064	0.724
CCI	T	1.000	0.095	0.051	0.579	CCI	T	1.000	0.096	0.051	0.608	CCI	T	1.000	0.094	0.050	0.611
DHI	CD	1.000	0.115	0.073	0.753	DHI	CD	1.000	0.117	0.075	0.761	DHI	CD	1.000	0.116	0.074	0.758
FCX	M	1.000	0.148	0.125	0.905	FCX	M	1.000	0.151	0.129	0.913	FCX	M	1.000	0.150	0.129	0.913
FITB	F	1.000	0.091	0.038	0.553	FITB	F	1.000	0.094	0.039	0.574	FITB	F	1.000	0.092	0.039	0.568
GILD	H	1.000	0.116	0.085	0.671	GS	F	1.000	0.093	0.039	0.558	GILD	H	1.000	0.116	0.084	0.679
GS	F	1.000	0.091	0.038	0.561	HRB	F	1.000	0.132	0.108	0.718	GS	F	1.000	0.092	0.039	0.571
HRB	F	1.000	0.130	0.107	0.724	HON	I	1.000	0.170	0.167	0.958	HRB	F	1.000	0.129	0.105	0.726
HON	I	1.000	0.166	0.162	0.950	ILMN	H	1.000	0.117	0.080	0.679	HON	I	1.000	0.168	0.165	0.955
ILMN	H	1.000	0.116	0.080	0.679	JNJ	H	1.000	0.124	0.094	0.776	ILMN	H	1.000	0.117	0.082	0.676
JNJ	H	1.000	0.121	0.090	0.768	MUR	E	1.000	0.146	0.123	0.934	JNJ	H	1.000	0.122	0.091	0.766
MUR	E	1.000	0.144	0.120	0.924	PNC	F	1.000	0.115	0.074	0.758	MUR	E	1.000	0.146	0.123	0.939
PNC	F	1.000	0.113	0.072	0.747	PRU	F	1.000	0.106	0.057	0.726	PNC	F	1.000	0.114	0.074	0.768
PRU	F	1.000	0.103	0.052	0.724	SHW	M	1.000	0.136	0.106	0.861	PRU	F	1.000	0.105	0.057	0.737
SHW	M	1.000	0.134	0.105	0.834	SYMC	T	1.000	0.148	0.131	0.855	SHW	M	1.000	0.135	0.106	0.855
SWKS	T	1.000	0.161	0.152	0.876	TIF	CD	1.000	0.149	0.128	0.926	SYMC	T	1.000	0.147	0.128	0.858
SYMC	T	1.000	0.147	0.128	0.861	UPS	I	1.000	0.163	0.157	0.921	TIF	CD	1.000	0.147	0.125	0.926
TSO	E	1.000	0.121	0.085	0.766	UTX	I	1.000	0.155	0.142	0.924	UPS	I	1.000	0.162	0.158	0.929
TIF	CD	1.000	0.146	0.124	0.921	WYNN	CD	1.000	0.147	0.126	0.913	UTX	I	1.000	0.154	0.142	0.929
UPS	I	1.000	0.160	0.154	0.924	AFL	F	1.000	0.125	0.091	0.839	WYNN	CD	1.000	0.145	0.123	0.911
UTX	I	1.000	0.152	0.138	0.939	GILD	H	0.997	0.119	0.088	0.678	MRK	H	1.000	0.113	0.082	0.666
WYNN	CD	1.000	0.144	0.123	0.908	EBAY	T	0.997	0.160	0.148	0.913	AFL	F	1.000	0.123	0.090	0.832
EBAY	T	0.997	0.156	0.143	0.905	MCK	H	0.995	0.126	0.100	0.743	EBAY	T	0.997	0.157	0.145	0.910
BBY	CD	0.995	0.150	0.134	0.897	MRK	H	0.995	0.114	0.083	0.664	BBY	CD	0.997	0.152	0.137	0.902
FAST	I	0.995	0.162	0.155	0.952	BBY	CD	0.992	0.154	0.139	0.891	MCK	H	0.984	0.127	0.103	0.754
GOOG	T	0.987	0.125	0.091	0.816	FAST	I	0.992	0.164	0.157	0.952	FAST	I	0.976	0.163	0.157	0.957
MCK	H	0.987	0.124	0.097	0.747	GOOG	T	0.989	0.127	0.093	0.814	COH	CD	0.968	0.137	0.111	0.864
SYK	H	0.979	0.141	0.124	0.812	SYK	H	0.971	0.144	0.127	0.808	NFLX	T	0.955	0.156	0.158	0.818
MRK	H	0.866	0.114	0.085	0.699	TSO	E	0.968	0.124	0.088	0.780	NDAQ	F	0.905	0.100	0.053	0.709

*Notes: See notes to Table 23.

Table 25: Factor Structure (IBM)

Sampling Frequency: 1 Minute

Stock			EN-1			Stock			EN-2			Stock			Lasso		
Ticker	Sector	Freq.	PCA	SPCA		Ticker	Sector	Freq.	PCA	SPCA		Ticker	Sector	Freq.	PCA	SPCA	
ADM	CS	1.000	0.103	0.040	0.587	AFL	F	1.000	0.140	0.112	0.961	AFL	F	1.000	0.141	0.113	0.961
AFL	F	1.000	0.137	0.110	0.966	ALL	F	1.000	0.111	0.051	0.726	ALL	F	1.000	0.111	0.051	0.726
ALL	F	1.000	0.108	0.049	0.721	AMT	T	1.000	0.117	0.062	0.824	AMT	T	1.000	0.118	0.063	0.808
AMT	T	1.000	0.114	0.062	0.829	ABC	H	1.000	0.137	0.118	0.755	ABC	H	1.000	0.138	0.117	0.758
ABC	H	1.000	0.133	0.115	0.761	AMGN	H	1.000	0.137	0.107	0.834	AMGN	H	1.000	0.138	0.109	0.826
AMGN	H	1.000	0.133	0.107	0.832	BBY	H	1.000	0.150	0.137	0.955	BBY	H	1.000	0.151	0.139	0.958
BBY	H	1.000	0.146	0.137	0.963	CPB	CS	1.000	0.111	0.058	0.666	CPB	CS	1.000	0.112	0.058	0.653
CPB	CS	1.000	0.109	0.058	0.674	CAH	H	1.000	0.146	0.131	0.861	CAH	H	1.000	0.148	0.132	0.855
CAH	H	1.000	0.142	0.129	0.874	CI	H	1.000	0.118	0.063	0.803	CI	H	1.000	0.118	0.062	0.800
CI	H	1.000	0.115	0.061	0.824	CAG	CS	1.000	0.120	0.080	0.632	CAG	CS	1.000	0.121	0.081	0.634
CAG	CS	1.000	0.117	0.077	0.624	GE	I	1.000	0.146	0.129	0.937	GE	I	1.000	0.146	0.130	0.929
GE	I	1.000	0.142	0.125	0.934	GRMN	CD	1.000	0.126	0.085	0.750	GRMN	CD	1.000	0.125	0.085	0.774
GRMN	CD	1.000	0.122	0.082	0.742	HBAN	F	1.000	0.070	0.012	0.189	HBAN	F	1.000	0.071	0.013	0.189
HBAN	F	1.000	0.070	0.013	0.176	ILMN	H	1.000	0.115	0.063	0.629	ILMN	H	1.000	0.116	0.064	0.642
ILMN	H	1.000	0.113	0.063	0.653	KR	CS	1.000	0.099	0.040	0.495	KR	CS	1.000	0.101	0.042	0.497
KR	CS	1.000	0.097	0.039	0.518	LM	F	1.000	0.103	0.036	0.639	LM	F	1.000	0.104	0.037	0.663
LM	F	1.000	0.101	0.037	0.684	LNC	F	1.000	0.127	0.081	0.879	LNC	F	1.000	0.128	0.081	0.892
LNC	F	1.000	0.124	0.078	0.879	MMC	F	1.000	0.100	0.037	0.629	MMC	F	1.000	0.101	0.038	0.613
MMC	F	1.000	0.098	0.036	0.621	MON	M	1.000	0.107	0.045	0.616	MON	M	1.000	0.108	0.045	0.626
MON	M	1.000	0.104	0.045	0.618	PEP	CS	1.000	0.122	0.075	0.771	PEP	CS	1.000	0.124	0.077	0.779
PEP	CS	1.000	0.120	0.076	0.789	PRU	F	1.000	0.113	0.052	0.784	PRU	F	1.000	0.114	0.052	0.782
PRU	F	1.000	0.111	0.051	0.795	SWN	E	1.000	0.129	0.091	0.855	SWN	E	1.000	0.130	0.091	0.858
SWN	E	1.000	0.126	0.089	0.855	SBUX	CD	1.000	0.165	0.168	0.989	SBUX	CD	1.000	0.167	0.169	0.989
SBUX	CD	1.000	0.161	0.165	0.992	TIF	CD	1.000	0.142	0.117	0.974	TIF	CD	1.000	0.143	0.117	0.976
TIF	CD	1.000	0.138	0.116	0.971	UNP	I	1.000	0.150	0.135	0.968	UNP	I	1.000	0.151	0.137	0.963
UNP	I	1.000	0.147	0.135	0.963	UNM	F	1.000	0.133	0.098	0.903	UNM	F	1.000	0.134	0.099	0.900
UNM	F	1.000	0.130	0.096	0.897	LLTC	T	1.000	0.196	0.245	0.995	LLTC	T	1.000	0.198	0.246	0.995
LLTC	T	0.997	0.191	0.239	0.997	EXPE	CD	1.000	0.180	0.204	0.989	EXPE	CD	1.000	0.181	0.206	0.992
CNX	U	0.989	0.143	0.123	0.968	CNX	U	0.984	0.147	0.127	0.965	CNX	U	0.987	0.147	0.126	0.971
SCHW	F	0.987	0.095	0.031	0.496	ADM	CS	0.945	0.106	0.042	0.604	TXN	T	0.916	0.186	0.220	0.994
GS	F	0.966	0.096	0.026	0.529	TXN	T	0.926	0.184	0.220	0.994	AXP	F	0.871	0.138	0.105	0.964
HOT	CD	0.945	0.154	0.149	0.964	AXP	F	0.861	0.136	0.103	0.969	GS	F	0.853	0.094	0.020	0.485
EXPE	CD	0.937	0.175	0.201	0.994	GS	F	0.845	0.096	0.023	0.526	ENDP	H	0.845	0.129	0.099	0.729
TXN	T	0.911	0.178	0.208	0.997	HOT	CD	0.834	0.159	0.152	0.962	HOT	CD	0.808	0.161	0.154	0.967

Sampling Frequency: 5 Minute

Stock			EN-1			Stock			EN-2			Stock			Lasso		
Ticker	Sector	Freq.	PCA	SPCA		Ticker	Sector	Freq.	PCA	SPCA		Ticker	Sector	Freq.	PCA	SPCA	
ALL	F	1.000	0.105	0.050	0.684	ALL	F	1.000	0.106	0.050	0.708	ALL	F	1.000	0.106	0.051	0.700
BK	F	1.000	0.090	0.025	0.508	BK	F	1.000	0.091	0.025	0.492	BK	F	1.000	0.091	0.024	0.487
BBY	CD	1.000	0.147	0.130	0.937	BBY	CD	1.000	0.149	0.132	0.939	BBY	CD	1.000	0.150	0.132	0.945
CNX	U	1.000	0.141	0.114	0.908	CNX	U	1.000	0.142	0.113	0.905	CNX	U	1.000	0.142	0.113	0.892
CSX	I	1.000	0.149	0.132	0.942	CSX	I	1.000	0.151	0.133	0.945	CSX	I	1.000	0.150	0.131	0.942
CPB	CS	1.000	0.110	0.066	0.618	CAH	H	1.000	0.147	0.130	0.855	CAH	H	1.000	0.147	0.131	0.842
CAH	H	1.000	0.148	0.135	0.847	CERN	T	1.000	0.169	0.173	0.924	CERN	T	1.000	0.169	0.172	0.913
CERN	T	1.000	0.165	0.169	0.903	CI	H	1.000	0.113	0.063	0.705	CI	H	1.000	0.113	0.063	0.718
CI	H	1.000	0.111	0.062	0.716	CSCO	T	1.000	0.161	0.157	0.942	CSCO	T	1.000	0.161	0.156	0.953
CSCO	T	1.000	0.159	0.157	0.945	CCI	T	1.000	0.100	0.050	0.632	CCI	T	1.000	0.100	0.049	0.624
CCI	T	1.000	0.098	0.048	0.621	DHI	CD	1.000	0.120	0.074	0.797	DHI	CD	1.000	0.120	0.074	0.795
DHI	CD	1.000	0.117	0.072	0.800	HSY	CS	1.000	0.111	0.064	0.600	HSY	CS	1.000	0.112	0.065	0.600
HSY	CS	1.000	0.110	0.064	0.611	JPM	F	1.000	0.113	0.060	0.768	JPM	F	1.000	0.113	0.060	0.776
JPM	F	1.000	0.112	0.058	0.782	LNC	F	1.000	0.121	0.075	0.803	LNC	F	1.000	0.121	0.076	0.824
LNC	F	1.000	0.119	0.074	0.808	NWSA	CD	1.000	0.168	0.174	0.947	NWSA	CD	1.000	0.167	0.172	0.945
NWSA	CD	1.000	0.164	0.171	0.950	PRU	F	1.000	0.108	0.050	0.763	PRU	F	1.000	0.108	0.050	0.750
PRU	F	1.000	0.106	0.049	0.753	DGX	H	1.000	0.130	0.093	0.768	DGX	H	1.000	0.129	0.093	0.766
DGX	H	1.000	0.128	0.095	0.755	SWN	E	1.000	0.126	0.084	0.824	SWN	E	1.000	0.126	0.084	0.824
SWN	E	1.000	0.126	0.086	0.837	SWKS	T	1.000	0.173	0.177	0.913	SWKS	T	1.000	0.171	0.175	0.916
SWKS	T	1.000	0.170	0.177	0.921	HOT	CD	1.000	0.157	0.150	0.929	HOT	CD	1.000	0.157	0.149	0.929
HOT	CD	1.000	0.154	0.148	0.924	TROW	F	1.000	0.122	0.072	0.850	TROW	F	1.000	0.121	0.071	0.858
TROW	F	1.000	0.120	0.072	0.847	TGT	CD	1.000	0.122	0.077	0.832	TGT	CD	1.000	0.123	0.077	0.821
TGT	CD	1.000	0.119	0.074	0.826	TIF	CD	1.000	0.145	0.120	0.942	TIF	CD	1.000	0.145	0.120	0.937
TIF	CD	1.000	0.142	0.118	0.950	WYNN	CD	1.000	0.154	0.137	0.937	WYNN	CD	1.000	0.153	0.136	0.939
TYC	I	1.000	0.121	0.082	0.811	CPB	CS	0.995	0.111	0.066	0.640	LOW	CD	0.997	0.145	0.124	0.916
WYNN	CD	1.000	0.151	0.135	0.937	TYC	I	0.995	0.122	0.082	0.804	TYC	I	0.995	0.123	0.082	0.825
XRX	T	0.992	0.153	0.154	0.836	LOW	CD	0.995	0.145	0.123	0.907	CPB	CS	0.992	0.112	0.068	0.629
LMT	I	0.987	0.121	0.085	0.787	THC	H	0.979	0.093	0.046	0.522	THC	H	0.987	0.093	0.045	0.541
ADP	T	0.982	0.157	0.151	0.965	ADP	T	0.961	0.160	0.152	0.964	ADP	T	0.961	0.159	0.151	0.970
THC	H	0.976	0.089	0.040	0.488	LMT	I	0.947	0.122	0.082	0.789	TXN	T	0.958	0.179	0.191	0.970
SCHW	F	0.968	0.101	0.046	0.603	TXN	T	0.947	0.179	0.192	0.967	MAS	I	0.953	0.173	0.177	0.928
MAS	I	0.958	0.170	0.177	0.934	MAS	I	0.942	0.172	0.177	0.922	BSX	H	0.926	0.109	0.066	0.531
LOW	CD	0.911	0.140	0.118	0.893	BSX	H	0.897	0.106	0.064	0.525	XRX	T	0.882	0.152	0.148	0.830

Sampling Frequency: 10 Minute

Stock			EN-1			Stock			EN-2			Stock			Lasso		
Ticker	Sector	Freq.	PCA	SPCA		Ticker	Sector	Freq.	PCA	SPCA		Ticker	Sector	Freq.	PCA	SPCA	
AFL	F	1.000	0.127	0.095	0.868	AFL	F	1.000	0.126	0.094	0.847	AFL	F	1.000	0.126	0.095	0.853
BK	F	1.000	0.089	0.031	0.532	BK	F	1.000	0.087	0.031	0.534	BK	F	1.000	0.088	0.031	0.529
BBY	CD	1.000	0.153	0.138	0.882	BBY	CD	1.000	0.152	0.135	0.882	BBY	CD	1.000	0.150	0.136	0.887
CNX	U	1.000	0.144	0.121	0.889	CNX	U	1.000	0.143	0.120	0.887	CNX	U	1.000	0.143	0.120	0.884
CAT	I	1.000	0.187	0.198	0.958	CAT	I	1.000	0.185	0.197	0.950	CAT	I	1.000	0.185	0.197	0.953
CERN	T	1.000	0.167	0.162	0.874	CERN	T	1.000	0.163	0.159	0.876	CERN	T	1.000	0.161	0.158	0.874
CI	H	1.000	0.118	0.078	0.745	CL	CS	1.000	0.104	0.064	0.597	CL	CS	1.000	0.103	0.068	0.592
CL	CS	1.000	0.099	0.062	0.597	CCI	T	1.000	0.097	0.052	0.629	CCI	T	1.000	0.097	0.052	0.629
CCI	T	1.000	0.095	0.052	0.618	DHI	CD	1.000	0.116	0.076	0.797	DHI	CD	1.000	0.116	0.076	0.789
DHI	CD	1.000	0.117	0.076	0.795	DVN	E	1.000	0.142	0.122	0.887	DVN	E	1.000	0.142	0.122	0.884
DVN	E	1.000	0.144	0.123	0.876	FCX	M	1.000	0.152	0.134	0.932	FCX	M	1.000	0.153	0.134	0.921
FCX	M	1.000	0.154	0.134	0.934	FAST	I	1.000	0.160	0.153	0.932	FAST	I	1.000	0.159	0.153	0.929
FAST	I	1.000	0.163	0.156	0.924	HUM	H	1.000	0.113	0.074	0.658	HUM	H	1.000	0.110	0.073	0.663
HUM	H	1.000	0.111	0.072	0.645	ILMN	H	1.000	0.109	0.078	0.611	ILMN	H	1.000	0.113	0.079	0.626
ILMN	H	1.000	0.114	0.080	0.605	JNJ	H	1.000	0.123	0.093	0.784	JNJ	H	1.000	0.123	0.094	0.789
JNJ	H	1.000	0.124	0.093	0.774	LM	F	1.000	0.105	0.060	0.761	LM	F	1.000	0.106	0.061	0.768
LM	F	1.000	0.106	0.060	0.774	PGR	F	1.000	0.111	0.067	0.758	PGR	F	1.000	0.112	0.068	0.758
MON	M	1.000	0.107	0.063	0.687	PRU	F	1.000	0.106	0.058	0.771	PRU	F	1.000	0.106	0.057	0.768
PGR	F	1.000	0.110	0.067	0.747	SWKS	T	1.000	0.162	0.150	0.871	SWKS	T	1.000	0.162	0.151	0.876
PRU	F	1.000	0.106	0.057	0.763	TSO	E	1.000	0.122	0.087	0.808	TSO	E	1.000	0.121	0.088	0.808
ROST	CD	1.000	0.135	0.109	0.847	TIF	CD	1.000	0.144	0.123	0.897	TIF	CD	1.000	0.144	0.123	0.895
SWKS	T	1.000	0.161	0.152	0.861	UPS	I	1.000	0.164	0.160	0.932	UPS	I	1.000	0.163	0.160	0.932
TSO	E	1.000	0.124	0.089	0.795	UTX	I	1.000	0.153	0.142	0.911	UTX	I	1.000	0.153	0.143	0.916
TIF	CD	1.000	0.146	0.125	0.905	WHR	CD	1.000	0.169	0.175	0.921	WHR	CD	1.000	0.168	0.174	0.921
UPS	I	1.000	0.165	0.162	0.942	MON	M	0.997	0.106	0.062	0.665	MON	M	0.997	0.105	0.062	0.675
UTX	I	1.000	0.154	0.143	0.916	WYNN	CD	0.997	0.146	0.130	0.897	WYNN	CD	0.997	0.146	0.129	0.889
WHR	CD	1.000	0.171	0.176	0.916	FISV	T	0.987	0.139	0.112	0.904	FISV	T	0.987	0.139	0.113	0.901
WYNN	CD	0.997	0.149	0.131	0.894	OXY	E	0.982	0.136	0.107	0.895	OXY	E	0.974	0.135	0.106	0.900
OXY	E	0.995	0.137	0.108	0.902	AET	H	0.976	0.115	0.079	0.674	ROST	CD	0.958	0.136	0.109	0.835
HD	CD	0.982	0.132	0.102	0.861	ROST	CD	0.963	0.136	0.107	0.839	HRB	F	0.958	0.129	0.102	0.692
AET	H	0.979	0.113	0.079	0.677	HRB	F	0.961	0.130	0.104	0.701	AET	H	0.942	0.112	0.079	0.679
USB	F	0.974	0.116	0.075	0.792	HD	CD	0.958	0.132	0.101	0.843	USB	F	0.929	0.117	0.076	0.779
HRB	F	0.963	0.132	0.105	0.699	CSX	I	0.939	0.154	0.135	0.910	FITB	F	0.918	0.094	0.044	0.593
FISV	T	0.942	0.141	0.116	0.902	USB	F	0.937	0.118	0.076	0.784	HD	CD	0.911	0.130	0.099	0.847

*Notes: See notes to Table 23.

References

AIOLFI, M., RODRIGUEZ, M., AND TIMMERMANN, A. 2009. Understanding analysts earnings expectations: Biases, nonlinearities, and predictability. *Journal of Financial Econometrics* 8:305–334.

AÏT-SAHALIA, Y. AND JACOD, J. 2012. Analyzing the spectrum of asset returns: Jump and volatility components in high frequency data. *Journal of Economic Literature* 50:1007–1050.

AÏT-SAHALIA, Y. AND JACOD, J. 2014. High-frequency financial econometrics. Princeton University Press: Princeton, NJ, USA.

AÏT-SAHALIA, Y., MYKLAND, P. A., AND ZHANG, L. 2011. Ultra high frequency volatility estimation with dependent microstructure noise. *Journal of Econometrics* 160:160–175.

AÏT-SAHALIA, Y. AND XIU, D. 2016a. Principal component analysis of high frequency data. *Chicago Booth Research Paper No. 15-39* .

AÏT-SAHALIA, Y. AND XIU, D. 2016b. Using principal component analysis to estimate a high dimensional factor model with high-frequency data. *Journal of Econometrics, forthcoming* .

ANDERSEN, T., BOLLERSLEV, T., DIEBOLD, F. X., AND LABYS, P. 2001. The distribution of realized exchange rate volatility. *Journal of the American Statistical Association* 96:42–55.

ANDERSEN, T. G., BOLLERSLEV, T., AND DIEBOLD, F. X. 2007. Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. *Review of Economics and Statistics* 89:701–720.

ANDERSEN, T. G., BOLLERSLEV, T., DIEBOLD, F. X., AND LABYS, P. 2003. Modeling and forecasting realized volatility. *Econometrica* 71:579–625.

ANDERSEN, T. G., BOLLERSLEV, T., AND MEDDAHI, N. 2004. Analytical evaluation of volatility forecasts. *International Economic Review* 45:1079–1110.

ANDERSEN, T. G., BOLLERSLEV, T., AND MEDDAHI, N. 2011. Realized volatility forecasting and market microstructure noise. *Journal of Econometrics* 160:220–234.

ANG, A. AND TIMMERMANN, A. 2012. Regime changes and financial markets. *Annu. Rev. Financ. Econ.* 4:313–337.

AUDRINO, F. AND HU, Y. 2016. Volatility forecasting: Downside risk, jumps and leverage effect. *Econometrics* 4:1–24.

BAI, J. AND NG, S. 2006a. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica* 74:1133–1150.

BAI, J. AND NG, S. 2006b. Evaluating latent and observed factors in macroeconomics and finance. *Journal of Econometrics* 131:507–537.

BAI, J. AND NG, S. 2008. Forecasting economic time series using targeted predictors. *Journal of Econometrics* 146:304–317.

BARNDORFF-NIELSEN, O. E., GRAVERSEN, S. E., JACOD, J., AND SHEPHARD, N. 2006. Limit theorems for bipower variation in financial econometrics. *Econometric Theory* 22:677–719.

BARNDORFF-NIELSEN, O. E. AND SHEPHARD, N. 2004. Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics* 2:1–37.

BOLLERSLEV, T., PATTON, A. J., AND QUAEDVLIEG, R. 2016. Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics* 192:1–18.

BRANDT, M. W. AND JONES, C. S. 2006. Volatility forecasting with range-based egarch models. *Journal of Business & Economic Statistics* 24:470–486.

CAMPBELL, J. Y. AND THOMPSON, S. B. 2008. Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies* 21:1509–1531.

CARRASCO, M. AND ROSSI, B. 2016. In-sample inference and forecasting in misspecified factor models. *Journal of Business and Economic Statistics* 34:313–338.

CORSI, F. 2009. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics* 7:174–196.

CORSI, F., PIRINO, D., AND RENO, R. 2010. Threshold bipower variation and the impact of jumps on volatility forecasting. *Journal of Econometrics* 159:276–288.

DUONG, D. AND SWANSON, N. R. 2015. Empirical evidence on the importance of aggregation, asymmetry, and jumps for volatility prediction. *Journal of Econometrics* 187:606–621.

GHYSELS, E., SANTA-CLARA, P., AND VALKANOV, R. 2006. Predicting volatility: getting the most out of return data sampled at different frequencies. *Journal of Econometrics* 131:59–95.

GHYSELS, E. AND SINKO, A. 2011. Volatility forecasting and microstructure noise. *Journal of Econometrics* 160:257–271.

HANSEN, P. R. AND LUNDE, A. 2005. A forecast comparison of volatility models: does anything beat a garch (1, 1)? *Journal of applied econometrics* 20:873–889.

JACOD, J., LI, Y., MYKLAND, P. A., PODOLSKIJ, M., AND VETTER, M. 2009. Microstructure noise in the continuous case: the pre-averaging approach. *Stochastic processes and their applications* 119:2249–2276.

JACOD, J. AND ROSENBAUM, M. 2013. Quarticity and other functionals of volatility: Efficient estimation. *The Annals of Statistics* 41:1462–1484.

KIM, H. H. AND SWANSON, N. R. 2017. Mining big data using parsimonious factor, machine learning, variable selection, and shrinkage methods. *International Journal of Forecasting, forthcoming* .

MANCINI, C. 2001. Disentangling the jumps of the diffusion in a geometric jumping brownian motion. *Giornale dell'Istituto Italiano degli Attuari* LXIV:19–47.

MANCINI, C. 2009. Non-parametric threshold estimation for models with stochastic diffusion coefficient and jumps. *Scandinavian Journal of Statistics* 36:270–296.

MEDDAHI, N. ET AL. 2001. An eigenfunction approach for volatility modeling. CIRANO.

PATTON, A. J. AND SHEPPARD, K. 2015. Good volatility, bad volatility: Signed jumps and the persistence of volatility. *Review of Economics and Statistics* 97:683–697.

PAYE, B. S. AND TIMMERMANN, A. 2006. Instability of return prediction models. *Journal of Empirical Finance* 13:274–315.

PODOLSKIJ, M. AND ZIGGEL, D. 2010. New tests for jumps in semimartingale models. *Statistical inference for stochastic processes* 13:15–41.

QI, X., LUO, R., AND ZHAO, H. 2013. Sparse principal component analysis by choice of norm. *Journal of Multivariate Analysis* 114:127–160.

STOCK, J. H. AND WATSON, M. W. 2002a. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97:1167–1179.

STOCK, J. H. AND WATSON, M. W. 2002b. Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics* 20:147–162.

STOCK, J. H. AND WATSON, M. W. 2006. Forecasting with many predictors. *Handbook of Economic Forecasting* 1:515–554.

SWANSON, N. R. AND XIONG, W. 2017. Big data analytics in economics: what have we learned so far, and where should we go from here? *Working Paper, Rutgers University* .

TIBSHIRANI, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

ZOU, H. AND HASTIE, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67:301–320.

ZOU, H., HASTIE, T., AND TIBSHIRANI, R. 2006. Sparse principal component analysis. *Journal of Computational and Graphical Statistics* 15:265–286.