Forecasting Volatility Using Double Shrinkage Methods*

Mingmian Cheng[1], Norman R. Swanson[2] and Xiye Yang[2]
[1]Sun Yat-sen University and [2]Rutgers University

March 2019

Abstract

In this paper, we propose and evaluate a shrinkage based methodology that is designed to improve the accuracy of forecasts of daily integrated volatility. Our approach is based on a two-step shrinkage procedure designed to extract latent common volatility factors from a large dimensional and high-frequency asset returns dataset. In the first step, we apply either LASSO or elastic net shrinkage on estimated integrated volatilities, in order to select a subset of assets that are informative about our target asset. In the second step, we utilize (sparse) principal component analysis on the selected assets, in order to estimate a latent return factor. This new factor is in turn utilized to construct a latent volatility factor. Although we find limited *in-sample* fit improvement, relative to various benchmark models currently used in the literature, all of our proposed factor-augmented forecasting models result in substantial predictive gains, as measured by *out-of-sample* R^2, and via the application of predictive accuracy tests. In particular, forecasting gains are observed at individual firm, sector, and market levels. Additionally, our empirical findings suggest that the first step of our procedure, which utilizes shrinkage, plays a crucial role in the success of our method, and the second step of our procedure also relies on shrinkage (via the use of SPCA) for optimal predictive performance.

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 Finance, Lingnan (University) College, Sun Yat-sen University, 135 Xingang West Road, Guangzhou, 510275, China, chengmm3@mail.sysu.edu.cn; 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@economics.rutgers.edu. The authors are grateful to Yacine Aït-Sahalia, Jianqing Fan, Guanhao (Gavin) Feng, Yuan Liao, Bruce Mizrach, Markus Pelger, Dacheng Xiu, and seminar participants at the 2017 China Meeting of the Econometric Society, the 2017 Canadian Economic Association meetings, the 23rd International Conference on Computing in Economics and Finance, and the 2nd International Conference on Econometrics and Statistics for comments that have been utilized in the preparation of this paper.

1 Introduction

Accurate volatility estimation and prediction is crucial to successful risk management and asset allocation. In light of this fact, it is not surprising that the seminal contribution of Jacod (2018), originally published as a 1994 working paper, has spurred the development of a veritable arsenal of realized integrated volatility (IV) estimators. A very few of these include 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 microstructure noise robust RV based on multi-scale variation and pre-averaging (Jacod et al. (2009) and Aït-Sahalia et al. (2011)). One important use of these sorts of IV estimators is in heterogeneous autoregressive (HAR) type forecasting models, as introduced in Corsi (2009), and built on by Andersen et al. (2007), Corsi et al. (2010), Duong and Swanson (2015), and Patton and Sheppard (2015), who augment the basic HAR model by the inclusion of a variety of jump variation related variables.[1] Although very parsimonious, the HAR-type models analyzed in the above papers only assess whether information derived from the target asset is useful for IV prediction. A less explored question is whether there are useful sources of information other than the target asset itself.

In this paper, we undertake to answer the above question by proposing and analyzing a shrinkage based methodology for extracting potentially useful predictive information from a large dimensional and high-frequency asset returns dataset. The methodology that we develop hinges on the construction of latent firm, sector, and industry level IV factors, which are used to augment standard HAR prediction models. The procedure that we propose in order to accomplish this is a two-step "double shrinkage" procedure. In the first step, we apply either LASSO or elastic net shrinkage on estimated integrated volatilities, in order to select a subset of assets that are informative about our target asset. In the second step, we apply either principal component analysis (PCA) or sparse principal component analysis (SPCA) on the selected assets, in order to estimate a latent return factor. This new factor is in turn utilized to construct a latent IV factor. It is

[1] More recently, Audrino and Hu (2016) explore the importance of leverage and downside risk in an HAR forecasting framework, and Bollerslev et al. (2016) further improve volatility forecasting by allowing HAR type model coefficients to evolve according to the degree of measurement error. Overall, there is a very deep and rich literature using HAR models for forecasting, and many key papers are cited in the above references. Other volatility forecasting models that are widely used in the financial econometrics literature include stochastic volatility (SV) models (see e.g., Meddahi (2001), Andersen et al. (2004), Andersen et al. (2011)), (G)ARCH-type models (see e.g., Andersen et al. (2003), Hansen and Lunde (2005), Brandt and Jones (2006)), and Mixed Data Sampling (MIDAS) models (see e.g., Ghysels et al. (2006), Ghysels and Sinko (2011)).

important to note that in the case where the second step involves using SPCA, we are essentially performing double shrinkage. In a first step we shrink the set of IV estimators, in order to "pare down" our original dataset. In the second step, we shrink the set of asset returns selected in the first step. One can immediately see that our second step indeed involves shrinkage by noting, as discussed in Zou et al. (2006), that SPCA can be interpreted as a variant of PCA, where the regression coefficients associated with interpreting PCA as a penalized regression problem are "shrunk" by imposing LASSO or elastic net constraints on them.[2] It is in this sense that our procedure involves "double shrinkage". Additionally, it is worth pointing out that the first step of our approach builds on methods developed in Bai and Ng (2008) in which "targeted predictors" are selected before the estimation of common factors; while the second step builds on the recent generalization of PCA to high-frequency data by Aït-Sahalia and Xiu (2017, 2018).

Interestingly, various alternative approaches to the construction of latent IV factors for inclusion in HAR forecasting models yield inferior results to those found using our procedure. In particular, one might imagine that a simpler procedure in which IV factors are directly constructed via PCA or SPCA on firm specific IV estimators will yield comparable predictions to ours. One might also surmise that using IV estimators constructed directly from S&P500 returns, or from sector specific return indices, or both, will yield comparable predictions. Finally, one might argue that simple HAR models utilizing only "own-asset" information will yield comparable predictions. This is not the case, however. Instead, our double shrinkage procedure yields significantly more accurate predictions than all of the above benchmark models. The reason for this finding appears to rest to a great extent on the importance of shrinkage. Namely, by first shrinking on IV in order to select a subset of "information rich" assets for use in our procedure, we are avoiding substantial information loss due to data noisiness and multicollinearity across asset returns and IVs.

Our empirical analysis points to a number of interesting findings. First, we show that for virtually all target assets (at firm, sector and market levels), except for the financial sector, gains associated with using our latent IV factors in HAR model prediction range from approximately 20% to 45%, compared with the best performing

[2]PCA has been extensively studied in the literature (see e.g., Stock and Watson (2002a,b, 2006), Bai and Ng (2006a,b, 2008), and the references cited therein). Additionally, the importance of targeting when using PCA and related dimension reduction methods in forecasting is discussed in Bai and Ng (2008), Carrasco and Rossi (2016), Swanson and Xiong (2018), and the references cited therein.

benchmark models, when comparing *out-of-sample* R^2 values.[3] These findings are robust to the use of different data frequencies (2.5-, 5-, and 10-minute), and hold over the daily forecasting sample period from June 1, 2009–December 31, 2010. Additionally, predictive accuracy tests comparing mean square forecast errors (MSFEs) indicate that *out-of-sample* predictive improvements are statistically significant. Second, we find that implementing our procedure using SPCA in the second step yields prediction models that dominate those associated with the use of PCA. For instance, when comparing predictions of Chevron IV using factors constructed using 2.5- and 5-minute frequency data, factor-augmented models with SPCA have 13%–18% larger *out-of-sample* R^2 values than those associated with the use of PCA.[4] The difference is even larger when predicting IV using 10-minute frequency data, although predictions for this frequency are worse, overall, than when predicting using higher frequency data. Taken together, these two findings underscore the importance of shrinkage in volatility prediction, particularly given that SPCA can be equated with the application of PCA (dimension reduction), followed by either LASSO or elastic net shrinkage on the weights estimated in the PCA step, as discussed above.

The prediction experiments that we conduct in the sequel also shed light on several other issues. For instance, we find that *in-sample* fit does not improve when our two-step procedure is used to extract IV factors, as compared with simpler benchmark modeling approaches, such as those discussed above. Hence, *in-sample* fit is not improved when IV factors are included in prediction models. Instead benefits accrue only when constructing ex ante predictions. We also find that the "best" IV predictions are associated with the use of a 5-minute sampling frequency, when comparing 2.5-, 5-, and 10-minute frequencies. A possible explanation for this is a trade-off between the effects of microstructure noise and jumps is that higher (lower) sampling frequencies may be associated with more microstructure noise (jumps), which contain less predictable content.

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 introduces our experimental setup, and includes a description of our forecasting models and the

[3] Our benchmark models include a variety of alternative specifications, including the models discussed in the previous paragraph.

[4] Similar findings characterize all of the firms and sectors examined in the sequel.

statistics used to analyze the models. Finally, Section 5 includes a discussion of the data used in our forecasting experiments, Section 6 summarizes our key empirical findings. and concluding remarks are contained in Section 7.

2 Setup

Consider a d-dimensional process, X, consisting of d asset log-prices. Following the high-frequency econometrics literature, assume that X follows an Itô-semimartingale defined on the 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 and σ_t are adapted, càdlàg, and locally bounded. Additionally, W_t is a multidimensional standard Brownian motion, μ is a random jump measure with compensator ν, and $\epsilon > 0$ is an arbitrary threshold. For more details on Itô-semimartingales and continuous-time asset price modeling, see Aït-Sahalia and Jacod (2014) and the references cited therein.

Various realized measures have been invented to estimate latent volatility on a fixed interval $[0, T]$, using high-frequency intraday data. For instance, realized volatility, one of the most widely known measures, is given by:

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

where $\lfloor \cdot \rfloor$ is the floor function, $X^\intercal$ denotes the transpose of X, and $\Delta_i^n X = X_{i\Delta_n} - X_{(i-1)\Delta_n}$ is the i^{th} intraday return, with Δ_n defined as an 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:

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

as $\Delta_n \to 0$, where $\sigma_s \sigma_s^\intercal$ will be referred to as the spot volatility matrix. However, when asset prices are discontinuous on $[0, T]$:

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

where the j-the element of ΔX_s is $\Delta X_s^j := X_s^j - X_{s-}^j$, which is non-zero if and only if the j^{th} asset, X^j, jumps at time s.

To separate integrated volatility from jump variation, one can use the threshold technique developed by Mancini (2001, 2009), to construct truncated realized volatility (TRV), defined as:

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

for some $\varpi \in (0, 1/2)$. The vector of truncation levels c is often set to be a multiplier of the volatility. To determine c_j, one can use the multipower variation (MPV) estimator developed by Barndorff-Nielsen and Shephard (2004) and Barndorff-Nielsen et al. (2006), as an initial estimator for the individual volatility. The general formula for the MPV estimator is given by:

$$\mathrm{MPV}_j(p^+) = \frac{\Delta_n^{1-p^+/2}}{m_{p_1} \cdots m_{p_k}} \sum_{i=1}^{\lfloor t/\Delta_n \rfloor - k + 1} |\Delta_i^n X^j|^{p_1} \cdots |\Delta_{i+k-1}^n X^j|^{p_k} \xrightarrow{\mathbb{P}} \int_0^t |\sigma_s^j|^{p^+} ds, \tag{6}$$

with $p_1, p_2, \ldots, p_k \ge 0$, $p^+ = p_1 + \cdots + p_k$, and $m_p = \mathbb{E}[|\mathcal{N}(0,1)|^p]$. Then we can set $c_j = a\sqrt{\mathrm{MPV}_j(2)/t}$, with a =3, 4, or 5 (these are standard choices in the literature).

We further assume that the continuous part of asset log-prices panel follows a continuous-time factor model on $[0, T]$. Namely, define $Y_t := X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s$ as the continuous part of X, and assume the following factor structure for Y_t:

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

where F_t is an r-dimensional ($r < d$) unobservable common factor, Z_t is an idiosyncratic component, and Λ_t is a $d \times r$ 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_t, and the idiosyncratic component, Z_t, are assumed to follow continuous

Itô-semimartingales, and are given by:

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

and

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

where B_s and $\widetilde{B}_s$ 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 model and general settings are also used in Aït-Sahalia and Xiu (2017).

3 Dimension Reduction and Forecasting Methods

The HAR model of Corsi (2009), upon which we build our analysis, is given as:

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

where RM represents a certain realized measure of the integrated volatility (IV) of the target asset, and $\mathrm{RM}_{[t,t-p]}$ is the average of RM's, over the most recent $p+1$ days. For instance, if RV is used as our realized measure in the above HAR model, then we define:

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

To eliminate the jump variation from total quadratic variation, we use truncated realized volatility (TRV) in our empirical application, as defined in (5), noting that TRV is a consistent estimate of IV.[5] The benchmark HAR model that we examine in our prediction experiments is called HAR-TRV.[6] This model is given by:

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

Furthermore, we examine the following factor-augmented model in our forecasting experiments:

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

[5] We actually combine the two estimators in (5) and (6), in the following sense. We first use bipower variation to get an initial consistent estimate of IV, and then use this to determine an initial choice for c in (5). Then, we obtain a second estimate of IV using truncation, and also a second choice of c. We iterate this procedure until the estimate of IV converges.

[6] In our forecasting experiments, several more benchmark models are also considered, as outlined in Section 4.

where y_{t+h} denotes h-step-ahead daily integrated volatility, h is the forecast horizon (set equal to unity in our experiments), 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 HAR-TRV model defined in (12)) and Ψ_t consists of r-dimensional unobservable predictors. Based on the structure of factors assumed in (7) and (8), 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 (also called our IV factor in the sequel). Note that one cannot disentangle Λ from η unless certain identification conditions, such as $\eta\eta^\intercal = I_r$, are imposed. However, as shown above, we don't have to disentangle these components from Ψ_t. This is because we are only interested in Ψ_t, which is the IV matrix of the r uncorrelated common factors in our setup. In summary, it is worth stressing that unlike many other applications of factor-augmented regression, we do not directly use weighted common factors, $\Lambda_t F_t$, extracted from a large panel of asset returns. Instead, what we actually use as predictors in forecasting models are the estimated IVs of these common factors (i.e. the Ψ_t).

As mentioned above, we propose a two-step shrinkage procedure to estimate the above latent common volatility factors. More specifically, we first use LASSO or elastic net shrinkage on the estimates IVs of all assets in order to obtain a parsimonious group of assets whose IVs are relevant to predicting the target asset's IV. We then apply PCA or SPCA to this group of assets (the panel of asset returns) in order to estimate common asset factors, from which our common IV factors are estimated. These factors are then inputted into (13) in order to forecast IV for the target asset. In the following two sections, we briefly outline the techniques used in this two-step procedure (i.e. the LASSO, elastic net, PCA, and SPCA).

3.1 LASSO and Elastic Net Shrinkage

In order to select stocks in the first step of our shrinkage procedure, we consider two shrinkage based variable selection, 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, consider a regression of y_{t+h} on w_t and χ_t, where y_{t+h} and w_t are defined in (13), and χ_t is a vector of IV estimates on day

t, for all assets in X_t. The LASSO estimator is the solution to the following problem:

$$\min_{\phi} \sum_t \left\{ \left\| y_{t+h} - \beta^{\intercal} w_t - \sum_j \phi_j \chi_{j,t} \right\|^2 + \lambda \sum_j |\phi_j| \right\}, \tag{14}$$

where the ϕ's are regression coefficients in a standard penalized regression. Similarly, the elastic net estimator is a solution to the following problem:

$$\min_{\phi} \sum_t \left\{ \left\| y_{t+h} - \beta^{\intercal} w_t - \sum_j \phi_j \chi_{j,t} \right\|^2 + \lambda \sum_j \Big(\frac{(1-\alpha)}{2} \phi_j^2 + \alpha |\phi_j| \Big) \right\}, \tag{15}$$

where $\alpha \in [0, 1]$. Of note is that when $\alpha = 1$, the elastic net is equivalent to the LASSO. Also, as α shrinks toward 0, elastic net estimators approach those obtained via ridge regression. Furthermore, note that the LASSO imposes an $\mathcal{L}_1$-norm penalty on coefficients in the model, while the elastic net induces a variety of double shrinkage, in the sense that it imposes $\mathcal{L}_1$-norm and $\mathcal{L}_2$-norm penalties on coefficients in the model. Finally, recall that it is the imposition of the $\mathcal{L}_1$-norm penalty that induces shrinkage to zero of some coefficients in the regression model; and it is the non-zero coefficients in the solution to these minimization problems that are used to select the final set of variables.

In our experiments, we set $\alpha = 0.2$ and 0.6. Furthermore, the regularization parameter, λ, is chosen via ten-fold cross validation. Only assets with nonzero ϕ's are retained in our final set of selected target predictor assets, say $\widetilde{X}_t$. For two different target assets, the selected pool of assets (i.e. $\widetilde{X}_t$) from which we construct the $\widehat{F}_t$'s and subsequently the $\widehat{\Psi}_t$'s can be substantially different. Intuitively, since the assets in $\widetilde{X}_t$ all have relatively large regression coefficients, they are potentially more informative about the target asset than other assets in X. Hence, the common volatility factors extracted using data from this selected pool of assets may be contaminated with less irrelevant information as would have been the case were the entire X dataset used in our analysis.

In closing this section, we stress that the above sorts of shrinkage are (potentially) carried out in both steps of our procedure. In particular, recall that the first step involves shrinkage on the set of all IV estimators of the assets in X; and this step is used to select a finer subset of asset returns for further analysis. In the second step, PCA or SPCA is used to construct common factor estimates from the new pared-down asset dataset. These common factor estimates are then used to construct our IV factors. Recall that SPCA can itself be interpreted as a two-step procedure, whereby PCA is first applied, followed by either LASSO or elastic net shrinkage of the PCA factor weights in a second step. Thus, our procedure (potentially) involves two layers of shrinkage, where the first layer involves shrinkage of a large set of IV estimators, and the second layer involves shrinkage of a parsimonious set of assets used to construct a latent common asset return factor (from which our IV factor is extracted). In addition, it is crucial to differentiate between these two layers of shrinkage, in the following sense. The LASSO and elastic net are supervised learning methods, while PCA and SPCA are unsupervised learning methods. Intuitively speaking, the first layer of shrinkage takes the ultimate goal (i.e., forecasting) into account, and carries out shrinkage through deletion of irrelevant variables; while the second layer focuses on the task of dimension reduction through combination of (almost) all variables in the dataset, without focus on forecasting the target variable.

3.2 (Sparse) Principal Component Analysis

In order to carry out the second step of our procedure, we utilize PCA or SPCA, both of which are briefly discussed in this section. To start, consider the following covariance matrix estimator, defined on a fixed interval, $[0, T]$:

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

One carries out PCA by applying an eigenvalue-eigenvector decomposition to $\widehat{\Sigma}_t$, yielding r estimated eigenvalues, in descending order, say $\widehat{\lambda}_1 > \widehat{\lambda}_2 > \cdots > \widehat{\lambda}_r$, and corresponding estimated eigenvectors, $\widehat{\xi}_1$, $\widehat{\xi}_2, \cdots$, $\widehat{\xi}_r$. The first r principal components on the fixed interval are estimated as follows:

$$\Delta_i^n \widehat{F}_j = \widehat{\xi}_j^\intercal (\Delta_i^n X) \mathbf{1}_{\{\|\Delta_i^n X\| \leq \boldsymbol{c} \Delta_n^{\varpi}\}}, \quad j = 1, \cdots, r. \tag{17}$$

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

$$\widehat{\Psi}_{j,t} = \frac{1}{t} \sum_{i=1}^{\lfloor t/\Delta_n \rfloor} (\Delta_i^n \widehat{F}_j)^2, \quad j = 1, \cdots, r. \tag{18}$$

Thus, for any $j = 1, \cdots, r$, we have $\widehat{\Psi}_{j,t} = \widehat{\xi}_j^\intercal \widehat{\Sigma}_t \widehat{\xi}_j = \widehat{\lambda}_j \widehat{\xi}_j^\intercal \widehat{\xi}_j$, which is equivalent to $\widehat{\lambda}_j$, if the eigenvector has unit-length.

Note that the above PCA procedure delivers the eigens (eigenvalues and eigenvectors) of the integrated volatility matrix. According to Aït-Sahalia and Xiu (2018), these eigens are different from the integrated eigens of the spot volatility matrix, when t does not shrink to zero. However, in finite samples, the time horizon t, which is one day in our empirical application, is relatively small compared to Δ_n, which we set equal to 2.5-, 5-, and 10-minute in our application. It is unpractical to further split our daily data into further blocks, and hence we do not distinguish eigens of integrated volatility versus integrated eigens of spot volatility differences in our empirical application, following the approach taken by Aït-Sahalia and Xiu (2017).

Also, it is well-known that eigens are nonlinear functions of the corresponding covariance matrix. Jacod and Rosenbaum (2013) show that various bias terms arise when estimating integrals of nonlinear functions of the spot volatility matrix. But when the local window size is relatively small, there is only one bias term, which can be consistently estimated. Moreover, according to Aït-Sahalia and Xiu (2018), 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, leading to a bias term that can be treated using the methods of Jacod and Rosenbaum (2013). These authors show that the estimator that we use is characterized by a higher order bias term; but nevertheless is consistent. Moreover, the higher order bias, which is an integral over $[0, t]$, is proportional to the true eigenvalue. Hence, in terms of forecasting, there is no distortion of information, since the main source of useful information comes from the eigenvalues. In view of these observations, we don't remove the bias term in our empirical application.

In general, PCA yields nonzero factor loadings for (almost) all variables, which exacerbates difficulty in interpretation, and can induce noisiness in estimated factors. To avoid these drawbacks, and to induce parsimony, we also utilize SPCA in the second step of our procedure. This technique is closely related to PCA (see Jolliffe et al. (2003) and Zou et al. (2006)), and involves estimating "sparse" eigenvectors $\widehat{\xi}_j$, for $j = 2, 3, ..., r$, which are solutions to the following optimization problem:

$$\max_{\|\xi_j\|_2=1,\ \xi_j \perp \xi_1, \cdots, \xi_{j-1}} \xi_j^\intercal \widehat{\Sigma}_t \xi_j, \tag{19}$$

subject to $\sum_{k=1}^{d} |\xi_{j,k}| \leq \delta$, where δ is a regularization parameter.

Note that the constraint in (19) imposes an $\mathcal{L}$-1 norm penalty on the eigenvectors, and hence induces sparsity. As discussed above, Zou et al. (2006) introduce other methods for carrying out SPCA based on an alternative interpretation linking SPCA with PCA followed by shrinkage. The key to SPCA, though, is that it yields sparse factor loadings, in the sense that loadings may be identically zero, a feature not feasible in the context of shrinkage on the $\mathcal{L}_2$-norm, such as that associated with ridge regression.[7]

4 Experimental Setup

The forecasting model in our IV prediction application is given above as (13). Given that we are interested in predicting TRV, we re-write this model as:

$$\widehat{\mathrm{TRV}}_{t+1} = \beta_0 + \beta_1 \widehat{\mathrm{TRV}}_t + \beta_2 \widehat{\mathrm{TRV}}_{[t,t-4]} + \beta_3 \widehat{\mathrm{TRV}}_{[t,t-21]} + \beta_\Psi^\intercal \widehat{\Psi}_t + \epsilon_t, \tag{20}$$

where latent IV factors (i.e., $\widehat{\Psi}_t$), are constructed by implementing the two-step procedure discussed in Section 3. We choose the number of latent factors, r, in our experiments by using 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 more informative one. Although the choice of cutoffs is somewhat arbitrary, our experimental findings are robust to various other cutoffs.

In total, we examine six factor-augmented models based on six "permutations" of our two-step procedure. These include the following models:

EN1-PCA: *Step 1: A pool of assets are selected based on elastic net shrinkage (with α = 0.2) of IV dataset. Step 2: Latent integrated volatility factors are extracted from latent asset factors constructed using PCA.*

EN2-PCA: *Step 1: A pool of assets are selected based on elastic net shrinkage (with α = 0.6) of IV dataset. Step 2: Latent integrated volatility factors are extracted from latent asset factors constructed using PCA.*

[7] As in the case of PCA, we use the r largest eigenvectors as common volatility factors in our prediction models.

LASSO-PCA: *Step 1: A pool of assets are selected based on LASSO shrinkage of IV dataset. Step 2: Latent integrated volatility factors are extracted from latent asset factors constructed using PCA.*

EN1-SPCA: *Step 1: A pool of assets are selected based on elastic net shrinkage (with* $\alpha = 0.2$*) of IV dataset. Step 2: Latent integrated volatility factors are extracted from latent asset factors constructed using SPCA.*

EN2-SPCA: *Step 1: A pool of assets are selected based on elastic net shrinkage (with* $\alpha = 0.6$*) of IV dataset. Step 2: Latent integrated volatility factors are extracted from latent asset factors constructed using SPCA.*

LASSO-SPCA: *Step 1: A pool of assets are selected based on LASSO shrinkage of IV dataset. Step 2: Latent integrated volatility factors are extracted from latent asset factors constructed using SPCA.*

Recall also that our main benchmark model, called BM-I is:

$$\widehat{\mathrm{TRV}}_{t+1} = \beta_0 + \beta_1 \widehat{\mathrm{TRV}}_t + \beta_2 \widehat{\mathrm{TRV}}_{[t,t-4]} + \beta_3 \widehat{\mathrm{TRV}}_{[t,t-21]} + \epsilon_t. \tag{21}$$

In addition to BM-I, we evaluate several other benchmark models. First, to demonstrate the importance of the first-step variable selection part of procedure, we construct common volatility factors using only the second step (i.e. PCA or SPCA) and include them in the benchmark BM-I model. This model is denoted BM-II, and is analogous to the widely used "diffusion index" model of Bai and Ng (2006a), Stock and Watson (2002a), Stock and Watson (2002b) and Stock and Watson (2006). Second, for cases where we are interested in forecasting IV of sector ETFs, we construct BM-III. This model adds a market volatility predictor measured by the TRV of market ETF (SPY) to BM-I. Third, for cases where we are interested in forecasting IV for individual firms, we construct BM-IV and BM-V. In BM-IV, an additional sector volatility predictor measured by TRV of the sector ETF is added to BM-I. In BM-V, additional sector and market volatility predictors, constructed as in BM-III and BM-IV, are added to BM-I. These models are utilized in our experiments as follows. When predicting IV for the market ETF (SPY), we compare the performance of our factor augmented models with BM-I and BM-II. When predicting IV for the sector ETFs, we compare the performance of our factor augmented models with BM-I and BM-II, and BM-III. Finally, when predicting IV for individual firms, we compare the performance of our factor augmented models with BM-I and BM-II, BM-III, BM-IV and BM-V.

Model estimation and volatility prediction is carried out each day, using a rolling-window estimation scheme, prior to the construction of each new daily IV prediction. The length of rolling window (i.e. the *in-sample* period), is 630 days (two and a half years). For example, we first estimate all models using data from November 27, 2006 to May 29, 2009 (630 trading days), and then construct one-day-ahead forecasts for June 1, 2009. Then, in order to forecast the volatility on June 2, 2009, we first estimate our models using data from November 28, 2006 to June 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 June 1, 2009 to December 31, 2010, which constitutes 402 trading days. Benchmark models are 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 and benchmark models, we consider two different criteria:

(a) *In-sample* R^2.

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

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

where y_t is the ex-post value of volatility, $\bar{y}_t$ is the historical average of volatility, and $\widehat{y}_t$ is our forecast. In addition, Diebold-Mariano-West (DMW) predictive accuracy tests of equal predictive accuracy are also carried out and reported. For a detailed discussion of this test, in which the null hypothesis is that two models have equal predictive accuracy, based on a given loss function, see Diebold and Mariano (1995)) and West (1996). In our implementation of the test, we utilized a quadratic loss function. In this sense, we are comparing the MSFEs of two competing models.

5 Data

We collect intraday observations on 480 constituents of the S&P 500 index[8]; 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). The sample period

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

is from January 3, 2006 to December 31, 2010. All data are collected from the TAQ database.

Each individual stock in our dataset is assigned to a sector, according to the Global Industry Classification Standard (GICS) code system. As shown in Figure 1, the largest 4 sectors are Consumer Discretionary (CD), Information Technology (IT), Industrials (I) and Health Care (HC). Approximately 55% individual stocks in our dataset belong to these four sectors, while the smallest sector, Telecommunication Services (TS), only contains 2.08% of stocks.

In our forecasting experiments, target assets include SPY, the 9 sector ETFs listed above, and 18 individual stocks, including: Ecolab Inc. (ECL), Dow Chemical Company (DOW), Chevron Corporation (CVX), Exxon Mobil Corporation (XOM), J.P. Morgan Chase & Co. (JPM), The Goldman Sachs Group, Inc. (GS), General Electric Company (GE), Minnesota Mining and Manufacturing Company (MMM), Microsoft Corporation (MSFT), International Business Machines Corporation (IBM), The Coca-Cola Company (KO), The Procter & Gamble Company (PG), Duke Energy Corporation (DUK), Southern Company (SO), Johnson & Johnson (JNJ), Merck & Company, Inc. (MRK), McDonald's Corporation (MCD), Lowe's Companies, Inc. (LOW).[9] For the sake of brevity, only select results are presented in the sequel. All results are included in a separate appendix. Finally, data cleaning, subsampling, etc., all follow standard procedures described in Aït-Sahalia and Jacod (2012); and overnight returns are excluded from our analysis.

6 Empirical Findings

6.1 Forecasting Performance

We begin by discussing the one-day-ahead predictive performance of the benchmark and various volatility factor-augmented models outlined in Section 4, for the forecasting sample period from June 1, 2009 to December 31, 2010. In our experiments, we include models with volatility factors constructed using three different sampling frequencies, including 2.5-, 5-, and 10-minutes. Recall also that all models are estimated with rolling data windows, prior to daily ex ante forecast construction. A number of clear-cut findings emerge upon inspection of the results contained in these Tables 1–19.

The most important finding is that our two-step shrinkage procedure results in

[9] Two stocks in this set of 18 individual stocks are from each of the 9 sectors in our dataset. For instance, CVX and XOM belong to the Energy sector, according to the GICS code.

notable and significant increases in *out-of-sample* R^2 values when predicting daily IV, at firm, sector, and market levels. In particular, for all assets except for those in the financial sector, increases range from approximately 20% to 45%, when compared with R^2 values associated with the best performed benchmark models, across all sampling frequencies. Additionally, DMW tests tell us that the volatility factor augmented models yield significantly smaller MSFEs than those of almost all benchmark models. In the following paragraphs, we discuss these (and related) findings in more detail.

Recall that we include five different benchmark models in our experiments. These benchmarks can be described as follows. BM-I is a non-augmented HAR model, BM-II augments BM-I by including an IV factor constructed using only step 2 of our procedure. BM-III and BM-IV include market and sector ETF volatility factors constructed without recourse to our procedure at all. Namely, IV factors are directly constructed from sector and market level asset return data. Finally, BM-V includes the market and sector ETF volatility factors included in BM-III and BM-IV, respectively. Interestingly, the use of BM-II – BM-V does not always equate with forecasting performance improvement, relative to BM-I. For instance, the *out-of-sample* R^2 of BM-II for the market ETF (SPY) is 0.319 using 5-minute frequency data, which is smaller than than that of BM-I, which is 0.347. As another example, note that for Duke Energy Corporation (DUK) (see Table 17), BM-III, BM-IV and BM-V generate either negative or approximately zero *out-of-sample* R^2 values, using 2.5- or 5-minute frequency data. Indeed, it is only for BM-IV, based on 10-minute frequency data, that we observe a slight increase in forecasting performance, relative to the benchmark HAR-TRV model. In this case there is a 9.8% improvement in *out-of-sample* R^2 (simply called R^2 hereafter). Similar results obtain for XLB, XLI, XLY, GE, and various individual assets.

Building on the above results, it is important to note that all of our benchmark models are dominated by our two-step IV factor-augmented models. For instance, for XLK (technology), the use of BM-II results in an improvement in R^2 from 0.266 to 0.271, when comparing with BM-I (see Table 6). This is a 1.88% improvement, and is based on using 5-minute frequency data. However, analogous R^2 values based on the use of our factor-augmented models range from 0.271 to 0.292 (corresponding to improvements ranging from 1.88% to 9.77%, relative to BM-I). As a second example consider the results in Table 9 for XLV (heath care). While the use of BM-II increases the R^2 0.206 to 0.267 (a 29.6% improvement, relative to BM-I), our factor-augmented models achieve R^2 values ranging from 0.306 to 0.332 (a 48.5% to 61.2% improvement, relative to BM-I). Similar patterns emerge when examining results for XLE, XLP, XLU, XLV, CVX, MSFT, KO, and various other individual assets reported on in Tables 1–19. The above findings suggest that it is not enough to simply use market or sector level data in order to directly extract common IV factors. Indeed, if the objective is to maximize predictive content, then there is much to be gained by applying shrinkage methods to disaggregate firm-level data in order to first build a parsimonious set of assets, from which IV factors can be extracted. This is the approach taken in our two-step procedure. Moreover, the more shrinkage the better. For example, recall that the first step of our procedure involves shrinkage applied to the firm level dataset in order to build a parsimonious set of assets, as just described. However, the second step may or may not utilize shrinkage. In particular, when PCA is used in this step, dimension reduction is achieved, but all assets from the first step still play a role, as weights used to construct the IV factor are all non-zero. On the other hand, if SPCA is used (which implies implementing further LASSO or elastic net shrinkage on PCA factor loading weights), then we are implementing a second layer of shrinkage in our procedure, as discussed in the introduction. This approach indeed yields the "best" results in our experiments. Consider the case of the energy sector (see Table 3). For this variable, called XLE, there is an improvement in predictive performance (as measured by R^2) if one uses our two-step method with PCA (see results for models EN1-PCA, EN2-PCA, and LASSO-PCA) instead of BM-I - BM-III. However, there is a further predictive improvement if one uses our two-step method with SPCA instead of PCA (see results for models EN1-SPCA, EN2-SPCA, and LASSO-SPCA). Indeed, there are no cases where the R^2 does not increase when SPCA is used in our procedure. This finding characterizes all of the results reported in Tables 1–19.

As a final indicator of the usefulness of the methods discussed in this paper, note that forecasting performance becomes unstable if the first "variable selection" step is removed from our procedure. For example, again in the case of XLE, although BM-II generates larger R^2 values than the benchmark HAR-TRV model using 2.5- and 5-minute frequency data, it performs significantly worse using 10-minute frequency data. In addition, the gains observed even at the 2.5- and 5-minute frequencies are very small. For example, the R^2 for BM-I and BM-II using 5-minute frequency data are very close (0.308 and 0.316, respectively). For individual stocks, the picture is even more stark. For example, in Table 11, BM-II performs worse than BM-I across all sampling frequencies. The *out-of-sample* R^2 for BM-II is even negative, when 10-minute frequency data are used. In contrast, all our factor-augmented models achieve much higher *out-of-sample* R^2 values than both BM-I and BM-II at all sampling frequencies.

There are a number of less important, although still interesting findings that also emerge upon inspection of the results reported in Tables 1–19.

First, *in-sample* fit is surprisingly stable across different models and the above three different data frequencies, no matter which asset is considered. Namely, *in-sample* fit changes little when common volatility factors are added to the benchmark HAR-TRV model, regardless of asset class. Thus, based solely on *in-sample* diagnostics, there appears to be little gain by adding volatility factors in the original HAR model. In fact, if only *in-sample* R^2 values were examined in order to assess the usefulness of common factors, then the story would change markedly. Take Johnson & Johnson as an example. The benchmark HAR-TRV model using 5-minute frequency data (without a common factor) achieves an *in-sample* R^2 value of 0.38, while *in-sample* R^2 values for our factor-augmented models are all between 0.42 and 0.43. 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.

One way to interpret the above disparity between *in-sample* and *out-of-sample* fit is given as follows. It is widely known that *in-sample* R^2 values tend 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 multiple explanations have been proposed, including the presence of (smooth) structural breaks and state transitions, as well as the general inability of linear models to capture inherently nonlinear interactions among financial variables and markets (see e.g., 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, consistent with the above general finding. For example, using the Coca-Cola Company (see the 5-minute panel in Table 16) to illustrate our findings, the BM-I model achieves an *in-sample* R^2 value of 0.56, as opposed to an *out-of-sample* R^2 value of 0.20. However, when the "best" factor-augmented *in-sample* and *out-of-sample* performances are compared in this example, the R^2 values are 0.60 and 0.36, 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 asset levels, although it is more starkly apparent at the individual stock level.

Second, our *out-of-sample* forecasting results vary with sampling frequency. Moreover, the "best" frequency varies across different assets and asset classes. However, in general, we recommend the use of 5-minute frequency, since our factor-augmented models generally yield the "best" predictions (see below for further discussion) using such data. The rationale for this finding is as follows. On one hand, using a higher frequency may result in contamination with a substantial amount of microstructure noise, which potentially deteriorates predictive performance. On the other hand, if the sampling frequency is relatively low, it is more difficult to eliminate the effects of jumps when estimating latent factors, leading to forecast deterioration (assuming that jumps are usually difficult to predict).

Third, there is an important wrinkle to the above story. For financial assets, *out-of-sample* R^2 values are negative in some cases, and in other cases predictive accuracy gains are negligible. A particularly interesting example of this is the financial sector ETF. For this ETF, simply adding the market ETF volatility to the benchmark HAR-TRV model (i.e. BM-III) results in the best forecasting performance, at 2.5- and 10-minute sampling frequencies (see Table 4), and at the 5-minute sampling frequency, BM-III also generates higher *out-of-sample* R^2 values than our volatility factor augmented models. At the individual stock level, the picture is even more stark. Consider J.P. Morgan Chase & Co. (see Table 13). *Out-of-sample* R^2 values for factor-augmented models are always less than 0, when 2.5- and 10-minute frequency data are used, and are less than 0.1 when 5-minute frequency data are used. Furthermore, simply adding both market and financial sector volatilities to the benchmark HAR-TRV model (i.e. BM-V) yields the most accurate forecasts. One possible explanation for the above finding is that the financial sector is a main driving force of the whole market. Hence, a crudely constructed factor is sufficient to capture most of the useful information in the sector. This argument is partly borne out in the data, since the datasets used to construct the volatility factors using our shrinkage methodology, always contain a significant number of financial sector stocks in them; and since these factors in turn lead to the impressive predictive gains when predicting IV, as discussed above. This point is discussed further in Section 6.2.

6.2 Latent Factor Structures

Figures 1–4 and Tables 20–25 summarize percentages of stocks (by sector) that comprise the variables selected in the first step of our procedure (Figures 1–4), and list the key stocks that are utilized in the construction of the latent IV factors in the second step of our procedure (Tables 20–25). For the sake of brevity, the above figures and tables only report results for a representative set of target stocks and sectors. Complete results are available upon request. A number of interesting findings based on these figures and tables are summarized below.

We begin by first noting that different variable selection methods (i.e., the LASSO and elastic net) used in the first step of our procedure select almost the same pools of stocks, at each sampling frequency. This is illustrated in Figures 2–4. In particular, each column of charts in these figures corresponds to results for a particular shrinkage method, while each row contains charts for a particular sampling frequency. Within each chart, absolute as well as relative percentages of stocks chosen in each sector are charted, with relative percentages calculated by rescaling according to the "size" of each sector, as depicted in Figure 1. The stated result can be seen to hold since the three charts in each row of these figures are virtually identical. Thus, there is little to choose between using LASSO or elastic net shrinkage in the first step of our double shrinkage procedure. However, it is worth noting that the pool of selected stocks does change with data frequency (compare each chart in a given column in the figures). For instance, consider the market ETF. Figure 2 indicates that when we use the elastic net with $\alpha = 0.2$, with 10-minute frequency data, almost 15% of selected stocks derive from the health care sector. This percentage jumps up to approximately 20% with 5-minute frequency data. As another example, note that the percentage of information technology stocks drops from approximately 15% to 10% when the sampling frequency decreases from 10-minutes to 5-minutes. Similar results can be seen upon inspection of Figures 3 (sector ETF) and 4 (individual stock).

Second, note that inspection of the results in Figures 2–4, suggests that consumer discretionary stocks tend to be selected most frequently, in the first step of our procedure, across all sampling frequencies. This is not surprising, given that the consumer discretionary sector is the largest amongst all sectors (see Figure 1). However, when we rescale the number of selected stocks by the size of each sector, then the picture changes. For example, consider the "Relative Ratio" results reported in Figure 3. Here, we see that the relative percentage of financial stocks is greater than the relative percentage of consumer discretionary stocks, at 2 of 3 data frequencies. This indicates the relative importance of financial stocks in the first step of our procedure.

Third, turning again to Figures 2–4, note that health care stocks tend to be selected quite frequently in the first step of our procedure. However, relatively small weights are placed on such stocks in the second step, when utilizing PCA and SPCA to estimate our IV factors. For instance, in Tables 20, 22 and 24, note that very few stocks from the health care sector are contained in our lists of stocks with the highest average factor weights, when averaging across all rolling windows in our prediction experiments.[10] Furthermore, examination of the results in Tables 21, 23 and 25 indicate that SPCA frequently places identically zero weights on health care stocks, particularly when higher frequency data are used in latent factor construction. For example, in Table 25, we report the frequency of stocks with a high likelihood of receiving a zero weight, and 7 of 15 listed stocks and 5 of 15 listed stocks belong to the health care sector, at 2.5-and 5-minute data frequencies, respectively. Indeed, almost 64% and 53% of daily weights are zero for IDXX (IDEXX Laboratories, Inc.) and for ILMN (Illumina, Inc.), respectively, under EN1-SPCA, at the 2.5-minute data frequency. Consequently, the average weight on IDXX and on ILMN decreases sharply from 0.068 to 0.021, and from 0.078 to 0.029, respectively, when we change from using PCA to using SPCA in the second step of our procedure. 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.

Finally, stocks in the consumer discretionary and financial sectors usually have larger factor loadings (weights), under both PCA and SPCA (see Tables 20, 22 and 24), particularly at higher data frequencies. For instance, in Table 20, only two or three stocks are not in the consumer discretionary and financial sectors, under EN2-PCA/SPCA, using 2.5-minute frequency data. CCL (Carnival Corporation), TGT (Target Corporation) and BBY (Best Buy Co., Inc.) – in the consumer discretionary sector, and TROW (T. Rowe Price Group, Inc.), AXP (The American Express Company) and PFG (The Principal Financial Group) – in the financial sector, all have average weights greater than 0.12. Similarly, in Table 22, these stocks again all have average weights greater than 0.12. Putting all of the above evidence together, we conclude that although health care stocks are frequently chosen in our first step shrinkage procedure, their contributions to common volatility factors appears to be less

[10]Tables 20, 22 and 24 list stocks that received the highest average weights in our IV factors, while Tables 21, 23 and 25 list stocks receiving the lowest average weights.

than that of consumer discretionary and financial stocks.

7 Concluding Remarks

This paper investigates whether latent common volatility factors extracted from a large-dimensional panel of high-frequency intraday stock returns can 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) or elastic net shrinkage, followed by factor estimation using (sparse) principal components analysis ((S)PCA)). The first step of our procedure involves variable selection based on examination of IV estimators, while the second step involves factor construction using asset returns, followed by IV estimation based on the estimated factor structure.

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 *out-of-sample* R^2. Diebold-Mariano-West test results support this finding. Second, our two step procedure "MSFE-dominates" a variety of benchmark models where latent factors are constructed using a variety of different methods. Third, *in-sample* model performance is not indicative of *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. Fourth, we recommend using high frequency datasets consisting of data sampled at 5-minute frequency, when constructing predictions of volatility using IV factor-augmented models. This choice offers a balance between reducing the impact of microstructure noise in higher frequency data, on the one hand, and reducing effects associated with the prevalence of jumps in lower frequency data, on the other hand. We also find that models utilizing SPCA perform better than those with PCA, when these methods are used to extract common volatility factors. This finding points to the usefulness of double shrinkage, given that SPCA can be interpreted as a form of shrinkage applied to PCA, and given that the first stage of our procedure also utilizes shrinkage.

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

(2017, 2018), 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, ir remains 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 also remains to assess whether the findings in this paper can be translated into profitable investment strategies, in real-time trading contexts.

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. 2017. Using principal component analysis to estimate a high dimensional factor model with high-frequency data. *Journal of Econometrics* 201:384–399.

Aït-Sahalia, Y. and Xiu, D. 2018. Principal component analysis of high-frequency data. *Journal of the American Statistical Association* pp. 1–17.

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. *Annual Review of Financial Economics* 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.

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 & 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 RENÒ, R. 2010. Threshold bipower variation and the impact of jumps on volatility forecasting. *Journal of Econometrics* 159:276–288.

Diebold, F. X. and Mariano, R. S. 1995. Comparing predictive accuracy. *Journal of Business & Economic Statistics* 20:134–144.

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. 2018. Limit of random measures associated with the increments of a brownian semimartingale. *Journal of Financial Econometrics* 16:526–569.

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.

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. 2003. A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics* 12:531–547.

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. 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.

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. 2018. Big data analytics in economics: What have we learned so far, and where should we go from here? *Canadian Journal of Economics* 51:695–746.

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

West, K. D. 1996. Asymptotic inference about predictive ability. *Econometrica* 64:1067–1084.

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.

Table 1: SPDR S&P 500 ETF (SPY)

Sampling Frequency	Model Specification							
	BM-I	BM-II	EN1-PCA	EN2-PCA	LASSO-PCA	EN1-SPCA	EN2-SPCA	LASSO-SPCA
	In-sample R^2							
2.5-minute	0.531	0.544	0.543	0.542	0.542	0.546	0.546	0.545
5-minute	0.600	0.601	0.603	0.603	0.602	0.606	0.605	0.605
10-minute	0.501	0.510	0.515	0.521	0.522	0.519	0.528	0.530
	Out-of-sample R^2							
2.5-minute	0.303	0.336	0.379 (3.17***) (1.89*)	0.372 (3.37***) (1.99**)	0.376 (3.03***) (1.76*)	0.396 (2.92***) (2.01**)	0.386 (3.19***) (2.29**)	0.394 (2.83***) (1.93*)
5-minute	0.347	0.319	0.353 (0.69) (2.75***)	0.371 (1.35) (2.53**)	0.369 (1.50) (2.84***)	0.369 (1.83*) (3.11***)	0.387 (2.33**) (3.25***)	0.383 (1.98**) (2.94***)
10-minute	0.251	0.313	0.410 (1.68*) (3.62***)	0.404 (1.53) (3.28***)	0.400 (1.84*) (3.38***)	0.413 (1.65*) (3.13***)	0.415 (1.55) (2.95***)	0.411 (1.76*) (3.66***)

*Note: Numerical entries in this table are *in-sample* R^2 and *out-of-sample* R^2 statistics associated with various models (listed in the second row of the table), for the target asset given in the title of the table. All models other than the benchmark models, denoted as "BM-I" and "BM-II", include latent volatility factors constructed using our two-step procedure. "EN1" and "EN2" denote models for which elastic net shrinkage is used in initial variable selection, with $\alpha = 0.2$ and 0.6, respectively. "LASSO" denotes the use of the least absolute shrinkage operator in initial variable selection. After the initial variable selection, either PCA or SPCA is utilized to obtain the latent volatility factor(s) used in all models (other than BM-I and BM-II). Entries in parentheses under each *out-of-sample* R^2 statistic are t-statistics from Diebold-Mariano-West (DMW) tests of equal predictive accuracy. The are two different tests carried out for each model, corresponding to the case where either BM-I (upper bracketed statistic) or BM-II (lower bracketed statistic) is used as the null model, against which the model listed in the first row of entries in the table is compared. The loss function used in the tests is the mean square forecast error, and ***, ** and * indicate rejection at 1%, 5% and 10% significance levels, respectively. Finally, positive statistic values indicate that the MSFE of the benchmark model is higher, so that rejection of the null hypothesis in this case implies that the IV factor model outperforms the benchmark. Complete details are given in Sections 3 and 6.

Table 2: Materials Select Sector SPDR ETF (XLB)

Sampling Frequency	Model Specification								
	BM-I	BM-II	BM-III	EN1-PCA	EN2-PCA	LASSO-PCA	EN1-SPCA	EN2-SPCA	LASSO-SPCA
	In-sample R^2								
2.5-minute	0.618	0.619	0.645	0.617	0.617	0.617	0.618	0.618	0.618
5-minute	0.625	0.626	0.625	0.627	0.627	0.627	0.627	0.627	0.627
10-minute	0.554	0.555	0.554	0.556	0.556	0.556	0.562	0.562	0.561
	Out-of-sample R^2								
2.5-minute	0.368	0.358	0.281	0.368 (2.27**) (1.18) (2.14**)	0.368 (2.27**) (1.18) (2.14**)	0.368 (2.27**) (1.18) (2.14**)	0.376 (1.70*) (1.53) (2.24**)	0.374 (2.07**) (1.43) (2.22**)	0.376 (1.70*) (1.44) (2.20**)
5-minute	0.310	0.305	0.310	0.330 (2.41**) (2.88***) (3.02***)	0.326 (2.05**) (2.43**) (2.49**)	0.328 (2.48**) (2.85***) (3.25***)	0.332 (3.81***) (2.60***) (3.34***)	0.330 (2.59***) (2.67***) (2.96***)	0.327 (2.23**) (2.29**) (2.44**)
10-minute	0.146	0.114	0.152	0.170 (3.15***) (4.75***) (2.12**)	0.174 (2.92***) (4.51***) (2.27**)	0.171 (3.17***) (5.03***) (1.99**)	0.195 (1.71*) (2.42**) (1.68*)	0.207 (1.95*) (2.62***) (1.95*)	0.188 (1.83*) (2.69***) (1.76*)

*Note: See notes to Table 1.

Table 3: Energy Select Sector SPDR ETF (XLE)

Sampling Frequency	Model Specification								
	BM-I	BM-II	BM-III	EN1-PCA	EN2-PCA	LASSO-PCA	EN1-SPCA	EN2-SPCA	LASSO-SPCA
	In-sample R^2								
2.5-minute	0.577	0.581	0.577	0.579	0.579	0.579	0.581	0.581	0.581
5-minute	0.615	0.614	0.616	0.617	0.618	0.617	0.621	0.621	0.621
10-minute	0.536	0.536	0.537	0.534	0.535	0.535	0.535	0.536	0.536
	Out-of-sample R^2								
2.5-minute	0.250	0.284	0.250	0.346 (5.15***) (2.85***) (5.14***)	0.345 (5.20***) (2.87***) (5.19***)	0.346 (5.32***) (2.88***) (5.31***)	0.384 (4.86***) (3.47***) (4.86***)	0.380 (5.01***) (3.56***) (5.01***)	0.378 (5.20***) (3.59***) (5.19***)
5-minute	0.308	0.316	0.313	0.369 (2.76***) (2.36**) (1.96**)	0.356 (3.30***) (2.60***) (1.87*)	0.364 (2.50**) (2.12**) (1.81*)	0.402 (3.21***) (2.96***) (2.68***)	0.393 (3.53***) (3.21***) (2.76***)	0.393 (2.90***) (2.60***) (2.31**)
10-minute	0.202	0.105	0.179	0.244 (3.77***) (5.68***) (2.81***)	0.273 (3.51***) (5.07***) (2.87***)	0.268 (3.49***) (5.13***) (2.77***)	0.255 (3.34***) (5.14***) (2.65***)	0.278 (2.93***) (4.48***) (2.58***)	0.278 (2.74***) (4.37***) (2.37**)

*Note: See notes to Table 1.

Table 4: Financial Select Sector SPDR ETF (XLF)

Sampling Frequency	Model Specification								
	BM-I	BM-II	BM-III	EN1-PCA	EN2-PCA	LASSO-PCA	EN1-SPCA	EN2-SPCA	LASSO-SPCA
	In-sample R^2								
2.5-minute	0.587	0.587	0.590	0.586	0.586	0.586	0.586	0.586	0.586
5-minute	0.585	0.588	0.586	0.589	0.589	0.588	0.590	0.591	0.591
10-minute	0.513	0.520	0.518	0.518	0.525	0.529	0.519	0.521	0.522
	Out-of-sample R^2								
2.5-minute	0.317	0.334	0.364	0.317 (0.33) (***-4.28) (***-3.94)	0.316 (-1.02) (***-3.61) (***-3.74)	0.316 (-1.21) (***-3.87) (***-3.83)	0.311 (*-1.91) (***-3.60) (***-3.82)	0.314 (-0.65) (**-2.50) (***-3.24)	0.314 (-1.26) (***-3.31) (***-3.62)
5-minute	0.286	0.324	0.311	0.304 (2.30**) (*-1.91) (-1.16)	0.304 (2.37**) (*-1.86) (-1.11)	0.305 (2.54**) (-1.59) (-0.81)	0.300 (2.00**) (**-2.07) (-1.57)	0.300 (1.79*) (**-2.35) (*-1.87)	0.300 (1.79*) (**-2.26) (*-1.83)
10-minute	0.149	0.122	0.187	0.167 (1.36) (1.34) (-1.23)	0.186 (2.44**) (1.64*) (-0.05)	0.185 (2.09**) (1.55) (-0.07)	0.164 (0.84) (0.94) (-0.87)	0.163 (1.31) (1.17) (-1.20)	0.161 (0.98) (1.03) (-1.27)

*Note: See notes to Table 1.

Table 5: Industrial Select Sector SPDR ETF (XLI)

Sampling Frequency	Model Specification								
	BM-I	BM-II	BM-III	EN1-PCA	EN2-PCA	LASSO-PCA	EN1-SPCA	EN2-SPCA	LASSO-SPCA
	In-sample R^2								
2.5-minute	0.536	0.583	0.555	0.557	0.557	0.557	0.560	0.560	0.560
5-minute	0.618	0.622	0.630	0.628	0.629	0.629	0.635	0.635	0.635
10-minute	0.527	0.531	0.526	0.544	0.544	0.545	0.544	0.546	0.546
	Out-of-sample R^2								
2.5-minute	0.286	0.161	0.074	0.343 (2.99***) (1.22) (1.67*)	0.345 (3.27***) (1.22) (1.67*)	0.344 (3.14***) (1.22) (1.67*)	0.361 (3.60***) (1.27) (1.69*)	0.365 (3.74***) (1.27) (1.69*)	0.363 (3.64***) (1.27) (1.68*)
5-minute	0.308	0.280	0.216	0.330 (1.62) (1.35) (1.37)	0.325 (1.07) (1.17) (1.30)	0.334 (1.80*) (1.37) (1.39)	0.343 (2.12**) (1.37) (1.39)	0.336 (1.56) (1.25) (1.33)	0.340 (1.99**) (1.34) (1.37)
10-minute	0.061	0.051	0.063	0.095 (1.89*) (1.92*) (1.86*)	0.099 (2.04**) (1.99**) (2.02**)	0.094 (2.17**) (2.09**) (2.16**)	0.093 (1.80*) (1.81*) (1.76*)	0.094 (1.98**) (1.94*) (1.95*)	0.093 (1.82*) (1.81*) (1.79*)

*Note: See notes to Table 1.

Table 6: Technology Select Sector SPDR ETF (XLK)

Sampling Frequency	Model Specification								
	BM-I	BM-II	BM-III	EN1-PCA	EN2-PCA	LASSO-PCA	EN1-SPCA	EN2-SPCA	LASSO-SPCA
	In-sample R^2								
2.5-minute	0.578	0.580	0.587	0.581	0.581	0.580	0.585	0.585	0.584
5-minute	0.618	0.618	0.618	0.623	0.622	0.622	0.623	0.623	0.623
10-minute	0.510	0.514	0.514	0.516	0.517	0.517	0.522	0.522	0.523
	Out-of-sample R^2								
2.5-minute	0.246	0.237	0.262	0.257	0.257	0.258	0.291	0.280	0.278
				(1.55)	(1.43)	(1.92*)	(2.79***)	(3.62***)	(3.35***)
				(1.58)	(1.55)	(1.46)	(1.82*)	(2.01**)	(1.94*)
				(-0.10)	(-0.11)	(-0.08)	(0.56)	(0.36)	(0.31)
5-minute	0.266	0.271	0.263	0.292	0.292	0.290	0.271	0.282	0.280
				(1.72*)	(1.84*)	(1.88*)	(0.36)	(1.52)	(1.40)
				(1.40)	(1.45)	(1.46)	(-0.02)	(0.59)	(0.52)
				(1.65*)	(1.72*)	(1.75*)	(0.55)	(1.56)	(1.42)
10-minute	0.079	0.027	0.103	0.105	0.109	0.101	0.137	0.138	0.127
				(1.85*)	(1.99**)	(1.75*)	(1.89*)	(1.70*)	(1.86*)
				(3.10***)	(3.23***)	(3.41***)	(2.53**)	(2.33**)	(2.58***)
				(0.05)	(0.17)	(-0.05)	(0.99)	(0.99)	(0.78)

*Note: See notes to Table 1.

Table 7: Consumer Staples Select Sector SPDR ETF (XLP)

Sampling Frequency	Model Specification								
	BM-I	BM-II	BM-III	EN1-PCA	EN2-PCA	LASSO-PCA	EN1-SPCA	EN2-SPCA	LASSO-SPCA
	In-sample R^2								
2.5-minute	0.284	0.348	0.372	0.329	0.329	0.329	0.321	0.321	0.322
5-minute	0.382	0.419	0.469	0.419	0.419	0.419	0.419	0.419	0.419
10-minute	0.402	0.414	0.408	0.423	0.420	0.419	0.425	0.420	0.420
	Out-of-sample R^2								
2.5-minute	0.100	0.163	0.120	0.267	0.271	0.278	0.289	0.306	0.308
				(3.19***)	(2.98***)	(2.91***)	(3.47***)	(3.15***)	(2.94***)
				(2.27**)	(2.19**)	(2.10**)	(2.34**)	(2.22**)	(2.08**)
				(1.73*)	(1.70*)	(1.68*)	(1.80*)	(1.77*)	(1.70*)
5-minute	0.218	0.441	0.415	0.454	0.456	0.461	0.486	0.485	0.491
				(2.62***)	(2.68***)	(2.65***)	(2.38**)	(2.46**)	(2.50**)
				(0.93)	(0.99)	(1.40)	(1.91*)	(2.08**)	(2.17**)
				(0.73)	(0.75)	(0.85)	(1.35)	(1.32)	(1.39)
10-minute	0.116	0.059	0.144	0.170	0.193	0.168	0.166	0.200	0.180
				(2.23**)	(3.20***)	(1.63)	(1.60)	(3.49***)	(2.19**)
				(3.95***)	(3.49***)	(2.71***)	(2.87***)	(4.36***)	(3.43***)
				(1.15)	(1.77*)	(1.06)	(0.66)	(1.77*)	(1.65*)

*Note: See notes to Table 1.

Table 8: Utilities Select Sector SPDR ETF (XLU)

Sampling Frequency	Model Specification								
	BM-I	BM-II	BM-III	EN1-PCA	EN2-PCA	LASSO-PCA	EN1-SPCA	EN2-SPCA	LASSO-SPCA
	In-sample R^2								
2.5-minute	0.482	0.506	0.517	0.499	0.499	0.499	0.496	0.496	0.496
5-minute	0.535	0.545	0.572	0.550	0.550	0.550	0.553	0.553	0.553
10-minute	0.394	0.414	0.421	0.404	0.405	0.405	0.403	0.406	0.406
	Out-of-sample R^2								
2.5-minute	0.256	0.249	0.067	0.355 (2.43**) (2.58***) (1.72*)	0.355 (2.46**) (2.51**) (1.70*)	0.357 (2.52**) (2.55**) (1.72*)	0.368 (2.91***) (2.13**) (1.65*)	0.370 (2.84***) (2.11**) (1.64*)	0.371 (2.87***) (2.15**) (1.65*)
5-minute	0.137	0.173	0.058	0.194 (2.63***) (2.76***) (1.72*)	0.195 (3.27***) (1.85*) (1.64*)	0.193 (2.93***) (1.87*) (1.64*)	0.214 (2.73***) (1.92*) (1.73*)	0.213 (2.95***) (1.66*) (1.65*)	0.211 (2.61***) (1.54) (1.63)
10-minute	0.119	0.252	0.194	0.351 (3.30***) (2.54**) (1.92*)	0.379 (2.67***) (2.23**) (1.80*)	0.361 (3.10***) (2.61***) (1.89*)	0.365 (3.58***) (2.59***) (2.01**)	0.412 (2.43**) (2.09**) (1.72*)	0.345 (3.35***) (2.09**) (1.80*)

*Note: See notes to Table 1.

Table 9: Health Care Select Sector SPDR ETF (XLV)

Sampling Frequency	Model Specification								
	BM-I	BM-II	BM-III	EN1-PCA	EN2-PCA	LASSO-PCA	EN1-SPCA	EN2-SPCA	LASSO-SPCA
	In-sample R^2								
2.5-minute	0.452	0.492	0.504	0.478	0.478	0.478	0.478	0.477	0.476
5-minute	0.470	0.494	0.533	0.491	0.490	0.491	0.497	0.496	0.497
10-minute	0.449	0.481	0.462	0.468	0.470	0.470	0.464	0.466	0.465
	Out-of-sample R^2								
2.5-minute	0.235	0.218	0.062	0.326 (3.01***) (1.85*) (1.86*)	0.329 (3.00***) (1.86*) (1.86*)	0.329 (3.02***) (1.86*) (1.86*)	0.324 (2.48**) (1.64*) (1.75*)	0.327 (2.60***) (1.71*) (1.79*)	0.329 (2.63***) (1.70*) (1.78*)
5-minute	0.206	0.267	0.188	0.332 (2.88***) (1.56) (1.95*)	0.300 (3.02***) (1.27) (1.91*)	0.306 (3.05***) (1.41) (1.96**)	0.349 (2.56**) (1.59) (1.96**)	0.322 (2.64***) (1.50) (1.96**)	0.334 (2.64***) (1.58) (1.99**)
10-minute	0.206	0.239	0.249	0.320 (3.05***) (2.31**) (1.48)	0.336 (3.06***) (3.25***) (1.62)	0.340 (2.82***) (3.23***) (1.51)	0.333 (3.03***) (3.00***) (1.58)	0.336 (3.52***) (3.04***) (1.83*)	0.337 (3.05***) (3.08***) (1.61)

*Note: See notes to Table 1.

Table 10: Consumer Discretionary Select Sector SPDR ETF (XLY)

Sampling Frequency	Model Specification								
	BM-I	BM-II	BM-III	EN1-PCA	EN2-PCA	LASSO-PCA	EN1-SPCA	EN2-SPCA	LASSO-SPCA
	In-sample R^2								
2.5-minute	0.470	0.542	0.497	0.498	0.498	0.498	0.501	0.502	0.502
5-minute	0.566	0.577	0.583	0.575	0.575	0.575	0.584	0.583	0.583
10-minute	0.486	0.503	0.490	0.514	0.517	0.516	0.523	0.523	0.523
	Out-of-sample R^2								
2.5-minute	0.328	0.356	0.232	0.451 (2.86***) (3.859***) (6.848***)	0.447 (2.872***) (3.757***) (6.702***)	0.449 (2.89***) (3.80***) (6.70***)	0.483 (2.85***) (4.14***) (7.00***)	0.473 (3.04***) (4.33***) (7.24***)	0.471 (3.16***) (4.44***) (7.19***)
5-minute	0.329	0.265	0.207	0.336 (0.42) (1.30) (1.84*)	0.339 (0.68) (1.35) (1.87*)	0.336 (0.50) (1.31) (1.84*)	0.368 (2.27**) (1.53) (1.94*)	0.372 (2.62***) (1.54) (1.94*)	0.368 (2.64***) (1.53) (1.94*)
10-minute	0.121	0.065	0.057	0.190 (2.32**) (1.54) (1.54)	0.194 (2.09**) (1.51) (1.51)	0.201 (2.12**) (1.54) (1.53)	0.207 (2.05**) (1.50) (1.50)	0.189 (2.00**) (1.55) (1.54)	0.221 (1.94*) (1.53) (1.52)

*Note: See notes to Table 1.

Table 11: Ecolab Inc. (ECL)

Sampling Frequency	Model Specification									
	BM-I	BM-II	BM-III	BM-IV	BM-V	EN1-PCA	EN2-PCA	LASSO-PCA	EN1-SPCA	EN2-SPCA	LASSO-SPCA
	In-sample R^2										
2.5-minute	0.561	0.571	0.567	0.595	0.596	0.569	0.569	0.569	0.571	0.571	0.571
5-minute	0.650	0.651	0.650	0.650	0.649	0.652	0.652	0.652	0.653	0.653	0.653
10-minute	0.482	0.506	0.502	0.514	0.515	0.489	0.490	0.490	0.506	0.505	0.509
	Out-of-sample R^2										
2.5-minute	0.251	0.231	0.173	-0.050	-0.050	0.298 (1.52) (4.34***) (5.54***) (5.25***) (4.67***)	0.301 (1.64*) (4.46***) (5.60***) (5.25***) (4.68***)	0.300 (1.64*) (4.55***) (5.68***) (5.24***) (4.66***)	0.317 (1.85*) (4.54***) (5.45***) (5.68***) (5.11***)	0.321 (2.02**) (4.50***) (5.43***) (5.67***) (5.14***)	0.320 (1.99**) (4.65***) (5.49***) (5.65***) (5.10***)
5-minute	0.277	0.246	0.283	0.282	0.280	0.314 (2.37**) (2.61***) (2.05**) (2.32**) (2.20**)	0.300 (2.30**) (3.02***) (1.35) (1.65*) (1.59)	0.315 (2.15**) (2.44**) (1.89*) (2.10**) (2.02**)	0.323 (2.36**) (2.49**) (2.26**) (2.41**) (2.37**)	0.316 (3.39***) (3.14***) (2.83***) (3.24***) (3.04***)	0.328 (2.26**) (2.42**) (2.15**) (2.29**) (2.24**)
10-minute	0.078	-0.079	0.012	-0.140	-0.111	0.209 (3.17***) (5.84***) (4.66***) (3.76***) (4.24***)	0.216 (2.66***) (6.02***) (4.21***) (3.54***) (3.94***)	0.221 (2.94***) (5.96***) (4.51***) (3.67***) (4.11***)	0.217 (2.44**) (4.47***) (3.41***) (3.49***) (3.79***)	0.238 (2.35**) (4.66***) (3.34***) (3.39***) (3.64***)	0.236 (2.23**) (4.56***) (3.26***) (3.37***) (3.63***)

*Note: See notes to Table 1. ECL belongs to the materials sector according to the GICS coding system.

Table 12: Chevron Corporation (CVX)

Sampling Frequency	Model Specification										
	BM-I	BM-II	BM-III	BM-IV	BM-V	EN1-PCA	EN2-PCA	LASSO-PCA	EN1-SPCA	EN2-SPCA	LASSO-SPCA
	In-sample R^2										
2.5-minute	0.504	0.525	0.532	0.563	0.567	0.505	0.505	0.505	0.507	0.507	0.507
5-minute	0.507	0.523	0.557	0.542	0.563	0.514	0.516	0.514	0.513	0.518	0.516
10-minute	0.470	0.476	0.470	0.491	0.491	0.470	0.470	0.470	0.477	0.476	0.476
	Out-of-sample R^2										
2.5-minute	0.075	0.138	0.015	-0.292	-0.245	0.287 (3.65***) (2.34**) (2.68***) (4.15***) (4.49***)	0.287 (3.72***) (2.37**) (2.69***) (4.17***) (4.51***)	0.292 (3.56***) (2.35**) (2.69***) (4.17***) (4.52***)	0.326 (3.71***) (2.67***) (2.87***) (4.37***) (4.77***)	0.323 (3.86***) (2.77***) (2.92***) (4.42***) (4.82***)	0.332 (3.55***) (2.64***) (2.85***) (4.37***) (4.79***)
5-minute	0.115	0.263	0.083	0.201	0.111	0.387 (2.98***) (1.87*) (2.02**) (2.81***) (2.43**)	0.328 (4.39***) (1.06) (1.91*) (1.54) (2.08**)	0.377 (3.06***) (1.87*) (2.02**) (2.81***) (2.44**)	0.442 (3.42***) (2.97***) (2.44**) (4.27***) (3.08***)	0.389 (5.17***) (2.18**) (2.35**) (2.51**) (2.69***)	0.432 (3.21***) (2.47**) (2.27**) (3.73***) (2.81***)
10-minute	0.030	-0.116	0.027	-0.149	-0.146	0.128 (4.13***) (5.29***) (2.90***) (1.88*) (1.90*)	0.144 (3.85***) (4.88***) (2.81***) (1.94*) (1.97**)	0.131 (3.66***) (5.08***) (2.68***) (1.85*) (1.86*)	0.162 (4.38***) (5.28***) (3.25***) (2.06**) (2.08**)	0.174 (3.70***) (4.70***) (2.88***) (2.01**) (2.03**)	0.166 (3.12***) (4.23***) (2.47**) (1.92*) (1.94*)

*Note: See notes to Table 1. CVX belongs to the Energy sector according to the GICS coding system.

Table 13: J.P. Morgan Chase & Co. (JPM)

Sampling Frequency	Model Specification										
	BM-I	BM-II	BM-III	BM-IV	BM-V	EN1-PCA	EN2-PCA	LASSO-PCA	EN1-SPCA	EN2-SPCA	LASSO-SPCA
	In-sample R^2										
2.5-minute	0.565	0.564	0.566	0.564	0.566	0.564	0.564	0.564	0.564	0.564	0.564
5-minute	0.555	0.555	0.554	0.555	0.556	0.557	0.558	0.557	0.561	0.561	0.560
10-minute	0.445	0.445	0.444	0.454	0.458	0.465	0.469	0.474	0.477	0.479	0.479
	Out-of-sample R^2										
2.5-minute	0.021	0.027	0.097	0.013	0.121	-0.014 (***-4.50) (***-4.80) (***-5.50) (***-5.69) (***-6.70)	-0.017 (***-4.15) (***-4.41) (***-5.18) (***-4.26) (***-5.84)	-0.017 (***-4.11) (***-4.39) (***-5.19) (***-4.12) (***-5.81)	-0.014 (***-3.59) (***-3.93) (***-5.25) (***-4.06) (***-6.76)	-0.027 (***-3.56) (***-3.78) (***-4.84) (***-3.80) (***-5.98)	-0.031 (***-3.34) (***-3.56) (***-4.63) (***-3.34) (***-5.44)
5-minute	0.047	0.083	0.069	0.045	0.122	0.074 (2.12**) (-0.71) (0.37) (1.62) (***-3.11)	0.081 (2.94***) (-0.18) (1.09) (2.09**) (***-2.84)	0.073 (2.35**) (-0.86) (0.36) (1.76*) (***-3.76)	0.085 (2.74***) (0.18) (1.16) (2.19**) (**-2.36)	0.087 (2.89***) (0.35) (1.34) (2.32**) (**-2.25)	0.085 (2.74***) (0.19) (1.17) (2.21**) (**-2.34)
10-minute	-0.167	-0.167	-0.180	-0.169	-0.053	-0.284 (-1.11) (-1.22) (-1.00) (-1.21) (**-2.07)	-0.520 (-1.11) (-1.14) (-1.07) (-1.14) (-1.46)	-0.240 (-1.55) (*-1.80) (-1.29) (-1.62) (***-3.40)	-0.684 (-1.14) (-1.16) (-1.11) (-1.16) (-1.39)	-0.498 (-1.25) (-1.28) (-1.20) (-1.29) (*-1.67)	-0.300 (*-1.92) (**-2.19) (*-1.77) (**-2.36) (***-3.65)

*Note: See notes to Table 1. JPM belongs to the Financials sector according to the GICS coding system.

Table 14: General Electric Company (GE)

Sampling Frequency	Model Specification										
	BM-I	BM-II	BM-III	BM-IV	BM-V	EN1-PCA	EN2-PCA	LASSO-PCA	EN1-SPCA	EN2-SPCA	LASSO-SPCA
	In-sample R^2										
2.5-minute	0.538	0.538	0.540	0.542	0.553	0.537	0.537	0.537	0.538	0.538	0.538
5-minute	0.511	0.511	0.510	0.511	0.523	0.511	0.511	0.511	0.519	0.520	0.520
10-minute	0.481	0.481	0.481	0.481	0.482	0.488	0.484	0.484	0.486	0.484	0.485
	Out-of-sample R^2										
2.5-minute	0.279	0.273	0.267	0.260	0.105	0.279 (0.44) (2.85***) (1.88*) (2.27**) (1.06)	0.279 (1.19) (2.89***) (1.89*) (2.28**) (1.06)	0.279 (1.31) (2.89***) (1.89*) (2.28**) (1.06)	0.285 (2.88***) (2.93***) (2.20**) (2.56**) (1.10)	0.285 (3.01***) (3.00***) (2.23**) (2.59***) (1.10)	0.285 (2.82***) (2.91***) (2.19**) (2.55**) (1.10)
5-minute	0.163	0.156	0.161	0.155	0.114	0.164 (1.51) (4.73***) (3.46***) (2.94***) (1.43)	0.163 (-0.40) (3.59***) (0.76) (2.67***) (1.38)	0.164 (0.22) (4.54***) (1.59) (3.25***) (1.40)	0.184 (4.30***) (6.37***) (4.83***) (7.24***) (1.91*)	0.183 (4.24***) (6.25***) (4.75***) (7.41***) (1.90*)	0.184 (4.58***) (6.66***) (5.11***) (7.47***) (1.92*)
10-minute	0.105	0.074	0.107	0.102	-0.006	0.158 (2.79***) (3.59***) (2.66***) (2.86***) (1.37)	0.152 (3.59***) (4.28***) (3.44***) (3.73***) (1.39)	0.161 (4.46***) (5.19***) (4.20***) (4.50***) (1.41)	0.170 (4.07***) (4.78***) (3.88***) (4.10***) (1.46)	0.150 (3.72***) (4.43***) (3.59***) (3.86***) (1.37)	0.157 (4.91***) (5.43***) (4.70***) (5.02***) (1.43)

*Note: See notes to Table 1. GE belongs to the Industrials sector according to the GICS coding system.

Table 15: Microsoft Corporation (MSFT)

Sampling Frequency	Model Specification										
	BM-I	BM-II	BM-III	BM-IV	BM-V	EN1-PCA	EN2-PCA	LASSO-PCA	EN1-SPCA	EN2-SPCA	LASSO-SPCA
	In-sample R^2										
2.5-minute	0.609	0.610	0.610	0.609	0.616	0.611	0.610	0.610	0.612	0.612	0.612
5-minute	0.603	0.606	0.610	0.609	0.610	0.609	0.609	0.609	0.610	0.610	0.610
10-minute	0.514	0.522	0.514	0.517	0.517	0.517	0.518	0.518	0.519	0.521	0.521
	Out-of-sample R^2										
2.5-minute	0.235	0.253	0.198	0.248	0.274	0.264 (1.76*) (2.18**) (1.64*) (2.24**) (-0.29)	0.263 (1.90*) (3.01***) (1.69*) (2.73***) (-0.33)	0.263 (1.78*) (2.27**) (1.65*) (2.35**) (-0.33)	0.278 (1.91*) (2.22**) (1.75*) (2.42**) (0.14)	0.274 (2.20**) (3.23***) (1.85*) (3.19***) (-0.01)	0.272 (1.98**) (2.43**) (1.76*) (2.58***) (-0.06)
5-minute	0.255	0.269	0.262	0.279	0.267	0.297 (2.28**) (3.10***) (2.57**) (0.70) (1.48)	0.296 (2.34**) (3.36***) (2.59***) (0.66) (1.44)	0.293 (2.33**) (2.92***) (2.19**) (0.51) (1.22)	0.298 (2.17**) (2.76***) (2.44**) (0.74) (1.50)	0.299 (2.22**) (3.24***) (2.75***) (0.80) (1.63)	0.298 (2.23**) (3.39***) (2.69***) (0.73) (1.54)
10-minute	0.136	0.096	0.141	0.177	0.192	0.192 (1.99**) (3.60***) (1.78*) (1.93*) (0.00)	0.193 (2.55**) (4.71***) (2.37**) (1.40) (0.04)	0.189 (2.49**) (4.62***) (2.30**) (0.96) (-0.14)	0.211 (1.91*) (2.98***) (1.73*) (1.80*) (0.84)	0.204 (2.11**) (3.31***) (1.92*) (1.67*) (0.49)	0.206 (2.06**) (3.33***) (1.88*) (1.87*) (0.61)

*Note: See notes to Table 1. MSFT belongs to the Information Technology sector according to the GICS coding system.

Table 16: The Coca-Cola Company (KO)

Sampling Frequency	Model Specification										
	BM-I	BM-II	BM-III	BM-IV	BM-V	EN1-PCA	EN2-PCA	LASSO-PCA	EN1-SPCA	EN2-SPCA	LASSO-SPCA
	In-sample R^2										
2.5-minute	0.413	0.463	0.484	0.418	0.485	0.437	0.436	0.436	0.440	0.440	0.440
5-minute	0.559	0.598	0.632	0.559	0.639	0.594	0.593	0.593	0.596	0.594	0.594
10-minute	0.505	0.515	0.511	0.507	0.510	0.530	0.531	0.531	0.534	0.534	0.536
	Out-of-sample R^2										
2.5-minute	0.066	0.107	-0.038	0.013	-0.003	0.284 (2.38**) (3.84***) (5.11***) (3.78***) (4.68***)	0.283 (2.37**) (3.78***) (5.07***) (3.79***) (4.61***)	0.287 (2.37**) (3.81***) (5.18***) (3.82***) (4.66***)	0.303 (2.36**) (4.19***) (5.33***) (3.68***) (4.75***)	0.305 (2.35**) (4.02***) (5.14***) (3.66***) (4.55***)	0.310 (2.37**) (4.08***) (5.31***) (3.70***) (4.64***)
5-minute	0.198	0.284	0.195	0.195	0.247	0.358 (1.84*) (2.97***) (2.69***) (1.86*) (1.90*)	0.318 (1.76*) (1.34) (1.78*) (1.80*) (1.05)	0.331 (1.89*) (1.90*) (1.99**) (1.93*) (1.25)	0.362 (1.67*) (2.74***) (3.04***) (1.69*) (2.20**)	0.331 (1.74*) (2.18**) (2.09**) (1.78*) (1.33)	0.341 (1.82*) (2.58***) (2.29**) (1.84*) (1.52)
10-minute	0.154	0.175	0.190	0.183	0.184	0.261 (1.83*) (3.01***) (1.40) (1.76*) (1.37)	0.234 (1.83*) (2.82***) (1.27) (1.53) (1.28)	0.274 (2.00**) (3.27***) (1.59) (2.00**) (1.55)	0.283 (1.96**) (2.78***) (1.53) (1.97**) (1.50)	0.243 (1.89*) (2.54**) (1.25) (1.75*) (1.24)	0.292 (2.24**) (3.26***) (1.82*) (2.33**) (1.76*)

*Note: See notes to Table 1. KO belongs to the Consumer Staples sector according to the GICS coding system.

Table 17: Duke Energy Corporation (DUK)

Sampling Frequency	Model Specification										
	BM-I	BM-II	BM-III	BM-IV	BM-V	EN1-PCA	EN2-PCA	LASSO-PCA	EN1-SPCA	EN2-SPCA	LASSO-SPCA
	In-sample R^2										
2.5-minute	0.450	0.489	0.509	0.472	0.509	0.477	0.476	0.476	0.475	0.476	0.476
5-minute	0.469	0.501	0.546	0.487	0.546	0.492	0.492	0.492	0.496	0.497	0.497
10-minute	0.412	0.426	0.421	0.425	0.460	0.422	0.423	0.423	0.422	0.423	0.423
	Out-of-sample R^2										
2.5-minute	0.280	0.208	-0.099	-0.244	-0.106	0.338 (1.43) (4.11***) (4.61***) (10.02***) (4.85***)	0.337 (1.42) (4.18***) (4.64***) (10.04***) (4.88***)	0.337 (1.41) (4.18***) (4.64***) (10.05***) (4.89***)	0.344 (1.42) (4.28***) (4.56***) (9.40***) (4.79***)	0.341 (1.36) (4.46***) (4.62***) (9.40***) (4.85***)	0.347 (1.55) (4.25***) (4.57***) (9.67***) (4.80***)
5-minute	0.262	0.251	0.007	-0.047	0.019	0.356 (2.54**) (2.84***) (3.40***) (8.61***) (3.25***)	0.351 (2.71***) (3.21***) (3.60***) (8.87***) (3.43***)	0.354 (2.61***) (3.19***) (3.59***) (8.58***) (3.42***)	0.380 (2.64***) (2.97***) (3.41***) (8.05***) (3.25***)	0.373 (2.68***) (3.13***) (3.50***) (8.21***) (3.34***)	0.379 (2.63***) (3.06***) (3.47***) (7.95***) (3.32***)
10-minute	0.255	0.152	0.254	0.280	0.223	0.379 (3.70***) (7.26***) (3.24***) (2.33**) (2.10**)	0.370 (3.61***) (6.80***) (3.14***) (2.10**) (1.99**)	0.379 (3.85***) (6.87***) (3.33***) (2.19**) (2.04**)	0.376 (3.18***) (6.78***) (2.87***) (2.18**) (2.03**)	0.377 (3.30***) (6.41***) (2.87***) (2.21**) (1.99**)	0.379 (3.36***) (6.23***) (2.89***) (2.14**) (1.96**)

*Note: See notes to Table 1. DUK belongs to the Utilities sector according to the GICS coding system.

Table 18: Johnson & Johnson (JNJ)

Sampling Frequency	Model Specification										
	BM-I	BM-II	BM-III	BM-IV	BM-V	EN1-PCA	EN2-PCA	LASSO-PCA	EN1-SPCA	EN2-SPCA	LASSO-SPCA
	In-sample R^2										
2.5-minute	0.319	0.404	0.432	0.399	0.432	0.374	0.374	0.374	0.368	0.370	0.370
5-minute	0.384	0.436	0.488	0.456	0.493	0.424	0.424	0.424	0.428	0.428	0.428
10-minute	0.399	0.477	0.435	0.408	0.439	0.428	0.428	0.428	0.427	0.427	0.428
	Out-of-sample R^2										
2.5-minute	0.122	0.017	-0.012	-0.374	-0.009	0.242 (2.05**) (3.55***) (2.73***) (5.73***) (4.00***)	0.236 (1.88*) (3.55***) (2.66***) (5.66***) (3.89***)	0.236 (1.83*) (3.55***) (2.64***) (5.63***) (3.86***)	0.254 (1.80*) (2.98***) (2.32**) (5.76***) (3.18***)	0.246 (1.62) (2.88***) (2.22**) (5.64***) (3.02***)	0.255 (1.64*) (2.80***) (2.19**) (5.64***) (2.93***)
5-minute	0.140	0.220	0.140	-0.079	0.101	0.333 (2.39**) (3.66***) (3.31***) (5.21***) (3.55***)	0.360 (2.50**) (3.74***) (3.60***) (5.57***) (3.94***)	0.337 (2.35**) (3.73***) (3.43***) (5.30***) (3.69***)	0.340 (2.29**) (3.66***) (3.37***) (5.14***) (3.61***)	0.364 (2.25**) (3.36***) (3.54***) (5.53***) (4.06***)	0.351 (2.11**) (3.18***) (3.41***) (5.48***) (3.95***)
10-minute	0.125	0.090	0.222	0.180	0.169	0.337 (2.21**) (3.52***) (1.49) (2.43**) (1.49)	0.337 (2.12**) (3.73***) (1.43) (2.33**) (1.44)	0.337 (2.25**) (3.58***) (1.53) (2.53**) (1.51)	0.344 (2.07**) (3.88***) (1.40) (2.20**) (1.42)	0.351 (2.08**) (4.13***) (1.45) (2.24**) (1.46)	0.349 (2.13**) (4.03***) (1.48) (2.32**) (1.48)

*Note: See notes to Table 1. JNJ belongs to the Health Care sector according to the GICS coding system.

Table 19: McDonald's Corporation (MCD)

Sampling Frequency	Model Specification										
	BM-I	BM-II	BM-III	BM-IV	BM-V	EN1-PCA	EN2-PCA	LASSO-PCA	EN1-SPCA	EN2-SPCA	LASSO-SPCA
	In-sample R^2										
2.5-minute	0.289	0.339	0.355	0.327	0.355	0.313	0.313	0.313	0.311	0.311	0.310
5-minute	0.380	0.428	0.469	0.436	0.470	0.395	0.395	0.395	0.401	0.400	0.401
10-minute	0.327	0.350	0.343	0.339	0.343	0.357	0.360	0.363	0.362	0.366	0.368
	Out-of-sample R^2										
2.5-minute	-0.116	-0.297	-0.831	-0.572	-0.855	0.082 (2.34**) (3.99***) (4.37***) (4.41***) (5.08***)	0.068 (2.14**) (4.11***) (4.47***) (4.50***) (5.21***)	0.066 (2.12**) (4.19***) (4.52***) (4.57***) (5.28***)	0.163 (2.97***) (3.96***) (4.24***) (4.33***) (4.85***)	0.144 (2.87***) (4.03***) (4.32***) (4.40***) (4.96***)	0.145 (2.91***) (4.12***) (4.38***) (4.47***) (5.03***)
5-minute	-0.062	-0.119	-0.453	-0.334	-0.513	0.188 (2.62***) (4.50***) (4.36***) (3.67***) (5.07***)	0.181 (2.52**) (4.43***) (4.31***) (3.63***) (5.01***)	0.185 (2.66***) (4.50***) (4.39***) (3.68***) (5.12***)	0.230 (2.79***) (5.38***) (4.68***) (4.02***) (5.37***)	0.234 (2.73***) (5.12***) (4.57***) (3.91***) (5.23***)	0.236 (2.85***) (5.23***) (4.63***) (4.00***) (5.31***)
10-minute	-0.107	-0.006	-0.193	-0.117	-0.164	0.089 (2.63***) (1.71*) (3.15***) (3.06***) (3.39***)	0.106 (2.85***) (1.90*) (3.22***) (3.11***) (3.38***)	0.115 (2.79***) (1.91*) (3.15***) (3.04***) (3.31***)	0.110 (2.27**) (1.62) (2.83***) (2.73***) (3.04***)	0.102 (2.57**) (1.70*) (3.02***) (2.87***) (3.17***)	0.101 (2.54**) (1.69*) (3.00***) (2.85***) (3.15***)

*Note: See notes to Table 1. MCD belongs to the Consumer Discretionary sector according to the GICS coding system.

Table 20: Factor Structure I (SPY)

Ticker	Sector	EN1-PCA	Ticker	Sector	EN1-SPCA	Ticker	Sector	EN2-PCA	Ticker	Sector	EN2-SPCA	Ticker	Sector	LASSO-PCA	Ticker	Sector	LASSO-SPCA
Sampling Frequency: 2.5 Minute																	
CCL	CD	0.133	CCL	CD	0.140	CCL	CD	0.133	CCL	CD	0.140	CCL	CD	0.132	CCL	CD	0.139
TROW	F	0.133	TROW	F	0.137	TROW	F	0.133	ADP	IT	0.137	TROW	F	0.132	ADP	IT	0.136
ADP	IT	0.132	ADP	IT	0.137	ADP	IT	0.133	TROW	F	0.137	ADP	IT	0.132	TROW	F	0.136
AXP	F	0.131	AXP	F	0.133	AXP	F	0.131	AXP	F	0.133	AXP	F	0.130	AXP	F	0.132
PFG	F	0.130	PFG	F	0.130	PFG	F	0.130	TGT	CD	0.130	PFG	F	0.129	TGT	CD	0.129
TGT	CD	0.129	TGT	CD	0.129	TGT	CD	0.130	PFG	F	0.130	TGT	CD	0.129	PFG	F	0.129
BBY	CD	0.128	LUK	F	0.128	BBY	CD	0.129	LUK	F	0.128	BBY	CD	0.128	LUK	F	0.127
LUK	F	0.128	BBY	CD	0.127	LUK	F	0.128	BBY	CD	0.128	LUK	F	0.127	BBY	CD	0.127
COST	CS	0.127	TIF	CD	0.124	COST	CS	0.128	TIF	CD	0.125	COST	CS	0.127	TIF	CD	0.124
TIF	CD	0.127	SBUX	CD	0.124	TIF	CD	0.127	TJX	CD	0.125	TIF	CD	0.126	SBUX	CD	0.124
SBUX	CD	0.126	TJX	CD	0.124	TJX	CD	0.127	SBUX	CD	0.125	SBUX	CD	0.126	TJX	CD	0.124
TJX	CD	0.126	COST	CS	0.123	SBUX	CD	0.127	COST	CS	0.125	TJX	CD	0.126	COST	CS	0.124
YUM	CD	0.126	YUM	CD	0.123	YUM	CD	0.126	YUM	CD	0.124	YUM	CD	0.125	YUM	CD	0.123
CMA	F	0.124	CMA	F	0.117	CMA	F	0.124	CMA	F	0.117	CMA	F	0.123	CMA	F	0.116
BK	F	0.121	BK	F	0.110	BK	F	0.121	STR	E	0.113	BK	F	0.120	STR	E	0.112
Sampling Frequency: 5 Minute																	
Ticker	Sector	EN1-PCA	Ticker	Sector	EN1-SPCA	Ticker	Sector	EN2-PCA	Ticker	Sector	EN2-SPCA	Ticker	Sector	LASSO-PCA	Ticker	Sector	LASSO-SPCA
DVN	E	0.111	DVN	E	0.108	DVN	E	0.111	UTX	I	0.109	UTX	I	0.110	UTX	I	0.108
UTX	I	0.109	UTX	I	0.108	UTX	I	0.109	DVN	E	0.107	DVN	E	0.109	DVN	E	0.106
JCP	CD	0.107	ECL	M	0.103	JCP	CD	0.108	ECL	M	0.104	JCP	CD	0.107	ECL	M	0.102
BBY	CD	0.107	JCP	CD	0.101	BBY	CD	0.107	ADP	IT	0.102	BBY	CD	0.107	ADP	IT	0.100
ECL	M	0.107	ADP	IT	0.101	ECL	M	0.107	JCP	CD	0.101	ECL	M	0.107	JCP	CD	0.100
TGT	CD	0.106	BBY	CD	0.100	ADP	IT	0.107	BBY	CD	0.101	ADP	IT	0.107	BBY	CD	0.099
ADP	IT	0.106	R	I	0.099	TGT	CD	0.107	R	I	0.100	TGT	CD	0.107	R	I	0.099
R	I	0.106	TGT	CD	0.098	R	I	0.106	TGT	CD	0.099	R	I	0.106	TGT	CD	0.098
TIF	CD	0.106	DO	E	0.098	TIF	CD	0.106	TIF	CD	0.098	TIF	CD	0.106	TIF	CD	0.097
AAPL	IT	0.105	TIF	CD	0.098	AAPL	IT	0.106	AAPL	IT	0.098	AAPL	IT	0.106	AAPL	IT	0.097
DO	E	0.105	AAPL	IT	0.097	DO	E	0.105	DO	E	0.097	DO	E	0.103	DO	E	0.096
DHI	CD	0.101	CSX	I	0.090	DHI	CD	0.101	CSX	I	0.091	DHI	CD	0.101	CSX	I	0.089
CSX	I	0.100	VLO	E	0.090	CSX	I	0.101	BK	F	0.090	BK	F	0.101	BK	F	0.089
VLO	E	0.100	BK	F	0.089	BK	F	0.101	VLO	E	0.089	CSX	I	0.101	VLO	E	0.088
BK	F	0.100	PVH	CD	0.088	PVH	CD	0.101	PVH	CD	0.089	PVH	CD	0.100	PVH	CD	0.088
Sampling Frequency: 10 Minute																	
Ticker	Sector	EN1-PCA	Ticker	Sector	EN1-SPCA	Ticker	Sector	EN2-PCA	Ticker	Sector	EN2-SPCA	Ticker	Sector	LASSO-PCA	Ticker	Sector	LASSO-SPCA
PEG	U	0.104	PEG	U	0.101	PEG	U	0.106	PEG	U	0.103	PEG	U	0.106	PEG	U	0.102
AEP	U	0.104	AEP	U	0.100	AEP	U	0.105	AEP	U	0.101	AEP	U	0.105	AEP	U	0.101
SRE	U	0.101	SRE	U	0.097	SRE	U	0.103	SRE	U	0.099	SRE	U	0.102	SRE	U	0.098
BBBY	CD	0.097	GPC	CD	0.096	NVLS	IT	0.099	NVLS	IT	0.096	PNC	F	0.097	BEN	F	0.094
GPC	CD	0.095	BBBY	CD	0.095	DGX	HC	0.096	BEN	F	0.094	DGX	HC	0.096	UTX	I	0.092
PNC	F	0.095	BEN	F	0.092	BEN	F	0.096	UTX	I	0.092	BEN	F	0.096	PNC	F	0.092
BEN	F	0.095	PFG	F	0.091	FCX	M	0.096	INTU	IT	0.092	FCX	M	0.096	INTU	IT	0.092
DGX	HC	0.095	UTX	I	0.090	INTU	IT	0.096	PFG	F	0.092	INTU	IT	0.096	PFG	F	0.091
FCX	M	0.095	PNC	F	0.090	TIF	CD	0.096	FCX	M	0.092	TIF	CD	0.096	FCX	M	0.091
INTU	IT	0.094	INTU	IT	0.090	K	CS	0.095	TIF	CD	0.091	K	CS	0.095	TIF	CD	0.091
TIF	CD	0.094	FCX	M	0.090	PFG	F	0.095	LUK	F	0.091	PFG	F	0.095	LUK	F	0.091
AZO	CD	0.094	TIF	CD	0.090	GIS	CS	0.095	ALL	F	0.091	GIS	CS	0.095	ALL	F	0.091
PFG	F	0.094	ALL	F	0.090	BBY	CD	0.095	BBY	CD	0.091	BBY	CD	0.095	BBY	CD	0.091
K	CS	0.094	LUK	F	0.090	BK	F	0.095	DGX	HC	0.090	BK	F	0.095	DGX	HC	0.090
BBY	CD	0.093	BBY	CD	0.089	ALL	F	0.095	UPS	I	0.090	ALL	F	0.094	UPS	I	0.090

*Note: Numerical entries in the columns denoted by "PCA" and "SPCA" indicate the sample averages of the factor loadings (weights) assigned to each stock in the construction of the principal component in the second step of our latent IV factor estimation procedure, based on these two alternative factor estimation methods. Stocks listed in the table are the most frequently chosen ones in the first step (variable selection) in our procedure, for the target asset given in the title of the table. In each sampling frequency panel, only stocks with the fifteen largest average factor loadings are included, in descending order, in the interests of space. Finally, averaging is done across all rolling windows in our prediction experiments. For complete details, refer to Sections 3 and 6.

Table 21: Factor Structure II (SPY)

Sampling Frequency: 2.5 Minute														
Stock		EN1			Stock		EN2			Stock		LASSO		
Ticker	Sector	PCA	SPCA	Zero Rate	Ticker	Sector	PCA	SPCA	Zero Rate	Ticker	Sector	PCA	SPCA	Zero Rate
IDXX	HC	0.069	0.023	0.609	IDXX	HC	0.069	0.024	0.610	IDXX	HC	0.069	0.023	0.610
F	CD	0.078	0.038	0.482	F	CD	0.079	0.038	0.480	F	CD	0.078	0.038	0.480
ILMN	HC	0.081	0.033	0.480	ILMN	HC	0.081	0.034	0.479	ILMN	HC	0.081	0.033	0.480
LSI	IT	0.084	0.042	0.430	LSI	IT	0.084	0.042	0.430	LSI	IT	0.083	0.042	0.429
SLM	F	0.086	0.045	0.421	SLM	F	0.086	0.045	0.424	SLM	F	0.085	0.044	0.425
BSX	HC	0.085	0.045	0.411	BSX	HC	0.085	0.045	0.411	BSX	HC	0.085	0.045	0.411
AMD	IT	0.086	0.044	0.389	AMD	IT	0.087	0.044	0.389	AMD	IT	0.086	0.044	0.390
HUM	HC	0.088	0.048	0.365	HUM	HC	0.088	0.048	0.366	HUM	HC	0.088	0.047	0.367
MHS	HC	0.090	0.049	0.341	MHS	HC	0.090	0.049	0.341	MHS	HC	0.090	0.048	0.343
EK	IT	0.097	0.066	0.334	EK	IT	0.097	0.066	0.331	EK	IT	0.096	0.066	0.331
AYE	U	0.094	0.061	0.330	AGN	HC	0.095	0.056	0.296	AGN	HC	0.094	0.056	0.296
AGN	HC	0.095	0.056	0.296	CERN	HC	0.094	0.056	0.288	CERN	HC	0.094	0.055	0.288
CERN	HC	0.094	0.055	0.293	CVH	F	0.094	0.056	0.281	CVH	F	0.093	0.055	0.283
CVH	F	0.094	0.056	0.284	AOS	I	0.096	0.060	0.255	AOS	I	0.096	0.060	0.256
AOS	I	0.096	0.060	0.255	YHOO	IT	0.099	0.065	0.247	YHOO	IT	0.098	0.064	0.248
Sampling Frequency: 5 Minute														
Stock		EN1			Stock		EN2			Stock		LASSO		
Ticker	Sector	PCA	SPCA	Zero Rate	Ticker	Sector	PCA	SPCA	Zero Rate	Ticker	Sector	PCA	SPCA	Zero Rate
GMCR	CS	0.068	0.048	0.490	GMCR	CS	0.068	0.048	0.492	GMCR	CS	0.067	0.047	0.497
IDXX	HC	0.075	0.055	0.400	IDXX	HC	0.076	0.055	0.399	IDXX	HC	0.074	0.054	0.403
ILMN	HC	0.079	0.056	0.385	ILMN	HC	0.079	0.056	0.383	ILMN	HC	0.077	0.054	0.386
UHS	HC	0.083	0.065	0.369	UHS	HC	0.083	0.065	0.368	UHS	HC	0.081	0.063	0.372
F	CD	0.079	0.059	0.367	F	CD	0.079	0.059	0.366	F	CD	0.078	0.058	0.368
SLM	F	0.082	0.059	0.363	SLM	F	0.082	0.059	0.360	SLM	F	0.081	0.058	0.363
CVC	TS	0.082	0.063	0.333	CVC	TS	0.082	0.063	0.330	CVC	TS	0.081	0.062	0.332
BSX	HC	0.082	0.063	0.321	BSX	HC	0.082	0.063	0.319	BSX	HC	0.081	0.062	0.320
MHS	HC	0.091	0.075	0.295	MHS	HC	0.091	0.074	0.293	MHS	HC	0.089	0.073	0.297
MRK	HC	0.091	0.076	0.287	MRK	HC	0.091	0.075	0.284	MRK	HC	0.089	0.074	0.287
DGX	HC	0.095	0.082	0.283	DGX	HC	0.095	0.082	0.279	DGX	HC	0.093	0.080	0.283
CERN	HC	0.086	0.066	0.281	CERN	HC	0.087	0.066	0.277	CERN	HC	0.086	0.065	0.279
PEP	CS	0.090	0.075	0.270	PEP	CS	0.090	0.075	0.267	PEP	CS	0.089	0.074	0.269
CVH	F	0.096	0.082	0.252	AGN	HC	0.091	0.074	0.251	AGN	HC	0.089	0.072	0.254
CCI	RE	0.088	0.069	0.251	CVH	F	0.096	0.082	0.250	CVH	F	0.094	0.080	0.253
Sampling Frequency: 10 Minute														
Stock		EN1			Stock		EN2			Stock		LASSO		
Ticker	Sector	PCA	SPCA	Zero Rate	Ticker	Sector	PCA	SPCA	Zero Rate	Ticker	Sector	PCA	SPCA	Zero Rate
SBR	E	0.074	0.070	0.500	SBR	E	0.076	0.071	0.499	SBR	E	0.075	0.071	0.499
LB	CD	0.075	0.070	0.495	LB	CD	0.077	0.071	0.493	LB	CD	0.077	0.071	0.493
THC	HC	0.083	0.075	0.354	THC	HC	0.085	0.077	0.353	THC	HC	0.085	0.077	0.353
IDXX	HC	0.087	0.080	0.296	IDXX	HC	0.089	0.081	0.292	IDXX	HC	0.088	0.081	0.291
F	CD	0.084	0.075	0.284	F	CD	0.085	0.076	0.282	F	CD	0.085	0.076	0.280
SLM	F	0.087	0.078	0.280	SLM	F	0.088	0.079	0.278	SLM	F	0.088	0.079	0.279
ILMN	HC	0.087	0.078	0.275	ILMN	HC	0.089	0.080	0.273	ILMN	HC	0.089	0.079	0.272
NYT	CD	0.087	0.078	0.273	NYT	CD	0.089	0.080	0.271	NYT	CD	0.089	0.080	0.271
DGX	HC	0.095	0.088	0.257	MYL	HC	0.087	0.078	0.271	MYL	HC	0.087	0.077	0.270
CERN	HC	0.089	0.080	0.250	DGX	HC	0.096	0.090	0.256	DGX	HC	0.096	0.090	0.255
HRL	CS	0.092	0.086	0.243	CERN	HC	0.090	0.082	0.249	CERN	HC	0.090	0.081	0.249
GIS	CS	0.093	0.087	0.241	HRL	CS	0.094	0.087	0.241	HRL	CS	0.094	0.087	0.241
JNJ	HC	0.092	0.086	0.224	GIS	CS	0.095	0.089	0.240	GIS	CS	0.095	0.089	0.239
CCI	RE	0.087	0.079	0.222	GPN	F	0.088	0.080	0.226	GPN	F	0.088	0.080	0.226
BLK	F	0.089	0.082	0.220	CCI	RE	0.089	0.080	0.219	JNJ	HC	0.094	0.087	0.220

*Note: See notes to Table 20. In this table, rather than tabulating sample averages of the factor loadings (weights) assigned to each stock that are the most frequently chosen, tabulated entries correspond to sample averages of factor loading associated with stocks that are the least frequently selected, as indicated by the percentage of times that SPCA assigns identically zero weights to them. This percentage is called "Zero Rate" in the table, and results are reported for stocks with the fifteen largest average "zero rates", in descending order.

Table 22: Factor Structure I (XLE)

Ticker	Sector	EN1-PCA	Ticker	Sector	EN1-SPCA	Ticker	Sector	EN2-PCA	Ticker	Sector	EN2-SPCA	Ticker	Sector	LASSO-PCA	Ticker	Sector	LASSO-SPCA
Sampling Frequency: 2.5 Minute																	
CCL	CD	0.132	CCL	CD	0.137	CCL	CD	0.131	CCL	CD	0.137	CCL	CD	0.131	CCL	CD	0.138
TROW	F	0.132	TROW	F	0.135	TROW	F	0.130	TROW	F	0.133	TROW	F	0.131	TROW	F	0.134
ADP	IT	0.131	ADP	IT	0.133	ADP	IT	0.130	ADP	IT	0.133	ADP	IT	0.130	ADP	IT	0.133
AXP	F	0.130	LLTC	IT	0.132	AXP	F	0.129	LLTC	IT	0.132	AXP	F	0.129	LLTC	IT	0.132
LLTC	IT	0.130	AXP	F	0.131	LLTC	IT	0.129	AXP	F	0.130	LLTC	IT	0.129	AXP	F	0.131
PFG	F	0.129	PFG	F	0.128	PFG	F	0.128	PFG	F	0.126	PFG	F	0.128	PFG	F	0.127
AAPL	IT	0.128	AAPL	IT	0.126	AAPL	IT	0.127	CSCO	IT	0.126	AAPL	IT	0.128	CSCO	IT	0.127
BBY	CD	0.126	BBY	CD	0.121	CSCO	IT	0.127	AAPL	IT	0.126	CSCO	IT	0.127	AAPL	IT	0.126
SBUX	CD	0.125	SBUX	CD	0.121	BBY	CD	0.125	SBUX	CD	0.121	BBY	CD	0.125	BBY	CD	0.120
COST	CS	0.124	TIF	CD	0.118	SBUX	CD	0.124	BBY	CD	0.120	COST	CS	0.123	YUM	CD	0.118
TIF	CD	0.124	YUM	CD	0.118	COST	CS	0.123	YUM	CD	0.118	TIF	CD	0.123	TIF	CD	0.118
YUM	CD	0.123	COST	CS	0.117	TIF	CD	0.123	TIF	CD	0.117	YUM	CD	0.123	COST	CS	0.116
CMA	F	0.123	CMA	F	0.115	YUM	CD	0.122	COST	CS	0.116	CMA	F	0.122	CMA	F	0.115
BK	F	0.120	BK	F	0.107	CMA	F	0.122	CMA	F	0.114	BK	F	0.119	BK	F	0.107
CSX	I	0.119	CSX	I	0.107	BK	F	0.119	BK	F	0.107	CSX	I	0.117	CSX	I	0.105
Sampling Frequency: 5 Minute																	
Ticker	Sector	EN1-PCA	Ticker	Sector	EN1-SPCA	Ticker	Sector	EN2-PCA	Ticker	Sector	EN2-SPCA	Ticker	Sector	LASSO-PCA	Ticker	Sector	LASSO-SPCA
TMK	F	0.110	TMK	F	0.109	TMK	F	0.110	TMK	F	0.108	TMK	F	0.110	TMK	F	0.107
BHI	E	0.108	UTX	I	0.106	UTX	I	0.107	UTX	I	0.105	UTX	I	0.107	UTX	I	0.104
UTX	I	0.107	BHI	E	0.105	TROW	F	0.107	TROW	F	0.101	TROW	F	0.106	TROW	F	0.100
TROW	F	0.106	TROW	F	0.101	LLTC	IT	0.106	LLTC	IT	0.100	LLTC	IT	0.105	LLTC	IT	0.099
JCP	CD	0.106	ECL	M	0.101	JCP	CD	0.105	ECL	M	0.100	ADP	IT	0.104	ECL	M	0.098
LLTC	IT	0.106	LLTC	IT	0.101	ADP	IT	0.105	ADP	IT	0.098	BBY	CD	0.104	ADP	IT	0.097
TGT	CD	0.106	JCP	CD	0.100	BBY	CD	0.105	JCP	CD	0.097	ECL	M	0.104	BBY	CD	0.095
BBY	CD	0.106	ADP	IT	0.099	TGT	CD	0.105	BBY	CD	0.097	AAPL	IT	0.103	TIF	CD	0.094
ECL	M	0.105	BBY	CD	0.099	ECL	M	0.105	TGT	CD	0.096	TIF	CD	0.103	TXN	IT	0.093
ADP	IT	0.105	TGT	CD	0.098	TIF	CD	0.104	TIF	CD	0.095	TXN	IT	0.102	AAPL	IT	0.093
TIF	CD	0.105	TIF	CD	0.097	AAPL	IT	0.104	TXN	IT	0.094	AFL	F	0.101	AFL	F	0.090
AAPL	IT	0.104	TXN	IT	0.096	TXN	IT	0.103	AAPL	IT	0.094	BK	F	0.099	PCP	I	0.089
TXN	IT	0.104	AAPL	IT	0.095	AFL	F	0.101	FMC	M	0.092	PCP	I	0.098	BK	F	0.087
FMC	M	0.101	FMC	M	0.094	BK	F	0.100	AFL	F	0.091	CSX	I	0.098	CSX	I	0.086
BK	F	0.100	PCP	I	0.091	FMC	M	0.099	PCP	I	0.090	DHI	CD	0.098	CTXS	IT	0.085
Sampling Frequency: 10 Minute																	
Ticker	Sector	EN1-PCA	Ticker	Sector	EN1-SPCA	Ticker	Sector	EN2-PCA	Ticker	Sector	EN2-SPCA	Ticker	Sector	LASSO-PCA	Ticker	Sector	LASSO-SPCA
PEG	U	0.104	PEG	U	0.100	PEG	U	0.107	PEG	U	0.104	PEG	U	0.108	PEG	U	0.104
AEP	U	0.103	AEP	U	0.099	AEP	U	0.107	AEP	U	0.103	AEP	U	0.107	AEP	U	0.103
SRE	U	0.102	SRE	U	0.098	SRE	U	0.106	SRE	U	0.102	SRE	U	0.106	SRE	U	0.102
NVLS	IT	0.101	NVLS	IT	0.097	NVLS	IT	0.103	BBBY	CD	0.099	CVH	F	0.102	BBBY	CD	0.100
CVH	F	0.100	BBBY	CD	0.097	CVH	F	0.101	NVLS	IT	0.099	BBBY	CD	0.101	BEN	F	0.096
BBBY	CD	0.098	CVH	F	0.094	BBBY	CD	0.101	BEN	F	0.096	FCX	M	0.100	FCX	M	0.095
VLO	E	0.096	BEN	F	0.093	FCX	M	0.099	FCX	M	0.095	GIS	CS	0.099	CVH	F	0.095
MAA	RE	0.096	INTU	IT	0.092	VLO	E	0.099	INTU	IT	0.095	TSO	E	0.099	UTX	I	0.095
INTU	IT	0.096	PFG	F	0.092	TSO	E	0.099	CVH	F	0.095	LM	F	0.099	PFG	F	0.095
TSO	E	0.096	UTX	I	0.092	LM	F	0.099	PFG	F	0.095	K	CS	0.099	INTU	IT	0.095
LM	F	0.096	MAA	RE	0.092	GIS	CS	0.099	UTX	I	0.095	TIF	CD	0.099	LM	F	0.094
TIF	CD	0.096	TIF	CD	0.091	TIF	CD	0.098	LM	F	0.094	BK	F	0.099	TIF	CD	0.094
BK	F	0.096	LM	F	0.091	INTU	IT	0.098	TIF	CD	0.094	INTU	IT	0.099	ALL	F	0.094
PFG	F	0.095	ALL	F	0.091	K	CS	0.098	ALL	F	0.094	PFG	F	0.099	UPS	I	0.094
GIS	CS	0.095	UPS	I	0.091	BK	F	0.098	UPS	I	0.093	BEN	F	0.098	GIS	CS	0.093

*Note: See notes to Table 20.

Table 23: Factor Structure II (XLE)

Sampling Frequency: 2.5 Minute														
Stock		EN1			Stock		EN2			Stock		LASSO		
Ticker	Sector	PCA	SPCA	Zero Rate	Ticker	Sector	PCA	SPCA	Zero Rate	Ticker	Sector	PCA	SPCA	Zero Rate
IDXX	HC	0.068	0.021	0.636	IDXX	HC	0.067	0.022	0.633	IDXX	HC	0.068	0.022	0.631
PRGO	HC	0.076	0.028	0.588	PRGO	HC	0.075	0.029	0.580	F	CD	0.077	0.036	0.494
F	CD	0.077	0.036	0.505	F	CD	0.077	0.036	0.496	ILMN	HC	0.080	0.032	0.493
ILMN	HC	0.080	0.031	0.502	ILMN	HC	0.080	0.032	0.495	SLM	F	0.084	0.043	0.440
SLM	F	0.085	0.043	0.447	SLM	F	0.084	0.042	0.442	HRL	CS	0.086	0.045	0.401
LSI	IT	0.084	0.041	0.439	LSI	IT	0.083	0.042	0.431	AMD	IT	0.086	0.043	0.397
BSX	HC	0.084	0.042	0.438	HRL	CS	0.086	0.044	0.403	HUM	HC	0.086	0.044	0.389
HRL	CS	0.086	0.043	0.416	AMD	IT	0.086	0.043	0.398	AGN	HC	0.093	0.053	0.316
AMD	IT	0.086	0.043	0.405	HUM	HC	0.086	0.044	0.392	DV	CD	0.094	0.056	0.313
HUM	HC	0.087	0.044	0.400	AGN	HC	0.092	0.052	0.319	CVH	F	0.092	0.052	0.302
MHS	HC	0.088	0.044	0.389	DV	CD	0.094	0.055	0.315	ABC	HC	0.095	0.061	0.275
EK	IT	0.095	0.062	0.345	CVH	F	0.091	0.052	0.306	AOS	I	0.094	0.057	0.264
AGN	HC	0.093	0.051	0.333	ABC	HC	0.095	0.060	0.278	SBAC	RE	0.095	0.058	0.259
DV	CD	0.094	0.055	0.323	AOS	I	0.094	0.057	0.266	OI	M	0.104	0.085	0.222
CVH	F	0.092	0.051	0.314	SBAC	RE	0.095	0.058	0.261	GPN	F	0.098	0.063	0.208
Sampling Frequency: 5 Minute														
Stock		EN1			Stock		EN2			Stock		LASSO		
Ticker	Sector	PCA	SPCA	Zero Rate	Ticker	Sector	PCA	SPCA	Zero Rate	Ticker	Sector	PCA	SPCA	Zero Rate
F	CD	0.078	0.059	0.376	UHS	HC	0.080	0.064	0.385	UHS	HC	0.079	0.063	0.387
UHS	HC	0.084	0.068	0.374	F	CD	0.076	0.057	0.381	F	CD	0.075	0.056	0.381
SLM	F	0.083	0.062	0.363	SLM	F	0.080	0.059	0.364	SLM	F	0.079	0.058	0.366
CVC	TS	0.082	0.064	0.341	CVC	TS	0.080	0.062	0.340	CVC	TS	0.079	0.062	0.343
BSX	HC	0.082	0.063	0.331	BSX	HC	0.080	0.061	0.335	BSX	HC	0.079	0.060	0.336
ABT	HC	0.089	0.076	0.322	ABT	HC	0.086	0.072	0.328	ABT	HC	0.084	0.071	0.330
MHS	HC	0.093	0.078	0.311	MHS	HC	0.088	0.073	0.317	MHS	HC	0.087	0.072	0.320
AMD	IT	0.083	0.062	0.310	AMD	IT	0.082	0.060	0.315	AMD	IT	0.081	0.059	0.318
CERN	HC	0.087	0.068	0.290	CERN	HC	0.085	0.066	0.292	CERN	HC	0.084	0.065	0.293
PEP	CS	0.089	0.075	0.274	CVH	F	0.094	0.082	0.270	CVH	F	0.092	0.080	0.272
CVH	F	0.098	0.086	0.268	CCI	RE	0.085	0.066	0.259	CCI	RE	0.084	0.065	0.260
CCI	RE	0.086	0.068	0.259	SRE	U	0.090	0.077	0.257	SRE	U	0.089	0.077	0.256
GPN	F	0.090	0.072	0.254	GPN	F	0.088	0.070	0.248	GPN	F	0.087	0.069	0.248
SRE	U	0.092	0.080	0.254	CAH	HC	0.092	0.079	0.245	CAH	HC	0.091	0.078	0.248
CAH	HC	0.095	0.082	0.244	PCLN	CD	0.088	0.067	0.241	PCLN	CD	0.087	0.066	0.242
Sampling Frequency: 10 Minute														
Stock		EN1			Stock		EN2			Stock		LASSO		
Ticker	Sector	PCA	SPCA	Zero Rate	Ticker	Sector	PCA	SPCA	Zero Rate	Ticker	Sector	PCA	SPCA	Zero Rate
SBR	E	0.077	0.072	0.500	SBR	E	0.080	0.075	0.498	SBR	E	0.081	0.075	0.498
LB	CD	0.078	0.072	0.495	BRK.B	F	0.087	0.079	0.372	BRK.B	F	0.087	0.079	0.371
BRK.B	F	0.085	0.077	0.376	IDXX	HC	0.091	0.083	0.289	IDXX	HC	0.091	0.083	0.289
IDXX	HC	0.088	0.080	0.294	ILMN	HC	0.092	0.082	0.272	ALXN	HC	0.090	0.081	0.289
SLM	F	0.088	0.078	0.280	SLM	F	0.091	0.081	0.271	ILMN	HC	0.092	0.082	0.271
NYT	CD	0.089	0.080	0.267	NYT	CD	0.092	0.083	0.268	CERN	HC	0.094	0.086	0.245
CERN	HC	0.091	0.083	0.248	CERN	HC	0.094	0.085	0.246	HOLX	HC	0.094	0.086	0.245
HOLX	HC	0.091	0.083	0.247	HOLX	HC	0.094	0.086	0.245	CVH	F	0.102	0.095	0.239
HRL	CS	0.093	0.087	0.242	HRL	CS	0.096	0.089	0.239	MRK	HC	0.097	0.090	0.237
CVH	F	0.100	0.094	0.240	CVH	F	0.101	0.095	0.238	HRL	CS	0.097	0.090	0.237
GIS	CS	0.095	0.089	0.232	MRK	HC	0.097	0.090	0.237	GIS	CS	0.099	0.093	0.225
EL	CS	0.093	0.086	0.221	GIS	CS	0.099	0.093	0.226	CCI	RE	0.092	0.083	0.217
GRMN	CD	0.091	0.083	0.221	CCI	RE	0.092	0.083	0.221	JNJ	HC	0.098	0.092	0.215
CCI	RE	0.089	0.080	0.220	JNJ	HC	0.097	0.092	0.213	SRE	U	0.106	0.102	0.209
SRE	U	0.102	0.098	0.216	SRE	U	0.106	0.102	0.212	PEG	U	0.108	0.104	0.208

*Note: See notes to Table 21.

Table 24: Factor Structure I (JNJ)

Ticker	Sector	EN1-PCA	Ticker	Sector	EN1-SPCA	Ticker	Sector	EN2-PCA	Ticker	Sector	EN2-SPCA	Ticker	Sector	LASSO-PCA	Ticker	Sector	LASSO-SPCA
Sampling Frequency: 2.5 Minute																	
CCL	CD	0.130	PH	I	0.139	CCL	CD	0.130	CCL	CD	0.136	CCL	CD	0.131	CCL	CD	0.136
PH	I	0.130	CCL	CD	0.136	AXP	F	0.129	AXP	F	0.129	AXP	F	0.129	AXP	F	0.130
AXP	F	0.129	AXP	F	0.129	PFG	F	0.128	ADP	IT	0.128	PFG	F	0.129	ADP	IT	0.129
PFG	F	0.128	ADP	IT	0.129	ADP	IT	0.128	PFG	F	0.127	ADP	IT	0.128	PFG	F	0.128
ADP	IT	0.128	PFG	F	0.127	TGT	CD	0.128	TGT	CD	0.126	TGT	CD	0.128	TGT	CD	0.127
TGT	CD	0.128	TGT	CD	0.126	JCP	CD	0.127	JCP	CD	0.125	JCP	CD	0.127	JCP	CD	0.125
JCP	CD	0.127	BBY	CD	0.124	BBY	CD	0.126	BBY	CD	0.124	BBY	CD	0.127	BBY	CD	0.125
BBY	CD	0.127	JCP	CD	0.124	LUK	F	0.125	LUK	F	0.123	LUK	F	0.126	LUK	F	0.123
LUK	F	0.125	LUK	F	0.122	TIF	CD	0.125	TIF	CD	0.122	TIF	CD	0.125	TIF	CD	0.122
TIF	CD	0.125	TIF	CD	0.122	COST	CS	0.125	SBUX	CD	0.121	COST	CS	0.125	SBUX	CD	0.122
COST	CS	0.125	SBUX	CD	0.121	SBUX	CD	0.125	YUM	CD	0.119	SBUX	CD	0.125	YUM	CD	0.120
SBUX	CD	0.125	COST	CS	0.119	YUM	CD	0.123	COST	CS	0.119	YUM	CD	0.124	COST	CS	0.120
YUM	CD	0.123	YUM	CD	0.119	CMA	F	0.122	CMA	F	0.114	CMA	F	0.123	CMA	F	0.115
CMA	F	0.122	CMA	F	0.114	GE	I	0.121	GE	I	0.111	GE	I	0.121	GE	I	0.111
GE	I	0.121	GE	I	0.111	CB	F	0.119	CB	F	0.108	CB	F	0.120	CB	F	0.108
Sampling Frequency: 5 Minute																	
Ticker	Sector	EN1-PCA	Ticker	Sector	EN1-SPCA	Ticker	Sector	EN2-PCA	Ticker	Sector	EN2-SPCA	Ticker	Sector	LASSO-PCA	Ticker	Sector	LASSO-SPCA
AXP	F	0.107	ECL	M	0.099	AXP	F	0.108	ECL	M	0.100	AXP	F	0.106	ECL	M	0.098
ECL	M	0.107	AXP	F	0.096	ECL	M	0.107	AXP	F	0.098	ECL	M	0.105	AXP	F	0.096
PFG	F	0.107	PFG	F	0.094	PFG	F	0.107	PFG	F	0.096	PFG	F	0.105	PFG	F	0.094
ADP	IT	0.106	ADP	IT	0.093	ADP	IT	0.106	ADP	IT	0.095	ADP	IT	0.104	ADP	IT	0.093
BBY	CD	0.106	BBY	CD	0.093	BBY	CD	0.106	BBY	CD	0.094	BBY	CD	0.104	BBY	CD	0.092
AAPL	IT	0.105	AAPL	IT	0.091	AAPL	IT	0.105	AAPL	IT	0.093	AAPL	IT	0.103	AAPL	IT	0.091
TGT	CD	0.105	TGT	CD	0.091	TGT	CD	0.105	TGT	CD	0.092	TGT	CD	0.102	TGT	CD	0.090
GE	I	0.103	DO	E	0.088	GE	I	0.103	GE	I	0.089	GE	I	0.101	GE	I	0.087
HOT	CD	0.102	GE	I	0.087	HOT	CD	0.102	HOT	CD	0.089	HOT	CD	0.100	HOT	CD	0.087
BK	F	0.100	HOT	CD	0.087	BK	F	0.100	BK	F	0.087	BK	F	0.099	BK	F	0.085
CB	F	0.099	BK	F	0.085	CB	F	0.100	CB	F	0.086	CB	F	0.098	CB	F	0.085
DHI	CD	0.098	CB	F	0.084	DHI	CD	0.099	CTXS	IT	0.083	DHI	CD	0.097	CTXS	IT	0.081
CTXS	IT	0.098	CTXS	IT	0.081	CTXS	IT	0.098	DHI	CD	0.083	CTXS	IT	0.097	DHI	CD	0.081
DO	E	0.095	DHI	CD	0.081	RJF	F	0.096	RJF	F	0.079	RJF	F	0.094	RJF	F	0.078
WYNN	CD	0.095	VLO	E	0.078	WYNN	CD	0.095	VLO	E	0.079	WYNN	CD	0.093	VLO	E	0.077
Sampling Frequency: 10 Minute																	
Ticker	Sector	EN1-PCA	Ticker	Sector	EN1-SPCA	Ticker	Sector	EN2-PCA	Ticker	Sector	EN2-SPCA	Ticker	Sector	LASSO-PCA	Ticker	Sector	LASSO-SPCA
DVN	E	0.071	DVN	E	0.067	DVN	E	0.072	DVN	E	0.069	DVN	E	0.071	DVN	E	0.067
BHI	E	0.070	BHI	E	0.067	BHI	E	0.071	BHI	E	0.068	BHI	E	0.069	BHI	E	0.066
AEP	U	0.069	AEP	U	0.066	AEP	U	0.070	AEP	U	0.067	AEP	U	0.068	AEP	U	0.065
SWN	E	0.068	DOV	I	0.065	XOM	E	0.070	XOM	E	0.067	SWN	E	0.067	DOV	I	0.065
SRE	U	0.068	SRE	U	0.064	SWN	E	0.069	DOV	I	0.067	SRE	U	0.067	SRE	U	0.064
DOV	I	0.066	SWN	E	0.064	SRE	U	0.069	SRE	U	0.066	DOV	I	0.066	SWN	E	0.064
ITW	I	0.065	ITW	I	0.062	DOV	I	0.068	SWN	E	0.065	INTU	IT	0.064	UTX	I	0.061
INTU	IT	0.064	UTX	I	0.061	ITW	I	0.066	ITW	I	0.064	TIF	CD	0.063	INTU	IT	0.060
TIF	CD	0.064	FISV	IT	0.061	INTU	IT	0.066	UTX	I	0.063	UTX	I	0.063	TIF	CD	0.059
FISV	IT	0.064	INTU	IT	0.060	TIF	CD	0.065	INTU	IT	0.061	MTB	F	0.063	DHR	HC	0.059
MTB	F	0.064	TIF	CD	0.060	UTX	I	0.065	TIF	CD	0.061	COH	CD	0.063	LUK	F	0.059
UTX	I	0.064	DHR	HC	0.059	MTB	F	0.065	DHR	HC	0.061	BK	F	0.063	UPS	I	0.059
COH	CD	0.064	UPS	I	0.059	COH	CD	0.065	LUK	F	0.061	AN	CD	0.063	MTB	F	0.059
BK	F	0.063	LUK	F	0.059	AN	CD	0.065	UPS	I	0.061	DHR	HC	0.063	COH	CD	0.059
AN	CD	0.063	MTB	F	0.059	BK	F	0.065	COH	CD	0.060	LUK	F	0.063	BK	F	0.058

*Note: See notes to Table 20.

Table 25: Factor Structure II (JNJ)

Sampling Frequency: 2.5 Minute														
Stock		EN1			Stock		EN2			Stock		LASSO		
Ticker	Sector	PCA	SPCA	Zero Rate	Ticker	Sector	PCA	SPCA	Zero Rate	Ticker	Sector	PCA	SPCA	Zero Rate
IDXX	HC	0.068	0.021	0.636	IDXX	HC	0.068	0.021	0.624	IDXX	HC	0.068	0.022	0.622
MO	CS	0.077	0.032	0.528	URI	I	0.076	0.035	0.549	URI	I	0.076	0.035	0.547
ILMN	HC	0.078	0.029	0.526	MO	CS	0.077	0.033	0.518	MO	CS	0.078	0.033	0.515
F	CD	0.076	0.034	0.510	ILMN	HC	0.078	0.029	0.516	ILMN	HC	0.078	0.030	0.514
LSI	IT	0.082	0.039	0.446	F	CD	0.076	0.035	0.504	F	CD	0.076	0.035	0.500
SLM	F	0.085	0.042	0.439	LSI	IT	0.082	0.039	0.443	LSI	IT	0.082	0.039	0.441
MHS	HC	0.087	0.044	0.403	SLM	F	0.085	0.042	0.433	SLM	F	0.085	0.043	0.430
HUM	HC	0.086	0.044	0.392	MHS	HC	0.087	0.044	0.392	MHS	HC	0.088	0.044	0.389
AGN	HC	0.091	0.050	0.335	HUM	HC	0.086	0.044	0.384	HUM	HC	0.087	0.045	0.383
CVH	F	0.092	0.051	0.330	AGN	HC	0.091	0.051	0.329	AGN	HC	0.092	0.051	0.326
CERN	HC	0.091	0.050	0.327	CVH	F	0.092	0.052	0.319	KR	CS	0.094	0.056	0.315
KR	CS	0.094	0.055	0.319	KR	CS	0.094	0.056	0.317	CVH	F	0.092	0.052	0.315
ABC	HC	0.094	0.058	0.295	CERN	HC	0.091	0.050	0.317	CERN	HC	0.092	0.051	0.315
AOS	I	0.094	0.055	0.281	ABC	HC	0.094	0.059	0.289	ABC	HC	0.095	0.059	0.287
YHOO	IT	0.096	0.059	0.270	AOS	I	0.094	0.056	0.274	AOS	I	0.094	0.056	0.273
Sampling Frequency: 5 Minute														
Stock		EN1			Stock		EN2			Stock		LASSO		
Ticker	Sector	PCA	SPCA	Zero Rate	Ticker	Sector	PCA	SPCA	Zero Rate	Ticker	Sector	PCA	SPCA	Zero Rate
IDXX	HC	0.069	0.050	0.444	IDXX	HC	0.070	0.050	0.443	IDXX	HC	0.069	0.050	0.446
URI	I	0.071	0.051	0.432	URI	I	0.072	0.051	0.431	URI	I	0.071	0.051	0.432
ILMN	HC	0.072	0.050	0.415	ILMN	HC	0.073	0.051	0.413	ILMN	HC	0.072	0.050	0.413
F	CD	0.073	0.052	0.388	F	CD	0.074	0.052	0.387	F	CD	0.073	0.052	0.390
SLM	F	0.077	0.055	0.374	SLM	F	0.078	0.055	0.371	SLM	F	0.076	0.055	0.374
CVC	TS	0.078	0.058	0.350	CVC	TS	0.079	0.059	0.348	CVC	TS	0.078	0.058	0.350
MHS	HC	0.081	0.067	0.338	MHS	HC	0.082	0.067	0.338	BSX	HC	0.076	0.057	0.342
DGX	HC	0.084	0.074	0.321	DGX	HC	0.085	0.073	0.321	MHS	HC	0.081	0.066	0.341
CERN	HC	0.083	0.062	0.297	CERN	HC	0.084	0.063	0.293	DGX	HC	0.084	0.072	0.324
CVH	F	0.086	0.073	0.292	CVH	F	0.087	0.073	0.292	CERN	HC	0.082	0.062	0.296
SLE	CD	0.081	0.063	0.292	SLE	CD	0.082	0.063	0.289	SLE	CD	0.081	0.062	0.292
SWY	CD	0.083	0.062	0.263	CCI	RE	0.084	0.062	0.259	CCI	RE	0.083	0.061	0.261
CCI	RE	0.084	0.061	0.261	SWY	CD	0.084	0.063	0.258	SWY	CD	0.083	0.062	0.260
PCLN	CD	0.086	0.063	0.244	PCLN	CD	0.087	0.064	0.241	PCLN	CD	0.085	0.063	0.242
GPN	F	0.087	0.066	0.240	GPN	F	0.087	0.067	0.237	GPN	F	0.086	0.065	0.239
Sampling Frequency: 10 Minute														
Stock		EN1			Stock		EN2			Stock		LASSO		
Ticker	Sector	PCA	SPCA	Zero Rate	Ticker	Sector	PCA	SPCA	Zero Rate	Ticker	Sector	PCA	SPCA	Zero Rate
SBR	E	0.042	0.041	0.546	EW	HC	0.059	0.054	0.319	SLM	F	0.055	0.049	0.315
EW	HC	0.057	0.053	0.320	SLM	F	0.057	0.050	0.313	NYT	CD	0.055	0.049	0.312
SLM	F	0.055	0.049	0.313	UNH	HC	0.062	0.058	0.293	MHS	HC	0.059	0.055	0.295
UNH	HC	0.061	0.057	0.293	MHS	HC	0.061	0.057	0.293	UNH	HC	0.060	0.056	0.295
MHS	HC	0.060	0.055	0.293	GIS	CS	0.063	0.059	0.273	GIS	CS	0.061	0.057	0.275
CERN	HC	0.058	0.052	0.279	HRL	CS	0.062	0.058	0.265	HRL	CS	0.060	0.057	0.266
GIS	CS	0.061	0.057	0.273	CCI	RE	0.058	0.052	0.259	CCI	RE	0.056	0.050	0.261
HRL	CS	0.061	0.057	0.266	GPN	F	0.059	0.053	0.250	GPN	F	0.057	0.051	0.251
CCI	RE	0.057	0.050	0.260	K	CS	0.064	0.060	0.247	K	CS	0.062	0.058	0.249
GPN	F	0.058	0.052	0.250	SRE	U	0.069	0.066	0.234	SRE	U	0.067	0.064	0.234
K	CS	0.062	0.058	0.249	MTB	F	0.065	0.060	0.230	MTB	F	0.063	0.059	0.232
SRE	U	0.068	0.064	0.232	AEP	U	0.070	0.067	0.220	WYNN	CD	0.061	0.055	0.221
MTB	F	0.064	0.059	0.230	WYNN	CD	0.063	0.057	0.219	AEP	U	0.068	0.065	0.220
WYNN	CD	0.061	0.056	0.219	AN	CD	0.065	0.060	0.218	AN	CD	0.063	0.058	0.219
AEP	U	0.069	0.066	0.219	SWN	E	0.069	0.065	0.216	BK	F	0.063	0.058	0.216

*Note: See notes to Table 21.

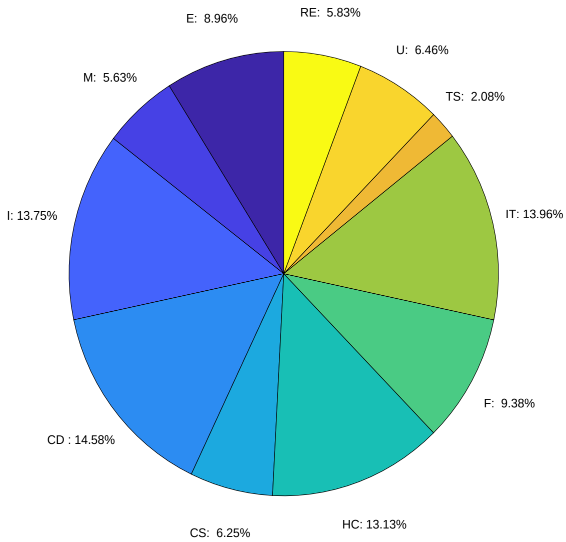

Figure 1: Percentage Shares of Sectors in the S&P 500 Index

*Notes: Following the Global Industry Classification Standard (GICS) coding system, 480 constituents of the S&P 500 index in our dataset are classified into 11 sectors. The percentage of each sector is reported in this figure. The 11 sectors are materials (M), energy (E), real estate (RE), utilities (U), telecommunication services (TS), information technology (IT), financials (F), health care (HC), consumer staples (CS), consumer discretionary (CD), and Industrials (I).

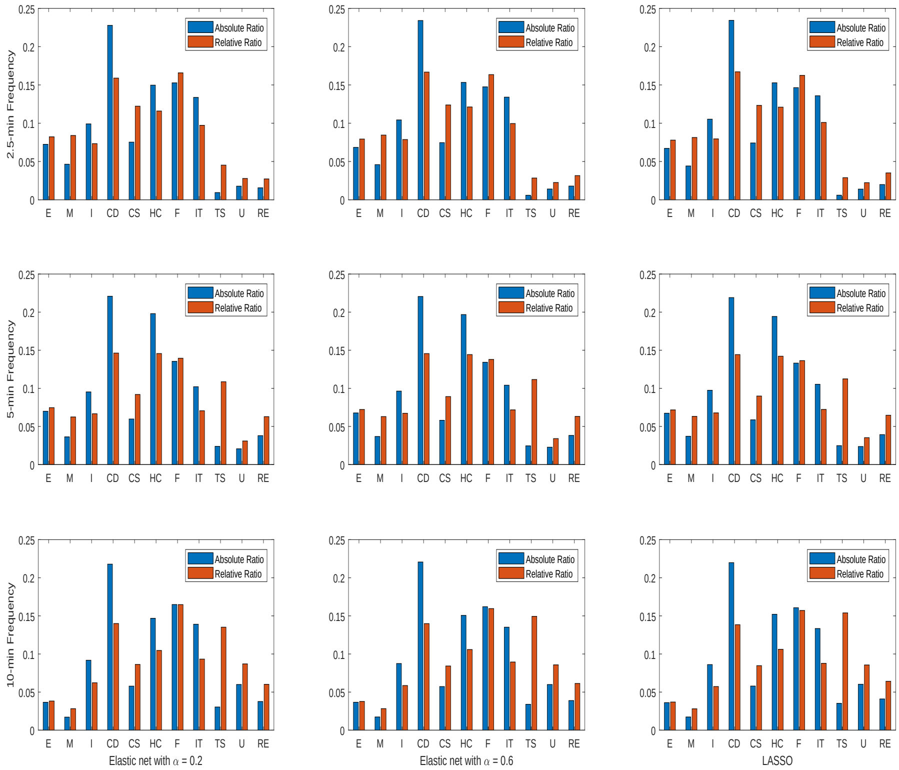

Figure 2: Average Rates of Selection (SPY)

*Notes: Charts in this figure indicate the percentages of stocks in each sector that are selected in the first step of our procedure using either the elastic net (first two columns of charts) or the LASSO (third column of charts), for the target asset given in the title of the figure. More specifically, for each rolling window, we calculate the ratio of the number of selected stocks in each sector to the total number of selected stocks, and take the average over all rolling windows in our out-of-sample prediction period. This is denoted "Absolute Ratio". We also chart the "Relative Ratio", for which the average ratios in "Absolute Ratio" are rescaled by the size of each sector, as given in 1. Finally, the different sectors are denoted along the horizontal axis of each chart. See Sections 3 and 4 for further details. 1.

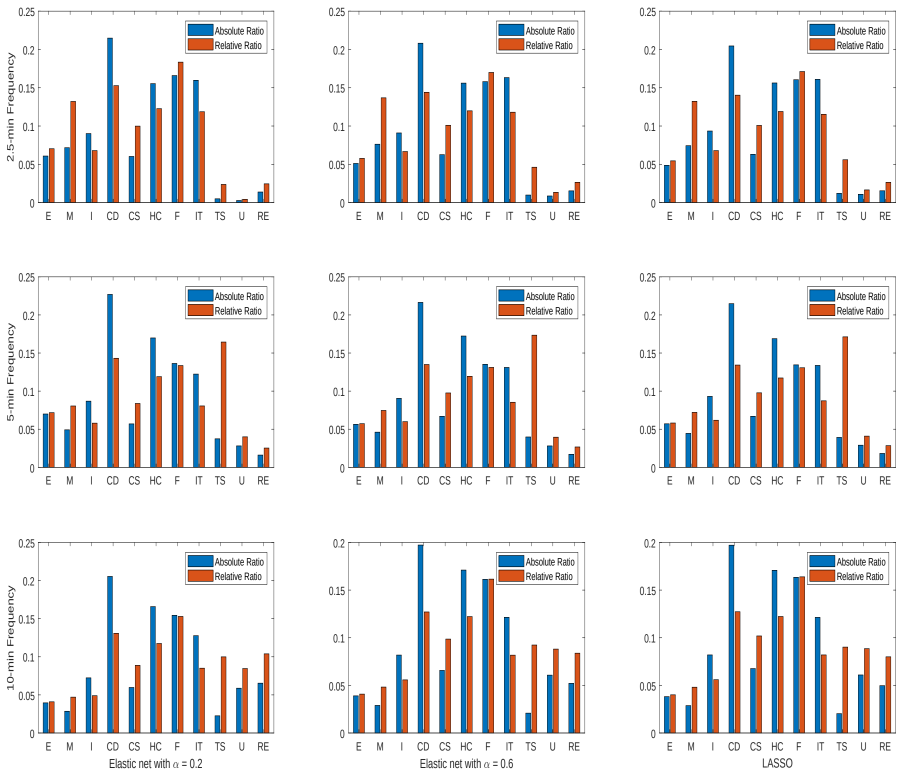

Figure 3: Average Rates of Selection (XLE)

*Notes: See notes to Figure 2.

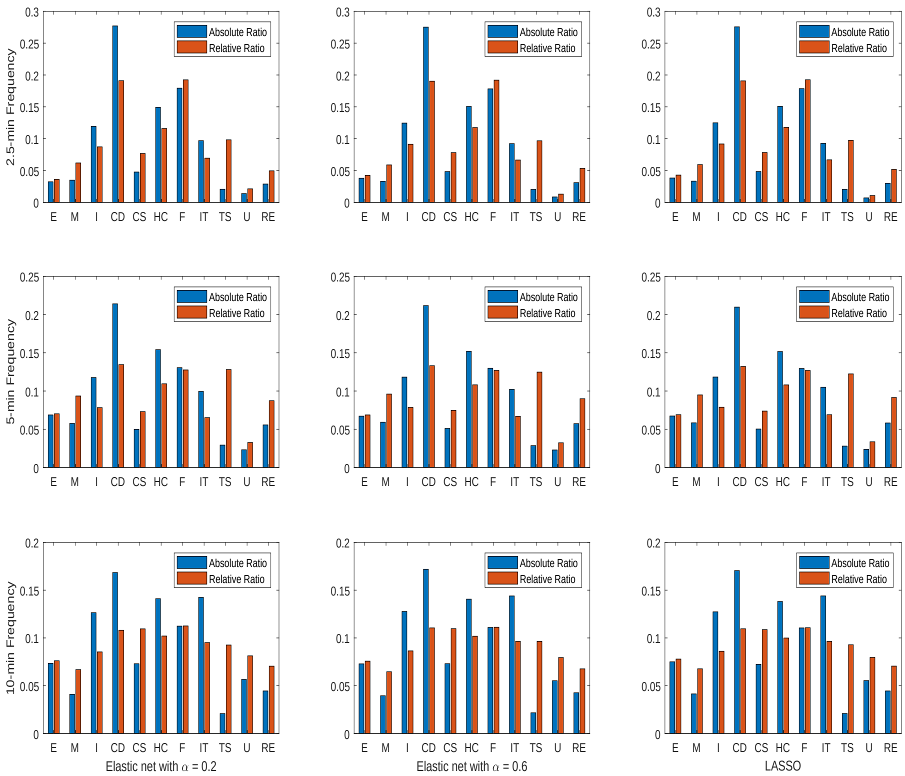

Figure 4: Average Rates of Selection (JNJ)

*Notes: See notes to Figure 2.