A Novel Cluster HAR-Type Model for Forecasting Realized Volatility

Xingzhi Yao*
Xi'an Jiaotong Liverpool University

Marwan Izzeldin †
Lancaster University

Zhenxiong Li ‡
Lancaster University

January 24, 2019

*Corresponding author. Address: International Business School Suzhou (IBSS), Xi'an Jiaotong Liverpool University, 215123, China; Phone: +86 0512 81883243. E-mail: Xingzhi.Yao@xjtlu.edu.cn.
†Address: Department of Economics, Lancaster University, LA1 4YD, UK
‡Address: Department of Economics, Lancaster University, LA1 4YD, UK

A Novel Cluster HAR-Type Model for Forecasting Realized Volatility

January 24, 2019

Abstract

This paper proposes a cluster HAR-type model that adopts the hierarchical clustering technique to form the cascade of heterogeneous volatility components. In contrast to the conventional HAR-type models, the proposed cluster models are based on the relevant lagged volatilities selected by the cluster group Lasso. Our simulation evidence suggests that the cluster group Lasso dominates other alternatives in terms of variable screening and that the cluster HAR serves as the top performer in forecasting the future realized volatility. The forecasting superiority of the cluster models are also demonstrated in an empirical application where the highest forecasting accuracy tends to be achieved by separating the jumps from the continuous sample path volatility process.

Keywords: Heterogeneous autoregressive model; Clustering; Lasso; Realized volatility.

JEL Classification: C58; C63; C49

1 Introduction

Modelling and forecasting volatility is of critical importance for asset and derivative pricing, asset allocation and risk management. With the increasing availability of high-frequency data, a new line of literature aims to exploit intraday information in the estimation and forecast of return volatility. This literature originates with Andersen and Bollerslev (1998) who pioneer the use of high-frequency data to derive realized variance (RV), a non-parametric measure defined as the cumulative sum of squared intraday returns. Compared with volatilities built on daily, weekly and monthly data, e.g. the parametric GARCH or stochastic volatility (SV) models, the RV provides model-free unbiased estimates of the ex post return variation under the conditions specified by Barndorff-Nielsen and Shephard (2002).

Studies of RV have documented compelling evidence of long memory, i.e. historical RVs have a persistent impact on the future RV, which has been traditionally modelled as an ARFIMA process, see Andersen et al. (2001) and Andersen et al. (2003) for examples. However, as a fractional integration model, the ARFIMA is non-trivial to estimate and lacks a clear economic interpretation. A good alternative is the heterogeneous autoregressive, or HAR, model proposed by Corsi (2009). The HAR model is an additive cascade of three volatility components realized over different time horizons, i.e. daily, weekly and monthly, so that the lag structure assumed in the HAR is fixed as (1,5,22). Although not formally a long-memory model, the HAR is able to reproduce the strong persistence of financial volatility by the sum of RV components aggregated at different interval sizes. In addition, the HAR is easy to implement, interpret and forecast. In spite of its wide application, little work has been undertaken in terms of determining the optimal lag structure implied by the HAR model, e.g. the selection of the relevant lags of the RV and the arrangement of volatility components.

A recent attempt to tackle this selection issue is the work of Audrino and Knaus (2016)

who apply the least absolute shrinkage and selection operator (Lasso) proposed by Tibshirani (1996). The motivation to use the Lasso is in that it produces estimated regression coefficients which are exactly zero, and therefore only predictors with nonzero estimates are perceived to be relevant. Audrino and Knaus (2016) show that the HAR-implied lag structure can be recovered asymptotically by the Lasso only where the HAR is the underlying data generating process (DGP). However, in the empirical study using selected stocks for the period 2001-2010, they find that the lag structure of the HAR model is not completely in accord with the one produced by the Lasso, which casts some doubt on the appropriateness of the HAR in volatility forecasts.

In addition, Audrino et al. (2015)[1] adopt the adaptive Lasso estimator introduced by Zou (2006) and examine the significance of the estimated coefficients. In confirming the results of Audrino and Knaus (2016), the lags selected by the adaptive Lasso are inconsistent with those implied by the HAR model. Moreover, the distant lags given by the adaptive Lasso, i.e. lags far beyond the 22nd, are generally statistically insignificant, which, to some extent, explains the excellent empirical performance of the HAR model. Distinct from Audrino and Knaus (2016) who concentrate on lassoing the AR terms, Audrino et al. (2016) consider flexible HAR specifications. They divide the lags in AR(50) into four groups, i.e. {1}, {2-5}, {6-22}, {23-50}, where the first three groups are implied by the lag structure of the HAR model. Applying the Group Lasso to select all the variables within a group if the group is considered active, Audrino et al. (2016) show that the hypothesis for the validity of the lag structure of the standard HAR is rejected in most cases. They argue that the primary reason for rejection might be the inappropriate arrangement of the groups and that a minor reason is the equality restrictions imposed on the AR coefficients. It is also worth noting that, in the out-of-sample (OOS) volatility forecasts, none of the models considered in Audrino and Knaus (2016) or Audrino et al. (2016) bring significant gains over the standard HAR.

[1] This can be seen as an extension of the earlier version of Audrino and Knaus (2016).

Against this background, this paper contributes to the existing literature by proposing a novel cluster HAR-type model. The proposed models utilize the technique of variable screening and clustering and are found to deliver important forecasting gains over the standard and alternative HAR specifications. With the use of the cluster group Lasso introduced by Bühlmann et al. (2013), the cluster models are constructed by the relevant variables only for the forecast of future RV. In the group of highly correlated variables (lagged RVs in our case), the Lasso considered in the existing forecasting literature often omits active variables, i.e. the so-called false negatives, since the Lasso selection is based on the strength of the individual variables as opposed to the strength of the groups of input variables. Unlike the Lasso, the cluster group Lasso splits the lagged RVs first and then selects whole clusters rather than single lagged RVs. In a Monte Carlo simulation study, we provide evidence for the better performance of the cluster group Lasso in regard to the variable screening and selection under different conditions.

In contrast to the fixed time scale, i.e. daily, weekly and monthly, assumed in the standard HAR, the cluster models are based on the volatility factors using the partition produced by the hierarchical clustering, i.e. an algorithm used in the procedure of the cluster group Lasso discussed above. In our simulation study with the mis-specified HAR as the underlying DGP, the hierarchical clustering often over selects the ingredients of the volatility components. This indicates that the effect of a volatility shock at time $t-1$ is identical to shocks at $t-2$, $t-3, \cdots, t-j_1$, where j_1 is often greater than five. To allow for a more flexible lag structure, we follow the work of Bollerslev et al. (2018) by including the first five daily lagged RVs with their own AR estimated coefficients. We refer to this model as the cluster HAR.

To examine the forecasting performance of the cluster HAR, we first conduct a simulation of asset prices based on the two-factor stochastic volatility diffusion with noise. In forecasting the future RV over different horizons, the simulation shows that the cluster HAR dominates various alternatives, e.g. the HAR, adaptive Lasso AR in Audrino and Knaus (2016) and

adaptive Lasso HAR in Audrino et al. (2016). We then consider an empirical application using the high-frequency data of the SPY and ten equities from 2000 to 2013 and implement the forecasting exercises in both pre- and post-crisis periods. To account for the relevance of jumps in volatility forecasts, we introduce the cluster HAR-TCJ model, which extends the HAR-TCJ of Corsi et al. (2010) by applying the cluster group Lasso on the continuous part of the quadratic variation. We find that the cluster HAR-TCJ even improves upon the cluster HAR and serves as the top performer in most cases considered. In addition, results of the Diebold and Mariano test indicate that the forecasting gains of the cluster HAR (cluster HAR-TCJ) over the HAR (HAR-TCJ) tend to be significant over monthly horizons, in which cases the superiority of the cluster models are also evident in the Model Confidence Set procedure introduced by Hansen et al. (2011).

It is worth pointing out that the cluster HAR-type model is motivated by the standard HAR and draws on insights from the model selection approach. The cluster model is straightforward to estimate by standard OLS, which indicates that its forecasting superiority goes along with the ease of implementation. The reasons for the cluster model to be the best performer can be summarized as follows: (a) it only includes the active predictors for the future RV; (b) it is based on a flexible arrangement of volatility factors determined by the hierarchical clustering; (c) it keeps the volatility cascade in the HAR structure to approximate the long memory observed in real data.

The rest of the paper is organized as follows. Section 2 introduces the models proposed in this study together with several Lasso-based estimators applied in the existing literature. Section 3 presents Monte Carlo simulations to illustrate the workings of the cluster models. Section 4 provides the data description and the empirical results of the in-sample estimation and OOS volatility forecasts given by different models. Robustness checks are discussed in section 5. A conclusion is presented in Section 6.

2 Methodology

We begin this section with an introduction of the cluster models proposed in this paper. We then present the various Lasso-based forecasting models considered in the existing literature before closing with a discussion of the forecasts over the multiperiod horizon.

2.1 Cluster HAR-type models

Our cluster models are based on the HAR introduced by Corsi (2009). The HAR model is among the most heavily adopted specifications for modelling and forecasting RV. For simplicity, we consider the RV estimator proposed by Barndorff-Nielsen and Shephard (2002), which is equal to the sum of intraday squared returns

$$RV_t = \sum_{j=1}^{M} r_{t,j}^2 \tag{1}$$

where $r_{t,j}$ stands for intraday returns within each time interval. To introduce the HAR model, we denote the average RV over the previous h days by

$$RV_t^h = \frac{1}{h} \sum_{i=1}^{h} RV_{t-i+1} \tag{2}$$

Thereby, $RV_t^W = \frac{1}{5} \sum_{i=1}^{5} RV_{t-i+1}$ and $RV_t^M = \frac{1}{22} \sum_{i=1}^{22} RV_{t-i+1}$ are respectively the weekly (5-day) and monthly (22-day) averages of daily RV. The standard HAR is given by

$$RV_{t+1} = \beta_0 + \beta_D RV_t + \beta_W RV_t^W + \beta_M RV_t^M + \varepsilon_{t+1} \tag{3}$$

where $\{\varepsilon_t\}$ is a sequence of independent and identically distributed (i.i.d.) innovations with zero mean. Estimates of coefficients, β_0, β_D, β_W and β_M, can be consistently obtained by a standard OLS regression.

As indicated by Corsi (2009), the HAR model in equation (3) is equivalent to an AR(22)

model with imposed equality constraints on the AR coefficients as follows

$$RV_{t+1} = \beta_0 + \sum_{i=1}^{22} \phi_i^{HAR} RV_{t-i+1} + \varepsilon_{t+1} \tag{4}$$

The coefficient restrictions implied by the HAR are given by

$$\phi_i^{HAR} = \begin{cases} \beta_D + \frac{1}{5}\beta_W + \frac{1}{22}\beta_M & \text{for } i = 1 \\ \frac{1}{5}\beta_W + \frac{1}{22}\beta_M & \text{for } i = 2,\ldots, 5 \\ \frac{1}{22}\beta_M & \text{for } i = 6,\ldots, 22 \end{cases} \tag{5}$$

The HAR is built on the assumption that market participants having different trading frequencies result in three types of volatility components, i.e. daily, weekly and monthly. However, it remains unclear whether the HAR-implied lag structure includes all relevant historical volatilities and whether the arrangement of three volatility components is appropriate for the purpose of forecasting. To address these issues, we consider a cluster group Lasso introduced by Bühlmann et al. (2013) to implement variable screening for volatility forecasts and construct the volatility components based on the selected clusters, which we now introduce.

A Cluster Group Lasso

Hierarchical clustering To implement the cluster group Lasso proposed by Bühlmann et al. (2013)[2], efforts must be first made to divide predictors into groups, i.e. homogenous clusters. Clustering methods are often adopted to split variables into groups so that elements in each group are strongly related to each other and contain similar information.

In this paper, we consider the hierarchical clustering algorithm in Chavent et al. (2012), which is based on a principal component method of constructing the synthetic variables of the clusters for dimension reduction[3]. We represent the daily RV_t by x_t with $(x_t,\cdots, x_{t-p+1})'$

[2]We also consider the cluster representative Lasso (CRL) introduced by Bühlmann et al. (2013) as a variable screening method. However, we find that the models based on the CRL are dominated by those using the cluster group Lasso in forecasting future RV. This can be due to the issue of false negatives often encountered in the use of the CRL.

[3]Other clustering methods are also considered in our simulation study, including k-means algorithm,

as predictor variables and let $P_K = (C_1, \cdots, C_K)$ be a partition of the p variables into K clusters. We then define the Homogeneity H of a cluster C_k as follows

$$H(C_k) = \sum_{\mathbf{x}_{t-j} \in C_k} r^2_{\mathbf{x}_{t-j},\, y_k} = \lambda_1^k \tag{6}$$

where r^2 denotes the squared Pearson correlation and y_k is the first component of the principal component analysis (PCA) applied to all the variables in C_k. For a partition P_K, the sum of the homogeneities of its clusters is given by

$$H(P_K) = \sum_{k=1}^{K} H(C_k) = \lambda_1^1 + \cdots + \lambda_1^K \tag{7}$$

where $\lambda_1^1 + \cdots + \lambda_1^K$ are the first eigenvalues of the PCA applied to the K clusters of the partition P_K.

The aim of the clustering algorithm is to find a partition which maximizes the homogeneity criterion in equation (7). Specifically, the procedure begins with the single variables, i.e. partition into p clusters. It then combines two clusters A and B with the smallest dissimilarity d defined below

$$d(A, B) = H(A) + H(B) - H(A \cup B) \tag{8}$$

The procedure is repeated until the single cluster $\{x_t, \cdots, x_{t-p+1}\}$ is obtained. Finally, a bootstrap procedure is conducted to examine the stability of the partitions and to help select a suitable number of clusters, see more details in Chavent et al. (2012).

Group Lasso With the group membership given by hierarchical clustering, we can select the relevant clusters by estimating the coefficients of predictors using the group Lasso. When

supervised clustering by Dettling and Bühlmann (2004), clustering based on canonical correlations by Bühlmann et al. (2013) and density-based clustering. However, these alternatives are found to encounter great difficulties in splitting a set of 22 lagged RVs into groups, where the true number of clusters in the simulation setting is three. The simulation code to evaluate the performances of the various clustering methods is provided in the supplemental files and results are not reported for brevity.

one group is active, all the past RVs within this group should be active. This is the key idea of the group Lasso proposed by Yuan and Lin (2006), who argue that the Lasso can only be used to select individual variables rather than a group of correlated variables.

We retain the notation used in the earlier subsection. As suggested by Yang and Zou (2015), the Group Lasso estimator can be obtained by solving the penalized least squares as follows

$$\widehat{\phi}^{Group} = \arg\min_{\phi} \left\{ \frac{1}{2} \sum_{t=p}^{T} \left(x_{t+1} - \sum_{j=1}^{p} \phi_j x_{t-j+1} \right)^2 + \lambda \sum_{k=1}^{K} \sqrt{p_k} \sqrt{\sum_{j \in I_k} \phi_j^2} \right\} \tag{9}$$

where the p lagged RVs are divided into K non-overlapping groups such that $(1, 2, \ldots, p) = \cup_{k=1}^{K} I_k$, and $I_k \cap I_{k'} = \varnothing$ for $k \neq k'$, and the cardinality of index set I_k is p_k. The selection of the tuning parameter λ causes shrinkage of the solutions towards zero, where some of the coefficients become exactly zero when λ is sufficiently large. To determine λ, we follow the "one-standard-error" rule via cross-validation throughout this paper[4]. We employ a simple unified algorithm-groupwise majorization descent (GMD)-as proposed by Yang and Zou (2015). This is used to solve the group Lasso learning problem if the loss function meets a quadratic majorization condition.

B the cluster HAR model

To implement the cluster HAR model, we apply the hierarchical clustering to the predictor variables $(x_t, \cdots, x_{t-p+1})$ and then estimate

$$x_{t+1} = c + \sum_{j=1}^{p} \phi_j x_{t-j+1} + \varepsilon_t \tag{10}$$

using the group Lasso based on the group structure implied by the obtained clusters, where $t = p, \cdots, T$. After removing the predictors considered irrelevant by the group Lasso, we

[4] We also attempted using alternative way of finding the tuning parameter, λ, e.g. by minimizing the BIC criterion. However, this does not alter the forecasting results reported below.

derive K volatility factors from the lagged RVs with nonzero estimated coefficients and obtain the following model

$$RV_{t+1} = \beta_0 + \sum_{k=1}^{K} \beta_{Cluster\ k} \left(\frac{1}{j_k} \sum_{j=1}^{j_k} RV_{t-j+1} \right) + \varepsilon_t \tag{11}$$

where j_1, j_2,$\cdots$, j_{K-1} denotes the partition point between the two non-overlapping clusters and j_K represents the number of relevant predictors. The implementation of the cluster HAR in equation (11) only utilizes the pre-specified volatility components and does not require a choice of the unknown tuning parameters, which suggests that the model is simple to estimate by standard OLS.

C the cluster HAR-TCJ model

We also consider an extension of the HAR, i.e. the HAR-TCJ model introduced by Corsi et al. (2010). With the aim of better accommodating jumps in the estimation of volatility models, the HAR-TCJ modifies the HAR-CJ model in Andersen et al. (2007) by considering a more powerful test for jump detection. The HAR-TCJ is defined as

$$RV_{t+1} = \beta_0 + \beta_{C,D}\widehat{TC}_t + \beta_{C,W}\widehat{TC}_t^W + \beta_{C,M}\widehat{TC}_t^M + \beta_{J,d}\widehat{TJ}_t + \varepsilon_{t+1} \tag{12}$$

where $\widehat{TC}_t^W$ and $\widehat{TC}_t^M$ are the weekly and monthly averages of $\widehat{TC}_t$ as in equation (2).

The jump component is estimated by the threshold bipower variation (TBPV) measure given by

$$\widehat{TJ}_t = I_{\{C-Tz>\Phi_\alpha\}} \left(RV_t - TBPV_t\right)^+ \tag{13}$$

where $C - Tz$ is the statistics of the jump test based on the corrected version of the TBPV and Φ_α denotes the cumulative distribution function of the normal distribution at level

$\alpha = 99.9\%$, see more details in Corsi et al. (2010). The TBPV is derived as

$$TBPV_t = \mu_1^{-2} \sum_{j=2}^{M} |r_{t,j-1}| \, |r_{t,j}| \, I_{\left\{r_{t,j-1}^2 \leq \upsilon_{j-1}\right\}} I_{\left\{r_{t,j}^2 \leq \upsilon_j\right\}} \tag{14}$$

with $\mu_1 = (2/\pi)^{0.5}$. The threshold is written as $\upsilon_t = c_\upsilon^2 \widehat{V}_t$ with $\widehat{V}_t$ being an estimator of the local variance and $c_\upsilon = 3$ considered in the empirical analysis of Corsi et al. (2010). Finally, the continuous part of variation corresponds to

$$\widehat{TC}_t = RV_t - \widehat{TJ}_t \tag{15}$$

In similar vein to the cluster HAR, we implement the cluster group Lasso on lags of the continuous sample path and employ the aggregation of the selected continuous parts to construct the cluster HAR-TCJ

$$RV_{t+1} = \beta_0 + \sum_{k=1}^{K} \beta_{Cluster\ k} \left(\frac{1}{j_k} \sum_{j=1}^{j_k} \widehat{TC}_{t-j+1} \right) + \beta_{J,d} \widehat{TJ}_t + \varepsilon_t \tag{16}$$

with j_1, j_2,$\cdots$, j_K defined earlier.

2.2 Lasso-based models

For a comparison with the proposed cluster models in regard to the accuracy of volatility forecasts, we introduce the adaptive Lasso AR in Audrino and Knaus (2016) and the adaptive Lasso HAR in Audrino et al. (2016) below.

Considering an $AR(p)$ process in equation (10), we obtain a sparse solution by solving the minimization problem as follows

$$\widehat{\phi}^{Lasso} = \arg\min_{\phi} \left\{ \frac{1}{2} \sum_{t=p}^{T} \left(x_{t+1} - \sum_{j=1}^{p} \phi_j x_{t-j+1} \right)^2 + \lambda \sum_{j=1}^{p} \left| \phi_j \right| \right\} \tag{17}$$

The solution for the constant c is $\widehat{c} = \overline{x}$. We remove c from the minimization by demeaning

the data, i.e. $\overline{x} = 0$. In the original Lasso, all the AR coefficients are penalized equally. Zou (2006) provides a refined version of the Lasso (the adaptive Lasso) allowing for a more flexible penalization, which helps to reduce false positives. The adaptive Lasso estimator is given by

$$\widehat{\phi}^{AL} = \arg\min_{\phi} \left\{ \frac{1}{2} \sum_{t=p}^{T} \left(x_{t+1} - \sum_{j=1}^{p} \phi_j x_{t-j+1} \right)^2 + \lambda \sum_{j=1}^{p} \lambda_j \left| \phi_j \right| \right\} \tag{18}$$

with λ_j as individual weights for each of the coefficients. We follow Zou (2006) by employing the weights as the inverse of the absolute values of the OLS coefficients. We then construct the adaptive Lasso HAR in Audrino et al. (2016) by applying the adaptive Lasso procedure to select the active terms in the equation below

$$RV_{t+1} = \beta_0 + \sum_{i=1}^{p} \beta_i \left(\frac{1}{i} \sum_{j=1}^{i} RV_{t-j+1} \right) + \varepsilon_{t+1} \tag{19}$$

2.3 Longer-horizon Forecasting

Following the work of Andersen et al. (2007), Corsi (2009), Bollerslev et al. (2016) and Bollerslev et al. (2018), among others, we extend the aforementioned models for one-day RV to longer horizons by replacing the daily RV on the left-hand-side of the different models with the aggregated RV over the multiperiod horizon h, i.e. RV_{t+h}^{h}. As a result, the h-step-ahead forecasts of the RV can be obtained as the one-step-ahead forecast for a given h (Chen et al. (2016)).

To generate forecasts over the multiperiod horizon, we split the sample containing T observations into an estimation period and evaluation period as follows

$$t = \underbrace{1, 2, \cdots, m,}_{\text{estimation period}} \underbrace{m+1, m+2, \cdots, T}_{\text{evaluation period}}$$

The forecasts are based on re-estimating the parameters of the different models each day with a rolling window of fixed length m. At time m, we estimate the model parameters using

the first m observations, and then generate the h-step-ahead OOS forecasts and compare them with the realization RV_{t+h}^h. In a similar fashion, the volatility point forecast at time $m+1$ can be obtained using the m observations ending at $m+1$. Iteratively applying this procedure, we finally produce $n = (T-h-m+1)$ number of OOS volatility forecasts.

3 Simulation Study

This section begins with a simulation to assess the performance of the cluster group Lasso in terms of the model selection and to evaluate the appropriateness of the hierarchical clustering in grouping or clustering variables. We then conduct another simulation experiment to demonstrate the forecasting superiority of the cluster HAR over alternative models including the HAR.

3.1 In-sample evaluation

Following the work of Audrino and Knaus (2016), we undertake the Monte Carlo study under the assumption that the true model is the extended HAR as follows

$$RV_{t+1} = \beta_0 + \beta_D RV_t + \beta_W RV_t^W + \beta_M RV_t^M + \gamma\kappa_{t+1} + \varepsilon_{t+1} \tag{20}$$

where $\varepsilon_{t+1} = h_{t+1}^{1/2}\omega_{t+1}$, $h_{t+1} = \alpha_0 + \alpha_1\varepsilon_t^2 + \beta_1 h_t$, with κ_t (jump) being an i.i.d. random variable following Poisson distribution with intensity $\lambda = 0.1$ and ω_t following a standard t distribution with five degrees of freedom. Settings of the parameters, i.e. α_0, α_1, β_1, β_0, β_D, β_W and β_M, and values of the unconditional mean $\widehat{\mu}$ and the unconditional variance $\widehat{\sigma}^2$ are consistent with those in Audrino and Knaus (2016)[5]. We then compute the implied AR coefficients using equation (5). Next, we simulate $x_1, \cdots, x_{22}$ from the normal distribution

[5] The HAR parameters, β_0, β_D, β_W and β_M, are averaged over the estimated HAR coefficients for the ten individual stocks under consideration in section 4. We employ the $\widehat{TC}_t$ of the empirical data when estimating the HAR parameters so that the estimated coefficients are not affected by the presence of jumps in the real data. The sample size is in line with that used in Audrino and Knaus (2016).

$N\left(\widehat{\mu}, \widehat{\sigma}^2\right)$ and obtain $x_{23}, \cdots, x_{2483}$ by equation (4).

The following subsection evaluates the ability of different estimators in selecting the relevant predictors for x_{t+1}, which are lags 1 to 22 suggested by equation (20). In applying the cluster group Lasso, we determine the number of clusters using the results of the bootstrap procedure of Chavent et al. (2012). With $p = 100$ in model (10), the number of clusters is set as five[6]. In addition to the Lasso and cluster group Lasso, we also consider the elastic net proposed for situations where high correlations or nearly linear dependence among a group of variables exist, see details in Zou and Hastie (2005). This procedure is replicated 1000 times and results are summarized in Figure 1 where the cases with ($\gamma = 1$) and without jumps ($\gamma = 0$) are reported separately.

Left panel of Figure 1 shows that, when jumps do not exist, the cluster group Lasso generally recovers the HAR structure in that the selection of lags $\phi_1, \cdots, \phi_{22}$ is abundant and that the false positives are rare and virtually disappear at lags greater than 27. However, the Lasso tends to treat lags greater than 6 as irrelevant and the elastic net results in the selection of monthly coefficients only at a moderate level of certainty. In addition, compared with the cluster group Lasso, the elastic net leads to more false positive selections, e.g. lag 33 is considered active by the elastic net in many cases. In our scenario which contains highly correlated variables, the poor performance of the elastic net in terms of variable selection and screening may be attributed to its failure in accounting for the correlation-structure among the variables[7] (Bühlmann et al. (2013)). Under the presence of jumps, the Lasso is clearly

[6] In the simulation and empirical study, our results in terms of the variable selection and volatility forecasts are found to be insensitive to the choice of the number of clusters.

[7] In the simulation study, we also consider the use of the elastic net in selecting the relevant predictors in forecasting future RV. Similar to the construction of the adaptive Lasso AR and adaptive Lasso HAR discussed earlier, we implement the elastic net AR and elastic net HAR in the OOS forecasts. We find that the elastic net-based models perform on a par with their adaptive Lasso counterparts and both are significantly inferior to the standard HAR in terms of the volatility forecast accuracy. Hence, we do not take the elastic net-based models into account when comparing the forecasting performances of the various models in the following analysis. Results of the elastic net-based models are not reported but can be obtained upon request.

inferior to the other alternatives in selecting the relevant lags, i.e. only the first few lags are included in the active set. The disappointing performance of the Lasso can be induced by its sensitivity to the signal-to-noise ratio, which greatly increases in our setting including jumps[8]. In addition, there is an increase in false positives for both the elastic net and cluster group Lasso. However, the latter does a slightly better job by making the correct selection, i.e. lags 1 to 22, more often.

We then employ this simulation containing jumps to evaluate whether the hierarchical clustering can produce a suitable partition of highly correlated variables. For this purpose, we consider two different scenarios, each with 1000 replications: (a) the true number of clusters is known but the arrangement of groups is unclear; (b) both the group structure and the number of clusters are unknown. To facilitate a more direct comparison, we let $p = 22$ and summarize the results in Figure 2. The left panel corresponds to the situation where the underlying HAR(1, 5, 22) lag structure is known and presents two partition points at lag 2 and 6, respectively.

We start with scenario (a) presented in the middle panel of Figure 2 where the true number of clusters is given. The hierarchical clustering often results in a partition with the first cluster being very large, i.e. the first partition point is around lag 7 and the second is identified at 7 lags beyond the first partition point. Similar performances of the clustering algorithm are observed in scenario (b) demonstrated in the right panel. We only report the cases where the number of clusters produced by the hierarchical clustering is less than 3, which accounts for 80% of the 1000 replications. The clustering algorithm is considered to perform well by correctly identifying the underlying number of clusters in more than half of the replications.

[8] The Lasso is found to be robust against the presence of jumps in the work of Audrino and Knaus (2016). This may be due to the inclusion of the intercept coefficient in the demeaned data prior to estimation. In our simulation, we consider the same specification of the Lasso in cases with and without jumps and do not clean the data to demonstrate the workings of the Lasso in the nearly real data dynamics.

In both scenarios (a) and (b), the clustering algorithm leads to the first two disjoint clusters $\{x_{t-1} \cdots x_{t-j_1};\ x_{t-j_1-1} \cdots\}$ where j_1 is around 7. Based on these partition results, we construct the first volatility factor as $\frac{1}{j_1}\sum_{j=1}^{j_1} RV_{t-j+1}$, suggesting that the role played by a volatility shock which occurred the previous day is identical to a shock which occurred j_1 days ago. The observation that j_1 is often greater than 5 clearly contradicts the underlying DGP and reduces the flexibility of the cluster models. To address this problem, we follow Bollerslev et al. (2018) by including the first five daily lagged RVs with their own estimated AR coefficients in the cluster models in the subsequent analysis.

3.2 Out-of-sample forecasts

To demonstrate the superior performance of the cluster HAR in OOS volatility forecasts, we conduct another simulation study which more closely mimics the dynamics of the realistic high-frequency prices. We follow Huang and Tauchen (2005) and Bollerslev et al. (2016) in employing the two-factor stochastic volatility diffusion with noise (SV2F) to simulate the log price level[9]. Although the SV2F model generates continuous sample paths, it can result in the rugged shape of the price series with the use of the volatility feedback effect and the exponential function (Huang and Tauchen (2005)). Our simulation is based on 78 intraday return observations, corresponding to 5-minute sampling frequency on which the RV is constructed. In the rest of this paper, we make use of the level RV[10] unless otherwise noted. We simulate 2000 observations in the sample and rely on a rolling window of 1000 observations for the OOS forecasts over daily ($h = 1$), weekly ($h = 5$) and monthly ($h = 22$)

[9] See details about the model specification and the parameter settings in Appendix A of Bollerslev et al. (2016).

[10] This is in contrast to the work of Audrino and Knaus (2016) who consider the log RV in comparing the forecasting performances between the HAR and Lasso AR. However, their approach is inconsistent with the standard construction of the HAR model cast in logarithmic form, see Andersen et al. (2007) for example, which applies the log transformation to the volatility components rather than the RV series directly. To make our results more comparable to those of the existing literature using the HAR, we employ the level RV throughout the paper.

horizons.

For the purpose of comparing the OOS forecasting performances, we consider the mean square error (MSE) and then standardize the MSE of each of the models by the MSE of the HAR model in order to highlight the relative gains. To examine the significance of difference in the squared forecasting errors between the HAR and its competitors, we employ the Diebold and Mariano (DM) test with the significance level of 5% corrected for autocorrelations and heteroskedasticity. In addition, we further assess the significance of differences of all the competing models considered using the model confidence set (MCS) by Hansen et al. (2011). This procedure sequentially eliminates models that are found to be inferior in the process of testing the null hypothesis of equal predictive ability (EPA), and the surviving models in the final set of the MCS contain the best-performing model with a given level of confidence, 90% in our case. Under the criterion of MSE, we rely on the range statistics, T_R, proposed by Hansen et al. (2011) in testing the EPA and obtain the p-values based on 5000 block bootstraps.

The main results are summarized in Table 1. It reports the average losses and the percentages of times that the model is found to display significant forecasting gains over the HAR based on the DM test and that the model survives all tests without being removed in the MCS[11]. On average, the cluster HAR outperforms the standard HAR in terms of the OOS volatility forecasts over various horizons. The cluster HAR survives in the MCS in almost all cases whereas the HAR is dropped more often as the forecasting horizon increases. Over monthly horizons, the superiority of the cluster HAR is greater and its gains over the

[11] When implementing the cluster models in forecasting RV, we apply the cluster group Lasso to the observations available for variable selection and clustering. Based on the selected variables and group structure, we construct the cluster models and then obtain the model estimates each day with a fixed length rolling window containing the previous 1000 days. Alternatively, one could apply the cluster group Lasso on the recent 1000 observations every day and update the specifications of the cluster models and re-estimate the model parameters for the h-step forecasts. This procedure will be repeated $n = (T - h - m + 1)$ times. The second approach is found to improve the accuracy of volatility forecasts. However, the price to pay for this is an increase in computation time induced by the implementation of the group Lasso n times in large datasets.

HAR are considered significant by the DM test in more replications conducted. However, the two models based on the adaptive Lasso are found inferior in various situations and are eliminated from the MCS in every single case over weekly and monthly horizons. As a result, the adaptive Lasso-based models are no longer under consideration in our following empirical application. Our results indicate that the idea of variable screening and clustering can improve forecasting accuracy over the HAR and that the gains are non-trivial, especially over long horizons.

4 Empirical Study

4.1 Data

In this study, we rely on 5-minute price data of SPY, which tracks the S&P 500 index closely, Citigroup Inc. (C), Microsoft (MSFT), PG&E Corp. (PCG), Pfizer (PFE), General Electric (GE), The Home Depot (HD), AT&T (T), ExxonMobil (XOM), Duke Energy (DUK) and Wal-Mart (WMT). Our sample covers the period from Jan 03, 2000 to Dec 31, 2013 with a total of 3521 observations. All data are obtained from Tick Data Inc. Table 2 provides summary statistics of the RV (1) and $\widehat{\text{TC}}$ (15) series for each of the stocks considered. The mean of the RV is greater than $\widehat{\text{TC}}$ due to the presence of jumps. All of the series are right skewed and exhibit positive kurtosis, indicative of their non-Gaussian distributions.

4.2 In-sample Estimation

Given the existence of jumps in the RV series considered, we account for the relevance of jumps by employing the HAR-TCJ and the cluster HAR-TCJ models in the subsequent analysis. The cluster HAR-TCJ applies the cluster group Lasso to the lags of the continuous part of the quadratic variation directly. This is motivated by the simulation evidence in section 3.1, which is indicative of the less desirable performance of the cluster group Lasso

in terms of the variable screening under the presence of jumps. Table 3 reports the adjusted R^2 of the HAR, HAR-TCJ and their cluster counterparts over the whole sample period. The cluster HAR-TCJ clearly outperforms the alternative models with regard to the in-sample fit over various horizons at both the stock and SPY level. To help gauge where the high R^2 of the cluster HAR-TCJ is coming from, we take the SPY as an example and compare the parameter estimates of the four models with the standard errors based on a Newey-West correction allowing for serial correlation.

Performing inference after model selection is always a difficult task. Although several approaches for the post-selection inference have been suggested by Lee et al. (2016), Tibshirani et al. (2016) and Tian and Taylor (2018), Liu et al. (2018) argue that these selection-adjusted intervals are often "too long to be useful". For simplicity, we employ the data splitting first introduced by Cox (1975), an approach to post-selection inference that most practitioners would agree is valid due to its transparent justification. By dividing the data into two halves, we select the variables for the cluster models using only the former and conduct inference based on the latter. Specifications of the cluster models for the SPY are discussed below[12].

Applying the cluster group Lasso to 100 lagged RVs, we obtain the active groups of the RV in the cluster HAR as follows: lags $\{1 \cdots 29\}$ for $h = 1$ and 5; lags $\{1 \cdots 29;\ 53 \cdots 68\}$ for $h = 22$, i.e. the optimal number of clusters is one for daily and weekly RV and two for monthly RV. The continuous volatility components in the cluster HAR-TCJ is constructed as: lags $\{1 \cdots 29\}$ for $h = 1$ and 5; lags $\{1 \cdots 29;\ 52 \cdots 67\}$ for $h = 22$. Table 4 presents the parameter estimates. For the one-day-ahead forecasts, both the HAR and HAR-TCJ place more weight on the weekly lag whereas the cluster models assign a greater weight to the daily lag, which is in line with the intuition that the shorter lags are more informative for the daily predictions. For the monthly horizon, longer lags are expected to increase in importance, which can be found in the cluster HAR-TCJ giving the largest weight to the second volatility

[12] We demean the data and thus drop the intercept of the different models.

component containing lags $\{1 \cdots 29;\ 52 \cdots 67\}$. However, this anticipation is not reflected in the HAR and HAR-TCJ where the weekly lag is considered the most important and receives the largest weight. In addition, our results for the positive and insignificant estimates of the jump coefficient are consistent with those in Corsi et al. (2010).

4.3 Out-of-sample Forecasts

This subsection provides a comparison of the HAR, HAR-TCJ and their corresponding cluster models in regard to the volatility forecasting accuracy. To examine the impact of the extreme conditions of the crisis on the forecasting performances of the different models[13], we divide our sample into pre-crisis period [Jan 03, 2000 to Aug. 31, 2007] and post-crisis period [Sep 04, 2007 to Dec 31, 2013][14]. Figure 3 depicts the evolution of the RV over the period 2000 to 2013 and the timing of estimated breakpoints using the sequential test introduced by Bai and Perron (1998). It shows that all the RV series are subject to at least one break and that the common break occurred across all the stocks around the end of 2007, which is associated with the onset of the financial crisis. We construct forecasts with a rolling window containing the previous 1000 days in both sub-samples and thus, for most stocks selected, the forecasts in each sub-sample are subject to one break in their in-sample estimation. In applying the cluster models, we consider 100 lagged volatilities, i.e. $p = 100$.

Forecasting results for the pre-crisis period are reported in Table 5, where the DM test examines the null hypothesis that the standard HAR (HAR-TCJ) and the cluster HAR (cluster HAR-TCJ) have the equal forecast accuracy. Over different horizons, the lowest loss is given by the cluster HAR-TCJ with the single exception of the case of the daily SPY

[13] We also implement the forecasting exercises using the whole sample with a rolling window comprised of the previous 2000 days. Although the gains of the cluster models are still evident in most cases, no models are removed from the MCS, indicating that all the models are of equal predictive ability. This could be due to the turbulent times the stocks experienced during the financial crisis, when all the models encounter difficulties in forecasting the future RV accurately.

[14] The choice for the date of the beginning of the crisis is in line with Audrino and Knaus (2016).

where the standard HAR dominates the others. The HAR and HAR-TCJ are outperformed by their corresponding cluster counterparts over weekly and monthly horizons at both the stock and SPY levels. This suggests that non-trivial forecasting gains over long horizons can be achieved by the construction of heterogenous volatility components using the active predictors. In addition, the gains afforded by the cluster models reach the maximum for $h = 22$, in which case the cluster HAR (cluster HAR-TCJ) is significantly superior to the HAR (HAR-TCJ) in 7/10 (8/10) stocks and the cluster model is the only object that survives all tests in the MCS procedure for 8/10 stocks.

The superiority of the cluster models remains unchanged in the case of the post-crisis period in Table 6. However, relative to the pre-crisis period, their gains over the HAR-type models are slightly dampened due to the presence of many extreme values of the RV triggered by the financial crisis of 2007-2008. The cluster HAR-TCJ continues to serve as the top performer on average for individual stocks over daily and weekly horizons as well as for the weekly and monthly SPY. Similar to the pre-crisis period, the greatest improvements of the cluster HAR (cluster HAR-TCJ) over the standard HAR (HAR-TCJ) are observed over monthly horizons when the MCS is purely constituted by the cluster model(s) in the cases of SPY and 6/10 stocks.

5 Robustness

Finally, we investigate the robustness of our simulation and empirical evidence by considering different numbers of lags in equation (10), i.e. $p = 50$ and $p = 150$. Figure 4 presents the comparison of the Lasso and cluster group Lasso with $p = 50$ in recovering lags for the simulated HAR process in equation (20). Consistent with the finding of Figure 1 with $p = 100$, cluster group Lasso dominates the Lasso in selecting the relevant lags, i.e. $\phi_1, \cdots, \phi_{22}$, whether the jump component is included or not. Although several false positives are

made by the cluster group Lasso, it rarely selects lags beyond 27. The Lasso is clearly inferior in that many active lags are omitted and that the issue of false negatives is even more severe under the presence of jumps. The results corresponding to $p = 150$ reported in Figure 5 remain qualitatively intact.

We then examine the sensitivity of the empirical forecasting results to different values of lag order. Given the earlier evidence for the good performances of the cluster models over long horizons, we concentrate on the monthly forecasting horizons only when comparing the various models with different numbers of lags, i.e. $p = 50$, 100 and 150. To render results comparable across different values of p, we consider the same evaluation window, that is the one implied by the case of $p = 150$. Forecasts are constructed in the same procedure as described in section 4.3. Table 7 reports the MSE of the different models relative to the standard HAR in the pre-crisis period. For different values of p, the cluster HAR-TCJ remains as the best-performing model at both the stock and SPY level. The gains of the cluster HAR (cluster HAR-TCJ) over the standard HAR (HAR-TCJ) are significant in most cases considered and the final set of the MCS only includes the cluster model(s), i.e. either the cluster HAR or the cluster HAR-TCJ, in 8/10 individual stocks. In addition, the cluster models display similar forecasting accuracy in the cases of $p = 50$ and 100 and their gains over the HAR-type models are even more substantial for $p = 150$.

Moving to the results for the post-crisis period in Table 8, the cluster models still dominate their competitors significantly in most cases with $p = 100$ and 150 but their gains are not observed for $p = 50$. This could be explained by an increase in volatility persistence during the crisis (Gagnon et al. (2016)), in which case the higher order lags still exhibit non-trivial predictive power for future RV. However, these distant lags relevant for the forecast of the post-crisis RV are not included in the case of $p = 50$. We therefore conclude that our results in support of the superiority of the cluster group Lasso in terms of the model selection and the cluster models in pre-crisis volatility forecasts remain robust to alternative numbers of lags

p. During the post-crisis period, distant lags appear more crucial in accurately forecasting the future RV and thus should not be omitted in the construction of the cluster models.

6 Conclusion

Based on the HAR framework, this paper introduces a cluster HAR model using the relevant lags of volatility selected by the cluster group Lasso. Construction of the volatility factors in the cluster HAR, i.e. cascade of realized variance (RV) aggregated over different horizons, is determined by a hierarchical clustering algorithm. To account for the relevance of jumps in volatility forecasts, we apply the same idea to the HAR-TCJ by Corsi et al. (2010) and obtain the cluster HAR-TCJ. In the simulation study, the cluster group Lasso is found to dominate the Lasso in terms of the model selection and the clustering algorithm often results in a partition with the first volatility component as being a very large cluster. To add to the flexibility of the proposed cluster models, we follow Bollerslev et al. (2018) by introducing the first five lagged RVs with their own estimated AR coefficients.

The forecasting superiority of the cluster HAR is first demonstrated using a Monte Carlo simulation where the standard HAR and the Lasso-based alternatives are considered. Furthermore, we conduct an empirical application using the daily RV data for the SPY and ten individual stocks from 2000 to 2013. Forecasting performances of the HAR, HAR-TCJ and their cluster counterparts are then evaluated in the pre- and post-crisis subsamples to highlight the impact of the extreme observations during the financial crisis. In line with the simulation evidence, the cluster models dominate the HAR-type models in various situations and the gains are best over long horizons. In particular, the cluster HAR-TCJ tends to work as the top performer by applying the technique of variable screening and clustering on the continuous part of the quadratic return variation.

References

Andersen, T., Bollersley, T., Diebold, F. X., and Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics*, 61(1):43–76.

Andersen, T. G. and Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39:885–905.

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. *The Review of Economics and Statistics*, 89(4):701–720.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625.

Audrino, F., Camponovo, L., and Roth, C. (2015). Testing the lag structure of assets' realized volatility dynamics. *Working Paper*.

Audrino, F., Huang, C., and Okhrin, O. (2016). Flexible HAR Model for Realized Volatility. *Working Paper*, pages 1–25.

Audrino, F. and Knaus, S. D. (2016). Lassoing the HAR model: A Model Selection Perspective on Realized Volatility Dynamics. *Econometric Reviews*, 35(8-10):1485–1521.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, pages 47–78.

Barndorff-Nielsen, O. E. and Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 64:253–280.

Bollerslev, T., Hood, B., Huss, J., and Pedersen, L. H. (2018). Risk Everywhere: Modeling and Managing Volatility. *The Review of Financial Studies*, 31(7):2729–2773.

Bollerslev, T., Patton, A. J., and Quaedvlieg, R. (2016). Exploiting the Errors: A Simple Approach for Improved Volatility Forecasting. *Journal of Econometrics*, 192(1):1–18.

Bühlmann, P., Rütimann, P., van de Geer, S., and Zhang, C. H. (2013). Correlated variables in regression: Clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143(11):1835–1858.

Chavent, M., Kuentz, V., Liquet, B., and Saracco, J. (2012). ClustOfVar: An R Package for the Clustering of Variables. *Journal of Statistical Software*, 50(13):1–16.

Chen, X. B., Gao, J., Li, D., and Silvapulle, P. (2016). Nonparametric Estimation and Forecasting for Time-Varying Coefficient Realized Volatility Models. *Journal of Business & Economic Statistics*, 0015(August):1–39.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):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(2):276–288.

Cox, D. R. (1975). A Note on Data-Splitting for the Evaluation of Significance Levels. *Biometrika*, 62(2):441–444.

Dettling, M. and Bühlmann, P. (2004). Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis*, 90(1 SPEC. ISS.):106–131.

Gagnon, M. H., Power, G. J., and Toupin, D. (2016). International stock market cointegration under the risk-neutral measure. *International Review of Financial Analysis*, 47:243–255.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The Model Confidence Set. *Econometrica*, 79(2):453–497.

Huang, X. and Tauchen, G. (2005). The relative contribution of jumps to total price variance. *Journal of Financial Econometrics*, 3(4):456–499.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44(3):907–927.

Liu, K., Markovic, J., and Tibshirani, R. (2018). More powerful post-selection inference, with application to the Lasso. *Working Paper*, pages 1–34.

Tian, X. and Taylor, J. (2018). Selective inference with a randomized response. *The Annals of Statistics*, 46(2):679–710.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact Post-Selection Inference for Sequential Regression Procedures. *Journal of the American Statistical Association*, 111(514):600–620.

Yang, Y. and Zou, H. (2015). A Fast Unified Algorithm for Solving Group-Lasso Penalized Learning Problems. *Statistics and Computing*, 25(6):1129–1141.

Yuan, M. and Lin, Y. (2006). Model Selection and Estimation in Regression With Grouped Variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

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(5):301–320.

Table 1

Simulation Results. The table reports the MSE for the different models relative to the MSE of the standard HAR model. DM test represents the percentage of times that the relative difference in squared forecasting errors between the model in question and the HAR is significant at 5% using the Diebold Mariano test (Newey-West heteroscedasticity consistent covariance matrix estimator). MCS demonstrates the percentage of times that a model is included in the superior models in the procedure of Model Confidence Set. Our simulations are based on the two-factor stochastic volatility diffusion with noise. Models are estimated using 2000 simulated daily observations and the forecasts are based on re-estimating the parameters of the different regressions each day with a fixed length Rolling Window (RW) made up of the previous 1000 days.

	horizon=1			**horizon=5**			**horizon=22**		
	MSE	DM test (%)	MCS (%)	MSE	DM test (%)	MCS (%)	MSE	DM test (%)	MCS (%)
HAR	1		80	1		55	1		51
adaptive Lasso AR	1.338	97	3	2.387	100	0	4.832	100	0
adaptive Lasso HAR	1.339	84	17	2.21	100	0	3.885	100	0
cluster HAR	0.987	15	100	0.929	29	99	0.787	34	100

Table 2

Summary Statistics of RV and $\widehat{TC}$. This table reports the descriptive statistics of the realized variance as well as the continuous part of variation for the SPY and 10 individual stocks, namely: Citigroup Inc. (C), Microsoft (MSFT), PG&E Corp (PCG), Pfizer (PFE), General Electric (GE), The Home Depot (HD), AT&T (T), ExxonMobil (XOM), Duke Energy (DUK), Wal-Mart (WMT). The time span is from Jan 03, 2000 to Dec 31, 2013 with a total of 3521 daily observations.

	C	MSFT	PCG	PFE	GE	HD	T	XOM	DUK	WMT	SPY
						RV_t					
Mean	7.839	2.960	5.333	2.569	3.453	3.549	3.051	2.158	2.733	2.297	1.168
SD	29.899	4.103	30.556	3.455	7.524	5.160	5.152	4.292	6.640	3.459	2.429
Kurtosis	403.191	42.678	1838.201	69.509	151.026	68.380	171.565	374.960	317.485	72.128	152.663
Skewness	16.416	4.932	39.058	6.140	9.697	6.002	8.846	14.894	14.721	5.954	9.691
Median	2.577	1.573	1.966	1.553	1.598	1.884	1.449	1.238	1.331	1.104	0.569
25%-quantile	1.084	0.903	1.023	0.912	0.762	1.068	0.796	0.733	0.724	0.636	0.274
75%-quantile	5.567	3.300	4.499	2.880	3.452	4.043	3.376	2.228	2.715	2.525	1.158
						$\widehat{TC}_t$					
Mean	6.803	2.720	3.323	2.207	3.107	3.146	2.719	2.007	2.284	2.030	1.091
SD	25.569	3.900	13.661	2.890	6.792	4.765	4.854	4.186	5.040	3.187	2.252
Kurtosis	638.161	44.192	1606.308	80.947	191.522	90.097	211.410	413.384	200.797	93.869	187.627
Skewness	20.121	4.987	35.057	6.313	10.648	6.859	9.832	15.774	11.422	6.724	10.506
Median	2.273	1.386	1.145	1.347	1.423	1.643	1.259	1.129	1.067	0.975	0.523
25%-quantile	0.949	0.791	0.617	0.784	0.678	0.911	0.666	0.652	0.561	0.552	0.251
75%-quantile	5.179	3.009	2.825	2.526	3.129	3.629	2.820	2.067	2.281	2.217	1.115

Table 3

In-Sample Fit. The table reports the measure of fit (adjusted R^2) for the various models. Average represents the in-sample losses averaged across the 10 individual stocks. The highest ratio in each column of the Average and the SPY is displayed in bold.

	C	MSFT	PCG	PFE	GE	HD	T	XOM	DUK	WMT	Average	SPY
horizon=1												
HAR	0.437	0.605	0.069	0.451	0.533	0.583	0.568	0.529	0.350	0.582	0.471	0.525
HAR-TCJ	0.460	0.611	0.089	0.471	0.534	0.589	0.575	0.532	0.386	0.585	0.483	0.536
cluster HAR	0.444	0.617	0.072	0.456	0.539	0.600	0.574	0.563	0.363	0.612	0.484	0.546
cluster HAR-TCJ	0.471	0.626	0.094	0.474	0.538	0.611	0.578	0.570	0.408	0.603	**0.497**	**0.565**
horizon=5	C	MSFT	PCG	PFE	GE	HD	S	XOM	DUK	WMT		SPY
HAR	0.446	0.701	0.137	0.590	0.582	0.691	0.686	0.619	0.523	0.679	0.565	0.639
HAR-TCJ	0.465	0.704	0.205	0.600	0.587	0.695	0.689	0.621	0.534	0.679	0.578	0.645
cluster HAR	0.471	0.707	0.166	0.597	0.602	0.696	0.702	0.629	0.533	0.689	0.579	0.646
cluster HAR-TCJ	0.501	0.708	0.243	0.608	0.603	0.700	0.699	0.632	0.553	0.686	**0.593**	**0.655**
horizon=22	C	MSFT	PCG	PFE	GE	HD	T	XOM	DUK	WMT		SPY
HAR	0.518	0.596	0.181	0.577	0.513	0.621	0.646	0.456	0.400	0.602	0.511	0.561
HAR-TCJ	0.544	0.587	0.263	0.578	0.507	0.620	0.641	0.454	0.405	0.600	0.520	0.561
cluster HAR	0.564	0.609	0.332	0.590	0.547	0.628	0.653	0.462	0.439	0.616	0.544	0.559
cluster HAR-TCJ	0.608	0.600	0.444	0.591	0.544	0.626	0.643	0.461	0.463	0.613	**0.559**	**0.563**

Table 4

In-Sample Model Estimates. The table reports the in-sample parameter estimates for the SPY for the standard HAR (HAR-TCJ) and the cluster HAR (cluster HAR-TCJ) for forecasting the daily (h=1), weekly (h=5) and monthly (h=22) RVs. β_1, β_2, β_3, β_4 and β_5 represent the coefficients of the first five daily lagged RVs and $\beta_{cluster1}$ and $\beta_{cluster2}$ stand for the coefficients corresponding to the first and second volatility components in the cluster models. Standard errors of the estimates are reported in parentheses and are adjusted using the Newey-West heteroscedasticity consistent covariance matrix estimator with 5, 10 and 44 lags for the daily, weekly and monthly regression estimates, respectively. *, ** and *** represent the significance of the coefficients at 10%, 5% and 1% level.

HAR				cluster HAR			HAR-TCJ				cluster HAR-TCJ		
	h=1	h=5	h=22		h=1	h=5	h=22		h=1	h=5	h=22		h=1	h=5	h=22
β_D	0.232**	0.194***	0.109***	β_1	0.331***	0.263***	0.184***	$\beta_{C,D}$	0.297*	0.255***	0.130***	$\beta_{C,1}$	0.414***	0.323***	0.195***
	(0.105)	(0.057)	(0.024)		(0.056)	(0.044)	(0.036)		(0.164)	(0.093)	(0.045)		(0.077)	(0.056)	(0.048)
β_W	0.462***	0.358***	0.339***	β_2	0.234***	0.148***	0.115***	$\beta_{C,W}$	0.487**	0.365***	0.292***	$\beta_{C,2}$	0.271**	0.169***	0.108***
	(0.152)	(0.115)	(0.110)		(0.086)	(0.042)	(0.030)		(0.208)	(0.140)	(0.095)		(0.110)	(0.049)	(0.028)
β_M	0.213**	0.303*	0.266***	β_3	-0.078	0.041**	0.059***	$\beta_{C,M}$	0.163*	0.265*	0.243**	$\beta_{C,3}$	-0.154***	0.028	0.039***
	(0.100)	(0.157)	(0.098)		(0.050)	(0.017)	(0.016)		(0.097)	(0.150)	(0.096)		(0.056)	(0.028)	(0.014)
				β_4	0.176**	0.075***	0.091***	β_J	0.059	0.213	1.499	$\beta_{C,4}$	0.212**	0.091***	0.088***
					(0.078)	(0.029)	(0.026)		(0.049)	(0.350)	(1.672)		(0.084)	(0.031)	(0.020)
				β_5	0.063	0.061	0.067***					$\beta_{C,5}$	0.069	0.029	0.059***
					(0.070)	(0.044)	(0.020)						(0.096)	(0.048)	(0.018)
				$\beta_{Cluster1}$	0.191**	0.278*	0.057*					$\beta_{Cluster1}$	0.149**	0.257*	-0.011
					(0.079)	(0.152)	(0.033)						(0.076)	(0.147)	(0.179)
				$\beta_{Cluster2}$			0.175***					$\beta_{Cluster2}$			0.227**
							(0.033)								(0.114)
												β_J	0.034	0.212	1.580
													(0.050)	(0.367)	(1.732)

Table 5

Out-Of-Sample Forecast Losses: pre-crisis period with $P = 100$. The table reports the MSE for the different models relative to the MSE of the standard HAR model using a rolling window of 1000 observations. Average represents the average MSE across all of the 10 individual stocks. The lowest loss ratio in each column of the Average and the SPY is displayed in bold. *, ** and *** indicate that the relative differences, squared forecasting errors, between the cluster HAR (cluster HAR-TCJ) and the standard HAR (HAR-TCJ) model are significant at 10%, 5% and 1% level using the Diebold-Mariano test (Newey-West heteroscedasticity consistent covariance matrix estimator). The numbers with underlines denote that the corresponding models are included in the MCS for the confidence of 90%. The time span is from Jan 03, 2000 to Aug 31, 2007 with a total of 1927 daily observations.

	C	MSFT	PCG	PFE	GE	HD	T	XOM	DUK	WMT	Average	SPY
horizon=1												
HAR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**1.000**
HAR-TCJ	1.019	1.050	0.658	0.915	1.113	1.030	0.986	1.019	0.942	0.963	0.969	1.005
cluster HAR	1.091	0.965**	1.002	1.008	0.944***	1.009	0.998	1.012	1.024**	0.985	1.004	1.042
cluster HAR-TCJ	1.141	0.999***	0.654	0.906	1.004***	1.021	1.008	1.036	0.943	0.945*	**0.966**	1.021
horizon=5												
HAR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
HAR-TCJ	1.031	1.022	0.377	0.821	1.102	1.005	0.984	1.029	0.880	1.008	0.926	0.970
cluster HAR	0.970	0.771***	0.533***	1.047	0.748***	0.900*	0.978	0.966	0.795***	0.873***	0.858	0.944**
cluster HAR-TCJ	1.040	0.793***	0.212***	0.796	0.849***	0.910**	1.010	1.015	0.772**	0.877***	**0.828**	**0.921*** **
horizon=22												
HAR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
HAR-TCJ	1.025	0.914	0.500	0.810	0.749	0.985	1.023	1.014	0.876	0.962	0.886	0.863
cluster HAR	0.697***	0.702***	0.425***	0.980	0.564***	0.707***	1.021	0.997	0.579***	0.637***	0.731	0.814***
cluster HAR-TCJ	0.711***	0.636***	0.144**	0.686**	0.521**	0.751***	1.000	1.009	0.497***	0.704***	**0.658**	**0.798**

Table 6

Out-Of-Sample Forecast Losses: post-crisis period with $P = 100$. The table reports the MSE for the different models relative to the MSE of the standard HAR model using a rolling window of 1000 observations. Average represents the average MSE across all of the 10 individual stocks. The lowest loss ratio in each column of the Average and the SPY is displayed in bold. *, ** and *** indicate that the relative differences, squared forecasting errors, between the cluster HAR (cluster HAR-TCJ) and the standard HAR (HAR-TCJ) model are significant at 10%, 5% and 1% level using the Diebold-Mariano test (Newey-West heteroscedasticity consistent covariance matrix estimator). The numbers with underlines denote that the corresponding models are included in the MCS for the confidence of 90%. The time span is from Sep 04, 2007 to Dec 31, 2013 with a total of 1594 daily observations.

	C	MSFT	PCG	PFE	GE	HD	T	XOM	DUK	WMT	Average	SPY
horizon=1												
HAR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
HAR-TCJ	0.888	0.992	1.019	1.026	1.204	0.989	0.864	1.038	0.988	0.984	0.999	**0.995**
cluster HAR	0.820***	1.056***	1.010	1.005	0.998	1.042***	1.054	0.962**	1.003	1.031*	1.019	1.001
cluster HAR-TCJ	0.791***	1.028*	1.065***	1.039	0.994***	0.996	0.867	1.014	1.020**	0.994	**0.981**	1.010
horizon=5												
HAR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
HAR-TCJ	0.702	0.971	0.960	1.034	0.948	1.041	0.788	0.982	1.051	1.014	0.949	0.971
cluster HAR	0.578***	0.997	0.973	0.977	0.676***	0.983	1.041	0.922***	1.007	1.025	0.918	0.981***
cluster HAR-TCJ	0.391***	0.982	0.992	1.069**	0.676***	1.025	0.797	0.952***	1.056	1.044*	**0.898**	**0.917*** **
horizon=22												
HAR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
HAR-TCJ	0.574	0.991	0.964	1.053	0.749	1.169	0.827	0.963	1.823	0.922	1.004	0.890
cluster HAR	0.652***	0.870***	1.021	0.870	0.644***	0.874**	0.834***	0.961	0.778***	0.889*	**0.839**	0.979***
cluster HAR-TCJ	0.294***	0.884	0.913	0.957	0.557*	1.239	0.842	0.933**	1.458**	0.809*	0.889	**0.749*** **

Table 7

Sensitivity of the Forecast Losses to the Choice of p: pre-crisis period. The table reports the MSE for the different models relative to the MSE of the standard HAR model using a rolling window of 1000 observations. Average represents the average MSE across all of the 10 individual stocks. The lowest loss ratio in each column of the Average and the SPY is displayed in bold. *, ** and *** indicate that the relative differences, squared forecasting errors, between the cluster HAR (cluster HAR-TCJ) and the standard HAR (HAR-TCJ) model are significant at 10%, 5% and 1% level using the Diebold-Mariano test (Newey-West heteroscedasticity consistent covariance matrix estimator). The numbers with underlines denote that the corresponding models are included in the MCS for the confidence of 90%.

	C	MSFT	PCG	PFE	GE	HD	T	XOM	DUK	WMT	Average	SPY
						$p = 50$						
HAR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
HAR-TCJ	1.025	0.902	0.472	0.812	0.726	0.984	1.045	1.016	0.872	0.954	0.881	0.863
cluster HAR	0.680***	0.640***	0.518***	1.009	0.563***	0.708***	1.021	1.004	0.585***	0.607***	0.733	0.828**
cluster HAR-TCJ	0.712***	0.555***	0.173**	0.697**	0.537	0.766***	1.033	1.026	0.492***	0.685***	**0.667**	**0.770**
						$p = 100$						
HAR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
HAR-TCJ	1.025	0.902	0.472	0.812	0.726	0.984	1.045	1.016	0.872	0.954	0.881	0.863
cluster HAR	0.704***	0.643***	0.445***	1.031	0.568***	0.728***	0.956	1.004	0.620***	0.634***	0.733	0.823**
cluster HAR-TCJ	0.708***	0.637***	0.165**	0.697**	0.531	0.761***	1.033	1.026	0.483***	0.696***	**0.674**	**0.782**
						$p = 150$						
HAR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
HAR-TCJ	1.025	0.902	0.472	0.812	0.726	0.984	1.045	1.016	0.872	0.954	0.881	0.863
cluster HAR	0.642***	0.540***	0.434***	1.025	0.579***	0.667***	0.956	1.007	0.608***	0.581***	0.704	0.788***
cluster HAR-TCJ	0.689***	0.499***	0.165**	0.685***	0.529	0.688***	1.033	1.047	0.483***	0.618***	**0.644**	**0.716***

Table 8

Sensitivity of the Forecast Losses to the Choice of p: post-crisis period. The table reports the MSE for the different models relative to the MSE of the standard HAR model using a rolling window of 1000 observations. Average represents the average MSE across all of the 10 individual stocks. The lowest loss ratio in each column of the Average and the SPY is displayed in bold. *, ** and *** indicate that the relative differences, squared forecasting errors, between the cluster HAR (cluster HAR-TCJ) and the standard HAR (HAR-TCJ) model are significant at 10%, 5% and 1% level using the Diebold-Mariano test (Newey-West heteroscedasticity consistent covariance matrix estimator). The numbers with underlines denote that the corresponding models are included in the MCS for the confidence of 90%.

	C	MSFT	PCG	PFE	GE	HD	T	XOM	DUK	WMT	Average	SPY
							$p = 50$					
HAR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
HAR-TCJ	0.528	0.821	0.979	1.041	0.686	1.315	0.805	0.954	1.600	0.952	**0.968**	**0.915**
cluster HAR	0.760***	1.105***	1.023	1.085***	0.895***	1.017***	1.034	1.125***	1.021	1.118***	1.018	1.144***
cluster HAR-TCJ	0.255***	0.923***	0.987*	1.111***	0.778***	1.303	0.804	1.074***	1.516	1.018***	0.977	1.038***
							$p = 100$					
HAR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
HAR-TCJ	0.528	0.821	0.979	1.041	0.686	1.315	0.805	0.954	1.600	0.952	0.968	0.915
cluster HAR	0.703***	0.880**	1.023	0.873	0.609***	0.864**	0.820***	1.125***	1.003	0.864**	0.876	0.922
cluster HAR-TCJ	0.260***	0.719***	0.987*	0.852	0.511	1.374	0.804	0.922**	1.206**	0.808*	**0.844**	**0.735*****
							$p = 150$					
HAR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
HAR-TCJ	0.528	0.821	0.979	1.041	0.686	1.315	0.805	0.954	1.600	0.952	0.968	0.915
cluster HAR	0.703***	0.880**	1.023	0.852	0.558***	0.935	0.820***	1.125***	0.795**	0.883*	0.857	0.916
cluster HAR-TCJ	0.260***	0.725**	0.987*	0.867	0.511	1.374	0.804	0.922**	1.206**	0.800*	**0.846**	**0.735*****

Figure 1

HAR Coefficients Selected with $P = 100$: the comparison of Lasso, Elastic Net and Cluster Group Lasso. This figure reports the times per 1000 replications a lag has been selected by the model selection device considered. Data is simulated by a mis-specified HAR model with stochastic volatility and stochastic volatility together with jumps.

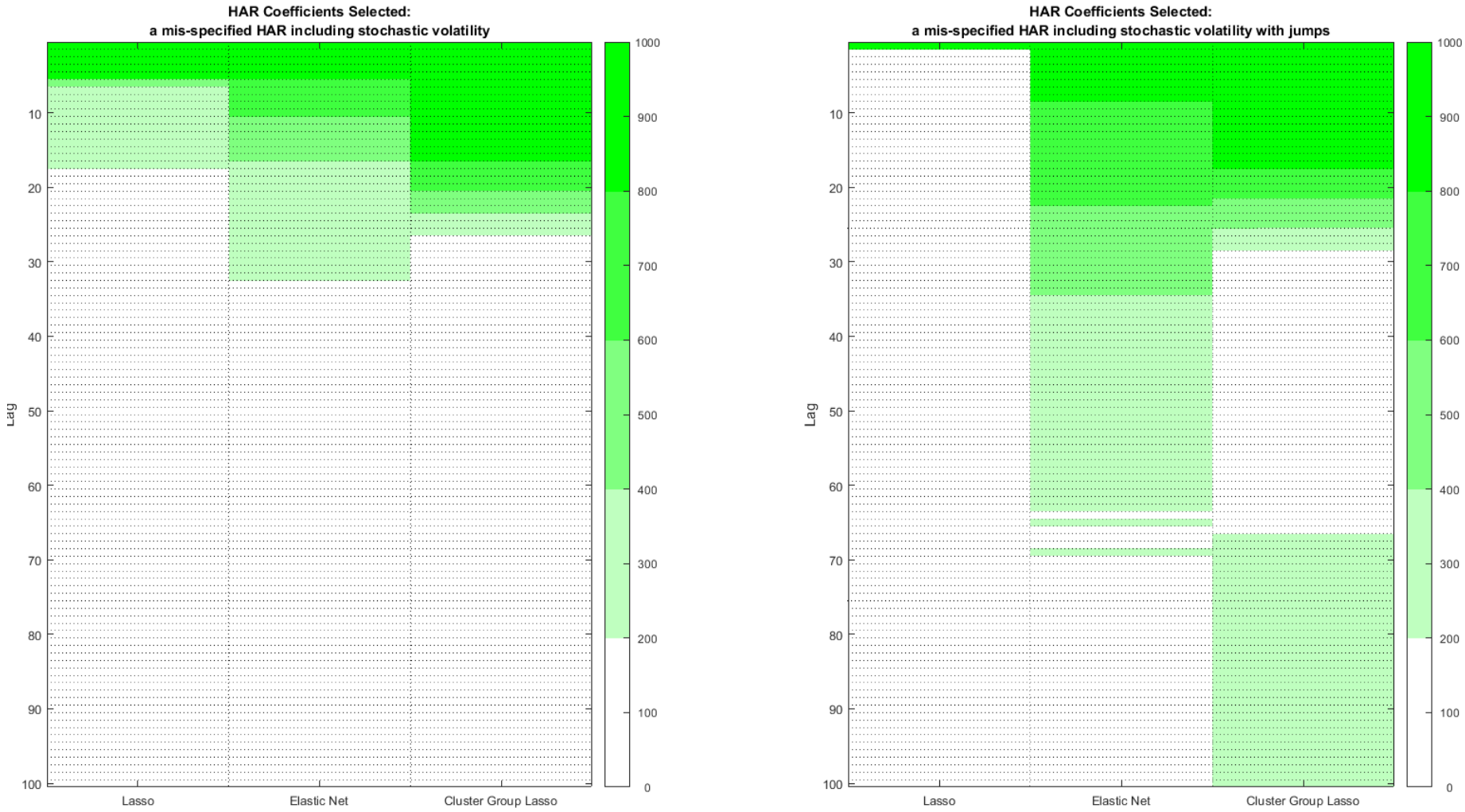

Figure 2

HAR Group Arrangements. This figure reports the times per 1000 replications a lag has been selected as a partition point. Data is simulated by a mis-specified HAR model with stochastic volatility containing jumps.

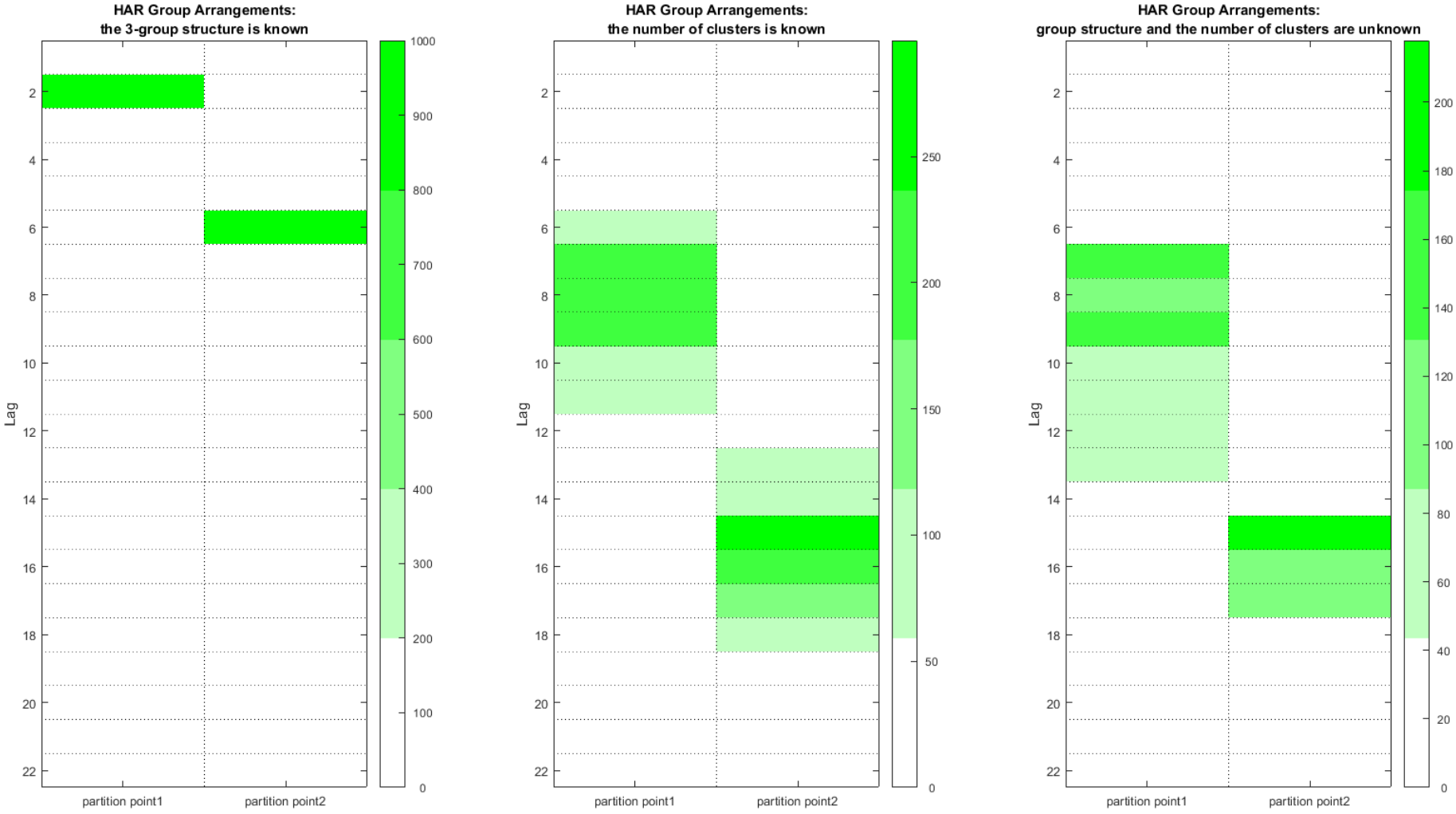

Figure 3

Breaks in the series of SPY and 10 Individual Stocks from 2000 till 2013. This figure plots the breaks given by the sequential estimation of multiple breaks in mean by Bai and Perron (1998).

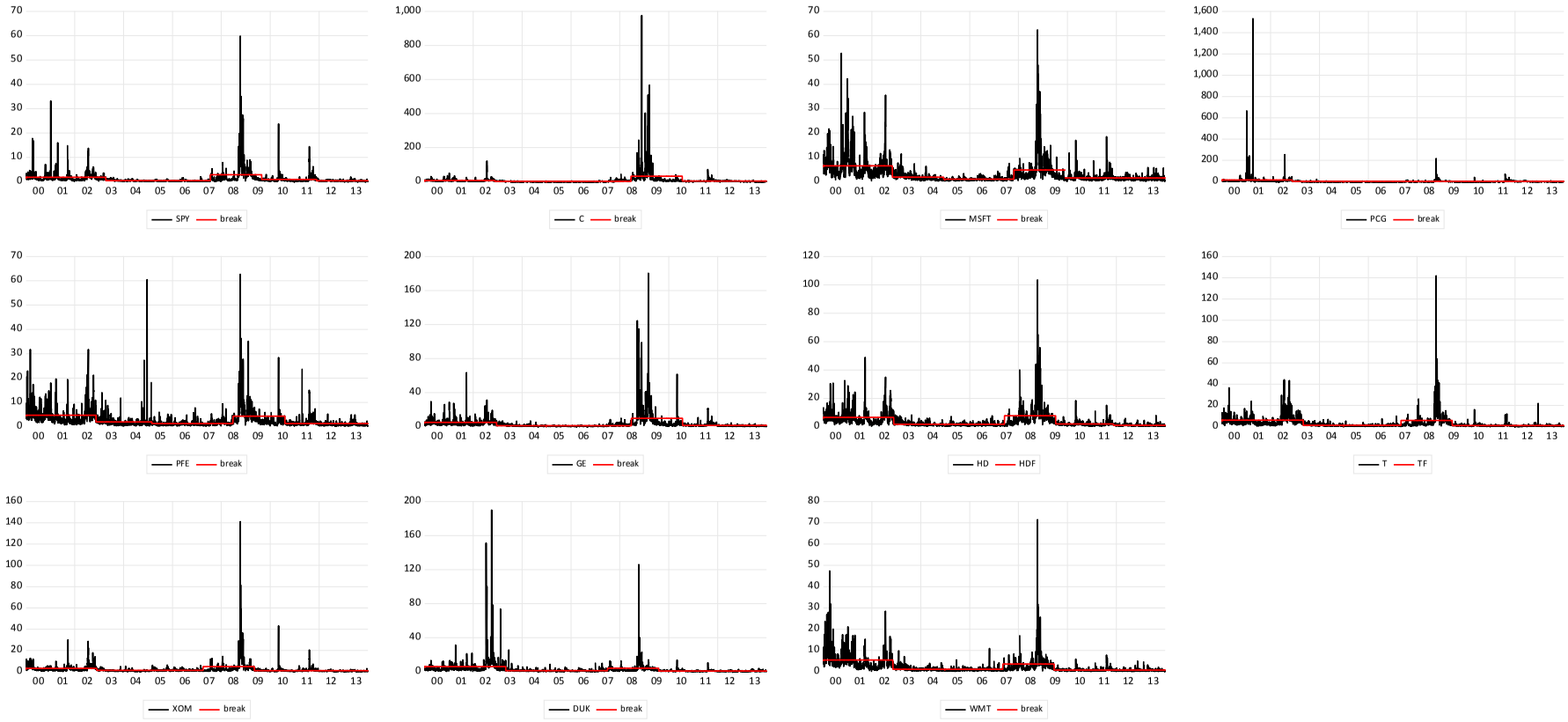

Figure 4

HAR Coefficients Selected with $P = 50$: the comparison of Lasso and Cluster Group Lasso. This figure reports the times per 1000 replications a lag has been selected by the model selection device considered. Data is simulated by a mis-specified HAR model with stochastic volatility and stochastic volatility together with jumps.

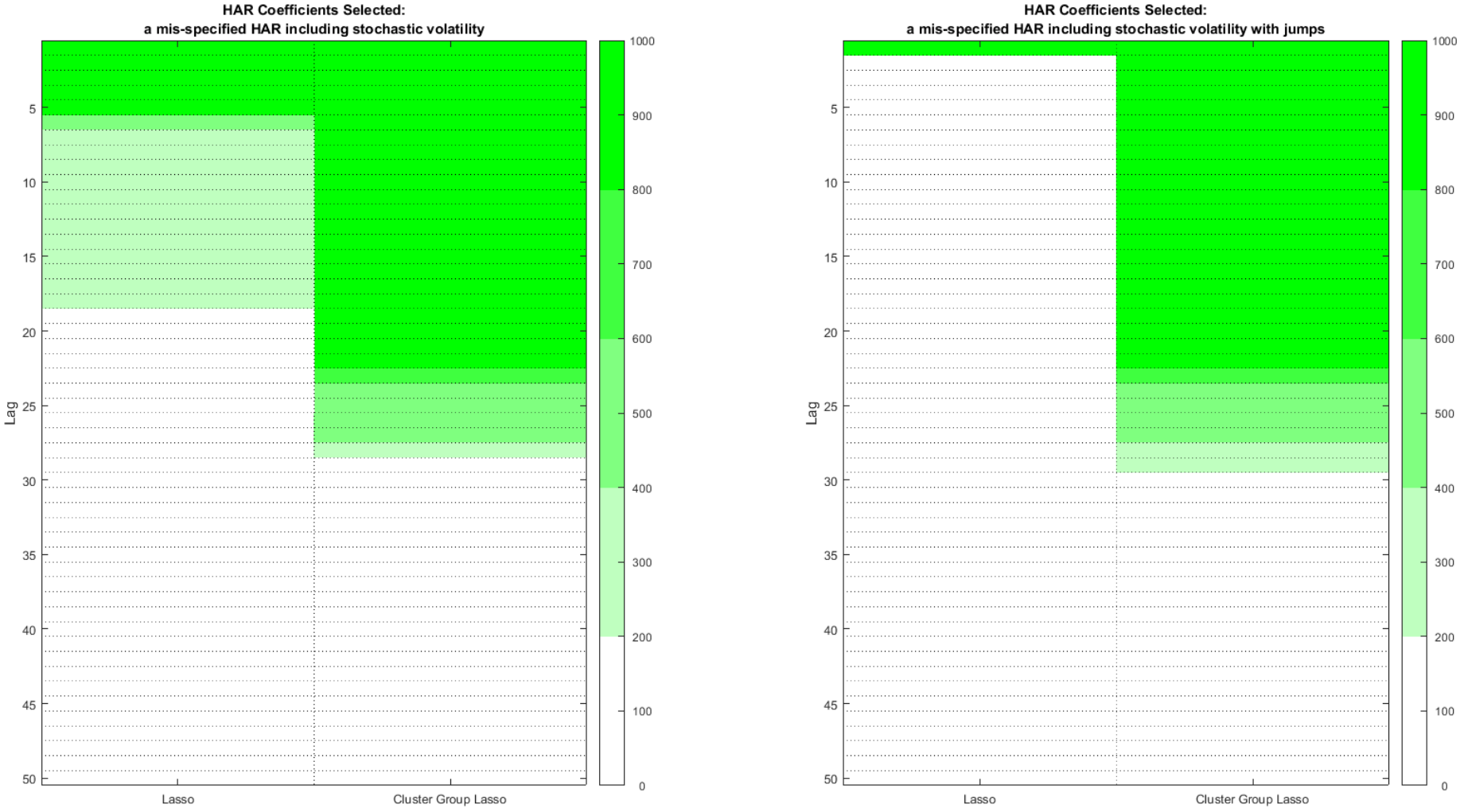

Figure 5

HAR Coefficients Selected with $P = 150$: the comparison of Lasso and Cluster Group Lasso. This figure reports the times per 1000 replications a lag has been selected by the model selection device considered. Data is simulated by a mis-specified HAR model with stochastic volatility and stochastic volatility together with jumps.