Predicting the Long-term Stock Market Volatility: A GARCH-MIDAS Model with Variable Selection

Tong Fang[1], Tae-Hwy Lee[2], Zhi Su[3]
1. School of Economics, Shandong University, Jinan, China
2. University of California, Riverside, USA
3. School of Finance, Central University of Finance and Economics, Beijing, China
May 2020

Abstract: We consider a GARCH-MIDAS model with short-term and long-term volatility components, in which the long-term volatility component depends on many macroeconomic and financial variables. We select the variables that exhibit the strongest effects on the long-term stock market volatility via maximizing the penalized log-likelihood function with an Adaptive-Lasso penalty. The GARCH-MIDAS model with variable selection enables us to incorporate many variables in a single model without estimating a large number of parameters. In the empirical analysis, three variables (namely, housing starts, default spread and realized volatility) are selected from a large set of macroeconomic and financial variables. The recursive out-of-sample forecasting evaluation shows that variable selection significantly improves the predictive ability of the GARCH-MIDAS model for the long-term stock market volatility.

Key words: Stock market volatility, GARCH-MIDAS model, Variable selection, Penalized maximum likelihood, Adaptive-Lasso

JEL classification: C32, C51, C53, G12

1. Introduction

Movements in aggregate financial volatility have significant impacts on capital investment, consumption, and economic activities (Fornari and Mele, 2013). What triggers the changes of aggregate financial volatility has drawn much attention among academics and practitioners. Many researchers relate aggregate volatility to macroeconomic variables (Officer, 1973; Schwert, 1989). Engle and Rangel (2008) find that links between economic fundamentals and aggregate financial volatility exist, but they are much weaker than seems reasonable. In addition to macroeconomic indicators, financial market variables including past volatility are also considered in predicting volatility (Ghysels et al., 2006; Christiansen et al., 2012). Recent studies are Engle et al. (2013), Asgharian et al. (2013) and Conrad and Loch (2015). Engle et al. (2013) propose a GARCH-MIDAS model, with the long-term component directly driven by inflation and industrial production. Asgharian et al. (2013) and Conrad and Loch (2015) both employ the GARCH-MIDAS framework to investigate the relationships between macroeconomic indicators and aggregate financial volatility.

The GARCH-MIDAS model, proposed by Engle et al. (2013), is a component model of volatility. The component GARCH models have been researched for more than 20 years. Ding and Granger (1996) consider a two component model, with an IGARCH(1,1) specification for the long-memory component, and a GARCH(1,1) process for the short-term component. Engle and Lee (1999) propose an additive component GARCH model, where the conditional variance is specified as the sum of two components: one is persistent with a near unit root, and the other component is mean-reverting with rapid time decay. They also indicate that decomposition of volatility into several components is useful in tests of economic and asset pricing hypothesis. Bauwens and Storti (2009) generalize the model of Ding and Granger (1996) by modeling the volatility as a convex combination of unobserved components where the combination weights are time varying. Engle and Rangel (2008) relax the assumption that the trend in the volatility process reverts to a constant level, and introduce the Spline-GARCH model, where the two components are separated using a multiplicative decomposition. The short-term component is modeled as a GARCH process evolving around a long-term component which reflects macroeconomic conditions, with the long-term component being specified using an exponential quadratic spline. The Spline-GARCH model is useful to understand the long-term or low-frequency volatility in a macroeconomic environment. However, these models do not relate macroeconomic variables with the long-term volatility component, until Engle et al. (2013) who propose a GARCH-MIDAS model. The GARCH-MIDAS model directly incorporates low-frequency macroeconomic variables in the long-term volatility component. The GARCH-MIDAS model has been the most popular model that is used to investigate the relationships between aggregate financial volatility and macroeconomic or financial variables (Asgharian et al., 2013; Conrad et al., 2014; Conrad and Loch, 2015; Pan et al., 2017; Su et al., 2017; Conrad and Kleen, 2019).

This paper characterizes the relationships between the long-term stock market volatility and macroeconomic & financial indicators, and employs a GARCH-MIDAS model that includes a variety of explanatory variables in the long-term volatility component. For a GARCH-MIDAS model with a large number of macroeconomic & financial variables, the “Adaptive-Lasso” of Zou (2006) is applied to determine which variables exhibit the strongest effects on the future long-term stock market volatility. Then we estimate the impacts of the selected variables on the long-term stock market volatility and analyze their "Beta weighting" schemes of the MIDAS model (Ghysels et al., 2007). We also compare the volatility predictive ability of our model with other GARCH-MIDAS models as in Conrad and Loch (2015) and Engle et al. (2013) in out-of-sample forecast evaluations.

Our contribution to the literature on stock market volatility predictability is twofold. First, we introduce variable selection in the long-term volatility component of the GARCH-MIDAS model, which helps us to determine the most important variables in predicting the long-term stock market volatility. Inspired by Engle et al. (2013) and Boffelli et al. (2017), we consider many covariates in a single model. However, the model with many covariates involves a large number of parameters, which increases estimation complexity and reduces estimation efficiency. It is difficult to give accurate interpretations on the parameter estimates. Therefore, we combine the Adaptive-Lasso with the log-likelihood function of the GARCH-MIDAS model, and estimate the parameters by maximizing the penalized log-likelihood function under linear constraints. An estimation procedure, which is similar to Ghysels and Qian (2019), is used to avoid an identification issue in variable selection. We choose the optimal tuning parameter of the Adaptive-Lasso using the Generalized Information Criteria (GIC).

Second, we select the variables that play the most important roles in predicting the long-term stock market volatility, and provide further evidence on the countercyclical pattern of financial volatility. We also estimate the GARCH-MIDAS model with the selected variables (Post-selection estimation), and analyze the parameter estimates and the dynamic structure of the estimated Beta weights. The empirical results show that the realized volatility (RV) is the most important determinant of stock market volatility. Previous studies show macroeconomic indicators play a significant role in predicting market volatility (Engle et al., 2013; Conrad and Loch, 2015; Conrad and Kleen, 2019). This paper provides novel evidence that the role of macroeconomic indicators may have been overstated.

Our main empirical results are summarized as follows:

(1) Three variables, which are housing starts, default spread and realized volatility, are selected for the full sample period. The realized volatility is the most important one among them. The real GDP and industrial production that have been always considered in previous studies are surprisingly not selected.

(2) Post-selection estimation results show that housing starts have a negative impact, and default spread and RV have positive impacts, on the long-term stock market volatility. The negative impact of

housing starts confirms the countercyclical phenomenon of stock market volatility.

(3) The out-of-sample forecast evaluations show that the model with selected variables significantly outperforms the other GARCH-MIDAS models, except for the model with RV. The results indicate that the GARCH-MIDAS model with variable selection reveals the best predictor of the long-term stock market volatility.

From an economic viewpoint, macroeconomic fundamentals have been considered to drive stock market volatility and to outperform RV in out-of-sample forecasting (Engle et al., 2013; Conrad and Loch, 2015; Conrad and Kleen, 2019). These related studies reveal a close linkage between aggregate volatility and macroeconomic conditions. However, this paper shows that historical volatility information (RV) significantly outperforms macroeconomic variables. Therefore, the role of macroeconomic variables in previous studies may have been overstated. Moreover, results in this paper reveal the volatility clustering phenomenon in the long-term volatility component. Volatility clustering phenomenon is a well-known stylized feature of financial asset returns. Most of previous studies mainly analyze volatility clustering within the class of ARCH and GARCH models in the short run (Engle, 1982; Bollerslev, 1986). We actually show the evidence for the long-term volatility clustering.

The remainder of this paper is organized as follows. Section 2 introduces the GARCH-MIDAS model with variable selection, and discusses the choice of the tuning parameter. Section 3 describes the data. Section 4 reports the empirical results, including the variable selection, post-selection estimation, out-of-sample forecast evaluations, and robustness checks. Section 5 concludes.

2. GARCH-MIDAS model and variable selection

2.1 The GARCH-MIDAS model

We first introduce the GARCH-MIDAS model proposed by Engle et al. (2013) and Conrad and Loch (2015). The model extracts two components of volatility, a short-term component following a mean reverting high-frequency daily GARCH process, and a long-term component incorporating explanatory variables of low frequency using the Beta weighting schemes.

The stock market daily log return $r_{i,t}$ at day $i = 1, \dots, N_t$ in a period $t = 1, \dots, T$ (e.g., month, quarter) is represented in the following specification:

$$r_{i,t} - E_{i-1,t}(r_{i,t}) = \sqrt{g_{i,t}\tau_t}\varepsilon_{i,t}, \tag{1}$$

where $E_{i-1,t}(\cdot)$ is the conditional expectation given $\Gamma_{i-1,t}$, the information set up to day $(i-1)$ of period t, and $\varepsilon_{i,t}|\Gamma_{i-1,t} \sim N(0,1)$. Daily expected returns are set to be constant μ. The total number of daily observations is denoted as $N_0 = \sum_{t=1}^{T} N_t$. Equation (1) shows that stock market volatility has the two components: the short-term volatility component $g_{i,t}$, and the long-term volatility component τ_t.

The short-term volatility component that accounts for daily fluctuations follows a mean-reverting asymmetric GARCH(1,1) process:

$$g_{i,t} = (1 - \alpha - \beta - \gamma/2) + \left(\alpha + \gamma \cdot \mathbf{1}_{\{r_{i-1,t} - \mu < 0\}}\right) \cdot \frac{(r_{i-1,t} - \mu)^2}{\tau_t} + \beta g_{i-1,t}, \tag{2}$$

with the constraints of $\alpha > 0$, $\beta > 0$ and $\alpha + \beta + \gamma/2 < 1$. This model ensures that $E[g_{i,t}] = 1$. The parameter γ contains the information of asymmetry.

The long-term volatility component with a single explanatory variable is given by:

$$\log(\tau_t) = m + \theta \sum_{k=1}^{K} \varphi_k(\omega_1, \omega_2) X_{t-k}, \tag{3}$$

where $\log(\tau_t)$ is considered rather than τ_t in order to ensure the positivity of the long-term volatility and $\varphi_k(\omega_1, \omega_2)$ is the Beta weighting scheme:

$$\varphi_k(\omega_1, \omega_2) = \frac{(k/(K+1))^{\omega_1 - 1} \cdot (1 - k/(K+1))^{\omega_2 - 1}}{\sum_{l=1}^{K} (l/(K+1))^{\omega_1 - 1} \cdot (1 - l/(K+1))^{\omega_2 - 1}}, \tag{4}$$

The weights φ_k are completely determined by two parameters ω_1 and ω_2. It is easy to find that $\varphi_k \geq 0$ for $k = 1, \cdots, K$, and $\sum_{k=1}^{K} \varphi_k = 1$. The Beta weighting schemes can generate decaying, hump-shaped, or U-shaped weights (Ghysels et al., 2007).

The GARCH-MIDAS model has been the most popular methodology for investigating the relationships between stock market volatility and economic variables of low frequency (Asgharian et al., 2013; Conrad et al., 2014; Conrad and Loch, 2015; Su et al., 2017; Pan et al., 2017; Boffelli et al., 2017). However, most of these studies focus on the effects of one variable at a time on the stock market volatility, while many economic and financial variables can lead to changes of stock market volatility. It would be desirable to include all these potentially useful predictors at once in a single model. In this regards, Engle et al. (2013) estimate a single model that combines four variables (the level and volatility of PPI, and the level and volatility of industrial production). Also, Boffelli et al. (2017) use six variables to estimate a GARCH-MIDAS model. Inspired these two papers, we consider including a "large" number of variables in a single GARCH-MIDAS model with modifying Equation (3) as follows:

$$\log(\tau_t) = m + \sum_{j=1}^{J} \theta_j \sum_{k=1}^{K} \varphi_k(\omega_{j,1}, \omega_{j,2}) X_{j,t-k}, \tag{5}$$

where J is the number of explanatory variables ($J = 20$ in this paper), and θ_j measures the impact of the jth variable on the long-term stock market volatility.

Therefore, our GARCH-MIDAS model consists of Equation (1), (2), (4) and (5) with *J* being large. The log-likelihood function is given by Equation (6), and Φ denotes all the parameters that are to be estimated. The model is usually estimated through the quasi-maximum likelihood estimation, and $\widehat{\Phi}$ denotes all the parameter estimates. The asymptotic standard errors are estimated consistently under the assumption of conditional normality.

$$LLF(\Phi) = -\frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\left[\log(2\pi) + \log\left(g_{i,t}(\Phi)\tau_t(\Phi)\right) + \frac{\left(r_{i,t}-\mu\right)^2}{g_{i,t}(\Phi)\tau_t(\Phi)}\right]. \quad (6)$$

2.2 GARCH-MIDAS with variable selection

The number of parameters is $3J + 1$ in Equation (5), which will be a very large number if J is large. With a large number of parameters to be estimated, it may become difficult to identify variables that exhibit the strongest effects, and it may be impossible to give accurate interpretations on the parameter estimates (Tibshirani, 1996). In this paper, we employ variable selection in the long-term volatility component in Equation (5), and use the Adaptive-Lasso of Zou (2006) for the penalized log-likelihood function:

$$PLLF_\lambda(\Phi) = -\frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\left[\log(2\pi) + \log\left(g_{i,t}(\Phi)\tau_t(\Phi)\right) + \frac{\left(r_{i,t}-\mu\right)^2}{g_{i,t}(\Phi)\tau_t(\Phi)}\right] - \lambda\sum_{j=1}^{J}\widehat{w}_j\left|\theta_j\right|, \quad (7)$$

where $\lambda > 0$ is the tuning parameter, and $PLLF_\lambda(\Phi)$ denotes the penalized log-likelihood function for a given λ, and $\widehat{w}_j$ is the adaptive weight for θ_j. Notably, the variable selection in GARCH-MIDAS model is not totally the same to that in linear regression of Tibshirani (1996). We actually select the Beta weighted average of lagged X_j, which is denoted as $\sum_{k=1}^{K}\varphi_k\left(\omega_{j,1},\omega_{j,2}\right)X_{j,t-k}$ in Equation (5), instead of X_j.

To obtain the adaptive weights, we first estimate a GARCH-MIDAS model with all J variables, and obtain the preliminary estimates $\hat{\theta}_j$ from maximizing Equations (6) under the constraints of $\alpha > 0$, $\beta > 0$ and $\alpha + \beta + \gamma/2 < 1$. Then the adaptive weight is calculated as $\widehat{w}_j = 1/\left|\hat{\theta}_j\right|^\eta$. In the simulation of Zou (2006), the probability of containing the true model is the highest when $\eta = 2$, so we take $\eta = 2$ in this paper.

For a given tuning parameter λ, we maximize $PLLF_\lambda(\Phi)$ subject to the linear constraints of $\alpha > 0$, $\beta > 0$ and $\alpha + \beta + \gamma/2 < 1$. The Broyden-Fletcher-Goldfard-Shanno (BFGS) algorithm is used for this optimization. $\widehat{\Phi}_\lambda$ denotes the parameter estimates from the maximization problem for the given λ.

2.3 Choosing the tuning parameter

Choosing the tuning parameter λ in the penalized log-likelihood estimation is important for determining the underlying true model. Cross-validate (CV) and information criteria (AIC and BIC) are widely used in model selection. Yang (2005) indicates that CV is asymptotically equivalent to AIC, implying that CV behaves similarly to AIC. Wang et al. (2009) propose a modified BIC which works for tuning parameter selection. In this paper, the tuning parameter is determined by Generalized Information Criteria (GIC), which is proposed by Fan and Tang (2012). GIC contains two components, the first component is used to evaluate the goodness of fit, and the second component is a penalty on the model complexity, which implies GIC trades off between model fitting and model complexity. GIC_λ denotes this information criteria for a given λ, and a GIC applied to the penalized log-likelihood function is shown in Equation (8):

$$GIC_\lambda = \frac{1}{N_0}\left\{2\left[LLF\left(\widehat{\Phi}\right) - PLLF_\lambda\left(\widehat{\Phi}_\lambda\right)\right] + a(N_0, p)\left|\hat{\theta}_\lambda\right|\right\}, \tag{8}$$

where $a(N_0, p)$ is a positive value depending on the number of total observations N_0, and the number of parameters $p = 3J + 1$ in the long-term volatility component. $LLF\left(\widehat{\Phi}\right)$ is the value of maximized log-likelihood function without variable selection, and $PLLF_\lambda\left(\widehat{\Phi}_\lambda\right)$ is the value of maximized penalized log-likelihood function. $2\left[LLF\left(\widehat{\Phi}\right) - PLLF_\lambda\left(\widehat{\Phi}_\lambda\right)\right]$ indicates the scaled deviation measure that is used to evaluate goodness of fit. $\left|\hat{\theta}_\lambda\right|$ denotes the number of non-zero elements in $\hat{\theta}_\lambda$, where $\hat{\theta}_\lambda$ is estimated from Equation (8) given the tuning parameter λ. Fan and Tang (2012) propose a uniform choice of $a(N_0, p) = \log\{\log(N_0)\} \cdot \log(p)$. For practical implementation, the tuning parameter is considered over a range from 0 to $\lambda_{\max}$, and we take the value of λ corresponding to the minimum GIC_λ as the optimal tuning parameter.

2.4 Estimation and identification

As the tuning parameter λ increases, some parameters θ will be shrunken to zero, and the corresponding variables will be dropped from the GARCH-MIDAS model. However, the variable selection procedure will lead to an identification problem. For example, once a parameter θ_j is shrunken to zero, the corresponding $\omega_{j,1}$ and $\omega_{j,2}$ in the Beta weighting schemes will not be entered in the penalized log-likelihood function. Thus the parameters $\omega_{j,1}$ and $\omega_{j,2}$ are not identified.

To avoid this identification problem, we use an estimation procedure which is similar to the estimation approach in Ghysels and Qian (2019), who estimate MIDAS regressions with polynomial parameter profiling. The idea of profiling has been widely discussed (Patefield, 1977; Barndorff-Nielsen, 1983; Barndorff-Nielsen and Cox, 1994). Suppose that the log-likelihood function depends on a parameter space

$\Phi = (\Phi_1, \Phi_2)$. The log-likelihood function may be difficult to maximize over the entire parameter space in many situations. However, if we first fix $\Phi_2 = \bar{\Phi}_2$, then maximizing the log-likelihood function with respect to Φ_1 may become easier.

For the GARCH-MIDAS model, let $\Phi_2 = \left(\omega_{1,1}, \omega_{1,2}, \omega_{2,1}, \omega_{2,2} \dots, \omega_{20,1}, \omega_{20,2}\right)$ be the Beta weighting parameters and $\Phi_1 = (\mu, \alpha, \beta, \gamma, m, \theta_1, \theta_2, \dots, \theta_{20})$ be the rest of the parameters. We fix the parameter $\Phi_2 = \hat{\Phi}_2$, where $\hat{\Phi}_2$ are the parameter estimates by estimating the GARCH-MIDAS model with all 20 variables. The GARCH-MIDAS model with variable selection is estimated in three steps, as shown in Table 1.

[INSERT TABLE 1 HERE]

3. Data description

We focus on the S&P500, and U.S. macroeconomic & financial data from 1969Q1 to 2018Q4. The stock market returns are in daily frequency, and the macroeconomic & financial variables are collected in quarterly frequency. The S&P500 index data are obtained from the Center for Research in Security Prices (CRSP), and we calculate the daily stock market returns as the natural logarithm of S&P500 index prices.

Data revision can be substantial for macroeconomic & financial variables (Conrad and Loch, 2015). Using revised instead of first release data (real-time) can be misleading in forecast evaluations, and the only safe way to evaluate forecasting models is with real-time data (Stark and Croushore, 2002; Stark, 2010; Croushore, 2011). Many studies confirm the importance of the use of real-time data. For example, Oh and Waldman (1990) show that economic activities sensitively respond to the announcements of economic indicators, meaning that the first release data definitely lead the economic activity. Diebold and Rudebusch (1991) find that the use of real-time data is crucial, since the variables included in the index of leading indicators are chosen *ex-post*. Robertson and Tallman (1998) show that real time data may be useful in forecasting real output. As employing the first release (real-time) data is of great necessity in forecast evaluations, we collect the first release macroeconomic & financial data from the Real-time Data Research Center (RDRC) of the Federal Reserve Bank of Philadelphia.[1] The other variables are obtained

[1] Actually, variables considered in this paper are not fully in real time. The real GDP growth rate, industrial production growth rate, unemployment rate, housing starts, real personal consumption, CPI, CFNAI, new orders index, consumer sentiment index, and term spread are collected from the Real-time Data Research Center. Some of the financial variables will not be revised after release, and they can also be seen as real time data. *Journal of Applied Econometrics* provides the dataset used in Conrad and Loch (2015). We use the same dataset as in Conrad and Loch (2015) and update the data to 2018Q4.

from the FRED database at the Federal Reserve Bank of St Louis, the Federal Reserve Bank of Chicago (FRBC), Quandl.com, the Survey of Consumers from University of Michigan (SCUM), and the personal website of K. R. French and A. Manela. For the macroeconomic & financial data that are available at a monthly or daily frequencies, we take quarterly averages. Descriptive statistics are reported in Table 2.[2]

[INSERT TABLE 2 HERE]

3.1 Macroeconomic variables

We consider the following macroeconomic variables that have been included in Christiansen et al. (2012), Engle et al. (2013), Asgharian et al. (2013), Conrad and Loch (2015) and Conrad and Kleen (2019): real GDP growth rate, industrial production growth rate, unemployment rate, housing starts, nominal corporate profits after tax, real personal consumption, CPI, PPI, the Chicago Fed national activity index (CFNAI), the new orders index of the Institute of Supply Management (ISM), monetary base, and the University of Michigan consumer sentiment index. CFNAI is an index designed to gauge overall economic activity and related inflationary pressure, which can be seen as a proxy of business cycles. The new orders index measures changes in new orders, supplier deliveries, inventories, production and employment, and it is a proxy of future activity in any industry, which can be seen as a leading economic indicator. The real GDP growth rate, industrial production growth rate, unemployment level, housing starts, corporate profits, personal consumption and new orders index are seasonally adjusted.

We consider CFNAI in levels, and take the first difference of level data for unemployment rate and consumer sentiment index. We take log difference of level data for new orders index. For the other variables, we take annualized quarter-over-quarter percentage changes as $100((X_t/X_{t-1})^4 - 1)$, following Engle et al. (2013) and Conrad and Loch (2015).

Volatilities of macroeconomic variables are also important determinants of stock market volatility (Schwert, 1989; Engle et al., 2013; Asgharian et al., 2013), and a GARCH(1,1) model is used to estimate quarterly volatility of macroeconomic variables as mentioned in Engle et al. (2013). The volatility of macroeconomic activity is measured by volatility of first release real GDP growth rate, and the volatility of inflation is measured by volatility of CPI.[3]

3.2 Financial variables

[2] Quandl.com is a database that offers financial and economic data.

[3] The volatility estimated by GARCH(1,1) model is also seen as macroeconomic or inflation uncertainty (Caporale and McKiernan, 1998; Fountas et al., 2006).

We employ six financial variables in this paper: term spread, default spread, equity market returns (MKT), short-term reversal factor (STR), implied volatility (IV) and realized volatility (RV). The term spread is calculated as the difference between the 10-year Treasury bond yield and the 3-month Treasury bill rate. Default spread is the yield spread between BAA and AAA rated bonds, which should affect aggregate volatility, according to Merton (1974). Equity market returns (MKT) in Fama and French (1992) can capture the leverage effect (Black, 1976; Glosten et al., 1993; Christiansen et al., 2012). Nagel (2012) finds that the short-term reversal factor (STR) can be related to market volatility. Realized volatility is also considered in volatility forecasts (Andersen et al., 2003; Ghysels et al., 2006; Andersen et al., 2011). The quarterly realized volatility is calculated as:

$$RV_t = \sum_{i=1}^{N_t} r_{i,t}^2. \tag{9}$$

The implied volatility indices, CBOE VIX and VXO, are used to measure market expectation of volatility conveyed by stock index option prices, and they are important in forecasting future financial volatility (Busch et al., 2011). The implied volatility is also a proxy of financial market uncertainty (Chung and Chuwonganant, 2014). Becker et al. (2009) find that VIX subsumes information relating to past jump contributions to aggregate volatility, and reflects information of future jump activity. Bekaert and Hoerova (2014) indicate that VIX has a high predictive power for financial instability. Thus, we consider the implied volatility as an explanatory variable in the long-term volatility component. However, VIX and VXO are only available since 1990 and 1986 respectively, and they do not match the time period of the other variables. Manela and Moreira (2017) propose a news-based implied volatility index (NVIX) that captures investors' perception of future uncertainty, and it is actually an estimated VXO index using the data from the front-page articles of the *Wall Street Journal*. NVIX is confirmed to be a source of financial aggregate volatility (Su et al., 2017), and we use NVIX as a proxy of implied volatility before 1986. We consider term spread, default spread, MKT and STR in levels, and take log difference of NVIX.[4]

4. Empirical analysis

4.1 Variable selection using Adaptive-Lasso

For practical implementations, we consider the tuning parameter λ over a 151-point grid of $[0, 15]$ with an increment of 0.1. Following Tibshirani (1996), we standardize all the macroeconomic & financial

[4] Since all of the macroeconomic & financial variables are included in Equation (5), it is necessary that all the variables have same frequency.

variables for the variable selection. We estimate the model with every $\lambda \in [0, 15]$, and we choose the tuning parameter corresponding to the minimum GIC.[5] Since the values of the GIC are not smooth over the different values of λ due to the estimation algorithm, we also use the smoothed GIC that is obtained using the Hodrick Prescott (HP) filter instead of the GIC to determine the optimal tuning parameter. Figure 1 shows the GIC and smoothed GIC as functions of λ, and we take $\lambda = 12.8$ as the optimal value of λ.

[INSERT FIGURE 1 HERE]

Figure 2 presents the parameter estimates of θ for each value of λ. When $\lambda = 12.8$, we maximize $PLLF_{\lambda=12.8}(\Phi)$ under linear constraints, and we find that the parameter estimates θ_j of the three variables, namely, the housing starts, default spread and RV, are not shrunken to zero. These three variables are selected from the 20 variables. We present the values of λ at which each parameter θ_j reaches zero in Figure 3. The parameters of four variables, which are the real GDP, consumption, monetary base and GDP volatility, first reach zero at $\lambda = 0.4$. The parameters of RV and housing starts do not reach zero at $\lambda = 15.0$. Figure 2 also shows that the parameter estimate for RV is larger than 0.2. RV will be the last variable that is dropped from the model when λ is very large.

[INSERT FIGURE 2 HERE]

The variable selection for the GARCH-MIDAS model provides us with a new perspective to determine which variable is the most important for predicting the long-term stock market volatility. Conrad and Loch (2015) find that the real GDP and industrial production growth rate are negatively associated with the future long-term aggregate volatility, which is the well-known countercyclical pattern that was mentioned by Officer (1973) and Schwert (1989). Choudhry et al. (2016) also reveal a strong relationship between the industrial production growth rate and stock market volatility. However, interestingly, these two variables are not selected in our estimation. The real GDP and industrial production growth rate have been emphasized in predicting volatility for more than 40 years, and the variable selection results indicate that their roles in predicting the stock market volatility may have been overestimated.

[INSERT FIGURE 3 HERE]

Although previous studies show that the implied volatility has a significant impact on the stock market volatility (Bekaert and Hoerova, 2014; Su et al., 2017), the implied volatility is not selected in the variable selection procedure. The volatility of real GDP and inflation, which are also known as economic

[5] The lag length $K = 12$ in Equation (5) is determined following Conrad & Loch (2015).

and inflation uncertainty, respectively, are not considered to be important variables for predicting the long-term stock market volatility. However, we are not saying that uncertainty is not a key variable in volatility predictions. There are many uncertainty indices that we do not consider in this paper due to data availability, including Economic Policy Uncertainty Index (Baker et al., 2016), macro uncertainty (Jurado et al., 2015), and financial uncertainty (Ludvigson et al., 2019). Whether these indices would be selected requires further research.

The housing starts, default spread and RV are selected. The literature on the relationships between the housing starts and financial volatility is quite limited. The most similar study is Löffler (2013), which considers the role of skyscraper construction starts in U.S. stock return predictions. Although skyscraper construction is different from housing construction, the research of Löffler (2013) inspires our reasons regarding why housing starts matter for the stock market volatility prediction: (1) Housing starts can be seen as a leading indicator of economic activities, and more housing construction implies positive expectations of the future economy, which motivates more financial investments. (2) The housing market is closely associated with the credit market, and a higher supply of new houses indicates an expansion of the credit market, which is a driving force of economic growth. (3) Housing starts are found to be strongly and positively affected by the monetary base which also stimulates the economy (Huang, 1973). The default spread and MKT, which are considered to be associated with the effects of leverage, are robust predictors of stock market volatility (Glosten et al., 1993; Christiansen et al., 2012). In our model estimation, the default spread performs better than the MKT when predicting the long-term stock market volatility. Conrad and Loch (2015) indicate that the model with macroeconomic variables does not yield lower mean squared errors (MSE) than the model with RV at 1-quarter ahead forecasting. The variable selection results show that RV is still a powerful predictor of the future financial volatility. The past market volatility contains the most important information that triggers the financial volatility.

4.2 Post-selection estimation

After the variable selection, we estimate a GARCH-MIDAS model with the selected variables (which we refer to as the post-selection estimation). The long-term volatility component with the three selected variables is given by:

$$\log(\tau_t) = m + \theta^{HS} \sum_{k=1}^{12} \varphi_k(\omega_1^{HS}, \omega_2^{HS}) HS_{t-k} + \theta^{DS} \sum_{k=1}^{12} \varphi_k(\omega_1^{DS}, \omega_2^{DS})\, DS_{t-k} + \theta^{RV} \sum_{k=1}^{12} \varphi_k(\omega_1^{RV}, \omega_2^{RV}) RV_{t-k}, \tag{10}$$

where *HS*, *DS*, and *RV* denote housing starts, default spread and realized volatility, respectively.

With respect to the post-selection estimator, Belloni et al. (2012) and Belloni and Chernozhukov (2013) show that the post-selection estimator is consistent for the true parameter. For the inference, we need to assess the significance of the parameter estimates. The variable selection can have a detrimental impact on the subsequently constructed inference procedures, such as confidence intervals, if these are constructed in a "naïve" way where the model selection is ignored (Berk et al., 2013; Leeb et al., 2015). The asymptotically valid confidence intervals may be obtained by following Belloni et al. (2016) and Belloni et al. (2018). In this paper, we use the naïve confidence intervals that are constructed as if the model with the selected variables is correct and fixed a priori (thus ignoring variable selection). Nevertheless, Leeb et al. (2015) show that the actual coverage probability of the naïve confidence interval only moderately deviates from the desired nominal coverage probability, which supports the use of the naïve confidence intervals in the post-selection inference.

[INSERT TABLE 3 HERE]

The post-selection estimation results are reported in Table 3. We are interested in the parameter estimates of θ, which measure the impacts of the selected variables on the long-term stock market volatility. The estimated parameter on housing starts is -0.0138 and significant at the 1% level, indicating that housing starts significantly predict the long-term stock market volatility. More new housing construction will lead to lower future aggregate volatility. The estimated parameter of the default spread is 0.4613, and it is positive and significant at the 1% level. The default spread captures the leverage effect, which makes it a driver of stock market volatility. The default spread tends to widen when firms' credit risk increases, which leads to higher stock market volatility (Black, 1967; Nelson, 1991; Christiansen et al., 2012). RV is positively associated with future market volatility.

Previous studies find that aggregate financial volatility is countercyclical (Officer, 1973; Schwert, 1989; Engle et al., 2013; Conrad and Loch, 2015). We confirm this conclusion in our new framework. Housing starts is a leading indicator that reflects the positive expectations and economic expansion of the future economy. Since it has a negative impact on the stock market volatility, we provide evidence for the countercyclical pattern of aggregate volatility from the perspective of housing starts.

[INSERT FIGURE 4 HERE]

In addition to parameter estimates, we plot the Beta weighting schemes for the three selected variables in Figure 4. The Beta weighting schemes help us to determine which lag of the variables has the strongest effect. The weighting scheme of housing starts is hump-shaped. The maximum weights are on the 3rd and 4th lags of housing starts. The weighting schemes of the default spread and RV follow extremely decaying patterns, as the weights on the 1st lagged default spread and RV are almost 1. The estimated weighting schemes show that considering one lagged default spread or RV is sufficient for predicting volatility. The estimated weighting schemes reveal the heterogeneous impacts of these three variables on the long-term stock market volatility.

4.3 Out-of-sample forecast evaluations

To evaluate the out-of-sample forecast performance of the GARCH-MIDAS model with the selected variables, we consider the 1-/2-/3-/4-quarter-ahead forecasts. All the models that are considered in the out-of-sample forecast evaluations are described as follows.

Model 1: GARCH-MIDAS model with the selected variables.

Model 2: GARCH-MIDAS model that incorporates one variable at a time in the long-term volatility component (Conrad and Loch, 2015).[6]

Model 3: GARCH-MIDAS model with all 20 macroeconomic & financial variables.[7]

Model 4: GARCH-MIDAS model that incorporates one principal component (PC) at a time following Conrad and Loch (2015). We employ the principal components obtained from all the 20 variables, and use the first three principal components in the long-term volatility components. The first, second and third principal component (PC_1, PC_2, PC_3) respectively accounts for 27.80%, 14.42% and 11.33% of the variation in the 20 variables.[8]

$$\log(\tau_t) = m + \theta \sum_{k=1}^{K} \varphi_k(\omega_1, \omega_2) PC_{j,t-k}\, , j = 1,2,3. \tag{11}$$

Model 5: GARCH-MIDAS model that incorporates the real-time variables and their corresponding median SPF forecasts as in Conrad and Loch (2015). The long-term component is given by Equation (12):[9]

$$\log(\tau_t) = m + \theta\left(\sum_{k=1}^{12} \varphi_k(\omega_1, \omega_2) X_{t-k} + \sum_{k=-3}^{0} \varphi_k(\omega_1, \omega_2) X_{t-k|t-1}^{SPF}\right), \tag{12}$$

where $X_{t-k|t-1}^{SPF}$ is the median SPF forecast of variable X_{t-k} that are based on information available in $t-1$.

Model 6: GARCH-MIDAS model that incorporates both the level and variance of the industrial

---

[6] There will be 20 different models for Model 2 due to each of the 20 variables being considered in this paper.
[7] Model 3 consists of Equation (1), (2), (4) and (5), as shown in Section 2.1.
[8] There will be 3 different models for Model 5.
[9] We consider the real time data and median SPF forecasts for real GDP, industrial production, unemployment rate, housing starts and corporate profits.

production (IP) and PPI in Engle et al. (2013). The long-term component is given by Equation (13):

$$\log(\tau_t) = m + \theta_1 \sum_{k=1}^{K} \varphi_k(\omega_1, \omega_2) X_{t-k}^{level} + \theta_2 \sum_{k=1}^{K} \varphi_k(\omega_1, \omega_2) X_{t-k}^{variance}, \quad (13)$$

where X is either the IP or PPI.

The full sample is divided into the estimation sample from 1969Q1 to 2006Q4 (152 quarters) and the out-of-sample forecasting period from 2007Q1 to 2018Q4 (48 quarters). To evaluate volatility forecasts, we compare the predicted volatility with the true conditional variance. Since the true conditional variance is unobservable, a proxy for it is required. The squared daily return is one of the common proxies used; however, it has been seen as a noisy proxy. Patton (2011) indicates that the realized volatility is better than the squared daily return. In addition, some studies find that microstructure noise can be ignored using a higher frequency data (Awartani et al., 2009). Ghysels and Sinko (2011) also show that the 5-minute frequency can be considered a low-noise environment. In this paper, we apply the S&P500 daily realized volatility $RV^{5\text{min}}$, which is calculated from the 5-minute S&P500 intraday returns, and evaluate the volatility forecasts by comparing the predicted volatility with $RV^{5\text{min}}$.[10]

A recursive out-of-sample forecast is employed. The s-quarter-ahead long-term volatility component forecast $\hat{\tau}_{t+s}$ for $s = 1,2,3,4$ remains the same within one quarter, and the short-term volatility component forecast $\hat{g}_{i,t+s}$ for $s = 1,2,3,4$ can be iteratively calculated by Equation (2). The daily volatility forecast is $\hat{\tau}_{t+s}\hat{g}_{i,t+s}$, and the forecast error is $RV_{i,t+s}^{5\text{min}} - \hat{\tau}_{t+s}\hat{g}_{i,t+s}$. The mean squared forecast error (MSFE) is given by:

$$MSFE = \frac{1}{N_0 - \sum_{t=1}^{152} N_t} \sum_{t=153}^{T} \sum_{i=1}^{N_t} \left(RV_{i,t}^{5\text{min}} - \hat{\tau}_t \hat{g}_{i,t}\right)^2, \quad (14)$$

where N_0 = 12,611 is the total number of daily observations, and T = 200 is the total number of quarters, as shown in Table 1.

To compare the forecasting performance of a model over the benchmark, we present the ratio of the corresponding MSFE: MSFE/MSFE^{benchmark}. A ratio lower than one indicates a forecasting improvement over the benchmark model. We consider Model 1 as the benchmark. In addition to the MSFE, we also compare the conditional predictive abilities of Model 1 and Models 2-6 via the Giacomini and White (GW, 2006) test. For each horizon, we conduct tests of the conditional predictive ability of Models 2-6 over Model 1 using a squared error loss function.[11] In the GW test results, a positive sign indicates that Models

[10] The S&P500 daily realized volatility data are calculated over 5-minute intraday returns, and the data are obtained from the Oxford-Man Institute's realized library (Version 0.3), University of Oxford.

[11] The loss function can be the squared error loss, the absolute error loss, the lin-lin loss, the linex loss, etc. (Giacomini and White, 2006). We also use the squared error loss, and the results with an absolute error loss function are similar.

2-6 outperform the benchmark, while a negative sign indicates the opposite.

We select the three variables (housing starts, default spread, and RV) that exhibit the strongest effects on the long-term stock market volatility. This result is obtained using the full sample period from 1969Q1 to 2018Q4. However, the estimation period changes over time, and the selected variables may also change. The selected variables that are considered in Model 1 should be the most important variables in the period from 1969Q1 to 2006Q4. We estimate the GARCH-MIDAS model with variable selection again using the data from 1969Q1 to 2006Q4. The selection results show that four variables, which are housing starts, PPI, default spread and RV, are selected.

The forecasting evaluation results are reported in Table 4. We first examine the MSFE ratio. Model 3 has the highest MSFE ratio among all the models, which implies that incorporating all the macroeconomic and financial information would overfit and destroy the out-of-sample predictive ability. Model 1 outperforms Models 2-6 for the volatility forecasts for 1/2/3/4 quarter forecasting horizons, except for the model with RV. The MSFE ratio comparisons indicate that RV is the most powerful stock market volatility predictor among the 20 macroeconomic & financial variables. In addition, considering the median SPF forecasts in the GARCH-MIDAS model yields a lower MSFE than the model without the SPF forecasts. For example, the MSFE for the model with real GDP is 1.0121, which is higher than the MSFE for the model with real GDP plus real GDP median SPF forecasts. Our results provide evidence for the out-of-sample forecasts using the median SPF forecasts and confirm the conclusion of Conrad and Loch (2015) that models with SPF forecasts outperform models that only consider past economic variables. Previous research using the SPF forecasts to predict the stock market volatility is limited, and more attention should still be paid to the SPF forecast data when predicting the stock market volatility.

[INSERT TABLE 4 HERE]

The GW statistics are significant with negative signs at 1/2/3 forecasting horizons, except for the model with RV. The benchmark model (Model 1) significantly outperforms almost all the other models that are considered in this paper. Conrad and Loch (2015) find that the GARCH-MIDAS model with macroeconomic variables could not significantly outperform the model with RV for the 1 quarter horizon, and this paper provides further evidence that past volatility performs when all the predictors are included in a GARCH-MIDAS model.

4.4 Robustness checks

4.4.1 Subsample variable selection

The four variables that are selected using the sample from 1969Q1-2006Q4 do not change for every out-of-sample estimation. However, the most important variables may change as the sample period changes. It is unreasonable to estimate the model with these four variables in the out-of-sample forecasting unless the variable selection results are stable over time. Therefore, we investigate whether the results are stable by using subsample analysis. We employ a recursive out-of-sample forecast with the initial estimation sample from 1969Q1 to 2006Q4 in Section 4.3. The data of a new quarter are iteratively added into the estimation sample, which provides 48 subsamples. The variable selection results remain almost the same when only one quarter's data are added. Therefore, we take three subsamples, with 12, 24 and 36 quarters, as examples: 1969Q1-2009Q4, 1969Q1-2012Q4 and 1969Q1-2015Q4, respectively. The subsample variable selection results show that the selected variables remain almost the same over time. The default spread and RV are still selected. The stability of the variable selection results means that out-of-sample forecast evaluations with three selected variables would not lead to much bias.

4.4.2 Estimation with restricted Beta weighting schemes

Under the constraint of $\omega_1 = 1$, the Beta weighting scheme, which generates a decaying pattern of weights, is specified as follows:

$$\varphi_k(\omega_2) = \frac{(1-k/(K+1))^{\omega_2-1}}{\sum_{l=1}^{K}(1-l/(K+1))^{\omega_2-1}}. \tag{15}$$

Conrad and Loch (2015) test the null hypothesis $H_0: \omega_1 = 1$ with a likelihood ratio test. We do not do the likelihood ratio test, because we are more likely to estimate the weighting functions with more choices of shapes via an unrestricted Beta weighting function as in Equation (4). The use of a restricted Beta weighting schemes can lead to new results. Therefore, we estimate the model with variable selection under the constraint of $\omega_1 = 1$ in this section, in order to see how the constraint changes the overall results. We consider the tuning parameter λ on a 101-point grid of $[0, 20]$ with an increment of 0.2. The variable selection results show that industrial production, housing starts, consumer sentiment and RV are selected. Housing starts and RV remain the most important variables for predicting the long-term stock market volatility. The industrial production and real GDP are the indicators that reflect the macroeconomic fundamentals, and it seems that industrial production is more important than real GDP growth when predicting the long-term stock market volatility.

4.4.3 Variable selection using revised data

In addition to real-time data, we also use revised data in the estimation of the GARCH-MIDAS model with variable selection. Four variables (housing starts, unemployment rate, term spread and MKT) are selected, as shown in Figure 5. If we use the restricted Beta weighting schemes as described in Equation (15), housing starts, term spread, MKT and RV are selected. The MKT instead of default spread is selected. We mentioned in Section 3 that using revised data in out-of-sample forecast evaluations can be misleading (Stark and Croushore, 2002; Stark, 2010; Croushore, 2011). This section indicates that using revised data in variable selection can also lead to different results.

[INSERT FIGURE 5 HERE]

5. Conclusions

This paper estimates a GARCH-MIDAS model with variable selection by combining the log-likelihood function with the Adaptive-Lasso penalty. By maximizing the penalized log-likelihood function under linear constraints, we could determine the variables that play the most important roles in predicting the long-term stock market volatility. We also use an estimate procedure which is similar to the parameter profiling in Ghysels and Qian (2019), in order to solve the identification issues and provide a computationally attractive estimation procedure.

Three variables, namely, housing starts, default spread and RV, are selected for the full sample period. The post-selection estimation results show a negative impact of housing starts, which confirms the countercyclical pattern of the stock market volatility. The real GDP and industrial production growth rates that are always considered in previous literature are not selected, implying that the role of them in predicting the long-term stock market volatility may be overestated. The MFSE ratio and Giacomini and White (2006) test are used in the out-of-sample forecasting evaluations. The results indicate that the model with the selected variables outperforms the other models that are considered in this paper, except for the model with RV. The overall empirical results show that RV is the most powerful predictor of the long-term stock market volatility. Our results also indicate the long-term volatility clustering phenomenon, whereas the volatility clustering phenomenon has been discussed in Engle (1982) and Bollerslev (1986). The GARCH-MIDAS model with variable selection successfully reveals the most important variables for predicting the long-term stock market volatility.

REFERENCES

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and Forecasting Realized Volatility. Journal of Econometrics 71(2), 579-625.

Andersen, T. G., Bollerslev, T., & Meddahi, N. (2011). Realized Volatility Forecasting and Market Microstructure Noise. Journal of Econometrics 160(1), 220-234.

Asgharian, H., Hou, A. J., & Javed, F. (2013). The Importance of the Macroeconomic Variables in Forecasting Stock Return Variance: A GARCH-MIDAS Approach. Journal of Forecasting 32(7), 600-612.

Awartani, B., Corradi, V., & Distaso, W. (2009). Assessing Market Microstructure Effects via Realized Volatility Measures with an Application to the Dow Jones Industrial Average Stocks. Journal of Business and Economic Statistics 27(2), 251-265.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty. The Quarterly Journal of Economics 131(4), 1593-1636.

Barndorff-Nielsen, O. (1983). On a Formula for the Distribution of the Maximum Likelihood Estimator. Biometrika 70, 343-365.

Barndorff-Nielsen, O., & Cox, D.R. (1994). Inference and Asymptotics. Chapman and Hall.

Bauwens, L., & Storti, G. (2009). A Component GARCH Model with Time Varying Weights. Studies in Nonlinear Dynamics & Econometrics 13(2), 702-711.

Becker, R., Clements, A. E., & McClelland, A. (2009). The Jump Component of S&P 500 Volatility and the VIX Index. Journal of Banking & Finance 33(6), 1033-1038.

Bekaert, G., & Hoerova, M. (2014). The VIX, the Variance Premium and Stock Market Volatility. Journal of Econometrics 183(2), 181-192.

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain. Econometrica 80(6): 2369-2429.

Belloni, A., & Chernozhukov, D. (2013). Least Squares after Model Selection in High-dimensional Sparse Models. Bernoulli 19(2): 521-547.

Belloni, A., Chernozhukov, V., Chetverikov, D., & Wei, Y. (2018). Uniformly Valid Post-regularization Confidence Regions for Many Functional Parameters in Z-estimation Framework. The Annals of Statistics 46(6B), 3643-3675.

Belloni, A., Chernozhukov, V., & Wei, Y. (2016). Post-selection Inference for Generalized Linear Models with Many Controls. Journal of Business & Economic Statistics 34(4), 606-619.

Berk, R., Brown, L., Buja, A., Zhang, K., & Zhao, L. (2013). Valid Post-selection Inference. The Annals of Statistics 41(2), 802-837.

Black, F. (1976). Studies in Stock Price Volatility Changes. In Proceedings of the American Statistical Association, Business and Economic Statistics Section 177-181.

Boffelli, S., Skintzi, V. D., & Urga, G. (2017). High- and Low-Frequency Correlations in European Government Bond Spreads and Their Macroeconomic Drivers. Journal of Financial Econometrics 15(1), 62-105.

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. Journal of Econometrics 31, 307-327.

Busch, T., Christensen, B. J., & Nielsen, M. Ø. (2011). The Role of Implied Volatility in Forecasting Future Realized Volatility and Jumps in Foreign Exchange, Stock, and Bond Markets. Journal of Econometrics 160(1), 48-57.

Caporale, T., & McKiernan, B. (1998). The Fischer Black Hypothesis: Some Time-Series Evidence.

Southern Economic Journal 64(3), 765-771.
Christiansen, C., Schmeling, M., & Schrimpf, A. (2012). A Comprehensive Look at Financial Volatility Prediction by Economic Variables. Journal of Applied Econometrics 27(6), 956-977.
Choudhry, T., Papadimitriou, F. I., & Shabi, S. (2016). Stock Market Volatility and Business Cycle: Evidence from Linear and Nonlinear Causality Tests. Journal of Banking & Finance 66: 89-101.
Chung, K. H., & Chuwanganant, C. (2014). Uncertainty, Market Structure, and Liquidity. Journal of Financial Economics 113(3), 476-499.
Conrad, C., and Kleen, O. (2019). Two are Better than One: Volatility Forecasting Using Multiplicative Component GARCH-MIDAS Models. Journal of Applied Econometrics 35(1): 19-45.
Conrad, C., and Loch, K. (2015). Anticipating Long-term Stock Market Volatility. Journal of Applied Econometrics 30(7), 1090-1114.
Conrad, C., Loch, K., & Rittler, D. (2014). On the Macroeconomic Determinants of Long-term Volatilities and Correlations in U.S. Stock and Crude Oil Markets. Journal of Empirical Finance 29, 26-40.
Croushore, D. (2011). Frontiers of Real-Time Data Analysis. Journal of Economic Literature 49(1), 72-100.
Diebold, F.X., & Rudebusch, G.D. (1991). Forecasting Output with the Composite Leading Index: A Real-time Analysis. Journal of the American Statistical Associations 86, 603-610.
Ding, Z., & Granger, C. W. J. (1996). Modeling Volatility Persistence of Speculative Returns: A New Approach. Journal of Econometrics 73(1), 185-215.
Engle, R. F. (1982). Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of UK Inflation. Econometrica 50, 987-1008.
Engle, R. F., & Lee, G. (1999). A Permanent and Transitory Component Model of Stock Return Volatility. In Cointegration, Causality, and Forecasting: A Festschrift in Honor of Clive W. J. Granger 475-497, New York: Oxford University Press.
Engle, R. F., & Rangel, J. G. (2008). The Spline-GARCH Model for Low-Frequency Volatility and Its Global Macroeconomic Causes. The Review of Financial Studies 21(3), 1187-1222.
Engle, R. F., Ghysels, E, & Sohn, B. (2013). Stock Market Volatility and Macroeconomic Fundamentals. The Review of Economics and Statistics 95(3), 776-797.
Fama, E. F., & French, K. R. (1992). The Cross-section of Expected Stock Returns. Journal of Finance 47(2), 427-465.
Fan, Y., and Tang, C. Y. (2012). Tuning Parameter Selection in High Dimensional Penalized Likelihood. Journal of the Royal Statistical Society Series B (Methodological) 75(3), 531-552.
Fornari, F., & Mele, A. (2010). Financial Volatility and Economic Activity. Journal of Financial Management, Markets and Institutions 1(2), 155-198.
Fountas, S., Karanasos, M., & Kim, J. (2006). Inflation Uncertainty, Output Growth Uncertainty and Macroeconomic Performance. Oxford Bulletin of Economics and Statistics 68(3), 319-343.
Ghysels, E., & Qian, H. (2019). Estimating MIDAS Regression via OLS with Polynomial Parameter Profiling. Econometrics and Statistics 9, 1-16.
Ghysels, E., Sinko, A. & Valkanov, R. (2007). MIDAS Regressions: Further Results and New Directions. Econometric Reviews 26(1), 53-90.
Ghysels, E., Santa-Clara, P., & Valkanov, R. (2006). Predicting Volatility: Getting the Most Out of Return Data Sampled at Different Frequencies. Journal of Econometrics 131, 59-95.
Ghysels, E., & Sinko, A. (2011). Volatility Forecasting and Microstructure Noise. Journal of Econometrics

160(1), 257-271.
Giacomini, R., & White, H. (2006). Tests of Conditional Predictive Ability. Econometrica 74(6), 1545-1578.
Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the Relationship between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. Journal of Finance 48(5), 1779-1801.
Huang, D. S. (1973). Short-Run Instability in Single-Family Housing Starts. Journal of the American Statistical Associations 68(344), 788-792.
Jurado, K., Ludvigson, S. C., & Ng, S. (2015). Measuring Uncertainty. American Economic Review 105(3), 1177-1216.
Leeb, H., Pötscher, B. M., & Ewald, K. (2015). On Various Confidence Intervals Post-Model-Selection. Statistical Science 30(2), 216-227.
Ludvigson, S. C., Ma, S., and Ng, S. (2019). Uncertainty and Business Cycles: Exogenous Impulse or Endogenous Response? American Economic Journal: Macroeconomics, forthcoming.
Löffler, G. (2013). Tower Building and Stock Market Returns. Journal of Financial Research 36(3), 413-434.
Manela, A., and Moreira, A. (2017). News Implied Volatility and Disaster Concerns. Journal of Financial Economics 123(1), 137-162.
Nagel, S. (2012). Evaporating Liquidity. Review of Financial Studies 25(7), 2005-2038.
Nelson, D.B. (1991). Conditional Heteroskedasticity in Asset Returns. Econometrica 59, 347-370.
Merton, R. C. (1974). On the Pricing of Corporate Debt: the Risk Structure of Interest Rates. Journal of Finance 29(2), 449-470.
Officer, R. R. (1973). The Variability of the Market Factor of the New York Stock Exchange. The Journal of Business 46(3), 434-453.
Oh, S., & Waldman, M. (1990). The Macroeconomic Effects of False Announcements. Quarterly Journal of Economics 105(4), 1017-1034.
Pan, Z., Wang, Y., Wu, C., & Yin, L. (2017). Oil Price Volatility and Macroeconomic Fundamentals: A Regime Switching GARCH-MIDAS Model. Journal of Empirical Finance 43, 130-142.
Patefield, W. M. (1977). On the Maximized Likelihood Function. Sankhyā: The Indian Journal of Statistics, Series B 39, 92-96.
Patton, A. J. (2011). Volatility Forecast Comparison Using Imperfect Volatility Proxies. Journal of Econometrics 160(1), 246-256.
Robertson, J. C., & Tallman, E. W. (1998). Data Vintages and Measuring Forecast Model Performance. Federal Reserve Bank of Atlanta Economic Review (Fourth Quarter), 4-20.
Schwert, G. W. (1989). Why Does Stock Market Volatility Change over Time? Journal of Finance 44(5), 1115-1153.
Stark, T. (2010). Realistic Evaluation of Real-Time Forecasts in the Survey of Professional Forecasters. Technical Report Research Rap Special Report, Federal Reserve Bank of Philadelphia.
Stark, T., & Croushore, D. (2002). Forecasting with a Real-Time Data Set for Macroeconomists. Journal of Macroeconomics 24, 507-531.
Su, Z., Fang, T., & Yin, L. (2017). The Role of News-based Implied Volatility among US Financial Markets. Economics Letters 157, 24-27.
Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58(1), 267-288.

Wang, H., Li, B., & Leng, C. (2009). Shrinkage Tuning Parameter Selection with a Diverging Number of Parameters. Journal of the Royal Statistical Society. Series B (Methodological) 71(3), 671-683.

Yang, Y. (2005). Can the Strengths of AIC and BIC be Shared? A Conflict between Model Identification and Regression Estimation. Biometrika 92(4), 937-950.

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society Series B (Methodological) 67(2), 301-320.

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

TABLE 1 Estimation procedure

Step 1	Estimate the GARCH-MIDAS model with all J variables by maximizing the log-likelihood function as in Equation (6) under linear constraints, and obtain the parameter estimates $\hat{\theta}_j$ and $\hat{\Phi} = (\hat{\Phi}_1, \hat{\Phi}_2)$. Calculate the adaptive weights as $\hat{w}_j = 1/\left|\hat{\theta}_j\right|^{\eta}$. Set $\Phi_2 = \hat{\Phi}_2$.
Step 2	Estimate the model with variable selection, by maximizing the penalized log-likelihood function as in Equation (7) under linear constraints conditional on $\hat{\Phi}_2$, with the tuning parameter λ on a grid of [0, $\lambda_{\max}$]. Obtain the parameter estimates $\hat{\Phi}_1$, and calculate the GIC for each value of λ.
Step 3	Determine the optimal tuning parameter by GIC, and obtain the selected variables.

TABLE 2 Descriptive statistics

Variable	Obs.	Min.	Max.	Mean.	Std.	Skew.	Kurt.	Database
Stock market data								
S&P 500 returns	12611	-9.94	4.76	0.01	0.46	-1.00	25.45	CRSP
Macroeconomic data								
real GDP	200	-10.37	11.16	2.39	3.00	-0.97	6.42	RDRC
industrial production	200	-29.03	21.16	2.18	6.28	-1.09	7.20	RDRC
unemployment rate	200	-0.97	1.77	0.01	0.37	1.49	7.36	RDRC
housing starts	200	-69.03	236.03	6.93	42.12	1.71	10.03	RDRC
corporate profits	200	-88.01	407.35	11.26	37.29	6.12	65.47	FRED
personal consumption	200	-13.18	10.05	2.89	2.97	-1.32	8.65	RDRC
CPI	200	-8.85	16.74	4.06	3.31	0.87	5.60	FRED
PPI	200	-37.79	31.60	3.89	7.61	-0.28	8.64	FRED
CFNAI	200	-3.41	1.92	-0.02	0.83	-1.50	7.12	FRBC
new orders	200	27.27	71.90	55.10	7.43	-0.85	4.45	Quandl
monetary base	200	-19.93	82.50	8.37	12.34	3.52	20.01	FRED
consumer sentiment	200	-14.70	16.50	0.03	5.15	0.08	3.80	SCUM
real GDP volatility	200	1.76	99.04	10.11	13.21	3.34	17.42	RDRC*
inflation volatility	200	1.39	182.53	13.74	26.69	3.40	15.82	RDRC*
Financial data								
term spread	200	-1.43	3.80	1.68	1.21	-0.47	2.55	RDRC
default spread	200	0.56	3.02	1.08	0.44	1.76	6.87	FRED*
MKT	200	-9.72	7.29	0.49	2.88	-0.68	3.91	French
STR	200	-8.66	7.66	0.46	1.88	-0.10	7.16	French
IV	200	14.16	50.38	23.32	4.92	1.50	7.66	FRED&Manela
RV	200	8.14	1143.40	70.10	114.22	6.80	58.06	CRSP*

Note: This table reports descriptive statistics for daily returns and quarterly macroeconomic & financial variables, including number of observations (Obs.), minimum (Min.), maximum (Max.), mean (Mean.), standard deviation (Std.), Skewness, (Skew.) and Kurtosis (Kurt.). The database with * indicates that the corresponding variable is calculated by the authors based on the data from the corresponding database.

TABLE 3 Post-selection estimation results

GARCH parameters and constant estimates					
μ	α	β	γ	m	
0.0118***	0.0120**	0.8853***	0.1360***	-2.2452***	
(0.0031)	(0.0042)	(0.0077)	(0.0099)	(0.0964)	
Parameter estimates for the long-term component					
θ^{HS}	ω_1^{HS}	ω_2^{HS}	θ^{DS}	ω_1^{DS}	ω_2^{DS}
-0.0138***	2.3277***	4.3622**	0.4613***	-20.6973	7.8284
(0.0021)	(0.6042)	(1.3492)	(0.0689)	(28.7439)	(40.8803)
θ^{RV}	ω_1^{RV}	ω_2^{RV}			
0.0008***	10.1092***	199.9871***			
(0.0003)	(21.5995)	(23.3460)			

Note: This table reports the post-variable-selection estimation results. The long-term volatility component is given by Equation (10). *HS*, *DS*, and *RV* indicate housing starts, default spread and realized volatility. The numbers in parentheses are the robust standard errors, and ***, **, * indicate 1%, 5% and 10% significant levels, respectively.

TABLE 4 Out-of-sample forecast evaluations

Models	1-quarter-ahead MSFE ratio (GW)	2-quarter-ahead MSFE ratio (GW)	3-quarter-ahead MSFE ratio (GW)	4-quarter-ahead MSFE ratio (GW)
Model 1 (Benchmark)	1	1	1	1
Model 2				
real GDP	1.0121(-)***	1.0131(-)**	1.0129(-)**	1.0135(-)*
industrial production	1.0124(-)***	1.0132(-)**	1.0130(-)***	1.0133(-)*
unemployment rate	1.0131(-)***	1.0127(-)**	1.0125(-)***	1.0127(-)*
housing starts	1.0147(-)***	1.0151(-)**	1.0148(-)**	1.0143(-)
corporate profits	1.0140(-)***	1.0143(-)***	1.0141(-)***	1.0142(-)**
personal consumption	1.0082(-)***	1.0088(-)**	1.0086(-)***	1.0090(-)**
CPI	1.0129(-)***	1.0141(-)**	1.0139(-)***	1.0140(-)*
PPI	1.0021(-)***	1.0038(-)***	1.0035(-)***	0.9917(-)***
CFNAI	1.0035(-)***	1.0040(-)***	1.0037(-)***	1.0042(-)***
new order	1.0130(-)***	1.0136(-)**	1.0133(-)**	1.0137(-)*
monetary base	1.0122(-)***	1.0135(-)**	1.0133(-)**	1.0131(-)*
consumer sentiment	1.0091(-)***	1.0094(-)***	1.0091(-)***	1.0064(-)***
real GDP volatility	1.0127(-)***	1.0133(-)*	1.0131(-)**	1.0131(-)
inflation volatility	1.0127(-)***	1.0139(-)*	1.0136(-)**	1.0139(-)
term spread	1.0294(-)***	1.0301(-)***	1.0298(-)***	1.0299(-)***
default spread	1.0092(-)***	1.0095(-)*	1.0092(-)**	1.0099(-)
MKT	1.0156(-)***	1.0159(-)*	1.0156(-)**	1.0148(-)
STR	1.0183(-)***	1.0187(-)	1.0184(-)	1.0188(-)
IV	1.0100(-)***	1.0077(-)***	1.0075(-)***	1.0066(-)***
RV	0.9988(+)***	0.9976(+)***	0.9973(+)***	0.9988(+)***
Model 3	1.0317(-)***	1.0329(-)***	1.0340(-)***	1.0330(-)***
Model 4				
PC1	1.0101(-)***	1.0108(-)***	1.0107(-)***	1.0114(-)**
PC2	1.0141(-)***	1.0137(-)**	1.0142(-)***	1.0152(-)
PC3	1.0119(-)***	1.0124(-)**	1.0122(-)***	1.0124(-)*
Model 5				
real GDP+SPF	1.0101(-)***	1.0110(-)**	1.0110(-)**	1.0114(-)*
industrial+SPF	1.0097(-)***	1.0105(-)**	1.0104(-)***	1.0099(-)*
unemployment+SPF	1.0085(-)***	1.0092(-)**	1.0091(-)***	1.0096(-)*
housing starts+SPF	1.0124(-)***	1.0128(-)**	1.0127(-)**	1.0126(-)*
profits+SPF	1.0115(-)***	1.0122(-)**	1.0123(-)**	1.0121(-)*
CPI+SPF	1.0108(-)***	1.0118(-)**	1.0117(-)**	1.0110(-)*
Model 6				
industrial (level+var)	1.0097(-)***	1.0101(-)**	1.0098(-)***	1.0105(-)*
PPI (level+var)	1.0087(-)***	1.0001(-)*	1.0115(-)*	1.0115(-)*

Note: This table reports the out-of-sample forecast evaluation. The number for each model is MSFE ratio relative to the benchmark model. A positive sign in parentheses indicates the corresponding model outperforms Model 1, and a negative sign indicates the opposite. ***, ** and * denotes the significant level at 1%, 5% and 10% in GW test, respectively.

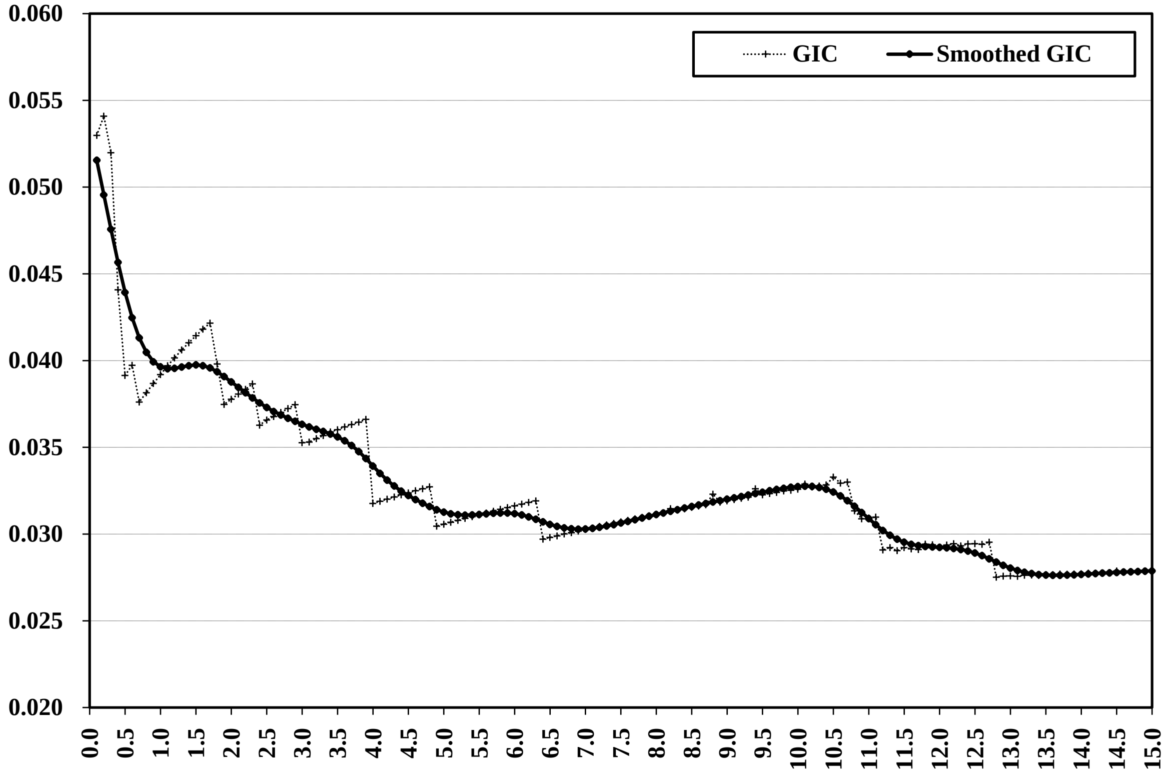

FIGURE 1 Generalized Information Criteria. *Notes*: Figure 1 reports GIC_{λ} as a function of the tuning parameter λ, where λ is considered on a 151-point grid on [0, 15] with an increment of 0.1. The black line denotes smoothed GIC by HP filter.

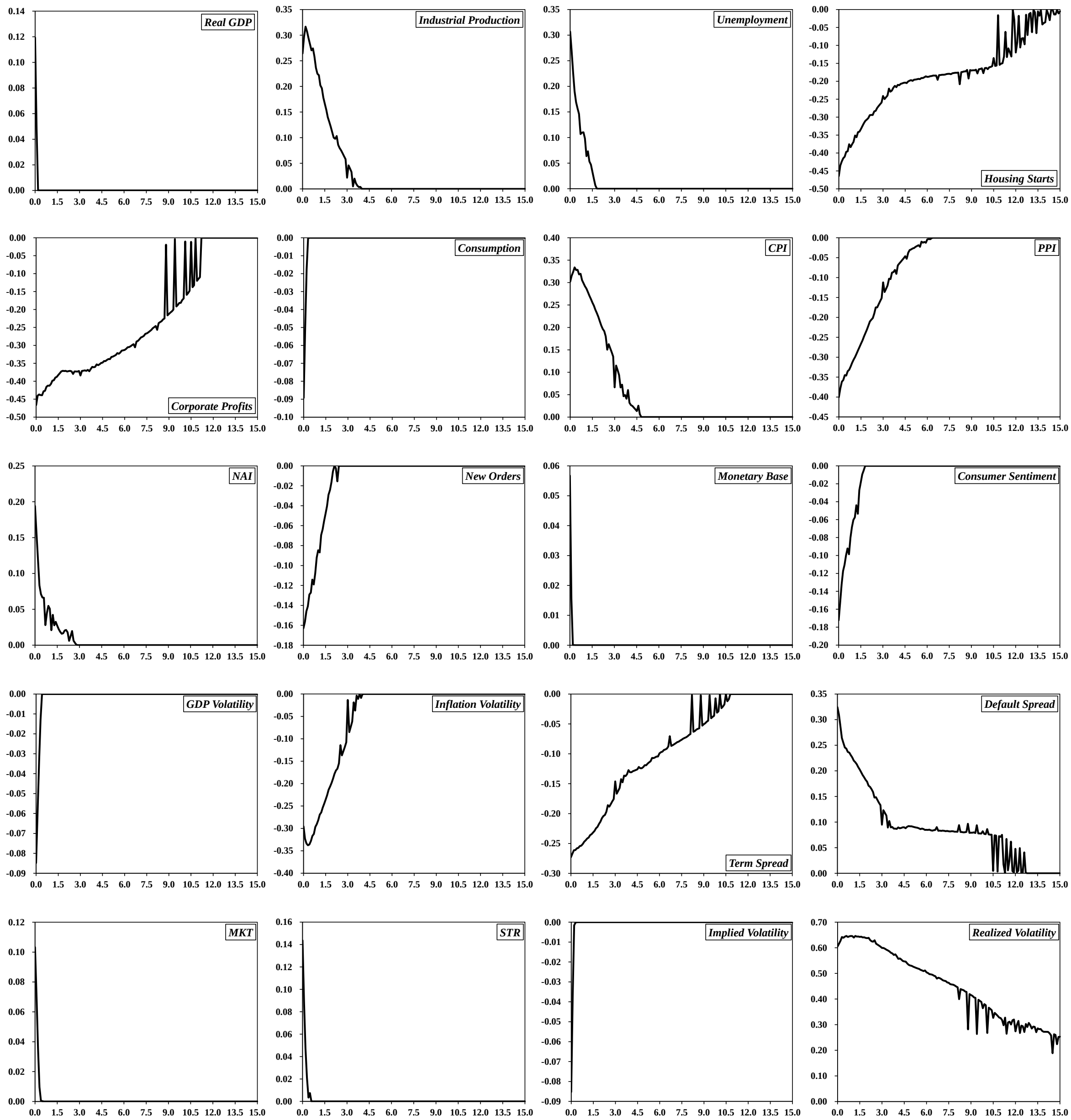

FIGURE 2 The Parameter Estimates $\hat{\theta}_j$ as a Function of Tuning Parameter λ. *Notes*: Figure 2 reports the parameter estimates as a function of the turning parameter λ, where λ is considered on a 151-point grid of [0, 15] with an increment of 0.1.

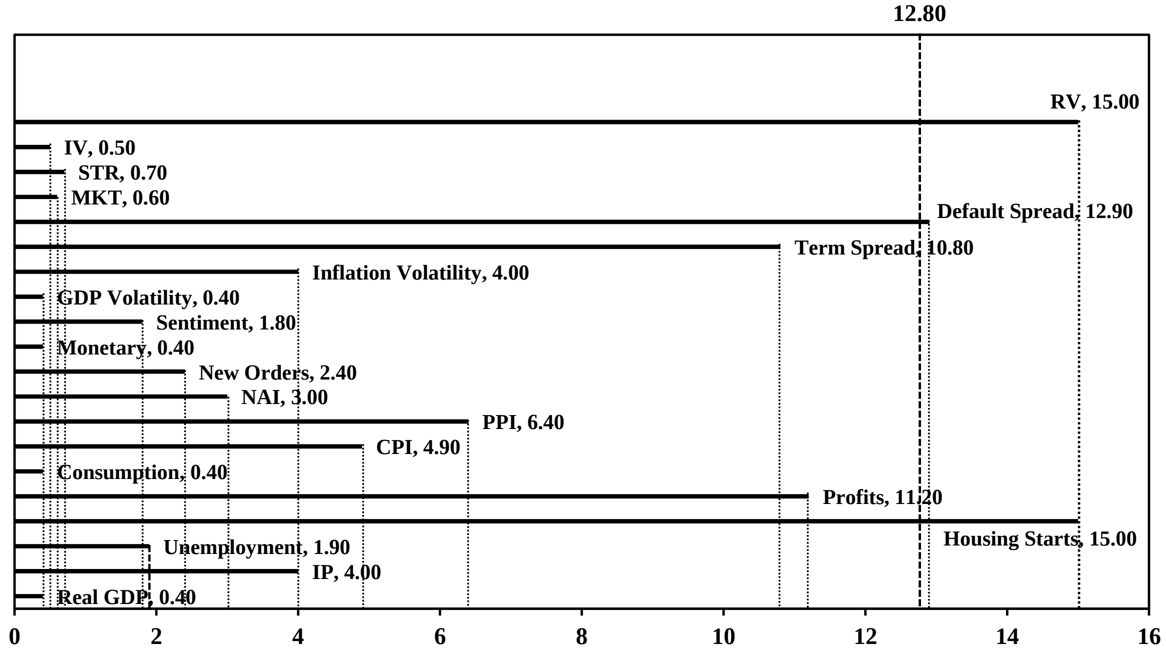

FIGURE 3 Values of λ at which Each Parameter $\hat{\theta}_j$ Reaches Zero. *Notes*: Figure 3 reports how the parameters $\hat{\theta}_j$ are shrunken to zero. The optimal tuning parameter determined by GIC is 12.8.

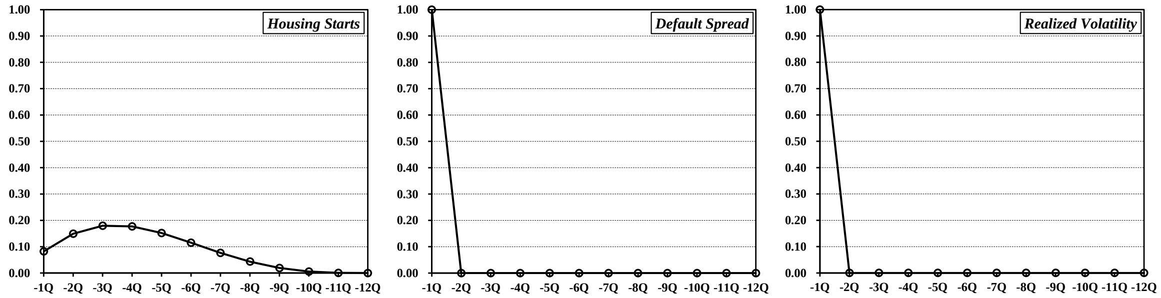

FIGURE 4 Estimated Beta Weights. *Notes*: Figure 4 reports the estimated Beta weights $\varphi_k(\omega_1, \omega_2)$ for the GARCH-MIDAS model with three selected variables.

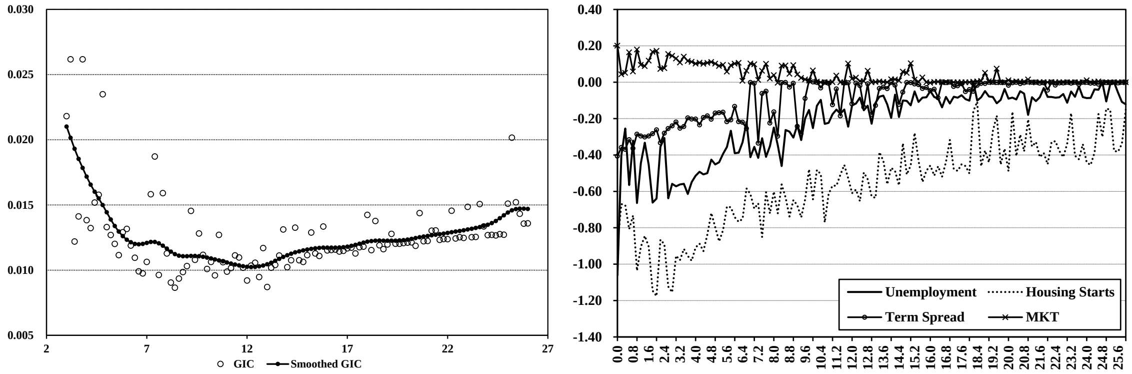

FIGURE 5 Variable Selection Using Revised Data. *Notes*: Figure 5 reports the variable selection results using revised data. The optimal tuning parameter is 12.2, and four variables are selected.