HIGH-DIMENSIONAL SPARSE MULTIVARIATE STOCHASTIC VOLATILITY MODELS

BENJAMIN POIGNARD\textsuperscript{a,b,*} AND MANABU ASAI\textsuperscript{c}

\textsuperscript{a}Graduate School of Economics, Osaka University, Toyonaka, Osaka, Japan
\textsuperscript{b}Riken-AIP, Chuo-ku, Tokyo, 103-0027, Japan
\textsuperscript{c}Faculty of Economics, Soka University, Hachioji, Tokyo, Japan

Although multivariate stochastic volatility models usually produce more accurate forecasts compared with the MGARCH models, their estimation techniques such as Bayesian MCMC typically suffer from the curse of dimensionality. We propose a fast and efficient estimation approach for MSV based on a penalized OLS framework. Specifying the MSV model as a multivariate state-space model, we carry out a two-step penalized procedure. We provide the asymptotic properties of the two-step estimator and the oracle property of the first-step estimator when the number of parameters diverges. The performances of our method are illustrated through simulations and financial data. Supplementary Material presenting technical proofs is available online.

Received 26 August 2021; Revised 28 March 2022; Accepted 29 March 2022

Keywords: Forecasting; multivariate stochastic volatility; penalized M-estimation.

\textbf{JEL.} C13; C32.

\textbf{MOS subject classification:} 62F12; 62P20.

1. INTRODUCTION

Over the past decades, various covariance models have been developed for describing dynamic structures for multivariate economic and financial time series. Within the multivariate GARCH (MGARCH) family, the dynamic conditional correlation (DCC) model of Engle (2002) and Tse and Tsui (2002), the BEKK model of Baba \textit{et al.} (1985) and Engle and Kroner (1995), and their variants are commonly used: see the survey of Bauwens \textit{et al.} (2006), for instance. As for the multivariate stochastic volatility (MSV) family, the MSV model of Harvey \textit{et al.} (1994) was extended, among others, by the factor model of Chib \textit{et al.} (2006) and the dynamic correlation model of Asai and McAleer (2009b); see Ghysels \textit{et al.} (1996), Asai \textit{et al.} (2006), and Chib \textit{et al.} (2009) for various univariate and multivariate SV models. Based on a thorough empirical analysis, Chib \textit{et al.} (2006) highlighted that the MSV models usually outperform MGARCH-based models in terms of out-of-sample forecasts.

Several methods for estimating the MSV models have been developed. In their seminal work, Harvey \textit{et al.} (1994) derived a state-space form based on the vector of the logarithm of squared returns. Based on this state-space setting, they performed a Kalman-based filtering technique to evaluate and optimize the quasi log-likelihood function. In the recent literature, a commonly used method is the Bayesian Markov Chain Monte Carlo (MCMC), as described, for example, in Chib \textit{et al.} (2009) and Kastner \textit{et al.} (2017), among others. An alternative estimation approach is the Monte Carlo likelihood (MCL) method suggested by Durbin and Koopman (1997, 2001) and applied by Asai \textit{et al.} (2015) and Asai and McAleer (2009a). However, empirical applications in the literature are typically limited to low-dimensional random vectors when methods based on MCMC or MCL are
performed, due to the severe costs in terms of computations, or the intricate choice of suitable priors (for the MCMC case), among others. In the same vein, MGARCH specifications also suffer from the so-called “curse of dimensionality” since the complexity is of order $O(p^3)$ in general, where $p$ corresponds to the problem dimension, as the specification of a general multivariate dynamic model often induces an explosion of the number of free parameters. Moreover, tricky conditions are required for the model parameters to satisfy the positive-definiteness of the variance–covariance process.

Another key hurdle of the aforementioned methods is the high nonlinearity of the models, which requires the use of likelihood-based estimation techniques. Therefore, strongly reduced versions of such multivariate models are most often considered as soon as $p$ is larger than four or five. The factor-model-based approach may be a solution to shrink the number of parameters. In particular, Kastner et al. (2017) considered factors in their stochastic volatility framework and provided a joint specification of a large number of covarying time series using a small number of latent factors. However, this factor-based method requires the identification of the corresponding factors together with the treatment of the rotational indeterminacy inherent to factor models.

The objective of this study consists in modeling high-dimensional variance–covariance matrices within the MSV framework in a flexible manner and breaking the curse of dimensionality without relying on standard procedures based on MCMC or MCL. To do so, we introduce a vector autoregressive and moving-average (VARMA) representation for the MSV model in the same spirit as Harvey et al. (1994) and apply an OLS-based two-step estimation approach extending the idea of Hannan and Rissanen (1982) and Hannan and Kavalieris (1984). More precisely, as a first step, we carry out an OLS estimation of a large dimensional VAR model with a sufficiently large number of lags to approximate the VARMA model. For the purpose of parsimony and to avoid over-fitting, we enforce the nullity of potentially numerous model coefficients using a penalization procedure on the model coefficients. Our study shares a similar spirit with Poignard and Fermanian (2021), who provided a framework for high-dimensional variance–covariance within the MGARCH family: they derived some parameterizations to directly generate positive-definite variance–covariance matrices based on multivariate ARCH processes allowing for a linear representation with respect to the parameters. However, our work differs from theirs in two main respects: our analysis lies within the MSV family; we consider a general penalization framework for efficient estimation, which includes a broad range of potentially non-convex penalty functions.

The main contributions of our method are as follows: using a penalized OLS framework, we can directly generate positive-definite variance–covariance matrices without relying on methods like MCMC/MCL and manage high-dimensional matrix processes; the large sample properties of the two-step estimator are provided; in particular, we prove the oracle property of the first-step estimator with a diverging dimension in the sense of Fan and Li (2001), which ensures the correct identification of the underlying set of non-zero coefficients.

The remainder of the article is organized as follows. In Section 2, we describe the framework and the new forecasting procedure based on a penalized OLS estimation framework. Section 3 contains the large sample properties of the penalized two-step OLS estimator. Section 4 reports simulation-based experiment results for in-sample estimates of covariance matrices together with out-of-sample forecasting results based on real financial portfolios. Finally, Section 5 concludes the article. All proofs and intermediary results are in the Supplementary Material.

**Notations:** Throughout this article, we denote the cardinality of a set $E$ by $\text{card}(E)$. For a vector $v \in \mathbb{R}^d$, the $\ell_p$ norm is $\|v\|_p = \left(\sum_{i=1}^{d} |v_i|^p\right)^{1/p}$ for $p > 0$, and $\|v\|_\infty = \max_i |v_i|$. Let the subset be $A \subseteq \{1, \ldots, d\}$; then, $v_A \in \mathbb{R}^{\text{card}(A)}$ is the vector $v$ restricted to $A$. $M_{m \times n}(\mathbb{R})$ denotes the space of $m \times n$ matrices with coefficients in $\mathbb{R}$. For a matrix $A$, $\|A\|_F$ is the Frobenius norm. We write $A^\top$ (resp. $v^\top$) to denote the transpose of the matrix $A$ (resp. the vector $v$). We write vec($A$) to denote the vectorization operator that stacks the columns of $A$ on top of each other into a vector. We denote by vech($A$) the $p(p + 1)/2$ vector that stacks the columns of the lower triangular part of the square and symmetric matrix $A$. $\lambda_{\min}(A)$ (resp. $\lambda_{\max}(A)$) denotes the minimum (resp. maximum) eigenvalue of $A$. We write tr($A$) to denote the trace of the square matrix $A$. The $I_p$ matrix is the $p$-dimensional identity matrix.

For a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, we denote the gradient or subgradient of $f$ by $\nabla f$ and the Hessian of $f$ by $\nabla^2 f$. 
We denote by $(\nabla^2f)_{\mathcal{A}}$ the Hessian of $f$ restricted to the block $\mathcal{A}$. We write $\mathcal{A}^c$ to denote the complement of the set $\mathcal{A}$.

## 2. Penalized OLS Framework for MSV

### 2.1. Framework

We consider a $p$-dimensional vectorial stochastic process $(y_t)_{t=1,...,T}$ and denote the vector of its model parameters by $\theta$. We then consider an MSV decomposition given as

$$
y_t = D \epsilon_t, \quad \epsilon_t \sim iid(0, \Gamma),
$$

(1)

$$
h_{t+1} = \mu + \Phi(h_t - \mu) + \eta_t, \quad \eta_t \sim \mathcal{N}_{\mathbb{R}^p}(0, \Sigma),
$$

(2)

where $\Gamma$ is a $p \times p$ correlation matrix, $\epsilon_t = (\epsilon_{1t}, ..., \epsilon_{pt})^\top$ is a $p \times 1$ random vector, which is i.i.d., centered with variance–covariance $\Gamma$, $h_t = (h_{1t}, ..., h_{pt})^\top$ is a $p \times 1$ vector of log-volatility, $D_t = \text{diag}(\exp(h_{1t}/2), ..., \exp(h_{pt}/2))$ is a diagonal matrix of volatility, $\mu = (\mu_1, ..., \mu_p)^\top$ is a $p \times 1$ vector, $\Phi$ is a $p \times p$ matrix, and $\Sigma$ is a $p \times p$ covariance matrix of $\eta_t$. The MSV model (1) and (2) reduces to the MSV model of Harvey et al. (1994) when $\Phi$ is diagonal and $\epsilon_{it}$ follows a $t$ distribution.

Subsequently, we define $y'_t = (\log(y_{1t}^2), ..., \log(y_{pt}^2))^\top$. Following Harvey et al. (1994), the MSV model can be formulated as a state-space model:

$$
y'_t = c + \alpha_t + \zeta_t,
$$

(3)

$$
\alpha_{t+1} = \Phi \alpha_t + \eta_t,
$$

(4)

where $c = (c_1, ..., c_p)^\top$, $\zeta_t = (\zeta_{1t}, ..., \zeta_{pt})^\top$, and $\alpha_t = h_t - \mu$ with $c_i = \mu_i + \mathbb{E}[(\epsilon_{i1}^2)]$ and $\zeta_{it} = \log(\epsilon_{it}^2) - \mathbb{E}[\log(\epsilon_{it}^2)]$. Assuming a $t$ distribution for $\epsilon_{it}$, Harvey et al. (1994) specified the covariance matrix of $\zeta_t$ as $\Sigma$. Note that $\mathbb{E}[\zeta_t] = 0$ by definition. Based on the state-space form, these authors suggested a quasi-maximum likelihood estimation of the MSV model using the Kalman filter. Alternative methods were proposed such as the Bayesian MCMC technique of Chib et al. (2006) and the Monte Carlo likelihood (MCL) method of Durbin and Koopman (1997, 2001). A significant drawback of these methods is the computational cost and, thus, the curse of dimensionality: most of the applications are restricted to small vector sizes and/or reduced forms are fostered.

In this article, we aim to tackle this issue for the MSV models using a penalized OLS estimation method. Although the MSV model (1) and (2) might be a basic model, the following advantages with respect to the MGARCH models can be highlighted: (i) relatively stable estimates and forecasts for variance–covariance matrices; (ii) simpler restrictions for stationarity conditions; and (iii) no intricate matrix parameterization and/or parameter restrictions to generate positive-definite matrices. Regarding (i), see the theoretical comparison of Taylor (1994, Section 5) and the empirical results of Danielsson (1998) and Ding and Vo (2012), for instance. As for (ii) and (iii), see Bauwens et al. (2006) and Chib et al. (2009) for the MGARCH and MSV models respectively.

As in Harvey et al. (1994) and Kim et al. (1998), we consider the log of squared returns. Harvey et al. (1994) suggested the quasi-maximum likelihood (QML) estimation based on the Kalman filter, by treating the distribution of $\log \chi^2(1)$ as a normal distribution. Ruiz (1994) analyzed the asymptotic properties of the QML estimator for the univariate case. Since their QML estimation depends on the numerical optimization algorithm, Shephard (1993) and So et al. (1997) developed simulated and standard expectation-maximization algorithms respectively. However, the inefficiency of the QML estimator comes from the fact that $\log \chi^2(1)$ is highly right-skewed. To fix this
issue, Kim et al. (1998) approximated $\log \chi^2(1)$ by a mixture of normal distributions to carry out a Bayesian MCMC estimation. Instead, our estimation procedure improves the efficiency using the penalized OLS regression, as previously detailed.

2.2. Our Proposed Approach

We propose a new procedure for estimating high-dimensional stochastic volatility models. Our approach starts from the measurement equation (3). Instead of the state-space form, we derive the VARMA representation of $y_t^\ell$ to apply the ideas of Hannan and Rissanen (1982) and Hannan and Kavalieris (1984) under the framework of a penalized OLS estimation. Our approach consists of four steps and can be summarized as follows:

**Step 1.** Consider a penalized OLS estimation to approximate the error terms in the VARMA representation;

**Step 2.** Using the approximated errors, obtain a regression-based estimator of $(c, \Phi)$;

**Step 3.** Conditional on the VARMA estimators, use an ad hoc estimator for $\Sigma_\xi$ such that the corresponding estimator is positive-definite;

**Step 4.** Obtain the estimator of $\Gamma$.

Let us now detail this four-step procedure. Since $y_t^\ell$ is the sum of a VAR(1) process and an i.i.d. noise by (3), the discussion of Granger and Morris (1976) suggests that $y_t^\ell$ has a VARMA(1,1) representation. By equations (3) and (4), we obtain

$$y_t^\ell = (I - \Phi)c + \Phi y_{t-1}^\ell + \zeta_t + \eta_{t-1} - \Phi \zeta_{t-1},$$

which can alternatively be written as

$$y_t^\ell = (I - \Phi)c + \Phi y_{t-1}^\ell + u_t + \Xi u_{t-1},$$

(5)

with $(u_t)$ being a $p$-dimensional white noise vector with moments $E[u_t] = 0$, $\text{Var}(u_t) = \Sigma_u$, $E[u_t u_s^T] = 0$ for $t \neq s$, where $\Xi$ and $\Sigma_u$ are obtained by matching moments of $w_t = (\zeta_t + \eta_{t-1}) - \Phi \zeta_{t-1}$ and $w_t^* = u_t + \Xi u_{t-1}$. Using $E[w_t w_s^T] = E[w_t^* w_{s-1}^T]$ and $E[w_t w_{t-1}^T] = E[w_t^* w_{t-1}^T]$, the relationship between $(\Xi, \Sigma_u)$ and other parameters is given as follows:

$$\Sigma_u + \Sigma_\zeta + \Phi \Sigma_\zeta \Phi^T = \Sigma_u + \Xi \Sigma_u \Xi^T,$$

(6)

$$-\Phi \Sigma_\zeta = \Xi \Sigma_u.$$

(7)

$\Sigma_u$ and $\Sigma_\zeta$ can be deduced from $\Phi, \Xi, \Sigma_u$ based on equations (6) and (7). Let $(x_t)$ denote the mean-subtracted process $x_t = y_t^\ell - E[y_t^\ell].$ Assuming a stable and invertible model, $(x_t)$ has an AR($\infty$) representation:

$$x_t = \sum_{i=1}^{\infty} \Psi_i x_{t-i} + u_t,$$

(8)

with $u_t$ defined in equation (5). Based on a penalized OLS estimation, we can obtain an approximation of $u_t$ in the first step, denoted as $\hat{u}_{t-1}^{(m)}$. The latter approximation depends on $m$: we empirically need to specify $m$ sufficiently large as a surrogate of $\infty$ in the summation in (8). Thus, for the sake of parsimony and to avoid the over-fitting issue, we assume sparsity among the $\Psi_i$s. In the second step, we calculate the OLS estimator of $(\widehat{\zeta}^*, \widehat{\Phi}, \widehat{\Sigma})$ by regressing $x_t$ on a constant, $x_{t-1}$, and $\hat{u}_{t-1}^{(m)}$. For the third step, we start from the decomposition of the unconditional variance–covariance matrix of $x_t$, which is given by

$$\Sigma_x = \Sigma_u + \Sigma_\zeta,$$

(9)
where $\Sigma_x = \mathbb{E}[x_t x_t^\top]$, $\Sigma_z = \mathbb{E}[z_t z_t^\top]$, and $\Sigma_u = \mathbb{E}[a_t a_t^\top]$ with
\[
\text{vec}(\Sigma_u) = [I_p - (\Phi \otimes \Phi)]^{-1}\text{vec}(\Sigma_y).
\]

Denoting the sample covariance matrix of $x_t$ and $u_t^{(m)}$ by $S_x$ and $S_{g(w)}$ respectively, we obtain an estimator of $\Sigma_z$ as
\[
S_z = -\frac{1}{2} \left[ \Phi^{-1} \hat{\Sigma}_{g(w)} + S_{g(w)} \hat{\Sigma}_{g(w)}^\top \Phi^{-1} \right],
\]
by the sample analogous of the mean of $\Sigma_x$ obtained by equation (7) and its transpose. As there is no guarantee for $S_z$ and $S_z - S_x$ to be positive-definite by the approach, we consider ad hoc estimators for $\Sigma_x$ and $\Sigma_u$ based on decomposition (9). Finally, in the fourth step, we estimate $\Gamma$ by a correlation matrix of $y_t$.

To summarize, our procedure can be broken down as follows:

**Step 1.** We approximate (8) as
\[
x_t = \sum_{i=1}^{m} \Psi_{ix_{t-i}} + u_t^{(m)}, \tag{10}
\]
with $u_t^{(m)} = u_t + \sum_{i>m} \Psi_i X_{t-i}$: under suitable parameter conditions, $\sum_{i>m} \Psi_i X_{t-i}$ is actually negligible when $m$ is large enough. Such conditions can be set in the same vein as Assumptions (15.2.2)-(15.2.4) of Lütkepohl (2006) or Assumption 2.1(b) of Chang et al. (2006) for the VAR models; as Assumption 2 of Chang and Park (2002) for AR models: if we consider a univariate process, based on their latter assumption, $\mathbb{E}[u_t^{(m)} - u_t^{(m)}] = o(m^{-r})$ assuming the existence of the $r$-th moment of $u_t$. Under the sparsity assumption for the VAR($m$) coefficients, we consider the penalized OLS problem
\[
\hat{\Psi}_{1:m} = \arg \min_{\Psi_{1:m}} \left\{ \frac{1}{2T} \sum_{t=1}^{T} \left| \left| x_t - \sum_{i=1}^{m} \Psi_{ix_{t-i}} \right| \right|_2^2 + \text{pen} \left( \frac{\lambda_T}{T} \cdot \text{vec}(\Psi_{1:m}) \right) \right\},
\]
where $\text{pen}(\cdot, \cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$ is a coordinate-separable penalty applied to the coefficients $\Psi_{1:m} = [\Psi_1 \cdots \Psi_m] \in \mathcal{M}_{p \times m}(\mathbb{R})$, $\lambda_T$ is the regularization parameter that depends on the sample size and enforces a particular type of sparse structure in the solution $\hat{\Psi}_{1:m}$. In vector form, $\text{vec}(\Psi_{1:m}) \in \mathbb{R}^d$, $d = mp^2$. In the asymptotic analysis detailed in Section 3.1, the dimension $d$ potentially diverges with the sample size $T$. In particular, this diverging property includes the case "$m$ large and $p$ fixed", which is pertinent when the objective is to suitably approximate $(u_t)$ by $(u_t^{(m)})$.

As the number of parameters $d$ increases with the sample size, we assume that the true parameter value is sparse, which refers to the condition that only $k < d$ elements of the true parameter are non-zero but allows the identities of these elements to be unknown. In other words, the true parameter contains a large number of zero coefficients. Moreover, when $m$ is large, the sparse property is pertinent in the context of time series with autoregressive components. Indeed, the most recent observations are likely to have a higher-level effect on the current ($x_t$) in contrast to older observations. Consequently, it is natural to assume that the parameters in $\Psi_i$ decay with $i$ and become negligible. Since the set of non-zero coefficients is unknown, we rely on the penalty function $\text{pen}(\cdot, \cdot)$ to estimate it. Importantly, the penalty function is non-differentiable at the origin to foster sparsity in the estimator. Furthermore, an additional merit for imposing sparsity is its ability to fix the so-called over-fitting issue: such a problem occurs when too many parameters must
be estimated in light of the sample size, which results in poor out-of-sample performances. Sparsity-based inference methods potentially fix this problem, as emphasized by, for example, Belloni et al. (2013) or Ng (2013).

**Step 2.** Let $\hat{u}^{(m)}_t = x_t - \sum_{j=1}^{\infty} \tilde{\Psi}_r y_{t-j}$. Conditional on $\hat{\Psi}_{1:m}$, we consider the regression

$$y'_t = c^* + \Phi y'_{t-1} + \Xi \hat{u}^{(m)}_{t-1} + \nu_t,$$

where the parameters are $(c^*, \Phi, \Xi)$, and since we replace $u_{t-1}$ by $\hat{u}^{(m)}_{t-1}$, $(\nu_t)$ is the error term for this auxiliary regression. The second-step objective function is

$$\hat{\xi} = \left\{ \frac{1}{2T} \sum_{t=1}^{T} \left| y'_t - \left( c^* + \Phi y'_{t-1} + \Xi \hat{u}^{(m)}_{t-1} \right) \right|^2 \right\},$$

such that we can obtain the estimator of $c$ by $\hat{c} = (I - \hat{\Phi})^{-1} \hat{\xi}$. In this step, the second-step parameter dimension is $p(1 + 2p)$.

**Step 3.** The estimators of $\Sigma_\zeta$ and $\Sigma_a$ are deduced as

$$\hat{\Sigma}_\zeta = rS_\zeta, \quad \hat{\Sigma}_a = (1-r)S_a,$$

where $r$ is a constant satisfying $0 < r < 1$. This ad hoc method aims to treat the positive-definiteness of the estimators and to deal with the high-dimensionality issue of $\hat{\Sigma}_\zeta$. While we consider a naive decomposition based on equation (9) for the former, we set $r = (\pi^2/2)(p^{-1}\text{tr}(S_\zeta))^{-1}$ in (11). Here, $\pi^2/2$ is the value of $\mathbb{E}[\zeta^2]$ when $\zeta$ follows the standard normal distribution. The ad hoc estimators yield $\text{tr}(\hat{\Sigma}_\zeta) = r\text{tr}(S_\zeta) = p\pi^2/2$ and $\text{tr}(\hat{\Sigma}_a) = \text{tr}(S_a) = p\pi^2/2$. Using such approach, we are able to estimate $\text{tr}(\Sigma_\zeta)$ and $\text{tr}(\Sigma_a)$ with accuracy and consistency respectively. More importantly, the computational cost is negligible, compared with alternative estimators (e.g. the GMM type method) that would require a numerical optimization with constraints on the positive-definiteness of $S_\zeta$ and $S_a - S_\zeta$.

**Step 4.** Estimate $\Gamma$ by a correlation matrix of $y_t$.

When the tuning parameter $\lambda_T$ shrinks to zero, Steps 1 and 2 reduce to the standard OLS estimation for low-dimensional VARMA models considered by Hannan and Rissanen (1982) and Hannan and Kavalieris (1984). Step 1 corresponds to a multivariate version of the AR($\infty$) representation of a log-GARCH model. Although Harvey et al. (1994) applied the Kalman filter, its computational cost is non-negligible for large $p$, since the cost evolves according to $O(Tp^3)$ for storing covariance matrices of a $p \times 1$ state vector for all $t = 1, \ldots, T$. For the estimators in **Step 3**, we may improve them by considering moment-matching methods using equations (6), (7), and (9) with restrictions on the positive-definiteness of the estimators of $\Sigma_\zeta$ and $\Sigma_a$. However, we use the above fast and efficient method described in **Step 3** without the need of a numerical optimization procedure. Finally, the fourth step can easily be adapted to a sparse correlation matrix setting, especially when the size $p/T$ is not negligible.

We now introduce our setting for generating the volatility process. For a low-dimensional case, we can calculate the minimum mean square linear estimator (MMSLE) of $\alpha$, based on the full sample $y' = (y'_1^T, \ldots, y'_T^T) \in \mathbb{R}^{pT}$ by the state-space smoothing algorithm. In the high-dimensional case, we consider the multivariate version of Harvey (1998)’s approach with the vector form of (3) as follows:

$$y' = c^* + \alpha + \zeta.$$

---

J. Time Ser. Anal. 44: 4–22 (2023) © 2022 John Wiley & Sons Ltd wileyonlinelibrary.com/journal/jtsa

DOI: 10.1111/jtsa.12647
We now provide the forecasts for variance–covariance based on our proposed method. The MMSLE for the
2.3. Volatility Forecasting
We call our proposed parameterization “penalized OLS-MSV”.

\[ V_x = V_a + V_\zeta, \]

where

\[
V_a = \begin{pmatrix}
\Sigma_x & \Sigma_x \Phi & \Sigma_x \Phi^2 & \ldots & \Sigma_x \Phi^{T-2} & \Sigma_x \Phi^{T-1} \\
\Phi \Sigma_x & \Sigma_x & \Sigma_x \Phi & \ldots & \Sigma_x \Phi^{T-3} & \Sigma_x \Phi^{T-2} \\
\Phi^2 \Sigma_x & \Phi \Sigma_x & \Sigma_x & \ldots & \Sigma_x \Phi^{T-4} & \Sigma_x \Phi^{T-3} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\Phi^{T-2} \Sigma_x & \Phi^{T-3} \Sigma_x & \Phi^{T-4} \Sigma_x & \ldots & \Sigma_x & \Sigma_x \Phi^T \\
\Phi^{T-1} \Sigma_x & \Phi^{T-2} \Sigma_x & \Phi^{T-3} \Sigma_x & \ldots & \Phi \Sigma_x & \Sigma_x
\end{pmatrix},
\]

and \( V_\zeta = I_T \otimes \Sigma_\zeta \). Then, the MMSLE can be calculated as follows:

\[ \tilde{\alpha} = V_a^{-1} (y^- - c^\prime) + c^\prime. \]

As in Harvey (1998), the covariance matrix is deduced from the relationship \( H_t = D_t \Gamma D_t \), such that the sample variance of the standardized variable of \( y_\alpha \) equals to one. We consider the estimator as \( \tilde{H}_t = \tilde{D}_t \Gamma \tilde{D}_t \), where

\[
\tilde{D}_t = \text{diag} (\tilde{a}_{1t}, \ldots, \tilde{a}_{pt}), \quad \tilde{a}_{it} = \bar{d}_i \exp (\bar{x}_u / 2), \quad \bar{d}_i = \sqrt{T^{-1} \sum_{t=1}^{T} \bar{x}_u^2 \exp (-\bar{x}_u)},
\]

for \( i = 1, \ldots, p \). The standardized variables are defined as \( \bar{z}_u = y_u / \bar{d}_u \), which implies \( T^{-1} \sum_{t=1}^{T} \bar{z}_u^2 = 1 \) by definition. We call our proposed parameterization “penalized OLS-MSV”.

2.3. Volatility Forecasting
We now provide the forecasts for variance–covariance based on our proposed method. The MMSLE for the
\( h \)th-step-ahead forecast of \( \alpha_T \) is given by

\[ \hat{\alpha}_{T+h} = R_x V_x (y^- - c^\prime) + c, \]

where \( R_x = [\Phi^{T+h-1} \Sigma_x \Phi^{T+h-2} \Sigma_x \ldots \Phi \Sigma_x] \). The \( h \)th-step-ahead forecast of the covariance matrix is given by

\[ \hat{H}_t = \tilde{D}_t \Gamma \tilde{D}_t, \]

where for \( i = 1, \ldots, p, \)

\[ \tilde{D}_{T+h} = \text{diag} (\tilde{a}_{1,T+h}, \ldots, \tilde{a}_{p,T+h}), \quad \tilde{a}_{i,T+h} = \bar{d}_i \exp (\bar{x}_{i,T+h} / 2). \]

By the structure of \( R_x \) and \( V_x \), the inconsistency in the off-diagonal elements of the third-step estimator may affect the forecasts. We assess its applicability via the Monte Carlo experiments performed in Section 4.

Using the covariance forecasts, we can construct time-varying portfolios for the asset returns, in which the portfolio weights are determined by past information. To assess the relevance of the volatility models in terms of
For the penalty function, we consider the convex penalty LASSO (vector of mean-subtracted series), \( \Psi_{1:m} = (\Psi_1, \ldots, \Psi_m) \in M_{p \times m}(\mathbb{R}) \), and \( Z_{i,m-1} = (x_{i,1}^\top, \ldots, x_{i,m-1}^\top) \in \mathbb{R}^{p(m-1)} \). For every \( T \), we assume there exists a unique pseudo-true parameter value \( \theta_{0,T} \): for every \( T \), the function \( \theta_T \rightarrow E[\ell(y_t, s \leq t; \theta_T)] \) is uniquely minimized on \( \Theta_{0,T} \) at \( \theta_T = \theta_{0,T} \) and the first-order conditions are satisfied, that is, \( E[\nabla_{\theta_T} G_T(y; \theta_{0,T})] = 0 \). In light of the possibly explosive number of parameters for a given \( T, \theta_{0,T} \) is assumed sparse so that the size of the true support \( k_T = \text{card}(A_T) \), with \( A_T := \{ i = 1, \ldots, d_T : \theta_{0,i,T} \neq 0 \} \), also diverges with \( T \). To estimate the latter support, we rely on the penalty function \( \text{pen}(\frac{\lambda}{T}, \cdot) \), which is assumed coordinate-separable, that is, \( \text{pen}(\frac{\lambda}{T}, \theta_T) = \sum_{i=1}^{d_T} p\left( \frac{\lambda}{T}, |\theta_{i,T}| \right) \). Then, the penalized problem becomes

\[
\hat{\theta}_T = \arg \min_{\theta_T \in \Theta_{0,T}} \left\{ G_T(y; \theta_T) + \frac{d_T}{T} \sum_{i=1}^{d_T} p\left( \frac{\lambda}{T}, |\theta_{i,T}| \right) \right\}. \tag{12}
\]

For the penalty function, we consider the convex penalty LASSO \( p(\lambda, |\theta|) = \lambda |\theta| \) of Tibshirani (1996) and the non-convex penalties SCAD and MCP. The SCAD of Fan and Li (2001) is defined as

\[
p(\lambda, |\theta|) = \begin{cases} 
|\theta|, & \text{for } |\theta| \leq \lambda, \\
(\frac{1}{2\alpha(\lambda-1)})(\theta^2 - 2\alpha\lambda|\theta| + \lambda^2), & \text{for } \lambda \leq |\theta| \leq a\lambda, \\
(a+1)\lambda^2/2, & \text{for } |\theta| > a\lambda.
\end{cases}
\]
where $a > 2$. The MCP due to Zhang (2010) is defined for $b > 0$ as
\[
p(\lambda, |\theta|) = \lambda \left[ |\theta| - \frac{\theta^2}{2b\lambda} \right] 1_{[0,\sqrt{b\lambda})} + \frac{b}{2} \frac{\lambda}{|\theta|} 1_{[\sqrt{b\lambda}, \infty)}.
\]
All assumptions we relied on for the large sample analysis are reported in Section C of the Supplementary Material. In particular, the sparsity assumption states that the true parameter vector is sparse, that is, the cardinality of the true sparse support $A_T$ is of size $k_T < d_T$. We assume the stability of the VARMA(1,1) process $(x_t)$ to apply the large sample theory for stationary processes. Finally, we assume suitable regularity conditions for both the non-penalized loss and the penalty function.

We first show the existence of the penalized estimator $\hat{\theta}_T$ for the three aforementioned penalty cases.

**Theorem 1.** Under Assumptions 1–6 given in the Supplementary Material, assume that the penalty function satisfies Assumptions 7(i–iii) in the Supplementary Material for the SCAD and MCP cases and satisfies $\lambda_T = o(T)$ for the LASSO case; then, under the scaling behavior $d_T^2 = o(T)$, there is a local optimum $\hat{\theta}_T$ of (12) satisfying
\[
\|\hat{\theta}_T - \theta_{0,T}\|_2 = O_p \left( \sqrt{d_T (T^{-1/2} + R_T)} \right),
\]
where $R_T = A_{1,T}$ for the SCAD and MCP defined in Assumption 7(ii), and $R_T = \frac{\lambda_T}{T}$ for the LASSO.

**Remark.** For a suitable choice of $\lambda_T$, we would obtain a $\sqrt{T/d_T}$-consistent $\hat{\theta}_T$. A diverging $d_T$ requires the use of an explicit norm: due to norm equivalences, some constants may appear that may depend on the size $d_T$ and, thus, on $T$.

Our second result is dedicated to the oracle property: we show that the penalization procedure in problem (12) asymptotically recovers the true underlying sparse subset $A_T$ and the non-zero estimated coefficients are normally distributed. We prove the oracle property for the SCAD and MCP only: these penalty functions are non-convex, a key property that enables to relax the incoherence/irrepresentable condition and/or avoid the specification of adaptive weights. The incoherence/irrepresentable condition – see inequality (3) of Zou (2006) regarding the irrepresentable condition – is necessary to prove the oracle property for the LASSO: such condition is non-trivial and difficult to empirically verify. Rather than assuming the incoherence/irrepresentable condition, Zou (2006) proposed the adaptive LASSO: stochastic weights are specified in the LASSO penalization to alter the convergence rate of the regularization parameter $\lambda_T$; such weights depend on a first-step $\sqrt{T/d_T}$-consistent estimator, typically a non-penalized OLS estimator: the adaptive LASSO is consequently a two-step procedure. In the same vein, Poignard (2020) specified adaptive weights in the Sparse Group LASSO penalty $\ell_1 + \ell_1/\ell_2$ penalty – since the convexity of the $\ell_1$ and $\ell_1/\ell_2$ norms prevents from satisfying the oracle property. The key advantage of non-convex penalization is the relaxation of the incoherence/irrepresentable condition and avoids a two-step procedure as in the adaptive LASSO.

**Theorem 2.** Under Assumptions 1–8 given in the Supplementary Material, assume $d_T^2 = o(T)$, $\lambda_T = o(T)$; then, the $\sqrt{T/d_T}$-consistent local estimator $\hat{\theta}_T$ of Theorem 1 satisfies
\[
\lim_{T \to \infty} P(\hat{A}_T = A_T) = 1, \quad \text{and} \quad \sqrt{T Q_T} \mathbb{V}^{-1/2} \left( \hat{\theta}_T - \theta_{0,T} \right)_{A_T} \xrightarrow{d} T \to \infty \mathbb{N}_{|A_T^c|} (0, \mathbb{C}),
\]
where $\mathbb{V}_{A_T} := (H^{-1} M H^{-1})_{A_T^c A_T}$, $H := E[\theta_T^2 \theta_T^T \ell(y_T, s \leq t; \theta_0, T)]_{1 \leq k \leq d_T}$ and $M := E[\theta_T \ell(y_T, s \leq t; \theta_0, T)]_{1 \leq k \leq d_T}$ and $\mathbb{V}_{\theta_T} := E[\theta_T \ell(y_T, s \leq t; \theta_0, T)]_{1 \leq k \leq d_T}$ and $\mathbb{V}_{\theta_T^2} := E[\theta_T^2 \ell(y_T, s \leq t; \theta_0, T)]_{1 \leq k \leq d_T}$.
However, our setting does not enable to simultaneously distinguish first-step estimator is through may derive a finite sample and explicit upper bound for the approximation error \( 1 \) identical to condition (15.2.5) of Proposition 15.1. (result of Lewis and Reinsel, 1985) of Lütkepohl (2006).

**Remark.** This result deserves a few comments:

(i) The scaling behavior \( d_T \) is given as \( d_T^2 = o(T) \). This is because the third-order term in the Taylor expansion vanishes for the least squares loss. If we consider a nonlinear-based non-penalized loss, this rate would become \( d_T^2 = o(T) \), as in Fan and Peng (2004) or Poignard (2020).

(ii) The cardinality of the true support \( A_T \) denoted by \( k_T \) also diverges with the sample size. Thus, the dimension of \( (\hat{\theta}_T - \theta_0)^A \) is diverging. This motivates the introduction of the matrix \( Q_T \) to obtain a finite dimensional Gaussian distribution.

The first-step estimator is deduced from the truncated VAR process (10): a VAR(\( m \)) is fitted to obtain \( (\hat{\psi}_m)^T \) as an approximation of \( (\psi) \). In this context, the specification of a diverging number of parameters is relevant to correctly approximate \( (\psi) \). When \( p \) is fixed and \( m := m_T \to \infty \), our scaling condition \( d_T^2 = o(T) \) for the oracle property is identical to condition (15.2.5) of Proposition 15.1. (result of Lewis and Reinsel, 1985) of Lütkepohl (2006). However, our setting does not enable to simultaneously distinguish \( m_T \to \infty \) and \( p := p_T \to \infty \). Furthermore, we may derive a finite sample and explicit upper bound for the approximation error \( \frac{1}{T} \sum_{t=1}^{T} \| \hat{\psi}_m(t) - (\psi_T + \sum_{s=1}^{T} \psi_s) \|_2^2 \) in the same spirit as in Proposition 2.4. of Wilms et al. (2021), who considered an approximation for VARMA(\( p,q \)) processes and relied on a LASSO penalization. The derivation of such bound would require additional assumptions on the non-convex penalty functions – such as the \( \mu \)-amenable assumption as in Loh and Wainwright (2017) – and the derivation of an exponential bound over \( \mathbb{V}_{\psi}(\mathbb{G}_T(\cdot; \hat{\theta}_T)) \). We leave this topic for future research.

### 3.2. Second-Step Estimator \( (\hat{\mathcal{c}}^*, \hat{\Phi}, \hat{\Xi}) \)

We consider the large sample properties of the second-step estimator \( \hat{\gamma} = (\hat{\mathcal{c}}^*, \text{vec}(\hat{\Phi})^T, \text{vec}(\hat{\Xi})^T)^T \), which is of size \( d_T = p(1 + 2p) \) assumed fixed. Conditional on \( \hat{\theta}_T \), we consider a second-step loss function \( L_T \) from \( \mathbb{R}^{pT} \times \Theta_2 \) to \( \mathbb{R} \) with \( \Theta_2 \subset \mathbb{R}^{(1+2p)} \) compact, and \( L_T(y; \hat{\theta}_T, \gamma) \) is the empirical loss associated to a continuous function \( f : \mathbb{R}^{pT} \times \Theta_1 \times \Theta_2 \to \mathbb{R} \), that is,

\[
L_T(y; \hat{\theta}_T, \gamma) = \frac{1}{T} \sum_{t=1}^{T} \| y_t - \mathcal{c}^* + \Phi y_{t-1}^e = \Xi \hat{\psi}_m(t) - g_T(\hat{\theta}_T) \|_2^2 = \frac{1}{T} \sum_{t=1}^{T} \| y_t - \mathcal{c}^* + \Phi y_{t-1}^e = \Xi \hat{\psi}_m(t) - g_T(\hat{\theta}_T) \|_2^2 = \frac{1}{T} \sum_{t=1}^{T} f(y_t, s \leq t; \hat{\theta}_T, \gamma),
\]

where \( K_{t-1}(\hat{\theta}_T) = (1, y_{t-1}, \hat{\psi}_m(t))^T \in \mathbb{R}^{1+2p} \) and \( \Gamma = (\mathcal{c}^*, \Phi, \Xi) \in \mathcal{M}_{p \times (1+2p)}(\mathbb{R}) \). The dependence with respect to the first-step estimator is through \( \mathcal{M}_T \). The problem of interest is

\[
\hat{\gamma} = \arg \min_{\gamma \in \Theta_2} \left\{ L_T(y; \hat{\theta}_T, \gamma) \right\}.
\]

The function \( y \to \mathbb{E}[f(y_t, s \leq t; \hat{\theta}_T, \gamma)] \) is assumed to be uniquely minimized at \( \gamma = \gamma_0 \); the true parameter vector \( \gamma_0 = (\mathcal{c}_0^*, \text{vec}(\Phi_0)^T, \text{vec}(\Xi_0)^T)^T \).

**Theorem 3.** Under the conditions of Theorem 1, under Assumptions 9–12 in the Supplementary Material, the sequence of second-step estimators \( \hat{\gamma} \) satisfies

\[
\| \hat{\gamma} - \gamma_0 \| = O_p \left( \frac{1}{\sqrt{T}} \right).
\]

[Correction added on 8 June, after first online publication: Theorems 3 and 4 were corrected in this version]
Remark. To derive this explicit convergence rate, the moment conditions in Assumptions 10 and 12 are key to control for the first-step estimator when the number of parameters is diverging. The framework can potentially be extended to a diverging $d_2 := d_{2,T}$, at the expense of more complicated assumptions. To keep our asymptotic arguments simple, we assumed $d_2$ to be fixed.

We now derive the asymptotic distribution of the second-step estimator $\hat{\gamma}$ conditional on $\hat{\theta}_T$ whose elements belong to $A_T$.

Theorem 4. Under the assumptions of Theorem 2, assume $d_2^2 = o(T)$ and $\hat{\theta}_T$ is the oracle estimator, under the conditions of Theorem 3 and under Assumptions 13–16 in the Supplementary Material, then

\[
\sqrt{T} (\hat{\gamma} - \gamma_0) \xrightarrow{d} N_{\mathbb{R}^{d_2}} (0, \Sigma),
\]

with $\Sigma_T = \mathbb{U}^{-1} \mathbb{V}_{A_T} \mathbb{V}_{A_T}^T \mathbb{U}^{-1} + \mathbb{U}^{-1} \mathbb{W} \mathbb{U}^{-1} - \mathbb{U}^{-1} \mathbb{V}_{A_T} \mathbb{H}^{-1}_{A_T} \mathbb{W}^{-1} \mathbb{H}^{-1}_{A_T} \mathbb{V}_{A_T}^T \mathbb{U}^{-1}$, where $\mathbb{V}_{A_T}$ and $\mathbb{H}_{A_T} \mathbb{V}_{A_T}$ is defined in Theorem 2, $\mathbb{W} := \mathbb{E} \{ (K_{\gamma_{t-1}}(\theta_{0,T}) \otimes (\gamma_t^e - \Gamma_0 K_{\gamma_{t-1}}(\theta_{0,T})) ) \} \sum_{\mathbb{Y}_{A_T}}$, $\mathbb{U} := \mathbb{E} \{ (I_p \otimes K_{\gamma_{t-1}}(\theta_{0,T}) K_{\gamma_{t-1}}(\theta_{0,T})^T ) \}$, and $\mathbb{V}_{A_T} := \mathbb{E} [ \hat{\gamma}_T^2 f(y,r,s \leq T; \theta_{0,T}, \gamma_0) ]_{\mathbb{I} \leq d_2, i \neq j}$ and $\mathbb{J}_{A_T} \mathbb{E} [ \hat{\gamma}_T^2 f(y,r,s \leq T; \theta_{0,T}, \gamma_0) ]_{\mathbb{I} \leq d_2, i \neq j}$.

Remark. The following comments can be emphasized:

(i) The effect of the first-step estimator is explicitly provided. It affects the variance of the second-step estimator through $\mathbb{U}^{-1} \mathbb{V}_{A_T} \mathbb{V}_{A_T}^T \mathbb{U}^{-1}$ and $\mathbb{U}^{-1} \mathbb{V}_{A_T} \mathbb{H}^{-1}_{A_T} \mathbb{V}_{A_T}^T \mathbb{U}^{-1}$.

(ii) The key difficulty is to establish that $\nabla_{\theta_0^T} (\gamma_0, \hat{\theta}_T)$ converges in probability to some deterministic counterpart while controlling for the diverging dimension of $\hat{\theta}_T$, justifying the moment conditions of Assumption 16.

The third-step estimator is accurate in the sense that $p^{-1} \text{tr}(\hat{\Sigma}_T)$ always takes the true value $\mathbb{E} [ p^{-1} \text{tr}(\Sigma_T) ] = \pi^2 / 2$ under the Gaussian assumption. To improve the estimator, we can consider the structure $\Sigma_T = \sigma^2_T \hat{\Sigma}$, where $P_T$ is a correlation matrix and $\sigma^2_T$ is the variance for non-Gaussian assumption. Neglecting the computational costs, we may estimate positive-definite $\Sigma_T$ and $\Sigma_0$ given $\hat{\Phi}$ and $\hat{\Sigma}$ under the restrictions discussed below equation (9).

4. EMPIRICAL ANALYSIS

4.1. Simulation Experiment

We empirically investigate the ability of the proposed penalization method to better capture complex variance–covariance processes. We simulate the $p$-dimensional stochastic process $(\varepsilon_t)$ based on two data generating processes (DGP): the multivariate ARCH and the BEKK processes. For the multivariate ARCH with $q^*$ lags – M-ARCH($q^*$) in the rest of the article – case, we consider the DGP

\[
\begin{align*}
\varepsilon_t &= H_t^{1/2} \eta_t, \\
H_t &= \Omega + \sum_{k=1}^{q^*} (I_p \otimes \varepsilon_{t-k}) A_k (I_p \otimes \varepsilon_{t-k}),
\end{align*}
\]

where $q^*$ is the number of lagged matrices being functions of $\varepsilon_{t-k}$ and the $p^2 \times p^2$ square matrices $A_k$ satisfy the stationarity conditions of Theorem 2 of Boussama (2006) together with the positivity condition given by Gouriéroux (1997). We generate the diagonal elements of $A_k$ from a uniform distribution $U([0,0.01,0.05])$ and the off-diagonal ones from $U([-0.01,0.01])$ under the ordering constraint $\forall k \geq 2, \forall i,j, |A_{i,j}| \leq |A_{k-1,i,j}|$. Note that these coefficients are more constrained (i.e. closer to zero) when the dimension $p$ increases. As for the matrix $\Omega$,
the diagonal and off-diagonal elements are simulated from $U((0.1, 0.2))$ and $U((-0.01, 0.01))$ respectively. As for the BEKK process, the DGP is based on

$$
\begin{align*}
& e_t = H_t^{1/2} \zeta_t, \\
& H_t = \Omega + AE_{t-1}e_{t-1}^T + BH_{t-1}B^T,
\end{align*}
$$

where $A, B$ are $p \times p$ matrices, satisfying the stationarity constraint $\|D_h^p((A \otimes A) + (B \otimes B))D_h\|_F < 1$, and $D_h^p$ and $D_h^q$ are the duplication matrix and elimination matrix respectively (see Section 11.3 “Stationarity of VEC and BEKK Models” of Francq and Zakoian (2010) for the stationarity condition and Remark 11.1 for the definition of the latter matrices). The entries of $A$ and $B$ are generated from the uniform distribution $U((-0.8, 0.8))$. The matrix $\Omega$ is generated as in the M-ARCH$(q')$ case. Unlike the M-ARCH$(q')$ case, the BEKK dynamic includes an autoregressive component through $B$, which motivated the use of larger lags when estimating our proposed parameters. In both proposed dynamics, we initialize the observations $(\zeta_1, \ldots, \zeta_t)$ with centered and unit variance multivariate Gaussian distribution, where $k = q'$ in the M-ARCH model and $k = 1$ in the BEKK model. Further, conditional on the past $k$ observations, we generate $H_t$ and, thus, $e_t$ according to a centered multivariate Gaussian distribution with variance–covariance $H_t$.

We consider the problem sizes, $p = 15, 50, 100$, and $T = 800$ observations for each of them. For the M-ARCH$(q')$-based data generating process, we considered $q' = 2$ when $p = 15$ and $q' = 1$ when $p = 50, 100$. Subsequently, we propose to compare the true variance–covariance processes – BEKK and M-ARCH$(q')$ – and the estimated ones through our proposed MSV model and the scalar DCC together with the constant correlation model (CCC). The estimation of the DCC model is based on the classic two-step Gaussian QMLE, where the marginal conditional volatility processes are specified as GARCH(1,1) and a correlation targeting procedure is applied in the second step, providing an estimated trajectory $\hat{\theta}_t^{dec}$. The CCC is estimated thanks to a joint estimation of the GARCH(1,1) parameters and correlation parameters through a Gaussian QML, which provides an estimated process $\hat{\theta}_t^{ccc}$. More details on the DCC and CCC can be found in Appendix D.

Regarding our proposed variance–covariance dynamic, the penalized OLS-MSV, denoted as $\hat{\theta}_t^{ols,scad}$ for the SCAD OLS-MSV, $\hat{\theta}_t^{ols,mcp}$ for the MCP OLS-MSV, and the non-penalized version of the OLS-MSV denoted as $\hat{\theta}_t^{ols}$, we considered two settings depending on the DGP. In the M-ARCH$(q')$ case, we set the number of lags in Step 1 in (10) as $m = 10$ when $p = 15$ and set it as $m = 5$ for a dimension $p = 50, 100$. In the BEKK case, due to the autoregressive nature of the latter dynamic, more lags were specified: we selected $m = 30$ (resp. $m = 15$, resp. $m = 5$) when $p = 15$ (resp. $p = 50$, resp. $p = 100$). For both DGPs, the correlation matrix of Step 4 is estimated as the sample correlation matrix estimator. In the SCAD and MCP cases, the coefficients $a$ and $b$ are set as $3.5 – a$ value close to the optimal one as in Fan and Li (2001) – and $3$ respectively.

We compare the true variance–covariance and the estimated variance–covariance processes through the aforementioned models. To do so, we specify a matrix distance, namely, the Frobenius norm, defined as $\|A - B\|_F := \sqrt{\text{tr}((A - B)^T(A - B))}$. We compute the previous norm for each $t$ and for $A = H_t$ and $B \in \{\hat{\theta}_t^{dec}, \hat{\theta}_t^{ccc}, \hat{\theta}_t^{ols}, \hat{\theta}_t^{ols,scad}, \hat{\theta}_t^{ols,mcp}\}$. We take the average of those quantities over $T = 800$ periods of time. Since we repeat this experiment 100 times, this provides an average gap for all those simulations.

By a cross-validation (CV) procedure – see, for example, Hastie et al. (2015, Chap. 2) – we selected the regularization parameter and emphasize that the standard CV developed for i.i.d. data cannot be used in our time series framework. To fix this issue, we used the hv-CV procedure devised by Racine (2000), which consists in leaving a gap between the test sample and the training sample, on both sides of the test sample.

The average difference results are reported in Table I for the M-ARCH$(q')$-based DGP and Table II for the BEKK-based DGP. First, our proposed method provides better in-sample results in terms of accuracy compared with standard MGARCH models. The results are closer to each other in the BEKK-based DGP case, essentially due to the presence of an autoregressive term, which is a priori in favor of the DCC/CCC model. Interestingly, our results emphasize the gain in considering a penalized MSV, especially when the dimension increases.
Table I. Average distance true/estimated covariance matrices – M-ARCH\((q^*)\) (100 replications)

|   | \(\hat{H}_{dcc}\) | \(\hat{H}_{ccc}\) | \(\hat{H}_{ols}\) | \(\hat{H}_{ols,scad}\) | \(\hat{H}_{ols,mcp}\) |
|---|-----------------|-----------------|-----------------|-----------------|-----------------|
| \(p = 15\) | 6.37            | 6.75            | 5.92            | 5.81            | 5.80            |
| \(p = 50\)  | 17.10           | 18.38           | 16.80           | 15.87           | 15.85           |
| \(p = 100\) | 64.05           | 70.51           | 62.09           | 59.84           | 59.66           |

Table II. Average distance true/estimated covariance matrices – BEKK (100 replications)

|   | \(\hat{H}_{dcc}\) | \(\hat{H}_{ccc}\) | \(\hat{H}_{ols}\) | \(\hat{H}_{ols,scad}\) | \(\hat{H}_{ols,mcp}\) |
|---|-----------------|-----------------|-----------------|-----------------|-----------------|
| \(p = 15\) | 21.75           | 22.02           | 21.60           | 20.57           | 21.18           |
| \(p = 50\)  | 114.40          | 115.63          | 115.78          | 111.16          | 113.14          |
| \(p = 100\) | 292.28          | 295.99          | 291.73          | 285.88          | 287.97          |

4.2. Application to Real Data

To assess the relevance of the proposed penalized method, we propose a real data experiment, where we focus on direct out-of-sample evaluation methods, which allow for pair-wise comparisons. They test whether some of the variance–covariance models provide better forecasts in terms of portfolio volatility behavior. Following the methodology of Engle and Colacito (2006), we develop a mean-variance portfolio approach to test the \(H_t\) forecasts.

Intuitively, if a conditional covariance process is misspecified, the minimum-variance portfolio should emphasize such a shortcoming, compared with other models. Here, consider an investor who allocates a fixed amount between \(p\) stocks, according to a minimum-variance strategy and independently at each time \(t\):

\[
\min_{w_t} w_t^\top H_t w_t, \quad \text{s.t. } i^\top w_t = 1, \tag{13}
\]

where \(w_t\) is the \(p \times 1\) vector of portfolio weights chosen at (the end of) time \(t - 1\), \(i\) is a \(p \times 1\) vector of 1, and \(H_t\) is the estimated conditional covariance matrix of the asset returns at time \(t\). The solution of (13) is given by the global minimum-variance portfolio

\[
\hat{w}_t = H_t^{-1} i / i^\top H_t^{-1} i.
\]

Engle and Colacito (2006) showed that the realized portfolio volatility is the smallest when the variance–covariance matrices are correctly specified. Consequently, if wealth is allocated using two different dynamic models \(i\) and \(j\), whose predicted covariance matrices are \((H_t^i)\) and \((H_t^j)\), the strategy providing the smallest portfolio variance will be considered as the best. To do so, we consider a sequence of minimum-variance portfolio weights \((w_{i,t})\) and \((w_{j,t})\), depending on the model. Further, we consider a distance based on the difference of the squared returns of the two portfolios, defined as

\[
u_{ij,t} = \left\{ w_{i,t}^\top e_t \right\}^2 - \left\{ w_{j,t}^\top e_t \right\}^2.
\]

The portfolio variances are the same if the predicted covariance matrices are the same. Thus, we test the null hypothesis \(H_0: \mathbb{E}\left[ u_{ij} \right] = 0\) by the Diebold and Mariano (1995) test. It consists of a least squares regression using HAC standard errors, given by \(u_{ij,t} = \alpha + \epsilon_{u,t}, \mathbb{E}[\epsilon_{u,t}] = 0\), and we test \(H_0: \alpha = 0\). If the mean of \(u_{ij,t}\) is significantly positive (resp. negative), the forecasts given by the covariance matrices of model \(j\) (resp. \(i\)) are preferred. Following Engle and Colacito (2006), we compute the test statistic

\[
\widehat{DM}_{ij} = \frac{\sqrt{\hat{h}} \bar{u}_{ij}}{\sqrt{\text{Var}(\sqrt{\hat{h}} \bar{u}_{ij})}}, \quad \bar{u}_{ij} = \frac{1}{h} \sum_{t=1}^{h} u_{ij,t},
\]
with \( h \) being the number of one-period ahead forecasts and \( \sqrt{\text{Var}(\sqrt{h} \ u_{ij})} \) is a heteroscedasticity and autocorrelation consistent estimator of the asymptotic variance of \( \sqrt{h} \ u_{ij} \). In particular, \( \bigcup M_j \overset{d}{\longrightarrow} \mathcal{N}(0, 1) \) under \( H_0 \).

We run the latter test to compare the scalar DCC (DCC), the orthogonal GARCH (O-G), the BEKK (BEKK), and our OLS-MSV method (MSV) together with its penalized counterpart (denoted as MSV-SCA and MSV-MCP for the SCAD and MCP respectively). We also consider the adaptive LASSO, denoted as MSV-AL, as an additional

Our OLS-MSV method (MSV) together with its penalized counterpart (denoted as MSV-SCA and MSV-MCP for the SCAD and MCP respectively). We also consider the adaptive LASSO, denoted as MSV-AL, as an additional

The results based on the Diebold–Mariano test are limited since they are pair-wise comparisons. It is not possible to ensure that an optimal test is clearly identified. To tackle this issue, Hansen et al. (2003, 2011) proposed the model confidence set (MCS) method, which is a testing framework for the null hypothesis of equivalence across subsets of models. Starting with a full set of candidate models, the MCS method sequentially trims the elements of this set, thus reducing the number of viable models. To be more precise, this approach performs an iterative selection procedure testing the null hypothesis of equal forecasting ability among all models included in a set \( \mathcal{M} \) (the starting set containing all candidate models) for a given loss function. The null hypothesis is

\[ H_0 : \ E[u_{ij,t}^2] = 0, \quad i > j \quad \text{for any} \quad i, j \in \mathcal{M}. \]

To test \( H_0 \), Hansen et al. (2003) proposed the following two statistics:

\[
t_R = \max_{i,j \in \mathcal{M}} \left| \frac{\bar{u}_{ij}}{\sqrt{\text{Var}(\bar{u}_{ij})}} \right|, \quad \text{and} \quad t_{SQ} = \sum_{i,j \in \mathcal{M}, i > j} \left( \frac{\bar{u}_{ij}}{\sqrt{\text{Var}(\bar{u}_{ij})}} \right)^2, \quad \bar{u}_{ij} = \frac{1}{h} \sum_{t=1}^{h} u_{ij,t},
\]
Table III. This table reports the out-of-sample t-statistics of the Diebold–Mariano test for the MSCI (a) and S&P 100 (b) portfolios that checks the equality between covariance matrix forecasts using the loss function $t_{ij}$ over the period December 2015–March 2018 and February 2018–January 2020 respectively. This loss function is defined as the difference between squared realized returns of alternative multivariate variance–covariance models. When the null hypothesis of equal predictive accuracy is rejected, a positive number is evidence in favor of the model in the column.

|                | DCC        | O-G         | BEKK        | MSV-AL10 | MSV-SCA10 | MSV-MCP10 | MSV-AL20 | MSV-SCA20 | MSV-MCP20 |
|----------------|------------|-------------|-------------|----------|-----------|-----------|----------|-----------|-----------|
| (a) MSCI portfolio |            |             |             |          |           |           |          |           |           |
| DCC            | –2.259c    | –4.302c     | 7.408c      | 7.541c   | 7.188c    | 7.441c    | 7.215c   | 7.496c    | 7.466c    | 7.399c    |
| O-G            | 2.259c     | –4.852c     | 6.207c      | 6.241c   | 6.056c    | 6.228c    | 6.196c   | 6.279c    | 6.298c    | 6.270c    |
| BEKK           | 4.302c     | 4.852c      | 6.249c      | 6.281c   | 6.204c    | 6.264c    | 6.256c   | 6.292c    | 6.323c    | 6.296c    |
| MSV-AL10       | –7.408c    | –6.207c     | –6.249c     | 0.667c   | –1.535c   | 1.257c    | 0.542c   | 0.641c    | 0.983c    | 1.276c    |
| MSV-SCA10      | –7.541c    | –6.241c     | –6.281c     | –0.667 –1.843c | –0.299c   | –0.096c   | –0.059c  | 0.237c    | 0.304c    |
| MSV-MCP10      | –7.441c    | –6.228c     | –6.264c     | –1.257 0.096c | –1.299c   | –0.157c   | 0.076c   | 0.547c    | 1.161c    |
| MSV-AL20       | –7.215c    | –6.196c     | –6.256c     | –0.542c | 0.096c   –1.299c   | –0.157c   | 0.076c   | 0.547c    | 1.161c    |
| MSV-SCA20      | –7.496c    | –6.279c     | –6.292c     | –0.641 0.059c | –1.521c   | –0.227c   | –0.076c  | 0.609c    | 0.826c    |
| MSV-MCP20      | –7.399c    | –6.270c     | –6.296c     | –1.276 –0.304c | –1.902c   | –0.839c   | –1.161c  | –0.826c   | –0.101c   |

|                | DCC        | O-G         | BEKK        | MSV-AL10 | MSV-SCA10 | MSV-MCP10 | MSV-AL20 | MSV-SCA20 | MSV-MCP20 |
|----------------|------------|-------------|-------------|----------|-----------|-----------|----------|-----------|-----------|
| (b) S&P 100 portfolio |            |             |             |          |           |           |          |           |           |
| DCC            | –4.785c    | –4.698c     | 3.694c      | 5.752c   | 3.652c    | 3.645c    | 2.791c   | 5.056c    | 4.587c    | 2.792c    |
| O-G            | 4.785c     | –0.732c     | 5.646c      | 8.562c   | 5.605c    | 5.609c    | 4.996c   | 6.496c    | 6.202c    | 4.995c    |
| BEKK           | 4.698c     | –0.732c     | 5.754c      | 8.781c   | 5.713c    | 5.716c    | 5.090c   | 6.651c    | 6.318c    | 5.088c    |
| MSV-AL10       | –3.694c    | –5.646c     | –5.754c     | 5.902c   | –1.117c   | –1.732c   | –7.651c  | 5.701c    | 5.235c    | 1.182c    |
| MSV-SCA10      | –7.572c    | –8.562c     | –8.781c     | –5.902c  | –5.936c   | –5.928c   | –6.870c  | –4.704c   | –4.803c   | –6.828c   |
| MSV-MCP10      | –3.652c    | –5.605c     | –5.713c     | 1.117c   | 5.936c    | –0.283c   | –7.735c  | 5.798c    | 5.427c    | –7.201c   |
| MSV-AL20       | –2.791c    | –4.996c     | –5.090c     | 7.651c   | 6.870c    | 7.735c    | 7.602c   | 7.905c    | 8.406c    | 0.312c    |
| MSV-SCA20      | –5.056c    | –6.496c     | –6.651c     | –5.701c  | 4.704c    | –5.798c   | –7.546c  | –7.905c   | –7.905c   | –7.781c   |
| MSV-MCP20      | –4.587c    | –6.202c     | –6.318c     | –5.235c  | 4.803c    | –5.427c   | –5.277c  | –8.406c   | 1.991c    | –8.366c   |
| MSV-MCP20      | –2.792c    | –4.995c     | –5.088c     | –1.182c  | 6.828c    | 7.201c    | 7.095c   | –0.312c   | 7.781c    | 8.366c    |

Note: a, b, c: Rejection of the hypothesis at 10%, 5%, and 1% respectively. The MSV models are indexed by the number of lags $m$.

where $h$ is the number of one-period ahead forecasts and $\sqrt{\text{Var}(\tilde{t}_{ij})}$ is the bootstrap estimate of the variance of $\tilde{t}_{ij}$. The $p$-values of the test statistics are obtained using a bootstrap method. For a given confidence level, if $H_0$ is rejected, the worst performing model is excluded from the set $M$, where such a model is identified using the following rule:

$$j = \arg \max_{j \in M} \left( \sum_{i \in M \setminus j} \tilde{t}_{ij} \right) \left( \sqrt{\text{Var} \left( \sum_{i \in M \setminus j} \tilde{t}_{ij} \right)} \right)^{-1/2},$$

where the variance is obtained again using a bootstrap approach. Table IV reports the MCS results for both MSCI and S&P 100 portfolios when applied to the loss function $t_{ij}$. We used the statistic $t_{SQ}$ to compute the $p$-values for three confidence levels (5%, 10%, and 20%). These $p$-values inform about the included/excluded models for a given confidence level. If a $p$-value is larger than the fixed confidence level, the corresponding model is included in the MCS test of statistically equivalent models. The higher the $p$-value, the better the model is in terms of prediction.
Table IV. These tables report the MCS results using the loss $u_{ij}$ and the statistic $t_{SQ}$ for different confidence levels

| Confidence level | 5%   | 10%  | 20%  |
|------------------|------|------|------|
| (a) MCS $p$-values, MSCI portfolio |      |      |      |
| DCC              | 0.001| 0.001| 0.001|
| O-G              | 0.001| 0.001| 0.001|
| BEKK             | 0.008| 0.008| 0.009|
| MSV$_{10}$       | 0.723| 0.704| 0.697|
| MSV-AL$_{10}$    | 0.934| 0.916| 0.929|
| MSV-SCA$_{10}$   | 0.103| 0.177| 0.111|
| MSV-MCP$_{10}$   | 0.723| 0.704| 0.697|
| MSV$_{20}$       | 0.723| 0.704| 0.697|
| MSV-AL$_{20}$    | 0.723| 0.865| 0.862|
| MSV-SCA$_{20}$   | 0.934| 0.916| 0.929|
| MSV-MCP$_{20}$   | 1.000| 1.000| 1.000|

| Confidence level | 5%   | 10%  | 20%  |
|------------------|------|------|------|
| (b) MCS $p$-values, S&P 100 portfolio |      |      |      |
| DCC              | 0.001| 0.001| 0.002|
| O-G              | 0.001| 0.001| 0.001|
| BEKK             | 0.001| 0.001| 0.001|
| MSV$_{10}$       | 0.001| 0.001| 0.001|
| MSV-AL$_{10}$    | 1.000| 1.000| 1.000|
| MSV-SCA$_{10}$   | 0.001| 0.001| 0.002|
| MSV-MCP$_{10}$   | 0.001| 0.001| 0.001|
| MSV$_{20}$       | 0.06 | 0.007| 0.05 |
| MSV-AL$_{20}$    | 0.301| 0.280| 0.250|
| MSV-SCA$_{20}$   | 0.001| 0.001| 0.001|

Note: Bold values indicate that the models are included in the confidence set, that is, they are statistically equivalent in terms of prediction accuracy. For each confidence level, the lowest $p$-value corresponds to the first model being excluded; the largest $p$-value corresponds to the best performing model. The MSV models are indexed by the number of lags $m$.

accuracy. For the MSCI portfolio results in Table IV(a), we can draw the following remarks: our MSV specifications are always included for any confidence level, contrary to the standard MGARCH models; among the MSV specifications, the penalized processes provide better forecasting performances. For the high-dimensional S&P 100 portfolio in Table IV(b), only the adaptive LASSO (with $m = 10$) and SCAD (with $m = 20$) penalized models are included in the test for all confidence levels. These results support our findings in the Diebold–Mariano test.

5. CONCLUSION

The focus of this study was on the estimation of high-dimensional MSV models. Our main contribution consisted in proposing an estimation framework that does not rely on standard MCMC/MCL methods but instead on a penalized OLS framework for state-space estimation. The corresponding large sample properties of the two-step estimator are derived. In particular, we considered a sparse first-step estimator for a broad range of penalty functions when the number of parameters is diverging. We derived an explicit convergence rate of the second-step estimator and its large sample distribution. The performances of our proposed method compared with standard MGARCH models are illustrated through simulated experiments together with an out-of-sample analysis for prediction accuracy, where our method clearly outperformed the competing MGARCH models. These results also emphasized the gain of penalization, which manages the over-fitting problem.

Various issues and extensions can be further considered. Our proposed model could be extended to accommodate a factor structure, long memory, and/or asymmetry, as discussed in Asai et al. (2006) and Chib et al. (2009).
Besides the factor setting considered by Chib et al. (2006) and So and Choi (2009), we can consider the rotation of the variables as in Harvey et al. (1994) and Hafner and Preminger (2009). For the long memory property, So and Kwok (2006) extended Harvey et al. (1994)’s model; hence, we can consider applying their work. For including asymmetric effects, we may extend Harvey and Shephard (1996)’s approach; see Asai and McAleer (2009a) for the multivariate case. Another direction would include modeling directly the variance–covariance matrix $H_t$ without relying on the decomposition $D_t \Gamma D_t$. To do so, a log-type dynamic on $H_t$ could be considered and the estimation could be managed through the development of a suitable state-space-based setting.

ACKNOWLEDGEMENTS

The authors are most grateful to Yoshihisa Baba, Shiqing Ling, the associate editor, and the two anonymous reviewers for their helpful comments and suggestions. This research is partially supported by the Japan Society for the Promotion of Science (19K01594, 19K23193).

DATA AVAILABILITY STATEMENT

The data that support the findings will be available through MSCI at https://www.msci.com/ and Yahoo Finance at https://finance.yahoo.com following an embargo from the date of publication to allow for the commercialization of the research findings.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

REFERENCES

Asai M, Caporin M, McAleer M. 2015. Forecasting value-at-risk using block structure multivariate stochastic volatility. International Review of Economics & Finance 40:60–50.

Asai M, McAleer M. 2009a. Multivariate stochastic volatility, leverage and news impact surfaces. Econometrics Journal 12:292–309.

Asai M, McAleer M. 2009b. The structure of dynamic correlations in multivariate stochastic volatility models. Journal of Econometrics 150:182–192.

Asai M, McAleer M, Yu J. 2006. Multivariate stochastic volatility: A review. Econometric Reviews 25:145–175.

Baba, Y., Engle, R.F., Kraft, D., and Kroner, K. (1985). Multivariate Simultaneous Generalized ARCH. Unpublished Paper, University of California, San Diego. [Published as Engle and Kroner (1995)]

Bauwens L, Laurent S, Rombouts JKV. 2006. Multivariate GARCH models: A survey. Journal of Applied Econometrics 21:79–109.

Belloni A, Chernozhukov V, Hansen C. 2011. Inference for high-dimensional sparse econometric models. Advances in Economics and Econometrics. 10th World Congress, Econometric Society. Vol. 3, Cambridge: Cambridge University Press, 245–295.

Boussama F. 2006. Ergodicitédes chaînes de Markov à valeursdansunevariétéalgébrique: Application aux modèlesGARCH-multivariés. Comptes Rendus de l’Académiedes Sciences Paris 343:275–278.

Candelon B, Colletaz G, Hurlin C, Tokpavi S. 2011. Backtesting value-at-risk: A GMM duration-based approach. Journal of Financial Econometrics 9:314–343.

Chang Y, Park JY. 2009. On the asymptotics of ADF tests for unit roots. Econometric Reviews 28:431–447.

Chang Y, Park JY, Song K. 2006. Bootstrapping cointegrating regressions. Journal of Econometrics 133:703–739.

Chib S, Nardari F, Shephard N. 2006. Analysis of high dimensional multivariate stochastic volatility models. Journal of Econometrics 134:341–371.

Chib S, Omori Y, Asai M. 2009. Multivariate stochastic volatility. In Handbook of Financial Time Series, Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds.). Springer-Verlag, New York, NY: 365–400.

Daniele J. 1998. Multivariate stochastic volatility models: Estimation and a comparison with VGARCH models. Journal of Empirical Finance 5:155–173.

Diebold F, Mariano R 1995. Comparing predictive accuracy. Journal of Business & Economic Statistics, 13, 253–263.
Ding L., Vo M. 2012. Exchange rates and oil prices: A multivariate stochastic volatility analysis. *Quarterly Review of Economics and Finance* 52:15–37.

Durbin J., Koopman SJ. 1997. Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika* 84:669–684.

Durbin J., Koopman SJ. 2001. *Time Series Analysis by State Space Methods* Oxford University Press, Oxford, UK.

Engle RF. 2002. Dynamic conditional correlation. *Journal of Business & Economic Statistics* 20:339–350.

Engle RF, Colacito R. 2006. Testing and valuing dynamic correlations for asset allocation. *Journal of Business & Economic Statistics* 24:238–253.

Engle RF, Kroner KF. 1995. Multivariate simultaneous generalized ARCH. *Econometric Theory* 11:122–150.

Fan J., Li R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456):1348–1360.

Fan J., Peng H. 2004. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* 32(3):928–961.

Françq C., Zakoïan J-M. 2010. *GARCH Models Structure, Statistical Inference and Financial Applications* John Wiley and Sons, Chichester, West Sussex.

Ghysels E, Harvey AC, Renault E. 1996. *Econometric Models Structure, Statistical Inference and Financial Applications* Springer, New York, NY.

Granger CWJ, Morris M. 1976. Time series modeling and interpretation. *Journal of the Royal Statistical Society, Series A* 139:246–257.

Hafner CM, Preminger A. 2009. Asymptotic theory for a factor GARCH model. *Biometrika* 96:81–94.

Hansen PR, Lunde A, Nason JM. 2003. Choosing the best volatility models: The model confidence set approach. *Oxford Bulletin of Economics and Statistics* 65:839–861.

Hansen PR, Lunde A, Nason JM. 2011. The model confidence set. *Econometrica* 79(2):453–497.

Harvey A. 1998. *Long memory in stochastic volatility*. In *Forecasting Volatility in Financial Markets*, Knight J, Satchell S (eds.). Butterworth-Haineman, Oxford, UK; 307–320.

Harvey AC, Ruiz E, Shephard N. 1994. Multivariate stochastic variance models. *Review of Economic Studies* 61:247–264.

Harvey AC, Shephard N. 1996. Estimation of an asymmetric stochastic volatility model for asset returns. *Journal of Business & Economic Statistics* 14:429–434.

Hastie T, Tibshirani R, Wainwright M. 2015. *Statistical Learning with Sparsity: The LASSO and Generalizations*. Monographs on Statistics and Applied Probability, Vol. 143 Chapman & Hall, Boca Raton, FL.

Kastner G, Frühwirth-Schnatter S, Lopes HF. 2017. High-dimensional sparse multivariate SV models. *Journal of Multivariate Analysis* 163:297–328.

Klopp T, Wainwright M. 2017. Support recovery without incoherence: A case for non-convex regularization. *The Annals of Statistics* 45(6):2455–2482.

Lütkepohl H. 2006. *New Introduction to Multiple Time Series Analyses* Springer-Verlag, New York, NY.

Ng S. 2013. Variable selection in predictive regressions. *Handbook of Economic Forecasting* 2:752–789.

Poignard B. 2020. Asymptotic theory of the adaptive sparse group lasso. *Annals of the Institute of Statistical Mathematics* 72:297–328.

Taylor SJ. 1994. *Modeling stochastic volatility: A review and comparative study*. *Mathematical Finance* 4:183–204.
Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B* **58**(1):267–288.

Tse YK, Tsui AKC. 2002. A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations. *Journal of Business & Economic Statistics* **20**:351–361.

Wilms I, Basu S, Bien J, Matteson DS. 2021. Sparse identification and estimation of large-scale vector autoregressive moving averages. *Journal of the American Statistical Association*.

Zhang C-H. 2010. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**:894–942.

Zou H. 2006. The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association* **101**(476):1418–1429.