Discussion Papers In Economics And Business

A Penalised OLS Framework for High-Dimensional Multivariate Stochastic Volatility Models

Benjamin Poignard

Manabu Asai

Discussion Paper 20-02

January 2020

Graduate School of Economics
Osaka University, Toyonaka, Osaka 560-0043, JAPAN

A Penalised OLS Framework for High-Dimensional Multivariate Stochastic Volatility Models*

Benjamin Poignard†
Graduate School of Economics, Osaka University and Riken-AIP, Japan

Manabu Asai‡
Faculty of Economics, Soka University, Japan

Abstract

Although multivariate stochastic volatility (MSV) models usually produce more accurate forecasts compared to multivariate GARCH models, their estimation techniques such as Monte Carlo likelihood or Bayesian Markov Chain Monte Carlo are computationally demanding and thus suffer from the so-called "curse of dimensionality": using such methods, the applications are typically restricted to low-dimensional vectors. In this paper, we propose a fast estimation approach for MSV models based on a penalised ordinary least squares framework. Specifying the MSV model as a multivariate state-space model, we propose a two-step penalised procedure for estimating the latter using a broad range of potentially non-convex penalty functions. In the first step, we approximate an EGARCH type dynamic using a penalised AR process with a sufficiently large number of lags, providing a sparse estimator. Conditionally on this first step estimator, we estimate the state vector based on a AR type dynamic. This two-step procedure relies on OLS based loss functions and thus easily accommodates high-dimensional vectors. We provide the large sample properties of the two-step estimator together with the so-called support recovery of the first step estimator. The empirical performances of our method are illustrated through in-sample simulations and out-of-sample variance-covariance matrix forecasts, where we consider as competitors commonly used MGARCH models.

Keywords: Forecasting; Multivariate Stochastic Volatility; Oracle Property; Penalised M-estimation.

JEL Classification: C13, C32.

*The authors are most grateful to Yoshi Baba for very helpful comments and suggestions. The second author acknowledges the financial support of the Japan Ministry of Education, Culture, Sports, Science and Technology, Japan Society for the Promotion of Science, and the Australian Academy of Science. The second author acknowledges the financial support of the Japan Society for the Promotion of Science.

†E-mail address: bpoignard@econ.osaka-u.ac.jp

‡E-mail address: m-asai@soka.ac.jp

1 Introduction

Over these past decades, various covariance models have been being 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 et al. (1985) and Engle and Kroner (1995), and their variants are commonly used: see the survey of Bauwens, Laurent, and Rombouts (2006) for instance. In the multivariate stochastic volatility (MSV) family, the MSV model of Harvey, Ruiz, and Shephard (1994) was extended, among others, by the factor model of Chib, Nardari, and Shephard (2006) and the dynamic correlation model of Asai and McAleer (2009): see, e.g., Ghysels, Harvey, and Renault (1996), Asai, McAleer, and Yu (2006), and Chib, Omori, and Asai (2009) for various univariate and multivariate stochastic volatility models. Based on a thorough empirical analysis, Chib, Nardari, and Shephard (2006) highlighted that the MSV models usually outperform MGARCH based models, such as, e.g, the DCC and the BEKK models in terms of out-of-sample forecasts.

For estimating MSV models, Harvey, Ruiz, and Shephard (1994) derived a state space form based on the vector of the log of squared returns. Using the corresponding state space form, they performed a Kalman based filtering technique to evaluate the quasi log likelihood function. In the recent literature, a commonly used method is the Bayesian MCMC one, as described, e.g., in Chib, Omori, and Asai (2009) and Kastner, Frühwirth-Schnatter, and Lopes (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, Caprion, and McAleer (2015) and Asai and McAleer (2009). Due to the severe cost in terms of computations, empirical applications in the literature are typically limited to low-dimensional random vectors when the MCMC and

MCL approaches are carried out. MGARCH specifications also suffer from the so-called "curse of dimensionality" since the complexity is in general of order $O(p^2)$, 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, inducing practical problems of inference and possibly overfitting. Moreover, tricky conditions are required on the model parameters to satisfy the positive-definiteness of the variance-covariance process.

Another key hurdle of the aforementioned methods is the high non-linearity of the models, which requires the use of likelihood based estimation approaches. Therefore, strongly reduced versions of such multivariate models are most often considered as soon as p is larger than four or five, typically. Another approach is given by factor modelling, which aims at reducing the model complexity. Among others, Fan et al. (2008) emphasised the relevance of factor models for high-dimensional precision matrix estimation. However, this approach requires the identification of the corresponding factors. An "expert" approach is based on some priors regarding the leading underlying factors. Otherwise, latent unobserved factors induce particular estimation issues and their number is questionable.

The objective of this paper consists in modelling high-dimensional variance-covariance matrices within the multivariate stochastic volatility framework in a flexible manner and breaking the curse of dimensionality without relying on standard MCMC or MCL based procedures. To do so, we introduce a vector autoregressive and moving-average (VARMA) representation for the MSV model of Harvey, Ruiz, and Shephard (1994) and apply an OLS-based two-step estimation approach extending the idea of Hannan and Rissanen (1982) and Hannan and Kavalieris (1984). Applying their approach to the MSV model requires an OLS estimation of a large dimensional VAR model with a sufficiently large number of lags in the first step. Nonetheless, for the purpose of parsimony and to avoid overfitting, we have to enforce the nullity of possibly numerous model coefficients. The OLS objective function is particularly adapted for regularisation/penalisation procedures and fast closed form algorithms can be applied. Our study shares a similar spirit with Poignard and Fermanian (2019), who provided a framework for high-dimensional variance-covariance within the MGARCH family: they derived some parameterizations to directly generate positive-definite covariance matrices based on Multivariate ARCH processes, which enables a linear representation with respect to the parameters and thus the use of a penalised OLS criterion at the estimation step. But our work differs from theirs in two main respects: our analysis lies within the multivariate stochastic volatility family; we consider a general penalisation framework, which includes a broad range of different penalty functions.

The main contributions of our method are as follows: using a parsimonious OLS framework, we can directly generate positive-definite variance-covariance matrices without relying on MCMC/MCL like methods and manage high-dimensional matrix processes; the large sample properties of the two-step estimator are provided together with the conditions for satisfying the so-called oracle property for the first step estimator in the sense of Fan and Li (2001), which ensures the correct identification of the underlying set of nonzero coefficients. In the first step, we consider a general penalised M-estimation framework, which encompasses potentially non-convex penalty functions.

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

Notations. Throughout this paper, we denote the cardinality of a set E by $\text{card}(E)$. For a vector $\boldsymbol{v} \in \mathbb{R}^d$, the ℓ_p norm is $\|\boldsymbol{v}\|_p = \left(\sum_{k=1}^{p} |\boldsymbol{v}_k|^p\right)^{1/p}$ for $p > 0$, and $\|\boldsymbol{v}\|_\infty = \underset{i}{\max}|\boldsymbol{v}_i|$. Let the subset $\mathcal{A} \subseteq \{1, \cdots, d\}$, then $\boldsymbol{v}_{\mathcal{A}} \in \mathbb{R}^{\text{card}(\mathcal{A})}$ is the vector $\boldsymbol{v}$ restricted to $\mathcal{A}$. $\mathcal{M}_{m\times n}(\mathbb{R})$ denotes the space of $m \times n$ matrices with coefficients in $\mathbb{R}$. For a matrix A, $\|A\|_s$ is the spectral norm. We write A' (resp. $\boldsymbol{v}'$) to denote the transpose of the matrix A (resp. the vector $\boldsymbol{v}$). We write $vec(A)$ to denote the vectorization operator that stacks the columns of A on top of one another into a vector. We denote by $\text{vech}(A)$ the $p(p+1)/2$ vector that stacks the columns of the lower triangular part of the square and symmetric matrix A. The I_p matrix is the p-dimensional identity matrix. For a function $f : \mathbb{R}^d \to \mathbb{R}$, we denote by ∇f the gradient or subgradient of f and $\nabla^2 f$ the Hessian of f. We denote by $(\nabla^2 f)_{\mathcal{A}\mathcal{A}}$ 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 Penalised OLS framework for MSV

2.1 Framework

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

$$y_t = D_t \varepsilon_t, \quad \varepsilon_t \sim iid(0, \Gamma), \tag{1}$$

$$h_{t+1} = \mu + \Phi(h_t - \mu) + \eta_t, \quad \eta_t \sim \mathcal{N}_{\mathbb{R}^p}(0, \Sigma_\eta), \tag{2}$$

where Γ is a $p \times p$ correlation matrix, $\varepsilon_t = (\varepsilon_{1t}, \ldots, \varepsilon_{pt})'$ is a $p \times 1$ random vector, which is independently and identically distributed (i.i.d.), centered with variance-covariance Γ, $h_t = (h_{1t}, \ldots, h_{pt})'$ is a $p \times 1$ vector of log-volatility, $D_t = \text{diag}\{\exp(h_{1t}/2), \ldots, \exp(h_{pt}/2)\}$ is a diagonal matrix of volatility, $\mu = (\mu_1, \dots, \mu_p)'$ is a $p \times 1$ vector, Φ is a $p \times p$ matrix, and Σ_η is a $p \times p$ covariance matrix of η_t. The MSV model (1) and (2) reduces to the MSV model of Harvey, Ruiz, and Shepard (1994), should we assume Φ diagonal and ε_{it} following the t distribution.

Then we define $x_t = (\log(y_{1t}^2), \dots, \log(y_{pt}^2))'$. As discussed in Harvey, Ruiz, and Shepard (1994), the MSV model can be formulated as a state space model, specified as

$$x_t = c + \alpha_t + \zeta_t, \tag{3}$$

$$\alpha_{t+1} = \Phi \alpha_t + \eta_t, \tag{4}$$

where $c = (c_1, \dots, c_p)'$, $\zeta_t = (\zeta_{1t}, \dots, \zeta_{pt})'$, and $\alpha_t = h_t - \mu$ with $c_i = \mu_i + \mathbb{E}[\log(\varepsilon_{it}^2)]$ and $\zeta_{it} = \log(\varepsilon_{it}^2) - \mathbb{E}[\log(\varepsilon_{it}^2)]$. Assuming a t distribution for ε_{it}, Harvey, Ruiz, and Shepard (1994) specified the covariance matrix of ζ_t as Σ_ζ. Note that $\mathbb{E}[\zeta_t] = 0$ by definition. Based on the state space form, these authors then 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, Nardari, and Shephard (2006) or 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 paper, we aim at tackling this issue for MSV models using a penalised OLS estimation method. Although the MSV model (1) and (2) might be a basic model, the following advantages with respect to MGARCH models can be highlighted: (i) relatively stable estimates and forecasts for variance-covariance matrices; (ii) simpler restrictions for stationarity conditions; (iii) no intricate matrix parameterization and/or parameter restrictions to generate positive-definite matrices.

2.2 Our proposed approach

In this section, we specify our proposed sparse OLS based MSV model - named as "penalised OLS-MSV" - and discuss how the latter can manage high-dimensional stochastic vectors. Our approach can be summarized as follows: derive a VARMA representation of x_t, and consider a regression-based estimator of (c, Φ), following the ideas of Hannan and Rissanen (1982) and Hannan and Kavalieris (1984); conditionally on the VARMA estimators, use an ad-hoc estimator for Σ_ζ such that the corresponding estimator is positive-definite; finally, obtain the estimator of Γ.

More precisely, the procedure is as follows. Since x_t 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 x_t has a VARMA(1,1) representation. By equations (3) and (4), we obtain

$$x_t = (I - \Phi)c + \Phi x_{t-1} + (\zeta_t + \eta_{t-1}) - \Phi\zeta_{t-1},$$

which has an alternative representation as

$$\begin{gathered} x_t = (I - \Phi)c + \Phi x_{t-1} + u_t + \Xi u_{t-1}, \\ \mathbb{E}[u_t] = 0, \quad \mathrm{Var}(u_t) = \Sigma_u, \quad \mathbb{E}[u_t u_s'] = \mathbf{0} \text{ for } t \neq s, \end{gathered} \tag{5}$$

where Ξ and Σ_u are obtained by matching moments of $w_t = (\zeta_t + \eta_{t-1}) - \Phi\zeta_{t-1}$ and $w_t^* = u_t + \Xi u_{t-1}$. By considering $\mathbb{E}[w_t w_t'] = \mathbb{E}[w_t^* w_t^{*\prime}]$ and $\mathbb{E}[w_t w_{t-1}'] = \mathbb{E}[w_t^* w_{t-1}^{*\prime}]$, the relationship between (Ξ, Σ_u) and other parameters is given as follows:

$$\Sigma_\eta + \Sigma_\zeta + \Phi\Sigma_\zeta\Phi' = \Sigma_u + \Xi\Sigma_u\Xi', \tag{6}$$

$$-\Phi\Sigma_\zeta = \Xi\Sigma_u. \tag{7}$$

Given the values of Φ, Ξ, and Σ_u we may obtain Σ_η and Σ_ζ by solving equations (6) and (7).

Under suitable stationarity conditions, x_t has an AR(∞) representation:

$$x_t = \sum_{i=1}^{\infty} \Psi_i x_{t-i} + u_t. \tag{8}$$

Based on a penalised OLS estimation, we can obtain a estimator of $\hat{u}_t$ in the first step. Indeed, we empirically need to specify m sufficiently large as a surrogate of ∞ in the summation in (8). Thus, for the sake of parsimony and to avoid the overfitting issue, we propose to enforce the nullity of possibly numerous model coefficients in the Φ_i's.

In the second step, we calculate the OLS estimator of $(\hat{c}^*, \hat{\Phi}, \hat{\Xi})$ by regressing x_t on a constant, x_{t-1}, and $\hat{u}_{t-1}$.

For the third step, we start from the decomposition of the unconditional variance-covariance matrix of x_t, which is given by

$$\Sigma_x = \Sigma_\alpha + \Sigma_\zeta, \tag{9}$$

where $\Sigma_x = \mathbb{E}[(x_t - c)(x_t - c)']$, $\Sigma_\zeta = \mathbb{E}[\zeta_t \zeta_t']$, and $\Sigma_\alpha = \mathbb{E}[\alpha_t \alpha_t']$ with

$$\text{vec}(\Sigma_\alpha) = [I_{p^2} - (\Phi \otimes \Phi)]^{-1} \text{vec}(\Sigma_\eta).$$

Denote the sample covariance matrix of x_t and $\hat{u}_t$ as S_x and $S_{\hat{u}}$, respectively. An estimator of Σ_ζ is given as

$$S_\zeta = -\frac{1}{2}\left[\hat{\Phi}^{-1}\hat{\Xi}S_{\hat{u}} + S_{\hat{u}}\hat{\Xi}'\hat{\Phi}'^{-1}\right],$$

by equation (7). As there is no guarantee for S_ζ and $S_x - S_\zeta$ to be positive definite, we consider ado hoc estimators for Σ_ζ and Σ_α based on decomposition (9).

In the fourth step, we estimate Γ by a correlation matrix of y_t.

We summarize our procedure as follows:

Step 1. Define $\Psi_{1:m} = [\Psi_1 \ \cdots \Psi_m] \in \mathcal{M}_{p\times pm}(\mathbb{R})$. Approximate (5) by

$$x_t = \sum_{i=1}^{m} \Psi_i x_{t-i} + u_t, \tag{10}$$

and consider the regularised OLS estimator

$$\hat{\Psi}_{1:m} = \arg\min_{\Psi_{1:m}} \left[\frac{1}{2T} \sum_{t=1}^{T} ||x_t - \sum_{i=1}^{m} \Psi_i x_{t-i}||_2^2 + \boldsymbol{p}(\frac{\lambda_T}{T}, \mathrm{vec}(\Psi_{1:m})) \right], \tag{11}$$

where $\boldsymbol{p}(\frac{\lambda_T}{T}, \cdot) : \mathbb{R}^d \to \mathbb{R}$ is a penalty function applied to each component of $\Psi_{1:m}$, where λ_T is the regularisation parameter, which depends on the sample size, and enforce a particular type of sparse structure in the solution $\hat{\Psi}_{1:m}$. The parameter dimension in the first step is denoted as d, which in Step 1 is $d = mp^2$. In this paper, as it will be detailed in Section 3, we consider the SCAD due to Fan and Li (2001), the MCP due to Zhang (2010), the Lasso of Tibshirani (1996) and the Bridge of Fu (1998).

Step 2. Define $\hat{u}_t = x_t - \sum_{i=1}^{m} \hat{\Psi}_i x_{t-i}$. Conditionally on $\hat{\Psi}_{1:m}$, we consider the regression

$$x_t = c^* + \Phi x_{t-1} + \Xi \hat{u}_{t-1} + v_t,$$

where the parameters are (c^*, Φ, Ξ). Thus, the second step objective function is

$$(\hat{c}^*, \hat{\Phi}, \hat{\Xi})|\hat{\Psi}_{1:m} = \arg\min_{(c^*,\Phi,\Xi)} \left[\frac{1}{2T} \sum_{t=1}^{T} ||x_t - \big(c^* + \Phi x_{t-1} + \Xi \hat{u}_{t-1}\big)||_2^2 \right],$$

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

Step 3. Consider the estimator of Σ_ζ and Σ_α as follows:

$$\hat{\Sigma}_\zeta = rS_x, \quad \hat{\Sigma}_\alpha = (1 - r)S_x, \tag{12}$$

where r is a constant which satisfies $0 < r < 1$. This ad hoc method aims at treating the positive definiteness of the estimators and dealing with the high-dimensionality issue of $\hat{\Sigma}_\zeta$. While we consider a naive decomposition based on equation (9) for the former, we set $r = (\pi^2/2)(p^{-1}\sum_{i=1}^{p} S_{x,ii})^{-1}$ in (12). Here, $\pi^2/2$ is the value of $\mathbb{E}[\zeta_{it}^2]$ when ε_{it} follows the standard normal distribution. Based on this specification, the average of the diagonal elements of $\hat{\Sigma}_\zeta$ becomes $\pi^2/2$. Using such approach, we are able to accurately estimate Σ_ζ and Σ_α and importantly the computational cost is negligible, compared to alternative estimators (e.g. GMM type method) that would require a numerical optimisation.

Step 4. Estimate Γ by a correlation matrix of y_t.

When the tuning parameter λ_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 actually corresponds to a multivariate version of the AR(∞) representation of a log-GARCH model, which is a special case of the exponential GARCH model of Nelson (1991). Thus, for significantly large m chosen ex-ante, (10) would be a relevant approximation of a log-GARCH type process. Our approach avoids the use of computationally demanding methods such as MCMC and MCL. Although Harvey, Ruiz, and Shephard (1994) applied the Kalman filter, its computational cost is non-negligible for larger p, since the cost increases with the speed of $O(Tp^2)$ for storing covariance matrices of $p \times 1$ state vector for all $t = 1, \dots, T$. 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 Σ_ζ and Σ_α. However, we use the above fast 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 α_t based on the full sample $\boldsymbol{x} = (x_1', \dots, x_T')'$ $(Tp \times 1)$ by the state space smoothing algorithm. In the high-dimensional case, we consider the multivariate version of the approach of Harvey (1998) with the vector form of (3), as follows:

$$\boldsymbol{x} = \boldsymbol{c}^{\dagger} + \boldsymbol{\alpha} + \boldsymbol{\zeta}$$

where $\boldsymbol{c}^{\dagger} = (\iota_T \otimes c)$ $(Tp \times 1)$, $\boldsymbol{\alpha} = (\alpha_1', \dots, \alpha_T')'$ $(Tp \times 1)$, and $\boldsymbol{\zeta} = (\zeta_1', \dots, \zeta_T')'$ $(Tp \times 1)$. By the model structure, the covariance matrix of $\boldsymbol{x}$ is given by

$$V_x = V_\alpha + V_\zeta,$$

where

$$V_\alpha = \begin{pmatrix} \Sigma_\alpha & \Sigma_\alpha \Phi' & \Sigma_\alpha (\Phi')^2 & \cdots & \Sigma_\alpha (\Phi')^{T-2} & \Sigma_\alpha (\Phi')^{T-1} \\ \Phi \Sigma_\alpha & \Sigma_\alpha & \Sigma_\alpha \Phi' & \cdots & \Sigma_\alpha (\Phi')^{T-3} & \Sigma_\alpha (\Phi')^{T-2} \\ \Phi^2 \Sigma_\alpha & \Phi \Sigma_\alpha & \Sigma_\alpha & \cdots & \Sigma_\alpha (\Phi')^{T-4} & \Sigma_\alpha (\Phi')^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \Phi^{T-2} \Sigma_\alpha & \Phi^{T-3} \Sigma_\alpha & \Phi^{T-4} \Sigma_\alpha & \cdots & \Sigma_\alpha & \Sigma_\alpha \Phi' \\ \Phi^{T-1} \Sigma_\alpha & \Phi^{T-2} \Sigma_\alpha & \Phi^{T-3} \Sigma_\alpha & \cdots & \Phi \Sigma_\alpha & \Sigma_\alpha \end{pmatrix},$$

and $V_\zeta = (I_T \otimes \Sigma_\zeta)$. Then the MMSLE is

$$\tilde{\boldsymbol{\alpha}} = V_\alpha V_x^{-1} (\boldsymbol{x} - \boldsymbol{c}^{\dagger}) + \boldsymbol{c}^{\dagger}. \tag{13}$$

As in Harvey (1998), we consider an estimator of the covariance matrix, $H_t = D_t \Gamma D_t$, such that the sample variance of the standardized variable of y_{it} equals to one. We consider the estimator as

$$\tilde{H}_t = \tilde{D}_t \hat{\Gamma} \tilde{D}_t, \tag{14}$$

where

$$\tilde{D}_t = \operatorname{diag}\left\{ \tilde{d}_{1t}, \dots, \tilde{d}_{pt} \right\}, \quad \tilde{d}_{it} = \bar{d}_i \exp(\tilde{x}_{it}/2), \quad \bar{d}_i = \sqrt{T^{-1} \sum_{t=1}^{T} y_{it}^2 \exp(-\tilde{x}_{it})},$$

for $i = 1, \dots, p$. The standardized variables are defined as $\tilde{z}_{it} = y_{it}/\tilde{d}_{it}$, which implies by definition $T^{-1} \sum_{t=1}^{T} \tilde{z}_{it}^2 = 1$. We call our proposed parameterisation "penalised OLS-MSV".

2.3 Volatility forecasting

We now provide the forecasts for variance-covariance based on our proposed method. The MMSLE for the lth-step-ahead forecast of α_T is given by:

$$\hat{\alpha}_{T+l} = R_l V_x^{-1}(\boldsymbol{x} - c^{\dagger}) + c, \tag{15}$$

where

$$R_l = \left[\Phi^{T+l-1}\Sigma_\alpha \;\; \Phi^{T+l-2}\Sigma_\alpha \;\; \cdots \;\; \Phi^{l}\Sigma_\alpha\right].$$

The lth-step-ahead forecast of the covariance matrix is given by:

$$\hat{H}_t = \hat{D}_t \hat{\Gamma} \hat{D}_t, \tag{16}$$

where

$$\hat{D}_{T+l} = \mathrm{diag}\left\{\hat{d}_{1,T+l}, \ldots, \hat{d}_{p,T+l}\right\}, \quad \hat{d}_{i,T+l} = \bar{d}_i \exp\left(\hat{x}_{i,T+l}/2\right),$$

for $i = 1, \ldots, p$.

3 Asymptotic properties

3.1 First step: penalised estimator $\hat{\Psi}_{1:m}$

In this section, we focus on the large sample properties of the first step penalised estimator $\hat{\Psi}_{1:m}$. We consider a loss function $\mathbb{G}_T$ from $\mathbb{R}^{pT} \times \Theta_1$ to $\mathbb{R}$. The value $\mathbb{G}_T(\underline{y};\theta)$, with $\theta = vec(\Psi_{1:m}) \in \Theta_1 \subset \mathbb{R}^d$ and $d = mp^2$, evaluates the quality of the "fit" for the realizations of y_t for every $t = 1, \cdots, T$ and under $\mathbb{P}_\theta$. $\mathbb{G}_T(\underline{y};\theta)$ is the empirical loss associated to a continuous function $\ell : \mathbb{R}^{pT} \times \Theta_1 \to \mathbb{R}$, idest

$$\mathbb{G}_T(\underline{y};\theta) := \frac{1}{T}\sum_{t=1}^{T}\frac{1}{2}\|x_t - \sum_{i=1}^{m}\Psi_i x_{t-i}\|_2^2 = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{2}\|x_t - \Psi_{1:m} Z_{m,t-1}\|_2^2 := \frac{1}{T}\sum_{t=1}^{T}\ell(y_s, s \leq t; \theta),$$

where x_t corresponds to the vector of continuous transforms of $\log(y_{it}^2)$, $\Psi_{1:m} = (\Psi_1, \cdots, \Psi_m) \in \mathcal{M}_{p\times pm}(\mathbb{R})$ and $Z_{m,t-1} = (x'_{t-1}, \cdots, x'_{t-m})' \in \mathbb{R}^{pm}$. Then, the problem of interest is

$$\hat{\theta} = \arg\min_{\theta\in\Theta_1} \{\mathbb{G}_T(\underline{y};\theta) + \sum_{i=1}^{d} \boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_i|)\}, \tag{17}$$

where $\boldsymbol{p}(\frac{\lambda_T}{T}, .) : \Theta_1 \rightarrow \mathbb{R}$ is a regulariser/penalty, λ_T is the regularisation parameter, which depends on the sample size, and enforce a particular type of sparse structure in the solution. The function $\theta \rightarrow \mathbb{E}[\ell(y_s, s \leq t; \theta)]$ is supposed to be uniquely minimized at $\theta = \theta_0$ so that $\mathbb{E}[\nabla_\theta \mathbb{G}_T(\underline{y}; \theta_0)] = 0$.

The penalisation is performed through the term $\sum_{i=1}^{d} \boldsymbol{p}\left(\frac{\lambda_T}{T}, |\theta_i|\right)$, which is a coordinate-separable penalty. In this paper, we consider the following set of penalties $\boldsymbol{p}(\lambda, \theta), \theta \in \mathbb{R}$:

$$\begin{array}{lll}
\textbf{Lasso} & : & \lambda|\theta|, \\
\textbf{Bridge} & : & \lambda|\theta|^q, q \in (0,1), \\
\textbf{MCP} & : & \text{sgn}(\theta)\lambda \int_0^{|\theta|} (1 - z/(\lambda b_{\text{mcp}}))_+ \text{d}z, \\
\textbf{SCAD} & : & \begin{cases} \lambda|\theta|, & \text{for } |\theta| \leq \lambda, \\ -\frac{(\theta^2 - 2b_{\text{scad}}\lambda|\theta| + \lambda^2)}{(2(b_{\text{scad}}-1))}, & \text{for } \lambda \leq |\theta| \leq b_{\text{scad}}\lambda, \\ (b_{\text{scad}}+1)\lambda^2/2, & \text{for } |\theta| > b_{\text{scad}}\lambda, \end{cases}
\end{array}$$

and $\lambda \geq 0$ is the regularisation parameter.

To carry out a sound asymptotic theory, we make the following assumptions.

Assumption 1. *card*$(\mathcal{A}) = k_0 < d$ *with* $\mathcal{A} = $ *supp*$(\theta) := \{i | \theta_{0,i} \neq 0\}$.

Assumption 2. *The parameter set* $\Theta_1 \subset \mathbb{R}^d$ *is compact and convex.*

Assumption 3. (y_t) *is a strictly stationary, non-anticipative and ergodic process.*

Assumption 4. *For any* $\theta \in \Theta_1$*, there exists a measurable function* $g(.)$ *such that* $|\ell(y_s, s \leq t; \theta)| \leq g(y_s, s \leq t)$ *with* $\mathbb{E}[g(y_s, s \leq t)] < \infty$.

Assumption 5. $\mathbb{H} = \mathbb{E}[\nabla^2_{\theta\theta'}\ell(y_s, s \leq t; \theta_0)]$ *and* $\mathbb{M} = \mathbb{E}[\nabla_\theta \ell(y_s, s \leq t; \theta_0) \nabla_{\theta'} \ell(y_s, s \leq t; \theta_0)]$ *exist and are positive-definite matrices.*

These assumptions deserve a few comments. First, assumption 1 corresponds to the sparsity assumption. As it is unknown, we rely on penalised M-estimation to recover this set. Assumption 2 is standard when analysing large sample properties. For the sake of clarity of our arguments, we assume 3, which refers to the probabilistic property of the process. Note that it can be relaxed in favor for mixing conditions. Should we consider e.g. strongly mixing time series, where $(y_t)_t$ would be a strongly mixing process with mixing coefficient $\alpha(.)$ satisfying $\alpha(\varsigma) \leq \kappa \rho^{\varsigma}$ with $\varsigma > 0$ and $0 < \rho < 1$, then we would rely on exponential bound adapted to the strongly mixing case to control each element of the Taylor expansions we propose to analyse. To do so, Theorem 2 of Merlevède, Peligrad and Rio (2009), could be used under strongly mixing and bounded random variables. Central limit theorem for strongly mixing processes, such as e.g. Theorem 5.2 of White (2001), could also be applied for that case. We highlight that relaxing such assumption for mixing processes would not alter the convergence rates we propose to derive. Assumption 4 corresponds to a domination condition on the likelihood function, which will be used when proving the consistency of the penalised estimator. Finally, assumption 5 is a standard regularity condition.

Theorem 1. *Under assumptions 1-4, if* $\frac{\lambda_T}{T} \rightarrow \lambda_0$*, then for any compact* $\boldsymbol{B} \subset \Theta_1$ *such that* $\theta_0 \in \boldsymbol{B}$*,*

$$\hat{\theta} \xrightarrow[T\rightarrow\infty]{\mathbb{P}} \arg\min_{\boldsymbol{x}\in\boldsymbol{B}}\{\mathbb{G}^{pen}_{\infty}(\underline{y};\boldsymbol{x})\} := \arg\min_{\boldsymbol{x}\in\boldsymbol{B}}\{\mathbb{G}_{\infty}(\underline{y};\boldsymbol{x}) + \sum_{i=1}^{d}\boldsymbol{p}(\lambda_0,|\boldsymbol{x}_i|)\} = \theta^*_0,$$

with $\mathbb{G}_{\infty}(\underline{y};\boldsymbol{x}) = \mathbb{E}[\ell(y_s, s \leq t;\boldsymbol{x})]$ *and for any scalar* x

$$\begin{array}{lll}
\textbf{Lasso}: \boldsymbol{p}(\lambda_0,|x|) & = & \lambda_0|x|, \\
\textbf{Bridge}: \boldsymbol{p}(\lambda_0,|x|) & = & \lambda_0|x|^q, \\
\textbf{MCP}: \boldsymbol{p}(\lambda_0,|x|) & = & b_{mcp}\lambda_0^2/2\mathbf{1}_{\{|x|>b_{mcp}\lambda_0\}} - (b_{mcp}\lambda_0-|x|)^2/(2b_{mcp})\mathbf{1}_{\{|x|\leq b_{mcp}\lambda_0\}}, \\
\textbf{SCAD}: \boldsymbol{p}(\lambda_0,|x|) & = & \begin{cases} \lambda_0|x|, & \text{for } |x| \leq \lambda_0, \\ -(x^2-2b_{scad}\lambda_0|x|+\lambda_0^2)/(2(b_{scad}-1)), & \text{for } \lambda_0 \leq |x| \leq b_{scad}\lambda_0, \\ (b_{scad}+1)\lambda_0^2/2, & \text{for } |x| > b_{scad}\lambda_0, \end{cases}
\end{array}$$

Hence if $\lambda_T = o(T)$*, then* $\hat{\theta}$ *is a consistent estimator.*

Remark. The penalised estimator does not converge to θ_0 when $\lambda_T = O(T)$ due to the bias originating from the penalisation. Part of the proof consists in proving a uniform convergence of the penalised criterion. To do so, we rely on Lemma 2.4 of Newey and McFadden (1994) to derive a uniform law of large numbers of the parameter dependent loss function.

We now propose an asymptotic probabilistic bound for the penalised estimator. First, we assume the following conditions on the penalty function.

Assumption 6. $\boldsymbol{p}(\frac{\lambda_T}{T}, |.|)$ *is twice continuously differentiable except at the origin. We define*

$$A_{1,T} = \max_{i \in \mathcal{A}} |\nabla_{\theta_i} \boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i}|)|, \;\; A_{2,T} = \max_{i \in \mathcal{A}} |\nabla^2_{\theta_i \theta_i} \boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i}|)|,$$

so that $A_{2,T} \to 0$.

Remark. The condition on the second derivative implies that the penalty has less influence than the non-penalised loss function in the regularised problem. Moreover, for the penalties of interest, the scaling of (λ_T, T) determines this rate.

Theorem 2. *Under assumption 1-6, the sequence of penalised estimators* $\hat{\theta}$ *satisfies*

$$\|\hat{\theta} - \theta_0\| = O_p(T^{-1/2} + \sqrt{card(\mathcal{A})} A_{1,T}).$$

Remark. This result highlights that for a suitable choice of λ_T, we would obtain a $\sqrt{T}$-consistent $\hat{\theta}$. This probability bound is similar to the so-called oracle bounds as it depends on the true sparse support card$(\mathcal{A})$.

We now derive the asymptotic distribution for the rate $\lambda_T = O(\sqrt{T})$ for the Lasso, SCAD and MCP and $\lambda_T = O(T^{q/2})$ in the Bridge case.

Theorem 3. *Under assumptions 1-6, suppose $\lambda_T = o(T)$, if the regularisation rate of the Lasso, SCAD and MCP satisfies $\lambda_T = O(\sqrt{T})$ and the Bridge regularisation rate satisfies $\lambda_T = O(T^{q/2})$, if $\lim\limits_{x \to 0^+} \nabla_x \boldsymbol{p}(\frac{\lambda_T}{T}, x) = \frac{\lambda_T}{T}$ for the SCAD and MCP, then*

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow[T\to\infty]{d} \arg \min_{\boldsymbol{u}\in\mathbb{R}^d} \{\mathbb{F}_\infty(\boldsymbol{u})\},$$

provided $\mathbb{F}_\infty(.)$ is the random function in $\mathbb{R}^d$ where

$$\mathbb{F}_\infty(\boldsymbol{u}) = \boldsymbol{u}'\boldsymbol{W} + \frac{1}{2}\boldsymbol{u}'\mathbb{H}\boldsymbol{u} + \phi(\lambda_0, \boldsymbol{u}, \theta_0),$$

where $\boldsymbol{W} \sim \mathcal{N}_{\mathbb{R}^d}(0, \mathbb{M})$ with $\mathbb{M} := \mathbb{M}(\theta_0) = \mathbb{E}[\nabla_\theta \ell(y_s, s \leq t; \theta_0)\nabla_{\theta'} \ell(y_s, s \leq t; \theta_0)]$, $\mathbb{H} = \mathbb{E}[\nabla^2_{\theta\theta'} \ell(y_s, s \leq t; \theta_0)]$ and

$$\begin{aligned}
\textbf{\textit{Lasso}} : \phi(\lambda_0, \boldsymbol{u}, \theta_0) &= \lambda_0 \sum_{k=1}^{d} \big(\boldsymbol{u}_i sgn(\theta_{0,i})\mathbf{1}_{\theta_{0,i}\neq 0} + |\boldsymbol{u}_i|\mathbf{1}_{\theta_{0,i}=0}\big), \\
\textbf{\textit{Bridge}} : \phi(\lambda_0, \boldsymbol{u}, \theta_0) &= \lambda_0 \sum_{k=1}^{d} |\boldsymbol{u}_i|^q \mathbf{1}_{\theta_{0,i}=0} \\
\textbf{\textit{MCP}} : \phi(\lambda_0, \boldsymbol{u}, \theta_0) &= \lambda_0 \sum_{k=1}^{d} |\boldsymbol{u}_i| \mathbf{1}_{\theta_{0,i}=0}, \\
\textbf{\textit{SCAD}} : \phi(\lambda_0, \boldsymbol{u}, \theta_0) &= \lambda_0 \sum_{k=1}^{d} |\boldsymbol{u}_i| \mathbf{1}_{\theta_{0,i}=0}.
\end{aligned}$$

Remark. The following comments can be noticed.

(i) To derive such distributions, we rely on specific theoretical results depending on the penalty function and thus on the (non-)convexity of the latter. These results are reported in Appendix B.

(ii) Theorem 3 establishes the $\sqrt{T}$-consistency of the penalised estimator. However, for $\lambda_T = O(\sqrt{T})$ in the Lasso case, the term in $\mathbf{1}_{\theta_{0,i}\neq 0}$ implies that the true active set $\mathcal{A}$ can not be recovered with high probability (See Proposition 1 of Zou, 2006). To fix this issue, Zou (2006) proposed the adaptive Lasso, which consists in penalising the coefficients differently through the introduction of stochastic weights. These weights are explicit functions of a first step estimator, which is $\sqrt{T}$-consistent. As a consequence, these weights alter the convergence rate of the penalisation parameter λ_T.

We now turn to the oracle property. It is well-known since Zou (2006) that the Lasso does not satisfy this property. The only method to fix this issue is to specify adaptive weights in the penalty function to penalise each coefficient differently. The key advantage of non-convex penalties (SCAD, MCP, Bridge) is that they actually allow for satisfying this property without the need of these stochastic weights.

Theorem 4. *Suppose $\lambda_T = o(T)$, for $\theta_{\mathcal{A}}$ satisfying $\|\theta_{\mathcal{A}} - \theta_{0,\mathcal{A}}\| = O_p(T^{-1/2})$, suppose assumptions 1-6, suppose the MCP and SCAD regularisation rates satisfy $\frac{\lambda_T}{T^{1/2}} \to \infty$ and*

$$\liminf_{T\to\infty} \liminf_{x\to 0^+} \frac{T}{\lambda_T} \nabla_x \boldsymbol{p}(\frac{\lambda_T}{T}, x) > 0,$$

and suppose the Bridge satisfies the regularisation rate $\frac{\lambda_T}{T^{q/2}} \to \infty, 0 < q < 1$ and $\lambda_T = O(\sqrt{T})$, then the $\sqrt{T}$-consistent local estimator $\hat{\theta}$ defined in Theorem 2 satisfies

$$\lim_{T\to\infty} \mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}) = 1, \quad \textit{and}$$

$$\sqrt{T}\big(\hat{\theta} - \theta_0\big)_{\mathcal{A}} \xrightarrow[T\to\infty]{d} \mathcal{N}_{\mathbb{R}^{k_0}}(0, \mathbb{H}^{-1}_{\mathcal{A}\mathcal{A}} \mathbb{M}_{\mathcal{A}\mathcal{A}} \mathbb{H}^{-1}_{\mathcal{A}\mathcal{A}}).$$

Remark. Theorem 4 deserves a few comments.

(i) This result establishes the conditions to satisfy the oracle property. Contrary to Theorem 3, where the rate for the SCAD/MCP is $\lambda_T = O(\sqrt{T})$ and for the Bridge is $\lambda_T = O(T^{q/2})$, we now require $\lambda_T/\sqrt{T} \to \infty$ for those ones and $\lambda_T/T^{q/2} \to \infty$ for the latter. Note that the adaptive Lasso is left aside as we report sparse estimation methods that do not require a two-step estimation. More details can be found in Zou (2006) on the adaptive Lasso and its properties.

(ii) The non-random quantities originating from the expansion of the penalty functions

$$\begin{array}{rcl}\mathbf{b}_{T,\mathcal{A}} & = & \big(\nabla_{\theta_1}\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,1}|)\text{sgn}(\theta_{0,1}),\cdots,\nabla_{\theta_{k_0}}\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,k_0}|)\text{sgn}(\theta_{0,k_0})\big)', \\ \mathbf{S}_{T,\mathcal{A}\mathcal{A}} & = & \text{diag}(\nabla^2_{\theta_i\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}|),i=1,\cdots,k_0),\end{array}$$

vanish when T is large enough due to the assumed convergence rates of the penalisation parameter λ_T for the Bridge, SCAD and MCP. A detailed discussion is provided in the proof.

3.2 Second step estimator $(\hat{c}^*,\hat{\Phi},\hat{\Xi})$

We now focus on the large sample properties of the second step estimator $\hat{\gamma}=(\hat{c}^*,vec(\hat{\Phi})',vec(\hat{\Xi}))'$, which is of size $p(1+2p)$. Conditionally on $\hat{\theta}$, we consider a second step loss function $\mathbb{L}_T$ from $\mathbb{R}^{pT}\times\Theta_2$ to $\mathbb{R}$ with $\Theta_2\subset\mathbb{R}^{p(1+2p)}$ and $\mathbb{L}_T(\underline{y};\hat{\theta},\gamma)$ is the empirical loss associated to a continuous function $f:\mathbb{R}^{pT}\times\Theta_1\times\Theta_2\to\mathbb{R}$, idest

$$\mathbb{L}_T(\underline{y};\hat{\theta},\gamma)=\frac{1}{T}\sum_{t=1}^{T}\frac{1}{2}\|x_t-\alpha-\Phi x_{t-1}-\Xi\hat{u}_{t-1}\|_2^2:=\frac{1}{T}\sum_{t=1}^{T}f(y_s,s\leq t;\hat{\theta},\gamma).$$

The problem of interest is

$$\hat{\gamma}=\arg\min_{\gamma\in\Theta_2}\{\mathbb{L}_T(\underline{y};\hat{\theta},\gamma)\}.$$

To derive the large sample properties of the second step estimator, we make the following assumptions.

Assumption 7. *The second step parameter set Θ_2 is compact.*

Assumption 8. *For any $\theta\in\Theta_1$ and $\gamma\in\Theta_2$, there exists a measurable function $k(.)$ such that $|f(y_s,s\leq t;\theta,\gamma)|\leq k(y_s,s\leq t)$ with $\mathbb{E}[k(y_s,s\leq t)]<\infty$.*

Assumption 9. *For any $(vec(\Phi)',vec(\Xi)',vec(\Psi_{1:m})')'\in\Theta_1\times\Theta_2$, the $pm\times pm$ matrix*

$$\Lambda=\begin{pmatrix}\Phi+\Xi & -\Xi\Psi_1 & -\Xi\Psi_2 & \cdots & -\Xi\Psi_m \\ I_p & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & I_p & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \cdots & \cdots & \vdots \\ \mathbf{0} & \cdots & \cdots & I_p & \mathbf{0}\end{pmatrix},$$

satisfies $\|\Lambda\|_s < 1$.

Assumption 9 is key to control for the first step estimator and allows for establishing

$$\sup_{\gamma \in \Theta_2} |\mathbb{L}_T(\underline{y}; \hat{\theta}, \gamma) - \mathbb{L}_T(\underline{y}; \theta_0^*, \gamma)| = o_p(1),$$

where $\hat{\theta}$ is a sequence in Θ_1 that tends to θ_0^* in probability.

Theorem 5. *Let $\hat{\beta} = (\hat{\theta}', \hat{\gamma}')'$ be a sequence of penalised OLS based and non-penalised OLS based estimators. Then, under the assumptions of Theorem 1 and assumptions 7-9, $\hat{\beta} \xrightarrow[T \to \infty]{\mathbb{P}} (\theta_0^{*'}, \gamma_0')'$.*

We now turn to the asymptotic distribution of the second step estimator, conditionally on $\hat{\theta}$.

Theorem 6. *Under the assumptions of Theorem 4 together with its conditions on the penalisation rate on λ_T and the penalty function's behaviour, under assumptions 7-9, then*

$$\sqrt{T} \begin{pmatrix} (\hat{\theta} - \theta_0)_{\mathcal{A}} \\ \hat{\gamma} - \gamma_0 \end{pmatrix} \xrightarrow[T \to \infty]{d} \mathcal{N}_{\mathbb{R}^{dim}}(\mathbf{0}, \mathbf{J}^{-1}\mathbf{I}\mathbf{J}^{-1}),$$

with $dim = k_0 + p(1 + 2p)$ and the variance-covariance is composed with

$$\mathbf{J} \;\; = \;\; \begin{pmatrix} \mathbb{E}[\nabla^2_{\theta\theta'} \ell(y_s, s \leq t; \theta_0)]_{\mathcal{A}\mathcal{A}} & 0 \\ \mathbb{E}[\nabla^2_{\theta\gamma'} f(y_s, s \leq t; \theta_0, \gamma_0)]_{\mathcal{A}\bullet} & \mathbb{E}[\nabla^2_{\gamma\gamma'} f(y_s, s \leq t; \theta_0, \gamma_0)] \end{pmatrix}$$

and

$$\mathbf{I} \;\; = \;\; \mathbb{E}\Big[\begin{pmatrix} \nabla_\theta \ell(y_s, s \leq t; \theta_0)_{\mathcal{A}} \nabla_{\theta'} \ell(y_s, s \leq t; \theta_0)_{\mathcal{A}} & \nabla_\theta \ell(y_s, s \leq t; \theta_0)_{\mathcal{A}} \nabla_\gamma f(y_s, s \leq t; \theta_0, \gamma_0) \\ \nabla_\gamma f(y_s, s \leq t; \theta_0, \gamma_0) \nabla_{\theta'} \ell(y_s, s \leq t; \theta_0)_{\mathcal{A}} & \nabla_\gamma f(y_s, s \leq t; \theta_0, \gamma_0) \nabla_{\gamma'} f(y_s, s \leq t; \theta_0, \gamma_0) \end{pmatrix} \Big].$$

Remark. Theorem 6 deserves some comments.

(i) The notation $\mathbb{E}[\nabla^2_{\theta\gamma'} f(y_s, s \leq t; \theta_0, \gamma_0)]_{\mathcal{A}\bullet}$ means that the expectation is restricted only to block $\mathcal{A}$ when evaluating the partial derivative with respect to $\theta \in \Theta_1$.

(ii) The effect of the two-step estimation can be measured through the asymptotic variance-covariance matrix. As a by-product, simple calculations provide the asymptotic variances

$\mathbb{V}(.)$ of $\hat{\theta}$ and $\hat{\gamma}$: with obvious notations on the block locations in J and I, $\mathbb{V}(\hat{\theta}) = J_{11}^{-1} I_{11} J_{11}^{-1}$, and

$$\mathbb{V}(\hat{\gamma}) = J_{22}^{-1} I_{22} J_{22}^{-1} - \Upsilon I_{12} J_{22}^{-1} - J_{22}^{-1} I_{21} \Upsilon' + \Upsilon I_{11} \Upsilon', \; \Upsilon := J_{22}^{-1} J_{21} J_{11}^{-1}.$$

4 Empirical analysis

4.1 Simulation experiment

In this section, we empirically investigate the ability of the proposed penalisation method to better capture complex variance-covariance processes. We simulate the p-dimensional stochastic process (ϵ_t) based on two data generating processes: the multivariate ARCH and the BEKK processes. For the multivariate ARCH with q^* lags - M-ARCH(q^*) in the rest of the paper - case, we consider

$$\begin{cases} \epsilon_t &= H_t^{1/2} \eta_t, \\ H_t &= \Omega + \sum_{k=1}^{q^*} (I_p \otimes \epsilon_{t-k}') A_k (I_p \otimes \epsilon_{t-k}), \end{cases}$$

where q^* is the number of lagged matrices being functions of ϵ_{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 $\mathcal{U}([0.01, 0.05])$ and the off-diagonal ones from $\mathcal{U}([-0.01, 0.01])$ under the ordering constraint $\forall k \geq 2, \forall i, j$, $|A_{k,ij}| \leq |A_{k-1,ij}|$. As for the matrix Ω, the diagonal and off-diagonal elements are simulated from $\mathcal{U}([0.1; 0.2])$ and $\mathcal{U}([-0.01, 0.01])$ respectively. As for the BEKK process, the data generating process is based on

$$\begin{cases} \epsilon_t &= H_t^{1/2} \eta_t, \\ H_t &= \Omega + A \epsilon_{t-1} \epsilon_{t-1}' A' + B H_{t-1} B', \end{cases}$$

where A, B are $p \times p$ matrices, satisfying the stationarity constraint $\|D_p^{+}\{(A \otimes A) + (B \otimes B)\} D_p\|_s < 1$, where D_p is the duplication matrix and D_p^{+} the elemination matrix (see subsection 11.3 "Stationarity of VEC and BEKK Models" of Francq and Zakoïan (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 $\mathcal{U}([-0.8, 0.8])$. The matrix Ω 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 parameterizations. In both proposed dynamics, we initialize the observations $(\epsilon_k, \cdots, \epsilon_1)$ with centered and unit variance multivariate Gaussian distribution, where $k = q^*$ in the M-ARCH model and $k = 1$ in the BEKK. Then conditionally on the past k observations, we generate H_t and thus ϵ_t according to a centered multivariate Gaussian distribution with variance-covariance H_t.

We consider the problem sizes, $p = 15, 50$, and $T = 2000$ 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 $p = 50$. Then 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 corresponding to process (25) with scalar matrix parameters 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 $\hat{H}_t^{dcc}$. 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 $\hat{H}_t^{ccc}$. More details on the DCC and CCC can be found in Appendix D.

Regarding our proposed variance-covariance dynamic, the penalised OLS-MSV, denoted as $\hat{H}_t^{ols,al}$ for the adaptive Lasso OLS-MSV, $\hat{H}_t^{ols,br}$ for the Bridge OLS-MSV, $\hat{H}_t^{ols,scad}$ for the SCAD OLS-MSV, $\hat{H}_t^{ols,mcp}$ for the MCP OLS-MSV, and the non-penalised version of the OLS-MSV denoted as $\hat{H}_t^{ols}$. We considered the same technique as in Zou (2006) for the adaptive Lasso, where we selected $\gamma = 1.5$ as the power entering in the stochastic weights and the first step estimator is the unpenalised OLS estimator. 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 as $m = 5$ for a dimension $p = 50$. In the BEKK case, due to the autoregressive nature of the latter dynamic, we selected $m = 30$ when $p = 15$ and $m = 15$ when the dimension is $p = 50$.

We compare the true variance-covariance process and the estimated variance-covariance processes through the aforementioned models. To do so, we specify a matrix distance, namely the Frobenius norm, defined as $||A - B||_F := \sqrt{\text{Trace}((A - B)'(A - B))}$. We compute the previous norm for each t and for $A = H_t$ and

$$B \in \{\hat{H}_t^{dcc}, \hat{H}_t^{ccc}, \hat{H}_t^{ols}, \hat{H}_t^{ols,br}, \hat{H}_t^{ols,al}, \hat{H}_t^{ols,scad}, \hat{H}_t^{ols,mcp}\}.$$

We take the average of those quantities over $T = 2000$ 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 e.g. Hastie and al. (2015, Chap. 2) - , we selected the regularisation parameter and emphasize that the standard CV developed for i.i.d. data can not 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.

Clearly, the relevance of the penalisation procedure reported in Table 1 for the M-ARCH(q^*) based DGP and Table 2 for the BEKK based DGP increases with the size p, for any penalisation method. Moreover, DCC/CCC models are always beaten by the penalised specifications, whatever p, whereas, interestingly, it is not the case when compared with the non-penalised MSV. Our results emphasize the clear gain when considering penalisation, especially in the case m, p large as in the BEKK case: the autoregressive nature of such process fostered the use of a large number of lags, that is a large m.

4.2 Application to real data

To assess the relevance of the proposed penalised method, we propose a real data experiment. To do so, we compare the forecasting performances of the covariance matrices H_t for a portfolio of daily financial returns composed of the MSCI stock index for the following 23 countries: Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Greece, Hong Kong, Ireland, Italy, Japan, Netherlands, New Zealand, Norway, Portugal, Singapore, Spain, Sweden, Switzerland, the United-Kingdom, the United-States. We focus on direct out-of-sample evaluation methods, which allow for pairwise 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, then the minimum variance portfolio should emphasize such a shortcoming, compared to other models. Then, consider an investor who allocates a fixed amount between p stocks, according to a minimum-variance strategy and independently at each time t. At each date t, he/she solves

$$\min_{w_t} \; w_t' H_t w_t, \;\; \text{s.t.} \;\; \iota' w_t = 1, \tag{18}$$

where w_t is the $p \times 1$ vector of portfolio weights chosen at (the end of) time $t-1$, ι is a $p \times 1$ vector of 1 and H_t is the estimated conditional covariance matrix of the asset returns at time t. They are deduced from some dynamics that have been estimated on the sub-sample December 1998 - November 2015. Once the latter process is estimated in-sample, out-of-sample predictions are plugged into the program (18) between December 2015 and March 2018. The solution of (18) is given by the global minimum variance portfolio $w_t = H_t^{-1}\iota / \iota' H_t^{-1} \iota$.

Engle and Colacito (2006) show that the realized portfolio volatility is the smallest one when the variance-covariance matrices are correctly specified. As a consequence, 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 one. To do so, we consider a sequence of minimum variance portfolio weights $(w_{i,t})$ and $(w_{j,t})$, depending on the model. Then, we consider a distance based on the difference of the squared returns of the two portfolios, defined as $u_{ij,t} = \left\{w_{i,t}'\epsilon_t\right\}^2 - \left\{w_{j,t}'\epsilon_t\right\}^2$. The portfolio variances are the same if the predicted covariance matrices are the same. Thus we test the null hypothesis $\mathcal{H}_0 : \mathbb{E}\left[u_{ij,t}\right] = 0$ 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 $\mathcal{H}_0 : \alpha = 0$. If the mean of $u_{ij,t}$ is significantly positive (resp. negative), then the forecasts given by the covariance matrices of model j (resp. i) are preferred.

We run the latter test to compare the scalar DCC (DCC), the Orthogonal GARCH (O-G), the BEKK (BEKK), our OLS-MSV method (MSV) together with its penalised counterpart (denoted as MSV-AL, MSV-BR, MSV-SCA, MSV-MCP for the adaptive Lasso, Bridge, SCAD, MCP respectively). The definition of the BEKK and O-GARCH processes are reported in Appendix D. The matrix forecast comparisons are provided in Table 3. The results emphasises that the proposed penalised OLS-MSV method outperforms the MGARCH based competitors. Interestingly, fostering sparsity yields to much better forecasting performances. Indeed, the non-penalised OLS-MSV is outperformed by all alternative specifications, whereas the sparsity based specifications outperform both MGARCH based models and the non-penalised OLS-MSV.

5 Conclusion

The focus of this paper was devoted to 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 penalised OLS framework for state-space estimation. The corresponding large sample properties of the two-step estimator are derived for a broad range of penalty functions. The performances of our proposed method compared to 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 penalisation, which manages the overfitting problem.

Various issues and extensions can be further considered. Our proposed model could be extended to a dynamic correlation setting and include long memory and/or asymmetry, as discussed in Asai, McAleer, and Yu (2006) and Chin, Omori, and Asai (2009). Another direction would consists in modelling 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 state-space based setting.

References

Abadir, K.M. and Magnus, J.R. (2005), *Matrix algebra.* Cambridge University Press.

Alexander, C. (2001). "Orthogonal GARCH" in Alexander, C. (Ed.), *Mastering Risk*, Financial Times-Prentice Hall, London, pp. 21-28.

Asai, M., M. Caporin, M. McAleer (2015), "Forecasting Value-at-Risk Using Block Structure Multivariate Stochastic Volatility", *International Review of Economics & Finance*, **40**, 40–50.

Asai, M. and M. McAleer (2009), "Multivariate Stochastic Volatility, Leverage and News Impact Surfaces", *Econometrics Journal*, **12**, 292–309.

Asai, M., M. McAleer, and J. Yu (2006), "Multivariate Stochastic Volatility: A Review", *Econometric Reviews*, **25**, 145–175.

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

Bauwens L., S. Laurent, and J. K. V. Rombouts (2006), "Multivariate GARCH Models: A Survey", *Journal of Applied Econometrics*, **21**, 79–109.

Billingsley, P. (1961), "The Lindeberg-Levy theorem for martingales", *Proceedings of the American Mathematical Society*, **12**, 788–792.

Billingsley (1995), *Probability and measure*, New York: John Wiley&Sons.

Boussama, F. (2006), "Ergodicité des chaînes de Markov à valeurs dans une variété algébrique: application aux modèles GARCH multivariés", *Comptes Rendus de l'Académie des Sciences Paris*, 343, 275–278.

Chernozhukov, V. and Hong, H. (2004), "Likelihood estimation and inference in a class of nonregular econometric models", *Econometrica*, **72**, No. 5, 1445-1480.

Chernozhukov, V., (2005), "Extremal quantile regression", *The Annals of Statistics*, **33**, No. 2, 806-839.

Chib, S., F. Nardari, and N. Shephard (2006), "Analysis of High Dimensional Multivariate Stochastic Volatility Models", *Journal of Econometrics* **134**, 341–371.

Chib, S., Y. Omori, and M. Asai (2009), "Multivariate Stochastic Volatility", In: Andersen, T.G., R.A. Davis, J.P. Kreiss, and T. Mikosch, T. (Eds.), *Handbook of Financial Time Series*, New York: Springer-Verlag, pp.365–400.

Durbin, J., and S.J. Koopman (1997), "Monte Carlo Maximum Likelihood Estimation for Non-Gaussian State Space Models", *Biometrika*, **84**, 669–684.

Durbin, J., and S.J. Koopman (2001), *Time Series Analysis by State Space Methods*, Oxford: Oxford University Press.

Engle, R. F. (1982), "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation", *Econornetrica*, **50**, 987–1007.

Engle, R. (2002). "Dynamic Conditional Correlation", *Journal of Business & Economic Statistics* **20**: 339-350.

Engle, R., and R. Colacito. (2006). "Testing and Valuing Dynamic Correlations for Asset Allocation", *Journal of Business & Economic Statistics* **24**: 238-253.

Engle, R.F. and Kroner, K.F. (1995). "Multivariate Simultaneous Generalized ARCH", *Econometric Theory* 11: 122-150.

Engle, R.F., O. Ledoit, and M. Wolf (2017), "Large Dynamic Covariance Matrices", *Journal of Business & Economic Statistics*, **37**, 363–375.

Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalised likelihood and its oracle properties", *Journal of the American Statistical Association*, **96**, No. 456, 1348-1360.

Fan, J., Y. Fan, and J. Lv. (2008). "Large dimensional covariance matrix estimation using a factor model", *Journal of Econometrics*, **147**, 186–197.

Francq, C., and Zakoïan, J.-M. (2010), *GARCH models structure, statistical inference and financial applications.* John Wiley and Sons, Chichester, West Sussex, U.K.

Fu, W. J. (1998). "penalised regressions: The bridge versus the Lasso." Journal of Computational and Graphical Statistics, **7**, 397–416.

Gouriéroux, C. (1997), "ARCH Models and financial Applications". Springer.

Granger, C. W. J. and M. Morris (1976), "Time Series Modeling and Interpretation", *Journal of the Royal Statistical Society, Series A*, **139**, 246–257.

Ghysels, E., A. C. Harvey, and E. Renault (1996), "Stochastic volatility", In: Rao, C. R., Maddala, G. S., eds. *Statistical Models in Finance (Handbook of Statistics)*, Amsterdam: North-Holland, pp. 119–191.

Hannan, E. J. and L. Kavalieris (1984), "Multivariate Linear Time Series Models", *Advances in Applied Probability*,**16**, 492–561.

Hannan, E. J. and J. Rissanen (1982), "Recursive Estimation of Mixed Autoregressive-Moving Average Order", *Biometrika*, **69**, 81–94.

Harvey, A. (1998), "Long Memory in Stochastic Volatility", In: Knight, J. and S. Satchell (eds.), *Forecasting Volatility in Financial Markets*, Oxford: Butterworth-Haineman, 307–320.

Harvey, A. C., E. Ruiz, and N. Shephard (1994), "Multivariate Stochastic Variance Models", *Review of Economic Studies*, **61**, 247–264.

Hastie,T., R. Tibshirani, and Wainwright, M. (2015), "Statistical Learning with Sparsity: The Lasso and Generalizations", *Monographs on Statistics and Applied Probability 143.* Chapman and Hall.

Hjort, N. and Pollard, D. (1993), "Asymptotics for minimisers of convex processes", *Statistical Research Report, Institute of Mathematics, University of Oslo.*

Kastner, G., S. Frü-Schnatter, and H. F. Lopes (2017), "Efficient Bayesian Inference for Multivariate Factor Stochastic Volatility Models", *Journal of Computational and Graphical Statistics*, **26**, 905–917.

Kim, J. and Pollard, D. (1990), "Cube root asymptotics", *The Annals of Statistics*, **18**, No. 1, 191-219.

Knight, K. and Fu, W. (2000), "Asymptotics for Lasso-type estimators", *The Annals of Statistics*, **28**, No. 5, 1356-1378.

Merlevède, F., Peligrad, M. and Rio, E. (2009). "Bernstein inequality and moderate deviations under strong mixing conditions", *Institute of Mathematical Statistics Collections, High Dimensional Probability*, **5**, 273–292.

Newey,W. K. and D. McFadden (1994), "Large Sample Estimation and Hypothesis Testing", In R. F. Engle and D. McFadden (Eds.), *Handbook of Econometrics, Volume 4*, 2111–2245. Amsterdam: Elsevier.

Poignard, B. and J.D. Fermanian (2019), “High-dimensional penalised ARCH processes”, To appear in *Econometric Reviews.*

Racine, J. (2000), “Consistent cross-validatory model-selection for dependent data: hv-block cross-validation”, *Journal of Econometrics.* 99: 39-61.

Tibshirani, R. (1996). “Regression shrinkage and selection via the Lasso”, *Journal of the Royal Statistical Society. Series B*, **58**, No. 1, pp. 267-288.

Tse, Y. K. and A. K. C. Tsui (2002), “A Multivariate Generalized Autoregressive Conditional Heteroscedasticity Model with Time-Varying Correlations”, *Journal of Business & Economic Statistics*, **20**, 351–361.

Umezu, Y., Shimizu, Y., Masuda, H. and Ninomiya, Y., (2018), ”AIC for the non-concave penalised likelihood method”, *Annals of the Institute of Statistical Mathematics*, **71**, No. 2, 247-274.

White, H. (2001), “Asymptotic Theory for Econometricians”, 2nd edition, Emerald, UK.

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**, No. 476, 1418-1429.

A Tables

Table 1: Average distance true/estimated covariance matrices - M-ARCH(q^*) (100 replications)

	$\hat{H}_t^{dcc}$	$\hat{H}_t^{ccc}$	$\hat{H}_t^{ols}$	$\hat{H}_t^{ols,br}$	$\hat{H}_t^{ols,al}$	$\hat{H}_t^{ols,scad}$	$\hat{H}_t^{mcp}$
$p = 15$	9.12	9.89	10.11	8.41	8.63	8.62	8.46
$p = 50$	13.27	14.25	15.37	12.13	12.44	12.66	12.54

Table 2: Average distance true/estimated covariance matrices - BEKK (100 replications)

	$\hat{H}_t^{dcc}$	$\hat{H}_t^{ccc}$	$\hat{H}_t^{ols}$	$\hat{H}_t^{ols,br}$	$\hat{H}_t^{ols,al}$	$\hat{H}_t^{ols,scad}$	$\hat{H}_t^{mcp}$
$p = 15$	14.76	14.91	18.06	14.24	14.07	14.38	14.12
$p = 50$	56.93	61.98	98.235	54.21	54.86	54.15	54.08

Table 3: Diebold Mariano Test of Multivariate Variance-Covariance models

	DCC	O-G	BEKK	MSV	MSV-AL	MSV-BR	MSV-SCA	MSV-MCP
DCC		$\mathbf{-2.351}^c$	-1.046	$\mathbf{-3.415}^c$	$\mathbf{7.671}^c$	$\mathbf{7.587}^c$	$\mathbf{7.673}^c$	$\mathbf{7.577}^c$
O-G	$\mathbf{2.351}^c$		$\mathbf{2.014}^b$	1.184	$\mathbf{6.339}^c$	$\mathbf{6.781}^c$	$\mathbf{6.451}^c$	$\mathbf{6.103}^c$
BEKK	1.046	$\mathbf{-2.014}^b$		$\mathbf{-2.092}^b$	$\mathbf{8.234}^c$	$\mathbf{8.175}^c$	$\mathbf{8.241}^c$	$\mathbf{8.164}^c$
MSV	$\mathbf{3.415}^c$	-1.184	$\mathbf{2.092}^b$		$\mathbf{1.895}^b$	$\mathbf{1.982}^b$	$\mathbf{2.445}^c$	$\mathbf{2.241}^b$
MSV-AL	$\mathbf{-7.671}^c$	$\mathbf{-6.339}^c$	$\mathbf{-8.234}^c$	$\mathbf{-1.895}^b$		-0.497	-0.196	-0.484
MSV-BR	$\mathbf{-7.587}^c$	$\mathbf{-6.340}^c$	$\mathbf{-8.175}^c$	$\mathbf{-1.982}^b$	0.497		$\mathbf{0.750}^a$	0.047
MSV-SCA	$\mathbf{-7.673}^c$	$\mathbf{-6.364}^c$	$\mathbf{-8.241}^c$	$\mathbf{-2.445}^c$	0.196	-0.750		-0.721
MSV-MCP	$\mathbf{-7.577}^c$	$\mathbf{-6.341}^c$	$\mathbf{-8.164}^c$	$\mathbf{-2.241}^c$	0.484	-0.047	0.721	

Table 4: This table reports the out-of-sample t-statistics of the Diebold-Mariano test that checks the equality between covariance matrix forecasts using the loss function $u_{ij,t}$ over the period December 2015 and March 2018. 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 favour of the model in the column. a, b, c: rejection of the null hypothesis at 10%, 5% and 1% respectively.

B Technical results

We used the *convexity argument* to derive the asymptotic distribution in the Lasso case. Chernozhukov and Huong (2004), Chernozhukhov (2005) use this convexity argument to obtain the asymptotic distribution of quantile regression type estimators. This argument relies on the convexity Lemma, which is a key result to obtain an asymptotic distribution when the objective function is not differentiable. It only requires the lower-semicontinuity and convexity of the empirical criterion. The convexity Lemma, as in Chernozhukov (2005), proof of Theorem 4.1, can be stated as follows.

Lemma 1. Convexity Lemma, Chernozhukov (2005)

Suppose

(i) a sequence of convex lower-semicontinuous $\mathbb{F}_T : \mathbb{R}^d \to \bar{\mathbb{R}}$ *marginally converges to* $\mathbb{F}_\infty : \mathbb{R}^d \to \bar{\mathbb{R}}$ *over a dense subset of* $\mathbb{R}^d$*;*

(ii) $\mathbb{F}_\infty$ *is finite over a nonempty open set* $E \subset \mathbb{R}^d$*;*

(iii) $\mathbb{F}_\infty$ *is uniquely minimized at a random vector* $\boldsymbol{u}_\infty$*.*

Then

$$\arg\min_{\boldsymbol{z}\in\mathbb{R}^d} \mathbb{F}_T(\boldsymbol{z}) \xrightarrow{d} \arg\min_{\boldsymbol{z}\in\mathbb{R}^d} \mathbb{F}_\infty(\boldsymbol{z}), \textit{ that is } \boldsymbol{u}_T \xrightarrow{d} \boldsymbol{u}_\infty.$$

As for the SCAD and MCP, due to the non-convexity of the penalty function, we used Lemma 3 of Umezu et al. (2018), which generalises Lemma 2 of Hjort and Pollard (1993) to the case of convex non-penalised loss functions with non-convex penalties. Lemma 2 of Hjort and Pollard (1993) allows for deriving consistency and asymptotic normality of estimators that are defined by minimisation of convex criterion functions.

Lemma 2. Umezu, Shimizu, Masuda, Ninomiya (2018)

Suppose that $\mathbb{G}_T(\boldsymbol{u})$ is a strictly convex random function that is approximated by $\tilde{\mathbb{G}}_T(\boldsymbol{u})$. Let $\bar{\mathbf{u}}$ be a subvector of $\boldsymbol{u}$, and let $\zeta(\boldsymbol{u})$ and $\eta(\bar{\boldsymbol{u}})$ be continuous functions such that $\zeta_T(\boldsymbol{u})$ and $\eta_T(\bar{\boldsymbol{u}})$ converge to $\zeta(\boldsymbol{u})$ and $\eta(\bar{\boldsymbol{u}})$ uniformly over $\boldsymbol{u}$ and $\bar{\mathbf{u}}$ in any compact set, respectively, and assume that $\zeta(\boldsymbol{u})$ is convex and $\eta(\mathbf{0}) = 0$. In addition, for

$$\nu_T(\boldsymbol{u}) = \mathbb{G}_T(\boldsymbol{u}) + \zeta_T(\boldsymbol{u}) + \eta_T(\bar{\boldsymbol{u}}), \text{ and } \tilde{\nu}_T(\boldsymbol{u}) = \tilde{\mathbb{G}}_T(\boldsymbol{u}) + \zeta(\boldsymbol{u}) + \eta(\bar{\boldsymbol{u}}),$$

let $\boldsymbol{u}_T$ and $\tilde{\boldsymbol{u}}_T$ be the arg min *of $\nu_T(\boldsymbol{u})$ and $\tilde{\nu}_T(\boldsymbol{u})$, respectively, and assume that $\tilde{\boldsymbol{u}}_T$ is unique and $\bar{\tilde{\boldsymbol{u}}}_T = 0$. Then, for any $\epsilon > 0, \delta > 0, \mu > \delta$, there exists $\gamma > 0$ such that*

$$\mathbb{P}(\|\boldsymbol{u}_T - \tilde{\boldsymbol{u}}_T\| \geq \delta) \leq \mathbb{P}(2\Delta_T(\delta) + \epsilon \geq \Upsilon_T(\delta)) + \mathbb{P}(\|\boldsymbol{u}_T - \tilde{\boldsymbol{u}}_T\| \geq \mu) + \mathbb{P}(\|\bar{\boldsymbol{u}}_T\| \geq \gamma),$$

where

$$\Delta_T(\delta) = \sup_{\boldsymbol{u}:\|\boldsymbol{u}-\tilde{\boldsymbol{u}}_T\|\leq\delta} |\nu_T(\boldsymbol{u}) - \tilde{\nu}_T(\boldsymbol{u})|, \ \Upsilon_T(\delta) = \inf_{\boldsymbol{u}:\|\boldsymbol{u}-\tilde{\boldsymbol{u}}_T\|=\delta} |\tilde{\nu}_T(\boldsymbol{u}) - \tilde{\nu}_T(\tilde{\boldsymbol{u}}_T)|.$$

Finally, for the large sample distribution of the Bridge penalised estimator, we relied on Theorem 2.7 of Kim and Pollard (1990).

Theorem 7. Kim and Pollard (1990)

Let $\{\mathbb{F}_T\}$ be a random function into the space of all locally bounded real functions on $\mathbb{R}^d$, and $\boldsymbol{u}_T$ random mapps into $\mathbb{R}^d$ such that

(i) $\mathbb{F}_T \xrightarrow{d} Q$ *for a Borel measure Q concentrated on $C_{\max}(\mathbb{R}^d)$*[1]*;*

(ii) $\boldsymbol{u}_T = O_p(1)$*;*

(iii) $\mathbb{F}_T(\boldsymbol{u}_T) \geq \sup_{\boldsymbol{u}} \{\mathbb{F}_T(\boldsymbol{u})\} - \alpha_T$ *for random variables (α_T) of order $o_p(1)$.*

Then $\boldsymbol{u}_T \xrightarrow{d} \arg\max_{\boldsymbol{u}} \{\mathbb{F}(\boldsymbol{u})\}$ *for a $\mathbb{F}(\boldsymbol{u})$ with distribution Q.*

[1] See their page 195 for a definition of this set

C Proofs

Proof of Theorem 1. In a first step, we prove the uniform convergence of $\mathbb{G}_T^{pen}(\underline{y};.)$ to the limit quantity $\mathbb{G}_\infty^{pen}(\underline{y};.)$ on any compact set $\boldsymbol{B} \subset \Theta$, idest

$$\sup_{\boldsymbol{x}\in\boldsymbol{B}} |\mathbb{G}_T^{pen}(\underline{y};\boldsymbol{x}) - \mathbb{G}_\infty^{pen}(\underline{y};\boldsymbol{x})| \xrightarrow[T\to\infty]{\mathbb{P}} 0. \tag{19}$$

We define $\mathcal{C} \subset \Theta$ an open convex set and pick $\boldsymbol{x} \in \mathcal{C}$. Then, under assumptions 2 and 4, by Lemma 2.4 of Newey and McFadden (1994), we have

$$\sup_{\boldsymbol{x}\in\boldsymbol{B}} |\mathbb{G}_T(\underline{y};\theta) - \mathbb{E}[\ell(\theta; y_s, s \leq t)]| = 0,$$

where the convergence of the sample criterion to the population level criterion is ensured by the ergodic Theorem of Billingsley (1995), which thus justifies assumption 3. Thus, using $\lambda_T/T \to \lambda_0$ and since the parameter is taken over a compact set, we obtain

$$\sup_{\boldsymbol{x}\in\boldsymbol{B}} |\mathbb{G}_T^{pen}(\underline{y};\boldsymbol{x}) - \mathbb{G}_\infty^{pen}(\underline{y};\boldsymbol{x})| \xrightarrow[T\to\infty]{\mathbb{P}} 0.$$

Now we would like

$$\text{arg min}\,\{\mathbb{G}_T(\underline{y};.) + \sum_{i=1}^{d} \boldsymbol{p}(\frac{\lambda_T}{T}, |.|)\} \xrightarrow[T\to\infty]{\mathbb{P}} \text{arg min}\,\{\mathbb{G}_\infty(\underline{y};.) + \sum_{i=1}^{d} \boldsymbol{p}(\lambda_0, |.|)\}.$$

First, $\mathbb{G}_T(\underline{y};\boldsymbol{x}) + \sum_{i=1}^{d} \boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_i|) \geq \mathbb{G}_T(\underline{y};\boldsymbol{x})$, and $\underset{\boldsymbol{x}\in\boldsymbol{B}}{\text{arg min}}\ \{\mathbb{G}_T(\underline{y};\boldsymbol{x})\} = O_p(1)$ by convexity of the criterion, it thus follows that $\underset{\boldsymbol{x}\in\boldsymbol{B}}{\text{arg min}}\ \{\mathbb{G}_T(\underline{y};\boldsymbol{x}) + \sum_{i=1}^{d} \boldsymbol{p}(\frac{\lambda_T}{T}, |\boldsymbol{x}_i|)\} = O_p(1)$ with probability one. □

Proof of Theorem 2. Let $\nu_T = T^{-1/2} + \sqrt{card(\mathcal{A})}A_{1,T}$. We would like to prove that for any $\epsilon > 0$, there exists $C_\epsilon > 0$ such that

$$\mathbb{P}(\frac{1}{\nu_T}\|\hat{\theta} - \theta_0\| > C_\epsilon) < \epsilon.$$

Following the reasoning of Fan and Li (2001), Theorem 1, and denoting $\mathbb{G}_T^{pen}(\underline{y};\theta) = \mathbb{G}_T(\underline{y};\theta) + \sum_{i=1}^{d} \boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_i|)$, we have

$$\mathbb{P}(\frac{1}{\nu_T}\|\hat{\theta} - \theta_0\| > C_\epsilon) \leq \mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : \mathbb{G}_T^{pen}(\underline{y};\theta_0 + \nu_T \boldsymbol{u}) \leq \mathbb{G}_T^{pen}(\underline{y};\theta_0)),$$

which implies that there is a local minimum in the ball $\{\theta_0 + \nu_T \boldsymbol{u}, \|\boldsymbol{u}\|_2 \leq C_\epsilon\}$ so that the minimum $\hat{\theta}$ satisfies $\|\hat{\theta} - \theta_0\| = O_p(\nu_T)$. Now by a Taylor expansion of the penalised loss function, we obtain

$$\begin{aligned} \mathbb{G}_T^{pen}(\underline{y};\theta_0 + \nu_T \boldsymbol{u}) - \mathbb{G}_T^{pen}(\underline{y};\theta_0) &= \nu_T \boldsymbol{u}' \nabla_\theta \mathbb{G}_T(\underline{y};\theta_0) + \tfrac{\nu_T^2}{2} \boldsymbol{u}' \nabla^2_{\theta\theta'} \mathbb{G}_T(\underline{y};\theta_0) \boldsymbol{u} \\ &+ \sum_{i=1}^{d} \{\boldsymbol{p}(\tfrac{\lambda_T}{T}, |\theta_{0,i} + \nu_T \boldsymbol{u}_i|) - \boldsymbol{p}(\tfrac{\lambda_T}{T}, |\theta_{0,i}|)\}, \end{aligned}$$

since the third derivative vanishes. We want to prove

$$\begin{aligned} \mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon &: \boldsymbol{u}' \nabla_\theta \mathbb{G}_T(\underline{y};\theta_0) + \tfrac{\nu_T}{2} \boldsymbol{u}' \mathbb{H} \boldsymbol{u} + \tfrac{\nu_T}{2} \mathcal{R}_T(\theta_0) \\ &+ \nu_T^{-1} \sum_{i=1}^{d} \{\boldsymbol{p}(\tfrac{\lambda_T}{T}, |\theta_{0,i} + \nu_T \boldsymbol{u}_i|) - \boldsymbol{p}(\tfrac{\lambda_T}{T}, |\theta_{0,i}|)\} \leq 0) < \epsilon, \end{aligned} \tag{20}$$

where $\mathcal{R}_T(\theta_0) = \boldsymbol{u}' \{\nabla^2_{\theta\theta'} \mathbb{G}_T(\underline{y};\theta_0) - \mathbb{H}\} \boldsymbol{u}$. First, using the matrix derivatives formula of Abadir and Magnus (2005), we obtain for the score

$$\begin{aligned} \nabla_\theta \mathbb{G}_T(\underline{y};\theta_0) &= \tfrac{1}{T} \sum_{t=1}^{T} \nabla_\theta \ell(y_s, s \leq t; \theta_0) \\ &= -\tfrac{1}{T} \sum_{t=1}^{T} \big(Z_{m,t-1} \otimes \{x_t - \Psi_{0,1:m} Z_{m,t-1}\}\big). \end{aligned}$$

As for the Hessian, we aim at extracting the form $\text{tr}(L(\boldsymbol{d}\Lambda)' M(\boldsymbol{d}\Lambda))$ for L (resp. M) any square $m \times m$ matrix (resp. $p \times p$). We thus obtain

$$\nabla^2_{\theta\theta'} \mathbb{G}_T(\underline{y};\theta_0) = \frac{1}{T} \sum_{t=1}^{T} (Z_{m,t-1} Z'_{m,t-1} \otimes I_p).$$

Under assumption 3, by the Central Limit Theorem of Billingsley (1961)

$$\boldsymbol{u}' \nabla_\theta \mathbb{G}_T(\underline{y};\theta_0) = O_p(n^{-1/2} \boldsymbol{u}' \mathbb{M} \boldsymbol{u}),$$

with $\mathbb{M} = \mathbb{E}[\nabla_\theta \ell(\underline{y};\theta_0) \nabla_{\theta'} \ell(\underline{y};\theta_0)]$ assumed well defined by assumption 5. By the ergodic Theorem of Billingsley (1995), we have

$$\nabla_{\theta\theta'} \mathbb{G}_T(\underline{y};\theta_0) \underset{T \to \infty}{\xrightarrow{\mathbb{P}}} \mathbb{H},$$

which implies $\mathcal{R}_T(\theta_0) = o_p(1)$. We now focus on the penalty terms. First, we show that the penalty functions satisfy assumption 6. The Lasso satisfies

$$\forall i \in \mathcal{A}, \nabla_{\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i}|) = \frac{\lambda_T}{T}\text{sgn}(\theta_{0,i}),\ \nabla^2_{\theta_i\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i}|) = 0.$$

For $0 < q < 1$, the Bridge satisfies

$$\begin{array}{ll} \forall i \in \mathcal{A}, & \nabla_{\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i}|) = \frac{\lambda_T}{T}q|\theta_{0,i}|^{q-1}\text{sgn}(\theta_{0,i}), \\ & \nabla^2_{\theta_i\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i}|) = \frac{\lambda_T}{T}q(q-1)|\theta_{0,i}|^{q-2}\text{sgn}(\theta_{0,i}), \end{array}$$

thus the second order derivative of the Bridge converges to 0 when $\lambda_T = o(T)$. As for the scad, we have

$$\forall i \in \mathcal{A}, \nabla_{\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i}|) = \frac{\lambda_T}{T}\big(\mathbf{1}_{\{|\theta_{0,i}| \leq \frac{\lambda_T}{T}\}} + \frac{(b_{scad}\frac{\lambda_T}{T} - |\theta_{0,i}|)_+}{(b_{scad}-1)\frac{\lambda_T}{T}}\mathbf{1}_{\{|\theta_{0,i}| > \frac{\lambda_T}{T}\}}\big).$$

As a consequence, the scad penalty is twice continuously differentiable for $\frac{\lambda_T}{T} < |\theta_{0,i}|$, which implies that $\forall i \in \mathcal{A}, \nabla^2_{\theta_i\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i}|) = o(1)$. In the MCP case, we have

$$\forall i \in \mathcal{A}, \nabla_{\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i}|) = (\frac{\lambda_T}{T} - \frac{|\theta_{0,i}|}{b_{mcp}})\text{sgn}(\theta_{0,i})\mathbf{1}_{\{|\theta_{0,i}| \leq b_{mcp}\frac{\lambda_T}{T}\}}.$$

Under $\lambda_T = o(T)$, we straightforwardly obtain $\nabla^2_{\theta_i\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i}|) = o(1)$ for non-zero components. Now for any $i \in \mathcal{A} \subset \{1, \cdots, d\}$, and since the penalties are coordinate-separable, we have

$$\boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i}+\nu_T\boldsymbol{u}_i|) - \boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i}|) = \nu_T\boldsymbol{u}_i\text{sgn}(\theta_{0,i})\nabla_{\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i}|) + \frac{\nu_T^2}{2}\boldsymbol{u}_i^2\nabla^2_{\theta_i\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i}|)\big(1+o(1)\big).$$

Hence, using $\boldsymbol{p}(\frac{\lambda_T}{T}, 0) = 0$, we have

$$|\sum_{i\in\mathcal{A}1}\{\boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i}+\nu_T\boldsymbol{u}_i|) - \boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i}|)\}| \leq \nu_T\|\boldsymbol{u}\|_1 A_{1,T} + \frac{\nu_T^2}{2}\|\boldsymbol{u}\|_2^2 A_{2,T}\big(1+o(1)\big).$$

Using $\|\boldsymbol{u}\|_1 \leq \sqrt{card(\mathcal{A})}\|\boldsymbol{u}\|_2$, and under assumption 6, the third derivative being dominated, we obtain

$$|\sum_{i\in\mathcal{A}}\{\boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i}+\nu_T\boldsymbol{u}_i|) - \boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i}|)\}| \leq \nu_T\sqrt{card(\mathcal{A})}\|\boldsymbol{u}\|_2 A_{1,T} + \frac{\nu_T^2}{2}\|\boldsymbol{u}\|_2^2 A_{2,T}.$$

Then, denoting $\delta_T = \lambda_{\min}(\mathbb{H})C_\epsilon^2\nu_T$, and using $\frac{\nu_T}{2}\mathbb{E}[\boldsymbol{u}'\nabla^2_{\theta\theta'}\ell(\underline{y};\theta_0)\boldsymbol{u}] \geq \delta_T$, we deduce that (20) can be bounded as

$$\begin{array}{l}\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : \boldsymbol{u}'\nabla_\theta \mathbb{G}_T(\underline{y};\theta_0) + \frac{\nu_T}{2}\boldsymbol{u}'\mathbb{E}[\nabla^2_{\theta\theta'}\ell(\underline{y};\theta_0)]\boldsymbol{u} + \frac{\nu_T}{2}\mathcal{R}_T(\theta_0) \\ +\nu_T^{-1}\sum\limits_{i=1}^d\{\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}+\nu_T\boldsymbol{u}_i|) - \boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}|)\} \leq 0) \\ \leq \mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : \boldsymbol{u}'|\nabla_\theta \mathbb{G}_T(\underline{y};\theta_0)| > \delta_T/6) + \mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon|\frac{\nu_T}{2}\mathcal{R}_T(\theta_0)| > \delta_T/6) \\ \mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\sum\limits_{i=1}^d\{\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}+\nu_T\boldsymbol{u}_i|) - \boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}|)\}| > \nu_T\delta_T/6).\end{array}$$

We have for n and C_ϵ sufficiently large enough

$$\begin{array}{l}\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\sum\limits_{i=1}^d\{\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}+\nu_T\boldsymbol{u}_i|) - \boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}|)\}| > \nu_T\delta_T/6) \\ \leq \mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : \nu_T\sqrt{card(\mathcal{A})}\|\boldsymbol{u}\|_2 A_{1,T} + \frac{\nu_T^2}{2}\|\boldsymbol{u}\|_2^2 A_{2,T} > \nu_T\delta_T/6) < \epsilon/3.\end{array}$$

Moreover, if $\nu_T = n^{-1/2} + \sqrt{card(\mathcal{A})}A_{1,T}$, for C_ϵ large enough

$$\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : \boldsymbol{u}'|\nabla_\theta \mathbb{G}_T(\underline{y};\theta_0)| > \delta_T/6) \leq \frac{C_\epsilon^2 C_{st}}{T\delta_T^2} \leq \frac{C_{st}}{C_\epsilon^4} < \epsilon/3,$$

where $C_{st} > 0$ is a generic constant. Finally, using $\mathcal{R}_T(\theta_0) = o_p(1)$, we obtain

$$\begin{array}{l}\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : \boldsymbol{u}'|\nabla_\theta \mathbb{G}_T(\underline{y};\theta_0)| > \delta_T/6) + \mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon|\frac{\nu_T}{2}\mathcal{R}_T(\theta_0)| > \delta_T/6) \\ \mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_0+\nu_T\boldsymbol{u}|) - \boldsymbol{p}(\frac{\lambda_T}{T},|\theta_0|)| > \nu_T\delta_T/6) \\ \leq \frac{C_{st}}{C_\epsilon^4} + 2\epsilon/3 \leq \epsilon,\end{array}$$

for T and C_ϵ large enough. We deduce $\|\hat{\theta} - \theta_0\| = O_p(\nu_T)$. □

Proof of Theorem 3. Let $\boldsymbol{u} \in \mathbb{R}^d$ such that $\theta = \theta_0 + \boldsymbol{u}/\sqrt{T}$, and the empirical criterion $\mathbb{F}_T(\boldsymbol{u}) = T\{\mathbb{G}_T^{pen}(\underline{y};\theta) - \mathbb{G}_T^{pen}(\underline{y};\theta_0)\}$. Note that $\mathbb{F}_T(\boldsymbol{u})$ is minimized at $\hat{u}_T = T^{1/2}(\hat{\theta} - \theta_0)$ because $\hat{\theta}$ minimizes $\mathbb{G}_T^{pen}(\boldsymbol{Z};\theta)$. Thus $\hat{u}_T = \arg\min\limits_{\boldsymbol{u}\in\mathbb{R}^d}\{\mathbb{F}_T(\boldsymbol{u})\}$.

We first establish the finite distributional convergence of $\mathbb{F}_T(.)$ to $\mathbb{F}_\infty(.)$. Then we separate the proof depending on what penalty function we consider. We have the expansion

$$\mathbb{F}_T(\boldsymbol{u}) = \sqrt{T}\boldsymbol{u}'\nabla_\theta\mathbb{G}_T(\underline{y};\theta_0) + \frac{1}{2}\boldsymbol{u}'\nabla^2_{\theta\theta'}\mathbb{G}_T(\underline{y};\theta_0)\boldsymbol{u} + T\sum_{i=1}^d\{\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}+\boldsymbol{u}_i/\sqrt{T}|) - \boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}|)\}.$$

By assumption 3, by the Central Limit Theorem of Billingsley (1961) and the ergodic Theorem of Billingsley (1995)

$$\sqrt{T}\nabla_\theta \mathbb{G}_T(\underline{y};\theta_0) \xrightarrow[T\to\infty]{d} \mathcal{N}_{\mathbb{R}^d}(0,\mathbb{M}), \ \ \nabla^2_{\theta\theta'}\mathbb{G}_T(\underline{y};\theta_0) \xrightarrow[T\to\infty]{\mathbb{P}} \mathbb{H}.$$

As for the penalty terms, in the MCP and SCAD cases, we proceed as follows. We have

$$T\sum_{i=1}^{d}\{\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}+\boldsymbol{u}_i/\sqrt{T}|)-\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}|)\} = \zeta_T(\boldsymbol{u})+\eta_T(\boldsymbol{u}),$$

where using the coordinate-separability property of the penalties

$$\zeta_T(\boldsymbol{u}) = T\sum_{i\in\mathcal{A}}\{\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}+\boldsymbol{u}_i/\sqrt{T}|)-\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}|)\}, \ \ \eta_T(\boldsymbol{u}) = T\sum_{i\in\mathcal{A}^c}\{\boldsymbol{p}(\frac{\lambda_T}{T},\boldsymbol{u}_i/\sqrt{T})\}.$$

Then we have by a Taylor expansion for the indices $i\in\mathcal{A}$

$$\zeta_T(\boldsymbol{u}) = \sum_{i\in\mathcal{A}}\sqrt{T}\nabla_{\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}|)\boldsymbol{u}_i\text{sgn}(\theta_{0,i}) + \nabla^2_{\theta_i\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}|)\boldsymbol{u}_i^2/2(1+o(1)).$$

Under assumption 6, we have $A_{2,T}=o(1)$. We then need to treat the first order term. For both SCAD and MCP, since their derivatives respectively vanish outside $[-b_{scad}\frac{\lambda_T}{T}, b_{scad}\frac{\lambda_T}{T}]$, $[-b_{mcp}\frac{\lambda_T}{T}, b_{mcp}\frac{\lambda_T}{T}]$ we have

$$\begin{array}{rcl}\nabla_{\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}|) & = & \nabla_{\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}|)\mathbf{1}_{\{|\theta_{0,i}|\leq b_{scad}\frac{\lambda_T}{T}\}},\\ \nabla_{\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}|) & = & \nabla_{\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}|)\mathbf{1}_{\{|\theta_{0,i}|\leq b_{mcp}\frac{\lambda_T}{T}\}},\end{array}$$

which implies for any $\epsilon>0$ and $i\in\mathcal{A}$ that

$$\begin{array}{rcl}\mathbb{P}(\sqrt{T}\nabla_{\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}|)\mathbf{1}_{\{|\theta_{0,i}|\leq b_{scad}\frac{\lambda_T}{T}\}}>\epsilon) & \leq & \mathbb{P}(|\theta_{0,i}|\leq b_{scad}\frac{\lambda_T}{T})\to 0,\\ \mathbb{P}(\sqrt{T}\nabla_{\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}|)\mathbf{1}_{\{|\theta_{0,i}|\leq b_{mcp}\frac{\lambda_T}{T}\}}>\epsilon) & \leq & \mathbb{P}(|\theta_{0,i}|\leq b_{mcp}\frac{\lambda_T}{T})\to 0.\end{array}$$

As a consequence, $\sqrt{T}\nabla_{\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}|)=o_p(1)$. This is a direct consequence of the unbiasedness property when regularising large coefficients. Thus $\zeta_T(\boldsymbol{u})\to 0$ as $T\to\infty$. As for $i\in\mathcal{A}^c$, we have

$$\eta_T(\boldsymbol{u}) = \sum_{i\in\mathcal{A}^c}\sqrt{T}\big(\nabla_{\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_i|)\big)_{\theta_i=0}|\boldsymbol{u}_i| + \frac{1}{2}\big(\nabla^2_{\theta_i\theta_i}(\frac{\lambda_T}{T},|\theta_i|)\big)_{\theta_i=0}\boldsymbol{u}_i^2(1+o(1)),$$

Based on the assumption that $\lim\limits_{x\to 0^+} \nabla_x \boldsymbol{p}(\frac{\lambda_T}{T}, x) = \frac{\lambda_T}{T}$ and $\lambda_T = O(\sqrt{T})$, we deduce

$$\eta_T(\boldsymbol{u}) \to \lambda_0 \sum_{i \in \mathcal{A}^c} |\boldsymbol{u}_i|.$$

As a consequence, by Lemma 2, where in the latter we take $\mathbb{G}_T(\boldsymbol{u}) = \tilde{\mathbb{G}}_T(\boldsymbol{u}) = \sqrt{T}\nabla_\theta \mathbb{G}_T(\underline{y};\theta_0)\boldsymbol{u} + \frac{1}{2}\boldsymbol{u}'\nabla^2_{\theta\theta'}\mathbb{G}_T(\underline{y};\theta_0)\boldsymbol{u}$, we obtain $\arg\min\limits_{\boldsymbol{u}}\{\mathbb{F}_T\} \xrightarrow{d} \arg\min\limits_{\boldsymbol{u}}\{\mathbb{F}_\infty\}$.

For the Bridge estimator, we have for any index $i \in \mathcal{A} \cup \mathcal{A}^c$ and under the rate $\lambda_T/T^{q/2} \to \lambda_0$, then

$$\begin{aligned} & T\{\boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i} + \boldsymbol{u}_i/\sqrt{T}|) - \boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i}|)\} \\ = \quad & \lambda_T \sum_{i=1}^{d} \{|\theta_{0,i} + \boldsymbol{u}_i/\sqrt{T}|^q - |\theta_{0,i}|^q\} \to \lambda_0 \sum_{k=1}^{d} |\boldsymbol{u}_i|^q \mathbf{1}_{\theta_{0,i}=0}. \end{aligned}$$

Now we need to prove that $\arg\min\limits_{\boldsymbol{u}}\{\mathbb{F}_T\} \xrightarrow{d} \arg\min\limits_{\boldsymbol{u}}\{\mathbb{F}_\infty\}$ for any $\boldsymbol{u}$ and n sufficiently large. To do so, we use Theorem 7 of Kim and Pollard (1990) (see Appendix B) and show $\arg\min\limits_{\boldsymbol{u}}\{\mathbb{F}_T\} = O_p(1)$. We first have the expansion

$$\begin{aligned} \mathbb{F}_T(\boldsymbol{u}) \quad &= \quad \mathbb{G}_T(\underline{y};\theta_0 + \boldsymbol{u}/\sqrt{T}) - \mathbb{G}_T(\underline{y};\theta_0) + \lambda_T \sum_{i=1}^{d} \{|\theta_{0,i} + \boldsymbol{u}_i/\sqrt{T}|^q - |\theta_{0,i}|^q\} \\ &\geq \quad \mathbb{G}_T(\underline{y};\theta_0 + \boldsymbol{u}/\sqrt{T}) - \mathbb{G}_T(\underline{y};\theta_0) - \lambda_T \sum_{i=1}^{d} |\boldsymbol{u}_i/\sqrt{T}|^q \\ &\geq \quad \mathbb{G}_T(\underline{y};\theta_0 + \boldsymbol{u}/\sqrt{T}) - \mathbb{G}_T(\underline{y};\theta_0) - \big(\lambda_0 + \epsilon\big) \sum_{i=1}^{d} |\boldsymbol{u}_i/\sqrt{T}|^q := \tilde{\mathbb{F}}_T(\boldsymbol{u}), \end{aligned}$$

where, following the proof of Theorem 3 of Knight and Fu (2000), ϵ is such that $\lambda/T^{q/2} \leq \lambda_0 + \epsilon$. Then, expanding $\mathbb{G}_T(\underline{y};\theta_0 + \boldsymbol{u}/\sqrt{T}) - \mathbb{G}_T(\underline{y};\theta_0)$ in $\tilde{\mathbb{F}}_T(\boldsymbol{u})$, we have

$$\tilde{\mathbb{F}}_T(\boldsymbol{u}) = \sqrt{T}\boldsymbol{u}'\nabla_\theta \mathbb{G}_T(\theta) + \frac{1}{2}\boldsymbol{u}'\nabla^2_{\theta\theta'}\mathbb{G}_T(\theta)\boldsymbol{u} - \big(\lambda_0 + \epsilon\big)\sum_{i=1}^{d} |\boldsymbol{u}_i/\sqrt{T}|^q.$$

The second term, which is quadratic in $\boldsymbol{u}$, dominates the term with $|\boldsymbol{u}_i|^q$. Hence $\arg\min\limits_{\boldsymbol{u}}\{\tilde{\mathbb{F}}_T\} = O_p(1)$, which in turns implies $\arg\min\limits_{\boldsymbol{u}}\{\mathbb{F}_T\} = O_p(1)$. We then obtain for the Bridge

$$\sqrt{T}(\hat{\theta} - \theta_0) = \arg\min_{\boldsymbol{u}}\{\mathbb{F}_T\} \xrightarrow{d} \arg\min_{\boldsymbol{u}}\{\mathbb{F}_\infty\}.$$

Finally, for the Lasso estimator, we have

$$
\begin{aligned}
& T\{\boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i} + \boldsymbol{u}_i/\sqrt{T}|) - \boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_{0,i}|)\} \\
= \quad & \lambda_T \sum_{k=1}^{d} \{|\theta_{0,i} + \boldsymbol{u}_i/\sqrt{T}| - |\theta_{0,i}|\} \to \lambda_0 \sum_{k=1}^{d} \big(\boldsymbol{u}_i \text{sgn}(\theta_{0,i}) \mathbf{1}_{\theta_{0,i} \neq 0} + |\boldsymbol{u}_i| \mathbf{1}_{\theta_{0,i}=0}\big).
\end{aligned}
$$

We thus proved that $\mathbb{F}_T(\boldsymbol{u}) \xrightarrow{d} \mathbb{F}_\infty(\boldsymbol{u})$, for a fixed $\boldsymbol{u}$. Let us observe that

$$
\boldsymbol{u}_T^* = \arg \min_{\boldsymbol{u}} \{\mathbb{F}_T(\boldsymbol{u})\},
$$

and $\mathbb{F}_T(.)$ admits as a minimizer $\boldsymbol{u}_T^* = \sqrt{T}(\hat{\theta} - \theta_0)$. As $\mathbb{F}_T$ is convex and $\mathbb{F}_\infty$ is continuous, convex and has a unique minimum, then by the convexity Lemma 1, we obtain

$$
\sqrt{T}(\hat{\theta} - \theta_0) = \arg \min_{\boldsymbol{u}} \{\mathbb{F}_T\} \xrightarrow{d} \arg \min_{\boldsymbol{u}} \{\mathbb{F}_\infty\}.
$$

□

Proof of Theorem 4. Let us define $\theta = (\theta'_{\mathcal{A}}, \theta'_{\mathcal{A}^c})'$. To prove the support recovery consistency, we show with probability tending to one when $T \to \infty$, under $\|\theta_{\mathcal{A}} - \theta_{0,\mathcal{A}}\| = O_p(T^{-1/2})$ and suitable regularisation rates depending on the penalty, that

$$
\mathbb{G}_T^{pen}(\underline{y}; \theta_{\mathcal{A}}, \mathbf{0}_{\mathcal{A}^c}) = \min_{\|\theta_{\mathcal{A}^c}\| \leq CT^{-1/2}} \{\mathbb{G}_T^{pen}(\underline{y}; \theta_{\mathcal{A}}, \theta_{\mathcal{A}^c})\}. \tag{21}
$$

To prove (21), for any $\sqrt{T}$-consistent $\theta_{\mathcal{A}}$, we show that over the set $\{i \in \mathcal{A}^c, \theta_i : |\theta_i| \leq T^{-1/2}C\}$ for $C > 0$

$$
\begin{aligned}
\nabla_{\theta_i} \mathbb{G}_T^{pen}(\underline{y}; \theta) > 0 \quad & \text{when} \quad 0 < \theta_i < T^{-1/2}C, \\
\nabla_{\theta_i} \mathbb{G}_T^{pen}(\underline{y}; \theta) < 0 \quad & \text{when} \quad -T^{-1/2}C < \theta_i < 0,
\end{aligned} \tag{22}
$$

with probability converging to 1. For any index $i \in \mathcal{A}^c$, by a Taylor expansion around the true parameter, we have

$$
\begin{aligned}
& \nabla_{\theta_i} \mathbb{G}_T^{pen}(\underline{y}; \theta) = \nabla_{\theta_i} \mathbb{G}_T(\underline{y}; \theta) + \nabla_{\theta_i} \boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_i|) \text{sgn}(\theta_i) \\
= \quad & \nabla_{\theta_i} \mathbb{G}_T(\underline{y}; \theta_0) + \nabla^2_{\theta_i \theta_i} \mathbb{G}_T(\underline{y}; \theta_0) \big(\theta_i - \theta_{0,i}\big) + \nabla_{\theta_i} \boldsymbol{p}(\frac{\lambda_T}{T}, |\theta_i|) \text{sgn}(\theta_i).
\end{aligned}
$$

By the Central Limit Theorem of Billingsley (1961) and the ergodic Theorem of Billingsley (1995), we have

$$\sqrt{T}\nabla_{\theta_i}\mathbb{G}_T(\underline{y};\theta_0) = O_p(1), \;\; \nabla^2_{\theta\theta'}\mathbb{G}_T(\underline{y};\theta_0) \underset{T\to\infty}{\xrightarrow{\mathbb{P}}} \mathbb{E}[\nabla^2_{\theta\theta'}\ell(y_s, s\leq t;\theta_0)].$$

We thus obtain for the SCAD and MCP penalties

$$\begin{array}{rcl}\nabla_{\theta_i}\mathbb{G}_T(\underline{y};\theta) & = & O_p(T^{-1/2}) + \nabla_{\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_i|)\text{sgn}(\theta_i) \\ & = & \frac{\lambda_T}{T}\{\frac{T}{\lambda_T}\nabla_{\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_i|)\text{sgn}(\theta_i) + O_p(\frac{\sqrt{T}}{\lambda_T})\}.\end{array}$$

As a consequence, under the condition $\underset{T\to\infty}{\lim}\ \underset{x\to 0^+}{\lim\inf}\ \frac{T}{\lambda_T}\nabla_x\boldsymbol{p}(\frac{\lambda_T}{T},x) > 0$ and if the regularisation parameter satisfies $\frac{\lambda_T}{T^{1/2}} \to \infty$, we deduce that the sign of the gradient entirely depends on the sign of $\hat{\theta}$. This this proves (22).

For the Bridge penalty, following the same reasoning as in the SCAD and MCP, the non-penalised terms are of order $O_p(T^{-1/2})$. As for the penalty, we have

$$\nabla_{\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_i|) = \frac{\lambda_T}{T}q|\theta_i|^{q-1}\text{sgn}(\theta_i) = \frac{\lambda_T}{T^{(q+1)/2}}q|T^{1/2}\theta_i|^{q-1}\text{sgn}(\theta_i).$$

As a consequence, we obtain

$$\nabla_{\theta_i}\mathbb{G}_T(\underline{y};\theta_0) = \frac{\lambda_T}{T^{(q+1)/2}}\{q|T^{1/2}\theta_i|^{q-1}\text{sgn}(\theta_i) + O_p(\frac{T^{q/2}}{\lambda_T})\}.$$

Thus, under the assumption that $\lambda_T/T^{q/2} \to \infty$, this this proves (22).

We now turn to the asymptotic distribution. We prove that $\hat{\theta}_{\mathcal{A}^c}$ degenerates at $\mathbf{0}_{\mathcal{A}^c}$ with probability approaching one. Now by a Taylor expansion around $\theta_{0,i}$, for $i\in\mathcal{A}$, we have

$$\begin{array}{rl} & \nabla_{\theta_i}\mathbb{G}_T(\underline{y};\hat{\theta}) + \nabla_{\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T},|\hat{\theta}_i|)\text{sgn}(\hat{\theta}_i) \\ = & \nabla_{\theta_i}\mathbb{G}_T(\underline{y};\theta_0) + \sum\limits_{j\in\mathcal{A}}\nabla^2_{\theta_i\theta_j}\mathbb{G}_T(\underline{y};\theta_0)(\hat{\theta}_j - \theta_{0,j}) + \nabla_{\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}|)\text{sgn}(\theta_{0,i}) \\ + & \nabla^2_{\theta_i\theta_i}\boldsymbol{p}(\frac{\lambda_T}{T},|\theta_{0,i}|)(\hat{\theta}_i - \theta_{0,i})(1+o(1)).\end{array}$$

Then inverting this relationship and multiplying by $\sqrt{T}$, we obtain in vector form with respect to the elements in $\mathcal{A}$

$$\big(\nabla^2_{\mathcal{A}\mathcal{A}}\mathbb{G}_T(\underline{y};\theta_0)+\mathbf{S}_{T,\mathcal{A}\mathcal{A}}\big)\sqrt{T}\{(\hat{\theta}-\theta_0)_{\mathcal{A}}+\big(\nabla^2_{\mathcal{A}\mathcal{A}}\mathbb{G}_T(\underline{y};\theta_0)+\mathbf{S}_{T,\mathcal{A}\mathcal{A}}\big)^{-1}\mathbf{b}_{T,\mathcal{A}}\}=-\sqrt{T}\nabla_{\mathcal{A}}\mathbb{G}_T(\underline{y};\theta_0),$$

which implies

$$\sqrt{T}(\hat{\theta}-\theta_0)_{\mathcal{A}}=-\big(\nabla^2_{\mathcal{A}\mathcal{A}}\mathbb{G}_T(\underline{y};\theta_0)+\mathbf{S}_{T,\mathcal{A}\mathcal{A}}\big)^{-1}\sqrt{T}\nabla_{\mathcal{A}}\mathbb{G}_T(\underline{y};\theta_0)-\sqrt{T}\mathbf{b}_{T,\mathcal{A}},$$

where

$$\begin{aligned}\mathbf{b}_{T,\mathcal{A}} &= \big(\nabla_{\theta_1}\boldsymbol{p}(\tfrac{\lambda_T}{T},|\theta_{0,1}|)\text{sgn}(\theta_{0,1}),\cdots,\nabla_{\theta_{k_0}}\boldsymbol{p}(\tfrac{\lambda_T}{T},|\theta_{0,k_0}|)\text{sgn}(\theta_{0,k_0})\big)',\\ \mathbf{S}_{T,\mathcal{A}\mathcal{A}} &= \text{diag}(\nabla^2_{\theta_i\theta_i}\boldsymbol{p}(\tfrac{\lambda_T}{T},|\theta_{0,i}|),i=1,\cdots,k_0).\end{aligned}$$

Now, since $\lambda_T/T\to 0$, this implies that $A_{1,T}\to 0$ for the SCAD and MCP. As for the Bridge, the $\sqrt{T}$-consistency requires $\lambda_T=O(\sqrt{T})$ by Theorem 2 and the oracle property requires $\lambda_T T^{-q/2}\to\infty$. As in the SCAD and MCP cases, $A_{1,T}\to 0$. Thus in all cases, $\mathbf{b}_{T,\mathcal{A}}$ and $\mathbf{S}_{T,\mathcal{A}\mathcal{A}}$ vanish for T large enough. We thus deduce that by the central limit theorem for U-statistics and the Slutsky theorem

$$\sqrt{T}\big(\hat{\theta}-\theta_0\big)_{\mathcal{A}}\underset{T\to\infty}{\xrightarrow{d}}\mathcal{N}_{\mathbb{R}^{k_0}}(0,\mathbb{H}^{-1}_{\mathcal{A}\mathcal{A}}\mathbb{M}_{\mathcal{A}\mathcal{A}}\mathbb{H}^{-1}_{\mathcal{A}\mathcal{A}}),$$

with

$$\mathbb{M}=\mathbb{E}[\nabla_{\theta}\ell(y_s,s\leq t;\theta_0)\nabla_{\theta'}\ell(y_s,s\leq t;\theta_0)],\ \mathbb{H}=\mathbb{E}[\nabla^2_{\theta\theta'}\ell(y_s,s\leq t;\theta_0)].$$

□

Proof of Theorem 5. Under the Theorem's assumption, $\hat{\theta}\underset{T\to\infty}{\xrightarrow{\mathbb{P}}}\theta_0^*$. Now let us prove

$$\forall\epsilon>0,\ \lim_{T\to\infty}\mathbb{P}(\|\hat{\gamma}-\gamma_0\|>\epsilon)=0.$$

First, note that we work under the assumption that $\beta_0=(\theta_0^{*'},\gamma_0')$ is a well-separated point. Second, let us first establish that for every $\bar{\beta}\in\Theta_{0\backslash 2}$ with $\|\bar{\gamma}-\gamma_0\|>0$, where $\Theta_{0\backslash 2}=\{\beta:\beta=$

$(\theta_0^*, \gamma) \in \Theta_1 \times \Theta_2\} = \{\theta_0^*\} \times \Theta_2$, and every $\pi > 0$, there exists an open ball $\mathcal{V}(\bar{\beta}, \pi)$ around $\bar{\beta}$ in the space $\Theta_{0\backslash 2}$ such that

$$\mathbb{E}_{\beta_0}[\inf_{\beta \in \mathcal{V}(\bar{\beta},\pi)} f(y_s, s \leq t; \beta)] \geq \mathbb{E}_{\beta_0}[f(y_s, s \leq t; \bar{\beta})] - \pi. \tag{23}$$

To prove this statement, for a given $\bar{\beta} = (\bar{\theta}, \bar{\gamma}) \in \Theta_{0\backslash 1}$, where $\Theta_{0\backslash 1} = \{\beta : \beta = (\theta, \gamma_0) \in \Theta\}$, and $\bar{\gamma} \neq \gamma_0$, consider a sequence of open balls with radius $1/k$, $k \in \mathbb{N}$ defined by $V_k(\beta_0) = \{\beta \in \Theta_{0\backslash 1} : \|\beta - \beta_0\| \leq 1/k\}$. Since the sequence of random variables $\big(\inf_{\beta \in V_k(\beta_0)} f(y_s, s \leq t; \beta)\big)_k$ is increasing, then by the Beppo-Levi Theorem

$$\lim_{k\to\infty} \mathbb{E}_{\beta_0}[\inf_{\beta \in V_k(\beta_0)} f(y_s, s \leq t; \beta)] = \mathbb{E}_{\beta_0}[f(y_s, s \leq t; \beta_0)],$$

which thus proves (23). Now, under the Theorem's assumption, $\hat{\theta} \xrightarrow[T\to\infty]{\mathbb{P}} \theta_0^*$. We thus need to prove $\forall \epsilon > 0$, $\lim_{T\to\infty} \mathbb{P}(\|\hat{\gamma} - \gamma_0\| > \epsilon) = 0$. Invoking (23), for any given $\pi > 0$ and $\bar{\beta} \in \Theta_{0\backslash 2}$, $\bar{\beta} \neq \beta_0$, with $\|\bar{\gamma} - \gamma_0\| \geq \epsilon/2$, we can find an open ball $\mathcal{U}(\bar{\beta}) \subset \Theta_{0\backslash 2}$ such that

$$\mathbb{E}_{\beta_0}[\inf_{\beta \in \mathcal{U}(\bar{\beta})} f(y_s, s \leq t; \beta)] \geq \mathbb{E}_{\beta_0}[f(y_s, s \leq t; \bar{\beta})] - \pi.$$

Since the function $\gamma \mapsto \mathbb{E}_{\beta_0}[f(y_s, s \leq t; \theta_0^*, \gamma)] - \mathbb{E}_{\beta_0}[f(y_s, s \leq t; \theta_0^*, \gamma_0)]$ defined on $\Theta_{0\backslash 2}$, is strictly positive because β_0 is a well separated point and continuous on the compact subset $\mathcal{C}_0(\epsilon) = \{\beta \in \Theta_{0\backslash 2} : \|\gamma - \gamma_0\| \geq \epsilon/2\}$, it reaches its minimum $2\mu > 0$. Hence, for any given $\bar{\beta} \in \mathcal{C}_0(\epsilon)$, we have

$$\pi := \pi(\bar{\beta}) = \mathbb{E}_{\beta_0}[f(y_s, s \leq t; \bar{\beta})] - \mathbb{E}_{\beta_0}[f(y_s, s \leq t; \beta_0\theta_0^*, \gamma_0)] - \mu > 0.$$

Moreover, define $\mathcal{U}(\beta_0) = \{\beta \in \Theta_{0\backslash 2} : \|\beta - \beta_0\| < \epsilon\}$. Then

$$\Theta_{0\backslash 2} \subset \mathcal{U}(\beta_0) \cup \bigcup_{\beta \in \mathcal{C}_0(\epsilon)} \mathcal{U}(\beta).$$

Since $\Theta_{0\backslash 2}$ can be covered by a finite set of open ball by sequential compactness, there is a finite set of points $\beta_1, \cdots, \beta_T$ in $\mathcal{C}_0(\epsilon)$ such that

$$\Theta_{0\backslash 2} \subset \mathcal{U}(\beta_0) \cup \bigcup_{i=1}^{n} \mathcal{U}(\beta_i).$$

We thus deduce

$$\mathbb{P}(\|\hat{\gamma}-\gamma_0\| > \epsilon) \leq \mathbb{P}((\theta_0^*,\hat{\gamma}) \in \bigcup_{i=1}^{n} \mathcal{U}(\beta_i)) \leq \sum_{i=1}^{n} \mathbb{P}((\theta_0^*,\hat{\gamma}) \in \mathcal{U}(\beta_i)).$$

By definition of $\hat{\beta}$, for any $i = 1, \cdots, n$,

$$\begin{aligned}
\mathbb{P}((\theta_0^*,\hat{\gamma}) \in \mathcal{U}(\beta_i)) &\leq \mathbb{P}(\inf_{\beta \in \mathcal{U}(\beta_i)} \mathbb{L}_T(\underline{y};\beta) \leq \mathbb{L}_T(\underline{y};\theta_0^*,\hat{\gamma})) \\
\leq\ & \mathbb{P}(\inf_{\beta \in \mathcal{U}(\beta_i)} \mathbb{L}_T(\underline{y};\beta) \leq \mathbb{L}_T(\underline{y};\hat{\theta},\hat{\gamma}) + |\mathbb{L}_T(\underline{y};\theta_0^*,\hat{\gamma}) - \mathbb{L}_T(\underline{y};\hat{\theta},\hat{\gamma})|) \\
\leq\ & \mathbb{P}(\inf_{\beta \in \mathcal{U}(\beta_i)} \mathbb{L}_T(\underline{y};\beta) \leq \mathbb{L}_T(\underline{y};\hat{\theta},\gamma_0) + |\mathbb{L}_T(\underline{y};\theta_0^*,\hat{\gamma}) - \mathbb{L}_T(\underline{y};\hat{\theta},\hat{\gamma})|) \\
\leq\ & \mathbb{P}(\mathbb{E}_{\beta_0}[\inf_{\beta \in \mathcal{U}(\beta_i)} f(y_s, s \leq t;\beta)] \leq \mathbb{E}_{\beta_0}[f(y_s, s \leq t;\beta_0)] + |\mathbb{L}_T(\underline{y};\beta_0) - \mathbb{E}_{\beta_0}[f(y_s, s \leq t;\beta_0)]| \\
& + |\mathbb{L}_T(\underline{y};\theta_0^*,\hat{\gamma}) - \mathbb{L}_T(\underline{y};\hat{\theta},\hat{\gamma})| + |\mathcal{R}_T(\beta_i)|),
\end{aligned}$$

where $\mathcal{R}_T(\beta_i) = \frac{1}{T}\sum_{t=1}^{T} \inf_{\beta \in \mathcal{U}(\beta_i)} f(y_s, s \leq t;\beta) - \mathbb{E}_{\beta_0}[\inf_{\beta \in \mathcal{U}(\beta_i)} f(y_s, s \leq t;\beta)]$. Using (23), we have

$$\mathbb{E}_{\beta_0}[\inf_{\beta \in \mathcal{U}(\beta_i)} f(y_s, s \leq t;\beta)] \geq \mathbb{E}_{\beta_0}[f(y_s, s \leq t;\beta_0)] + \mu.$$

Thus

$$\begin{aligned}
&\mathbb{P}((\theta_0^*,\hat{\gamma}) \in \mathcal{U}(\beta_i)) \\
&\leq\ \mathbb{P}(\mu \leq |\mathcal{R}_T(\beta_i)| + |\mathbb{L}_T(\underline{y};\beta_0) - \mathbb{E}_{\beta_0}[f(y_s, s \leq t;\beta_0)]| + |\mathbb{L}_T(\underline{y};\theta_0^*,\hat{\gamma}) - \mathbb{L}_T(\underline{y};\hat{\theta},\hat{\gamma})|) \\
&\leq\ \mathbb{P}(\mu/3 \leq |\mathcal{R}_T(\beta_i)|) + \mathbb{P}(\mu/3 \leq |\mathbb{L}_T(\underline{y};\beta_0) - \mathbb{E}_{\beta_0}[f(y_s, s \leq t;\beta_0)]|) + \mathbb{P}(\mu/3 \leq |\mathbb{L}_T(\underline{y};\theta_0^*,\hat{\gamma}) - \mathbb{L}_T(\underline{y};\hat{\theta},\hat{\gamma})|).
\end{aligned}$$

Let us focus on $\mathbb{P}(\mu/3 \leq |\mathcal{R}_T(\beta_i)|)$. Although the quantity $f(y_s, s \leq t;\beta)$ is not necessarily integrable, the Ergodic Theorem can still be applied to $\mathbb{E}_{\beta_0}[f(y_s, s \leq t;\beta)]$. Moreover, $f(y_s, s \leq t;\beta)$ is a measurable function of an ergodic process, by assumption 8, the Ergodic Theorem can be applied to the process $(\inf_{\beta \in \mathcal{U}(\beta_i)} f(y_s, s \leq t;\beta))$, that is

$$\lim_{T\to\infty}\inf \frac{1}{T}\sum_{t=1}^{T} \inf_{\beta \in \mathcal{U}(\beta_i)} f(y_s, s \leq t;\beta) = \mathbb{E}_{\beta_0}[\inf_{\beta \in \mathcal{U}(\beta_i)} f(y_s, s \leq t;\beta)].$$

Thus, for $T > T_1$, we have

$$\mathbb{P}(\mu/3 \leq |\mathcal{R}_T(\beta_i)|) \leq \epsilon/3.$$

By the Ergodic Theorem, for $T > T_2$, then

$$\mathbb{P}(\mu/3 \leq |\mathbb{L}_T(\underline{y};\beta_0) - \mathbb{E}_{\beta_0}[f(y_s, s \leq t;\beta_0)]|) \leq \epsilon/3.$$

Finally, we need to control for the last probability. To do so, we need to prove

$$\sup_{\gamma\in\Theta} |\mathbb{L}_T(\underline{y};\hat{\theta},\gamma) - \mathbb{L}_T(\underline{y};\theta_0^*,\gamma)| = o_p(1).$$

To do so, by a Taylor expansion of $\mathbb{L}_T(\underline{y};\hat{\theta},\gamma)$ around θ_0^*, we have

$$\mathbb{L}_T(\underline{y};\hat{\theta},\gamma) = \mathbb{L}_T(\underline{y};\theta_0^*,\gamma) + (\hat{\theta}-\theta_0^*)'\nabla_\theta \mathbb{L}_T(\underline{y};\tilde{\theta},\gamma),$$

where $\|\tilde{\theta}-\hat{\theta}\| \leq \|\hat{\theta}-\theta_0^*\|$. Using the consistency of $\hat{\theta}$, it is sufficient to prove that

$$\frac{1}{T}\sum_{t=1}^{T} \sup_{\theta\in\Theta:\|\theta-\theta_0^*\|\leq\alpha} \|\nabla_\theta f(y_s, s\leq t;\theta,\gamma)\| = O_p(1),$$

for a small $\alpha > 0$. The score (applied here with respect to $\theta = vec(\Psi_{1:m})$) is given by

$$\begin{aligned}
&\nabla_\theta f(y_s, s\leq t;\theta,\gamma)\\
&= \nabla_\theta \frac{1}{2}\big([x_t - c^* - \Phi x_{t-1} - \Xi\{x_{t-1} - \Psi_{1:m}Z_{m,t-2}\}]'[x_t - c^* - \Phi x_{t-1} - \Xi\{x_{t-1} - \Psi_{1:m}Z_{m,t-2}\}]\big)\\
&= \big(\{Z_{m,t-2}\} \otimes \{\Xi'[x_t - c^* - \Phi x_{t-1} - \Xi\{x_{t-1} - \Psi_{1:m}Z_{m,t-2}\}]\}\big).
\end{aligned}$$

As a consequence, there exists some positive constant $C > 0$ such that, for any $\alpha > 0$,

$$\sup_{\theta\in\Theta:\|\theta-\theta_0^*\|\leq\alpha} \|\nabla_\theta f(y_s, s\leq t;\theta,\gamma)\| \leq C \sup_{\theta\in\Theta:\|\theta-\theta_0^*\|\leq\alpha} \|Z_{m,t-2}\|\|\Xi\|\|x_t - c^* - \Phi x_{t-1} - \Xi\{x_{t-1} - \Psi_{1:m}Z_{m,t-2}\}\|.$$

Based on $x_t = c^* + \Phi x_{t-1} + \Xi\{x_{t-1} - \Psi_{1:m}Z_{m,t-2}\} + v_t$, we obtain

$$X_t = \underline{c}^* + \Lambda X_{t-1} + \underline{v}_t,$$

where

$$X_t = \begin{pmatrix} x_t \\ x_{t-1} \\ x_{t-2} \\ \vdots \\ x_{t-m} \end{pmatrix}, \underline{v}_t = \begin{pmatrix} v_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \underline{c} = \begin{pmatrix} c^* \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \Lambda = \begin{pmatrix} \Phi + \Xi & -\Xi\Psi_1 & -\Xi\Psi_2 & \cdots & -\Xi\Psi_m \\ I_p & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & I_p & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \cdots & \cdots & \vdots \\ \mathbf{0} & \cdots & \cdots & I_p & \mathbf{0} \end{pmatrix}.$$

Assumption 9 provides a sufficient condition for stationarity, that is $\|\Lambda\|_s < 1$, with $\|.\|_s$ the spectral radius of Λ. Hence, under $\|\Lambda\|_s < 1$,

$$\frac{1}{T}\sum_{t=1}^{T} \sup_{\theta \in \Theta : \|\theta - \theta_0^*\| \leq \alpha} \|\nabla_\theta f(y_s, s \leq t; \theta, \gamma)\| = O_p(1).$$

As a consequence, for $T > T_3$, we deduce

$$\mathbb{P}(\mu/3 \leq |\mathbb{L}_T(\underline{y}; \theta_0^*, \hat{\gamma}) - \mathbb{L}_T(\underline{y}; \hat{\theta}, \hat{\gamma})|) < \epsilon/3.$$

We can conclude that for $T > T_1 \vee T_2 \vee T_3$, we have

$$\mathbb{P}((\theta_0^*, \hat{\gamma}) \in \mathcal{U}(\beta_i)) < \epsilon.$$

This thus proves the desired consistency result. □

Proof of Theorem 6. Through a Taylor expansion around β_0, we obtain for the $\hat{\gamma}$ component

$$0 = \nabla_\gamma \mathbb{L}_T(\underline{y}; \hat{\theta}), \hat{\gamma}) = \nabla_\gamma \mathbb{L}_T(\underline{y}; \beta_0) + \nabla_\theta \mathbb{L}_T(\underline{y}; \bar{\beta})_{\mathcal{A}} (\hat{\theta} - \theta_0)_{\mathcal{A}} + \nabla_\gamma \mathbb{L}_T(\underline{y}; \bar{\beta})(\hat{\gamma} - \gamma_0),$$

where $\|\bar{\beta} - \beta_0\| < \|\hat{\beta} - \beta_0\|$. Then inverting this relationship, multiplying by $\sqrt{T}$ and using the asymptotic expansion of the first step estimator, we obtain

$$\begin{aligned} & \sqrt{T}(\hat{\gamma} - \gamma_0) \\ = \; & (-\nabla_\gamma \mathbb{L}_T(\underline{y}; \bar{\beta}))^{-1} \nabla_\theta \mathbb{L}_T(\underline{y}; \bar{\beta})_{\mathcal{A}} \sqrt{T}(\hat{\theta} - \theta_0)_{\mathcal{A}} + (-\nabla_\gamma \mathbb{L}_T(\underline{y}; \bar{\beta}))^{-1} \sqrt{T} \nabla_\gamma \mathbb{L}_T(\underline{y}; \beta_0) \\ = \; & (-\nabla_\gamma \mathbb{L}_T(\underline{y}; \bar{\beta}))^{-1} \nabla_\theta \mathbb{L}_T(\underline{y}; \bar{\beta})_{\mathcal{A}} \{-\mathbf{b}_{T,\mathcal{A}} + (-\big(\nabla^2_{\mathcal{A}\mathcal{A}} \mathbb{G}_T(\underline{y}; \theta_0) + \mathbf{S}_{T,\mathcal{A}\mathcal{A}}\big))^{-1} \sqrt{T} \nabla_{\mathcal{A}} \mathbb{G}_T(\underline{y}; \theta_0)\} \\ & + (-\nabla_\gamma \mathbb{L}_T(\underline{y}; \bar{\beta}))^{-1} \sqrt{T} \nabla_\gamma \mathbb{L}_T(\underline{y}; \beta_0). \end{aligned}$$

Hence, by Slutsky's Theorem and the Central Limit Theorem, we obtain the desired asymptotic distribution since $\mathbf{S}_{T,\mathcal{A}\mathcal{A}} = \mathbf{0}$ and $\mathbf{b}_{T,\mathcal{A}} = \mathbf{0}$ for T sufficiently large. □

D Some competing M-GARCH models

The BEKK model directly generates a variance-covariance process. Developed by Baba, Engle, Kraft and Kroner, in a preliminary version of Engle and Kroner (1995), the BEKK is specified for a p-dimensional random vector ϵ_t as

$$\begin{cases} \epsilon_t &= H_t^{1/2}\eta_t, \text{ with } H_t := \mathbb{E}[\epsilon_t\epsilon_t'|\mathcal{F}_{t-1}] \succ 0 \text{ so that} \\ H_t &= \Omega + \sum_{k=1}^{q}\sum_{j=1}^{K} A_{kj}\epsilon_{t-k}\epsilon_{t-k}'A_{kj}' + \sum_{i=1}^{r}\sum_{i=1}^{K} B_{ij}H_{t-i}B_{ij}', \end{cases}$$

where K is an integer, Ω, A_{kj} and B_{kj} are square $p \times p$ matrices and $\Omega \succ 0$. One advantage of the BEKK model is there is no positive semi-definite constraint on the A_{kj} and B_{kj} matrices. However, it imposes highly artificial constraints on the volatilities and covariances of the components. As a consequence, the coefficients of a BEKK representation are difficult to interpret. In our application, a scalar BEKK was considered, where A_{kj} and B_{kj} are scalar with $K = 1$, $q = r = 1$, together with a Gaussian QMLE estimation.

Beside BEKK type dynamics, factor models provide rather natural alternatives. The O-GARCH assumes the decomposition $H_t = P\Lambda_t P'$, where $\Lambda_t = \text{diag}(\lambda_{1,t}, \cdots, \lambda_{K,t})$, with K the number of factors. Here, we choose $K = p$ factors and each λ_t is supposed to follow a univariate GARCH(1,1) process that is estimated by maximum likelihood. The matrix P is nonsingular and it is estimated by PCA on the empirical variance-covariance matrix of ϵ_t: see Alexander (2001), e.g.

Beside the latter direct specification of the covariance matrices (H_t) dynamics, an alternative road is to split the task into two parts: individual volatility dynamics on one side, and correlation dynamics on the other side. The most commonly used correlation process is the Dynamic Conditional Correlation (DCC) of Engle (2002). In its BEKK form, the general DCC model is specified as

$$\begin{cases} \epsilon_t &= H_t^{1/2}\eta_t, \text{ with } H_t := \mathbb{E}[\epsilon_t\epsilon_t'|\mathcal{F}_{t-1}] \succ 0 \text{ so that} \\ H_t &= D_t R_t D_t, \ R_t = Q_t^{\star -1/2} Q_t Q_t^{\star -1/2}, \\ Q_t &= \Omega + \sum_{k=1}^{q} M_k Q_{t-k} M_k' + \sum_{l=1}^{r} W_l u_{t-l} u_{t-l}' W_l', \end{cases} \tag{24}$$

where $D_t = \text{diag}\left(\sqrt{h_{11,t}}, \sqrt{h_{22,t}}, \dots, \sqrt{h_{pp,t}}\right)$, $u_t = (u_{1,t}, \dots, u_{p,t})$ with $u_{i,t} = \epsilon_{i,t}/\sqrt{h_{ii,t}}$, $Q_t = [q_{ij,t}]$, $Q_t^\star = \text{diag}\left(q_{11,t}, q_{22,t}, \dots, q_{pp,t}\right)$. The model is parameterized by some deterministic matrices $(M_k)_{k=1,\cdots,q}$, $(W_l)_{l=1,\cdots,r}$ and a positive definite $p \times p$ matrix Ω. Alternatively, Engle (2002) considered a VEC-type specification too. Denoting by $\odot$ the Hadamard matrix product, the (Q_t)-dynamics become

$$Q_t = \Omega^* + \sum_{k=1}^{q} B_k \odot Q_{t-k} + \sum_{l=1}^{r} A_l \odot u_{t-l} u_{t-l}', \tag{25}$$

where the deterministic matrices $(B_k)_{k=1,\cdots,q}$ and $(A_l)_{l=1,\cdots,r}$ must be positive semi-definite.

Since the number of parameters of the latter models is of order $O(p^2)$, the matrices M_k and W_l (resp. B_k's and A_l) are often assumed to be scalar. This is typically a strong and questionable constraint, particularly when the dimension p increases or when the variables in (ϵ_t) are heterogeneous. Furthermore, their inference is usually carried out trough the QML method, based on a Gaussian or Student quasi likelihood function. Under this methodology, applying a regularisation method, even possible, is numerically arduous and no general asymptotic results exist in this case (to the best of our knowledge), due to the non-convexity of the QML criterion.

If $R_t = R$ a constant correlation matrix, then (24) becomes the Constant Conditional Correlation (CCC) model.