Penalized Convex Estimation in Dynamic Location-Scale models

ALAMI CHENTOUFI, Reda

CREST-ENSAE

May 2025

Penalized Convex Estimation in Dynamic Location-Scale Models

Reda ALAMI CHENTOUFI
Crest-ENSAE, 5, avenue Henry Le Chatelier, 91120 Palaiseau cedex, France
reda.alamichentoufi@ensae.fr

Abstract

This paper introduces a two-step convex estimator for dynamic location–scale models. Step 1 relies on a $\sqrt{T}$-consistent preliminary estimator. Step 2 minimizes an adaptive L^1-penalized weighted least squares (WLS) criterion, yielding a sparse estimator. The objective is convex, avoiding the local-optima issues of non-convex optimizations. Consistency, asymptotic distribution, and model-selection consistency are proven. Simulations confirm finite-sample performance. A financial data set illustrates practical utility.

Keywords: Weighted LSE; Adaptive LASSO estimation; variable selection; GARCH models; ARMA models; Location–scale dynamics
MSC subject classification: 62M10, 62F12, 62J07, 91B84, 60G10
JEL classification: C01, C22, C51, C52, C58

1 Introduction

Model-selection complexity rises exponentially with the number of parameters. When the parameter count is small (N), each of the 2^N possible submodels can be ranked with the Akaike information criterion (AIC) (Akaike, 1974, 1998) or the Bayesian information criterion (BIC) (Schwarz, 1978). Once N grows, exhaustive search becomes infeasible; for example, $N = 20$ already implies roughly one million submodels. Penalized estimators embed selection in the estimation step and thus bypass the search. The least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) illustrates this approach and has been examined in depth by (Bunea et al., 2007; Zhang and Huang, 2008; Chan et al., 2015; Schweikert, 2022; Bhattacharjee et al., 2023), among many others.

Time-series applications of LASSO have largely focused on Markovian settings. Least squares (LS) LASSO has been used for autoregressive (AR) and autoregressive conditional heteroskedasticity (ARCH) (Engle, 1982) models. In linear regression, Fu and Knight (2000) derived the asymptotics of the LASSO estimator. Wang et al. (2007) applied LASSO to regression with AR errors and penalized both regression and AR coefficients. Nardi and Rinaldo (2011) studied the LS estimator for AR models whose order increases with sample size. Basu and Michailidis (2015) extended LASSO to vector AR models. Kock (2016) proved oracle properties for the adaptive LASSO (Zou, 2006) in non-stationary AR processes. Adamek et al. (2023) used LASSO in high-dimensional time-series settings, and Poignard and Fermanian (2021) investigated grouped LASSO for multivariate ARCH models. All of these techniques rely on convex quadratic LS criteria, which simplify optimization and asymptotic analysis.

Convexity breaks down in non-Markovian models with persistence, such as autoregressive moving average (ARMA) and generalized ARCH (GARCH) models (Bollerslev, 1986). The conditional LS loss for ARMA models is non-convex because of recursive moving-average (MA) terms. GARCH models are usually estimated by quasi-maximum likelihood (QML); QML is non-convex and can face boundary-parameter problems (Francq and Thieu, 2019). Non-convexity complicates asymptotic theory, raises computational cost, and introduces local optima (Wang et al., 2014; Loh, 2017).

Non-convex penalized likelihood for time series was analyzed by Nielsen and Rahbek (2024), who extended the non-concave penalties of Fan and Li (2001); Fan and Peng (2004) from independent and identically distributed (i.i.d.) data to temporally dependent data and allowed parameters on the boundary of the parameter space. Chan and Chen (2011) proposed a two-step adaptive LASSO for ARMA models: residuals from an initial AR fit serve as exogenous regressors to convexify the LS problem, and the AR order is chosen by an information criterion. Chan et al. (2020) introduced a non-convex penalized estimator for ARMA models with unit roots. Their iterative algorithm minimizes a non-convex loss and selects the LASSO penalty through a BIC-like rule, but at high computational cost.

Motivated by these efforts, this paper tackles dynamic location-scale models in which standard estimators are non-convex. The procedure is two-step. The first step relies on a $\sqrt{T}$-consistent estimator. The second step applies an adaptive L^1-penalized WLS criterion to recover sparsity while keeping the objective convex. Unlike Chan and Chen (2011), no auxiliary AR-order selection is required. Unlike Nielsen and Rahbek (2024) and Chan et al. (2020), the second step remains convex, so the global optimum is guaranteed. As shown in Section 4, computation time is markedly shorter on real data compared to Chan et al. (2020).

The method has five advantages. First, it covers general dynamic location-scale models. Second, it relies on a $\sqrt{T}$-consistent preliminary estimator, eliminating auxiliary model selection and needing only mild regularity conditions. Third, it does not require high-order moment assumptions, which is crucial for heavy-tailed financial data. Fourth, it sidesteps boundary-parameter issues. Finally, its convex second step allows fast computation through the least-angle-regression LASSO (LARS–LASSO) algorithm (Efron et al., 2004).

The remainder of the paper is organized as follows. Section 2 formalizes the model and the two-step estimator. Section 3 presents assumptions and proves consistency, asymptotic distribution, and model-selection consistency for the location parameter estimator. Analogous results for scale are given in Appendix I to keep the text light. Section 4 reports Monte Carlo evidence and an empirical study with financial data. All proofs are in Appendix II.

2 Convexification of the Penalized Estimation

Let $\{y_t, t \in \mathbb{Z}\}$ be a real-valued process decomposed as:

$$y_t = \mu_t + \epsilon_t, \tag{1}$$

$$\epsilon_t = \sigma_t \eta_t, \tag{2}$$

where $\{\eta_t, t \in \mathbb{Z}\}$ is an i.i.d. innovation process with zero mean and unit variance. The location μ_t and scale σ_t are defined as:

$$\mu_t = m\left(\epsilon_{t-1}, \epsilon_{t-2}, \ldots, y_{t-1}, y_{t-2}, \ldots, \boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \ldots; \boldsymbol{\phi}_0\right), \tag{3}$$

$$\sigma_t^2 = h\left(\epsilon_{t-1}, \epsilon_{t-2}, \ldots, \sigma_{t-1}^2, \sigma_{t-2}^2, \ldots, \boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \ldots; \boldsymbol{\theta}_0\right) > 0, \tag{4}$$

where $\{\boldsymbol{X}_t, t \in \mathbb{Z}\}$ is a vector of exogenous covariates. The parameters satisfy $\boldsymbol{\phi}_0 \in \Phi \subset \mathbb{R}^\nu$ and $\boldsymbol{\theta}_0 \in \Theta \subset \mathbb{R}^n$, with Φ and Θ compact and convex. The mappings:

$$m : \mathbb{R}^\infty \times \Phi \mapsto \mathbb{R}, \quad h : \mathbb{R}^\infty \times \Theta \mapsto \mathbb{R}$$

are measurable. We estimate $\boldsymbol{\phi}_0$ and $\boldsymbol{\theta}_0$ separately. The procedure for $\boldsymbol{\phi}_0$ is presented first and the procedure for $\boldsymbol{\theta}_0$ is left to the Appendix I.

2.1 Construction of the Location Parameter Estimator

Define the auxiliary mappings:

$$\boldsymbol{\phi} \mapsto \mu_t(\boldsymbol{\phi}) = m\left(\epsilon_{t-1}(\boldsymbol{\phi}), \epsilon_{t-2}(\boldsymbol{\phi}), \ldots, y_{t-1}, y_{t-2}, \ldots, \boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \ldots; \boldsymbol{\phi}\right), \tag{5}$$

$$\boldsymbol{\phi} \mapsto \epsilon_t(\boldsymbol{\phi}) = y_t - \mu_t(\boldsymbol{\phi}), \tag{6}$$

$$(\boldsymbol{\phi}, \boldsymbol{\varphi}) \mapsto f_t(\boldsymbol{\phi}, \boldsymbol{\varphi}) = m\left(\epsilon_{t-1}(\boldsymbol{\phi}), \epsilon_{t-2}(\boldsymbol{\phi}), \ldots, y_{t-1}, y_{t-2}, \ldots, \boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \ldots; \boldsymbol{\varphi}\right). \tag{7}$$

Equation (7) splits the recursion (5) in two parts: $\boldsymbol{\phi}$ generates the lagged residuals, while $\boldsymbol{\varphi}$ is the argument to be penalized. Let $\boldsymbol{\alpha} = (\boldsymbol{\phi}', \boldsymbol{\varphi}')'$ and $\boldsymbol{\alpha}_0 = (\boldsymbol{\phi}_0', \boldsymbol{\phi}_0')'$. The penalized WLS criterion is:

$$Q_T(\boldsymbol{\alpha}) = \frac{1}{T} L_T(\boldsymbol{\alpha}) + \sum_{j=1}^{\nu} \lambda_{j,T} |\varphi_j|, \tag{8}$$

$$L_T(\boldsymbol{\alpha}) = \sum_{t=1}^{T} l_t(\boldsymbol{\alpha}), \quad l_t(\boldsymbol{\alpha}) = \left(\frac{y_t - f_t(\boldsymbol{\alpha})}{w_t}\right)^2. \tag{9}$$

The weight $w_t = w(y_{t-1}, y_{t-2}, \ldots, \boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \ldots) \geq \underline{w} > 0$ is a measurable mapping from $\mathbb{R}^\infty$ to $[\underline{w}, \infty)$. The penalty sequence $(\boldsymbol{\lambda}_T)_{T \in \mathbb{N}} = \left((\lambda_{1,T}, \ldots, \lambda_{\nu,T})'\right)_{T \in \mathbb{N}}$ is deterministic with non-negative components.

Let us denote $\boldsymbol{O}_t = (y_t, \boldsymbol{X}_t')'$. Because only a finite sample $\{\boldsymbol{O}_t, 1 \leq t \leq T\}$ is observed, we work with truncated versions of the mappings defined previously. Let initial values $\widetilde{\boldsymbol{O}}_0, \widetilde{\boldsymbol{O}}_{-1}, \ldots$ be fixed, and set:

$$\boldsymbol{\phi} \mapsto \widetilde{\mu}_t(\boldsymbol{\phi}) = m\left(\widetilde{\epsilon}_{t-1}(\boldsymbol{\phi}), \ldots, y_{t-1}, \ldots, \widetilde{y}_0, \ldots, \boldsymbol{X}_{t-1}, \ldots, \widetilde{\boldsymbol{X}}_0, \ldots; \boldsymbol{\phi}\right), \tag{10}$$

$$\boldsymbol{\phi} \mapsto \widetilde{\epsilon}_t(\boldsymbol{\phi}) = y_t - \widetilde{\mu}_t(\boldsymbol{\phi}), \tag{11}$$

$$\boldsymbol{\alpha} \mapsto \widetilde{f}_t(\boldsymbol{\alpha}) = m\left(\widetilde{\epsilon}_{t-1}(\boldsymbol{\phi}), \ldots, y_{t-1}, \ldots, \widetilde{y}_0, \ldots, \boldsymbol{X}_{t-1}, \ldots, \widetilde{\boldsymbol{X}}_0, \ldots; \boldsymbol{\varphi}\right), \tag{12}$$

$$\widetilde{w}_t = w\left(y_{t-1}, \ldots, \widetilde{y}_0, \ldots, \boldsymbol{X}_{t-1}, \ldots, \widetilde{\boldsymbol{X}}_0, \ldots\right). \tag{13}$$

Write $\widetilde{Q}_T$, $\widetilde{L}_T$, and $\widetilde{l}_t$ for the criteria obtained by substituting (10)-(13) into (8)-(9).

For many models, such as ARMA and GARCH, $\widetilde{L}_T(\boldsymbol{\phi}, \cdot)$ is convex almost surely (a.s.). Hence, with any consistent preliminary estimator $\widehat{\boldsymbol{\phi}}_T$ of $\boldsymbol{\phi}_0$ we keep the convexity in the second step based on i.e. $\widetilde{L}_T\left(\widehat{\boldsymbol{\phi}}_T, \cdot\right)$. The second stage LASSO estimator of $\boldsymbol{\varphi}_0$ is:

$$\widehat{\boldsymbol{\varphi}}_T = \underset{\boldsymbol{\varphi}\in\Phi}{\arg\min}\, \widetilde{Q}_T\left(\widehat{\boldsymbol{\phi}}_T, \boldsymbol{\varphi}\right). \tag{14}$$

Section 3.1 studies $\widehat{\boldsymbol{\varphi}}_T$. Section 3.2 covers the adaptive version:

$$Q_T^{AL}(\boldsymbol{\alpha}) = \begin{cases} \frac{1}{T}L_T(\boldsymbol{\alpha}) + \sum_{j=1}^{\nu} \frac{\lambda_{j,T}}{\left|\widehat{\phi}_{j,T}\right|^{\tau}} |\varphi_j| & \text{if } \forall j \in \{1,\dots,\nu\}, \widehat{\phi}_{j,T} \neq 0 \\ \infty \quad \text{otherwise,} \end{cases} \tag{15}$$

$$\widehat{\boldsymbol{\varphi}}_T^{AL} = \underset{\boldsymbol{\varphi}\in\Phi}{\arg\min}\, \widetilde{Q}_T^{AL}\left(\widehat{\boldsymbol{\phi}}_T, \boldsymbol{\varphi}\right), \tag{16}$$

where τ is a fixed positive constant.

Example 1 (ARMAX convexification)**.** Consider an ARMA process with exogenous components (ARMAX):

$$y_t = a_{1,0}y_{t-1} + b_{1,0}\epsilon_{t-1} + \varsigma_{1,0}X_{1,t-1} + \epsilon_t, \tag{17}$$

with $\epsilon_t = \sigma_t\eta_t$ as in (2). In practice the true orders are unknown. A common tactic fixes large integers (p, q, r) and estimates:

$$y_t = \sum_{i=1}^{p} a_{i,0}y_{t-i} + \sum_{j=1}^{q} b_{j,0}\epsilon_{t-j} + \sum_{k=1}^{r} \varsigma_{k,0}X_{k,t-1} + \epsilon_t, \tag{18}$$

where most coefficients are zero at the truth. Exhaustive search over the 2^{p+q+r} sub-models is infeasible. We therefore penalize. Define recursively:

$$\widetilde{\epsilon}_t(\boldsymbol{\phi}) = y_t - \sum_{i=1}^{p} a_i y_{t-i} - \sum_{j=1}^{q} b_j\widetilde{\epsilon}_{t-j}(\boldsymbol{\phi}) - \sum_{k=1}^{r} \varsigma_k X_{k,t-1},$$

where $\boldsymbol{\phi} = (a_1, \dots, a_p, b_1, \dots, b_q, \varsigma_1, \dots, \varsigma_r)$ and $\boldsymbol{\phi}_0 = (a_{1,0}, 0, \dots, 0, b_{1,0}, 0, \dots, 0, \varsigma_{1,0}, 0, \dots, 0)$. A naive L^1-penalized criterion is non-convex because of the MA terms:

$$\boldsymbol{\phi} \mapsto \sum_{t=1}^{T} \tilde{\epsilon}_t^2(\boldsymbol{\phi}) + \sum_{i=1}^{p+q+r} \lambda_{i,T}|\phi_i|,$$

Our two-step procedure convexifies this function as follows:

- **Step 1:** Obtain a consistent preliminary $\widehat{\boldsymbol{\phi}}_T$ (e.g. QMLE or conditional LS).
- **Step 2:** Solve the convex problem

$$\boldsymbol{\varphi} \mapsto ||\boldsymbol{Y}_T - \boldsymbol{\Omega}_T\boldsymbol{\varphi}||^2 + \sum_{i=1}^{p+q+r} \lambda_{i,T}|\varphi_i|, \quad \boldsymbol{Y}_T = (y_1, \dots, y_T)' \in \mathbb{R}^T,$$

$$\boldsymbol{\Omega}_T = \begin{pmatrix} \widetilde{y}_0 & \dots & \widetilde{y}_{1-p} & \widetilde{\epsilon}_0\left(\widehat{\boldsymbol{\phi}}_T\right) & \dots & \widetilde{\epsilon}_{1-q}\left(\widehat{\boldsymbol{\phi}}_T\right) & \widetilde{X}_{1,0} & \dots & \widetilde{X}_{r,0} \\ y_1 & \dots & \widetilde{y}_{2-p} & \widetilde{\epsilon}_1\left(\widehat{\boldsymbol{\phi}}_T\right) & \dots & \widetilde{\epsilon}_{2-q}\left(\widehat{\boldsymbol{\phi}}_T\right) & X_{1,1} & \dots & X_{r,1} \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ y_{T-1} & \dots & y_{T-p} & \widetilde{\epsilon}_{T-1}\left(\widehat{\boldsymbol{\phi}}_T\right) & \dots & \widetilde{\epsilon}_{T-q}\left(\widehat{\boldsymbol{\phi}}_T\right) & X_{1,T-1} & \dots & X_{r,T-1} \end{pmatrix}.$$

The objective is a standard LASSO. The y_ts are regressed on their lagged values, the residuals $\widetilde{\epsilon}_t\left(\widehat{\boldsymbol{\phi}}_T\right)$ and the exogenous covariates. The residuals are considered as exogenous in this regression. When there is no penalization (i.e. $\boldsymbol{\lambda}_T = \mathbf{0}$) the solution is explicit $\widehat{\boldsymbol{\varphi}}_T = \left(\boldsymbol{\Omega}_T'\boldsymbol{\Omega}_T\right)^{-1}\boldsymbol{\Omega}_T'\boldsymbol{Y}_T$. Otherwise, the LARS–LASSO algorithm yields the full penalty path efficiently. In this case, the minimizer of this objective function $\widehat{\boldsymbol{\varphi}}_T$ is sparse in general.

Example 2 (GARCHX convexification)**.** Consider a GARCH process with exogenous components (GARCHX):

$$\sigma_t^2 = \omega_0 + \alpha_{1,0}\epsilon_{t-1}^2 + \beta_{1,0}\sigma_{t-1}^2 + \pi_{1,0}X_{1,t-1}, \tag{19}$$

with positive coefficients and exogenous covariates. In this example, we assume that there is no location component (the scale process is observed). Squaring (2) gives:

$$\epsilon_t^2 = \omega_0 + \alpha_{1,0}\epsilon_{t-1}^2 + \beta_{1,0}\sigma_{t-1}^2 + \pi_{1,0}X_{1,t-1} + \sigma_t^2\left(\eta_t^2 - 1\right).$$

The squared process ϵ_t^2 is now a location process. Fix large orders (P, Q, R) and write the expanded model:

$$\sigma_t^2 = \omega_0 + \sum_{i=1}^{Q}\alpha_{i,0}\epsilon_{t-i}^2 + \sum_{j=1}^{P}\beta_{j,0}\sigma_{t-j}^2 + \sum_{k=1}^{R}\pi_{k,0}X_{k,t-1}^2,$$

where the true parameter is $\boldsymbol{\theta}_0 = (\omega_0, \alpha_{1,0}, \dots, \alpha_{Q,0}, \beta_{1,0}, \dots, \beta_{P,0}, \pi_{1,0}, \dots, \pi_{R,0})$ with $\alpha_{i,0} = 0$, $\beta_{j,0} = 0$, $\pi_{k,0} = 0$ for $i, j, k \geq 2$. Given a consistent preliminary estimator $\widehat{\boldsymbol{\theta}}_T$ of $\boldsymbol{\theta}_0$, compute the fitted variances $\widetilde{\sigma}_t^2\left(\widehat{\boldsymbol{\theta}}_T\right)$, and regress ϵ_t^2 on the lagged squares, the fitted variances, and the exogenous terms with an L^1 penalty. The resulting problem is convex. The fitted variances can also serve as weights for the regression. Detailed Monte Carlo illustrations follow in Section 4.

2.2 Construction of the Scale Parameter Estimator

Section 2.1 dealt only with the location parameter $\boldsymbol{\phi}_0$, so the first-step estimator $\widehat{\boldsymbol{\phi}}_T$ was enough. When the scale parameter $\boldsymbol{\theta}_0$ is of interest, a global estimator $\widehat{\boldsymbol{\rho}}_T := \left(\widehat{\boldsymbol{\phi}}_T', \widehat{\boldsymbol{\theta}}_T'\right)'$ of $\boldsymbol{\rho}_0 := (\boldsymbol{\phi}_0', \boldsymbol{\theta}_0')'$ is necessary. The scale parameter appears in the latent process ϵ_t. Because ϵ_t is unobserved, we work with the residuals:

$$\widetilde{\epsilon}_t\left(\widehat{\boldsymbol{\phi}}_T\right) = y_t - \widetilde{\mu}_t\left(\widehat{\boldsymbol{\phi}}_T\right).$$

and use the auxiliary mappings below to complete the estimation scheme:

$$(\boldsymbol{\phi}, \boldsymbol{\theta}) \in \Phi \times \Theta \times \Theta, \quad \sigma_t^2(\boldsymbol{\phi}, \boldsymbol{\theta}) = h\left(\epsilon_{t-1}(\boldsymbol{\phi}), \dots, \sigma_{t-1}^2(\boldsymbol{\phi}, \boldsymbol{\theta}), \dots, \boldsymbol{X}_{t-1}, \dots; \boldsymbol{\theta}\right). \tag{20}$$

$$(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\psi}) \in \Phi \times \Theta \times \Theta, \quad g_t(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\psi}) = h\left(\epsilon_{t-1}(\boldsymbol{\phi}), \dots, \sigma_{t-1}^2(\boldsymbol{\phi}, \boldsymbol{\theta}), \dots, \boldsymbol{X}_{t-1}, \dots; \boldsymbol{\psi}\right). \tag{21}$$

These functions are equivalent to (5) and (7). Since ϵ_t is not observed, we introduce $\epsilon_t(\boldsymbol{\phi})$ instead of ϵ_t in the recursion (4). Set $\boldsymbol{\rho} = (\boldsymbol{\phi}', \boldsymbol{\theta}')'$, $\boldsymbol{\beta} = (\boldsymbol{\rho}', \boldsymbol{\vartheta}')'$ and $\boldsymbol{\beta}_0 = (\boldsymbol{\rho}_0', \boldsymbol{\theta}_0')'$. Introduce:

$$\mathcal{Q}_T(\boldsymbol{\beta}) = \frac{1}{T}\mathcal{L}_T(\boldsymbol{\beta}) + \sum_{j=1}^{n} \iota_{j,T} |\vartheta_j|, \tag{22}$$

$$\mathcal{L}_T(\boldsymbol{\beta}) = \sum_{t=1}^{T} \ell_t(\boldsymbol{\beta}), \quad \ell_t(\boldsymbol{\beta}) = \left(\frac{\epsilon_t(\boldsymbol{\phi}) - g_t(\boldsymbol{\beta})}{\omega_t}\right)^2, \tag{23}$$

The weight $\omega_t = \omega(y_{t-1}, y_{t-2}, \ldots, \boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \ldots) \geq \underline{\omega} > 0$ is a measurable mapping from $\mathbb{R}^\infty$ to $[\underline{\omega}, \infty)$. The penalties $(\boldsymbol{\iota}_T)_{T\in\mathbb{N}} = \left((\iota_{1,T}, \ldots, \iota_{n,T})'\right)_{T\in\mathbb{N}}$ are deterministic and non-negative. Using the truncated counterparts (tilde notation) and the preliminary estimator $\widehat{\boldsymbol{\rho}}_T$, we define the second-step estimator of $\boldsymbol{\theta}_0$ as:

$$\widehat{\boldsymbol{\vartheta}}_T = \arg\min_{\boldsymbol{\vartheta}\in\Theta} \widetilde{\mathcal{Q}}_T(\widehat{\boldsymbol{\rho}}_T, \boldsymbol{\vartheta}).$$

The adaptive version is:

$$\mathcal{Q}_T^{AL}(\boldsymbol{\beta}) = \begin{cases} \frac{1}{T}\mathcal{L}_T(\boldsymbol{\beta}) + \sum_{j=1}^{n} \frac{\iota_{j,T}}{\left|\widehat{\theta}_{j,T}\right|^\tau} |\vartheta_j| & \text{if } \forall j \in \{1,\ldots,n\}, \widehat{\theta}_{j,T} \neq 0 \\ \infty \quad \text{otherwise,} & \end{cases}$$

$$\widehat{\boldsymbol{\vartheta}}_T^{AL} = \arg\min_{\boldsymbol{\vartheta}\in\Theta} \widetilde{\mathcal{Q}}_T^{AL}\left(\widehat{\boldsymbol{\theta}}_T, \boldsymbol{\vartheta}\right).$$

To preserve clarity and avoid overloading the notation, the same symbol τ is used for the adaptive power as in the previous section. Appendix I studies $\widehat{\boldsymbol{\vartheta}}_T$ and $\widehat{\boldsymbol{\vartheta}}_T^{AL}$.

Example 3 (ARMAX-GARCHX)**.** Compute $\widehat{\boldsymbol{\rho}}_T = \left(\widehat{\boldsymbol{\phi}}_T', \widehat{\boldsymbol{\theta}}_T'\right)'$ and use $\widehat{\boldsymbol{\phi}}_T$ as in the first example of ARMAX model to compute $\widehat{\boldsymbol{\varphi}}_T$. All the procedure of the ARMAX example remains independent of the dynamics of ϵ_t. Then use $\widetilde{\epsilon}_t\left(\widehat{\boldsymbol{\phi}}_T\right)$ instead of ϵ_t in the second example of GARCHX model. The squared residuals $\widetilde{\epsilon}_t^2\left(\widehat{\boldsymbol{\phi}}_T\right)$ can be regressed on their lagged values $\widetilde{\epsilon}_{t-1}^2\left(\widehat{\boldsymbol{\phi}}_T\right), \ldots, \widetilde{\epsilon}_{t-Q}^2\left(\widehat{\boldsymbol{\phi}}_T\right)$, the fitted variances $\widetilde{\sigma}_{t-1}^2(\widehat{\boldsymbol{\rho}}_T), \ldots, \widetilde{\sigma}_{t-P}^2(\widehat{\boldsymbol{\rho}}_T)$, and the exogenous components. The optimization problem becomes:

$$\boldsymbol{\vartheta} \mapsto ||\boldsymbol{Y}_T - \boldsymbol{\Omega}_T\boldsymbol{\vartheta}||^2 + \sum_{i=1}^{P+Q+R} \iota_{i,T}|\vartheta_i|, \quad \boldsymbol{Y}_T = \left(\widetilde{\epsilon}_1^2\left(\widehat{\boldsymbol{\phi}}_T\right), \ldots, \widetilde{\epsilon}_T^2\left(\widehat{\boldsymbol{\phi}}_T\right)\right)' \in \mathbb{R}^T,$$

$$\boldsymbol{\Omega}_T = \begin{pmatrix} \widetilde{\epsilon}_0^2\left(\widehat{\boldsymbol{\phi}}_T\right) & \ldots & \widetilde{\epsilon}_{1-Q}^2\left(\widehat{\boldsymbol{\phi}}_T\right) & \widetilde{\sigma}_0^2\left(\widehat{\boldsymbol{\phi}}_T\right) & \ldots & \widetilde{\sigma}_{1-P}^2\left(\widehat{\boldsymbol{\phi}}_T\right) & \widetilde{X}_{1,0} & \ldots & \widetilde{X}_{R,0} \\ \widetilde{\epsilon}_1^2\left(\widehat{\boldsymbol{\phi}}_T\right) & \ldots & \widetilde{\epsilon}_{2-Q}^2\left(\widehat{\boldsymbol{\phi}}_T\right) & \widetilde{\sigma}_1^2\left(\widehat{\boldsymbol{\phi}}_T\right) & \ldots & \widetilde{\sigma}_{2-P}^2\left(\widehat{\boldsymbol{\phi}}_T\right) & X_{1,1} & \ldots & X_{R,1} \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ \widetilde{\epsilon}_{T-1}^2\left(\widehat{\boldsymbol{\phi}}_T\right) & \ldots & \widetilde{\epsilon}_{T-Q}^2\left(\widehat{\boldsymbol{\phi}}_T\right) & \widetilde{\sigma}_{T-1}^2\left(\widehat{\boldsymbol{\phi}}_T\right) & \ldots & \widetilde{\sigma}_{T-P}^2\left(\widehat{\boldsymbol{\phi}}_T\right) & X_{1,T-1} & \ldots & X_{R,T-1} \end{pmatrix}.$$

3 Theoretical Results

This section gives the asymptotic properties of the LASSO estimator $\widehat{\boldsymbol{\varphi}}_T$. It then treats the adaptive version $\widehat{\boldsymbol{\varphi}}_T^{AL}$, derives its limit law, and proves selection consistency. Because the scale estimator obeys the same logic, its results are deferred to Appendix I.

Throughout, let $\mathcal{F}_t$ be the sigma-field generated by $\{\boldsymbol{O}_u, u \leq t\}$. Write $\mathcal{V}(\boldsymbol{\phi}_0)$ for a neighborhood of $\boldsymbol{\phi}_0$, and $\mathcal{V}(\boldsymbol{\alpha}_0) = \mathcal{V}(\boldsymbol{\phi}_0) \times \mathcal{V}(\boldsymbol{\phi}_0)$. The interior of Φ is denoted $\overset{\circ}{\Phi}$.

3.1 LASSO Estimator

Under the following assumptions, inspired by Aknouche and Francq (2023), we show the strong consistency of the estimator $\widehat{\boldsymbol{\varphi}}_T$.

A1 The process $\left\{(y_t, \epsilon_t)', t \in \mathbb{Z}\right\}$ is a solution to (1)-(4) and $\left\{(y_t, \epsilon_t, \boldsymbol{X}_t')', t \in \mathbb{Z}\right\}$ is strictly stationary, ergodic, and $\mathcal{F}_t$-measurable at time t, with η_t independent of $\mathcal{F}_{t-1}$.

A2 $\left(1 + |y_t| + \sup_{\boldsymbol{\alpha} \in \Phi \times \Phi} |f_t(\boldsymbol{\alpha})|\right) a_t \xrightarrow[t \to \infty]{a.s.} 0$, with $a_t = \sup_{\boldsymbol{\alpha} \in \Phi \times \Phi} \left|\widetilde{f}_t(\boldsymbol{\alpha}) - f_t(\boldsymbol{\alpha})\right|$.

A3 $\left(1 + y_t^2 + \sup_{\boldsymbol{\alpha} \in \Phi \times \Phi} f_t^2(\boldsymbol{\alpha})\right) d_t \xrightarrow[t \to \infty]{a.s.} 0$, with $d_t = \left|\widetilde{w}_t^2 - w_t^2\right|$.

A4 $\mathbb{E}\left[\left(\frac{\sigma_1}{w_1}\right)^2\right] < \infty$.

A5 $\mathcal{V}(\boldsymbol{\phi}_0) \subset \Phi$.

A6 The function f_1 is a.s. of class C^1 on $\overset{\circ}{\Phi} \times \overset{\circ}{\Phi}$, and $w_1^{-1} \sup_{\boldsymbol{\alpha} \in \mathcal{V}(\boldsymbol{\phi}_0) \times \overset{\circ}{\Phi}} \left\|\frac{\partial f_1}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha})\right\|$ belongs to L^2.

A7 $\mathbf{0} \in \overset{\circ}{\Phi}$.

A8 $\widehat{\boldsymbol{\phi}}_T \xrightarrow[T \to \infty]{a.s.} \boldsymbol{\phi}_0$.

A9 $\forall \boldsymbol{\phi} \in \mathcal{V}(\boldsymbol{\phi}_0)$, the functions $l_t(\boldsymbol{\phi}, \cdot)$ and $\widetilde{l}_t(\boldsymbol{\phi}, \cdot)$ are a.s. strictly convex on Φ.

Assumptions A2 and A3 ensure that initial values have no impact asymptotically. These assumptions are broadly applicable since the influence of initial values diminishes exponentially in many models. The choice of $\{w_t, t \in \mathbb{Z}\}$ is guided by the Assumption A4 so that no high-order moments of the Data Generating Process (DGP) are required. Assumption A5 places $\boldsymbol{\phi}_0$ in $\overset{\circ}{\Phi}$. This interior condition is natural for WLS; it is not always satisfied by QMLEs, which may reach the boundary. Assumption A7 keeps $\mathbf{0}$ in $\overset{\circ}{\Phi}$. This condition ensures that the estimator belongs to $\overset{\circ}{\Phi}$ because Φ is convex and the L^1 penalty shrinks the estimator towards $\mathbf{0}$. Lastly, Assumption A9 holds true in models like ARMA, GARCH, or GJR-GARCH.

Remark 3.1. If the observed process is $\{\epsilon_t, t \in \mathbb{Z}\}$, following a scale model defined by (2) and (4), it follows that $\epsilon_t^2 = \sigma_t^2 + \sigma_t^2\left(\eta_t^2 - 1\right)$ is a location process. In that case our results apply to the squared series provided $\mathbb{E}\left[\eta_t^4\right] < \infty$.

Theorem 3.1. *Assume that $\boldsymbol{\lambda}_T \xrightarrow[T \to \infty]{} \boldsymbol{\lambda}_\infty < +\infty$ component-wise. Then, under Assumptions A1-A9:*

$$\widehat{\boldsymbol{\varphi}}_T \xrightarrow[T \to \infty]{a.s.} \arg\min_{\boldsymbol{\varphi} \in \Phi} Q_\infty(\boldsymbol{\phi}_0, \boldsymbol{\varphi}),$$

where $Q_\infty(\boldsymbol{\phi}_0, \boldsymbol{\varphi}) := \mathbb{E}\left[l_1(\boldsymbol{\phi}_0, \boldsymbol{\varphi})\right] + \sum_{j=1}^{\nu} \lambda_{j,\infty} |\varphi_j|$ exists and is finite on $\overset{\circ}{\Phi}$. If $\boldsymbol{\lambda}_\infty = \mathbf{0}$, then:

$$\widehat{\boldsymbol{\varphi}}_T \xrightarrow[T \to \infty]{a.s.} \boldsymbol{\phi}_0.$$

Theorem 3.1 demonstrates that the estimator converges to a biased limit when $\boldsymbol{\lambda}_\infty$ is non-zero. To derive the asymptotic distribution, $\boldsymbol{\lambda}_T$ must converge to $\mathbf{0}$ at an appropriate rate. The asymptotic distribution of the estimator is established under the following assumptions.

A10 $\mathbb{E}\left[\left(\frac{\sigma_1}{w_1}\right)^4\right] < \infty$.

A11 The function f_t is a.s. of class C^2 on $\mathcal{V}(\boldsymbol{\alpha}_0)$, $w_t^{-1} \sup_{\boldsymbol{\alpha}\in\mathcal{V}(\boldsymbol{\alpha}_0)} \left|\left|\frac{\partial f_t}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha})\right|\right|^2$ and $w_t^{-1} \sup_{\boldsymbol{\alpha}\in\mathcal{V}(\boldsymbol{\alpha}_0)} \left|\left|\frac{\partial^2 f_t}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}'}(\boldsymbol{\alpha})\right|\right|$ belong to L^2.

A12 There is a closed and convex subset C of $\mathbb{R}^\nu$ and a sequence of symmetric positive definite $\nu\times\nu$ matrices $(\boldsymbol{J}_T)_{T\in\mathbb{N}}$ converging a.s. to a symmetric positive definite matrix $\boldsymbol{J}$ such that:

$$\sqrt{T}\left(\widehat{\boldsymbol{\phi}}_T - \boldsymbol{\phi}_0\right) = \arg\min_{\boldsymbol{z}\in C} \left(\boldsymbol{Z}_T - \boldsymbol{z}\right)' \boldsymbol{J}_T \left(\boldsymbol{Z}_T - \boldsymbol{z}\right) + o_{\mathbb{P}}(1), \quad \boldsymbol{Z}_T = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \boldsymbol{\Delta}_t \gamma(\eta_t),$$

where $\boldsymbol{\Delta}_t$ is an $\mathcal{F}_{t-1}$-measurable $\nu\times k$ matrix for some positive integer k and $\gamma:\mathbb{R}\to\mathbb{R}^k$ is a measurable function such that $\boldsymbol{\Delta}_t$ and $\gamma(\eta_t)$ belong to L^2, $\mathbb{E}[\gamma(\eta_t)] = \mathbf{0}$ and $\mathbb{V}[\gamma(\eta_t)] =: \boldsymbol{\Gamma}$.

A13 Letting $b_t = \sup_{\boldsymbol{\alpha}\in\Phi\times\Phi} \left|\left|\frac{\partial f_t}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}) - \frac{\partial \widetilde{f}_t}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha})\right|\right|$, the sequences:

$$d_t \sup_{\boldsymbol{\alpha}\in\Phi\times\Phi} \left|\left|\frac{\partial f_t}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha})\right|\right| \left(1 + |y_t| + \sup_{\boldsymbol{\alpha}\in\Phi\times\Phi} |f_t(\boldsymbol{\alpha})|\right), \quad a_t \sup_{\boldsymbol{\alpha}\in\Phi\times\Phi} \left|\left|\frac{\partial f_t}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha})\right|\right|,$$
$$b_t \left(1 + |y_t| + \sup_{\boldsymbol{\alpha}\in\Phi\times\Phi} |f_t(\boldsymbol{\alpha})|\right),$$

are a.s. of order $O(t^{-\kappa})$ for some $\kappa > \frac{1}{2}$.

Assumption A12, based on Francq and Zakoïan (2018), includes cases where the true parameter $\boldsymbol{\phi}_0$ lies on the boundary of the domain for the first-step estimator $\widehat{\boldsymbol{\phi}}_T$. This situation arises, for example, when the first-stage estimator is a QMLE for a GARCH model with at least one parameter equal to zero. Our second step estimator does not have any boundary issues as discussed previously. Additional examples can be found in Francq and Zakoïan (2019) for GARCH models, Francq and Thieu (2019) for APARCHX models, and Andrews (1999) for more general cases. Since this study focuses on a penalized estimator, the true parameter $\boldsymbol{\phi}_0$ is expected to be sparse, making boundary issues in the first stage likely. When the first stage estimator does not have boundary issues, the assumptions reduces to a Bahadur expansion $\sqrt{T}\left(\widehat{\boldsymbol{\phi}}_T - \boldsymbol{\phi}_0\right) = \boldsymbol{Z}_T$. Assumption A13 ensures that initial values are asymptotically irrelevant when deriving the estimator's asymptotic distribution.

To derive the asymptotic distribution, we introduce the following functions. The truncated versions are denoted by tildes in the same way as (10)-(13):

$$E_T(\boldsymbol{v}) = G_T(\boldsymbol{v}) + \sum_{j=1}^{\nu} T\lambda_{j,T}\left(\left|\frac{v_j}{\sqrt{T}} + \phi_{j,0}\right| - |\phi_{j,0}|\right),$$

$$G_T(\boldsymbol{v}) = L_T\left(\widehat{\boldsymbol{\phi}}_T, \frac{\boldsymbol{v}}{\sqrt{T}} + \boldsymbol{\phi}_0\right) - L_T(\boldsymbol{\phi}_0, \boldsymbol{\phi}_0) + \mathbb{K}_T(\boldsymbol{v}), \quad \mathbb{K}_T(\boldsymbol{v}) = \begin{cases} 0 & \text{if } \boldsymbol{v}\in\Upsilon_T, \\ \infty & \text{otherwise} \end{cases}.$$

The function $\boldsymbol{v} \mapsto L_T\left(\widehat{\boldsymbol{\phi}}_T, \frac{\boldsymbol{v}}{\sqrt{T}} + \boldsymbol{\phi}_0\right)$ is defined on $\Upsilon_T := \left\{\sqrt{T}\left(\boldsymbol{x} - \boldsymbol{\phi}_0\right), \boldsymbol{x} \in \Phi\right\}$ and it is clear that $\sqrt{T}\left(\widehat{\boldsymbol{\varphi}}_T - \boldsymbol{\phi}_0\right) = \underset{\boldsymbol{v} \in \Upsilon_T}{\arg\min}\, \widetilde{E}_T\left(\boldsymbol{v}\right)$. We extend the definition domain of the function G_T outside Υ_T with an infinite value by adding $\mathbb{K}_T$. We show that under the previous assumptions, the sequence $\left(\Upsilon_T\right)_{T\geq 1}$ is an exhaustion of $\mathbb{R}^\nu$. This result ensures that G_T is lower semi-continuous then we show that when $\sqrt{T}\boldsymbol{\lambda}_T \underset{T\to\infty}{\longrightarrow} \boldsymbol{\lambda}_\infty < \infty$:

$$\widetilde{E}_T \underset{T\to\infty}{\overset{d}{\longrightarrow}} E_\infty.$$

The previous convergence is to be understood in the sens of epi-convergence in distribution (see Knight (1999)). Since $\left(\Upsilon_T\right)_{T\geq 1}$ spans $\mathbb{R}^\nu$, the limit process is defined on $\mathbb{R}^\nu$ as:

$$E_\infty\left(\boldsymbol{v}\right) = G_\infty\left(\boldsymbol{v}\right) + \sum_{j=1}^{\nu} \lambda_{j,\infty}\left(v_j \operatorname{sign}\left(\phi_{j,0}\right)\mathbb{I}_{\phi_{j,0}\neq 0} + \left|v_j\right|\mathbb{I}_{\phi_{j,0}=0}\right),$$

$$G_\infty\left(\boldsymbol{v}\right) = \begin{pmatrix}\boldsymbol{W}\\ \boldsymbol{Z}^C\\ \boldsymbol{v}\end{pmatrix}' \boldsymbol{\Pi}'_{2,2\nu}\left(\boldsymbol{\Pi}_{1,2\nu} + 2\mathbb{E}\left[P\left(\frac{1}{w_1}\frac{\partial f_1}{\partial \boldsymbol{\alpha}}\left(\boldsymbol{\alpha}_0\right)\right)\right]\boldsymbol{\Pi}_{2,2\nu}\right)\begin{pmatrix}\boldsymbol{W}\\ \boldsymbol{Z}^C\\ \boldsymbol{v}\end{pmatrix},$$

where $P\left(\boldsymbol{x}\right) := \boldsymbol{x}\boldsymbol{x}'$, for some matrix or column vector $\boldsymbol{x}$, and for an integer k, $\boldsymbol{\Pi}_{1,k} = \begin{pmatrix}\boldsymbol{I}_{k\times k} & \boldsymbol{0}_{k\times k}\end{pmatrix}$ and $\boldsymbol{\Pi}_{2,k} = \begin{pmatrix}\boldsymbol{0}_{k\times k} & \boldsymbol{I}_{k\times k}\end{pmatrix}$. The random vector of the limit process is characterized as follows:

$$\begin{pmatrix}\boldsymbol{W}\\ \boldsymbol{Z}\end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix} I\left(\boldsymbol{\phi}_0\right) & R\left(\boldsymbol{\phi}_0\right)\\ R'\left(\boldsymbol{\phi}_0\right) & \boldsymbol{\Sigma}\end{pmatrix}\right),$$

$$I\left(\boldsymbol{\phi}_0\right) = 4\mathbb{E}\left[P\left(\frac{\sigma_1}{w_1^2}\frac{\partial f_1}{\partial \boldsymbol{\alpha}}\left(\boldsymbol{\alpha}_0\right)\right)\right], \quad R\left(\boldsymbol{\phi}_0\right) = -2\mathbb{E}\left[\frac{\sigma_1}{w_1^2}\frac{\partial f_1}{\partial \boldsymbol{\alpha}}\left(\boldsymbol{\alpha}_0\right)\mathbb{E}\left[\eta_1\gamma'\left(\eta_1\right)\right]\boldsymbol{\Delta}_1'\right],$$

$$\boldsymbol{\Sigma} = \mathbb{V}\left[\boldsymbol{\Delta}_1\gamma\left(\eta_1\right)\right] = \mathbb{V}\left[\boldsymbol{\Delta}_1\boldsymbol{\Gamma}^{1/2}\right].$$

The boundary issues of the estimator $\widehat{\boldsymbol{\phi}}_T$, introduced in Assumption A12, lead to the following projection $\boldsymbol{Z}^C = \underset{\boldsymbol{z}\in C}{\arg\min}\, \left(\boldsymbol{Z} - \boldsymbol{z}\right)'\boldsymbol{J}\left(\boldsymbol{Z} - \boldsymbol{z}\right)$.

Remark 3.2. Note that if the components of the matrix function $\gamma\left(\cdot\right)$ are even, and the distribution of η_t is symmetric then $R\left(\boldsymbol{\phi}_0\right) = \boldsymbol{0}$. It is the case for the QMLE for the GARCH model with gaussian η_t.

Theorem 3.2. *Assume that $\sqrt{T}\boldsymbol{\lambda}_T \underset{T\to\infty}{\longrightarrow} \boldsymbol{\lambda}_\infty < +\infty$ component-wise. Then, under Assumptions A1-A3, A5-A13:*

$$\sqrt{T}\left(\widehat{\boldsymbol{\varphi}}_T - \boldsymbol{\phi}_0\right) \underset{T\to\infty}{\overset{d}{\longrightarrow}} \underset{\boldsymbol{v}\in\mathbb{R}^\nu}{\arg\min}\, E_\infty\left(\boldsymbol{v}\right).$$

To study the recovery of the sparse support of the parameter $\boldsymbol{\phi}_0$, we define the sets of active and inactive components. Without loss of generality, we assume that they have the following form:

$$\mathbb{A} = \left\{j \in \left\{1,\ldots,\nu\right\}, \phi_{j,0} \neq 0\right\} = \left\{1,\ldots,\nu_0\right\}, \quad \overline{\mathbb{A}} = \left\{j \in \left\{1,\ldots,\nu\right\}, \phi_{j,0} = 0\right\} = \left\{\nu_0 + 1,\ldots,\nu\right\}.$$

Similarly, we define the sets of active and inactive components of the estimator $\widehat{\boldsymbol{\varphi}}_T$ as:

$$\mathbb{A}_T = \{j \in \{1, \ldots, \nu\}, \widehat{\varphi}_{j,T} \neq 0\}, \quad \overline{\mathbb{A}}_T = \{j \in \{1, \ldots, \nu\}, \widehat{\varphi}_{j,T} = 0\}.$$

A14 $\mathbb{E}\left[P\left(\frac{1}{w_1}\frac{\partial f_1}{\partial \boldsymbol{\varphi}}(\boldsymbol{\alpha}_0)\right)\right]$ is invertible.

Proposition 3.1. *Under the Assumptions of Theorem 3.2 and Assumption A14:*

$$\limsup_{T\to\infty} \mathbb{P}[\mathbb{A}_T = \mathbb{A}] < 1.$$

Proposition 3.1, analogue to Zou (2006), shows that the parameter selection is not consistent with the standard LASSO estimator. In the next section, we introduce the adaptive LASSO estimator and derive its asymptotic properties.

3.2 Adaptive LASSO Estimator

The theory for the adaptive LASSO estimator builds on a similar loss function:

$$E_T^{AL}(\boldsymbol{v}) = \begin{cases} G_T(\boldsymbol{v}) + \sum_{j=1}^{\nu} \frac{T\lambda_{j,T}}{|\widehat{\phi}_{j,T}|^{\tau}}\left(\left|\frac{v_j}{\sqrt{T}} + \phi_{j,0}\right| - |\phi_{j,0}|\right) & \text{if } \forall j \in \{1, \ldots, \nu\}, \widehat{\phi}_{j,T} \neq 0 \\ \infty \quad \text{otherwise.} \end{cases} \tag{24}$$

where τ is a fixed positive constant. The truncated versions of these functions are defined following the same logic as previously. Let $\boldsymbol{A}$ and $\overline{\boldsymbol{A}}$ be the selection matrices defined by removing the rows of $\boldsymbol{I}_{\nu\times\nu}$ corresponding to the inactive and the active components, respectively. In the following, for a vector $\boldsymbol{x} \in \mathbb{R}^{\nu}$ we denote by $\boldsymbol{x}_{\mathbb{A}}$ the sub-vector $\boldsymbol{A}\boldsymbol{x}$, i.e. $\boldsymbol{x} = \left(\boldsymbol{x}_{\mathbb{A}}', \boldsymbol{x}_{\overline{\mathbb{A}}}'\right)'$. We define for any $\boldsymbol{y}$ such that $\boldsymbol{A}'\boldsymbol{y} \in \Upsilon_T$:

$$E_{\mathbb{A},T}^{AL}(\boldsymbol{y}) = E_T^{AL}(\boldsymbol{A}'\boldsymbol{y}), G_{\mathbb{A},T}^{AL}(\boldsymbol{y}) = G_T^{AL}(\boldsymbol{A}'\boldsymbol{y}),$$

with their truncated version defined following the same logic as previously. These functions correspond to (24) with the components $\overline{\mathbb{A}}$ constrained to be 0. To simplify the expressions we define:

$$\mathbb{E}\left[P\left(\frac{1}{w_1}\frac{\partial f_1}{\partial \boldsymbol{\varphi}}(\boldsymbol{\alpha}_0)\right)\right] = \boldsymbol{M} = \begin{pmatrix} \boldsymbol{M}_{11} & \boldsymbol{M}_{12} \\ \boldsymbol{M}_{21} & \boldsymbol{M}_{22} \end{pmatrix}, \quad \mathbb{E}\left[\frac{1}{w_1^2}\frac{\partial f_1}{\partial \boldsymbol{\varphi}}(\boldsymbol{\alpha}_0)\frac{\partial f_1}{\partial \boldsymbol{\phi}'}(\boldsymbol{\alpha}_0)\right]\boldsymbol{Z}^C = \boldsymbol{V},$$

where $\boldsymbol{M}_{11}$ is a $\nu_0 \times \nu_0$ matrix, $\boldsymbol{M}_{22}$ is a $(\nu - \nu_0) \times (\nu - \nu_0)$ matrix and the other matrices are conformable with the partitioning.

Theorem 3.3. *Assume that* $\sqrt{T}\boldsymbol{\lambda}_T \underset{T\to\infty}{\longrightarrow} \boldsymbol{0}$ *and* $T^{\frac{\tau}{2}}\boldsymbol{\lambda}_T \underset{T\to\infty}{\longrightarrow} \infty$*, Under the Assumptions of Theorem 3.2 and Assumption A14:*

$$\sqrt{T}\left(\widehat{\boldsymbol{\varphi}}_{\mathbb{A},T}^{AL} - \boldsymbol{\phi}_{\mathbb{A},0}\right) \underset{T\to\infty}{\overset{d}{\longrightarrow}} -\frac{1}{4}\boldsymbol{M}_{11}^{-1}(\boldsymbol{W}_{\mathbb{A}} + 4\boldsymbol{V}_{\mathbb{A}}),$$

$$\widehat{\boldsymbol{\varphi}}_{\overline{\mathbb{A}},T}^{AL} \underset{T\to\infty}{\overset{\mathbb{P}}{\longrightarrow}} \boldsymbol{0}, \quad \mathbb{P}[\mathbb{A}_T = \mathbb{A}] \underset{T\to\infty}{\longrightarrow} 1.$$

Theorem 3.3 shows that the adaptive LASSO recovers the sparse support with probability tending to one, just as in the oracle property of Fan and Li (2001). However, its asymptotic law is not Gaussian in general. The extra term $\boldsymbol{V}$ in the limit reflects the projection onto the boundary-affected set of the first-step estimator. It leads to a non-standard distribution. If the preliminary estimator has no boundary constraints, then the limit reduces to a multivariate normal as in the classical oracle case. In general, the boundary issues arise when some components of the parameter $\boldsymbol{\phi}_0$ are zero. Therefore, the vector $\boldsymbol{V}_{\mathbb{A}}$ is gaussian in general.

3.3 Application to GARCHX Model

We now apply the two-step estimator to a GARCH model with exogenous covariates $\{\boldsymbol{X}_t, t \in \mathbb{Z}\}$ of dimension r. In practice, it is usual to model financial returns with one lag of persistence. Therefore, we allow the ARCH components to be over-parametrized and we assume that the DGP follows a GARCHX(p,1) dynamic:

$$\epsilon_t = \sigma_t \eta_t, \tag{25}$$

$$\sigma_t^2 = \underline{w}_0 + \sum_{i=1}^{p} a_{i,0}\epsilon_{t-i}^2 + b_0\sigma_{t-1}^2 + \boldsymbol{\varsigma}_0'\boldsymbol{X}_{t-1} > 0, \tag{26}$$

where $\boldsymbol{a}_0 = (a_{1,0}, \ldots, a_{p,0})$, and $\boldsymbol{\phi}_0 = (\omega_0, \boldsymbol{a}_0', b_0, \boldsymbol{\varsigma}_0')' \in \overset{\circ}{\Phi} \subset \mathbb{R}^{1+p+1+r}$ with $\boldsymbol{\phi}_0$ having non-negative components. Squaring $\epsilon_t = \sigma_t\eta_t$ gives:

$$\epsilon_t^2 = \underline{w}_0 + \sum_{i=1}^{p} a_{i,0}\epsilon_{t-i}^2 + b_0\sigma_{t-1}^2 + \boldsymbol{\varsigma}_0'\boldsymbol{X}_{t-1} + \sigma_t^2\left(\eta_t^2 - 1\right). \tag{27}$$

Equation (27) defines a location model. So the framework of Section 2.1 applies with $y_t = \epsilon_t^2$ and $\mu_t = \sigma_t^2$. In this case:

$$f_t(\boldsymbol{\alpha}) = \left(1, \epsilon_{t-1}^2, \ldots, \epsilon_{t-p}^2, \sigma_{t-1}^2(\boldsymbol{\phi}), \boldsymbol{X}_{t-1}'\right)\boldsymbol{\varphi}, \quad l_t(\boldsymbol{\alpha}) = \left(\frac{\epsilon_t^2 - f_t(\boldsymbol{\alpha})}{w_t}\right)^2.$$

An explicit form of the weight process is given later. The following assumptions are sufficient to prove the theorems of Section 3 in the case of GARCHX(p,1).

A15 The process $\left\{(\eta_t, \boldsymbol{X}_t')', t \in \mathbb{Z}\right\}$ is strictly stationary and ergodic such that $\exists s > 0, \mathbb{E}\left[||\boldsymbol{X}||^s\right] < \infty$ and the usual top-Lyapunov condition is satisfied.

A16 $\exists \rho > 0, \forall \boldsymbol{\phi} \in \Phi, |b| \leq \rho < 1$.

A17 $\forall \boldsymbol{x} \in \mathbb{R}^r \setminus \{\boldsymbol{0}\}$, $\boldsymbol{x}'\boldsymbol{X}_1$ is not degenerated.

As shown in Lemmas 1 and 2 of Francq and Thieu (2019), under the previous Assumptions, there exist a unique stationary, ergodic and non-anticipative solution to (25)-(26) with a small order moment $2s$. In the following, we assume that $\{\epsilon_t, t \in \mathbb{Z}\}$ is the solution. Assumption A17 suffices to ensure the condition of Assumption A14. Under Assumption A16 and since Φ is compact, we let $\sup_{b \in \Phi}|b| := \rho < c < 1$ and define $w_t = 1 + \sum_{i \in \mathbb{N}^*} c^i \left(\epsilon_{t-i}^2 + ||\boldsymbol{X}_{t-i}||_1\right)$. We also assume that a first step estimator $\widehat{\boldsymbol{\phi}}_T$ is available.

Remark 3.3. The first step estimator $\widehat{\boldsymbol{\phi}}_T$ could be the QMLE for example. Note that the parameter set Φ used for the second-stage estimator does not coincide with the parameter set of the QML estimation context. The QMLE is constrained whereas the WLSE is not.

Corollary 3.1. *Under Assumptions A15-A16, if $\widehat{\boldsymbol{\phi}}_T$ satisfies Assumption A8 then the Theorem 3.1 holds. Moreover, if $\mathbb{E}\left[\eta_t^4\right] < \infty$ and $\widehat{\boldsymbol{\phi}}_T$ satisfies Assumption A12 then the Theorem 3.2 holds. Under Assumption A17 gives Theorem 3.3.*

4 Illustration Based on Financial Data and Simulations

In this section, we illustrate the two-steps method by an application to real data. We compare the results and the execution time of the two-stage approach to Chan et al. (2020). We also describe the tuning of the LASSO hyper-parameter. Then we study the finite-sample properties with Monte Carlo experiments.

4.1 Monthly Interest Rate on Three-Month US Government Treasury Bills

We compare our convex penalized estimation to the non-convex iterative approach of Chan et al. (2020) for ARMA models. We use the same data-set (length 461) of the log-differential of the monthly interest rate on three-month government Treasury bills for the period 1950 to 1988. The data is represented in Figure 1. In the light of Figure 2 an ARMA(7,7) can be suggested as an

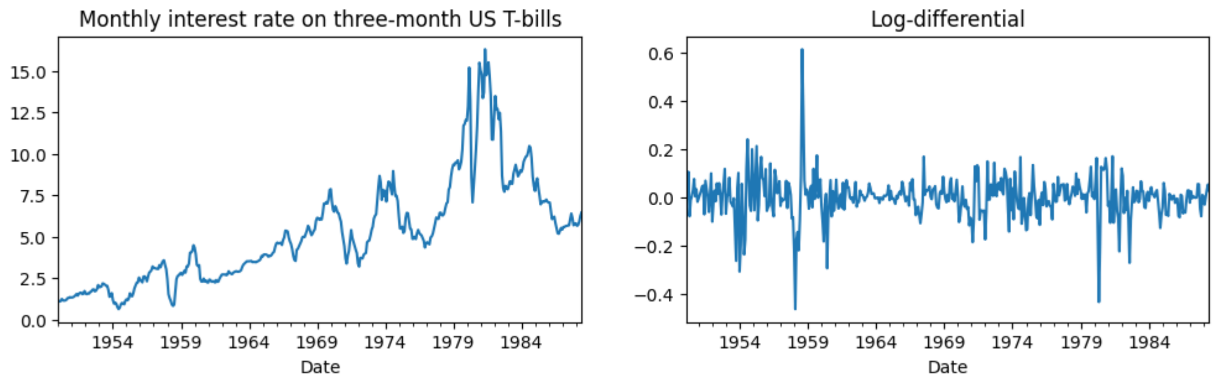

Figure 1: Monthly interest rate on three-month government Treasury bills, 1950–1988.

initial over-parameterized model. This over-parameterization leads to 16384 possible sub-models.

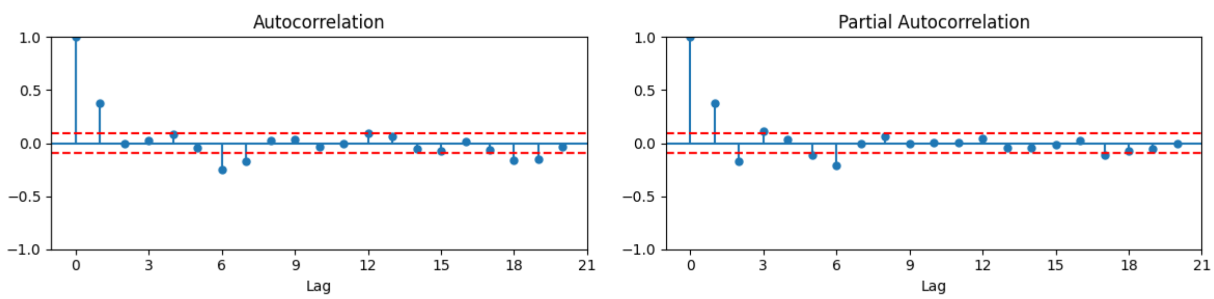

Figure 2: ACF (left) and PACF (right) of the log-differential series.

We use our two stage procedure to perform a sparse estimation. The first-stage estimator is a conditional LSE, then the LARS-LASSO algorithm is computed. We use the adaptive power $\tau = 2$ and assume that the tuning parameter $\boldsymbol{\lambda}_T$ is the same for all the parameters $\lambda_{1,T} = \ldots = \lambda_{14,T}$. Then, to avoid tuning this hyper-parameter, we perform a post-LASSO estimation at each LARS-LASSO step (i.e. the gray vertical lines in Figure 3). The post-LASSO calculations can be done

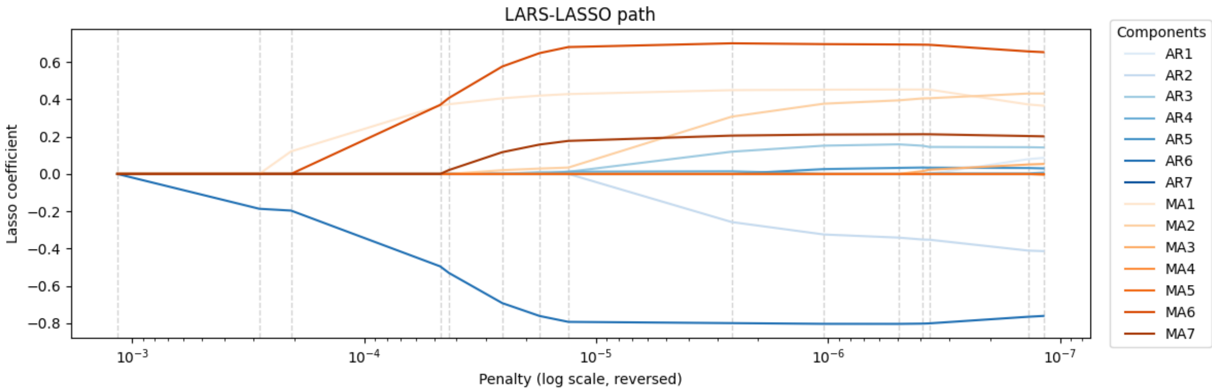

Figure 3: LARS-LASSO path of the second-step estimator.

in parallel when it is advantageous. The best model is selected according to the BIC criterion. The non-convex method of Chan et al. (2020) is used as described in their paper with the choice of the hyper-parameter guided by a BIC-like information criterion. We stop the algorithm if the 1000^{th} iteration is reached or if the error is bellow 10^{-6}. Both estimators select AR6, MA1 and MA6 coefficients. The non-convex method gives AR6 $= -0.38$, MA1 $= 0.43$ and MA6 $= 0.19$ whereas our convex method gives AR6 $= -0.42$, MA1 $= 0.43$ and MA6 $= 0.23$. The results of the non-convex method are slightly different of the those in the original paper due to the sensitivity to the initial conditions. The algorithm stops after reaching the maximum number of iterations. However, the selected set of parameters is the same as our convex approach. The runtime of our approach is 485 milliseconds whereas the non-convex approach takes 5.2 seconds to produce the results. Our approach is more than 10 times faster. The experiment was done 100 times to compare the average runtime of the two methods on a MacBook Pro with Apple M2 Pro processor and 16 GB or RAM. The model selected by our method produces non-correlated residuals (see Table 1 and Figure 4) With this comparison we show that our method produces the same results

Table 1: Ljung-Box tests of the model residuals.

Lags	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Stat	0.32	3.07	3.10	3.10	3.27	3.28	3.98	3.99	4.73	4.88	5.23	5.37	10.98	11.58	13.62	15.31	20.01	20.03	21.08	25.96	26.17	26.58
PVal	0.96	0.55	0.68	0.80	0.86	0.92	0.91	0.95	0.94	0.96	0.97	0.98	0.75	0.77	0.69	0.64	0.39	0.46	0.45	0.25	0.29	0.32

as Chan et al. (2020) with minimal computational cost, leading to a valid sparse model. We also show how our approach reduces the dimension of the selection problem: Reducing the number of possible sub-models from 16384 to 15. Moreover, the method is designed to take advantage of the efficiency of the LARS-LASSO algorithm and, for large models, the post-LASSO estimation can be done in parallel leading to a shorter computational time. Here, the post-LASSO is done sequentially.

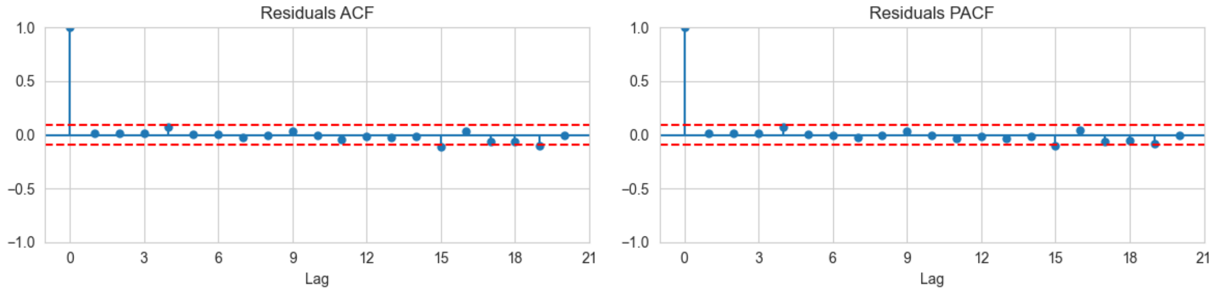

Figure 4: ACF and PACF of the residuals of the model estimated by the two-step method.

4.2 Monte Carlo Experiment

This section provides Monte Carlo simulations with 500 trajectories to investigate the finite sample properties of the penalized two-stage estimator. In the following, the adaptive LASSO power is $\tau = 2$ for both location and scale processes.

4.2.1 ARMAX-GARCHX

We start by estimating an ARMAX(3,3)-GARCHX(2,2) model with two exogenous components in ARMA and GARCH parts. The true DGP is a sparse ARMAX(1,2)-GARCHX(1,1) model. The parameters of this experiment are described in Table 2. The exogenous components of the ARMA part, denoted $\boldsymbol{Y}$, are i.i.d. $\mathcal{N}(0,1)$ and the exogenous components of the GARCH part, denoted $\boldsymbol{X}$, are i.i.d. $\chi^2(1)$. The first step estimation is given by the QML. As in the previous experiment, the tuning of the hyper-parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\iota}$ is avoided by performing a post-LASSO. Table 3 shows that the true sparse parameter is accurately recovered as the sample size grows.

Table 2: Non-zero simulation coefficients (all remaining AR, MA, ARCH/PERSISTENCE, Y and X terms are zero). Coefficients marked † are not penalized.

AR1	MA1	Y0	INTERCEPT†	ARCH1†	PERSISTENCE1†	X0
0.90	–0.80	–0.50	0.01	0.09	0.84	0.30

The accuracy is higher for the ARMAX parameters than for the GARCHX parameters. This Monte Carlo experiment demonstrates that the finite sample performance of the procedure aligns with theoretical expectations. The sparse support is accurately identified as the sample size grows.

4.2.2 GJR-GARCHX

We perform a second experiment on a GJR-GARCHX(2,2,1) model with 5 exogenous components. The parameters of this experiment are described in Table 4 with the same adaptive power $\tau = 2$. The exogenous parameters are the squares of independent GARCH(1,1) trajectories, with parameters estimated on the log-returns of the S&P 500 index, NASDAQ Composite, Russell 2000, FTSE 100 and DAX 30, on the period from 2000-01-01 to 2019-01-01. All the parameters are penalized, except the intercept. Table 5 shows the percentage of times each coefficient is picked by BIC along the LARS-LASSO path for sample sizes $T = 500, 1000, 4000, 5000$. Selection of

Table 3: Selection rates (%) for penalized coefficients in the ARMAX–GARCHX experiment.

	Parameter	$T = 500$	$T = 1000$	$T = 2000$	$T = 4000$
ARMA	AR_1	100.00	100.00	100.00	100.00
	AR_2	0.00	0.00	0.00	0.00
	AR_3	0.00	3.60	0.00	0.00
	MA_1	0.40	0.40	0.00	0.00
	MA_2	100.00	100.00	100.00	100.00
	MA_3	0.20	0.00	0.00	0.00
ARMA Exogenous	Y_0	100.00	100.00	100.00	100.00
	Y_1	0.80	0.60	0.80	0.80
GARCH	$ARCH_2$	0.00	1.60	3.00	4.40
	$PERSISTENCE_2$	48.60	37.60	33.80	14.20
GARCH Exogenous	X_0	55.00	78.40	98.00	100.00
	X_1	0.60	0.40	1.60	2.20

true coefficients rises steadily with n: all active terms exceed 80% for $n \geq 4000$. False-positive rates stay bellow 8% once $n \geq 2000$. LEVERAGE2 is falsely selected in 30% of trajectories at $n = 1000$, but this over-selection disappears as n grows. The selection of ARCH1 and X0 terms are slower but reaches satisfactory rates for $n \geq 4000$. Overall, the probability of recovering the correct model grows when n grows.

Table 4: Non-zero simulation coefficients for the GJR–GARCHX experiment (all other terms = 0). Coefficients marked † are not penalized.

INTERCEPT†	ARCH1	LEVERAGE1	PERSISTENCE1	X0	X1
0.03	0.02	0.15	0.90	0.30	0.50

Table 5: Selection rates (%) for penalized coefficients in the GJR–GARCHX experiment.

	Parameter	$T = 500$	$T = 1000$	$T = 2000$	$T = 4000$	$T = 5000$
GJR-GARCH	INTERCEPT	100.00	100.00	100.00	100.00	100.00
	$ARCH_1$	0.00	0.00	0.00	0.00	100.00
	$ARCH_2$	0.00	0.00	0.00	0.00	0.00
	$LEVERAGE_1$	100.00	100.00	0.00	100.00	100.00
	$LEVERAGE_2$	0.00	0.00	100.00	0.00	0.00
	$PERSISTENCE_1$	100.00	100.00	100.00	100.00	100.00
Exogenous	X_0	0.00	0.00	0.00	0.00	100.00
	X_1	0.00	100.00	0.00	100.00	100.00
	X_2	0.00	0.00	0.00	0.00	0.00
	X_3	0.00	0.00	0.00	0.00	0.00
	X_4	0.00	0.00	0.00	0.00	0.00

5 Conclusion

This paper develops a two-step, convex L^1-penalized estimator for dynamic location–scale models. The first step relies on a preliminary $\sqrt{T}$-consistent estimator. The second step performs a weighted least-squares LASSO while preserving convexity. Under mild assumptions the estimator is consistent, and its asymptotic distribution is obtained. The adaptive version consistently recovers the true support. Convexity lets the LARS algorithm trace the full penalty path in milliseconds. Simulations report selection accuracy above 95 % for sample sizes typical of financial data. On real financial data the method matches published coefficients while reducing computation time by a factor of ten.

Future work will extend the two-step scheme to multivariate models, enabling joint estimation of cross-asset volatility and improving portfolio-level risk metrics such as multivariate Value-at-Risk.

6 Acknowledgments

The author thanks Christian FRANCQ and Jean-Michel ZAKOÏAN for their guidance and feedback.

Appendix

The appendix is organized in two parts: The first part presents the theoretical results for the scale-parameter estimator $\widehat{\boldsymbol{\varphi}}_T$ and its adaptive version, while the second part gathers the technical proofs and supporting lemmas.

I Theoretical Results for the Scale Parameter

In the following, write $\mathcal{V}(\boldsymbol{\theta}_0)$ for a neighborhood of $\boldsymbol{\theta}_0$, $\mathcal{V}(\boldsymbol{\rho}_0) = \mathcal{V}(\boldsymbol{\phi}_0) \times \mathcal{V}(\boldsymbol{\theta}_0)$ and $\mathcal{V}(\boldsymbol{\beta}_0) = \mathcal{V}(\boldsymbol{\rho}_0) \times \mathcal{V}(\boldsymbol{\theta}_0)$. We show the strong consistency of $\widehat{\boldsymbol{\vartheta}}_T$ under the following assumptions.

A18 $\left(1 + y_t^2 + \sup_{\boldsymbol{\phi}\in\Phi} \mu_t^2(\boldsymbol{\phi}) + \sup_{\boldsymbol{\beta}\in\Phi\times\Theta\times\Theta} |g_t(\boldsymbol{\beta})|\right) e_t \xrightarrow[t\to\infty]{a.s.} 0$, with

$e_t = \sup_{\boldsymbol{\phi}\in\Phi} |\mu_t(\boldsymbol{\phi}) - \widetilde{\mu}(\boldsymbol{\phi})| \left(1 + |y_t| + \sup_{\boldsymbol{\phi}\in\Phi} |\mu_t(\boldsymbol{\phi})|\right) + \sup_{\boldsymbol{\beta}\in\Phi\times\Theta\times\Theta} |g_t(\boldsymbol{\beta}) - \widetilde{g}_t(\boldsymbol{\beta})|.$

A19 $\left(1 + y_t^4 + \sup_{\boldsymbol{\phi}\in\Phi} \mu_t^4(\boldsymbol{\phi}) + \sup_{\boldsymbol{\beta}\in\Phi\times\Theta\times\Theta} g_t^2(\boldsymbol{\beta})\right) c_t \xrightarrow[t\to\infty]{a.s.} 0$, with $c_t = \left|\widetilde{\omega}_t^2 - \omega_t^2\right|$.

A20 $\mathbb{E}\left[\left(\frac{\sigma_1^2}{\omega_1}\right)^2\right] < \infty$.

A21 $\mathcal{V}(\boldsymbol{\rho}_0) \subset \Phi \times \Theta$.

A22 The functions $g_1(\cdot)$ and $\mu_1(\cdot)$ are a.s. of class C^1 on $\mathring{\Phi} \times \mathring{\Theta} \times \mathring{\Theta}$ and $\mathring{\Phi}$, respectively, and $\omega_1^{-1} \sup_{\boldsymbol{\beta}\in\mathcal{V}(\boldsymbol{\rho}_0)\times\mathring{\Theta}} \left|\left|\frac{\partial g_1}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta})\right|\right|$ and $\omega_1^{-1} \sup_{\boldsymbol{\phi}\in\mathcal{V}(\boldsymbol{\phi}_0)} \left|\left|\frac{\partial \mu_1}{\partial \boldsymbol{\phi}}(\boldsymbol{\phi})\right|\right|^2$ belong to L^2.

A23 $\mathbf{0} \in \mathring{\Theta}$.

A24 $\widehat{\boldsymbol{\rho}}_T \xrightarrow[T\to\infty]{a.s.} \boldsymbol{\rho}_0$.

A25 $\forall \boldsymbol{\rho} \in \mathcal{V}(\boldsymbol{\rho}_0)$, the functions $\ell_t(\boldsymbol{\rho}, \cdot)$ and $\widetilde{\ell}_t(\boldsymbol{\rho}, \cdot)$ are a.s. strictly convex on Θ.

The assumptions above align with those in Section 3, but add regularity conditions on μ_t. Because the estimator now uses the residuals $\widetilde{\epsilon}_t\left(\widehat{\boldsymbol{\phi}}_T\right)$ instead of the unobserved process ϵ_t, these extra conditions ensure adequate control of the residual terms.

Theorem I.1. *Assume that $\boldsymbol{\iota}_T \xrightarrow[T\to\infty]{} \boldsymbol{\iota}_\infty < +\infty$ component-wise. Then, under Assumptions A1, A18-A25:*

$$\widehat{\boldsymbol{\vartheta}}_T \xrightarrow[T\to\infty]{a.s.} \arg\min_{\boldsymbol{\vartheta}\in\Theta} \mathcal{Q}_\infty(\boldsymbol{\rho}_0, \boldsymbol{\vartheta}),$$

where $\mathcal{Q}_\infty(\boldsymbol{\rho}_0, \boldsymbol{\vartheta}) := \mathbb{E}[\ell_1(\boldsymbol{\rho}_0, \boldsymbol{\vartheta})] + \sum_{j=1}^n \iota_{j,\infty} |\vartheta_j|$ exists and is finite on $\mathring{\Theta}$. If $\boldsymbol{\iota}_\infty = \mathbf{0}$, then:

$$\widehat{\boldsymbol{\vartheta}}_T \xrightarrow[T\to\infty]{a.s.} \boldsymbol{\theta}_0.$$

A26 $\mathbb{E}\left[\left(\frac{\sigma_1^2}{\omega_1}\right)^4 + \eta_1^6\right] < \infty$ and $\mathbb{E}\left[\eta_1^3\right] = 0$.

A27 The functions $g_1(\cdot)$ and $\mu_1(\cdot)$ are a.s. of class C^2 on $\mathcal{V}(\boldsymbol{\beta}_0)$ and $\mathcal{V}(\boldsymbol{\phi}_0)$, respectively, and $\omega_1^{-1}\sup_{\boldsymbol{\beta}\in\mathcal{V}(\boldsymbol{\beta}_0)}\left|\left|\frac{\partial g_1}{\partial\boldsymbol{\beta}}(\boldsymbol{\beta})\right|\right|$, $\omega_1^{-1}\sup_{\boldsymbol{\beta}\in\mathcal{V}(\boldsymbol{\beta}_0)}\left|\left|\frac{\partial^2 g_1}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}(\boldsymbol{\beta})\right|\right|$, $\omega_1^{-1}\sup_{\boldsymbol{\phi}\in\mathcal{V}(\boldsymbol{\phi}_0)}\left|\left|\frac{\partial\mu_1}{\partial\boldsymbol{\phi}}(\boldsymbol{\phi})\right|\right|^2$, and $\omega_1^{-1}\sup_{\boldsymbol{\phi}\in\mathcal{V}(\boldsymbol{\phi}_0)}\left|\left|\frac{\partial^2\mu_1}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'}(\boldsymbol{\phi})\right|\right|^2$ belong to L^4.

A28 There is a closed and convex subset $\mathcal{C}$ of $\mathbb{R}^{\nu+n}$ and a sequence of symmetric positive definite $(\nu+n)\times(\nu+n)$ matrices $(\boldsymbol{\mathcal{J}}_T)_{T\in\mathbb{N}}$ converging a.s. to a symmetric positive definite matrix $\boldsymbol{\mathcal{J}}$ such that:

$$\sqrt{T}\left(\widehat{\boldsymbol{\rho}}_T-\boldsymbol{\rho}_0\right)=\arg\min_{\boldsymbol{z}\in\mathcal{C}}\ \left(\boldsymbol{\mathcal{Z}}_T-\boldsymbol{z}\right)'\boldsymbol{\mathcal{J}}_T\left(\boldsymbol{\mathcal{Z}}_T-\boldsymbol{z}\right)+o_{\mathbb{P}}(1),\quad \boldsymbol{\mathcal{Z}}_T=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{\Lambda}_t\xi(\eta_t),$$

where $\boldsymbol{\Lambda}_t$ is an $\mathcal{F}_{t-1}$-measurable $(\nu+n)\times k$ matrix for some positive integer k and $\xi:\mathbb{R}\to\mathbb{R}^k$ is a measurable function such that $\boldsymbol{\Lambda}_t$ and $\xi(\eta_t)$ belong to L^2, $\mathbb{E}[\xi(\eta_t)]=\mathbf{0}$ and $\mathbb{V}[\xi(\eta_t)]=:\boldsymbol{\chi}$.

A29 Letting

$$\begin{aligned}\varrho_t=&\sup_{\boldsymbol{\phi}\in\Phi}\left|\mu_t(\boldsymbol{\phi})-\widetilde{\mu}_t(\boldsymbol{\phi})\right|\sup_{\boldsymbol{\phi}\in\Phi}\left|\left|\frac{\partial\mu_t}{\partial\boldsymbol{\phi}}(\boldsymbol{\phi})\right|\right|+\sup_{\boldsymbol{\beta}\in\Phi\times\Theta\times\Theta}\left|\left|\frac{\partial g_t}{\partial\boldsymbol{\beta}}(\boldsymbol{\beta})-\frac{\partial\widetilde{g}_t}{\partial\boldsymbol{\beta}}(\boldsymbol{\beta})\right|\right|\\&+\left(1+|y_t|+\sup_{\boldsymbol{\phi}\in\Phi}|\mu_t(\boldsymbol{\phi})|\right)\sup_{\boldsymbol{\phi}\in\Phi}\left|\left|\frac{\partial\mu_t}{\partial\boldsymbol{\phi}}(\boldsymbol{\phi})-\frac{\partial\widetilde{\mu}_t}{\partial\boldsymbol{\phi}}(\boldsymbol{\phi})\right|\right|,\end{aligned}$$

the sequences

$$\begin{aligned}&c_t\left(1+|y_t|^2+\sup_{\boldsymbol{\phi}\in\Phi}|\mu_t(\boldsymbol{\phi})|^2+\sup_{\boldsymbol{\beta}\in\Phi\times\Theta\times\Theta}|g_t(\boldsymbol{\beta})|\right)\left(1+|y_t|+\sup_{\boldsymbol{\phi}\in\Phi}|\mu_t(\boldsymbol{\phi})|\right)\sup_{\boldsymbol{\phi}\in\Phi}\left|\left|\frac{\partial\mu_t}{\partial\boldsymbol{\phi}}(\boldsymbol{\phi})\right|\right|,\\&c_t\left(1+|y_t|^2+\sup_{\boldsymbol{\phi}\in\Phi}|\mu_t(\boldsymbol{\phi})|^2+\sup_{\boldsymbol{\beta}\in\Phi\times\Theta\times\Theta}|g_t(\boldsymbol{\beta})|\right)\sup_{\boldsymbol{\beta}\in\Phi\times\Theta\times\Theta}\left|\left|\frac{\partial g_t}{\partial\boldsymbol{\beta}}(\boldsymbol{\beta})\right|\right|,\\&e_t\left(\left(1+|y_t|+\sup_{\boldsymbol{\phi}\in\Phi}|\mu_t(\boldsymbol{\phi})|\right)\sup_{\boldsymbol{\phi}\in\Phi}\left|\left|\frac{\partial\mu_t}{\partial\boldsymbol{\phi}}(\boldsymbol{\phi})\right|\right|+\sup_{\boldsymbol{\beta}\in\Phi\times\Theta\times\Theta}\left|\left|\frac{\partial g_t}{\partial\boldsymbol{\beta}}(\boldsymbol{\beta})\right|\right|\right),\\&\varrho_t\left(1+|y_t|^2+\sup_{\boldsymbol{\phi}\in\Phi}|\mu_t(\boldsymbol{\phi})|^2+\sup_{\boldsymbol{\beta}\in\Phi\times\Theta\times\Theta}|g_t(\boldsymbol{\beta})|\right),\end{aligned}$$

are a.s. of order $O\left(t^{-\kappa}\right)$ for some $\kappa>\frac{1}{2}$.

In this context, we need Assumption A26 to ensure that the noise η_t is symmetric. It is needed to derive the asymptotic distribution of the estimator $\widehat{\boldsymbol{\vartheta}}_T$. We introduce the following functions:

$$\mathcal{E}_T(\boldsymbol{\psi})=\mathcal{G}_T(\boldsymbol{\psi})+\sum_{j=1}^{n}T\iota_{j,T}\left(\left|\frac{\psi_j}{\sqrt{T}}+\theta_{j,0}\right|-|\theta_{j,0}|\right),$$

$$\mathcal{G}_T(\boldsymbol{\psi})=\mathcal{L}_T\left(\widehat{\boldsymbol{\rho}}_T,\frac{\boldsymbol{\psi}}{\sqrt{T}}+\boldsymbol{\theta}_0\right)-\mathcal{L}_T(\boldsymbol{\rho}_0,\boldsymbol{\theta}_0)+\mathbb{S}_T(\boldsymbol{\psi}),\quad \mathbb{S}_T(\boldsymbol{\psi})=\begin{cases}0 & \text{if }\boldsymbol{\psi}\in\Psi_T,\\ \infty & \text{otherwise}\end{cases}.$$

The function $\boldsymbol{\psi}\mapsto L_T\left(\widehat{\boldsymbol{\rho}}_T,\frac{\boldsymbol{\psi}}{\sqrt{T}}+\boldsymbol{\theta}_0\right)$ is defined on $\Psi_T:=\left\{\sqrt{T}\left(\boldsymbol{x}-\boldsymbol{\theta}_0\right),\boldsymbol{x}\in\Theta\right\}$. It is clear that

$\sqrt{T}\left(\widehat{\boldsymbol{\vartheta}}_T - \boldsymbol{\theta}_0\right) = \underset{\boldsymbol{\psi}\in\Psi_T}{\arg\min}\, \widetilde{\mathcal{E}}_T(\boldsymbol{\psi})$. The epi-limits are defined on $\mathbb{R}^n$ as:

$$\mathcal{E}_\infty(\boldsymbol{\psi}) = \mathcal{G}_\infty(\boldsymbol{\psi}) + \sum_{j=1}^{n} \iota_{j,\infty}\left(\psi_j \operatorname{sign}(\theta_{j,0})\, \mathbb{I}_{\theta_{j,0}\neq 0} + |\psi_j|\, \mathbb{I}_{\theta_{j,0}=0}\right),$$

$$\mathcal{G}_\infty(\boldsymbol{\psi}) = \begin{pmatrix} \mathcal{W} \\ \mathcal{Z}^C \\ \boldsymbol{\psi} \end{pmatrix}' \boldsymbol{\Pi}_{2,\nu+2n}' \left(\boldsymbol{\Pi}_{1,\nu+2n} + 2\mathbb{E}\left[\frac{1}{\omega_1^2} P\left(2\epsilon_1 \frac{\partial \mu_1}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0) + \frac{\partial g_1}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0)\right)\right] \boldsymbol{\Pi}_{2,\nu+2n}\right) \begin{pmatrix} \mathcal{W} \\ \mathcal{Z}^C \\ \boldsymbol{\psi} \end{pmatrix}.$$

The random vector of the limit process is characterized as follows:

$$\begin{pmatrix} \mathcal{W} \\ \mathcal{Z} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathcal{I}(\boldsymbol{\rho}_0) & \mathcal{R}(\boldsymbol{\rho}_0) \\ \mathcal{R}'(\boldsymbol{\rho}_0) & \boldsymbol{\Sigma} \end{pmatrix}\right),$$

$$\mathcal{I}(\boldsymbol{\rho}_0) = 4\mathbb{E}\left[P\left(\frac{\epsilon_1^2 - \sigma_1^2}{\omega_1^2}\left(2\epsilon_1 \frac{\partial \mu_1}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0) + \frac{\partial g_1}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0)\right)\right)\right],$$

$$\mathcal{R}(\boldsymbol{\rho}_0) = -2\mathbb{E}\left[\frac{\epsilon_1^2 - \sigma_1^2}{\omega_1^2}\left(2\epsilon_1 \frac{\partial \mu_1}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0) + \frac{\partial g_1}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0)\right) \xi'(\eta_1)\, \boldsymbol{\Lambda}_1'\right], \quad \boldsymbol{\Xi} = \mathbb{V}\left[\boldsymbol{\Lambda}_1 \xi(\eta_1)\right] = \mathbb{V}\left[\boldsymbol{\Lambda}_1 \boldsymbol{\chi}^{1/2}\right].$$

Remark I.1. We can write:

$$\mathcal{R}(\boldsymbol{\rho}_0) = -4\mathbb{E}\left[\frac{\sigma_1^3}{\omega_1^2} \frac{\partial \mu_1}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0)\, \mathbb{E}\left[\eta_1\left(\eta_1^2 - 1\right) \xi'(\eta_1)\right] \boldsymbol{\Lambda}_1'\right] - 2\mathbb{E}\left[\frac{\sigma_1^2}{\omega_1^2} \frac{\partial g_1}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0)\, \mathbb{E}\left[\left(\eta_1^2 - 1\right) \xi'(\eta_1)\right] \boldsymbol{\Lambda}_1'\right].$$

Depending on the parity of the components of the function $\xi(\cdot)$, either one term or the other vanishes.

Theorem I.2. *Assume that $\sqrt{T}\boldsymbol{\iota}_T \underset{T\to\infty}{\longrightarrow} \boldsymbol{\iota}_\infty < +\infty$ component-wise. Then, under Assumptions A1, A18-A19, A21-A29:*

$$\sqrt{T}\left(\widehat{\boldsymbol{\vartheta}}_T - \boldsymbol{\theta}_0\right) \underset{T\to\infty}{\overset{d}{\longrightarrow}} \underset{\boldsymbol{\psi}\in\mathbb{R}^n}{\arg\min}\, \mathcal{E}_\infty(\boldsymbol{\psi}).$$

To study the recovery of the sparse support of the parameter $\boldsymbol{\theta}_0$, we define the sets of active and inactive components. Without loss of generality, we assume that they have the following form:

$$\mathbb{B} = \{j \in \{1,\ldots,n\}, \theta_{j,0} \neq 0\} = \{1,\ldots,n_0\}, \quad \overline{\mathbb{B}} = \{j \in \{1,\ldots,n\}, \theta_{j,0} = 0\} = \{n_0+1,\ldots,n\}.$$

Similarly, we define the sets of active and inactive components of the estimator $\widehat{\boldsymbol{\vartheta}}_T$ as:

$$\mathbb{B}_T = \left\{j \in \{1,\ldots,n\}, \widehat{\vartheta}_{j,T} \neq 0\right\}, \quad \overline{\mathbb{B}}_T = \left\{j \in \{1,\ldots,n\}, \widehat{\vartheta}_{j,T} = 0\right\}.$$

A30 $\mathbb{E}\left[P\left(\frac{1}{\omega_1}\frac{\partial g_1}{\partial \boldsymbol{\vartheta}}(\boldsymbol{\beta}_0)\right)\right]$ is invertible.

Proposition I.1. *Under the Assumptions of Theorem I.2 and Assumption A30:*

$$\limsup_{T\to\infty} \mathbb{P}\left[\mathbb{B}_T = \mathbb{B}\right] < 1.$$

To consistently recover the sparse support of $\boldsymbol{\theta}_0$, we define the following modified loss function to perform an adaptive LASSO:

$$\mathcal{E}_T^{AL}(\boldsymbol{\psi}) = \begin{cases} \mathcal{G}_T(\boldsymbol{\psi}) + \sum_{j=1}^{n} \frac{T\iota_{j,T}}{\left|\widehat{\theta}_{j,T}\right|^{\tau}} \left(\left| \frac{\psi_j}{\sqrt{T}} + \theta_{j,0} \right| - |\theta_{j,0}| \right) & \text{if } \forall j \in \{1,\dots,n\}, \widehat{\theta}_{j,T} \neq 0 \\ \infty \quad \text{otherwise.} \end{cases} \tag{28}$$

Let $\boldsymbol{B}$ and $\overline{\boldsymbol{B}}$ be the selection matrices defined by removing the rows of $\boldsymbol{I}_{n\times n}$ corresponding to the inactive and the active components, respectively. In the following, for a vector $\boldsymbol{x} \in \mathbb{R}^n$ we denote by $\boldsymbol{x}_{\mathbb{B}}$ sub-vector $\boldsymbol{B}\boldsymbol{x}$, i.e. $\boldsymbol{x} = \left(\boldsymbol{x}'_{\mathbb{B}}, \boldsymbol{x}'_{\overline{\mathbb{B}}}\right)$. We define for any $\boldsymbol{y}$ such that $\boldsymbol{B}'\boldsymbol{y} \in \Psi_T$:

$$\mathcal{E}_{\mathbb{B},T}^{AL}(\boldsymbol{y}) = \mathcal{E}_T^{AL}(\boldsymbol{B}'\boldsymbol{y}), \quad \mathcal{G}_{\mathbb{B},T}^{AL}(\boldsymbol{y}) = \mathcal{G}_T^{AL}(\boldsymbol{B}'\boldsymbol{y}),$$

with their truncated version defined following the same logic as previously. These functions correspond to (28) with the components $\overline{\mathbb{B}}$ constrained to be 0. To simplify the expressions we define:

$$\mathbb{E}\left[P\left(\frac{1}{\omega_1}\frac{\partial g_1}{\partial \boldsymbol{\vartheta}}(\boldsymbol{\beta}_0)\right)\right] = \boldsymbol{\mathcal{M}} = \begin{pmatrix} \boldsymbol{\mathcal{M}}_{11} & \boldsymbol{\mathcal{M}}_{12} \\ \boldsymbol{\mathcal{M}}_{21} & \boldsymbol{\mathcal{M}}_{22} \end{pmatrix}, \quad \mathbb{E}\left[\frac{1}{\omega_1^2}\frac{\partial g_1}{\partial \boldsymbol{\vartheta}'}(\boldsymbol{\beta}_0)\left(2\epsilon_1 \frac{\partial \mu_1}{\partial \boldsymbol{\rho}'}(\boldsymbol{\rho}_0) + \frac{\partial g_1}{\partial \boldsymbol{\rho}'}(\boldsymbol{\beta}_0)\right)\right]\boldsymbol{\mathcal{Z}}^{\mathcal{C}} = \boldsymbol{\zeta},$$

where $\boldsymbol{\mathcal{M}}_{11}$ is a $n_0 \times n_0$ matrix, $\boldsymbol{\mathcal{M}}_{22}$ is a $(n-n_0)\times(n-n_0)$ matrix and the other matrices are conformable with the partitioning.

Theorem I.3. *Assume that* $\sqrt{T}\boldsymbol{\iota}_T \underset{T\to\infty}{\longrightarrow} \mathbf{0}$ *and* $T^{\frac{\tau}{2}}\boldsymbol{\iota}_T \underset{T\to\infty}{\longrightarrow} \infty$*, Under the Assumptions of Theorem I.2 and Assumption A30:*

$$\sqrt{T}\left(\widehat{\boldsymbol{\vartheta}}_{\mathbb{B},T}^{AL} - \boldsymbol{\theta}_{\mathbb{B},0}\right) \underset{T\to\infty}{\overset{d}{\longrightarrow}} -\frac{1}{4}\boldsymbol{\mathcal{M}}_{11}^{-1}\left(\boldsymbol{\mathcal{W}}_{\mathbb{B}} + 4\boldsymbol{\zeta}_{\mathbb{B}}\right),$$
$$\widehat{\boldsymbol{\vartheta}}_{\overline{\mathbb{B}},T}^{AL} \underset{T\to\infty}{\overset{\mathbb{P}}{\longrightarrow}} \mathbf{0}, \quad \mathbb{P}\left[\mathbb{B}_T = \mathbb{B}\right] \underset{T\to\infty}{\longrightarrow} 1.$$

The proof of this Theorem is the same as that of Theorem 3.3.

II Proofs and technical lemmas

The appendix provides the proofs of the main results. In the following, K denotes a generic positive constant whose value can vary from line to line.

We introduce several Lemmas as a toolbox for proving all the theorems. Their notations are independent of the rest of the paper.

Lemma II.1. *Let* $\{\boldsymbol{O}_u, u \leq t\}$ *be a strictly stationary and ergodic process and* $\boldsymbol{x} \mapsto F_t(\boldsymbol{x}) = F(\boldsymbol{O}_t, \boldsymbol{O}_{t-1}, \dots; \boldsymbol{x})$ *a measurable function of the past observations, defined on an open and convex set* $\Omega \subset \mathbb{R}^n$*. Let* $\boldsymbol{x}_0 \in \Omega$ *and assume that* $\sup_{\boldsymbol{x}\in\mathcal{V}(\boldsymbol{x}_0)} ||F_t(\boldsymbol{x})||$ *belongs to* L^1 *for some neighborhood* $\mathcal{V}(\boldsymbol{x}_0) \subset \Omega$ *of* $\boldsymbol{x}_0$*, then for any sequence* $(\boldsymbol{x}_T^*)$ *such that* $\boldsymbol{x}_T^* \underset{T\to\infty}{\overset{a.s.}{\longrightarrow}} \boldsymbol{x}_0$*:*

$$\frac{1}{T}\sum_{t=1}^{T} F_t(\boldsymbol{x}_T^*) \underset{T\to\infty}{\overset{a.s.}{\longrightarrow}} \mathbb{E}\left[F_1(\boldsymbol{x}_0)\right].$$

Proof. For any $\boldsymbol{x} \in \Omega$ the process $\{F_t(\boldsymbol{x}), t \in \mathbb{Z}\}$ is strictly stationary and ergodic and at $\boldsymbol{x}_0$, $F_t(\boldsymbol{x}_0)$ belongs to L^1. Therefore, by the ergodic Theorem, $\frac{1}{T}\sum_{t=1}^T F_t(\boldsymbol{x}_0) \xrightarrow[T\to\infty]{a.s.} \mathbb{E}[F_1(\boldsymbol{x}_0)]$. Let $\mathcal{B}_{1/z}(\boldsymbol{x}_0) := \left\{\boldsymbol{x} \in \mathbb{R}^n ; ||\boldsymbol{x} - \boldsymbol{x}_0|| \leq \frac{1}{z}\right\}$ where $z \in \mathbb{N}^*$. Assume that z and T are large enough such that $\boldsymbol{x}_T^* \in \mathcal{B}_{1/z}(\boldsymbol{x}_0) \subset \mathcal{V}(\boldsymbol{x}_0)$ almost surely:

$$\left|\left|\frac{1}{T}\sum_{t=1}^T F_t(\boldsymbol{x}_T^*) - \mathbb{E}[F_1(\boldsymbol{x}_0)]\right|\right| \leq \frac{1}{T}\sum_{t=1}^T Z_{z,t} + \left|\left|\frac{1}{T}\sum_{t=1}^T F_t(\boldsymbol{x}_0) - \mathbb{E}[F_1(\boldsymbol{x}_0)]\right|\right|,$$

where $Z_{z,t} = \sup\limits_{\mathcal{B}_{1/z}(\boldsymbol{x}_0)} ||F_t(\boldsymbol{x}) - F_t(\boldsymbol{x}_0)||$. On the one hand $\left|\left|\frac{1}{T}\sum_{t=1}^T F_t(\boldsymbol{x}_0) - \mathbb{E}[F_1(\boldsymbol{x}_0)]\right|\right| \xrightarrow[T\to\infty]{a.s.} 0$. On the other hand $Z_{z,t} \leq \sup\limits_{\mathcal{V}(\boldsymbol{x}_0)} ||F_t(\boldsymbol{x}) - F_t(\boldsymbol{x}_0)|| \leq 2\sup\limits_{\mathcal{V}(\boldsymbol{x}_0)} ||F_t(\boldsymbol{x})||$. The right hand side of the inequality belongs to L^1, then by the ergodic theorem $\frac{1}{T}\sum_{t=1}^T Z_{z,t} \xrightarrow[T\to\infty]{a.s.} \mathbb{E}[Z_{z,1}]$. By the dominated convergence Theorem $\mathbb{E}[Z_{z,1}] \xrightarrow[z\to\infty]{} 0$. This concludes the proof. □

The lemma below extracts the almost sure convergence part of Lemma 2.2 in Davis et al. (1992) for convex processes. In the original proof this step is obtained via a Skorokhod representation on the way to convergence in distribution; we present it separately here to keep the paper self-contained.

Lemma II.2. *Let $\{F_T(\cdot)\}$ and $F(\cdot)$ be stochastic processes continuous and strictly convex on an open convex set $A \subset \mathbb{R}^p$ and suppose that for each $\boldsymbol{x} \in A$, $F_T(\boldsymbol{x}) \xrightarrow[T\to\infty]{a.s.} F(\boldsymbol{x})$. Let $\boldsymbol{y}_T$ minimize $F_T(\cdot)$ and $\boldsymbol{y}$ minimize $F(\cdot)$ such that $\boldsymbol{y}_T, \boldsymbol{y} \in A$. Then*

$$\boldsymbol{y}_T \xrightarrow[T\to\infty]{a.s.} \boldsymbol{y}.$$

Proof. The strict convexity ensures the uniqueness of the argmins in the following. Using Rockafellar (1970) Theorem 10.8, for any given compact set $K \subset A$:

$$\sup_{\boldsymbol{u}\in K} |F_T(\boldsymbol{u}) - F(\boldsymbol{u})| \xrightarrow[T\to\infty]{\text{a.s.}} 0.$$

For $\gamma > 0$, let $B_\gamma = \{\boldsymbol{u} : ||\boldsymbol{u} - \boldsymbol{y}|| = \gamma\}$ and suppose that $||\boldsymbol{y}_T - \boldsymbol{y}|| > \gamma$ for infinitely many T. Since $F_T \to F$ uniformly on B_γ and $F_T(\boldsymbol{y}) \to F(\boldsymbol{y})$, it follows that for infinitely many T and all $\boldsymbol{u} \in B_\gamma$

$$F_T(\boldsymbol{u}) > F_T(\boldsymbol{y}) > F_T(\boldsymbol{y}_T).$$

But this contradicts the convexity of F_T by choosing $\boldsymbol{u} \in B_\gamma$ such that the points $\boldsymbol{u}, \boldsymbol{y}, \boldsymbol{y}_T$ are collinear. □

Proof of Theorem 3.1. We will establish the following intermediate results.

(a) $\sup\limits_{\boldsymbol{\alpha}\in\Phi\times\Phi} \frac{1}{T}\left|L_T(\boldsymbol{\alpha}) - \widetilde{L}_T(\boldsymbol{\alpha})\right| \xrightarrow[T\to\infty]{a.s.} 0.$

(b) $\forall \boldsymbol{\alpha} \in \mathcal{V}(\boldsymbol{\alpha}_0) \times \mathring{\Phi}, Q_\infty(\boldsymbol{\alpha}) = \mathbb{E}[l_1(\boldsymbol{\alpha})] + \sum_{j=1}^{\nu} \lambda_{j,\infty}|\varphi_j|$ exists and is finite.

(c) $\widehat{\boldsymbol{\varphi}}_T \xrightarrow[T\to\infty]{a.s.} \arg\min\limits_{\boldsymbol{\varphi}\in\Phi} Q_\infty(\boldsymbol{\phi}_0, \boldsymbol{\varphi})$.

(d) $\underset{\boldsymbol{\varphi}\in\Phi}{\arg\min}\, Q_\infty(\boldsymbol{\phi}_0,\boldsymbol{\varphi}) = \boldsymbol{\phi}_0$ if $\boldsymbol{\lambda}_\infty = \mathbf{0}$.

(a) Asymptotic irrelevance of the initial values. Let $\boldsymbol{\alpha}\in\Phi\times\Phi$:

$$\begin{aligned}\left|l_t(\boldsymbol{\alpha}) - \widetilde{l}_t(\boldsymbol{\alpha})\right| &= \left|\frac{\widetilde{w}_t^2\left(y_t - f_t(\boldsymbol{\alpha})\right)^2 - w_t^2\left(y_t - \widetilde{f}_t(\boldsymbol{\alpha})\right)^2}{w_t^2\widetilde{w}_t^2}\right| \\ &= \left|\frac{\left(w_t^2 - \widetilde{w}_t^2\right)\left(y_t - f_t(\boldsymbol{\alpha})\right)^2 - w_t^2\left(f_t(\boldsymbol{\alpha}) - \widetilde{f}_t(\boldsymbol{\alpha})\right)\left(2y_t - f_t(\boldsymbol{\alpha}) - \widetilde{f}_t(\boldsymbol{\alpha})\right)}{w_t^2\widetilde{w}_t^2}\right| \\ &\le \frac{d_t\left|y_t - f_t(\boldsymbol{\alpha})\right|^2 + w_t^2 a_t\left|2y_t - f_t(\boldsymbol{\alpha}) - \widetilde{f}_t(\boldsymbol{\alpha})\right|}{w_t^2\widetilde{w}_t^2}.\end{aligned}$$

Under Assumption A2, for t large enough we have almost surely $\sup_{\boldsymbol{\alpha}\in\Phi\times\Phi}\left|\widetilde{f}_t(\boldsymbol{\alpha})\right| \le 1 + \sup_{\boldsymbol{\alpha}\in\Phi\times\Phi}\left|f_t(\boldsymbol{\alpha})\right|$. For $u,v\in\mathbb{R}$ we have $(u+v)^2 \le 2\left(u^2+v^2\right)$. Using these two results gives:

$$\sup_{\boldsymbol{\alpha}\in\Phi\times\Phi}\left|l_t(\boldsymbol{\alpha}) - \widetilde{l}_t(\boldsymbol{\alpha})\right| \le \frac{2d_t\left(1 + y_t^2 + \sup_{\boldsymbol{\alpha}\in\Phi\times\Phi}\left|f_t^2(\boldsymbol{\alpha})\right|\right)}{\underline{w}^4} + \frac{2a_t\left(1 + |y_t| + \sup_{\boldsymbol{\alpha}\in\Phi\times\Phi}\left|f_t(\boldsymbol{\alpha})\right|\right)}{\underline{w}^2}$$

By Assumption A3, the right hand side of the inequality goes to 0 almost surely as t goes to infinity, therefore $\sup_{\boldsymbol{\alpha}\in\Phi\times\Phi}\left|l_t(\boldsymbol{\alpha}) - \widetilde{l}_t(\boldsymbol{\alpha})\right| \xrightarrow[t\to\infty]{a.s.} 0$. By Cesàro's Lemma, (a) is established.

(b) Existence of the limit loss function. We start by noting that:

$$l_t(\boldsymbol{\alpha}) = \left(\frac{y_t - f_t(\boldsymbol{\alpha})}{w_t}\right)^2 = \left(\frac{\sigma_t\eta_t}{w_t} + \frac{\mu_t - f_t(\boldsymbol{\alpha})}{w_t}\right)^2.$$

Under Assumptions A5-A6, using the mean value Theorem and the compactness of Φ:

$$\sup_{\boldsymbol{\alpha}\in\mathcal{V}(\boldsymbol{\phi}_0)\times\mathring{\Phi}}\left|\frac{\mu_t - f_t(\boldsymbol{\alpha})}{w_t}\right| \le \sup_{\boldsymbol{\alpha}\in\mathcal{V}(\boldsymbol{\phi}_0)\times\mathring{\Phi}}\frac{||\boldsymbol{\alpha} - \boldsymbol{\alpha}_0||}{w_t}\sup_{\boldsymbol{\alpha}\in\mathcal{V}(\boldsymbol{\phi}_0)\times\mathring{\Phi}}\left|\left|\frac{\partial f_t}{\partial\boldsymbol{\alpha}}(\boldsymbol{\alpha})\right|\right| \le \frac{K}{w_t}\sup_{\boldsymbol{\alpha}\in\mathcal{V}(\boldsymbol{\phi}_0)\times\mathring{\Phi}}\left|\left|\frac{\partial f_t}{\partial\boldsymbol{\alpha}}(\boldsymbol{\alpha})\right|\right|. \tag{29}$$

The last term belongs to L^2. Under Assumption A4 and using the independence of η_t and $\mathcal{F}_{t-1}$ gives:

$$\mathbb{E}\left[\sup_{\boldsymbol{\alpha}\in\mathcal{V}(\boldsymbol{\phi}_0)\times\mathring{\Phi}} l_1(\boldsymbol{\alpha})\right] < \infty. \tag{30}$$

Moreover, $\boldsymbol{\lambda}_T \xrightarrow[T\to\infty]{} \boldsymbol{\lambda}_\infty < \infty$. We obtain the existence and the finiteness of the limit loss $Q_\infty(\boldsymbol{\alpha})$.

(c) Convergence of the minimizers. Under Assumption A1 and using (30), the conditions of Lemma II.1 are satisfied, and we have for $\boldsymbol{\varphi}\in\mathring{\Phi}$, $Q_T\left(\widehat{\boldsymbol{\phi}}_T,\boldsymbol{\varphi}\right) \xrightarrow[T\to\infty]{a.s.} Q_\infty(\boldsymbol{\phi}_0,\boldsymbol{\varphi})$. Using the result of part (a) we obtain $\widetilde{Q}_T\left(\widehat{\boldsymbol{\phi}}_T,\boldsymbol{\varphi}\right) \xrightarrow[T\to\infty]{a.s.} Q_\infty(\boldsymbol{\phi}_0,\boldsymbol{\varphi})$. Under Assumption A9, the functions $\widetilde{Q}_T\left(\widehat{\boldsymbol{\phi}}_T,\cdot\right)$ and $Q_\infty(\boldsymbol{\phi}_0,\cdot)$ are strictly convex on Φ, thus having unique minima. Under Assumptions A5 and A7, these minima lie in $\mathring{\Phi}$. Lemma II.2 gives $\underset{\boldsymbol{\varphi}\in\mathring{\Phi}}{\arg\min}\,\widetilde{Q}_T\left(\widehat{\boldsymbol{\phi}}_T,\boldsymbol{\varphi}\right) \xrightarrow[T\to\infty]{a.s.} \underset{\boldsymbol{\varphi}\in\mathring{\Phi}}{\arg\min}\, Q_\infty(\boldsymbol{\phi}_0,\boldsymbol{\varphi})$. Since the strict convexity ensures the uniqueness of the minima, the argmins can be taken over the whole compact set. The result (c) is established.

(d) Asymptotic unbiasedness under vanishing penalty. Let $\boldsymbol{\varphi} \in \mathcal{V}(\boldsymbol{\phi}_0)$, if $\boldsymbol{\lambda}_\infty = \mathbf{0}$ then $Q_\infty(\boldsymbol{\phi}_0, \boldsymbol{\varphi}) = \mathbb{E}[l_1(\boldsymbol{\phi}_0, \boldsymbol{\varphi})] = \mathbb{E}\left[\left(\frac{\mu_t + \sigma_t\eta_t - f_t(\boldsymbol{\phi}_0, \boldsymbol{\varphi})}{w_t}\right)^2\right]$. Taking the derivative with respect to $\boldsymbol{\varphi}$ at $\boldsymbol{\phi}_0$ gives $\frac{\partial Q_\infty}{\partial \boldsymbol{\varphi}}(\boldsymbol{\alpha}_0) = -2\mathbb{E}\left[\frac{y_1 - f_1(\boldsymbol{\alpha}_0)}{w_1^2}\frac{\partial f_1}{\partial \boldsymbol{\varphi}}(\boldsymbol{\alpha}_0)\right] = -2\mathbb{E}[\eta_1]\,\mathbb{E}\left[\frac{\sigma_1}{w_1^2}\frac{\partial f_1}{\partial \boldsymbol{\varphi}}(\boldsymbol{\alpha}_0)\right] = \mathbf{0}$. The conclusion follows. □

Proof of Theorem 3.2. We split the proof into the following parts:

(a) $(\Upsilon_T)_{T\geq 1}$ is an exhaustion of $\mathbb{R}^\nu$.

(b) $\forall k \in \mathbb{N}^*,\ \sup_{\boldsymbol{v}\in\Upsilon_k}\left|G_T(\boldsymbol{v}) - \widetilde{G}_T(\boldsymbol{v})\right| \xrightarrow[T\to\infty]{\mathbb{P}} 0$.

(c) $\mathbb{E}\left[\sup_{\boldsymbol{\alpha}\in\mathcal{V}(\boldsymbol{\alpha}_0)}\left|\left|\frac{\partial l_1}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha})\right|\right|^2\right] < \infty$ and $\mathbb{E}\left[\sup_{\boldsymbol{\alpha}\in\mathcal{V}(\boldsymbol{\alpha}_0)}\left|\left|\frac{\partial^2 l_1}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}'}(\boldsymbol{\alpha})\right|\right|\right] < \infty$.

(d) $\begin{pmatrix}\frac{1}{\sqrt{T}}\frac{\partial L_T}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}_0)\\ \sqrt{T}\left(\widehat{\boldsymbol{\phi}}_T - \boldsymbol{\phi}_0\right)\end{pmatrix} \xrightarrow[T\to\infty]{d} \begin{pmatrix}\boldsymbol{W}\\ \boldsymbol{Z}^C\end{pmatrix}$.

(e) $\sqrt{T}\left(\widehat{\boldsymbol{\varphi}}_T - \boldsymbol{\phi}_0\right) \xrightarrow[T\to\infty]{d} \arg\min_{\boldsymbol{v}\in\mathbb{R}^\nu} E_\infty(\boldsymbol{v})$.

(a) Exhaustion of $\mathbb{R}^\nu$ by an increasing sequence of compacta. Let $p, q \in \mathbb{N}^*, p < q$ and $\boldsymbol{x} \in \Upsilon_p$ then $\exists \boldsymbol{y} \in \Phi$ such that:

$$\boldsymbol{x} = \sqrt{p}(\boldsymbol{y} - \boldsymbol{\phi}_0) = \sqrt{q}\left(\sqrt{\frac{p}{q}}\boldsymbol{y} + \left(1 - \sqrt{\frac{p}{q}}\right)\boldsymbol{\phi}_0 - \boldsymbol{\phi}_0\right).$$

The set Φ is convex so $\sqrt{\frac{p}{q}}\boldsymbol{y} + \left(1 - \sqrt{\frac{p}{q}}\right)\boldsymbol{\phi}_0 \in \Phi$ and under Assumption A7, $\boldsymbol{\phi}_0 \in \overset{\circ}{\Phi}$ therefore $\sqrt{\frac{p}{q}}\boldsymbol{y} + \left(1 - \sqrt{\frac{p}{q}}\right)\boldsymbol{\phi}_0 \in \overset{\circ}{\Phi}$ and $\boldsymbol{x} \in \overset{\circ}{\Upsilon}_q$. Moreover, since $\mathbf{0} \in \overset{\circ}{\Upsilon}_q, \exists\delta > 0, \mathcal{B}_\delta(\mathbf{0}) \subset \overset{\circ}{\Upsilon}_q$. Let $\boldsymbol{z} \in \mathbb{R}^\nu, \boldsymbol{z} \neq \mathbf{0}$ then $\frac{\delta}{2||\boldsymbol{z}||}\boldsymbol{z} \in \mathcal{B}_\delta(\mathbf{0}) \subset \overset{\circ}{\Upsilon}_q$. By taking k such that $\sqrt{k} > \frac{\delta}{2||\boldsymbol{z}||}$ we have $\boldsymbol{z} \in \Upsilon_k$. The conclusion follows.

(b) Asymptotic decrease of the effect of the initial values. Let $k \in \mathbb{N}^*$. Under Assumptions A5, A8, A11 and for T large enough such that almost surely $\begin{pmatrix}\widehat{\boldsymbol{\phi}}_T\\ \frac{\boldsymbol{v}}{\sqrt{T}} + \boldsymbol{\phi}_0\end{pmatrix} \in \mathcal{V}(\boldsymbol{\alpha}_0)$ we have the following:

$$\sup_{\boldsymbol{v}\in\Upsilon_k}\left|G_T(\boldsymbol{v}) - \widetilde{G}_T(\boldsymbol{v})\right| \leq K\left(\left|\left|\sqrt{T}\left(\widehat{\boldsymbol{\phi}}_T - \boldsymbol{\phi}_0\right)\right|\right| + 1\right)\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\sup_{\boldsymbol{\alpha}\in\mathcal{V}(\boldsymbol{\alpha}_0)}\left|\left|\frac{\partial l_t}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}) - \frac{\partial \widetilde{l}_t}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha})\right|\right|.$$

Under Assumption A13 and for t large enough:

$$\begin{aligned}
&\sup_{\boldsymbol{\alpha}\in\mathcal{V}(\boldsymbol{\alpha}_0)}\left|\left|\frac{\partial l_t}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}) - \frac{\partial \widetilde{l}_t}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha})\right|\right|\\
&\leq K\sup_{\boldsymbol{\alpha}\in\mathcal{V}(\boldsymbol{\alpha}_0)}\left|\left|\frac{\left(\widetilde{w}_t^2 - w_t^2\right)\left(y_t - f_t(\boldsymbol{\alpha})\right)}{w_t^2\widetilde{w}_t^2}\frac{\partial f_t}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha})\right|\right| + K\sup_{\boldsymbol{\alpha}\in\mathcal{V}(\boldsymbol{\alpha}_0)}\left|\left|\frac{y_t - f_t(\boldsymbol{\alpha})}{\widetilde{w}_t^2}\frac{\partial f_t}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}) - \frac{y_t - \widetilde{f}_t(\boldsymbol{\alpha})}{\widetilde{w}_t^2}\frac{\partial \widetilde{f}_t}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha})\right|\right|\\
&\leq Kd_t\left(1 + |y_t| + \sup_{\boldsymbol{\alpha}\in\Phi\times\Phi}|f_t(\boldsymbol{\alpha})| + a_t\right)\sup_{\boldsymbol{\alpha}\in\Phi\times\Phi}\left|\left|\frac{\partial f_t}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha})\right|\right| + Kb_t\left(1 + |y_t| + \sup_{\boldsymbol{\alpha}\in\Phi\times\Phi}|f_t(\boldsymbol{\alpha})|\right).
\end{aligned}$$

The last term is almost surely of order $O\left(t^{-\kappa}\right)$ with $\kappa > \frac{1}{2}$. Under Assumption A12, $\left|\left|\sqrt{T}\left(\widehat{\boldsymbol{\phi}}_T - \boldsymbol{\phi}_0\right)\right|\right| = O_{\mathbb{P}}\left(1\right)$. These two results give (b).

(c) Integrability of the suprema of the loss function's first and second derivatives. Under Assumption A11 and for $\boldsymbol{\alpha} \in \mathcal{V}\left(\boldsymbol{\alpha}_0\right)$, we have $\frac{\partial l_t}{\partial \boldsymbol{\alpha}}\left(\boldsymbol{\alpha}\right) = -\frac{2}{w_t^2}\left(y_t - f_t\left(\boldsymbol{\alpha}\right)\right)\frac{\partial f_t}{\partial \boldsymbol{\alpha}}\left(\boldsymbol{\alpha}\right)$. Under Assumptions A10 and A11 and with the same arguments as in the proof of Theorem 3.1 part (b) with Equation (29) we obtain:

$$\mathbb{E}\left[\sup_{\boldsymbol{\alpha}\in\mathcal{V}\left(\boldsymbol{\alpha}_0\right)} l_1^2\left(\boldsymbol{\alpha}\right)\right] < \infty. \tag{31}$$

Cauchy-Schwartz's inequality along with the moments given by Assumption A11 and Equation (31) give:

$$\mathbb{E}\left[\sup_{\boldsymbol{\alpha}\in\mathcal{V}\left(\boldsymbol{\alpha}_0\right)} \left|\left|\frac{\partial l_1}{\partial \boldsymbol{\alpha}}\left(\boldsymbol{\alpha}\right)\right|\right|^2\right] < \infty. \tag{32}$$

The second derivative is $\frac{\partial^2 l_1}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}'}\left(\boldsymbol{\alpha}\right) = \frac{2}{w_t^2}\left[P\left(\frac{\partial f_1}{\partial \boldsymbol{\alpha}}\left(\boldsymbol{\alpha}\right)\right) - \left(y_1 - f_1\left(\boldsymbol{\alpha}\right)\right)\frac{\partial^2 g_1}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}'}\left(\boldsymbol{\alpha}\right)\right]$. With Equations (32) and (31) we obtain:

$$\mathbb{E}\left[\sup_{\boldsymbol{\alpha}\in\mathcal{V}\left(\boldsymbol{\alpha}_0\right)} \left|\left|\frac{\partial^2 l_1}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}'}\left(\boldsymbol{\alpha}\right)\right|\right|\right] < \infty. \tag{33}$$

(d) C.L.T. for martingale increments. Under Assumption A11, $\frac{\partial l_t}{\partial \boldsymbol{\alpha}}\left(\boldsymbol{\alpha}_0\right)$ exists and using the Bahadur-type expansion given by Assumption A12, we define:

$$\boldsymbol{U}_t = \begin{pmatrix} \frac{\partial l_t}{\partial \boldsymbol{\alpha}}\left(\boldsymbol{\alpha}_0\right) \\ \boldsymbol{\Delta}_t \gamma\left(\eta_t\right) \end{pmatrix}.$$

On the one hand, $\boldsymbol{\Delta}_t$ and $\gamma\left(\eta_t\right)$ are independent and belong to L^2. On the other hand, using the Equation (32) of part (c), $\boldsymbol{U}_t$ belongs to L^2 and we have $\mathbb{E}\left[\boldsymbol{U}_t|\mathcal{F}_{t-1}\right] = \begin{pmatrix} -\mathbb{E}\left[\eta_t\right]\frac{2\sigma_t}{w_t^2}\frac{\partial f_t}{\partial \boldsymbol{\alpha}}\left(\boldsymbol{\alpha}_0\right) \\ \boldsymbol{\Delta}_t \mathbb{E}\left[\gamma\left(\eta_t\right)\right] \end{pmatrix} = \boldsymbol{0}$. Under Assumption A1, the process is a strictly stationary and ergodic L^2 martingale increments. We can apply the martingale C.L.T. of Billingsley (1961) yielding $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{U}_t \xrightarrow[T\to\infty]{d} \begin{pmatrix} \boldsymbol{W} \\ \boldsymbol{Z} \end{pmatrix}$. We also have $\boldsymbol{J}_T \xrightarrow[T\to\infty]{a.s.} \boldsymbol{J}$. Slutsky's Lemma gives the joint convergence in distribution:

$$\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{U}_t, \boldsymbol{J}_T\right) \xrightarrow[T\to\infty]{d} \left(\begin{pmatrix} \boldsymbol{W} \\ \boldsymbol{Z} \end{pmatrix}, \boldsymbol{J}\right)$$

Given that the set C is closed and convex, the projection $\mathrm{Proj}_C\left(\boldsymbol{x}, \boldsymbol{M}\right) := \arg\min_{\boldsymbol{z}\in C}\left(\boldsymbol{z} - \boldsymbol{x}\right)'\boldsymbol{M}\left(\boldsymbol{z} - \boldsymbol{x}\right)$ is a continuous mapping where $\boldsymbol{x} \in \mathbb{R}^{\nu}$ and $\boldsymbol{M}$ is a symmetric positive definite matrix (see Rockafellar and Wets (2009) Theorem 1.17 for a general result or Francq and Zakoïan (2019) Section 8.2 in GARCH context). Therefore, the continuous mapping Theorem gives:

$$\begin{pmatrix} \frac{1}{\sqrt{T}}\frac{\partial L_T}{\partial \boldsymbol{\alpha}}\left(\boldsymbol{\alpha}_0\right) \\ \sqrt{T}\left(\widehat{\boldsymbol{\phi}}_T - \boldsymbol{\phi}_0\right) \end{pmatrix} \xrightarrow[T\to\infty]{d} \begin{pmatrix} \boldsymbol{W} \\ \boldsymbol{Z}^C \end{pmatrix}.$$

The components of the covariance matrix of the vector $(\boldsymbol{W}', \boldsymbol{Z}')'$ are given by:

$$\mathbb{V}\left[\frac{\partial l_1}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}_0)\right] = \mathbb{E}\left[P\left(\eta_1 \frac{2\sigma_1}{w_1^2}\frac{\partial f_1}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}_0)\right)\right] = 4\mathbb{E}\left[P\left(\frac{\sigma_1}{w_1^2}\frac{\partial f_1}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}_0)\right)\right] = I(\boldsymbol{\phi}_0),$$

$$\mathbb{E}\left[\frac{\partial l_1}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}_0)\gamma'(\eta_t)\boldsymbol{\Delta}_t'\right] = \mathbb{E}\left[\frac{-2\eta_1\sigma_1}{w_1^2}\frac{\partial f_1}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}_0)\gamma'(\eta_1)\boldsymbol{\Delta}_1'\right] = R(\boldsymbol{\phi}_0),$$

$$\mathbb{V}[\boldsymbol{\Delta}_1\gamma(\eta_1)] = \mathbb{E}\left[\boldsymbol{\Delta}_1\gamma(\eta_1)\gamma'(\eta_1)\boldsymbol{\Delta}_1'\right] = \mathbb{E}\left[\boldsymbol{\Delta}_1\boldsymbol{\Gamma}\boldsymbol{\Delta}_1'\right] = \mathbb{V}\left[\boldsymbol{\Delta}_1\boldsymbol{\Gamma}^{1/2}\right] = \boldsymbol{\Sigma}.$$

(e) **Asymptotic distribution of $\widehat{\varphi}_T$.** Let $\boldsymbol{v} \in \mathbb{R}^\nu$. Under Assumptions A8, A11 and with T large enough such that $\begin{pmatrix} \widehat{\boldsymbol{\phi}}_T \\ \frac{\boldsymbol{v}}{\sqrt{T}} + \boldsymbol{\phi}_0 \end{pmatrix} \in \mathcal{V}(\boldsymbol{\alpha}_0)$ we have:

$$G_T(\boldsymbol{v}) = \begin{pmatrix} \sqrt{T}\left(\widehat{\boldsymbol{\phi}}_T - \boldsymbol{\phi}_0\right) \\ \boldsymbol{v} \end{pmatrix}' \frac{1}{\sqrt{T}}\frac{\partial L_T}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}_0) + \begin{pmatrix} \sqrt{T}\left(\widehat{\boldsymbol{\phi}}_T - \boldsymbol{\phi}_0\right) \\ \boldsymbol{v} \end{pmatrix}' \frac{1}{2T}\frac{\partial^2 L_T}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}'}(\boldsymbol{\alpha}_T)\begin{pmatrix} \sqrt{T}\left(\widehat{\boldsymbol{\phi}}_T - \boldsymbol{\phi}_0\right) \\ \boldsymbol{v} \end{pmatrix},$$

where $\boldsymbol{\alpha}_T$ lies between $\boldsymbol{\alpha}_0$ and $\begin{pmatrix} \widehat{\boldsymbol{\phi}}_T \\ \frac{\boldsymbol{v}}{\sqrt{T}} + \boldsymbol{\phi}_0 \end{pmatrix}$. We can easily see that:

$$\begin{pmatrix} \sqrt{T}\left(\widehat{\boldsymbol{\phi}}_T - \boldsymbol{\phi}_0\right) \\ \boldsymbol{v} \end{pmatrix} = \boldsymbol{\Pi}_{2,2\nu}\begin{pmatrix} \frac{1}{\sqrt{T}}\frac{\partial L_T}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}_0) \\ \sqrt{T}\left(\widehat{\boldsymbol{\phi}}_T - \boldsymbol{\phi}_0\right) \\ \boldsymbol{v} \end{pmatrix} \text{ and } \frac{1}{\sqrt{T}}\frac{\partial L_T}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}_0) = \boldsymbol{\Pi}_{1,2\nu}\begin{pmatrix} \frac{1}{\sqrt{T}}\frac{\partial L_T}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}_0) \\ \sqrt{T}\left(\widehat{\boldsymbol{\phi}}_T - \boldsymbol{\phi}_0\right) \\ \boldsymbol{v} \end{pmatrix}.$$

Rewriting the previous expansion gives:

$$\begin{aligned} G_T(\boldsymbol{v}) &= \begin{pmatrix} \sqrt{T}\left(\widehat{\boldsymbol{\phi}}_T - \boldsymbol{\phi}_0\right) \\ \boldsymbol{v} \end{pmatrix}'\left(\frac{1}{\sqrt{T}}\frac{\partial L_T}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}_0) + \frac{1}{2T}\frac{\partial^2 L_T}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}'}(\boldsymbol{\alpha}_T)\begin{pmatrix} \sqrt{T}\left(\widehat{\boldsymbol{\phi}}_T - \boldsymbol{\phi}_0\right) \\ \boldsymbol{v} \end{pmatrix}\right) \\ &= \begin{pmatrix} \frac{1}{\sqrt{T}}\frac{\partial L_T}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}_0) \\ \sqrt{T}\left(\widehat{\boldsymbol{\phi}}_T - \boldsymbol{\phi}_0\right) \\ \boldsymbol{v} \end{pmatrix}' \boldsymbol{\Pi}_{2,2\nu}'\left(\boldsymbol{\Pi}_{1,2\nu} + \frac{1}{2T}\frac{\partial^2 L_T}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}'}(\boldsymbol{\alpha}_T)\boldsymbol{\Pi}_{2,2\nu}\right)\begin{pmatrix} \frac{1}{\sqrt{T}}\frac{\partial L_T}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}_0) \\ \sqrt{T}\left(\widehat{\boldsymbol{\phi}}_T - \boldsymbol{\phi}_0\right) \\ \boldsymbol{v} \end{pmatrix}. \end{aligned}$$

Using Equation (33) and given that $\boldsymbol{\alpha}_T \xrightarrow[T\to\infty]{a.s.} \boldsymbol{\alpha}_0$, Lemma II.1 applies and we obtain:

$$\frac{1}{T}\frac{\partial^2 L_T}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}'}(\boldsymbol{\alpha}_T) \xrightarrow[T\to\infty]{a.s.} \mathbb{E}\left[\frac{\partial^2 l_1}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}'}(\boldsymbol{\alpha}_0)\right] = 4\mathbb{E}\left[P\left(w_1^{-1}\frac{\partial f_1}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}_0)\right)\right].$$

Slutsky's Lemma and the result of part (d) give $E_T(\boldsymbol{v}) \xrightarrow[T\to\infty]{d} E_\infty(\boldsymbol{v})$. It follows that, using (b), $\widetilde{E}_T(\boldsymbol{v}) \xrightarrow[T\to\infty]{d} E_\infty(\boldsymbol{v})$. The finite dimensional convergence holds trivially using the Cramér-Wold device. In the following, we use the concept of epi-convergence in distribution and the results of Knight (1999). Note that by the result of part (a) the sequence $(\Upsilon_T)_{T\geq 1}$ spans $\mathbb{R}^\nu$, therefore the limit function E_∞ in defined on $\mathbb{R}^\nu$. Under Assumption A9, the functions $\widetilde{E}_T$ and E_∞ are strictly convex on their domains, thus having unique argmins. Their extension on $\mathbb{R}^\nu$ is convex and lower-semicontinuous (l.s.c.) since $\mathbb{K}_T$ is l.s.c.. Using Theorem 5 (a) of Knight (1999) gives:

$$\widetilde{E}_T \xrightarrow[T\to\infty]{d} E_\infty,$$

in the sens of epi-convergence in distribution. The same Theorem 5 (b) gives also the convergence of the argmins and the conclusion follows. □

Proof of Proposition 3.1. We have $\mathbb{P}\left[\mathbb{A}_T = \mathbb{A}\right] \leq \mathbb{P}\left[\forall j \in \overline{\mathbb{A}} : \widehat{\varphi}_{j,T} = 0\right]$. By Theorem 3.2 we have $\sqrt{T}\left(\widehat{\boldsymbol{\varphi}}_T - \boldsymbol{\phi}_0\right) \xrightarrow[T\to\infty]{d} \boldsymbol{v}^* := \underset{\boldsymbol{v}\in\mathbb{R}^\nu}{\arg\min}\, E_\infty\left(\boldsymbol{v}\right)$. Using the portmanteau Theorem:

$$\limsup_{T\to\infty} \mathbb{P}\left[\forall j \in \overline{\mathbb{A}} : \sqrt{T}\widehat{\varphi}_{j,T} = 0\right] \leq \mathbb{P}\left[\forall j \in \overline{\mathbb{A}} : \boldsymbol{v}_j^* = 0\right].$$

Since $\boldsymbol{v}^*$ is the unique minimizer of the convex function E_∞ then $\mathbf{0} \in \frac{\partial E_\infty}{\partial \boldsymbol{v}}\left(\boldsymbol{v}^*\right)$ where $\frac{\partial E_\infty}{\partial \boldsymbol{v}}$ is the sub-gradient of E_∞. The previous equation implies that:

$$\begin{cases} \frac{\partial G_\infty}{\partial v_j}\left(\boldsymbol{v}^*\right) + \lambda_{j,\infty}\,\mathrm{sign}\left(\phi_{j,0}\right) = 0 \text{ if } j \in \mathbb{A}, \\ \left|\frac{\partial G_\infty}{\partial v_j}\left(\boldsymbol{v}^*\right)\right| \leq \lambda_{j,\infty} \text{ if } j \in \overline{\mathbb{A}}. \end{cases} \tag{34}$$

For $\boldsymbol{v} \in \mathbb{R}^\nu$, rewriting the function G_∞ gives:

$$\begin{aligned} G_\infty\left(\boldsymbol{v}\right) &= \begin{pmatrix} \boldsymbol{W} \\ \boldsymbol{Z}^C \\ \boldsymbol{v} \end{pmatrix}' \boldsymbol{\Pi}'_{2,2\nu}\left(\boldsymbol{\Pi}_{1,2\nu} + 2\mathbb{E}\left[P\left(\frac{1}{w_1}\frac{\partial f_1}{\partial \boldsymbol{\alpha}}\left(\boldsymbol{\alpha}_0\right)\right)\right]\boldsymbol{\Pi}_{2,2\nu}\right)\begin{pmatrix} \boldsymbol{W} \\ \boldsymbol{Z}^C \\ \boldsymbol{v} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{Z}^C \\ \boldsymbol{v} \end{pmatrix}' \boldsymbol{W} + 2\begin{pmatrix} \boldsymbol{Z}^C \\ \boldsymbol{v} \end{pmatrix}' \mathbb{E}\left[\begin{pmatrix} P\left(\frac{1}{w_1}\frac{\partial f_1}{\partial \boldsymbol{\phi}}\left(\boldsymbol{\alpha}_0\right)\right) & \frac{1}{w_1^2}\frac{\partial f_1}{\partial \boldsymbol{\phi}}\left(\boldsymbol{\alpha}_0\right)\frac{\partial f_1}{\partial \boldsymbol{\varphi}'}\left(\boldsymbol{\alpha}_0\right) \\ \frac{1}{w_1^2}\frac{\partial f_1}{\partial \boldsymbol{\varphi}}\left(\boldsymbol{\alpha}_0\right)\frac{\partial f_1}{\partial \boldsymbol{\phi}'}\left(\boldsymbol{\alpha}_0\right) & P\left(\frac{1}{w_1}\frac{\partial f_1}{\partial \boldsymbol{\varphi}}\left(\boldsymbol{\alpha}_0\right)\right) \end{pmatrix}\right]\begin{pmatrix} \boldsymbol{Z}^C \\ \boldsymbol{v} \end{pmatrix} \\ &= \boldsymbol{W}'\boldsymbol{Z}^C + \boldsymbol{W}'\boldsymbol{v} + 2\boldsymbol{Z}^{C'}\mathbb{E}\left[P\left(\frac{1}{w_1}\frac{\partial f_1}{\partial \boldsymbol{\phi}}\left(\boldsymbol{\alpha}_0\right)\right)\right]\boldsymbol{Z}^C + 2\boldsymbol{v}'\mathbb{E}\left[P\left(\frac{1}{w_1}\frac{\partial f_1}{\partial \boldsymbol{\varphi}}\left(\boldsymbol{\alpha}_0\right)\right)\right]\boldsymbol{v} \\ &+ 4\boldsymbol{v}'\mathbb{E}\left[\frac{1}{w_1^2}\frac{\partial f_1}{\partial \boldsymbol{\varphi}}\left(\boldsymbol{\alpha}_0\right)\frac{\partial f_1}{\partial \boldsymbol{\phi}'}\left(\boldsymbol{\alpha}_0\right)\right]\boldsymbol{Z}^C. \end{aligned}$$

The derivative with respect to $\boldsymbol{v}$ is:

$$\frac{\partial G_\infty}{\partial \boldsymbol{v}}\left(\boldsymbol{v}\right) = \boldsymbol{W} + 4\mathbb{E}\left[P\left(\frac{1}{w_1}\frac{\partial f_1}{\partial \boldsymbol{\varphi}}\left(\boldsymbol{\alpha}_0\right)\right)\right]\boldsymbol{v} + 4\mathbb{E}\left[\frac{1}{w_1^2}\frac{\partial f_1}{\partial \boldsymbol{\varphi}}\left(\boldsymbol{\alpha}_0\right)\frac{\partial f_1}{\partial \boldsymbol{\phi}'}\left(\boldsymbol{\alpha}_0\right)\right]\boldsymbol{Z}^C. \tag{35}$$

Therefore, Equation (34) becomes:

$$\begin{cases} \boldsymbol{W}_{\mathbb{A}} + 4\left(\mathbb{E}\left[P\left(\frac{1}{w_1}\frac{\partial f_1}{\partial \boldsymbol{\varphi}}\left(\boldsymbol{\alpha}_0\right)\right)\right]\boldsymbol{v}^* + \mathbb{E}\left[\frac{1}{w_1^2}\frac{\partial f_1}{\partial \boldsymbol{\varphi}}\left(\boldsymbol{\alpha}_0\right)\frac{\partial f_1}{\partial \boldsymbol{\phi}'}\left(\boldsymbol{\alpha}_0\right)\right]\boldsymbol{Z}^C\right)_{\mathbb{A}} = -\boldsymbol{\lambda}_{\mathbb{A},\infty} \circ \mathrm{sign}\,\boldsymbol{\phi}_{\mathbb{A},0}, \\ \left|\left(\frac{\partial G_\infty}{\partial \boldsymbol{v}}\left(\boldsymbol{v}^*\right)\right)_{\overline{\mathbb{A}}}\right| \leq \boldsymbol{\lambda}_{\overline{\mathbb{A}},\infty}, \end{cases}$$

where the inequality and the absolute value are component-wise. Let us consider the event $\left\{\forall j \in \overline{\mathbb{A}} : \boldsymbol{v}_j^* = 0\right\}$. We have $\boldsymbol{v}_{\overline{\mathbb{A}}}^* = \mathbf{0}$ and $\left(\mathbb{E}\left[P\left(\frac{1}{w_1}\frac{\partial f_1}{\partial \boldsymbol{\varphi}}\left(\boldsymbol{\alpha}_0\right)\right)\right]\boldsymbol{v}^*\right)_{\mathbb{A}} = \boldsymbol{M}_{11}\boldsymbol{v}_{\mathbb{A}}^*$. Therefore, under Assumption A14:

$$\begin{cases} \boldsymbol{v}_{\mathbb{A}}^* = -\frac{1}{4}\boldsymbol{M}_{11}^{-1}\left(\boldsymbol{W}_{\mathbb{A}} + \boldsymbol{\lambda}_{\mathbb{A},\infty} \circ \mathrm{sign}\,\boldsymbol{\phi}_{\mathbb{A},0} + 4\boldsymbol{V}_{\mathbb{A}}\right), \\ \left|\boldsymbol{W}_{\overline{\mathbb{A}}} - \boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1}\left(\boldsymbol{W}_{\mathbb{A}} + \boldsymbol{\lambda}_{\mathbb{A},\infty} \circ \mathrm{sign}\,\boldsymbol{\phi}_{\mathbb{A},0} + 4\boldsymbol{V}_{\mathbb{A}}\right) + 4\boldsymbol{V}_{\overline{\mathbb{A}}}\right| \leq \boldsymbol{\lambda}_{\overline{\mathbb{A}},\infty}. \end{cases}$$

Finally:

$$\begin{aligned}&\limsup_{T\to\infty}\mathbb{P}\left[\forall j\in\overline{\mathbb{A}}:\sqrt{T}\widehat{\varphi}_{j,T}=0\right]\leq\mathbb{P}\left[\forall j\in\overline{\mathbb{A}}:\boldsymbol{v}_j^*=0\right]\\&\leq\mathbb{P}\left[\left|\boldsymbol{W}_{\overline{\mathbb{A}}}-\boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1}\left(\boldsymbol{W}_{\mathbb{A}}+\boldsymbol{\lambda}_{\mathbb{A},\infty}\circ\operatorname{sign}\boldsymbol{\phi}_{\mathbb{A},0}+4\boldsymbol{V}_{\mathbb{A}}\right)+4\boldsymbol{V}_{\overline{\mathbb{A}}}\right|\leq\boldsymbol{\lambda}_{\overline{\mathbb{A}},\infty}\right]<1.\end{aligned}$$

□

Theorem 3.3. We split the proof of this Theorem into three parts:

(a) $\exists\underline{\lambda}_T>0,\underline{\lambda}_T=O_{\mathbb{P}}(1)$ and $\left(\forall j\in\overline{\mathbb{A}},\frac{\lambda_{j,T}}{|\widehat{\phi}_{j,T}|^{\tau}}\geq\underline{\lambda}_T\implies\widehat{\varphi}_{j,T}^{AL}=0\right)$.

(b) $\sqrt{T}\left(\widehat{\boldsymbol{\varphi}}_{\mathbb{A},T}^{AL}-\boldsymbol{\phi}_{\mathbb{A},0}\right)\xrightarrow[T\to\infty]{d}-\frac{1}{4}\boldsymbol{M}_{11}^{-1}\left(\boldsymbol{W}_{\mathbb{A}}+4\boldsymbol{V}_{\mathbb{A}}\right)$ and $\widehat{\boldsymbol{\varphi}}_{\overline{\mathbb{A}},T}^{AL}\xrightarrow[T\to\infty]{\mathbb{P}}\mathbf{0}$.

(c) $\mathbb{P}\left[\mathbb{A}_T=\mathbb{A}\right]\xrightarrow[T\to\infty]{}1$.

Throughout, the cases where $\widehat{\phi}_{j,T}=0$ for some j are not mentioned. They lead to obvious situations.

(a) Sparsity threshold. The estimator $\widehat{\boldsymbol{\varphi}}_T^{AL}$ minimizes $\widetilde{Q}_T^{AL}$ and belongs to $\overset{\circ}{\Phi}$. Therefore, under Assumptions A6, A8 and for T large, the function $\widetilde{L}_T$ is differentiable on $\mathcal{V}(\boldsymbol{\phi}_0)\times\overset{\circ}{\Phi}$ and we have for $j\in\{1,\ldots,\nu\}$:

$$\begin{cases}\frac{1}{T}\frac{\partial\widetilde{L}_T}{\partial\varphi_j}\left(\widehat{\boldsymbol{\phi}}_T,\widehat{\boldsymbol{\varphi}}_T^{AL}\right)+\frac{\lambda_{j,T}}{|\widehat{\phi}_{j,T}|^{\tau}}\operatorname{sign}\left(\widehat{\varphi}_{j,T}^{AL}\right)=0\text{ if }\widehat{\varphi}_{j,T}^{AL}\neq0,\\\left|\frac{1}{T}\frac{\partial\widetilde{L}_T}{\partial\varphi_j}\left(\widehat{\boldsymbol{\phi}}_T,\widehat{\boldsymbol{\varphi}}_T^{AL}\right)\right|\leq\frac{\lambda_{j,T}}{|\widehat{\phi}_{j,T}|^{\tau}}\text{ otherwise.}\end{cases}$$

Taking $\underline{\lambda}_T:=1+\sum_{j=1}^{\nu}\sup_{\boldsymbol{\alpha}\in\mathcal{V}(\boldsymbol{\phi}_0)\times\overset{\circ}{\Phi}}\left|\frac{1}{T}\frac{\partial L_T}{\partial\varphi_j}(\boldsymbol{\alpha})\right|$, we have:

$$\underline{\lambda}_T\leq1+\sum_{j=1}^{\nu}\frac{1}{T}\sum_{t=1}^{T}\sup_{\boldsymbol{\alpha}\in\mathcal{V}(\boldsymbol{\phi}_0)\times\overset{\circ}{\Phi}}\left|\frac{\partial l_t}{\partial\varphi_j}(\boldsymbol{\alpha})\right|\xrightarrow[T\to\infty]{a.s.}1+\sum_{j=1}^{\nu}\mathbb{E}\left[\sup_{\boldsymbol{\alpha}\in\mathcal{V}(\boldsymbol{\phi}_0)\times\overset{\circ}{\Phi}}\left|\frac{\partial l_1}{\partial\varphi_j}(\boldsymbol{\alpha})\right|\right].$$

It follows that $\underline{\lambda}_T=O_{\mathbb{P}}(1)$. Using the result of part (b) of the proof of Theorem 3.2, the impact of the initial values is an $o_{\mathbb{P}}(1)$ uniformly on $\mathcal{V}(\boldsymbol{\phi}_0)\times\overset{\circ}{\Phi}$.

(b) Asymptotic distribution of the active components. Let us denote:

$$\boldsymbol{y}_T=\operatorname*{arg\,min}_{\boldsymbol{y}\in\mathbb{R}^{\nu_0}}\widetilde{E}_{\mathbb{A},T}(\boldsymbol{y})=\sqrt{T}\left(\widehat{\boldsymbol{\varphi}}_{\mathbb{A},T}^{AL}-\boldsymbol{\phi}_{\mathbb{A},0}\right).$$

For the active components, the penalty part converges almost surely to a finite limit under Assumption A8. A slight adaptation of the proof of Theorem 3.2 with Slutsky's Lemma gives:

$$\boldsymbol{y}_T\xrightarrow[T\to\infty]{d}\boldsymbol{y}^*:=\operatorname*{arg\,min}_{\boldsymbol{y}\in\mathbb{R}^{\nu_0}}E_{\mathbb{A},\infty}(\boldsymbol{y})=\operatorname*{arg\,min}_{\boldsymbol{y}\in\mathbb{R}^{\nu_0}}G_{\mathbb{A},\infty}(\boldsymbol{y}).$$

And since it is an optimum we get $\mathbf{0}=\frac{\partial G_{\mathbb{A},\infty}}{\partial\boldsymbol{y}}(\boldsymbol{y}^*)=\boldsymbol{W}_{\mathbb{A}}+4\boldsymbol{M}_{11}\boldsymbol{y}^*+4\boldsymbol{V}_{\mathbb{A}}$. Therefore:

$$\boldsymbol{y}^*=-\frac{1}{4}\boldsymbol{M}_{11}^{-1}\left(\boldsymbol{W}_{\mathbb{A}}+4\boldsymbol{V}_{\mathbb{A}}\right).\qquad(36)$$

We have $\underline{\lambda}_T = O_{\mathbb{P}}(1)$ and for $j \in \overline{\mathbb{A}}$, $\frac{\lambda_{j,T}}{\left|\widehat{\phi}_{j,T}\right|^{\tau}} = \frac{T^{\frac{\tau}{2}}\lambda_{j,T}}{\left|\sqrt{T}\widehat{\phi}_{j,T}\right|^{\tau}} \xrightarrow[T\to\infty]{\mathbb{P}} \infty$. For $\xi > 0$, using the results of part (a):

$$\mathbb{P}\left[\left|\left|\widehat{\boldsymbol{\varphi}}_T^{AL} - \begin{pmatrix} \frac{\boldsymbol{y}_T}{\sqrt{T}} + \boldsymbol{\phi}_{\mathbb{A},0} \\ \boldsymbol{0} \end{pmatrix}\right|\right|_1 > \xi\right] = \mathbb{P}\left[\left|\left|\widehat{\boldsymbol{\varphi}}_{\mathbb{A},T}^{AL} - \left(\frac{\boldsymbol{y}_T}{\sqrt{T}} + \boldsymbol{\phi}_{\mathbb{A},0}\right)\right|\right|_1 + \left|\left|\widehat{\boldsymbol{\varphi}}_{\overline{\mathbb{A}},T}^{AL}\right|\right|_1 > \xi\right]$$

$$= \mathbb{P}\left[\left|\left|\widehat{\boldsymbol{\varphi}}_{\mathbb{A},T}^{AL} - \left(\frac{\boldsymbol{y}_T}{\sqrt{T}} + \boldsymbol{\phi}_{\mathbb{A},0}\right)\right|\right|_1 + \left|\left|\widehat{\boldsymbol{\varphi}}_{\overline{\mathbb{A}},T}^{AL}\right|\right|_1 > \xi \text{ and } \forall j \in \overline{\mathbb{A}}, \underline{\lambda}_T < \frac{\lambda_{j,T}}{\left|\widehat{\phi}_{j,T}\right|^{\tau}}\right]$$

$$+ \mathbb{P}\left[\left|\left|\widehat{\boldsymbol{\varphi}}_{\mathbb{A},T}^{AL} - \left(\frac{\boldsymbol{y}_T}{\sqrt{T}} + \boldsymbol{\phi}_{\mathbb{A},0}\right)\right|\right|_1 + \left|\left|\widehat{\boldsymbol{\varphi}}_{\overline{\mathbb{A}},T}^{AL}\right|\right|_1 > \xi \text{ and } \exists j \in \overline{\mathbb{A}}, \underline{\lambda}_T \geq \frac{\lambda_{j,T}}{\left|\widehat{\phi}_{j,T}\right|^{\tau}}\right]$$

$$= \mathbb{P}\left[\left|\left|\widehat{\boldsymbol{\varphi}}_{\mathbb{A},T}^{AL} - \left(\frac{\boldsymbol{y}_T}{\sqrt{T}} + \boldsymbol{\phi}_{\mathbb{A},0}\right)\right|\right|_1 > \xi \text{ and } \forall j \in \overline{\mathbb{A}}, \underline{\lambda}_T < \frac{\lambda_{j,T}}{\left|\widehat{\phi}_{j,T}\right|^{\tau}}\right]$$

$$+ \mathbb{P}\left[\left|\left|\widehat{\boldsymbol{\varphi}}_{\mathbb{A},T}^{AL} - \left(\frac{\boldsymbol{y}_T}{\sqrt{T}} + \boldsymbol{\phi}_{\mathbb{A},0}\right)\right|\right|_1 + \left|\left|\widehat{\boldsymbol{\varphi}}_{\overline{\mathbb{A}},T}^{AL}\right|\right|_1 > \xi \text{ and } \exists j \in \overline{\mathbb{A}}, \underline{\lambda}_T \geq \frac{\lambda_{j,T}}{\left|\widehat{\phi}_{j,T}\right|^{\tau}}\right].$$

On the one hand $\mathbb{P}\left[\forall j \in \overline{\mathbb{A}}, \underline{\lambda}_T < \frac{\lambda_{j,T}}{\left|\widehat{\phi}_{j,T}\right|^{\tau}}\right] \xrightarrow[T\to\infty]{} 1$ giving:

$$\mathbb{P}\left[\left|\left|\widehat{\boldsymbol{\varphi}}_{\mathbb{A},T}^{AL} - \left(\frac{\boldsymbol{y}_T}{\sqrt{T}} + \boldsymbol{\phi}_{\mathbb{A},0}\right)\right|\right| + \left|\left|\widehat{\boldsymbol{\varphi}}_{\overline{\mathbb{A}},T}^{AL}\right|\right|_1 > \xi \text{ and } \exists j \in \overline{\mathbb{A}}, \underline{\lambda}_T \geq \frac{\lambda_{j,T}}{\left|\widehat{\phi}_{j,T}\right|^{\tau}}\right] \xrightarrow[T\to\infty]{} 0.$$

On the other hand, in the event $\left\{\forall j \in \overline{\mathbb{A}}, \underline{\lambda}_T < \frac{\sqrt{T}\lambda_{j,T}}{\left|\widehat{\phi}_{j,T}\right|^{\tau}}\right\}$ we have $\widehat{\boldsymbol{\varphi}}_{\mathbb{A},T}^{AL} = \frac{\boldsymbol{y}_T}{\sqrt{T}} + \boldsymbol{\phi}_{\mathbb{A},0}$. The conclusion follows.

(c) Selection consistency. The results of part (b) gives $\widehat{\boldsymbol{\varphi}}_{\mathbb{A},T}^{AL} \xrightarrow[T\to\infty]{\mathbb{P}} \boldsymbol{\phi}_{\mathbb{A},0}$. Therefore, $\mathbb{P}\left[\mathbb{A} \subset \mathbb{A}_T\right] \xrightarrow[T\to\infty]{} 1$. We also have for $j \in \overline{\mathbb{A}}$:

$$\mathbb{P}\left[j \in \mathbb{A}_T\right] = \mathbb{P}\left[j \in \mathbb{A}_T \text{ and } \underline{\lambda}_T \geq \frac{\lambda_{j,T}}{\left|\widehat{\phi}_{j,T}\right|^{\tau}}\right] + \mathbb{P}\left[j \in \mathbb{A}_T \text{ and } \underline{\lambda}_T < \frac{\lambda_{j,T}}{\left|\widehat{\phi}_{j,T}\right|^{\tau}}\right].$$

By the results of part (a) we have $\mathbb{P}\left[j \in \mathbb{A}_T \text{ and } \underline{\lambda}_T < \frac{\lambda_{j,T}}{\left|\widehat{\phi}_{j,T}\right|^{\tau}}\right] = 0$ and $\mathbb{P}\left[\underline{\lambda}_T < \frac{\lambda_{j,T}}{\left|\widehat{\phi}_{j,T}\right|^{\tau}}\right] \xrightarrow[T\to\infty]{} 1$. The conclusion follows. □

Proof of Corollary 3.1. The proof of this Corollary relies on standard arguments. We only provide a sketch of prove to show that the assumptions of the theorems of Section 3 are satisfied. For $\boldsymbol{\alpha} \in \Phi \times \Phi$, under Assumptions A15-A16 we have:

$$\sigma_t^2(\boldsymbol{\phi}) = \sum_{j\in\mathbb{N}} b^j \left(\underline{w} + \sum_{i=1}^{p} a_i \epsilon_{t-i-j}^2 + \boldsymbol{\varsigma}'\boldsymbol{X}_{t-1-j}\right),$$

$$\frac{\partial \sigma_t^2}{\partial \boldsymbol{\phi}}(\boldsymbol{\phi}) = \sum_{j\in\mathbb{N}} b^j \left(1, \epsilon_{t-1-j}^2, \ldots, \epsilon_{t-p-j}^2, \sigma_{t-j}^2(\boldsymbol{\phi}), \boldsymbol{X}_{t-1-j}'\right)',$$

$$f_t(\boldsymbol{\alpha}) = \left(1, \epsilon_{t-1}^2, \ldots, \epsilon_{t-p}^2, \sigma_{t-1}^2(\boldsymbol{\phi}), \boldsymbol{X}_{t-1}'\right)\boldsymbol{\varphi}.$$

We show that Assumptions A2-A3 are satisfied. Since $\sup_{\boldsymbol{\phi}\in\Phi}|b| < \rho < c < 1$, we have:

$$a_t = \sup_{\boldsymbol{\alpha}\in\Phi\times\Phi}\left|\widetilde{f}_t(\boldsymbol{\alpha}) - f_t(\boldsymbol{\alpha})\right| \leq \rho \sup_{\boldsymbol{\phi}\in\Phi}\left|\widetilde{\sigma}_{t-1}^2(\boldsymbol{\alpha}) - \sigma_{t-1}^2(\boldsymbol{\alpha})\right| \leq K\rho^t,$$

$$d_t = |w_t - \widetilde{w}_t| \leq Kc^t,$$

$$\sup_{\boldsymbol{\alpha}\in\Phi\times\Phi}|f_t(\boldsymbol{\alpha})| \leq K\left(1 + \sum_{i=1}^{p}\epsilon_{t-i}^2 + ||\boldsymbol{X}_{t-1}||_1 + \sum_{j\in\mathbb{N}}\rho^j\left(1 + \sum_{i=1}^{p}\epsilon_{t-i-j}^2 + ||\boldsymbol{X}_{t-1-j}||_1\right)\right).$$

The final term admits finite low-order moments, since $|\epsilon_t|^{2s}$ and $||\boldsymbol{X}_t||^s$ belong to L^1. Hence, Assumptions A2 and A3 are fulfilled. Moreover, because $c > \rho$, the same bound establishes Assumption A4. Finally, Assumption A17 guarantees that Assumption A14 holds. Verification of the remaining assumptions in Section 3 follows by standard arguments and is therefore omitted. □

Proof of Theorem I.1. To prove the Theorem, we will establish the following intermediate results.

(a) $\sup_{\boldsymbol{\beta}\in\Phi\times\Theta\times\Theta}\frac{1}{T}\left|\mathcal{L}_T(\boldsymbol{\beta}) - \widetilde{\mathcal{L}}_T(\boldsymbol{\beta})\right| \xrightarrow[T\to\infty]{a.s.} 0.$

(b) $\forall\boldsymbol{\beta}\in\mathcal{V}(\boldsymbol{\rho}_0)\times\overset{\circ}{\Theta}$, $\mathcal{Q}_\infty(\boldsymbol{\beta}) = \mathbb{E}[\ell_1(\boldsymbol{\beta})] + \sum_{j=1}^n \iota_\infty|\vartheta_j|$ exists and is finite.

(c) $\widehat{\boldsymbol{\vartheta}}_T \xrightarrow[T\to\infty]{a.s.} \arg\min_{\boldsymbol{\vartheta}\in\Theta}\mathcal{Q}_\infty(\boldsymbol{\rho}_0, \boldsymbol{\vartheta})$.

(d) $\arg\min_{\boldsymbol{\vartheta}\in\Theta}\mathcal{Q}_\infty(\boldsymbol{\rho}_0, \boldsymbol{\vartheta}) = \boldsymbol{\theta}_0$ if $\boldsymbol{\iota}_\infty = \mathbf{0}$.

(a) Asymptotic irrelevance of the initial values. Let $\boldsymbol{\beta}\in\Phi\times\Theta\times\Theta$:

$$\begin{aligned}
\left|\ell_t(\boldsymbol{\beta}) - \widetilde{\ell}_t(\boldsymbol{\beta})\right| &= \left|\frac{\widetilde{\omega}_t^2\left(\epsilon_t^2(\boldsymbol{\phi}) - g_t(\boldsymbol{\beta})\right)^2 - \omega_t^2\left(\widetilde{\epsilon}_t^2(\boldsymbol{\phi}) - \widetilde{g}_t(\boldsymbol{\beta})\right)^2}{\omega_t^2\widetilde{\omega}_t^2}\right| \\
&\leq \frac{2c_t\left(1 + y_t^4 + \mu_t^4(\boldsymbol{\phi}) + g_t^2(\boldsymbol{\beta})\right)}{\underline{\omega}^4} + \frac{\left|\left(\widetilde{\epsilon}_t^2(\boldsymbol{\phi}) - \widetilde{g}_t(\boldsymbol{\beta})\right)^2 - \left(\epsilon_t^2(\boldsymbol{\phi}) - g_t(\boldsymbol{\beta})\right)^2\right|}{\underline{\omega}^2} \\
&\leq \frac{2c_t\left(1 + y_t^4 + \mu_t^4(\boldsymbol{\phi}) + g_t^2(\boldsymbol{\beta})\right)}{\underline{\omega}^4} \\
&+ \frac{|\mu_t(\boldsymbol{\phi}) - \widetilde{\mu}_t(\boldsymbol{\phi})|\,|-2y_t + \mu_t(\boldsymbol{\phi}) + \widetilde{\mu}_t(\boldsymbol{\phi})|\left(4y_t^2 + 2\mu_t^2(\boldsymbol{\phi}) + 2\widetilde{\mu}_t^2(\boldsymbol{\phi}) + |g_t(\boldsymbol{\beta})| + |\widetilde{g}_t(\boldsymbol{\beta})|\right)}{\underline{\omega}^2} \\
&+ \frac{|g_t(\boldsymbol{\beta}) - \widetilde{g}_t(\boldsymbol{\beta})|\left(4y_t^2 + 2\mu_t^2(\boldsymbol{\phi}) + 2\widetilde{\mu}_t^2(\boldsymbol{\phi}) + |g_t(\boldsymbol{\beta})| + |\widetilde{g}_t(\boldsymbol{\beta})|\right)}{\underline{\omega}^2}
\end{aligned}$$

Under Assumption A18, for t large enough we have almost surely $\sup_{\boldsymbol{\beta}\in\Phi\times\Theta\times\Theta}|\widetilde{g}_t(\boldsymbol{\beta})| \leq 1 + \sup_{\boldsymbol{\beta}\in\Phi\times\Theta\times\Theta}|g_t(\boldsymbol{\beta})|$ and $\sup_{\boldsymbol{\phi}\in\Phi}|\widetilde{\mu}_t(\boldsymbol{\beta})| \leq 1 + \sup_{\boldsymbol{\phi}\in\Phi}|\mu_t(\boldsymbol{\beta})|$. Therefore:

$$\sup_{\boldsymbol{\beta}\in\Phi\times\Theta\times\Theta}\left|\ell_t(\boldsymbol{\beta}) - \widetilde{\ell}_t(\boldsymbol{\beta})\right|$$

$$\leq K\frac{c_t\left(1+y_t^4+\underset{\boldsymbol{\phi}\in\Phi}{\sup}\mu_t^4\left(\boldsymbol{\phi}\right)+\underset{\boldsymbol{\beta}\in\Phi\times\Theta\times\Theta}{\sup}g_t^2\left(\boldsymbol{\beta}\right)\right)}{\underline{\omega}^4}+K\frac{e_t\left(1+y_t^2+\underset{\boldsymbol{\phi}\in\Phi}{\sup}\mu_t^2\left(\boldsymbol{\phi}\right)+\underset{\boldsymbol{\beta}\in\Phi\times\Theta\times\Theta}{\sup}\left|g_t\left(\boldsymbol{\beta}\right)\right|\right)}{\underline{\omega}^2}$$

Under Assumption A19, the right hand side of the inequality goes to 0 almost surely as t goes to infinity. Therefore, $\underset{\boldsymbol{\beta}\in\Phi\times\Theta\times\Theta}{\sup}\left|\ell_t\left(\boldsymbol{\beta}\right)-\widetilde{\ell}_t\left(\boldsymbol{\beta}\right)\right|\xrightarrow[t\to\infty]{a.s.}0$. Cesàro's Lemma gives (a).

(b) Existence of the limit loss function. We start by noting that:

$$\ell_t\left(\boldsymbol{\beta}\right)=\left(\frac{\epsilon_t^2\left(\boldsymbol{\phi}\right)-g_t\left(\boldsymbol{\beta}\right)}{\omega_t}\right)^2=\left(\frac{\left(\mu_t-\mu_t\left(\boldsymbol{\phi}\right)\right)\left(2\sigma_t\eta_t+\mu_t-\mu_t\left(\boldsymbol{\phi}\right)\right)}{\omega_t}+\frac{\sigma_t^2}{\omega_t}\left(\eta_t^2-1\right)+\frac{\sigma_t^2-g_t\left(\boldsymbol{\beta}\right)}{\omega_t}\right)^2.$$

Under Assumptions A21-A22, using the mean value Theorem and the compactness of $\Phi\times\Theta$:

$$\underset{\boldsymbol{\phi}\in\mathcal{V}(\boldsymbol{\phi}_0)}{\sup}\left|\frac{\mu_t-\mu_t\left(\boldsymbol{\phi}\right)}{\sqrt{\omega_t}}\right|\leq\frac{K}{\sqrt{\omega_t}}\underset{\boldsymbol{\phi}\in\mathcal{V}(\boldsymbol{\phi}_0)}{\sup}\left|\left|\frac{\partial\mu_t}{\partial\boldsymbol{\phi}}\left(\boldsymbol{\phi}\right)\right|\right|,\tag{37}$$

$$\underset{\boldsymbol{\beta}\in\mathcal{V}(\boldsymbol{\rho}_0)\times\mathring{\Theta}}{\sup}\left|\frac{\sigma_t^2-g_t\left(\boldsymbol{\beta}\right)}{\omega_t}\right|\leq\frac{K}{\omega_t}\underset{\boldsymbol{\beta}\in\mathcal{V}(\boldsymbol{\rho}_0)\times\mathring{\Theta}}{\sup}\left|\left|\frac{\partial g_t}{\partial\boldsymbol{\beta}}\left(\boldsymbol{\beta}\right)\right|\right|,\tag{38}$$

By Assumption A20, using the independence of η_t and $\mathcal{F}_{t-1}$ gives:

$$\mathbb{E}\left[\underset{\boldsymbol{\beta}\in\mathcal{V}(\boldsymbol{\rho}_0)\times\mathring{\Theta}}{\sup}\ell_1\left(\boldsymbol{\beta}\right)\right]<\infty.\tag{39}$$

Moreover, $\boldsymbol{\iota}_T\underset{T\to\infty}{\longrightarrow}\boldsymbol{\iota}_T<\infty$. We obtain the existence and the finiteness of $\mathcal{Q}_\infty\left(\boldsymbol{\beta}\right)$.

(c) Convergence of the minimizers. Under Assumption A1 and Equation (39), the conditions of Lemma II.1 are satisfied and we have for $\boldsymbol{\vartheta}\in\mathring{\Theta}$, $\mathcal{Q}_T\left(\widehat{\boldsymbol{\rho}}_T,\boldsymbol{\vartheta}\right)\xrightarrow[T\to\infty]{a.s.}\mathcal{Q}_\infty\left(\boldsymbol{\rho}_0,\boldsymbol{\vartheta}\right)$. Using (a) we obtain $\widetilde{\mathcal{Q}}_T\left(\widehat{\boldsymbol{\rho}}_T,\boldsymbol{\vartheta}\right)\xrightarrow[T\to\infty]{a.s.}\mathcal{Q}_\infty\left(\boldsymbol{\rho}_0,\boldsymbol{\vartheta}\right)$. Under Assumption A25, the functions $\widetilde{\mathcal{Q}}_T\left(\widehat{\boldsymbol{\rho}}_T,\cdot\right)$ and $\mathcal{Q}_\infty\left(\boldsymbol{\rho}_0,\cdot\right)$ are strictly convex on Θ, thus having a unique minima. Under Assumption A23, these minima lie in $\mathring{\Theta}$. Lemma II.2 gives (c).

(d) Asymptotic unbiasedness under vanishing penalty. Let $\boldsymbol{\vartheta}\in\mathcal{V}\left(\boldsymbol{\theta}_0\right)$, if $\boldsymbol{\iota}_\infty=\mathbf{0}$ then $\mathcal{Q}_\infty\left(\boldsymbol{\rho}_0,\boldsymbol{\vartheta}\right)=\mathbb{E}\left[\ell_1\left(\boldsymbol{\rho}_0,\boldsymbol{\vartheta}\right)\right]=\mathbb{E}\left[\left(\frac{\sigma_t^2\eta_t^2-g_t(\boldsymbol{\rho}_0,\boldsymbol{\vartheta})}{\omega_t}\right)^2\right]$. Taking the derivative with respect to $\boldsymbol{\vartheta}$ at $\boldsymbol{\theta}_0$ gives $\frac{\partial\mathcal{Q}_\infty}{\partial\boldsymbol{\vartheta}}\left(\boldsymbol{\beta}_0\right)=-2\mathbb{E}\left[\frac{\sigma_t^2\left(\eta_t^2-1\right)}{\omega_1^2}\frac{\partial g_1}{\partial\boldsymbol{\vartheta}}\left(\boldsymbol{\beta}_0\right)\right]=\mathbf{0}$. Which concludes the proof. □

Proof of Theorem I.2. We split the proof into the following parts:

(a) $\left(\Psi_T\right)_{T\geq1}$ is an exhaustion of $\mathbb{R}^n$.

(b) $\forall k\in\mathbb{N}^*$, $\underset{\boldsymbol{\psi}\in\Psi_k}{\sup}\left|\mathcal{G}_T\left(\boldsymbol{\psi}\right)-\widetilde{\mathcal{G}}_T\left(\boldsymbol{\psi}\right)\right|\xrightarrow[T\to\infty]{\mathbb{P}}0$.

(c) $\mathbb{E}\left[\underset{\boldsymbol{\beta}\in\mathcal{V}(\boldsymbol{\beta}_0)}{\sup}\left|\left|\frac{\partial\ell_1}{\partial\boldsymbol{\beta}}\left(\boldsymbol{\beta}\right)\right|\right|^2\right]<\infty$ and $\mathbb{E}\left[\underset{\boldsymbol{\beta}\in\mathcal{V}(\boldsymbol{\beta}_0)}{\sup}\left|\left|\frac{\partial^2\ell_1}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\left(\boldsymbol{\beta}\right)\right|\right|\right]<\infty$.

(d) $\begin{pmatrix}\frac{1}{\sqrt{T}}\frac{\partial\mathcal{L}_T}{\partial\boldsymbol{\beta}}\left(\boldsymbol{\beta}_0\right)\\\sqrt{T}\left(\widehat{\boldsymbol{\rho}}_T-\boldsymbol{\rho}_0\right)\end{pmatrix}\xrightarrow[T\to\infty]{d}\begin{pmatrix}\boldsymbol{\mathcal{W}}\\\boldsymbol{\mathcal{Z}}^C\end{pmatrix}$.

(e) $\sqrt{T}\left(\widehat{\boldsymbol{\vartheta}}_T - \boldsymbol{\theta}_0\right) \xrightarrow[T\to\infty]{d} \arg\min_{\boldsymbol{\psi}\in\mathbb{R}^n} \mathcal{E}_\infty\left(\boldsymbol{\psi}\right)$.

(a) Exhaustion of $\mathbb{R}^n$ by an increasing sequence of compacta. The proof of this part follows by the same arguments as the proof of Theorem 3.2 part (a).

(b) Asymptotic decrease of the effect of the initial values. Let $k \in \mathbb{N}^*$. Under Assumptions A21, A24, A27 and for T large enough such that almost surely $\begin{pmatrix} \widehat{\boldsymbol{\rho}}_T \\ \frac{\boldsymbol{\psi}}{\sqrt{T}} + \boldsymbol{\theta}_0 \end{pmatrix} \in \mathcal{V}\left(\boldsymbol{\beta}_0\right)$ we have the following:

$$\sup_{\boldsymbol{\psi}\in\Psi_k} \left|\mathcal{G}_T\left(\boldsymbol{\psi}\right) - \widetilde{\mathcal{G}}_T\left(\boldsymbol{\psi}\right)\right| \leq K\left(\left|\left|\sqrt{T}\left(\widehat{\boldsymbol{\rho}}_T - \boldsymbol{\rho}_0\right)\right|\right| + 1\right)\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\sup_{\boldsymbol{\beta}\in\mathcal{V}(\boldsymbol{\beta}_0)}\left|\left|\frac{\partial \ell_t}{\partial \boldsymbol{\beta}}\left(\boldsymbol{\beta}\right) - \frac{\partial \widetilde{\ell}_t}{\partial \boldsymbol{\beta}}\left(\boldsymbol{\beta}\right)\right|\right|.$$

Under Assumption A29 and for t large enough:

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathcal{V}(\boldsymbol{\beta}_0)}\left|\left|\frac{\partial \ell_t}{\partial \boldsymbol{\beta}}\left(\boldsymbol{\beta}\right) - \frac{\partial \widetilde{\ell}_t}{\partial \boldsymbol{\beta}}\left(\boldsymbol{\beta}\right)\right|\right| \\
&\leq 2\sup_{\boldsymbol{\beta}\in\mathcal{V}(\boldsymbol{\beta}_0)}\left|\left|\frac{2c_t\left(\epsilon_t^2\left(\boldsymbol{\phi}\right) - g_t\left(\boldsymbol{\beta}\right)\right)\left(\left(\mu_t\left(\boldsymbol{\phi}\right) - y_t\right)\frac{\partial \mu_t}{\partial \boldsymbol{\beta}}\left(\boldsymbol{\beta}\right) - \frac{\partial g_t}{\partial \boldsymbol{\beta}}\left(\boldsymbol{\beta}\right)\right)}{\underline{\omega}^4}\right|\right| \\
&+ 4\sup_{\boldsymbol{\beta}\in\mathcal{V}(\boldsymbol{\beta}_0)}\left|\left|\frac{\left(\epsilon_t^2\left(\boldsymbol{\phi}\right) - g_t\left(\boldsymbol{\beta}\right)\right)\left[\left(\mu_t\left(\boldsymbol{\phi}\right) - \widetilde{\mu}_t\left(\boldsymbol{\phi}\right)\right)\frac{\partial \mu_t}{\partial \boldsymbol{\beta}}\left(\boldsymbol{\beta}\right) - \left(\widetilde{\mu}_t\left(\boldsymbol{\phi}\right) - y_t\right)\left(\frac{\partial \widetilde{\mu}_t}{\partial \boldsymbol{\beta}}\left(\boldsymbol{\beta}\right) - \frac{\partial \mu_t}{\partial \boldsymbol{\beta}}\left(\boldsymbol{\beta}\right)\right)\right]}{\underline{\omega}^2}\right|\right| \\
&+ 2\sup_{\boldsymbol{\beta}\in\mathcal{V}(\boldsymbol{\beta}_0)}\left|\left|\frac{\left(\epsilon_t^2\left(\boldsymbol{\phi}\right) - g_t\left(\boldsymbol{\beta}\right)\right)\left[\frac{\partial g_t}{\partial \boldsymbol{\beta}}\left(\boldsymbol{\beta}\right) - \frac{\partial \widetilde{g}_t}{\partial \boldsymbol{\beta}}\left(\boldsymbol{\beta}\right)\right]}{\underline{\omega}^2}\right|\right| \\
&+ 2\sup_{\boldsymbol{\beta}\in\mathcal{V}(\boldsymbol{\beta}_0)}\left|\left|\frac{\left[\left(\mu_t\left(\boldsymbol{\phi}\right) - \widetilde{\mu}_t\left(\boldsymbol{\phi}\right)\right)\left(2y_t - \widetilde{\mu}_t\left(\boldsymbol{\phi}\right) - \mu_t\left(\boldsymbol{\phi}\right)\right) - \left(g_t\left(\boldsymbol{\beta}\right) - \widetilde{g}_t\left(\boldsymbol{\beta}\right)\right)\right]\left(2\left(\widetilde{\mu}_t\left(\boldsymbol{\phi}\right) - y_t\right)\frac{\partial \widetilde{\mu}_t}{\partial \boldsymbol{\beta}}\left(\boldsymbol{\beta}\right) - \frac{\partial \widetilde{g}_t}{\partial \boldsymbol{\beta}}\left(\boldsymbol{\beta}\right)\right)}{\underline{\omega}^2}\right|\right| \\
&\leq K\frac{c_t\left(1 + y_t^2 + \sup_{\boldsymbol{\phi}\in\mathcal{V}(\boldsymbol{\phi}_0)}\mu_t^2\left(\boldsymbol{\phi}\right) + \sup_{\boldsymbol{\beta}\in\mathcal{V}(\boldsymbol{\beta}_0)}\left|g_t\left(\boldsymbol{\beta}\right)\right|\right)\left(1 + \left|y_t\right| + \sup_{\boldsymbol{\phi}\in\mathcal{V}(\boldsymbol{\phi}_0)}\left|\mu_t\left(\boldsymbol{\phi}\right)\right|\right)\sup_{\boldsymbol{\phi}\in\mathcal{V}(\boldsymbol{\phi}_0)}\left|\left|\frac{\partial \mu_t}{\partial \boldsymbol{\phi}}\left(\boldsymbol{\phi}\right)\right|\right|}{\underline{\omega}^4} \\
&+ K\frac{c_t\left(1 + y_t^2 + \sup_{\boldsymbol{\phi}\in\mathcal{V}(\boldsymbol{\phi}_0)}\mu_t^2\left(\boldsymbol{\phi}\right) + \sup_{\boldsymbol{\beta}\in\mathcal{V}(\boldsymbol{\beta}_0)}\left|g_t\left(\boldsymbol{\beta}\right)\right|\right)\sup_{\boldsymbol{\beta}\in\mathcal{V}(\boldsymbol{\beta}_0)}\left|\left|\frac{\partial g_t}{\partial \boldsymbol{\beta}}\left(\boldsymbol{\beta}\right)\right|\right|}{\underline{\omega}^4} \\
&+ K\frac{\varrho_t\left(1 + \left|y_t\right|^2 + \sup_{\boldsymbol{\phi}\in\Phi}\left|\mu_t\left(\boldsymbol{\phi}\right)\right|^2 + \sup_{\boldsymbol{\beta}\in\Phi\times\Theta\times\Theta}\left|g_t\left(\boldsymbol{\beta}\right)\right|\right)}{\underline{\omega}^2} \\
&+ K\frac{e_t\left(\left(1 + \left|y_t\right| + \sup_{\boldsymbol{\phi}\in\Phi}\left|\mu_t\left(\boldsymbol{\phi}\right)\right|\right)\sup_{\boldsymbol{\phi}\in\Phi}\left|\left|\frac{\partial \mu_t}{\partial \boldsymbol{\phi}}\left(\boldsymbol{\phi}\right)\right|\right| + \sup_{\boldsymbol{\beta}\in\Phi\times\Theta\times\Theta}\left|\left|\frac{\partial g_t}{\partial \boldsymbol{\beta}}\left(\boldsymbol{\beta}\right)\right|\right|\right)}{\underline{\omega}^2}
\end{aligned}$$

The right hand term of the last inequality is almost surely of order $O\left(t^{-\kappa}\right)$ with $\kappa > \frac{1}{2}$. Under Assumption A28, $\left|\left|\sqrt{T}\left(\widehat{\boldsymbol{\rho}}_T - \boldsymbol{\rho}_0\right)\right|\right| = O_{\mathbb{P}}\left(1\right)$. The result (a) follows.

(c) Integrability of the suprema of the loss function's first and second derivatives.

Under Assumption A27 and for $\boldsymbol{\beta} \in \mathcal{V}(\boldsymbol{\beta}_0)$, we have:

$$\begin{aligned}\frac{\partial \ell_t}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) &= \frac{2}{\omega_t^2}\left(\epsilon_t^2(\boldsymbol{\phi}) - g_t(\boldsymbol{\beta})\right)\left(2\left(\mu_t(\boldsymbol{\phi}) - y_t\right)\frac{\partial \mu_t}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) - \frac{\partial g_t}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta})\right)\\ &= \frac{2}{\omega_t^2}\left(\epsilon_t^2(\boldsymbol{\phi}) - g_t(\boldsymbol{\beta})\right)\left(2\left(\mu_t(\boldsymbol{\phi}) - \mu_t - \sigma_t\eta_t\right)\frac{\partial \mu_t}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) - \frac{\partial g_t}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta})\right).\end{aligned}$$

Under Assumptions A26-A27 and with the same arguments as in the proof of Theorem I.1 part (b) with Equations (37)-(38) we obtain:

$$\mathbb{E}\left[\sup_{\boldsymbol{\alpha} \in \mathcal{V}(\boldsymbol{\beta}_0)} \ell_1^2(\boldsymbol{\beta})\right] < \infty. \tag{40}$$

Cauchy-Schwartz's inequality along with the moments given by Assumptions A26-A27 and Equation (40) give:

$$\mathbb{E}\left[\sup_{\boldsymbol{\beta} \in \mathcal{V}(\boldsymbol{\beta}_0)} \left|\left|\frac{\partial \ell_1}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta})\right|\right|^2\right] < \infty. \tag{41}$$

And the second derivative is:

$$\begin{aligned}\frac{\partial^2 \ell_1}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}(\boldsymbol{\beta}) = &\frac{2}{\omega_t^2}P\left(2\left(\mu_t(\boldsymbol{\phi}) - y_t\right)\frac{\partial \mu_t}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) - \frac{\partial g_t}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta})\right)\\ &+ \frac{2}{\omega_t^2}\left(\epsilon_t^2(\boldsymbol{\phi}) - g_t(\boldsymbol{\beta})\right)\left(2P\left(\frac{\partial \mu_t}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta})\right) + 2\left(\mu_t(\boldsymbol{\phi}) - y_t\right)\frac{\partial^2 \mu_t}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}(\boldsymbol{\beta}) - \frac{\partial^2 g_t}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}(\boldsymbol{\beta})\right)\\ = &\frac{2}{\omega_t^2}P\left(2\left(\mu_t(\boldsymbol{\phi}) - \mu_t - \sigma_t\eta_t\right)\frac{\partial \mu_t}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) - \frac{\partial g_t}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta})\right)\\ &+ \frac{2}{\omega_t^2}\left(\epsilon_t^2(\boldsymbol{\phi}) - g_t(\boldsymbol{\beta})\right)\left(2P\left(\frac{\partial \mu_t}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta})\right) + 2\left(\mu_t(\boldsymbol{\phi}) - \mu_t - \sigma_t\eta_t\right)\frac{\partial^2 \mu_t}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}(\boldsymbol{\beta}) - \frac{\partial^2 g_t}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}(\boldsymbol{\beta})\right).\end{aligned}$$

With Equations (41) and (40) we obtain:

$$\mathbb{E}\left[\sup_{\boldsymbol{\beta} \in \mathcal{V}(\boldsymbol{\beta}_0)} \left|\left|\frac{\partial^2 \ell_1}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}(\boldsymbol{\beta})\right|\right|\right] < \infty. \tag{42}$$

(d) C.L.T. for martingale increments. Under Assumption A27, $\frac{\partial \ell_t}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0)$ exists and using the Bahadur-type expansion given by Assumption A28, we define $\boldsymbol{\mathcal{U}}_t = \begin{pmatrix} \frac{\partial \ell_t}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0) \\ \boldsymbol{\Lambda}_t \xi(\eta_t) \end{pmatrix}$. On the one hand, $\boldsymbol{\Lambda}_t$ and $\xi(\eta_t)$ are independent and belong to L^2. On the other hand, using the Equation (41) of part (c), $\boldsymbol{\mathcal{U}}_t$ belongs to L^2 and we have under Assumption A26, $\mathbb{E}\left[\eta_t^3\right] = 0$ therefore:

$$\mathbb{E}\left[\boldsymbol{\mathcal{U}}_t | \mathcal{F}_{t-1}\right] = \begin{pmatrix} \left[-\frac{4\sigma_t^3 \mathbb{E}\left[\eta_t^3 - \eta_t\right]}{\omega_t^2}\frac{\partial \mu_t}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0) - \frac{2\sigma_t^2 \mathbb{E}\left[\eta_t^2 - 1\right]}{\omega_t^2}\frac{\partial g_t}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0)\right] \\ \boldsymbol{\Lambda}_t \mathbb{E}\left[\xi(\eta_t)\right] \end{pmatrix} = \mathbf{0}.$$

Under Assumption A1, the process is a strictly stationary and ergodic L^2 martingale increments. We can apply the C.L.T. of Billingsley (1961), yielding $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{\mathcal{U}}_t \xrightarrow[T\to\infty]{d} \begin{pmatrix} \boldsymbol{\mathcal{W}} \\ \boldsymbol{\mathcal{Z}} \end{pmatrix}$. We also have

$\mathcal{J}_T \xrightarrow[T\to\infty]{a.s.} \mathcal{J}$. Slutsky's Lemma gives the joint convergence in distribution:

$$\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathcal{U}_t, \mathcal{J}_T\right) \xrightarrow[T\to\infty]{d} \left(\begin{pmatrix}\mathcal{W}\\ \mathcal{Z}\end{pmatrix}, \mathcal{J}\right)$$

The continuous mapping Theorem gives $\begin{pmatrix}\frac{1}{\sqrt{T}}\frac{\partial \mathcal{L}_T}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0)\\ \sqrt{T}\left(\widehat{\boldsymbol{\rho}}_T - \boldsymbol{\rho}_0\right)\end{pmatrix} \xrightarrow[T\to\infty]{d} \begin{pmatrix}\mathcal{W}\\ \mathcal{Z}^{\mathcal{C}}\end{pmatrix}$. The components of the covariance matrix are given by:

$$\mathbb{V}\left[\frac{\partial \ell_1}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0)\right] = 4\mathbb{E}\left[P\left(\frac{\epsilon_1^2-\sigma_1^2}{\omega_1^2}\left(2\epsilon_1\frac{\partial \mu_1}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0) + \frac{\partial g_1}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0)\right)\right)\right] = \mathcal{I}(\boldsymbol{\rho}_0),$$

$$\mathbb{E}\left[\frac{\partial \ell_1}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0)\,\xi'(\eta_t)\,\boldsymbol{\Lambda}_t'\right] = -2\mathbb{E}\left[\frac{\epsilon_1^2-\sigma_1^2}{\omega_1^2}\left(2\epsilon_1\frac{\partial \mu_1}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0) + \frac{\partial g_1}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0)\right)\xi'(\eta_t)\,\boldsymbol{\Lambda}_t'\right] = \mathcal{R}(\boldsymbol{\rho}_0),$$

$$\mathbb{V}\left[\boldsymbol{\Lambda}_1\xi(\eta_1)\right] = \mathbb{E}\left[\boldsymbol{\Lambda}_1\xi(\eta_1)\,\xi'(\eta_1)\,\boldsymbol{\Lambda}_1'\right] = \mathbb{E}\left[\boldsymbol{\Lambda}_1\boldsymbol{\chi}\boldsymbol{\Lambda}_1'\right] = \mathbb{V}\left[\boldsymbol{\Lambda}_1\boldsymbol{\chi}^{1/2}\right] = \boldsymbol{\Xi}.$$

(e) **Asymptotic distribution of $\widehat{\boldsymbol{\vartheta}}_T$.** The same arguments as Theorem 3.2 part (e) lead to the result. Note that in this new context we have:

$$\mathbb{E}\left[\frac{\partial \ell_1}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}(\boldsymbol{\beta}_0)\right] = 2\mathbb{E}\left[\frac{1}{\omega_1^2}P\left(2\epsilon_1\frac{\partial \mu_t}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0) + \frac{\partial g_t}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0)\right)\right].$$

□

Proof of Proposition I.1. We have $\mathbb{P}\left[\mathbb{B}_T = \mathbb{B}\right] \le \mathbb{P}\left[\forall j \in \overline{\mathbb{B}} : \widehat{\vartheta}_{j,T} = 0\right]$. By Theorem I.2 we have $\sqrt{T}\left(\widehat{\boldsymbol{\vartheta}}_T - \boldsymbol{\theta}_0\right) \xrightarrow[T\to\infty]{d} \boldsymbol{\psi}^* := \arg\min_{\boldsymbol{\psi}\in\mathbb{R}^n} \mathcal{E}_\infty(\boldsymbol{\psi})$. Using portmanteau Theorem:

$$\limsup_{T\to\infty} \mathbb{P}\left[\forall j \in \overline{\mathbb{B}} : \sqrt{T}\widehat{\vartheta}_{j,T} = 0\right] \le \mathbb{P}\left[\forall j \in \overline{\mathbb{B}} : \boldsymbol{\psi}_j^* = 0\right].$$

Since $\boldsymbol{\psi}^*$ is the unique minimizer of the convex function $\mathcal{E}_\infty$ then $\mathbf{0} \in \frac{\partial \mathcal{E}_\infty}{\partial \boldsymbol{\psi}}(\boldsymbol{\psi}^*)$. Therefore:

$$\begin{cases} \frac{\partial \mathcal{G}_\infty}{\partial \psi_j}(\boldsymbol{\psi}^*) + \iota_{j,\infty}\,\mathrm{sign}(\theta_{j,0}) = 0 \text{ if } j \in \mathbb{B}, \\ \left|\frac{\partial \mathcal{G}_\infty}{\partial \psi_j}(\boldsymbol{\psi}^*)\right| \le \iota_{j,\infty} \text{ if } j \in \overline{\mathbb{B}}. \end{cases} \tag{43}$$

For $\boldsymbol{\psi} \in \mathbb{R}^n$, rewriting the function $\mathcal{G}_\infty$ gives:

$$\begin{aligned}\mathcal{G}_\infty(\boldsymbol{\psi}) = &\begin{pmatrix}\mathcal{W}\\ \mathcal{Z}^{\mathcal{C}}\\ \boldsymbol{\psi}\end{pmatrix}' \boldsymbol{\Pi}_{2,2\nu}'\left(\boldsymbol{\Pi}_{1,2\nu} + 2\mathbb{E}\left[\frac{1}{\omega_1^2}P\left(2\epsilon_1\frac{\partial \mu_1}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0) + \frac{\partial g_1}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0)\right)\right]\boldsymbol{\Pi}_{2,2\nu}\right)\begin{pmatrix}\mathcal{W}\\ \mathcal{Z}^{\mathcal{C}}\\ \boldsymbol{\psi}\end{pmatrix} \\ &= \mathcal{W}'\mathcal{Z}^{\mathcal{C}} + \mathcal{W}'\boldsymbol{\psi} + 2\mathcal{Z}^{\mathcal{C}\prime}\mathbb{E}\left[\frac{1}{\omega_1^2}P\left(2\epsilon_1\frac{\partial \mu_1}{\partial \boldsymbol{\rho}}(\boldsymbol{\rho}_0) + \frac{\partial g_1}{\partial \boldsymbol{\rho}}(\boldsymbol{\beta}_0)\right)\right]\mathcal{Z}^{\mathcal{C}} \\ &+ 2\boldsymbol{\psi}'\mathbb{E}\left[\frac{1}{\omega_1^2}P\left(\frac{\partial g_1}{\partial \boldsymbol{\vartheta}}(\boldsymbol{\beta}_0)\right)\right]\boldsymbol{\psi} + 4\boldsymbol{\psi}'\mathbb{E}\left[\frac{1}{\omega_1^2}\frac{\partial g_1}{\partial \boldsymbol{\vartheta}'}(\boldsymbol{\beta}_0)\left(2\epsilon_1\frac{\partial \mu_1}{\partial \boldsymbol{\rho}'}(\boldsymbol{\rho}_0) + \frac{\partial g_1}{\partial \boldsymbol{\rho}'}(\boldsymbol{\beta}_0)\right)\right]\mathcal{Z}^{\mathcal{C}}.\end{aligned}$$

The derivative with respect to $\boldsymbol{\psi}$ is:

$$\frac{\partial \mathcal{G}_\infty}{\partial \boldsymbol{\vartheta}}(\boldsymbol{\psi}) = \boldsymbol{\mathcal{W}} + 4\mathbb{E}\left[\frac{1}{\omega_1^2} P\left(\frac{\partial g_1}{\partial \boldsymbol{\vartheta}}(\boldsymbol{\beta}_0)\right)\right]\boldsymbol{\psi} + 4\mathbb{E}\left[\frac{1}{\omega_1^2}\frac{\partial g_1}{\partial \boldsymbol{\vartheta}'}(\boldsymbol{\beta}_0)\left(2\epsilon_1 \frac{\partial \mu_1}{\partial \boldsymbol{\rho}'}(\boldsymbol{\rho}_0) + \frac{\partial g_1}{\partial \boldsymbol{\rho}'}(\boldsymbol{\beta}_0)\right)\right]\boldsymbol{\mathcal{Z}}^{\mathcal{C}}. \quad (44)$$

The remaining of the proof is similar to Proposition 3.1. □

References

Adamek, R., Smeekes, S., Wilms, I., 2023. Lasso inference for high-dimensional time series. Journal of Econometrics 235, 1114–1143.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716–723.

Akaike, H., 1998. Information theory and an extension of the maximum likelihood principle, in: Parzen, E., Tanabe, K., Kitagawa, G. (Eds.), Selected Papers of Hirotugu Akaike. Springer, pp. 199–213.

Aknouche, A., Francq, C., 2023. Two-stage weighted least squares estimator of the conditional mean of observation-driven time series models. Journal of Econometrics 237, 105174.

Andrews, D.W.K., 1999. Estimation when a parameter is on a boundary. Econometrica 67, 1341–1383.

Basu, S., Michailidis, G., 2015. Regularized estimation in sparse high-dimensional time series models. The Annals of Statistics 43, 1535–1567.

Bhattacharjee, M., Chakraborty, N., Koul, H.L., 2023. Weighted l1-penalized corrected quantile regression for high-dimensional temporally dependent measurement errors. Journal of Time Series Analysis 44, 442–473.

Billingsley, P., 1961. The lindeberg-levy theorem for martingales. Proceedings of the American Mathematical Society 12, 788–792.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327.

Bunea, F., Tsybakov, A., Wegkamp, M., 2007. Sparsity oracle inequalities for the lasso. Electronic Journal of Statistics 1, 169–194.

Chan, K.S., Chen, K., 2011. Subset arma selection via the adaptive lasso. Statistics and Its Interface 4, 197–205.

Chan, N.H., Ling, S., Yau, C.Y., 2020. Lasso-based variable selection of arma models. Statistica Sinica 30, 1925–1948.

Chan, N.H., Yau, C.Y., Zhang, R.M., 2015. Lasso estimation of threshold autoregressive models. Journal of Econometrics 189, 285–296.

Davis, R.A., Knight, K., Liu, J., 1992. M-estimation for autoregressions with infinite variance. Stochastic Processes and Their Applications 40, 145–180.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. The Annals of Statistics 32, 407–499.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. Econometrica 50, 987–1007.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association 96, 1348–1360.

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

Francq, C., Thieu, 2019. Qml inference for volatility models with covariates. Econometric Theory 35, 37–72.

Francq, C., Zakoïan, J.M., 2018. Estimation risk for the var of portfolios driven by semi-parametric multivariate models. Journal of Econometrics 205, 381–401.

Francq, C., Zakoïan, J.M., 2019. GARCH Models: Structure, Statistical Inference and Financial Applications. John Wiley & Sons.

Fu, W., Knight, K., 2000. Asymptotics for lasso-type estimators. The Annals of Statistics 28, 1356–1378.

Knight, K., 1999. Epi-convergence in distribution and stochastic equi-semicontinuity. Unpublished manuscript 37, 14.

Kock, A.B., 2016. Consistent and conservative model selection with the adaptive lasso in stationary and nonstationary autoregressions. Econometric Theory 32, 243–259.

Loh, P.L., 2017. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. Annals of Statistics 45, 866–896.

Nardi, Y., Rinaldo, A., 2011. Autoregressive process modeling via the lasso procedure. Journal of Multivariate Analysis 102, 528–549.

Nielsen, H.B., Rahbek, A., 2024. Penalized quasi-likelihood estimation and model selection with parameters on the boundary of the parameter space. The Econometrics Journal 27, 107–125.

Poignard, B., Fermanian, J.D., 2021. High-dimensional penalized arch processes. Econometric Reviews 40, 86–107.

Rockafellar, R.T., 1970. Convex Analysis. Princeton University Press.

Rockafellar, R.T., Wets, R.J.B., 2009. Variational analysis. volume 317. Springer Science & Business Media.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6, 461–464.

Schweikert, K., 2022. Oracle efficient estimation of structural breaks in cointegrating regressions. Journal of Time Series Analysis 43, 83–104.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58, 267–288.

Wang, H., Li, G., Tsai, C.L., 2007. Regression coefficient and autoregressive order shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69, 63–78.

Wang, Z., Liu, H., Zhang, T., 2014. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. Annals of Statistics 42, 2164–2201.

Zhang, C.H., Huang, J., 2008. The sparsity and bias of the lasso selection in high-dimensional linear regression. The Annals of Statistics 36, 1567–1594.

Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–1429.