Penalized Convex Estimation in Dynamic Location-Scale models

ALAMI CHENTOUFI, Reda

CREST-ENSAE

December 2024

Penalized Convex Estimation in Dynamic Location-Scale Models*

Reda ALAMI CHENTOUFI[a]

[a]*Crest-ENSAE, 5, avenue Henry Le Chatelier, 91120 Palaiseau cedex, France, reda.alamichentoufi@ensae.fr*

Abstract

This paper introduces a two-step procedure for convex penalized estimation in dynamic location-scale models. The method uses a consistent, non-sparse first-step estimator to construct a convex Weighted Least Squares (WLS) optimization problem compatible with the Least Absolute Shrinkage and Selection Operator (LASSO), addressing challenges associated with non-convexity and enabling efficient, sparse estimation.

The consistency and asymptotic distribution of the estimator are established, with finite-sample performance evaluated through Monte Carlo simulations. The method's practical utility is demonstrated through an application to electricity prices in France, Belgium, the Netherlands, and Switzerland, effectively capturing seasonal patterns and external covariates while ensuring model sparsity.

Keywords: Weighted LSE; LASSO estimation; variable selection; GARCH models
JEL classification: C01, C22, C51, C52, C58

1. Introduction

Submodel selection in time series modeling becomes more difficult as the number of model parameters increases. When the parameter dimension N is small, all 2^N submodels can be evaluated using criteria such as the Akaike Information Criterion (AIC) (Akaike, 1974, 1998) or the Bayesian Information Criterion (BIC) (Schwarz, 1978). However, as N increases, exhaustive evaluation becomes computationally infeasible and risks overfitting.

Penalized estimation methods address these issues by automating model selection. Among these, the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) is widely used, combining estimation and model selection. It has been extensively studied, for example, in (Fan and Peng, 2004; Zou, 2006; Bunea et al., 2007; Zhang and Huang, 2008; Chan et al., 2015; Adamek et al., 2023; Nielsen and Rahbek, 2024), among others. Studies have been done on the sparse estimation for Autoregressive (AR) models and Autoregressive Conditional Heteroskedasticity (ARCH) models (Engle, 1982) with the Least Squares (LS) approach: Wang et al. (2007) implemented LASSO techniques to linear regression model with

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

AR errors, penalizing both regression and AR coefficients. Nardi and Rinaldo (2011) developed a LS estimator (LSE) for AR models under a double asymptotic framework, allowing the maximal AR order to grow with sample size. Kock (2016) studied Adaptive LASSO (Zou, 2006) for non-stationary AR processes and established its "oracle property". Grouped LASSO for multivariate ARCH models was explored in Poignard and Fermanian (2021). These approaches aim to estimate and select model parameters simultaneously to achieve sparsity. The AR nature of such models allows the construction of estimators from convex LS problems, simplifying the derivation of the asymptotic properties of M-estimators and facilitating numerical optimization. Still, these models often fail to adequately capture persistence effects in time series.

To account for the complexity of real-world data, models incorporating persistent components, such as Auto-Regressive Moving Average (ARMA) and Generalized ARCH (GARCH) (Bollerslev, 1986), are essential. At the same time, persistence introduces non-convexity in estimation, complicating theoretical analysis and numerical optimization. For instance, the LS loss function for ARMA(p,q) models is non-convex due to the residual structure, which combines AR terms and recursive MA terms, with the latter introducing non-convexity. Similarly, GARCH models often rely on Quasi-Maximum Likelihood (QML) estimation, which is non-convex and prone to boundary issues (Francq and Thieu, 2019). These approaches inherently involve non-convex loss functions, creating challenges such as establishing estimator asymptotic properties, addressing local optima, and increasing computational costs (Wang et al., 2014; Loh, 2017).

To mitigate these challenges, Chan and Chen (2011) proposed a two-step adaptive LASSO estimator for ARMA models. Their method uses residuals from an initial AR model as exogenous variables, effectively convexifying the LS optimization problem, achieving asymptotic normality and the "oracle property". The initial AR order is selected using information criteria. Extending this work, Chan et al. (2020) introduced a non-convex LS approach with adaptive LASSO penalty for ARMA models with a unit root. They established the asymptotic properties using piecewise arguments from Ling and McAleer (2010). To address local optima caused by non-convexity, Chan et al. (2020) proposed an iterative algorithm to identify the global optimum. Yet, this approach increases computational cost. These approaches either rely on an auxiliary AR model selected via information criteria, shifting the complexity from ARMA model selection to AR model selection by information criteria, or face the drawbacks of non-convex loss functions, necessitating computationally intensive iterative procedures.

Building on the two-step approach of Chan et al. (2020), this work introduces an estimation method for dynamic location-scale models to address these non-convexity issue. In the first step, a consistent, non-sparse estimator provides initial parameter estimates. In the second step, a penalized Weighted LS (WLS) optimization is performed, achieving sparsity while addressing non-convexity by using the first-step estimate of persistence terms. This approach

draws on Hannan and McDougall (1988), where AR model residuals were used as exogenous variables, and Aknouche and Francq (2023), who employed WLS to avoid imposing high-order moment assumptions on the Data Generating Process (DGP).

The advantages of the proposed method are threefold. First, the proposed method applies to a broad class of dynamic location-scale models, allowing penalized estimation of the scale component even when the scale process is indirectly observed through residuals. Second, the two-step approach, relying on a consistent first-step estimator, addresses non-convexity from persistent components without depending on auxiliary (information criteria selected) models to provide proxies for these components. This ensures consistency and asymptotic properties under mild conditions, as in Fu and Knight (2000), and eliminates the need for high-order moment assumptions on the DGP, making the method particularly suitable for financial applications. Third, by convexifying the optimization problem, the approach makes use of the LARS-LASSO algorithm (Efron et al., 2004) for linear location or scale specifications, avoiding iterative procedures and significantly reducing computational complexity.

The method's practical utility is demonstrated through an application to electricity day-ahead prices. The approach achieves sparsity in the Integrated AR model with GARCH innovations and exogenous covariates (IARX-GARCHX), identifies seasonal effects known in electricity markets (Liu and Shi, 2013), and validates the relevance of realized measures highlighted by Frömmel et al. (2014). Stability analysis reflects the dynamic behavior of electricity prices, consistent with observations by Janczura and Weron (2012) and Samitas and Armenatzoglou (2014).

The paper begins with a presentation of the model, notations, and estimator in Section 2. Section 3 introduces the assumptions and main theorems, proving the strong consistency of the estimator and deriving its asymptotic distribution in the style of Fu and Knight (2000) under mild conditions. Section 4 presents Monte Carlo experiments, followed by an application to electricity prices in France, Belgium, the Netherlands, and Switzerland in Section 5. Detailed proofs are provided in Appendix I.

2. Convexification with a Two-Step Procedure

This section introduces the mathematical tools and outlines a two-step procedure for convexifying parameter estimation in dynamic location-scale models. The approach is illustrated with examples of specific models.

2.1. Two-Step Estimator

The reference model is defined as follows:

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

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

where $\{y_t, t \in \mathbb{Z}\}$ and $\{\epsilon_t, t \in \mathbb{Z}\}$ denote real-valued processes, and $\{\eta_t, t \in \mathbb{Z}\}$ is an independent and identically distributed (i.i.d.) innovation process with zero mean and unit variance. Let $\{\boldsymbol{Y}_t, t \in \mathbb{Z}\}$ and $\{\boldsymbol{X}_t, t \in \mathbb{Z}\}$ represent two vector-valued exogenous processes. The location μ_t and the scale σ_t are assumed to follow the parametric forms:

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

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

where $\boldsymbol{\phi}_0$ and $\boldsymbol{\theta}_0$ are respectively ν-dimensional and n-dimensional parameter vectors that describe the model. Each belongs to a corresponding compact and convex set: $\Phi \subset \mathbb{R}^\nu$ and $\Theta \subset \mathbb{R}^n$. The functions m and h are measurable mappings, where $m : \mathbb{R}^\infty \times \Phi \mapsto \mathbb{R}$ and $h : \mathbb{R}^\infty \times \Theta \mapsto \mathbb{R}$. To build the two-step optimization problem based on the parameteric forms (3)-(4), the following recursive functions are assumed to exist:

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

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

$$(\boldsymbol{\phi}, \boldsymbol{\theta}) \mapsto \sigma_t^2(\boldsymbol{\phi}, \boldsymbol{\theta}) = h\left(\epsilon_{t-1}(\boldsymbol{\phi}), \ldots, \sigma_{t-1}^2(\boldsymbol{\phi}, \boldsymbol{\theta}), \ldots, \boldsymbol{X}_{t-1}, \ldots; \boldsymbol{\theta}\right). \tag{7}$$

The functions (5), (6) and (7) are referred to as the "recursive mean", "recursive error", and "recursive variance", respectively. Since ϵ_t is not directly observed, the recursive variance function depends on two parameters, $(\boldsymbol{\phi}, \boldsymbol{\theta})$: the parameter $\boldsymbol{\phi}$ is used to compute the recursive errors, providing a proxy for ϵ_t in the parametric form (4).

Consider the simple case of an ARMA(1,1) for which $\epsilon_t(\boldsymbol{\phi}_0) = y_t - a_0 y_{t-1} - b_0 \epsilon_{t-1}(\boldsymbol{\phi}_0)$ with $\boldsymbol{\phi}_0 = (a_0, b_0)'$. The function $\epsilon_t^2(\cdot)$ is non-convex, and thus the LSE of $\boldsymbol{\phi}_0$ is a non-convex optimization problem:

$$\boldsymbol{\phi} \mapsto \sum_{t=1}^T \epsilon_t^2(\boldsymbol{\phi}). \tag{8}$$

Now, if a first-step estimator $\widehat{\boldsymbol{\phi}}^{(1)}$ is available, one can define a second-step penalized LSE of $\boldsymbol{\phi}_0$ by minimizing the convex objective function:

$$\boldsymbol{\phi} \mapsto \sum_{t=1}^T \left(y_t - a y_{t-1} - b \epsilon_{t-1}\left(\widehat{\boldsymbol{\phi}}^{(1)}\right)\right)^2 + p_T(\boldsymbol{\phi}),$$

where $\boldsymbol{\phi} = (a, b)'$ and $p_T(\cdot)$ is a LASSO-type penalty term. Applying this idea in a general context, a "two-step mean" function and a "two-step variance" function are introduced and defined as follows:

$$\boldsymbol{\phi}, \boldsymbol{\upsilon} \in \Phi, \quad f_t(\boldsymbol{\phi}, \boldsymbol{\upsilon}) = m\left(\epsilon_{t-1}(\boldsymbol{\phi}), \ldots, y_{t-1}, \ldots, \boldsymbol{Y}_{t-1}, \ldots; \boldsymbol{\upsilon}\right), \tag{9}$$

$$\left(\boldsymbol{\phi}', \boldsymbol{\theta}', \boldsymbol{\psi}'\right)' \in \Phi \times \Theta \times \Theta, \quad g_t(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\psi}) = h\left(\epsilon_{t-1}(\boldsymbol{\phi}), \ldots, \sigma_{t-1}^2(\boldsymbol{\phi}, \boldsymbol{\theta}), \ldots, \boldsymbol{X}_{t-1}, \ldots; \boldsymbol{\psi}\right). \tag{10}$$

A more intuitive understanding of these two-step functions is provided in the next section through illustrative examples. The following WLS loss functions are defined to target the estimation of specific components of the model:

$$L_T(\boldsymbol{\phi},\boldsymbol{\upsilon}) = \sum_{t=1}^{T} l_t(\boldsymbol{\phi},\boldsymbol{\upsilon}) \quad \text{with} \quad l_t(\boldsymbol{\phi},\boldsymbol{\upsilon}) = \left(\frac{y_t - f_t(\boldsymbol{\phi},\boldsymbol{\upsilon})}{w_t}\right)^2, \tag{11}$$

$$L_T^*(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}) = \sum_{t=1}^{T} l_t^*(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}) \quad \text{with} \quad l_t^*(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}) = \left(\frac{\epsilon_t^2(\boldsymbol{\phi}) - g_t(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi})}{w_t^*}\right)^2, \tag{12}$$

where the weight processes w_t and w_t^*, introduced to control the moments required for the procedure to converge, are defined as two measurable functions of the past observations, mapping $\mathbb{R}^\infty$ to $[\underline{\omega},\infty)$ and $[\underline{\omega}^*,\infty)$, respectively, such that:

$$w_t = w(y_{t-1},\ldots,\boldsymbol{Y}_{t-1},\ldots,\boldsymbol{X}_{t-1},\ldots) \geqslant \underline{\omega} > 0,$$
$$w_t^* = w^*(y_{t-1},\ldots,\boldsymbol{Y}_{t-1},\ldots,\boldsymbol{X}_{t-1},\ldots) \geqslant \underline{\omega}^* > 0.$$

Adding LASSO penalties to (11)-(12) gives the penalized loss functions:

$$Q_T(\boldsymbol{\phi},\boldsymbol{\upsilon}) = \frac{L_T(\boldsymbol{\phi},\boldsymbol{\upsilon}) + p_T(\boldsymbol{\upsilon})}{T} \quad \text{with} \quad p_T(\boldsymbol{\upsilon}) = \sum_{j\in\mathcal{S}} \lambda_{T,j} |\upsilon_j|, \tag{13}$$

$$Q_T^*(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}) = \frac{L_T^*(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}) + p_T^*(\boldsymbol{\psi})}{T} \quad \text{with} \quad p_T^*(\boldsymbol{\psi}) = \sum_{j\in\mathcal{S}^*} \lambda_{T,j}^* |\psi_j|, \tag{14}$$

where $\mathcal{S} \subset \{1,\ldots,\nu\}$ and $\mathcal{S}^* \subset \{1,\ldots,n\}$ denote the sets of indices of the vectors $\boldsymbol{\upsilon}$ and $\boldsymbol{\psi}$ (respectively) to penalize. The sequences $(\boldsymbol{\lambda}_T)_{T\in\mathbb{N}} = \left((\lambda_{T,i})_{i\in\mathcal{S}}\right)_{T\in\mathbb{N}}$ and $(\boldsymbol{\lambda}_T^*)_{T\in\mathbb{N}} = \left(\left(\lambda_{T,i}^*\right)_{i\in\mathcal{S}^*}\right)_{T\in\mathbb{N}}$ are vectors with non-negative, deterministic components.

These loss functions assume that the process is fully observed, meaning all its infinite past values are known. In practice, however, only a finite sample $\{\boldsymbol{O}_t, t\in\{1,\ldots,T\}\}$ is available, where $\boldsymbol{O}_t = (y_t, \boldsymbol{Y}_t', \boldsymbol{X}_t')'$. Consequently, the previously defined functions can only be computed with truncation, as observations are unavailable for the distant past. Let $\widetilde{\boldsymbol{O}}_0, \widetilde{\boldsymbol{O}}_{-1}, \ldots$ be initial values used to replace $\boldsymbol{O}_0, \boldsymbol{O}_{-1}, \ldots$. The truncated versions of μ_t, ϵ_t, σ_t^2, g_t, f_t, w_t, and w_t^*, denoted as $\widetilde{\mu}_t$, $\widetilde{\epsilon}_t$, $\widetilde{\sigma}_t^2$, $\widetilde{g}_t$, $\widetilde{f}_t$, $\widetilde{w}_t$, and $\widetilde{w}_t^*$, are obtained by substituting $\boldsymbol{O}_0, \boldsymbol{O}_{-1}, \ldots$ with the initial values in equations (5)-(14). Similarly, $\widetilde{Q}_T$, $\widetilde{L}_T$, $\widetilde{l}_t$, $\widetilde{Q}_T^*$, $\widetilde{L}_T^*$ and $\widetilde{l}_t^*$ are defined by replacing their respective functions with these truncated forms. Under mild assumptions, it is shown that these initial values become asymptotically irrelevant in the estimation framework.

The convexified WLS problem can now be formulated. First, note that for a wide range of model specifications, the functions $\widetilde{L}_T(\boldsymbol{\phi},\cdot)$ and $\widetilde{L}_T^*(\boldsymbol{\phi},\boldsymbol{\theta},\cdot)$ are convex. Given a strongly consistent estimator $\left(\widehat{\boldsymbol{\phi}}_T^{(1)}, \widehat{\boldsymbol{\theta}}_T^{(1)}\right)$ of $(\boldsymbol{\phi}_0,\boldsymbol{\theta}_0)$ obtained from the observed sample, the functions $\widetilde{L}_T\left(\widehat{\boldsymbol{\phi}}_T^{(1)},\cdot\right)$ and $\widetilde{L}_T^*\left(\widehat{\boldsymbol{\phi}}_T^{(1)}, \widehat{\boldsymbol{\theta}}_T^{(1)},\cdot\right)$ also remain convex. This leads to the following two-step convex estimation procedure:

Algorithm 1 Two-Step Estimation Procedure

Step 1: Compute $\left(\widehat{\boldsymbol{\phi}}_T^{(1)}, \widehat{\boldsymbol{\theta}}_T^{(1)}\right)$.

Step 2-1: Minimize $\boldsymbol{\upsilon} \mapsto \widetilde{Q}_T\left(\widehat{\boldsymbol{\phi}}_T^{(1)}, \boldsymbol{\upsilon}\right)$, yielding $\widehat{\boldsymbol{\phi}}_T$.

Step 2-2: Minimize $\boldsymbol{\psi} \mapsto \widetilde{Q}_T^*\left(\widehat{\boldsymbol{\phi}}_T^{(1)}, \widehat{\boldsymbol{\theta}}_T^{(1)}, \boldsymbol{\psi}\right)$, yielding $\widehat{\boldsymbol{\theta}}_T$.

The objective of this study is to establish the asymptotic properties of the resulting estimators.

2.2. Examples for Specific Models

This section gives examples to explain the two-step procedure. It shows how to estimate the location parameter using (13) and why the scale parameter needs the more general loss function (14).

2.2.1. ARMAX(1,1) Model

Suppose the DGP follows the ARMAX model:

$$y_t = \left(y_{t-1}, \epsilon_{t-1}, \boldsymbol{Y}'_{t-1}\right) \boldsymbol{\phi}_0 + \epsilon_t.$$

Assume a first step estimator $\widehat{\boldsymbol{\phi}}_T^{(1)}$ of $\boldsymbol{\phi}_0$, such as the LSE, is available. Using this estimator, the residuals are:

$$\widetilde{\epsilon}_0\left(\widehat{\boldsymbol{\phi}}_T^{(1)}\right) = \widetilde{\epsilon}_0 \quad \text{and} \quad \widetilde{\epsilon}_t\left(\widehat{\boldsymbol{\phi}}_T^{(1)}\right) = y_t - \left(y_{t-1}, \widetilde{\epsilon}_{t-1}\left(\widehat{\boldsymbol{\phi}}_T^{(1)}\right), \boldsymbol{Y}'_{t-1}\right) \widehat{\boldsymbol{\phi}}_T^{(1)}. \tag{15}$$

Using the two-step mean function, the following expression is obtained:

$$\widetilde{f}_t\left(\widehat{\boldsymbol{\phi}}_T^{(1)}, \boldsymbol{\upsilon}\right) = \left(y_{t-1}, \widetilde{\epsilon}_{t-1}\left(\widehat{\boldsymbol{\phi}}_T^{(1)}\right), \boldsymbol{Y}'_{t-1}\right) \boldsymbol{\upsilon}.$$

The two-step mean function is linear in $\boldsymbol{\upsilon}$, which ensures that the minimization problem in **Step 2-1** is convex. This convexity arises because the residuals, computed from $\widehat{\boldsymbol{\phi}}_T^{(1)}$, are treated as exogenous in the optimization problem.

2.2.2. GARCHX(1,1) Model

Suppose the DGP follows the GARCHX(1,1) model:

$$\begin{aligned} \sigma_t^2 &= \left(1, \epsilon_{t-1}^2, \sigma_{t-1}^2, \boldsymbol{X}'_{t-1}\right) \boldsymbol{\theta}_0 > 0, \\ \epsilon_t &= \sigma_t \eta_t, \end{aligned}$$

where the parameter $\boldsymbol{\theta}_0$ and the exogenous process $\{\boldsymbol{X}_t, t \in \mathbb{Z}\}$ are non-negative to ensure that the conditional variance remains positive with probability one. The squared process $\left\{\epsilon_t^2, t \in \mathbb{Z}\right\}$ can be expressed as:

$$\epsilon_t^2 = \left(1, \epsilon_{t-1}^2, \sigma_{t-1}^2, \boldsymbol{X}_{t-1}'\right) \boldsymbol{\theta}_0 + \sigma_t^2 \left(\eta_t^2 - 1\right).$$

This shows that $\left\{\epsilon_t^2, t \in \mathbb{Z}\right\}$ behaves as a location model, where the term σ_t^2 acts as a location component. With appropriate assumptions on the moments of η_t^2, this problem can be adapted to **Step 2-1**. A natural choice for the first-step estimator $\widehat{\boldsymbol{\theta}}_T^{(1)}$ of $\boldsymbol{\theta}_0$ is the QMLE.

2.2.3. ARMAX(1,1)-GARCHX(1,1) Model

Suppose the DGP follows the ARMAX(1,1)-GARCHX(1,1) model:

$$\begin{aligned} y_t &= \left(y_{t-1}, \epsilon_{t-1}, \boldsymbol{Y}_{t-1}'\right) \boldsymbol{\phi}_0 + \epsilon_t, \\ \sigma_t^2 &= \left(1, \epsilon_{t-1}^2, \sigma_{t-1}^2, \boldsymbol{X}_{t-1}'\right) \boldsymbol{\theta}_0 > 0, \\ \epsilon_t &= \sigma_t \eta_t, \end{aligned}$$

The penalized estimator $\widehat{\boldsymbol{\phi}}_T$ of $\boldsymbol{\phi}_0$ remains the same as in the ARMAX(1,1) case. However, the penalized estimation of $\boldsymbol{\theta}_0$ requires a different approach. After computing the first-step estimator $\left(\widehat{\boldsymbol{\phi}}_T^{(1)}, \widehat{\boldsymbol{\theta}}_T^{(1)}\right)$ of $(\boldsymbol{\phi}_0, \boldsymbol{\theta}_0)$, only the residuals (15) are observed rather than the true values of the underlying GARCHX(1,1) process. As a result, the loss function (13) is no longer suitable, as it must account for the approximation $\widetilde{\epsilon}_t(\cdot)$. To address this, the loss function (14) is formulated as an adaptation of (13). This formulation leads to the minimization problem defined in **Step 2-2**, enabling estimation of $\boldsymbol{\theta}_0$ while accounting for the residual-based approximation.

All the previously discussed examples are compatible with the LARS-LASSO algorithm, which efficiently computes the full LASSO path. The best sub-model is selected in a final step using criteria such as AIC or BIC. This method reduces computational complexity and achieves significant efficiency gains.

3. Theoretical Results

This section establishes the strong consistency and derives the asymptotic distributions of the two-step estimators. Let $\mathcal{F}_t$ denote the sigma-field generated by $\{\boldsymbol{O}_u, u \leqslant t\}$, and consider the following assumption.

A0 The process $\left\{(y_t, \epsilon_t)', t \in \mathbb{Z}\right\}$ is a solution to (1)-(4) and $\left\{(y_t, \epsilon_t, \boldsymbol{Y}_t', \boldsymbol{X}_t')', t \in \mathbb{Z}\right\}$ is strictly stationary, ergodic, and non-anticipative with respect to the filtration $\mathcal{F}$, with η_t independent of $\mathcal{F}_{t-1}$.

This assumption is maintained throughout. The discussion begins with the results for location parameter estimator, followed by those of the scale parameter estimator.

3.1. Location Parameter Estimator

The consistency of $\widehat{\boldsymbol{\phi}}_T$ is established under the following assumptions.

A1 $\left(1 + |y_t| + \sup_{\Phi\times\Phi} |f_t|\right) a_t \xrightarrow[t\to\infty]{\text{a.s.}} 0$, with $a_t = \sup_{\Phi\times\Phi} \left|\widetilde{f}_t - f_t\right|$.

A2 $\left(1 + y_t^2 + \sup_{\Phi\times\Phi} f_t^2\right) d_t \xrightarrow[t\to\infty]{\text{a.s.}} 0$, with $d_t = \left|\widetilde{w}_t^2 - w_t^2\right|$.

A3 $\mathbb{E}\left[\left(\frac{\sigma_t}{w_t}\right)^2\right] < \infty$.

In the following, the gradient and Hessian with respect to $(x_1, \ldots, x_k)$ in $\mathbb{R}^k, k \in \mathbb{N}$, are denoted by $\nabla_{x_1,\ldots,x_k}$ and $\nabla^2_{x_1,\ldots,x_k}$, respectively. When taken with respect to all parameters of a function, they are written as ∇ and ∇^2. The norm $||\cdot||$ refers to a deterministic vector or matrix norm, depending on the context.

A4 There exists a neighborhood $\mathcal{V}(\boldsymbol{\phi}_0) \subset \Phi$ of $\boldsymbol{\phi}_0$ such that f_t is a.s. of class C^1 on $\mathcal{V}(\boldsymbol{\phi}_0) \times \overset{\circ}{\Phi}$ and $\mathbb{E}\left[\left(\frac{M_t}{w_t}\right)^2\right] < \infty$ with $M_t = \sup_{\mathcal{V}(\boldsymbol{\phi}_0)\times\overset{\circ}{\Phi}} ||\nabla f_t||$ where $\overset{\circ}{\Phi}$ stands for the interior of Φ.

A5 $(\boldsymbol{\phi}_{0,1}\mathbb{I}_{1\notin\mathcal{S}}, \ldots, \boldsymbol{\phi}_{0,\nu}\mathbb{I}_{\nu\notin\mathcal{S}})' \in \overset{\circ}{\Phi}$, with $\mathbb{I}$ denoting the indicator function.

A6 $\widehat{\boldsymbol{\phi}}_T^{(1)} \xrightarrow[T\to\infty]{\text{a.s.}} \boldsymbol{\phi}_0$.

A7 $\forall \boldsymbol{\phi} \in \mathcal{V}(\boldsymbol{\phi}_0)$, the functions $l_t(\boldsymbol{\phi}, \cdot)$ and $\widetilde{l}_t(\boldsymbol{\phi}, \cdot)$ are a.s. strictly convex on Φ.

Assumptions A1 and A2 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 the weight process $\{w_t, t \in \mathbb{Z}\}$ is guided by the Assumption A3, which avoids the need for high-order moments of the DGP. Assumption A4 requires that the true parameter $\boldsymbol{\phi}_0$ lie in the interior of the parameter space Φ and Assumption A5 is sufficient to avoid boundary issues. Since the L^1 penalty induces a shrinkage effect on the estimator, there is a non-zero probability that it lies on the segment between $\mathbf{0}$ and $\boldsymbol{\phi}_0$. Φ must be defined so that $\boldsymbol{\phi}_0 \in \overset{\circ}{\Phi}$ and for each penalized component $i \in \mathcal{S}$, the corresponding dimension in Φ includes 0 in its interior. This is not restrictive since the estimator is based on WLS, it suffices to choose Φ large enough. Lastly, Assumption A7 is generally straightforward; for instance, in models like ARMAX, GARCHX, or GJR-GARCHX, the loss function $\boldsymbol{\upsilon} \mapsto l_t(\boldsymbol{\phi}, \boldsymbol{\upsilon})$ is convex.

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 $\epsilon_t^2 = \sigma_t^2 + \sigma_t^2\left(\eta_t^2 - 1\right)$. It is a location process with a weak white noise as error term. To apply the results of this section in the context of scale models, an additional moment assumption on η_t is required: $\left(\eta_t^2 - 1\right)$ must have a finite second order moment.

Under the previous assumptions, the following consistency result can be stated.

Theorem 3.1. *Assume that* $\frac{1}{T}\boldsymbol{\lambda}_T \xrightarrow[T\to\infty]{} \boldsymbol{\iota}_\infty < +\infty$ *component-wise. Then, under Assumptions A0, A3-A4, the function* $Q_\infty(\boldsymbol{\phi},\boldsymbol{\upsilon}) := \mathbb{E}[l_1(\boldsymbol{\phi},\boldsymbol{\upsilon})] + \sum_{j\in\mathcal{S}} \iota_{\infty,j}|\upsilon_j|$ *exists. Adding Assumptions A1-A2 and A5-A7, the following holds:*

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

If $\boldsymbol{\iota}_\infty = \mathbf{0}$*, then* $\widehat{\boldsymbol{\phi}}_T \xrightarrow[T\to\infty]{a.s.} \boldsymbol{\phi}_0$.

Theorem 3.1 demonstrates that the estimator converges to a biased limit when $\boldsymbol{\iota}_\infty$ is non-zero. To derive the asymptotic distribution, $\frac{1}{T}\boldsymbol{\lambda}_T$ must converge to $\mathbf{0}$ at an appropriate rate. The asymptotic distribution of the estimator is established under the following assumptions.

A8 $\mathbb{E}\left[\left(\frac{\sigma_t}{w_t}\right)^4\right] < \infty$.

A9 The function f_t is a.s. of class C^2 on $\mathcal{V}(\boldsymbol{\phi}_0)\times\mathcal{V}(\boldsymbol{\phi}_0)$ and $\mathbb{E}\left[\left(\frac{\mathcal{M}_t}{w_t}\right)^4 + \left(\frac{\mathcal{K}_t}{w_t}\right)^2\right] < \infty$, with

$$\mathcal{M}_t = \sup_{\mathcal{V}(\boldsymbol{\phi}_0)\times\mathcal{V}(\boldsymbol{\phi}_0)} ||\nabla f_t(\boldsymbol{\phi},\boldsymbol{\upsilon})||,$$
$$\mathcal{K}_t = \sup_{\mathcal{V}(\boldsymbol{\phi}_0)\times\mathcal{V}(\boldsymbol{\phi}_0)} \left|\left|\nabla^2 f_t(\boldsymbol{\phi},\boldsymbol{\upsilon})\right|\right|.$$

Throughout, for a symmetric positive definite matrix $\boldsymbol{\mathcal{J}}$, the norm $\boldsymbol{x}\mapsto\sqrt{\boldsymbol{x}'\boldsymbol{\mathcal{J}}\boldsymbol{x}}$ is denoted by $||\cdot||_{\boldsymbol{\mathcal{J}}}$.

A10 There is a closed and convex subset $\mathcal{C}$ of $\mathbb{R}^\nu$ and a sequence of symmetric positive definite $\nu\times\nu$ matrices $(\boldsymbol{\mathcal{J}}_T)_{T\in\mathbb{N}}$ converging a.s. to a symmetric positive definite matrix $\boldsymbol{\mathcal{J}}$ such that:

$$\sqrt{T}\left(\widehat{\boldsymbol{\phi}}_T^{(1)} - \boldsymbol{\phi}_0\right) = \arg\min_{\boldsymbol{\xi}\in\mathcal{C}} ||\boldsymbol{Z}_T - \boldsymbol{\xi}||_{\boldsymbol{\mathcal{J}}_T} + o_{\mathbb{P}}(1) \quad \text{with} \quad \boldsymbol{Z}_T = \frac{1}{\sqrt{T}}\sum_{t=1}^T \boldsymbol{\Delta}_t\gamma(\eta_t),$$

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 $\mathbb{V}[\gamma(\eta_t)] =: \boldsymbol{\Gamma}$.

A11 Letting $b_t = \sup_{\Phi\times\Phi}\left|\left|\nabla f_t - \nabla\widetilde{f}_t\right|\right|$, the sequences:

$$d_t \sup||\nabla f_t||\left(1 + |y_t| + \sup_{\Phi\times\Phi}|f_t|\right),$$
$$a_t \sup\left|\left|\nabla\widetilde{f}_t\right|\right|,$$
$$b_t\left(1 + |y_t| + \sup_{\Phi\times\Phi}|f_t|\right),$$

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

Assumption A10, based on Francq and Zakoïan (2018), addresses cases where the true parameter $\boldsymbol{\phi}_0$ lies on the boundary of the domain for the first-stage estimator $\widehat{\boldsymbol{\phi}}_T^{(1)}$. 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. 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. Assumption A11 ensures that initial values are asymptotically irrelevant when deriving the estimator's asymptotic distribution. The remaining assumptions extend those in Theorem 3.1.

The asymptotic behavior of $\widehat{\boldsymbol{\phi}}_T$ is derived by studying the asymptotic properties of the arg min of the following function (and its truncated version). This function, inspired by Fu and Knight (2000), is defined for $\boldsymbol{\varphi}$ in $\mathbb{R}^{\nu}$ by:

$$\Lambda_T(\boldsymbol{\varphi}) = \mathcal{Y}_T(\boldsymbol{\varphi}) + p_T\left(\frac{\boldsymbol{\varphi}}{\sqrt{T}} + \boldsymbol{\phi}_0\right) - p_T(\boldsymbol{\phi}_0) \text{ where } \mathcal{Y}_T(\boldsymbol{\varphi}) = L_T\left(\boldsymbol{\phi}_T^{(1)}, \frac{\boldsymbol{\varphi}}{\sqrt{T}} + \boldsymbol{\phi}_0\right) - L_T(\boldsymbol{\phi}_0, \boldsymbol{\phi}_0).$$

It is straightforward to observe that $\sqrt{T}\left(\widehat{\boldsymbol{\phi}}_T - \boldsymbol{\phi}_0\right)$ is the arg min of $\widetilde{\Lambda}_T$, the truncated version of Λ_T. The analysis proceeds by expanding L_T on a neighborhood of $(\boldsymbol{\phi}_0, \boldsymbol{\phi}_0)$. This expansion is valid under Assumption B9 and since $\frac{1}{T}\boldsymbol{\lambda}_T \xrightarrow[T\to\infty]{} \mathbf{0}$, which means that $\widehat{\boldsymbol{\phi}}_T \xrightarrow[T\to\infty]{\text{a.s.}} \boldsymbol{\phi}_0$ according to Theorem 3.1. With the rate of convergence of $\frac{1}{T}\boldsymbol{\lambda}_T$ discussed later. The expansion yields a function that is a continuous transformation of a Gaussian vector, leading to the asymptotic distribution result.

Before stating the next result, define $P(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{x}'$, for a matrix or column vector $\boldsymbol{x}$. Under the previous assumptions, the following quantities are well-defined:

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

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

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

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

The boundary issues of the initial estimator $\widehat{\boldsymbol{\phi}}_T^{(1)}$, as described in Assumption A10, lead to the following projection of the Gaussian vector:

$$\boldsymbol{\mathcal{W}} = \begin{pmatrix} \boldsymbol{W} \\ \arg\min_{\boldsymbol{\xi}\in\mathcal{C}} \|\boldsymbol{Z} - \boldsymbol{\xi}\|_{\mathcal{J}} \end{pmatrix}.$$

For further details, see Lemma I.6. Lemma I.7 establishes that this projected vector appears in the quadratic form characterizing the asymptotic distribution of $\mathcal{Y}_T$, specifically:

$$\mathcal{Y}_\infty(\boldsymbol{\varphi}) := \begin{pmatrix} \boldsymbol{\mathcal{W}} \\ \boldsymbol{\varphi} \end{pmatrix}' M(\boldsymbol{\phi}_0) \begin{pmatrix} \boldsymbol{\mathcal{W}} \\ \boldsymbol{\varphi} \end{pmatrix} \text{ with } M(\boldsymbol{\phi}_0) = \boldsymbol{D}_2' \left[\boldsymbol{D}_1 + \frac{1}{2} J(\boldsymbol{\phi}_0) \boldsymbol{D}_2 \right],$$

where $\boldsymbol{D}_1 = \begin{pmatrix} \boldsymbol{I}_{2\nu} & \boldsymbol{0}_{2\nu} \end{pmatrix}$, $\boldsymbol{D}_2 = \begin{pmatrix} \boldsymbol{0}_{2\nu} & \boldsymbol{I}_{2\nu} \end{pmatrix}$, and $\boldsymbol{I}_{2\nu}$ denotes the $2\nu \times 2\nu$ identity matrix, while $\boldsymbol{0}_{2\nu}$ denotes the $2\nu \times 2\nu$ zero matrix and $J(\boldsymbol{\phi}_0) = 2\mathbb{E}\left[P\left(\frac{\nabla f_1(\boldsymbol{\phi}_0, \boldsymbol{\phi}_0)}{w_1} \right) \right]$ exists as in Lemma I.5.

Theorem 3.2. *Assume that* $\frac{1}{\sqrt{T}} \boldsymbol{\lambda}_T \xrightarrow[T\to\infty]{} \boldsymbol{\iota}_\infty^* < +\infty$ *component-wise. Then, under Assumptions A0-A2, A6-A11, the following holds:*

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

with $\Lambda_\infty(\boldsymbol{\varphi}) = \mathcal{Y}_\infty(\boldsymbol{\varphi}) + \sum_{j\in\mathcal{S}} \iota_{\infty,j} \left[\varphi_j sign(\phi_{0,j}) \mathbb{I}_{\phi_{0,j}\neq 0} + |\varphi_j| \mathbb{I}_{\phi_{0,j}=0} \right]$.

Theorem 3.2 demonstrates that if $\frac{1}{T}\boldsymbol{\lambda}_T$ converges to zero at an appropriate rate, the limiting distribution remains influenced by the LASSO shrinkage effect, even when the estimator is asymptotically unbiased. However, if $\frac{1}{\sqrt{T}}\boldsymbol{\lambda}_T \xrightarrow[T\to\infty]{} \boldsymbol{0}$, the asymptotic distribution coincides with that of a WLSE without penalty.

In this section, it was shown that $\widehat{\boldsymbol{\phi}}_T$ is strongly consistent and its asymptotic distribution was derived under standard assumptions, assuming direct observation of process realizations without proxies. The next section focuses on the scale part, where the scale parameter is estimated from the residuals computed from the first-step estimator.

3.2. Scale Parameter Estimator

In this section, it is assumed that a first-stage estimator $\left(\widehat{\boldsymbol{\phi}}_T^{(1)}, \widehat{\boldsymbol{\theta}}_T^{(1)} \right)$ is available. The parameters of the location component can be estimated separately, as previously described, without accounting for the scale component. However, estimating the scale component requires certain adaptations. Specifically, in the location-scale case, the process $\{\epsilon_t, t \in \mathbb{Z}\}$ is not directly observed; only residuals derived from the first-step estimation (via $\widehat{\boldsymbol{\phi}}_T^{(1)}$) are available. Thus, assumptions on the functions describing these residuals must be introduced.

The consistency of $\widehat{\boldsymbol{\theta}}_T$ is established under the following assumptions.

B1 $a_t^* \left[1 + y_t^2 + \sup_\Phi \mu_t^2 + \sup_{\Phi\times\Theta\times\Theta} |g_t| \right] \xrightarrow[t\to\infty]{\text{a.s.}} 0$, with
$a_t^* = \sup |\mu_t(\boldsymbol{\phi}) - \widetilde{\mu}_t(\boldsymbol{\phi})| \left(1 + \sup |\mu_t(\boldsymbol{\phi})| \right) + \sup |\widetilde{g}_t - g_t|$.

B2 $d_t^* \left[1 + y_t^4 + \sup_\Theta \mu_t^4 + \sup_{\Phi\times\Theta\times\Theta} g_t^2 \right] \xrightarrow[t\to\infty]{\text{a.s.}} 0$, with $d_t^* = \left| \widetilde{w}_t^{*^2} - w_t^{*^2} \right|$.

B3 $\mathbb{E}\left[\left(\frac{\sigma_t^2}{w_t^*} \right)^2 \right] < \infty$.

B4 There exists two neighborhoods $\mathcal{V}(\boldsymbol{\phi}_0) \subset \Phi$ of $\boldsymbol{\phi}_0$ and $\mathcal{V}(\boldsymbol{\theta}_0) \subset \Theta$ of $\boldsymbol{\theta}_0$ such that g_t and $\mu_t(\cdot)$ are a.s. of class C^1 on $\mathcal{V}(\boldsymbol{\phi}_0) \times \mathcal{V}(\boldsymbol{\theta}_0) \times \overset{\circ}{\Theta}$ and $\mathcal{V}(\boldsymbol{\phi}_0)$, respectively, and $\mathbb{E}\left[\left(\frac{G_t}{w_t^*}\right)^2 + \left(\frac{E_t^2}{w_t^*}\right)^2\right] < \infty$, with

$$G_t = \sup_{\mathcal{V}(\boldsymbol{\phi}_0)\times\mathcal{V}(\boldsymbol{\theta}_0)\times\overset{\circ}{\Theta}} ||\nabla g_t||,$$
$$E_t = \sup_{\mathcal{V}(\boldsymbol{\phi}_0)} \left|\left|\frac{\partial \mu_t}{\partial \boldsymbol{\phi}}(\boldsymbol{\phi})\right|\right|.$$

B5 $\left(\boldsymbol{\theta}_{0,1}\mathbb{I}_{1\notin\mathcal{S}^*}, \ldots, \boldsymbol{\theta}_{0,\nu}\mathbb{I}_{\nu\notin\mathcal{S}^*}\right)' \in \overset{\circ}{\Theta}$.

B6 $\left(\widehat{\boldsymbol{\phi}}_T^{(1)}, \widehat{\boldsymbol{\theta}}_T^{(1)}\right) \xrightarrow[T\to\infty]{\text{a.s.}} (\boldsymbol{\phi}_0, \boldsymbol{\theta}_0)$.

B7 $\forall (\boldsymbol{\phi}, \boldsymbol{\theta}) \in \mathcal{V}(\boldsymbol{\phi}_0) \times \mathcal{V}(\boldsymbol{\theta}_0)$, the functions $l_t^*(\boldsymbol{\phi}, \boldsymbol{\theta}, \cdot)$ and $\widetilde{l}_t^*(\boldsymbol{\phi}, \boldsymbol{\theta}, \cdot)$ are strictly convex on Θ.

The equivalence between the assumptions in the previous sections and those stated here is straightforward.

Theorem 3.3. *Assume that* $\frac{1}{T}\boldsymbol{\lambda}_T^* \xrightarrow[T\to\infty]{} \boldsymbol{\iota}_\infty^* < +\infty$ *component-wise. Then, under Assumptions A0, B3-B4, the function* $Q_\infty^*(\boldsymbol{\phi}_0, \boldsymbol{\theta}_0, \boldsymbol{\psi}) := \mathbb{E}\left[l_1^*(\boldsymbol{\phi}_0, \boldsymbol{\theta}_0, \boldsymbol{\psi})\right] + \sum_{j\in\mathcal{S}^*} \iota_{j,\infty}^* |\psi_j|$ *exists. Adding B1-B2, B5-B7, the following holds:*

$$\widehat{\boldsymbol{\theta}}_T \xrightarrow[T\to\infty]{a.s.} \arg\min_{\boldsymbol{\psi}\in\Theta} Q_\infty^*(\boldsymbol{\phi}_0, \boldsymbol{\theta}_0, \boldsymbol{\psi}).$$

If $\boldsymbol{\iota}_\infty^* = \mathbf{0}$, *then* $\widehat{\boldsymbol{\theta}}_T \xrightarrow[T\to\infty]{a.s.} \boldsymbol{\theta}_0$.

Remark 3.2. When $\frac{1}{T}\boldsymbol{\lambda}_T \xrightarrow[T\to\infty]{} \mathbf{0}$, the penalized location estimator $\widehat{\boldsymbol{\phi}}_T$ converges almost surely to $\boldsymbol{\phi}_0$ as stated in Theorem 3.1. As a result, Assumption B6 is satisfied. This allows $\widehat{\boldsymbol{\theta}}_T$ to be computed using the residuals $\widetilde{\epsilon}_t\left(\widehat{\boldsymbol{\phi}}_T\right)$. In other words, $\widehat{\boldsymbol{\phi}}_T$ replaces $\widehat{\boldsymbol{\phi}}_T^{(1)}$ in $\boldsymbol{\psi} \mapsto \widetilde{Q}_T^*\left(\widehat{\boldsymbol{\phi}}_T^{(1)}, \widehat{\boldsymbol{\theta}}_T^{(1)}, \boldsymbol{\psi}\right)$ during **Step 2-2**.

The consistency of $\widehat{\boldsymbol{\theta}}_T$ follows from arguments nearly identical to those in the previous section. However, deriving the asymptotic distribution requires additional assumptions on the distribution of η_t.

B8 $\mathbb{E}\left[\left(\frac{\sigma_t^2}{w_t^*}\right)^4 + \eta_t^6\right] < \infty$ and $\mathbb{E}\left[\eta_t^3\right] = 0$.

B9 The functions $\epsilon_t(\cdot)$ and g_t are almost surely of class C^2 on $\mathcal{V}(\boldsymbol{\phi}_0)$ and $\mathcal{V}(\boldsymbol{\phi}_0) \times \mathcal{V}(\boldsymbol{\theta}_0) \times$

$\mathcal{V}(\boldsymbol{\theta}_0)$, respectively, and $\mathbb{E}\left[\left(\frac{\mathcal{G}_t}{w_t^*}\right)^4+\left(\frac{\mathcal{H}_t}{w_t^*}\right)^2+\left(\frac{\mathcal{E}_t^2}{w_t^*}\right)^4+\left(\frac{\zeta_t^2}{w_t^*}\right)^2\right]<\infty$, with

$$\begin{aligned}
\mathcal{E}_t &= \sup_{\mathcal{V}(\boldsymbol{\phi}_0)} ||\nabla \epsilon_t(\boldsymbol{\phi})||, \\
\zeta_t &= \sup_{\mathcal{V}(\boldsymbol{\phi}_0)} \left|\left|\nabla^2 \epsilon_t(\boldsymbol{\phi})\right|\right|, \\
\mathcal{G}_t &= \sup_{\mathcal{V}(\boldsymbol{\phi}_0)\times\mathcal{V}(\boldsymbol{\theta}_0)\times\mathcal{V}(\boldsymbol{\theta}_0)} ||\nabla g_t(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi})||, \\
\mathcal{H}_t &= \sup_{\mathcal{V}(\boldsymbol{\phi}_0)\times\mathcal{V}(\boldsymbol{\theta}_0)\times\mathcal{V}(\boldsymbol{\theta}_0)} \left|\left|\nabla^2 g_t(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi})\right|\right|.
\end{aligned}$$

B10 Let $\tau=\nu+n$. There is a closed and convex subset $\mathcal{C}^*$ of $\mathbb{R}^\tau$ and a sequence of symmetric positive definite $\tau\times\tau$ matrices $(\boldsymbol{\mathcal{J}}_T^*)_{T\in\mathbb{N}}$ converging a.s. to a symmetric positive definite matrix $\boldsymbol{\mathcal{J}}^*$ such that:

$$\sqrt{T}\begin{pmatrix}\hat{\boldsymbol{\phi}}_T^{(1)}-\boldsymbol{\phi}_0\\ \hat{\boldsymbol{\theta}}_T^{(1)}-\boldsymbol{\theta}_0\end{pmatrix} = \arg\min_{\boldsymbol{\xi}\in\mathcal{C}^*} ||\boldsymbol{Z}_T^*-\boldsymbol{\xi}||_{\boldsymbol{\mathcal{J}}_T^*} + o_{\mathbb{P}}(1) \quad \text{with} \quad \boldsymbol{Z}_T^* = \tfrac{1}{\sqrt{T}}\sum_{t=1}^T \boldsymbol{\Delta}_t^* \gamma^*(\eta_t),$$

where $\boldsymbol{\Delta}_t^*$ is an $\mathcal{F}_{t-1}$-measurable $\tau\times k$ 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 $\mathbb{V}[\gamma^*(\eta_t)]=:\boldsymbol{\Gamma}^*$.

B11 Letting

$$b_t^* = \sup|\mu_t-\widetilde{\mu}_t|\sup||\nabla\mu_t|| + (1+|y_t|+\sup|\mu_t|)\,||\nabla\mu_t-\nabla\widetilde{\mu}_t|| + \sup||\nabla g_t-\nabla\widetilde{g}_t||,$$

the sequences

$$\begin{aligned}
&d_t^*\left(1+|y_t|^2+\sup_\Phi|\mu_t|^2+\sup_{\Phi\times\Theta\times\Theta}|g_t|\right)\left((1+|y_t|+\sup_\Phi|\mu_t|)\sup_\Phi||\nabla\mu_t(\boldsymbol{\phi})||+\sup_{\Phi\times\Phi}||\nabla g_t||\right),\\
&a_t^*\left[(1+|y_t|+\sup_\Phi|\mu_t|)\sup_\Phi||\nabla\mu_t(\boldsymbol{\phi})||+\sup_{\Phi\times\Phi}||\nabla g_t||\right],\\
&b_t^*\left(1+|y_t|^2+\sup_\Phi|\mu_t|^2+\sup_{\Phi\times\Theta\times\Theta}|g_t|\right),
\end{aligned}$$

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

As in the previous section, the asymptotic behavior of $\widehat{\boldsymbol{\theta}}_T$ is derived using the function defined for $\boldsymbol{\vartheta}\in\mathbb{R}^n$ as follows:

$$\Lambda_T^*(\boldsymbol{\vartheta}) = \mathcal{Y}_T^*(\boldsymbol{\vartheta})+p_T^*\left(\frac{\boldsymbol{\vartheta}}{\sqrt{T}}+\boldsymbol{\theta}_0\right)-p_T^*(\boldsymbol{\theta}_0) \text{ where } \mathcal{Y}_T^*(\boldsymbol{\vartheta}) = L_T^*\left(\boldsymbol{\phi}_T^{(1)},\boldsymbol{\theta}_T^{(1)},\frac{\boldsymbol{\vartheta}}{\sqrt{T}}+\boldsymbol{\theta}_0\right)-L_T^*(\boldsymbol{\phi}_0,\boldsymbol{\theta}_0,\boldsymbol{\theta}_0).$$

The following objects exist under the previous assumptions:

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

$$I^*(\boldsymbol{\phi}_0,\boldsymbol{\theta}_0) = \mathbb{E}\left[P\left(\frac{-2\left(\epsilon_1^2-\sigma_1^2\right)\left(2\epsilon_1\nabla_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}}\mu_1(\boldsymbol{\phi}_0)+\nabla g_1(\boldsymbol{\phi}_0,\boldsymbol{\theta}_0,\boldsymbol{\theta}_0)\right)}{w_1^{*2}}\right)\right],$$

$$R^*(\boldsymbol{\phi}_0,\boldsymbol{\theta}_0) = \mathbb{E}\left[\frac{-2\left(\epsilon_1^2-\sigma_1^2\right)\left(2\epsilon_1\nabla_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}}\mu_1(\boldsymbol{\phi}_0)+\nabla g_1(\boldsymbol{\phi}_0,\boldsymbol{\theta}_0,\boldsymbol{\theta}_0)\right)\gamma^{*'}(\eta_1)\boldsymbol{\Delta}_1^{*'}}{w_1^{*2}}\right],$$

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

Similar to the previous section, the boundary issues of $\begin{pmatrix} \hat{\boldsymbol{\phi}}_T^{(1)} \\ \hat{\boldsymbol{\theta}}_T^{(1)} \end{pmatrix}$ are accounted for through Assumption B10 and results in the following projection of the Gaussian vector:

$$\boldsymbol{\mathcal{W}}^* = \begin{pmatrix} \boldsymbol{W}^* \\ \arg\min_{\boldsymbol{\xi}\in\mathcal{C}^*} ||\boldsymbol{Z}^*-\boldsymbol{\xi}||_{\boldsymbol{\mathcal{J}}^*} \end{pmatrix}.$$

The function $\mathcal{Y}_T^*$ is shown to converge in distribution to:

$$\mathcal{Y}_\infty^*(\boldsymbol{\vartheta}) = \begin{pmatrix} \boldsymbol{\mathcal{W}}_{\boldsymbol{\varphi}}^* \end{pmatrix}' M^*(\boldsymbol{\phi}_0,\boldsymbol{\theta}_0)\begin{pmatrix} \boldsymbol{\mathcal{W}}_{\boldsymbol{\varphi}}^* \end{pmatrix} \text{ with } M^*(\boldsymbol{\phi}_0,\boldsymbol{\theta}_0) = \boldsymbol{D}_2^{*'}\left[\boldsymbol{D}_1^* + \frac{1}{2}J^*(\boldsymbol{\phi}_0,\boldsymbol{\theta}_0)\boldsymbol{D}_2^*\right],$$

where $\boldsymbol{D}_1^* = \begin{pmatrix} \boldsymbol{I}_{(2\tau+n)} & \mathbf{0}_{(2\tau+n)} \end{pmatrix}$, $\boldsymbol{D}_2^* = \begin{pmatrix} \mathbf{0}_{(2\tau+n)} & \boldsymbol{I}_{(2\tau+n)} \end{pmatrix}$ and

$$J^*(\boldsymbol{\phi}_0,\boldsymbol{\theta}_0) = 2\mathbb{E}\left[P\left(\frac{2\epsilon_1\nabla_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\vartheta}}\epsilon_t(\boldsymbol{\phi_0}) - \nabla g_1(\boldsymbol{\phi}_0,\boldsymbol{\theta}_0,\boldsymbol{\theta}_0)}{w_1^*}\right)\right].$$

Theorem 3.4. *Assume that* $\frac{1}{\sqrt{T}}\boldsymbol{\lambda}_T^* \xrightarrow[T\to\infty]{} \boldsymbol{\iota}_\infty^* < +\infty$ *component-wise. Then, under Assumptions A0, B1-B2 and B6-B11, the following holds:*

$$\sqrt{T}\left(\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\right) \xrightarrow[T\to\infty]{d} \arg\min_{\boldsymbol{\vartheta}\in\mathbb{R}^n} \Lambda_\infty^*(\boldsymbol{\vartheta}),$$

with $\Lambda_\infty^*(\boldsymbol{\vartheta}) = \mathcal{Y}_\infty^*(\boldsymbol{\vartheta}) + \sum_{j\in\mathcal{S}^*}\iota_{\infty,j}^*\left[\vartheta_j sign(\theta_{0,j})\mathbb{I}_{\theta_{0,j}\neq 0} + |\vartheta_j|\mathbb{I}_{\theta_{0,j}=0}\right]$.

Remark 3.3. Under the conditions of Theorem 3.1, if the estimator $\widehat{\boldsymbol{\phi}}_T$ is asymptotically unbiased, it can replace $\widehat{\boldsymbol{\phi}}_T^{(1)}$ in **Step 2-2**. However, the non-differentiability of $\widetilde{\Lambda}_\infty$ results in an estimator that does not satisfy Assumption B10, making Theorem 3.4 inapplicable in such cases.

3.3. Application to ARMAX(1,1)-GARCHX(1,1)

This section establishes sufficient conditions under which the previous results are valid for the ARMAX(1,1)-GARCHX(1,1) model:

$$y_t = \phi_{0,1} y_{t-1} + \phi_{0,2} \epsilon_{t-1} + \boldsymbol{\varsigma}_0' \boldsymbol{Y}_{t-1} + \epsilon_t, \tag{16}$$

$$\epsilon_t = \sigma_t \eta_t, \tag{17}$$

$$\sigma_t^2 = \omega_0 + \alpha_0 \epsilon_{t-1}^2 + \beta_0 \sigma_t^2 + \boldsymbol{\pi}_0' \boldsymbol{X}_{t-1}. \tag{18}$$

The noise process $\{\eta_t, t \in \mathbb{Z}\}$ is assumed to be an i.i.d. process with zero mean and unit variance, and the process $\left\{(\eta_t, \boldsymbol{Y}_t', \boldsymbol{X}_t')', t \in \mathbb{Z}\right\}$ is strictly stationary and ergodic. To ensure that σ_t^2 is positive with probability one, the components of $\boldsymbol{\pi}_0$ and those of $\{\boldsymbol{X}_t, t \in \mathbb{Z}\}$ are assumed to be positive.

Let $\boldsymbol{\theta} = (\omega, \alpha, \beta, \boldsymbol{\pi}')' \in \Theta$ and $\boldsymbol{\phi} = (\phi_1, \phi_2, \boldsymbol{\varsigma}')' \in \Phi$, where both parameter sets are compact, convex, and contain the true parameters. The two-step recursive functions for this model are defined as follows:

$$f_t(\boldsymbol{\phi}, \boldsymbol{\upsilon}) = \left(y_{t-1}, \epsilon_{t-1}, \boldsymbol{Y}_{t-1}'\right) \boldsymbol{\upsilon}, \tag{19}$$

$$g_t(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\psi}) = \left(1, \epsilon_{t-1}^2(\boldsymbol{\phi}), \sigma_{t-1}^2(\boldsymbol{\phi}, \boldsymbol{\theta}), \boldsymbol{X}_{t-1}'\right) \boldsymbol{\psi}. \tag{20}$$

The convexity conditions A7 and B7 are satisfied. The next step is to select the weight processes to ensure that the moment assumptions A3-A4, A8-A9, B3-B4, and B8-B9 are met. A unique strictly stationary and ergodic solution to equations (16)-(18) exists under the following assumptions.

C1 $\mathbb{E}\left[\ln\left(\alpha_0 \eta_1^2 + \beta_0\right)\right] < 0$.

C2 $\phi_{0,1} \neq -\phi_{0,2}$.

C3 $\sup_{\boldsymbol{\phi} \in \Phi} |\phi_1| < 1$ and $\sup_{\boldsymbol{\phi} \in \Phi} |\phi_2| < 1$.

C4 $\exists \delta > 0 : \mathbb{E}\left[||\boldsymbol{Y}_1||^\delta + ||\boldsymbol{X}_1||^\delta\right] < \infty$.

C5 $\sup_{\boldsymbol{\theta} \in \Theta} |\beta| < 1$.

Assumption C1 and the existence of a small order moment C4 are standard requirements for the existence of a strictly stationary, ergodic, and causal solution to equations (17)-(18) with small order moments, see Lemma 2 in Francq and Thieu (2019). Adding Assumptions C2 and C3, Proposition 1 in Pan et al. (2007) ensures the existence of a strictly stationary, ergodic, causal, and invertible solution for (16). Under Assumption C5, since the parameter set $\Phi \times \Theta$ is compact, there exists a constant $0 < \rho < 1$ such that $\sup_{\boldsymbol{\phi} \in \Phi} |\phi_1| < \rho$, $\sup_{\boldsymbol{\phi} \in \Phi} |\phi_2| < \rho$, and

$\sup_{\boldsymbol{\theta}\in\Theta} |\beta| < \rho$. The weight process is defined as follows:

$$w_t = 1 + \sum_{t\in\mathbb{N}} c^i \left|\left|\boldsymbol{O}_{t-i-1}\right|\right|_1,$$

where $||\cdot||_1$ denotes the L^1 norm, and c is a decay coefficient satisfying:

C6 $\rho < c < 1$.

The scale weight process is given by $w_t^* = w_t^2$. Alternatively, the scale weight process can be defined as $w_t^* = 1 + \sum_{t\in\mathbb{N}} c^i \left|\left|\boldsymbol{O}_{t-i-1}\right|\right|_2^2$, where $||\cdot||_2$ denotes the L^2 norm. Both definitions provide the necessary control over the moments of the GARCHX process.

The first-step estimator is assumed to be the QMLE, which is strongly consistent and satisfying Assumptions A10 and B10. The following result applies.

Corollary 3.1. *Under C1-C6 Theorems 3.1, 3.2 and 3.3 hold. Furthermore, if* $\mathbb{E}\left[\eta_t^6\right] < +\infty$ *and* $\mathbb{E}\left[\eta_t^3\right] = 0$*, Theorem 3.4 also holds.*

In practice, the fitted conditional variance of the first-step QMLE is used as the weight process for the second-step penalized WLSE of the scale component, and the fitted conditional volatility is used as the weight process for the second-step penalized WLSE of the location component.

4. Numerical Experiments

In this section, a series of Monte Carlo experiments is presented to illustrate the convergence properties of the estimators and their effectiveness in selecting relevant variables. The model used for these experiments is specified as follows:

$$\begin{aligned}
y_t &= \phi_{0,1} y_{t-1} + \phi_{0,2} y_{t-2} + \psi_{0,1}\epsilon_{t-1} + \psi_{0,2}\epsilon_{t-2} + \boldsymbol{\varsigma}_0' \boldsymbol{Y}_{t-1} + \epsilon_t, \\
\epsilon_t &= \sigma_t \eta_t, \\
\sigma_t^2 &= \omega_0 + \alpha_0 \epsilon_{t-1}^2 + \beta_0 \sigma_{t-1}^2 + \boldsymbol{\pi}_0' \boldsymbol{X}_{t-1}.
\end{aligned}$$

The parameters are defined as $\boldsymbol{\phi}_0 = (\phi_{0,1}, \phi_{0,2}, \psi_{0,1}, \psi_{0,2}, \boldsymbol{\varsigma}_0')'$ and $\boldsymbol{\theta}_0 = (\omega_0, \alpha_0, \beta_0, \boldsymbol{\pi}_0')'$. The exogenous components are specified as:

$$\boldsymbol{Y}_t = (v_{1,t}, v_{2,t}, v_{1,t-1}, v_{2,t-1}, v_{1,t-2}, v_{2,t-2})' \ , \boldsymbol{X}_t = (x_{1,t}, x_{2,t}, x_{1,t-1}, x_{2,t-1}, x_{1,t-2}, x_{2,t-2})',$$

where:

$$\forall t \in \mathbb{Z} : \begin{pmatrix} v_{1,t} \\ v_{2,t} \\ z_{1,t} \\ z_{2,t} \end{pmatrix} = 0.7 \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \\ z_{1,t-1} \\ z_{2,t-1} \end{pmatrix} + \boldsymbol{e}_t, \quad \begin{pmatrix} x_{1,t} \\ x_{2,t} \end{pmatrix} = \begin{pmatrix} e^{z_{1,t}} \\ e^{z_{2,t}} \end{pmatrix}, \quad \boldsymbol{e}_t \overset{i.i.d.}{\sim} \mathcal{N}\left(\boldsymbol{0}_4, \boldsymbol{I}_4\right). \tag{21}$$

The parameter values used in the simulations are:

$$\boldsymbol{\varsigma}_0 = (1, 0.9, -0.5, 0, 0, 0)', \boldsymbol{\phi}_0 = \left(0.9, 0, -0.3, 0, \boldsymbol{\varsigma}_0'\right)', \boldsymbol{\pi}_0 = (0.15, 0.3, 0, 0, 0, 0)', \boldsymbol{\theta}_0 = \left(0.1, 0.09, 0.84, \boldsymbol{\pi}_0'\right)'.$$

The DGP is ARMAX(1,1)-GARCHX(1,1), the true AR(2) and MA(2) coefficients are zero. An ARMAX(2,2)-GARCHX(1,1) model is fitted, applying penalization to the AR(2), MA(2), and all exogenous coefficients. The penalized and non-penalized index sets are:

$$\mathcal{S} = \{2, 4, 5, \ldots, 10\} \ , \overline{\mathcal{S}} = \{1, 3\} \ , \mathcal{S}^* = \{4, \ldots, 9\} \ , \overline{\mathcal{S}^*} = \{1, 2, 3\}.$$

The sets of active and inactive index sets are:

$$\mathcal{A} = \{1, 3, 5, 6, 7\} \ , \overline{\mathcal{A}} = \{2, 4, 8, 9, 10\} \ , \mathcal{A}^* = \{1, \ldots, 5\} \ , \overline{\mathcal{A}^*} = \{6, \ldots, 9\}.$$

A total of 10^3 trajectories, each of length 5500, are generated. The initial 500 observations are treated as a burn-in period and excluded from the analysis. To examine convergence, the procedure is applied to progressively larger observation windows of sizes 250, 500, 1500, 2500, and 5000. For each window size and trajectory, the QMLE is first computed, followed by the LARS-LASSO paths for the ARMAX and GARCHX components (see **Step 2-1** and **Step 2-2**, Section 2.1). Since these paths are computed separately, two hyperparameters, $\boldsymbol{\lambda}_T$ and $\boldsymbol{\lambda}_T^*$, must be tuned.

The first approach, referred to as "separate selection", proceeds as follows. For each step in the LARS-LASSO path for the ARMAX component, the post-LASSO QMLE is computed while keeping all GARCHX parameters active, and the AIC or BIC is used to select the subset of active ARMAX parameters. Similarly, for each step in the LARS-LASSO path for the GARCHX component, the post-LASSO QMLE is computed while keeping all ARMAX parameters active, and the AIC or BIC is used to select the subset of active GARCHX parameters. A final post-LASSO QMLE is then performed using the active parameter subsets identified in the previous steps.

The second approach, called "nested selection", evaluates all combinations of steps in the LARS-LASSO paths for both the ARMAX and GARCHX components. For each combination, the post-LASSO QMLE is computed using the parameters which are active at the corresponding ARMAX-GARCHX (joint) LARS-LASSO step, and the AIC or BIC is evaluated. The post-LASSO QMLE with the best AIC or BIC score is then selected.

In the experiments, both methods produced nearly identical results, with selection rate differences of approximately 2% in the worst cases. The "separate selection" method was preferred due to its lower computational cost. The number of criteria evaluated is of order $|\mathcal{S}| + \left|\mathcal{S}^{scale}\right|$ for the "separate selection" method, compared to $\left(|\mathcal{S}| + \left|\mathcal{S}^{scale}\right|\right)^2$ for the "nested selection". The procedure was applied across all window sizes, reusing the same set

of trajectories while progressively expanding the observation windows as defined in the grid.

Remark 4.1. The hyperparameters $\boldsymbol{\lambda}_T$ and $\boldsymbol{\lambda}_T^*$ are used to select the optimal subset of parameters, improving model sparsity. Instead of directly tuning these hyperparameters, the active parameters were identified using the AIC and BIC criteria. Consequently, these hyperparameters are not explicitly referenced, as they function solely as tools for sub-model selection.

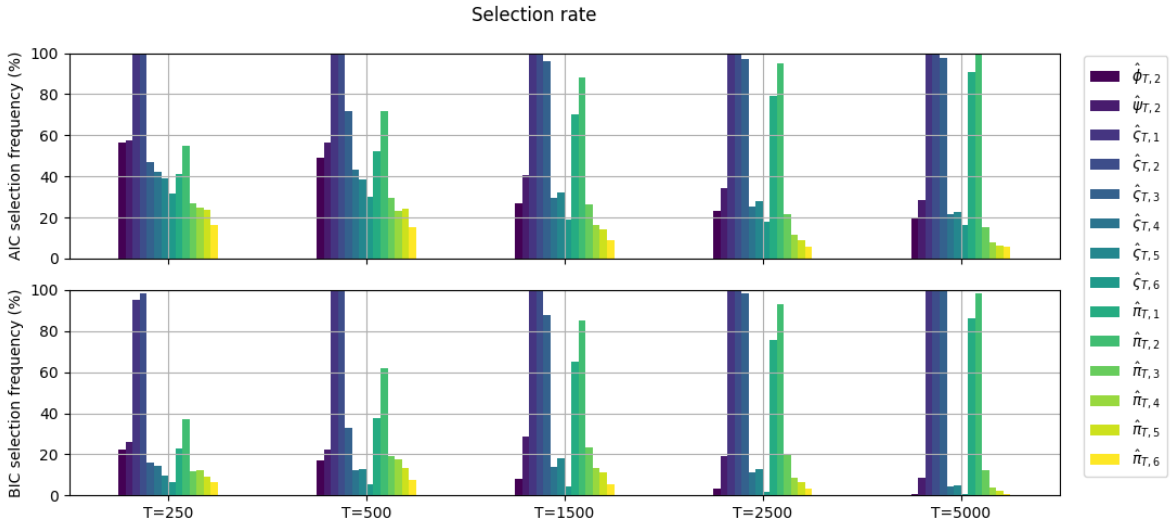

Figure 1: Selection rates using AIC and BIC criteria for varying sample sizes (10^3 trajectories).

The selection of parameters in $\mathcal{S}$ and $\mathcal{S}^*$ is conducted with the aim of correctly identifying active variables ($\mathcal{S} \cap \mathcal{A}$ and $\mathcal{S}^* \cap \mathcal{A}^*$) while minimizing the inclusion of inactive variables ($\overline{\mathcal{S}} \cap \overline{\mathcal{A}}$ and $\overline{\mathcal{S}^*} \cap \overline{\mathcal{A}^*}$).

As shown in Figure 1, both AIC and BIC improve in distinguishing active from inactive variables as sample sizes increase. BIC, being more conservative, emphasizes model simplicity and is particularly effective at excluding irrelevant variables, especially with larger sample sizes. This aligns well with our focus on achieving model sparsity to avoid overfitting. In contrast, AIC is more inclusive, capturing a broader set of variables, which can be beneficial for detecting subtle signals but may include a few irrelevant ones.

The method shows high accuracy in identifying active variables in the ARMAX component. For small samples (e.g., 250 observations), the selection rate for active variables is nearly 100%, with inactive variables mostly excluded. In the GARCHX component, selection accuracy is lower due to the reliance on residuals instead of observed variables. However, the accuracy for GARCHX parameters improves as the sample size increases, reflecting the method's consistency.

Figures 2 and 3 illustrate the convergence of the two-step and post-LASSO estimators for non-penalized coefficients ($\overline{\mathcal{S}}$ and $\overline{\mathcal{S}^*}$) and penalized coefficients ($\mathcal{S}$ and $\mathcal{S}^*$), respectively.

Figure 2 shows significant variability in the estimation of the GARCHX intercept $\widehat{\omega}_T$. This occurs because, when LASSO shrinks GARCHX coefficients to zero–classifying them as $\mathcal{S}^* \cap \overline{\mathcal{A}^*}$ instead of $\mathcal{S}^* \cap \mathcal{A}^*$–the intercept absorbs the mean effect of the omitted variables. In Figure 3, this shrinkage mechanism is shown to affect the convergence of active penalized coefficients ($\mathcal{S} \cap \mathcal{A}$ and $\mathcal{S}^* \cap \mathcal{A}^*$), particularly for smaller samples. This interdependence between penalized coefficients and the intercept demonstrates how shrinkage of relevant variables destabilizes the intercept through compensatory adjustments.

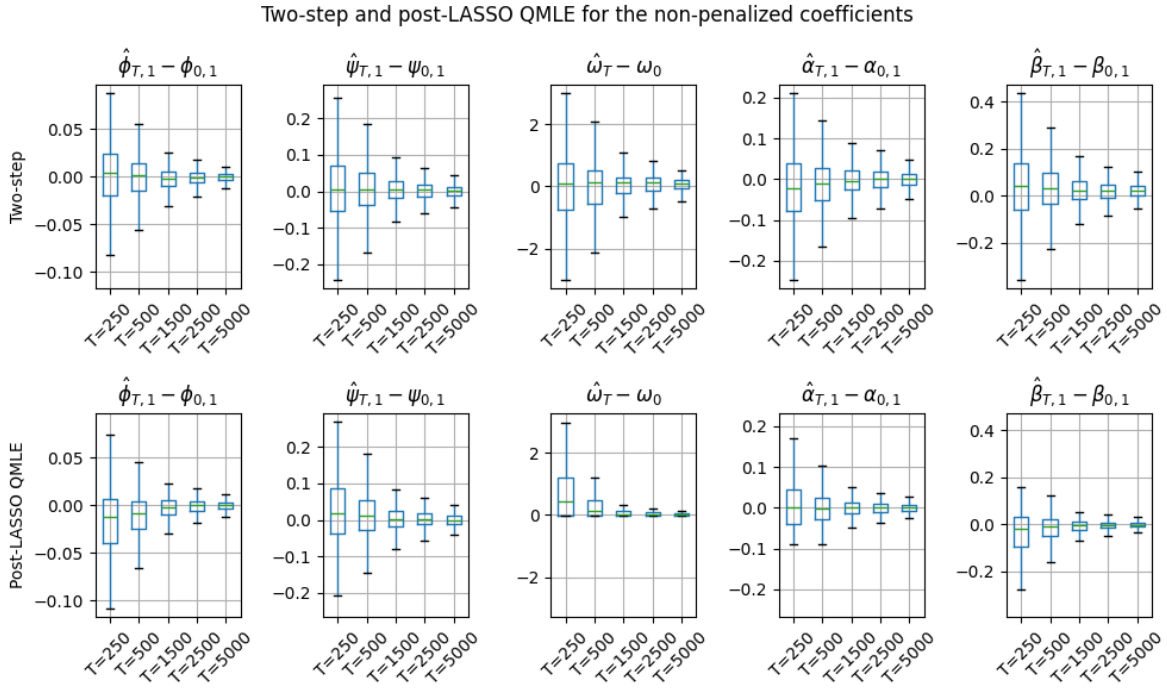

Figure 2: Box plots of two-step and post-LASSO QMLE estimation errors, across 10^3 trajectories, for the non-penalized parameters.

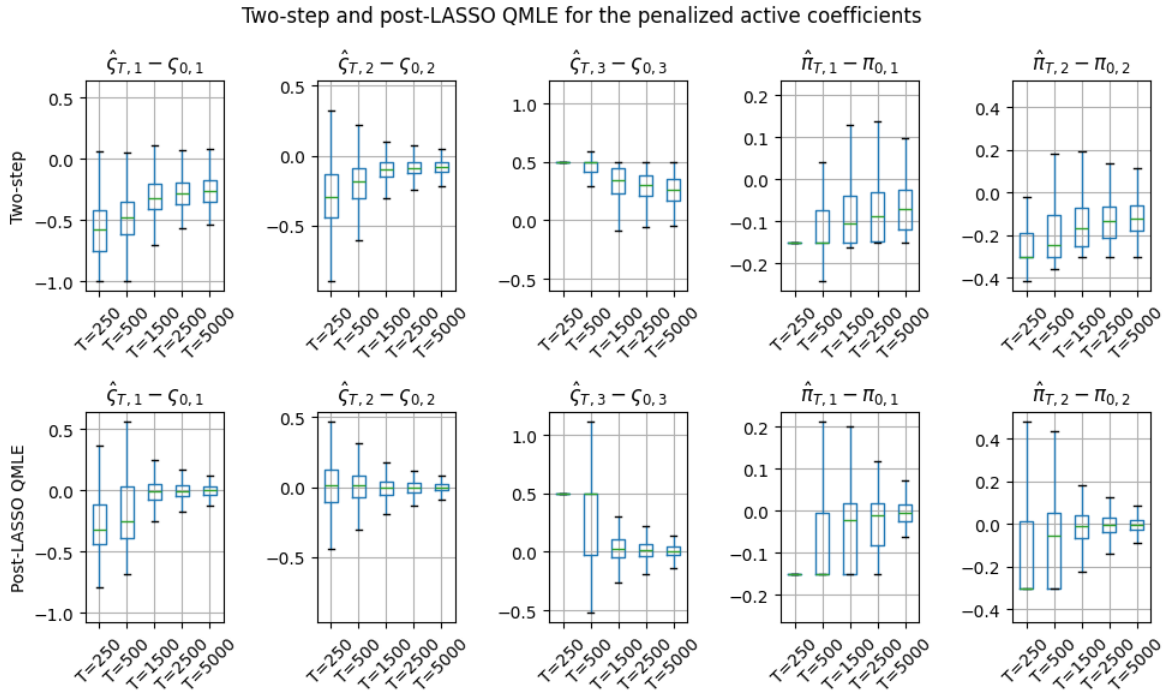

Figure 3: Box plots of two-step and post-LASSO QMLE estimation errors, across 10^3 trajectories, for the penalized active parameters.

This Monte Carlo experiment demonstrates that the finite sample performance of the procedure aligns with theoretical expectations. Active and inactive variables are correctly identified, and the active variables are estimated with increasing accuracy as the sample size grows. Furthermore, the number of fitted sub-models required to identify the optimal model is significantly reduced thanks to the LARS-LASSO algorithm.

5. Application to Electricity Data

In this real-data application, day-ahead electricity prices at 6 PM are modeled for France (FR), Belgium (BE), Switzerland (CH), and the Netherlands (NL). These prices are determined the day before production and delivery through hourly auctions for the following day. Unlike other commodities, electricity prices exhibit mean reversion, price spikes, and negative values due to the inability to store electricity.

The analysis covers the period from January 5, 2015, to June 30, 2021, using data from the ENTSO-E website. The dataset consists of hourly time series for each country, denoted as $p_{t,j}^{NL}$, $p_{t,j}^{BE}$, $p_{t,j}^{FR}$, and $p_{t,j}^{CH}$, where t represents the day and $j \in \{1, \ldots, 24\}$ the hour. By convention, $p_{t,0}$ corresponds to the 6 PM price of day $t-1$. To address non-stationarity and capture day-to-day fluctuations, the 6 PM price increments for each country are computed as the difference in prices at 6 PM between consecutive days and are denoted as NL_t, BE_t, FR_t,

and CH_t, as the difference in prices at 6 PM between consecutive days. Figure 4 presents the 6 PM prices and their corresponding increments.

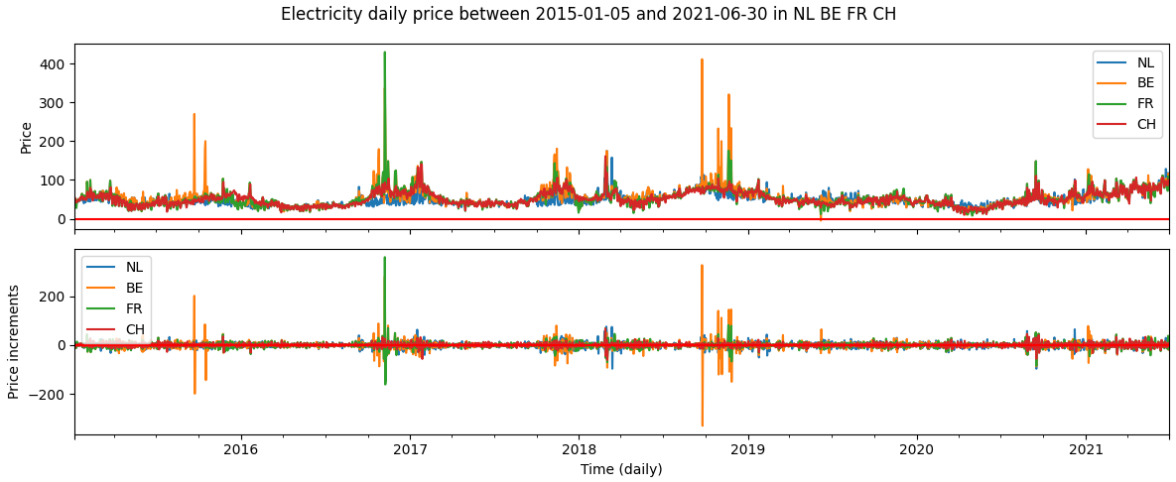

Figure 4: Electricity prices and increments at 6 PM from January 5, 2015 to June 30, 2021 in France, Belgium, Switzerland and Netherlands

In the work of Liu and Shi (2013) and Frömmel et al. (2014), ARIMA-GARCH models with limited lag orders have been shown to effectively describe electricity price dynamics. Liu and Shi (2013) used an ARMA-GARCH model with five-month AR lags and various GARCH specifications to model hourly electricity prices, achieving residuals without significant autocorrelation. Frömmel et al. (2014) extended this approach by including realized measures as exogenous variables in the GARCHX component and adding seasonal indicators, such as holidays, weekends, and broader seasonal cycles, in the AR component to account for seasonality.

This study adopts an Integrated AR model of order 30, one differentiation order and exogenous covariates (IARX) with GARCHX(1,1) innovations, applying penalization to all coefficients except the GARCHX intercept (ω) to ensure a parsimonious specification. The IAR order 30 allows the two-step estimator to shrink irrelevant parameters, potentially leading to a sparse model consistent with a lower-order seasonal structure. For each country, an IARX(30)-GARCHX(1,1) model is estimated using QMLE. The exogenous variables in the IARX component include the log-returns of the Dutch TTF Natural Gas Calendar (sourced from Macrobond), the log-returns of the USD/EUR exchange rate (from Yahoo Finance), and lagged price increments from neighboring countries. Additionally, two realized measures are included:

- The realized variance:

$$RV_t = \sum_{j=1}^{23} r_{t,j}^2 \quad \text{where} \quad r_{t,j} = p_{t,j} - p_{t,j-1}, \tag{22}$$

- The intra-day range:

$$IR_t = \left(\max_j p_{t,j} - \min_j p_{t,j} \right)^2. \tag{23}$$

The stationarity of the dataset was assessed using the Phillips-Perron, KPSS, and ADF tests, all of which confirmed stationarity. The notation follows the definitions provided in equations (16)–(18), with the exogenous variables specified as follows:

$$\begin{aligned}
\boldsymbol{Y}_t^{FR} &= (USD/EUR_t, Gas_t, NL_t, BE_t, CH_t)', \quad \boldsymbol{X}_t^{FR} = \left(RV_t^{FR}, IR_t^{FR}, USD/EUR_t^2, Gas_t^2, NL_t^2, BE_t^2, CH_t^2\right)', \\
\boldsymbol{Y}_t^{NL} &= (USD/EUR_t, Gas_t, BE_t, FR_t, CH_t)', \quad \boldsymbol{X}_t^{NL} = \left(RV_t^{NL}, IR_t^{NL}, USD/EUR_t^2, Gas_t^2, BE_t^2, FR_t^2, CH_t^2\right)', \\
\boldsymbol{Y}_t^{BE} &= (USD/EUR_t, Gas_t, NL_t, FR_t, CH_t)', \quad \boldsymbol{X}_t^{BE} = \left(RV_t^{BE}, IR_t^{BE}, USD/EUR_t^2, Gas_t^2, NL_t^2, FR_t^2, CH_t^2\right)', \\
\boldsymbol{Y}_t^{CH} &= (USD/EUR_t, Gas_t, NL_t, BE_t, FR_t)', \quad \boldsymbol{X}_t^{CH} = \left(RV_t^{CH}, IR_t^{CH}, USD/EUR_t^2, Gas_t^2, NL_t^2, BE_t^2, FR_t^2\right)'.
\end{aligned}$$

Here, the variable USD/EUR_t represents the log-return of the USD/EUR exchange rate, and Gas_t denotes the log-return of the Dutch TTF Natural Gas Calendar. The realized measures specific to each country are denoted by their corresponding superscripts.

The robustness of the variable selection procedure was further evaluated by applying the two-step estimation method to the full dataset and to a 730-day rolling window, updated every 15 days.

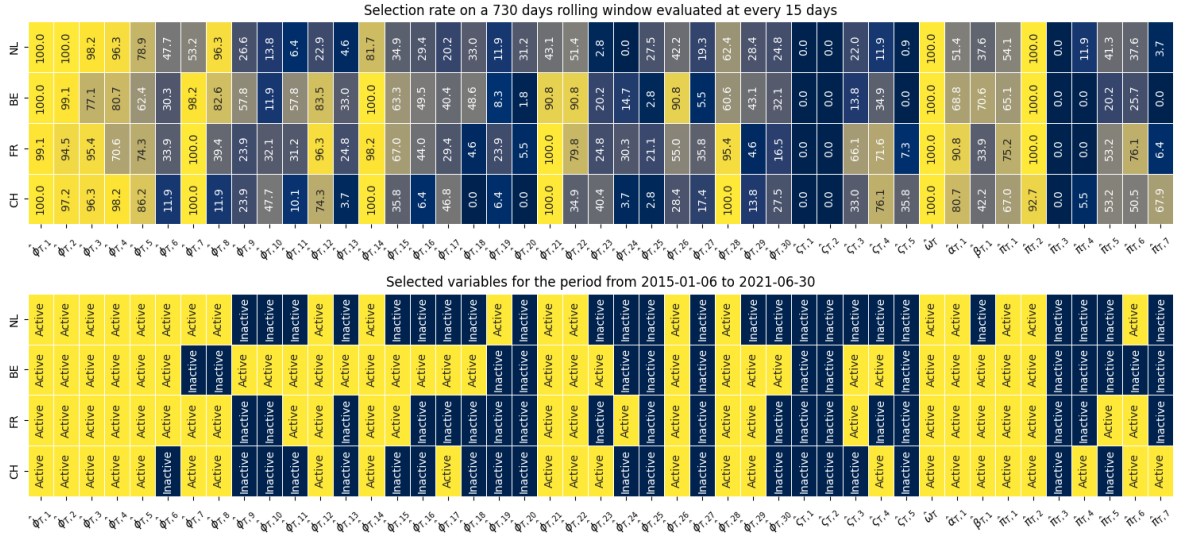

Figure 5: Frequency of selection across rolling windows (top) and selection results on the full sample (bottom).

Figure 5 presents the variable selection results for the proposed estimator. Autoregressive coefficients for lags 1 through 5, as well as 7 (one week), 14 (two weeks), 21 (three weeks), and 28 (four weeks), are consistently retained across the rolling window and the entire period. This stability suggests a significant and persistent seasonal effect in electricity prices. Real-

ized measures, particularly the intra-day range, ARCH and GARCH components, are also frequently selected, indicating their relevance in explaining volatility.

These results are consistent with findings in the existing literature. For example, Frömmel et al. (2014) highlighted the significance of realized measures in explaining electricity price volatility using a different ARMA-GARCH framework. Similarly, Liu and Shi (2013) employed SARIMA models on hourly electricity prices, utilizing seasonal structures similar to those identified in this application.

The observed temporal variability in variable selection may indicate structural changes in electricity markets. While regime-switching dynamics are not explicitly modeled in this analysis, such variability could suggest their presence—a topic frequently addressed in the literature through Markov Regime Switching models. For instance, Janczura and Weron (2012) and Samitas and Armenatzoglou (2014) demonstrated that regime-switching frameworks effectively capture price spike regimes, which may align with the parameter evolution identified here. The frequent selection of the intra-day range could also reflect its role in signaling spike-driven regime switches. These observations suggest that regime-switching dynamics warrant further investigation in future research.

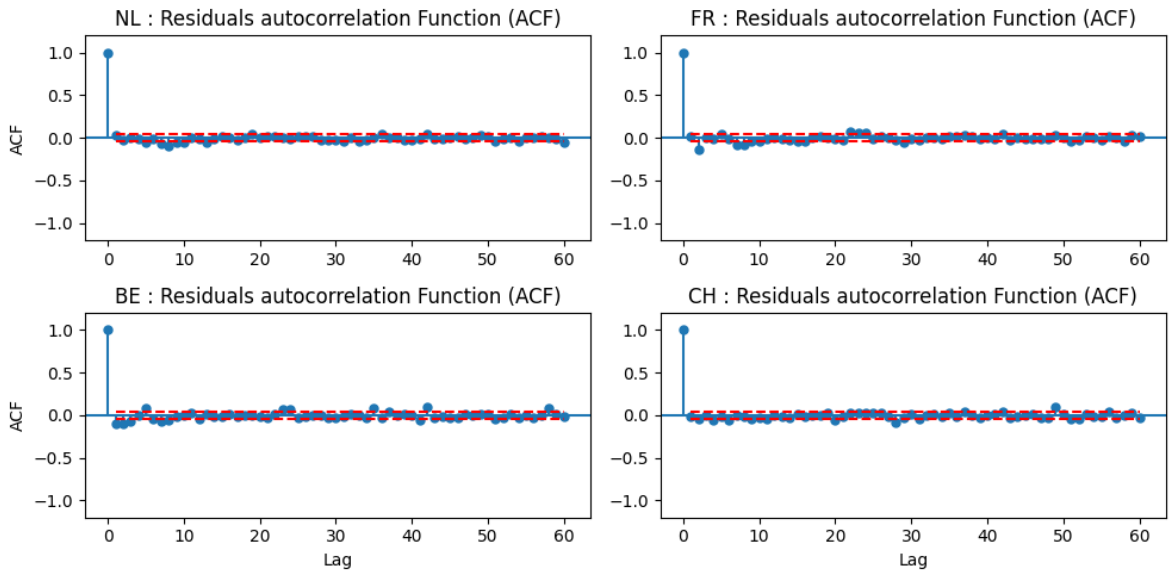

Figure 6: Parsimonious model residuals with the usual significance band.

Figure 6 confirms that the residuals of the selected model exhibit auto-correlations consistent with a GARCH process. The used significance bands are approximations using the usual $\pm\frac{1.96}{\sqrt{T}}$. A general significance band formula for non-i.i.d. white noise can calculated using generalized Bartlett formula (see Francq and Zakoïan (2009)).

6. Conclusion

In conclusion, this paper develops a penalized two-stage estimator for dynamic location-scale models. The first stage relies on an initial estimator with standard properties, such as strong consistency and a Bahadur expansion, while accounting for potential boundary issues, as is common with QMLE. In the second stage, a penalized WLSE is employed, avoiding higher-order moment assumptions. Under mild conditions, the penalized two-stage estimator is shown to converge to the true parameters, with its asymptotic distribution derived. For standard models, these conditions reduce to simple and standard assumptions. The computational efficiency of the method is enhanced by the LARS-LASSO algorithm, which computes the full solution path for standard GARCH and ARMA-GARCH models at low computational cost. Future extensions to adaptive LASSO remain a potential avenue for further research. The theoretical properties are supported by simulations and a real-world application. Simulation results show that, as sample sizes increase, the selection of the most parsimonious model using AIC and BIC criteria approaches 100%. Post-selection, WLS and post-LASSO QMLE estimators demonstrate convergence. The methodology is applied to electricity prices in France, Belgium, the Netherlands, and Switzerland through an IARX-GARCHX model. The results highlight the model's ability to achieve sparsity while identifying the seasonal effects typical of electricity markets.

Future work could extend the two-step penalization framework to multidimensional GARCHX processes. A multidimensional extension would allow joint estimation of inter-dependencies across assets, improving applications in risk management and portfolio optimization.

I. Proofs and technical lemmas

The appendix provides the proofs of the main results. Throughout, C represents a positive deterministic constant that may vary between lines.

The following Lemmas support the proofs of Theorems 3.1 and 3.2.

Lemma I.1. *Under Assumptions A1-A2, the following holds:*

$$\sup_{\boldsymbol{\phi},\boldsymbol{\upsilon}\in\Phi} \tfrac{1}{T}\left|L_T\left(\boldsymbol{\phi},\boldsymbol{\upsilon}\right)-\widetilde{L}_T\left(\boldsymbol{\phi},\boldsymbol{\upsilon}\right)\right| \xrightarrow[T\to\infty]{a.s.} 0.$$

Proof. To simplify, function parameters are omitted in this proof, with the assumption that $\boldsymbol{\phi},\boldsymbol{\upsilon}\in\Phi$. Define $u_t = f_t - y_t$ and $\widetilde{u}_t = \widetilde{f}_t - y_t$, then:

$$\left|l_t-\widetilde{l}_t\right| = \left|\tfrac{\widetilde{w}_t^2u_t^2-w_t^2\widetilde{u}_t^2}{w_t^2\widetilde{w}_t^2}\right| = \left|\tfrac{\left(\widetilde{w}_t^2-w_t^2\right)u_t^2-w_t^2(\widetilde{u}_t-u_t)(\widetilde{u}_t+u_t)}{w_t^2\widetilde{w}_t^2}\right| \leqslant C\left[d_tu_t^2+|\widetilde{u}_t-u_t|\left(|\widetilde{u}_t|+|u_t|\right)\right]. \tag{A.1}$$

Using $u_t^2 \leqslant 2\left(y_t^2+f_t^2\right) \leqslant 2\left(1+y_t^2+\sup f_t^2\right)$, Assumption A1 ensures that for sufficiently large t:

$$\sup_{\Phi\times\Phi}\left|\widetilde{f}_t\right| \leqslant 1+\sup_{\Phi\times\Phi}|f_t| \quad \text{almost surely.}$$

This leads to:

$$\left|l_t-\widetilde{l}_t\right| \leqslant C\left[d_t\left(1+y_t^2+\sup_{\Phi\times\Phi}f_t^2\right)+a_t\left(1+|y_t|+\sup_{\Phi\times\Phi}|f_t|\right)\right]. \tag{A.2}$$

Under Assumptions A1-A2, the right term in (A.2) converges almost surely to 0 as $t\to\infty$. The result follows by applying Cesàro's lemma. □

Lemma I.2. *Under Assumptions A0, A3-A4, the expectation* $\mathbb{E}\left[l_1\left(\boldsymbol{\phi},\boldsymbol{\upsilon}\right)\right]$ *exists and is finite for any* $\left(\boldsymbol{\phi},\boldsymbol{\upsilon}\right)$ *in* $\mathcal{V}\left(\boldsymbol{\phi}_0\right)\times\overset{\circ}{\Phi}$.

Proof. Rewriting l_1 using $\epsilon_1=\sigma_1\eta_1$ gives:

$$l_1\left(\boldsymbol{\phi},\boldsymbol{\upsilon}\right) = \left(\tfrac{\sigma_1\eta_1}{w_1}+\tfrac{\mu_t-f_1(\boldsymbol{\phi},\boldsymbol{\upsilon})}{w_1}\right)^2.$$

Under Assumptions A0 and A3, the term $\frac{\sigma_1}{w_1}\eta_1$ belongs to L^2. Additionally, by Assumption A4 and the mean value inequality:

$$\left|\tfrac{\mu_1-f_1(\boldsymbol{\phi},\boldsymbol{\upsilon})}{w_1}\right| \leqslant \mathcal{D}\left(\Phi^2\right)\tfrac{M_1}{w_1},$$

where $\mathcal{D}\left(A\right)$ represents the diameter of a set A, defined as:

$$\mathcal{D}\left(A\right) = \sup_{\boldsymbol{x},\boldsymbol{y}\in A}||\boldsymbol{x}-\boldsymbol{y}|| \in \overline{\mathbb{R}}^+.$$

The compactness of Φ ensures that $\mathcal{D}\left(\Phi^2\right)$ is finite. Assumption A4 guarantees that $\frac{M_1}{w_1}$ belongs to L^2. This establishes the result. □

Lemma I.3. *Under Assumptions A0, A3-A6, the following holds:*

$$\sup_{(\boldsymbol{\phi},\boldsymbol{\upsilon})\in\mathcal{V}(\boldsymbol{\phi}_0)\times\overset{\circ}{\Phi}}\left|\frac{1}{T}\frac{\partial L_T}{\partial\boldsymbol{\phi}'}(\boldsymbol{\phi},\boldsymbol{\upsilon})\left(\widehat{\boldsymbol{\phi}}_T^{(1)}-\boldsymbol{\phi}_0\right)\right|\xrightarrow[T\to\infty]{a.s.}0.$$

Proof. For $(\boldsymbol{\phi},\boldsymbol{\upsilon})\in\mathcal{V}(\boldsymbol{\phi}_0)\times\overset{\circ}{\Phi}$ and under Assumption A4, the derivative $\frac{\partial L_T}{\partial\boldsymbol{\phi}}(\boldsymbol{\phi},\boldsymbol{\upsilon})$ exists. Using the equivalence of norms in finite-dimensional spaces and the Cauchy-Schwarz inequality:

$$\left|\frac{1}{T}\frac{\partial L_T}{\partial\boldsymbol{\phi}'}(\boldsymbol{\phi},\boldsymbol{\upsilon})\left(\widehat{\boldsymbol{\phi}}_T^{(1)}-\boldsymbol{\phi}_0\right)\right|\leqslant\frac{C}{T}\left\|\frac{\partial L_T}{\partial\boldsymbol{\phi}}(\boldsymbol{\phi},\boldsymbol{\upsilon})\right\|\left\|\widehat{\boldsymbol{\phi}}_T^{(1)}-\boldsymbol{\phi}_0\right\|\leqslant C\left\|\widehat{\boldsymbol{\phi}}_T^{(1)}-\boldsymbol{\phi}_0\right\|\frac{1}{T}\sum_{t=1}^{T}\left\|\frac{\partial l_t}{\partial\boldsymbol{\phi}}(\boldsymbol{\phi},\boldsymbol{\upsilon})\right\|.$$

Applying the mean value theorem:

$$\left\|\frac{\partial l_t}{\partial\boldsymbol{\phi}}(\boldsymbol{\phi},\boldsymbol{\upsilon})\right\|\leqslant 2\frac{|f_t(\boldsymbol{\phi},\boldsymbol{\upsilon})-\mu_t|+|\eta_t|\sigma_t}{w_t^2}\left\|\frac{\partial f_t}{\partial\boldsymbol{\phi}}(\boldsymbol{\phi},\boldsymbol{\upsilon})\right\|\leqslant C\frac{M_t^2+|\eta_t|\sigma_t M_t}{w_t^2}.$$

Therefore:

$$\left|\frac{1}{T}\frac{\partial L_T}{\partial\boldsymbol{\phi}'}(\boldsymbol{\phi},\boldsymbol{\upsilon})\cdot\left(\widehat{\boldsymbol{\phi}}_T^{(1)}-\boldsymbol{\phi}_0\right)\right|\leqslant C\left\|\widehat{\boldsymbol{\phi}}_T^{(1)}-\boldsymbol{\phi}_0\right\|\frac{1}{T}\sum_{t=1}^{T}\frac{M_t^2+|\eta_t|\sigma_t M_t}{w_t^2}. \tag{A.3}$$

Under Assumption A0, the process $\left\{\frac{M_t^2+|\eta_t|\sigma_t M_t}{w_t^2},t\in\mathbb{Z}\right\}$ is strictly stationary and ergodic. Furthermore, under Assumptions A3-A4 and using Hölder's inequality, this process belongs to L^1. Applying the ergodic theorem ensures that the average in the right side of (A.3) converges almost surely to a finite value. Assumption A6 guarantees the strong consistency of the first-step estimator, which leads to the conclusion. □

The proof of Theorem 3.1 relies on specific results from Rockafellar (1970) and Davis et al. (1992), which will be stated.

Theorem I.1 (Rockafellar (1970) Theorem 10.8)**.** *Let C be a relatively open convex set, and let $f_1,f_2,\ldots$ be a sequence of finite convex functions on C. Suppose that the sequence converges pointwise on a dense subset of C, i.e. that there exists a subset C' of C such that $C\subset$ closure (C') and, for each $\boldsymbol{x}\in C'$, the limit of $f_1(\boldsymbol{x}),f_2(\boldsymbol{x}),\ldots$, exists and is finite. The limit then exists for every $\boldsymbol{x}\in C$, and the function f, where*

$$f(\boldsymbol{x})=\lim_{i\to\infty}f_i(\boldsymbol{x}),$$

is finite and convex on C. Moreover the sequence $f_1,f_2,\ldots$, converges to f uniformly on each closed bounded subset of C.

The following Lemma is an adaptation of Davis et al. (1992) Lemma 2.2.. Its statement and proof are presented below with slight modifications to fit the context of this paper. The original lemma establishes the convergence in distribution of argmins, and the same arguments are employed here to justify this property in an almost sure framework.

Lemma I.4. *Let $\{V_T(\cdot)\}$ and $V(\cdot)$ be stochastic processes continuous and strictly convex on an open convex set $A \subset \mathbb{R}^p$ and suppose that for each $\boldsymbol{x} \in A$*

$$V_T(\boldsymbol{x}) \xrightarrow[T\to\infty]{a.s.} V(\boldsymbol{x}).$$

Let $\boldsymbol{\xi}_T$ minimize $V_T(\cdot)$ and $\boldsymbol{\xi}$ minimize $V(\cdot)$ such that $\boldsymbol{\xi}_T, \boldsymbol{\xi} \in A$. Then

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

Proof. The strict convexity insures the uniqueness of the argmins in the following. Using Theorem I.1, for any given compact set $K \subset A$:

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

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

$$V_T(\boldsymbol{u}) > V_T(\boldsymbol{\xi}) > V_T(\boldsymbol{\xi}_T).$$

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

Proof of Theorem 3.1. Under Assumptions A4-A6, and for sufficiently large T, $\widehat{\boldsymbol{\phi}}_T^{(1)} \in \mathcal{V}(\boldsymbol{\phi}_0)$ almost surely. Using a Taylor-Lagrange expansion for L_T with respect to $\boldsymbol{\phi}$, the following holds for any $\boldsymbol{v} \in \overset{\circ}{\Phi}$:

$$L_T\left(\widehat{\boldsymbol{\phi}}_T^{(1)}, \boldsymbol{v}\right) = L_T(\boldsymbol{\phi}_0, \boldsymbol{v}) + \tfrac{1}{T}\tfrac{\partial L_T}{\partial \boldsymbol{\phi}'}(\boldsymbol{\phi}_T^*, \boldsymbol{v})\left(\widehat{\boldsymbol{\phi}}_T^{(1)} - \boldsymbol{\phi}_0\right), \tag{A.4}$$

where $\boldsymbol{\phi}_T^*$ lies between $\boldsymbol{\phi}_0$ and $\widehat{\boldsymbol{\phi}}_T^{(1)}$. By Assumption A0, the process $\{l_t(\boldsymbol{\phi}, \boldsymbol{v}), t \in \mathbb{Z}\}$ is stationary and ergodic. Applying Lemma I.2 gives:

$$\forall \boldsymbol{v} \in \overset{\circ}{\Phi} : \tfrac{1}{T} L_T(\boldsymbol{\phi}_0, \boldsymbol{v}) \xrightarrow[T\to\infty]{\text{a.s.}} \mathbb{E}[l_1(\boldsymbol{\phi}_0, \boldsymbol{v})].$$

Since $\boldsymbol{\phi}_T^*$ lies between $\boldsymbol{\phi}_0$ and $\widehat{\boldsymbol{\phi}}_T^{(1)}$, it is necessarily contained within $\mathcal{V}(\boldsymbol{\phi}_0)$. Using Lemmas

I.1 and I.3, the following holds:

$$\forall \boldsymbol{\upsilon} \in \overset{\circ}{\Phi} : \frac{1}{T}\widetilde{L}_T\left(\widehat{\boldsymbol{\phi}}_T^{(1)}, \boldsymbol{\upsilon}\right) \xrightarrow[T\to\infty]{\text{a.s.}} \mathbb{E}\left[l_1\left(\boldsymbol{\phi}_0, \boldsymbol{\upsilon}\right)\right].$$

The penalty term $\frac{1}{T}p_T$ converges to a convex, finite function $p_\infty\left(\boldsymbol{\psi}\right) := \sum_{i\in\mathcal{S}} \iota_{\infty,i}\left|\psi_i\right|$. It remains to establish the convergence of $\widehat{\boldsymbol{\phi}}_T$ to $\arg\min_{\boldsymbol{\upsilon}\in\Phi} Q_\infty\left(\boldsymbol{\phi}_0, \boldsymbol{\upsilon}\right)$. Define $C_T\left(\cdot\right) = \widetilde{Q}_T\left(\widehat{\boldsymbol{\phi}}_T^{(1)}, \cdot\right)$ on $\overset{\circ}{\Phi}$. Pointwise almost sure convergence $C_T\left(\cdot\right) \xrightarrow[T\to\infty]{\text{a.s.}} C_\infty\left(\cdot\right) := Q_\infty\left(\boldsymbol{\phi}_0, \cdot\right)$ holds on $\overset{\circ}{\Phi}$.

Assumption A7 and the positivity of p_T ensure that C_T is almost surely convex and continuous for each T. Since C_∞ is almost surely continuous, Theorem I.1 implies uniform convergence of $C_T\left(\cdot\right)$ to $C_\infty\left(\cdot\right)$ on any compact subset $K \subset \overset{\circ}{\Phi}$.

Assumption A7 further ensures strict convexity of C_∞, guaranteeing uniqueness of the arg min. Assumption A5 guarantees that the arg min belongs to $\overset{\circ}{\Phi}$. By Lemma I.4, $\widehat{\boldsymbol{\phi}}_T$ converges almost surely to the unique arg min of $Q_\infty\left(\boldsymbol{\phi}_0, \cdot\right)$.

If $\boldsymbol{\iota}_\infty = \mathbf{0}$, the vector $\boldsymbol{\phi}_0$ minimizes C_∞. This follows from the following equation:

$$\frac{\partial C_\infty}{\partial \boldsymbol{\upsilon}}\left(\boldsymbol{\phi}_0\right) = \mathbb{E}\left[2\eta_t \frac{\sigma_t}{w_t^2}\left(\frac{\partial f_t(\boldsymbol{\phi}_0, \cdot)}{\partial \boldsymbol{\upsilon}}\right)_{\boldsymbol{\upsilon}=\boldsymbol{\phi}_0}\right] = \mathbb{E}\left[\eta_t\right]\mathbb{E}\left[2\frac{\sigma_t}{w_t^2}\left(\frac{\partial f_t(\boldsymbol{\phi}_0, \cdot)}{\partial \boldsymbol{\upsilon}}\right)_{\boldsymbol{\upsilon}=\boldsymbol{\phi}_0}\right] = \mathbf{0}.$$

Given that $\mathbb{E}\left[\eta_t\right] = 0$, the gradient at $\boldsymbol{\phi}_0$ is zero. The strict convexity of C_∞ guarantees that $\boldsymbol{\phi}_0$ is the unique global minimizer. □

Lemma I.5. *Under Assumptions A0 and A8-A9, the matrix $J\left(\boldsymbol{\phi}_0\right)$ exists, and:*

$$\frac{1}{T}\nabla^2 L_T\left(\boldsymbol{\phi}_T^*, \boldsymbol{\varphi}_T^*\right) \xrightarrow[T\to\infty]{a.s.} J\left(\boldsymbol{\phi}_0\right),$$

for any sequence $\begin{pmatrix}\boldsymbol{\phi}_T^* \\ \boldsymbol{\varphi}_T^*\end{pmatrix} \xrightarrow[T\to\infty]{a.s.} \begin{pmatrix}\boldsymbol{\phi}_0 \\ \boldsymbol{\phi}_0\end{pmatrix}$.

Proof. Using Assumptions A8-A9 and the mean value theorem:

$$\left|\left|\nabla^2 l_t\right|\right| = \left|\left|2\frac{P(\nabla f_t)+(f_t-y_t)\nabla^2 f_t}{w_t^2}\right|\right| = \left|\left|2\frac{P(\nabla f_t)-\sigma_t\eta_t\nabla^2 f_t+(f_t-\mu_t)\nabla^2 f_t}{w_t^2}\right|\right| \leqslant C\frac{\mathcal{M}_t^2+(\mathcal{M}_t+\sigma_t|\eta_t|)\mathcal{K}_t}{w_t^2}. \tag{A.5}$$

Thus, $J\left(\boldsymbol{\phi}_0\right)$ exists.

Let $\mathcal{B}\left(\begin{pmatrix}\boldsymbol{\phi}_0 \\ \boldsymbol{\phi}_0\end{pmatrix}, \frac{1}{j}\right)$ denote the $\left(\mathbb{R}^{2\nu}, \left|\left|\cdot\right|\right|\right)$-ball centered at $\begin{pmatrix}\boldsymbol{\phi}_0 \\ \boldsymbol{\phi}_0\end{pmatrix}$ with radius $\frac{1}{j}$, where j is a sufficiently large integer insuring $\mathcal{B}\left(\begin{pmatrix}\boldsymbol{\phi}_0 \\ \boldsymbol{\phi}_0\end{pmatrix}, \frac{1}{j}\right) \subset \mathcal{V}\left(\boldsymbol{\phi}_0\right) \times \mathcal{V}\left(\boldsymbol{\phi}_0\right)$. For T large, $\begin{pmatrix}\boldsymbol{\phi}_T^* \\ \boldsymbol{\upsilon}_T^*\end{pmatrix} \in \mathcal{B}\left(\begin{pmatrix}\boldsymbol{\phi}_0 \\ \boldsymbol{\phi}_0\end{pmatrix}, \frac{1}{j}\right)$ almost surely. This leads to:

$$\left|\left|\frac{1}{T}\nabla^2 L_T\left(\boldsymbol{\phi}_T^*, \boldsymbol{\upsilon}_T^*\right) - J\left(\boldsymbol{\phi}_0\right)\right|\right| \leqslant \frac{1}{T}\sum_{t=1}^T Z_{t,j} + \left|\left|\frac{1}{T}\nabla^2 L_T\left(\boldsymbol{\phi}_0, \boldsymbol{\phi}_0\right) - J\left(\boldsymbol{\phi}_0\right)\right|\right|, \tag{A.6}$$

where $Z_{t,j} = \sup_{\mathcal{B}\left(\binom{\boldsymbol{\phi}_0}{\boldsymbol{\phi}_0}, \frac{1}{j}\right)} \left|\left|\nabla^2 l_t\left(\boldsymbol{\phi}, \boldsymbol{v}\right) - \nabla^2 l_t\left(\boldsymbol{\phi}_0, \boldsymbol{\phi}_0\right)\right|\right|$. The triangle inequality gives:

$$Z_{t,j} \leqslant 2 \sup_{\mathcal{B}\left(\binom{\boldsymbol{\phi}_0}{\boldsymbol{\phi}_0}, \frac{1}{j}\right)} \left|\left|\nabla^2 l_t\right|\right| \leqslant 2 \sup_{\mathcal{V}(\boldsymbol{\phi}_0)\times\mathcal{V}(\boldsymbol{\phi}_0)} \left|\left|\nabla^2 l_t\right|\right|.$$

The right term of the inequality belongs to L^1 and is independent of j. Under Assumption A0 and by the ergodic theorem, $\frac{1}{T}\sum_{t=1}^{T} Z_{t,j} \xrightarrow[T\to\infty]{\text{a.s.}} \mathbb{E}\left[Z_{1,j}\right]$. Using the dominated convergence theorem, $\mathbb{E}\left[Z_{1,j}\right] \xrightarrow[j\to\infty]{} 0$. Moreover, $\mathbb{E}\left[\nabla^2 l_1\left(\boldsymbol{\phi}_0, \boldsymbol{\phi}_0\right)\right] = J\left(\boldsymbol{\phi}_0\right) + \mathbb{E}\left[\frac{2\eta_1\sigma_1}{w_1^2}\nabla^2 f_1\left(\boldsymbol{\phi}_0, \boldsymbol{\phi}_0\right)\right]$, where the second term equals $\mathbf{0}$. By the ergodic theorem:

$$\left|\left|\tfrac{1}{T}\nabla^2 L_T\left(\boldsymbol{\phi}_0, \boldsymbol{\phi}_0\right) - J\left(\boldsymbol{\phi}_0\right)\right|\right| \xrightarrow[T\to\infty]{\text{a.s.}} 0.$$

□

Lemma I.6. *Under Assumptions A0, A8-A10, the matrices $I\left(\boldsymbol{\phi}_0\right)$, $R\left(\boldsymbol{\phi}_0\right)$, and $\boldsymbol{\Sigma}$ exist. Defining:*

$$\boldsymbol{\mathcal{W}}_T = \begin{pmatrix} \frac{1}{\sqrt{T}}\nabla L_T(\boldsymbol{\phi}_0,\boldsymbol{\phi}_0) \\ \sqrt{T}\left(\boldsymbol{\phi}_T^{(1)} - \boldsymbol{\phi}_0\right) \end{pmatrix}, \tag{A.7}$$

it holds that:

$$\boldsymbol{\mathcal{W}}_T \xrightarrow[T\to+\infty]{d} \boldsymbol{\mathcal{W}}.$$

Proof. Using Assumption A10, substitute the Bahadur expansion into (A.7) gives:

$$\boldsymbol{\mathcal{W}}_T = \begin{pmatrix} \frac{1}{\sqrt{T}}\nabla L_T(\boldsymbol{\phi}_0,\boldsymbol{\phi}_0) \\ \arg\min_{\boldsymbol{\xi}\in\mathcal{C}} ||\boldsymbol{Z}_T - \boldsymbol{\xi}||_{\boldsymbol{\mathcal{J}}_T} + o_{\mathbb{P}}(1) \end{pmatrix}$$

Define:

$$\boldsymbol{U}_T = \tfrac{1}{\sqrt{T}}\sum_{t=1}^{T} \boldsymbol{\mathcal{U}}_t + o_{\mathbb{P}}\left(1\right) \quad \text{where} \quad \boldsymbol{\mathcal{U}}_t = \begin{pmatrix} \frac{2\eta_t\sigma_t}{w_t^2}\nabla f_t(\boldsymbol{\phi}_0,\boldsymbol{\phi}_0) \\ \boldsymbol{\Delta}_t\gamma(\eta_t) \end{pmatrix}.$$

Here, $\boldsymbol{\Delta}_t\gamma\left(\eta_t\right)$ belongs to L^1, with $\boldsymbol{\Delta}_t$ being $\mathcal{F}_{t-1}$-measurable and $\mathbb{E}\left[\gamma\left(\eta_t\right)\right] = \mathbf{0}$. Additionally, under Assumptions A0, A8-A9, the process $\left\{\frac{2\eta_t\sigma_t}{w_t^2}\nabla f_t\left(\boldsymbol{\phi}_0, \boldsymbol{\phi}_0\right), t \in \mathbb{Z}\right\}$ is strictly stationary, ergodic and belongs to L^2 because:

$$\left|\left|\tfrac{2\eta_t\sigma_t}{w_t^2}\nabla f_t\left(\boldsymbol{\phi}_0, \boldsymbol{\phi}_0\right)\right|\right| \leqslant \tfrac{C[\mathcal{M}_t + \sigma_t|\eta_t|]\mathcal{M}_t}{w_t^2}.$$

Since $\nabla f_t\left(\boldsymbol{\phi}_0, \boldsymbol{\phi}_0\right)$ is $\mathcal{F}_{t-1}$-measurable, $\{\boldsymbol{\mathcal{U}}_t, t \in \mathbb{Z}\}$ forms a L^2 martingale difference sequence. Applying the martingale central limit theorem from Billingsley (1961) and Slutsky's lemma gives:

$$\boldsymbol{U}_T \xrightarrow[T\to+\infty]{d} \binom{\boldsymbol{W}}{\boldsymbol{Z}} \sim \mathcal{N}\left(\mathbf{0}, \mathbb{V}\left[\boldsymbol{\mathcal{U}}_1\right]\right).$$

The variance calculation is as follows:

$$I(\boldsymbol{\phi}_0) = \mathbb{V}\left[\frac{2\eta_1\sigma_1}{w_1^2}\nabla f_1(\boldsymbol{\phi}_0, \boldsymbol{\phi}_0)\right] = 4\mathbb{E}\left[P\left(\frac{\sigma_1}{w_1^2}\nabla f_1(\boldsymbol{\phi}_0, \boldsymbol{\phi}_0)\right)\right],$$
$$R(\boldsymbol{\phi}_0) = Cov\left[\frac{2\eta_1\sigma_1}{w_1^2}\nabla f_1(\boldsymbol{\phi}_0, \boldsymbol{\phi}_0), \boldsymbol{\Delta}_1\gamma(\eta_1)\right] = \mathbb{E}\left[\frac{2\sigma_1}{w_1^2}\nabla f_1(\boldsymbol{\phi}_0, \boldsymbol{\phi}_0)\mathbb{E}\left[\eta_1(\gamma(\eta_1))'\right]\boldsymbol{\Delta}_1'\right]$$
$$\mathbb{V}\left[\boldsymbol{\Delta}_1\gamma(\eta_1)\right] = \mathbb{E}\left[P\left(\boldsymbol{\Delta}_1\boldsymbol{\Gamma}^{\frac{1}{2}}\right)\right] = \mathbb{V}\left[\boldsymbol{\Delta}_1\boldsymbol{\Gamma}^{\frac{1}{2}}\right] = \boldsymbol{\Sigma}.$$

Using Francq and Zakoïan (2019) Section 8.2, $\arg\min_{\boldsymbol{\xi}\in\mathcal{C}}||\boldsymbol{Z}_T - \boldsymbol{\xi}||_{\boldsymbol{\mathcal{J}}_T}$ is continuous with respect to $\begin{pmatrix}\boldsymbol{Z}_T\\ \boldsymbol{\mathcal{J}}_T\end{pmatrix}$. Therefore, $\boldsymbol{\mathcal{W}}_T$ is a continuous transformation of $\begin{pmatrix}\boldsymbol{U}_T\\ \boldsymbol{\mathcal{J}}_T\end{pmatrix}$. By Slutsky's theorem and the continuous mapping theorem:

$$\begin{pmatrix}\boldsymbol{\mathcal{W}}_T\\ \boldsymbol{\mathcal{J}}_T\end{pmatrix} \xrightarrow[T\to+\infty]{d} \begin{pmatrix}\boldsymbol{W}\\ \boldsymbol{\mathcal{J}}\end{pmatrix}.$$

□

For the following, the definition and theorems from Van der Vaart and Wellner (1996) are required.

Definition I.1 (Space of locally bounded functions (Van der Vaart and Wellner (1996))**.** Let $T_1 \subset T_2 \subset \ldots$ be arbitrary sets and $T = \bigcup_{i=1}^{\infty} T_i$. The space $l^{\infty}(T_1, T_2, \ldots)$ is defined as the set of all functions $\boldsymbol{z} : T \mapsto R$ that are uniformly bounded on every T_i (but not necessarily on T). This is a complete metric space with respect to the metric $d(\boldsymbol{z}_1, \boldsymbol{z}_2) = \sum_{i=1}^{\infty}\left(||\boldsymbol{z}_1 - \boldsymbol{z}_2||_{T_i} \wedge 1\right)2^{-i}$. A sequence converges in this metric if it converges uniformly on each T_i.

Theorem I.2 (Van der Vaart and Wellner (1996) Theorem 1.6.1)**.** *Let $T_1 \subset T_2, \ldots$ be arbitrary sets and $T = \bigcup_{i=1}^{\infty} T_i$. Let $X_\alpha(\omega) : \Omega_\alpha \mapsto l^{\infty}(T_1, T_2, \ldots)$ be arbitrary maps (random function). Then, the net X_α converges weakly to a tight limit if and only if the nets of restrictions $X_{\alpha|T_i} : \Omega \mapsto l^{\infty}(T_i)$ converges weakly $(i \in \mathbb{N})$ to a tight limit.*

Corollary I.1 (Corollary of Van der Vaart and Wellner (1996) Theorem 1.6.1 for convex processes)**.** *Let $\boldsymbol{X}_\alpha$ be a net of stochastic processes indexed by a convex, open subset C of $\mathbb{R}^k$ such that every path $t \mapsto \boldsymbol{X}_\alpha(t)$ is convex on C. If the net converges marginally in distribution to a limit, then it converges in distribution to a tight limit in the space $l^{\infty}(K_1, K_2, \ldots)$ for any sequence of compact sets $K_1 \subset K_2 \subset \ldots \subset C$.*

Lemma I.7. *For $\boldsymbol{\varphi}$ in $\mathbb{R}^{\nu}$ and under Assumptions A0, A4-A6, A8-A10,*

$$\mathcal{Y}_T(\cdot) \xrightarrow[T\to+\infty]{d} \mathcal{Y}_\infty(\cdot) \quad in \quad l^{\infty}(K_1, K_2, \ldots),$$

for any sequence of compact sets $K_1 \subset K_2 \subset \ldots \subset \mathbb{R}^{\nu}$.

Proof. Using Assumptions A6 and A9, a Taylor-Lagrange expansion of L_T gives:

$$\mathcal{Y}_T(\boldsymbol{\varphi}) = \begin{pmatrix} \sqrt{T}\left(\hat{\boldsymbol{\varphi}}_T^{(1)} - \boldsymbol{\varphi}_0\right) \\ \boldsymbol{\varphi} \end{pmatrix}' \frac{1}{\sqrt{T}} \nabla L_T(\boldsymbol{\phi}_0, \boldsymbol{\phi}_0) + \frac{1}{2} \begin{pmatrix} \sqrt{T}\left(\hat{\boldsymbol{\varphi}}_T^{(1)} - \boldsymbol{\varphi}_0\right) \\ \boldsymbol{\varphi} \end{pmatrix}' \frac{1}{T} \nabla^2 L_T(\boldsymbol{\phi}_T^*, \boldsymbol{\upsilon}_T^*) \begin{pmatrix} \sqrt{T}\left(\hat{\boldsymbol{\varphi}}_T^{(1)} - \boldsymbol{\varphi}_0\right) \\ \boldsymbol{\varphi} \end{pmatrix}, \tag{A.8}$$

where $\begin{pmatrix} \boldsymbol{\phi}_T^* \\ \boldsymbol{\upsilon}_T^* \end{pmatrix}$ lies between $\begin{pmatrix} \boldsymbol{\phi}_0 \\ \boldsymbol{\phi}_0 \end{pmatrix}$ and $\begin{pmatrix} \boldsymbol{\phi}_T^{(1)} \\ \frac{\boldsymbol{\varphi}}{\sqrt{T}} + \boldsymbol{\phi}_0 \end{pmatrix}$. Using Lemma I.6, define $\boldsymbol{\mathcal{W}}_T$ as in the lemma. Then:

$$\begin{pmatrix} \sqrt{T}\left(\hat{\boldsymbol{\varphi}}_T^{(1)} - \boldsymbol{\varphi}_0\right) \\ \boldsymbol{\varphi} \end{pmatrix} = D_2 \begin{pmatrix} \boldsymbol{\mathcal{W}}_T \\ \boldsymbol{\varphi} \end{pmatrix}, \quad \frac{1}{\sqrt{T}} \nabla L_T(\boldsymbol{\phi}_0, \boldsymbol{\phi}_0) = D_1 \begin{pmatrix} \boldsymbol{\mathcal{W}}_T \\ \boldsymbol{\varphi} \end{pmatrix}.$$

Rewriting $\mathcal{Y}_T(\boldsymbol{\varphi})$:

$$\begin{aligned} \mathcal{Y}_T(\boldsymbol{\varphi}) &= \left(\boldsymbol{D}_2 \begin{pmatrix} \boldsymbol{\mathcal{W}}_T \\ \boldsymbol{\varphi} \end{pmatrix}\right)' \left(\boldsymbol{D}_1 \begin{pmatrix} \boldsymbol{\mathcal{W}}_T \\ \boldsymbol{\varphi} \end{pmatrix}\right) + \frac{1}{2} \left(\boldsymbol{D}_2 \begin{pmatrix} \boldsymbol{\mathcal{W}}_T \\ \boldsymbol{\varphi} \end{pmatrix}\right)' \frac{1}{T} \nabla^2 L_T(\boldsymbol{\phi}_T^*, \boldsymbol{\upsilon}_T^*) \left(\boldsymbol{D}_2 \begin{pmatrix} \boldsymbol{\mathcal{W}}_T \\ \boldsymbol{\varphi} \end{pmatrix}\right) \\ &= \begin{pmatrix} \boldsymbol{\mathcal{W}}_T \\ \boldsymbol{\varphi} \end{pmatrix}' \boldsymbol{D}_2' \boldsymbol{D}_1 \begin{pmatrix} \boldsymbol{\mathcal{W}}_T \\ \boldsymbol{\varphi} \end{pmatrix} + \frac{1}{2} \begin{pmatrix} \boldsymbol{\mathcal{W}}_T \\ \boldsymbol{\varphi} \end{pmatrix}' \boldsymbol{D}_2' \frac{1}{T} \nabla^2 L_T(\boldsymbol{\phi}_T^*, \boldsymbol{\upsilon}_T^*) \boldsymbol{D}_2 \begin{pmatrix} \boldsymbol{\mathcal{W}}_T \\ \boldsymbol{\varphi} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\mathcal{W}}_T \\ \boldsymbol{\varphi} \end{pmatrix}' \boldsymbol{D}_2' \left[\boldsymbol{D}_1 + \frac{1}{2} \frac{1}{T} \nabla^2 L_T(\boldsymbol{\phi}_T^*, \boldsymbol{\upsilon}_T^*) \boldsymbol{D}_2\right] \begin{pmatrix} \boldsymbol{\mathcal{W}}_T \\ \boldsymbol{\varphi} \end{pmatrix}. \end{aligned}$$

Set $M_T(\boldsymbol{\phi}_T^*, \boldsymbol{\upsilon}_T^*) = \boldsymbol{D}_2' \left[\boldsymbol{D}_1 + \frac{1}{2} \frac{1}{T} \nabla^2 L_T(\boldsymbol{\phi}_T^*, \boldsymbol{\upsilon}_T^*) \boldsymbol{D}_2\right]$, so:

$$\mathcal{Y}_T(\boldsymbol{\varphi}) = \begin{pmatrix} \boldsymbol{\mathcal{W}}_T \\ \boldsymbol{\varphi} \end{pmatrix}' M_T(\boldsymbol{\phi}_T^*, \boldsymbol{\upsilon}_T^*) \begin{pmatrix} \boldsymbol{\mathcal{W}}_T \\ \boldsymbol{\varphi} \end{pmatrix}.$$

For $k \in \mathbb{N}$, let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k \in \mathbb{R}$ and $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_k \in \mathbb{R}^\nu$. Then:

$$\sum_{i=1}^k \boldsymbol{u}_i \mathcal{Y}_T(\boldsymbol{\varphi}_i) = \begin{pmatrix} \boldsymbol{\mathcal{W}}_T \\ \boldsymbol{\varphi}_1 \\ \vdots \\ \boldsymbol{\mathcal{W}}_T \\ \boldsymbol{\varphi}_k \end{pmatrix}' \begin{pmatrix} \boldsymbol{u}_1 M_T\left(\boldsymbol{\phi}_T^{*,1}, \boldsymbol{\upsilon}_T^{*,1}\right) & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \ddots & & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \ldots & & \boldsymbol{u}_k M_T\left(\boldsymbol{\phi}_T^{*,k}, \boldsymbol{\upsilon}_T^{*,k}\right) \end{pmatrix} \begin{pmatrix} \boldsymbol{\mathcal{W}}_T \\ \boldsymbol{\varphi}_1 \\ \vdots \\ \boldsymbol{\mathcal{W}}_T \\ \boldsymbol{\varphi}_k \end{pmatrix}, \tag{A.9}$$

where for all $i \in \{1, \ldots, k\}$, $\begin{pmatrix} \boldsymbol{\phi}_T^{*,k} \\ \boldsymbol{\upsilon}_T^{*,k} \end{pmatrix}$ lies between $\begin{pmatrix} \boldsymbol{\phi}_0 \\ \boldsymbol{\upsilon}_0 \end{pmatrix}$ and $\begin{pmatrix} \hat{\boldsymbol{\phi}}_T^{(1)} \\ \frac{\boldsymbol{\varphi}_i}{\sqrt{T}} + \boldsymbol{\phi}_0 \end{pmatrix}$.

First, Lemma I.5 ensures almost sure convergence of the Hessian term and Lemma I.6 gives the asymptotic distribution of $\boldsymbol{\mathcal{W}}_T$. Second, by the continuous mapping theorem, Slutsky's lemma, and Lemma I.5, $\sum_{i=1}^k \boldsymbol{u}_i \mathcal{Y}_T(\boldsymbol{\varphi}_i)$ converges weakly to $\sum_{i=1}^k \boldsymbol{u}_i \mathcal{Y}_\infty(\boldsymbol{\varphi}_i)$. Then, by the Cramér-Wold theorem, the finite-dimensional distributions of $\mathcal{Y}_T$ converge to those of $\mathcal{Y}_\infty$. Finally, using Corollary I.1, $\mathcal{Y}_T$ converges weakly to a tight limit in $l^\infty(K_1, K_2, \ldots)$ for any sequence of compact sets $K_1 \subseteq K_2 \subseteq \ldots \subseteq \mathbb{R}^\nu$. Since a measure is determined by its finite-dimensional marginals, $\mathcal{Y}_T$ converges weakly to $\mathcal{Y}_\infty$ in $l^\infty(K_1, K_2, \ldots)$. □

Proof of Theorem 3.2. Lemma I.7 establishes the weak convergence of $\mathcal{Y}_T$ to $\mathcal{Y}_\infty$ in $l^\infty(K_1, K_2, \ldots)$ for any sequence of compact sets $K_1 \subseteq K_2 \subseteq \ldots \subseteq \mathbb{R}^\nu$. The deterministic limit of the penalty term follows from Fu and Knight (2000), Theorem 2.

For large T, $\begin{pmatrix} \hat{\boldsymbol{\phi}}_T^{(1)} \\ \frac{\boldsymbol{\varphi}}{\sqrt{T}} + \boldsymbol{\phi}_0 \end{pmatrix}$ belongs to $\mathcal{V}(\boldsymbol{\phi}_0) \times \mathcal{V}(\boldsymbol{\phi}_0)$ almost surely. Applying the mean value

theorem to $L_T - \widetilde{L}_T$ gives:

$$\left|\Lambda_T(\boldsymbol{\varphi}) - \widetilde{\Lambda}_T(\boldsymbol{\varphi})\right| \leqslant C \left|\left| \begin{pmatrix} \sqrt{T}\left(\hat{\boldsymbol{\phi}}_T^{(1)} - \boldsymbol{\phi}_0\right) \\ \boldsymbol{\varphi} \end{pmatrix} \right|\right| \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sup_{\mathcal{V}(\boldsymbol{\phi}_0)\times\mathcal{V}(\boldsymbol{\phi}_0)} \left|\left| \nabla l_t - \nabla \widetilde{l}_t \right|\right|.$$

Using $u_t = y_t - f_t$, the following holds:

$$\frac{||\nabla l_t - \nabla \widetilde{l}_t||}{2} = \left|\left| \frac{\widetilde{w}_t^2 u_t \nabla u_t - w_t^2 \widetilde{u}_t \nabla \widetilde{u}_t}{w_t^2 \widetilde{w}_t^2} \right|\right| = \left|\left| \frac{\left[\widetilde{w}_t^2 - w_t^2\right] u_t \nabla u_t - w_t^2 \left[\widetilde{u}_t \nabla \widetilde{u}_t - u_t \nabla u_t\right]}{w_t^2 \widetilde{w}_t^2} \right|\right|,$$

yielding:

$$\begin{aligned} \left|\left| \nabla l_t - \nabla \widetilde{l}_t \right|\right| \leqslant\ & C\left[d_t \sup|u_t| \sup||\nabla u_t|| + \sup|\widetilde{u}_t - u_t| \sup||\nabla u_t||\right] \\ & + C\left[\sup|\widetilde{u}_t| \sup||\nabla u_t - \nabla \widetilde{u}_t||\right], \end{aligned} \tag{A.10}$$

with

$$\begin{aligned} \sup_{\Phi\times\Phi}|u_t| &\leqslant 1 + |y_t| + \sup_{\Phi\times\Phi}|f_t| \\ \sup_{\Phi\times\Phi}||\nabla u_t|| &= \sup_{\Phi\times\Phi}||\nabla f_t|| \\ \sup|\widetilde{u}_t - u_t| &= a_t \\ \sup||\nabla u_t - \nabla \widetilde{u}_t|| &= b_t. \end{aligned}$$

Under Assumption A11, $a_t \to 0$ almost surely. For large t, $\sup|\widetilde{u}_t| \leqslant 1 + |y_t| + \sup_{\Phi\times\Phi}|f_t|$. Substituting:

$$\begin{aligned} \left|\left| \nabla l_t - \nabla \widetilde{l}_t \right|\right| \leqslant\ & C\left[d_t \sup||\nabla f_t|| \left(1 + |y_t| + \sup_{\Phi\times\Phi}|f_t|\right)\right] \\ & + C\left[a_t \sup||\nabla u_t|| + b_t\left(1 + |y_t| + \sup_{\Phi\times\Phi}|f_t|\right)\right]. \end{aligned} \tag{A.11}$$

Assumption A11 ensures that $\left|\Lambda_T(\boldsymbol{\varphi}) - \widetilde{\Lambda}_T(\boldsymbol{\varphi})\right| \to 0$ almost surely as $T \to \infty$, uniformly on any compact $K \subset \mathbb{R}^\nu$. Thus, $\Lambda_T - \widetilde{\Lambda}_T \xrightarrow[T\to\infty]{\text{a.s.}} 0$ in $l^\infty(K_1, K_2, \ldots)$ for any sequence of compact sets $K_1 \subseteq K_2 \subseteq \ldots \subseteq \mathbb{R}^\nu$. Under Assumptions A7 and A9, $\widetilde{\Lambda}_T$ is strictly convex and continuous on $\mathbb{R}^\nu$ for all T, as is Λ_∞, ensuring a unique $\arg\min$. Since $\widetilde{\Lambda}_T(\cdot) \xrightarrow[T\to+\infty]{\text{d}} \Lambda_\infty(\cdot)$ marginally, Corollary I.1 gives convergence in $l^\infty(K_1, K_2, \ldots)$.

Using Theorem 1.10.3 of Van der Vaart and Wellner (1996) and Skorokhod's representation theorem, Lemma 2.2 of Davis et al. (1992) establishes the asymptotics of $\sqrt{T}\left(\widehat{\boldsymbol{\phi}}_T - \boldsymbol{\phi}_0\right)$ as the $\arg\min$ of $\widetilde{\Lambda}_T$. This concludes the proof of Theorem 3.2. □

The following Lemma supports the proofs of Theorems 3.3 and 3.4.

Lemma I.8. *Under Assumptions B1-B2, the following holds:*

$$\sup_{\Phi\times\Theta\times\Theta}\tfrac{1}{T}\left|L_T^*\left(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}\right)-\widetilde{L}_T^*\left(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}\right)\right|\xrightarrow[T\to\infty]{a.s.}0.$$

Proof. To simplify, function parameters are omitted in this proof, with the assumption that $(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\vartheta})\in\Phi\times\Theta\times\Theta$. Define $u_t^*=g_t-\epsilon_t^2\left(\cdot\right)$ and $\widetilde{u}_t^*=\widetilde{g}_t-\tilde{\epsilon}_t^2\left(\cdot\right)$. Under Assumption B1, for sufficiently large t, the equivalent of the equation (A.1), and the following bounds are established:

$$\begin{aligned}
\left|u_t^*\right|&\leqslant 2\left(y_t^2+\mu_t^2\left(\boldsymbol{\phi}\right)\right)+\left|g_t\right|\leqslant C\left(1+y_t^2+\mu_t^2\left(\boldsymbol{\phi}\right)+\left|g_t\right|\right),\\
u_t^{*^2}&\leqslant 2\left(\epsilon_t^4\left(\boldsymbol{\phi}\right)+g_t^2\right)\leqslant 2\left(\epsilon_t^4\left(\boldsymbol{\phi}\right)+g_t^2\right)\leqslant 8\left(y_t^2+\mu_t^2\left(\boldsymbol{\phi}\right)\right)^2+2g_t^2\leqslant C\left(1+y_t^4+\mu_t^4\left(\boldsymbol{\phi}\right)+g_t^2\right),\\
\left|\widetilde{u}_t^*-u_t^*\right|&\leqslant\sup\left|\mu_t\left(\boldsymbol{\phi}\right)-\widetilde{\mu}_t\left(\boldsymbol{\phi}\right)\right|\left(\sup\left|\mu_t\left(\boldsymbol{\phi}\right)\right|+\sup\left|\widetilde{\mu}_t\left(\boldsymbol{\phi}\right)\right|\right)+\sup\left|\widetilde{g}_t-g_t\right|,
\end{aligned}$$

yielding:

$$\begin{aligned}
\left|l_t^*-\widetilde{l}_t^2\right|&\leqslant Cd_t^*\left(1+y_t^4+\mu_t^4\left(\boldsymbol{\phi}\right)+g_t^2\right)\\
&\quad+C\left(\sup\left|\mu_t\left(\boldsymbol{\phi}\right)-\widetilde{\mu}_t\left(\boldsymbol{\phi}\right)\right|\left(1+\sup\left|\mu_t\left(\boldsymbol{\phi}\right)\right|\right)+\sup\left|\widetilde{g}_t-g_t\right|\right)\left(1+y_t^2+\mu_t^2\left(\boldsymbol{\phi}\right)+\left|g_t\right|\right)\\
&\leqslant C\left[d_t^*\left(1+y_t^4+\mu_t^4\left(\boldsymbol{\phi}\right)+g_t^2\right)+a_t^*\left(1+y_t^2+\mu_t^2\left(\boldsymbol{\phi}\right)+\left|g_t\right|\right)\right].
\end{aligned}$$

The conclusion follows by Assumptions B1-B2, and Cesàro's lemma. □

Remark I.1. The uniform bound in (A.3) is the key argument in the proof of Lemma I.3. The same reasoning applies to both the location parameter estimator and the scale parameter estimator. In the proof of Theorem 3.3, it suffices to establish a uniform bound analogous to (A.3). Similarly, the inequality in (A.5) is central to the proof of Lemma I.5, and an equivalent result ensures the same conclusion for the scale parameter estimator.

Proof of Theorem 3.3. The proof of this Theorem follows the same reasoning as for Theorem 3.1. A sketch of proof is provided following these steps: First, expanding L_T^* as in (A.4), establishing the equivalents of Lemmas I.2 and I.3, then proving the convergence of the arg min of $C_T^*\left(\cdot\right):=\widetilde{Q}_T^*\left(\widehat{\boldsymbol{\phi}}_T^{(1)},\widehat{\boldsymbol{\theta}}_T^{(1)},\cdot\right)$ to the arg min of $C_\infty^*\left(\cdot\right):=Q_\infty^*\left(\boldsymbol{\phi}_0,\boldsymbol{\theta}_0,\cdot\right)$.

Lemma I.8 corresponds to Lemma I.1, and the counterparts of Lemmas I.2 and I.3 are derived using the following results and similar arguments as in the previous section. First, the expectation $\mathbb{E}\left[l_1^*\left(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}\right)\right]$ exists and is finite for $(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi})\in\mathcal{V}\left(\boldsymbol{\phi}_0\right),\mathcal{V}\left(\boldsymbol{\theta}_0\right)\times\overset{\circ}{\Theta}$. This is demonstrated as follows:

$$\begin{aligned}
\left|\tfrac{\epsilon_t^2(\boldsymbol{\phi})-g_t(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi})}{w_t^*}\right|&=\left|\tfrac{\epsilon_t^2(\boldsymbol{\phi})-\epsilon_t^2+\epsilon_t^2-\sigma_t^2+\sigma_t^2-g_t(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi})}{w_t^*}\right|=\left|\tfrac{(\mu_t-\mu_t(\boldsymbol{\phi}))(2\sigma_t\eta_t+\mu_t-\mu_t(\boldsymbol{\phi}))+\sigma_t^2\left(\eta_t^2-1\right)+\sigma_t^2-g_t(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi})}{w_t^*}\right|\\
&\leqslant C\tfrac{E_t(2\sigma_t|\eta_t|+E_t)+\sigma_t^2\left|\eta_t^2-1\right|+G_t}{w_t^*}.
\end{aligned}$$

By Assumptions B3-B4 and Hölder's inequality, the dominant term belongs to L^2. This establishes the equivalent of Lemma I.2. Second, the key point for the counterpart of Lemma I.3 is the uniform bound (A.3) on $||\nabla l_t^*(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi})||$ for $(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}) \in \mathcal{V}(\boldsymbol{\phi}_0), \mathcal{V}(\boldsymbol{\theta}_0) \times \overset{\circ}{\Theta}$. Specifically:

$$
\begin{aligned}
||\nabla l_t^*(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi})|| &= 2\left|\left|\frac{\left(\epsilon_t^2(\boldsymbol{\phi})-g_t(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi})\right)\left(-2\epsilon_t(\boldsymbol{\phi})\nabla_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}}\mu_t(\boldsymbol{\phi})-\nabla g_t(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi})\right)}{w_t^{*2}}\right|\right| \\
&\leqslant \frac{E_t(2\sigma_t|\eta_t|+E_t)+\sigma_t^2\left|\eta_t^2-1\right|+G_t}{w_t^*}\frac{2(E_t+\sigma_t|\eta_t|)E_t+G_t}{w_t^*}.
\end{aligned}
$$

Under Assumptions B3-B4, the dominant term belongs to L^2. Since the gradient is uniformly bounded by an L^1 process, it also belongs to L^1. Under these assumptions and Assumption B6, the same reasoning as for Lemma I.3 applies. Finally, the consistency follows by the same arguments as for Theorem 3.1. When $\boldsymbol{\iota}_\infty^* = \mathbf{0}$, the estimator $\widehat{\boldsymbol{\theta}}_T$ converges to the true parameter $\boldsymbol{\theta}_0$ since it satisfies the following equation:

$$
\mathbb{E}\left[\left(\frac{\partial l_t^*}{\partial \boldsymbol{\psi}}(\boldsymbol{\phi}_0,\boldsymbol{\theta}_0,\cdot)\right)_{\boldsymbol{\psi}=\boldsymbol{\theta}_0}\right] = \mathbb{E}\left[-2\frac{\sigma_t^2\left(\eta_t^2-1\right)}{w_t^{*2}}\left(\frac{\partial g_t}{\partial \boldsymbol{\psi}}(\boldsymbol{\phi}_0,\boldsymbol{\theta}_0,\cdot)\right)_{\boldsymbol{\psi}=\boldsymbol{\theta}_0}\right] = \mathbf{0}.
$$

□

Proof of Theorem 3.4. This Theorem's proof follows the structure of Theorem 3.2, considering $\begin{pmatrix}\boldsymbol{\phi}\\ \boldsymbol{\theta}\\ \boldsymbol{\psi}\end{pmatrix} \in \mathcal{V}(\boldsymbol{\phi}_0) \times \mathcal{V}(\boldsymbol{\phi}_0) \times \mathcal{V}(\boldsymbol{\phi}_0)$. With an expansion equivalent to (A.8), the Hessian term can be bounded by:

$$
\begin{aligned}
||\nabla^2 l_t^*|| \leqslant\ & 2\left|\left|\frac{P\left(2\epsilon_t(\boldsymbol{\phi})\nabla_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}}\mu_t(\boldsymbol{\phi})+\nabla g_t(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi})\right)}{w_t^{*2}}\right|\right| \\
&+ 2\left|\left|\frac{\left(\epsilon_t^2(\boldsymbol{\phi})-g_t(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi})\right)\left(2P\left(\nabla_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}}\mu_t(\boldsymbol{\phi})\right)-2\epsilon_t(\boldsymbol{\phi})\nabla^2_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}}\mu_t(\boldsymbol{\phi})-\nabla^2 g_t(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi})\right)}{w_t^{*2}}\right|\right|.
\end{aligned}
$$

Each term is uniformly bounded as follows:

$$
\begin{aligned}
\left|\left|\frac{2\epsilon_t(\boldsymbol{\phi})\nabla_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}}\mu_t(\boldsymbol{\phi})+\nabla g_t(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi})}{w_t^*}\right|\right| &\leqslant C\frac{\mathcal{E}_t^2+\sigma_t|\eta_t|\mathcal{E}_t+\mathcal{G}_t}{w_t^*}, \\
\left|\left|\frac{\epsilon_t^2(\boldsymbol{\phi})-g_t(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi})}{w_t^*}\right|\right| &\leqslant C\frac{(\mathcal{E}_t+\sigma_t|\eta_t|)\mathcal{E}_t+\sigma_t^2\left|\eta_t^2-1\right|+\mathcal{G}_t}{w_t^*}, \\
\left|\left|\frac{2P\left(\nabla_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}}\mu_t(\boldsymbol{\phi})\right)-2\epsilon_t(\boldsymbol{\phi})\nabla^2_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}}\mu_t(\boldsymbol{\phi})-\nabla^2 g_t(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi})}{w_t^*}\right|\right| &\leqslant C\frac{\mathcal{E}_t^2+(E_t+\sigma_t|\eta_t|)\zeta_t+\mathcal{H}_t}{w_t^*}.
\end{aligned}
$$

Under Assumptions B8-B9, these terms are bounded by L^2 processes. Thus, $\left|\left|\nabla^2 l_t^*\right|\right|$ is uniformly bounded by an L^2 process, giving the equivalent of (A.5). The equivalent of Lemma I.5 holds with the same arguments. The martingale Central Limit Theorem is also used in

this context. Define:

$$\boldsymbol{\mathcal{W}}_T^* = \begin{pmatrix} \frac{1}{\sqrt{T}} \nabla L_T^*(\boldsymbol{\phi}_0, \boldsymbol{\theta}_0, \boldsymbol{\theta}_0) \\ \sqrt{T}\left(\boldsymbol{\phi}_T^{(1)} - \boldsymbol{\phi}_0\right) \\ \sqrt{T}\left(\boldsymbol{\theta}_T^{(1)} - \boldsymbol{\theta}_0\right) \end{pmatrix},$$

and

$$\boldsymbol{U}_T^* = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \boldsymbol{\mathcal{U}}_t^* + o_{\mathbb{P}}(1) \quad \text{with} \quad \boldsymbol{\mathcal{U}}_t^* = \begin{pmatrix} \frac{-2\sigma_t^2\left(\eta_t^2 - 1\right)\left(2\sigma_t \eta_t \nabla_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}} \mu_t(\boldsymbol{\phi}_0) + \nabla g_t(\boldsymbol{\phi}_0, \boldsymbol{\theta}_0, \boldsymbol{\theta}_0)\right)}{w_t^{*2}} \\ \boldsymbol{\Delta}_t^* \gamma^*(\eta_t) \end{pmatrix}.$$

Under Assumptions B8-B9, the bound:

$$\left\| \frac{-2\sigma_t^2\left(\eta_t^2 - 1\right)\left(2\sigma_t \eta_t \nabla_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}} \mu_t(\boldsymbol{\phi}_0) + \nabla g_t(\boldsymbol{\phi}_0, \boldsymbol{\theta}_0, \boldsymbol{\theta}_0)\right)}{w_t^{*2}} \right\| \leqslant C \frac{\left|\eta_t^2 - 1\right| \sigma_t^2 (\sigma_t |\eta_t| \mathcal{E}_t + \mathcal{G}_t)}{w_t^{*2}},$$

belongs to L^2. Furthermore, since $\mathbb{E}\left[\eta_t^3\right] = 0$, $\boldsymbol{\mathcal{U}}_t^*$ is an L^2 martingale difference. By Billingsley's Central Limit Theorem:

$$\boldsymbol{U}_T^* \xrightarrow[T \to +\infty]{d} \begin{pmatrix} \boldsymbol{W}^* \\ \boldsymbol{Z}^* \end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \mathbb{V}\left[\boldsymbol{\mathcal{U}}_1^*\right]\right).$$

The variance terms are:

$$\begin{aligned}
I^*(\boldsymbol{\phi}_0, \boldsymbol{\theta}_0) &= \mathbb{V}\left[\frac{-2\left(\epsilon_1^2 - \sigma_1^2\right)\left(2\epsilon_1 \nabla_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}} \mu_1(\boldsymbol{\phi}_0) + \nabla g_1(\boldsymbol{\phi}_0, \boldsymbol{\theta}_0, \boldsymbol{\theta}_0)\right)}{w_1^{*2}} \right] \\
&= \mathbb{E}\left[P\left(\frac{-2\sigma_1^2\left(\eta_1^2 - 1\right)\left(2\sigma_1 \eta_1 \nabla_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}} \mu_1(\boldsymbol{\phi}_0) + \nabla g_1(\boldsymbol{\phi}_0, \boldsymbol{\theta}_0, \boldsymbol{\theta}_0)\right)}{w_1^{*2}} \right) \right], \\
R^*(\boldsymbol{\phi}_0, \boldsymbol{\theta}_0) &= Cov\left[\frac{-2\left(\epsilon_1^2 - \sigma_1^2\right)\left(2\epsilon_1 \nabla_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}} \mu_1(\boldsymbol{\phi}_0) + \nabla g_1(\boldsymbol{\phi}_0, \boldsymbol{\theta}_0, \boldsymbol{\theta}_0)\right)}{w_1^2}, \boldsymbol{\Delta}_1^* \gamma^*(\eta_1) \right] \\
&= \mathbb{E}\left[\frac{-2\sigma_1^2\left(\eta_1^2 - 1\right)\left(2\sigma_1 \eta_1 \nabla_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}} \mu_1(\boldsymbol{\phi}_0) + \nabla g_1(\boldsymbol{\phi}_0, \boldsymbol{\theta}_0, \boldsymbol{\theta}_0)\right) \gamma^{*'}(\eta_1) \boldsymbol{\Delta}_1^{*'}}{w_1^{*2}} \right] \\
\mathbb{V}\left[\boldsymbol{\Delta}_1^* \gamma^*(\eta_1)\right] &= \mathbb{E}\left[P\left(\boldsymbol{\Delta}_1^* \boldsymbol{\Gamma}^{*\frac{1}{2}} \right) \right] = \mathbb{V}\left[\boldsymbol{\Delta}_1^* \boldsymbol{\Gamma}^{*\frac{1}{2}} \right] = \boldsymbol{\Sigma}^*.
\end{aligned}$$

The counterpart for Lemma I.7 holds with the exact same arguments. Finally, for large t

under Assumption B11:

$$
\begin{aligned}
\sup_{\Phi\times\Phi}|u_t^*| &\leqslant 1+2|y_t|^2+2\sup_{\Phi}|\mu_t|^2+\sup_{\Phi\times\Theta\times\Theta}|g_t|\,,\\
\sup_{\Phi\times\Phi}||\nabla u_t^*|| &\leqslant 2\left(1+|y_t|+\sup_{\Phi}|\mu_t|\right)\sup_{\Phi}||\nabla\mu_t(\boldsymbol{\phi})||+\sup_{\Phi\times\Phi}||\nabla g_t||\,,\\
\sup|\widetilde{u}_t^*-u_t^*| &\leqslant \sup_{\Phi}|\mu_t-\widetilde{\mu}_t|\left(1+\sup_{\Phi}|\mu|_t+\sup_{\Phi}|\widetilde{\mu}_t|\right)+\sup|g_t-\widetilde{g}_t|\\
&\leqslant 2\sup_{\Phi}|\mu_t-\widetilde{\mu}_t|\left(1+\sup_{\Phi}|\mu|_t\right)+\sup|g_t-\widetilde{g}_t|\leqslant 2a_t^*,\\
\sup||\nabla u_t^*-\nabla\widetilde{u}_t^*|| &\leqslant 2\sup||\epsilon_t(\boldsymbol{\phi})\nabla\epsilon_t-\widetilde{\epsilon}_t(\boldsymbol{\phi})\nabla\widetilde{\epsilon}_t-(\nabla g_t-\nabla\widetilde{g}_t)||\\
&\leqslant 2\left[\sup|\mu_t-\widetilde{\mu}_t|\sup||\nabla\mu_t||+(1+|y_t|+\sup|\mu_t|)||\nabla\mu_t-\nabla\widetilde{\mu}_t||\right]+\sup||\nabla g_t-\nabla\widetilde{g}_t||\\
&\leqslant 2\left[\sup|\mu_t-\widetilde{\mu}_t|\sup||\nabla\mu_t||+(1+|y_t|+\sup|\mu_t|)||\nabla\mu_t-\nabla\widetilde{\mu}_t||+\sup||\nabla g_t-\nabla\widetilde{g}_t||\right]\\
&\leqslant 2b_t^*.
\end{aligned}
$$

The conclusion follows by the same arguments as Theorem 3.2. □

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.

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., 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 semiparametric multivariate models. Journal of Econometrics 205, 381–401.

Francq, C., Zakoïan, J.M., 2009. Bartlett's formula for a general class of nonlinear processes. Journal of Time Series Analysis 30, 449–465.

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

Frömmel, M., Han, X., Kratochvil, S., 2014. Modeling the daily electricity price volatility with realized measures. Energy Economics 44, 492–502.

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

Hannan, E.J., McDougall, A.J., 1988. Regression procedures for arma estimation. Journal of the American Statistical Association 83, 490–498.

Janczura, J., Weron, R., 2012. Efficient estimation of markov regime-switching models: An application to electricity spot prices. AStA Advances in Statistical Analysis 96, 385–407.

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

Ling, S., McAleer, M., 2010. A general asymptotic theory for time-series models. Statistica Neerlandica 64, 97–111.

Liu, H., Shi, J., 2013. Applying arma–garch approaches to forecasting short-term electricity prices. Energy Economics 37, 152–166.

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.

Pan, J., Wang, H., Yao, Q., 2007. Weighted least absolute deviations estimation for arma models with infinite variance. Econometric Theory 23, 852–879.

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.

Samitas, A., Armenatzoglou, A., 2014. Regression tree model versus markov regime switching: A comparison for electricity spot price modelling and forecasting. Operational Research 14, 319–340.

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

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

Van der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes: With Applications to Statistics. Springer.

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.