TI 2018-013/III
Tinbergen Institute Discussion Paper

Missing Observations in Observation-Driven Time Series Models

Francisco (F.) Blasques[1,2]
Paolo Gorgi[1,2]
Siem Jan (S.J.) Koopman[1,2,3]

[1] VU Amsterdam, The Netherlands
[2] Tinbergen Institute, The Netherlands
[3] CREATES, Aarhus University, Denmark

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

Contact: discussionpapers@tinbergen.nl

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 598 4580

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900

Missing Observations in Observation-Driven Time Series Models*

F. Blasques[a,b], P. Gorgi[a,b], and S. J. Koopman[a,b,c]

[a]Vrije Universiteit Amsterdam, The Netherlands

[b]Tinbergen Institute, The Netherlands

[c]CREATES, Aarhus University, Denmark

February 9, 2018

Abstract

We argue that existing methods for the treatment of missing observations in observation-driven models lead to inconsistent inference. We provide a formal proof of this inconsistency for a Gaussian model with time-varying mean. A Monte Carlo simulation study supports this theoretical result and illustrates how the inconsistency problem extends to score-driven and, more generally, to observation-driven models, which include well-known models for conditional volatility. To overcome the problem of inconsistent inference, we propose a novel estimation procedure based on indirect inference. This easy-to-implement method delivers consistent inference. The asymptotic properties are formally derived. Our proposed method shows a promising performance in both a Monte Carlo study and an empirical study concerning the measurement of conditional volatility from financial returns data.

Keywords: missing data, observation-driven models, consistency, indirect inference, volatility.

JEL codes: C22, C58

---

*Corresponding author: P. Gorgi. Email address: p.gorgi@vu.nl. Koopman acknowledges support from CREATES, Aarhus University, Denmark, funded by Danish National Research Foundation, (DNRF78).

1 Introduction

Missing observations are often encountered in empirical studies and are typically treated as a nuisance. They can occur for several reasons. For example, financial markets are closed during holidays and stock prices are not recorded during these days. However, the underlying values of the stocks may still be changing due to external events, even if no trading takes place. Missing observations can also be due to specific events such as computer failures, loss of records, and budget constraints. A more systematic pattern of missing observations typically occurs in the analysis of unequally spaced time series. A practical solution is to base the analysis on an equally spaced time series and to insert "missing observations" in time periods which are artificially introduced and for which no observations are available. The literature on the treatment of missing observations in statistical inference is extensive; see, for example, Pigott (2001) for a review and for many references on the subject.

Observation-driven time series models are widely employed to describe the time-variation in economic and financial time series. Such models feature time-varying parameters that are driven by past observed values of the time series. This is in contrast with parameter-driven models, where time-varying parameters are driven by stochastic processes with their own source of error (Cox, 1981). A notable example of an observation-driven model is the generalised autoregressive conditional heteroskedasticity (GARCH) model of Engle (1982) and Bollerslev (1986). Creal et al. (2013) and Harvey (2013) recently introduced the class of generalised autoregressive score (GAS) models that encompasses a wide range of observation-driven models. Among others, the GARCH model, the exponential GARCH model of Nelson (1991) and the Poisson autoregressive model of Davis et al. (2003) are special cases of GAS models. The peculiarity of GAS models is that time-varying parameters are driven by the score of the predictive likelihood function. The GAS approach has also delivered several novel specifications. Examples include the fat-tailed location model of Harvey and Luati (2014), the copula model of Salvatierra and Patton (2015) and the spatial model of Blasques et al. (2016b).

The handling of missing data in observation-driven models is widely discussed in empirical studies where these models are implemented. The most common approach employed by practitioners is to set the innovation of the observation-driven time varying parameter to zero, that is, to set the innovation to its conditional expectation. This solution originates in the context of score-driven or GAS models. In this case, the score innovation is set to zero when a missing observation occurs. Statistical inference is then simply based on the maximization of the resulting pseudo likelihood function. We refer to this method as the "setting-to-zero" strategy; see, for instance, Creal et al. (2014), Koopman et al. (2015), Lucas et al. (2016), Delle Monache et al. (2016) and Buccheri et al. (2017). The "setting-to-zero" approach is appealing from a practical point of view as it is easy-to-implement and computationally not demanding, given the closed form expression of the pseudo likelihood function. Furthermore, in analogy with the Kalman filter for parameter-driven models, this approach can be justified by some intuitive arguments; see Lucas et al. (2016). However, there is no formal discussion in the literature on the asymptotic properties of the method. In this paper we show that the "setting-to-zero" strategy delivers inconsistent inference. We formally prove the inconsistency of the pseudo maximum likelihood (pseudo ML) estimator for a GAS model defined for a Gaussian distribution with a time-varying mean. We perform simulation experiments that show how the inconsistency problem extends to other observation-driven models, including the GARCH model and the Student-t GAS conditional volatility (t-GAS) model of Creal et al. (2013) and Harvey (2013).

We emphasise that a straightforward solution to missing observations in observation-driven models is not available. This is in sharp contrast to the treatment of missing observations for parameter-driven models that poses no additional challenges from an estimation perspective: missing observations can be integrated out of the likelihood and exact maximum likelihood estimation can be performed. Most earlier contributions on inference with missing observations has focused on linear time series models. For example, it is well documented that for analyses based on the autoregressive moving average (ARMA) model with Gaussian disturbances, missing observations can be handled within the Kalman filter; see Harvey and Pierse (1984). However, we argue that no consistent procedure has been designed for observation-driven models, only except for a special case such as the estimator of Bondon and Bahamonde (2012) for the ARCH model. Our aim is to bridge this gap by developing an indirect inference method that delivers consistent inference in this context.

Our indirect inference method for the treatment of missing observations can be adopted for general classes of observation-driven models, but we focus on score-driven models for simplicity of exposition. The proposed method is easy-to-implement and delivers a general approach to statistical inference for observation-driven models with missing observations. The intuition behind using indirect inference in this setting is the ability to replicate missing observations in the simulation step of the indirect inference method. Therefore, under the assumption that the data are missing at random, we can exactly replicate the generating process of the observed time series. The auxiliary model we consider is the one obtained by setting the score innovation to zero. The asymptotic properties of the proposed estimator are formally derived. The finite sample accuracy is studied in a Monte Carlo simulation experiment. We show that the finite sample performance of the proposed estimator is comparable to that of the infeasible but efficient exact maximum likelihood estimator. Finally, we compare the performance of our estimator in an empirical application with financial data. In particular, we study the performance of alternative estimators in the context of a conditional volatility Student's t model applied to the daily S&P500 stock index.

The remainder of the paper is organised as follows. Section 2 presents the modeling setting and describes the "setting-to-zero" approach. Section 3 shows the inconsistency for the Gaussian GAS model with time varying mean and presents simulation-based evidence of inconsistent behaviour of the pseudo MLE in the context of a Gaussian score model of the conditional mean as well as other observation-driven models. Section 4 introduces the new estimator and establishes its asymptotic properties. Section 5 presents a Monte Carlo simulation study to evaluate the finite sample performance of the new estimator. Section 6 presents an empirical illustration with financial data that compares our estimator against available alternatives in the context of the conditional volatility Student's t model. Section 7 concludes.

2 Pseudo ML for score-driven models with missing observations

For clarity of exposition we focus the discussion on the class of GAS models. However, since the score can be regarded as the innovation of the time varying parameter, the arguments do not rely on a score-driven parameter update. It follows that the "setting-to-zero" method is applicable to the wider class of observation-driven models by rewriting the updating equation of the time-varying parameter as the sum of a memory term and a zero-mean innovation term. Therefore, all results discussed in this section and the following sections are applicable to observation-driven models in general.

We start our treatment for missing observations in observation-driven models by formally introducing the "setting-to-zero" method. Given a univariate time series $\{y_t\}_{t\in\mathbb{Z}}$, the class of score-driven models or GAS models of Creal et al. (2013) and Harvey (2013) can be represented as

$$y_t \sim p(y_t|f_t;\theta), \qquad f_{t+1} = \omega + \beta f_t + \alpha s_t, \qquad t \in \mathbb{Z}, \tag{1}$$

where $p(\cdot|f_t;\theta)$ is a conditional density function, f_t is the time-varying parameter that is specified as an autoregressive process with innovation s_t and θ is the vector containing all static parameters, including the coefficients ω, β and α. The score innovation s_t is specified as

$$s_t = S_t u_t, \qquad u_t = \partial \log p(y_t|f_t;\theta)/\partial f_t, \qquad t \in \mathbb{Z}, \tag{2}$$

where u_t is the score and S_t is a scaling factor that is typically taken as a transformation of the Fisher information; see Creal et al. (2013) for a more detailed discussion. The formulation is straightforward and simple. We consider some specific examples in the next section.

We assume that the time series $\{y_t\}_{t\in\mathbb{Z}}$ is subject to missing observations. In particular, in each time period $t \in \mathbb{Z}$, the random variable y_t is observed if $I_t = 1$ and not observed if $I_t = 0$. The process $\{I_t\}_{t\in\mathbb{Z}}$

is assumed to be a stationary and ergodic process such that $I_t = 1$ with probability π and $I_t = 0$ with probability $1 - \pi$. Finally, we assume that the observations are missing at random: the data generating process $\{y_t\}_{t\in\mathbb{Z}}$ is independent of $\{I_t\}_{t\in\mathbb{Z}}$.

The "setting-to-zero" method consists of setting the score innovation equal to zero $s_t = 0$ when the corresponding observation is missing, that is when $I_t = 0$. Hence the time varying parameter is available for all time points t and is recovered using the observed data only. The pseudo likelihood function is obtained by using this filtered time-varying parameter for computing the conditional log-density function. The estimation of the parameters in the model is carried out by maximising the resulting pseudo log-likelihood function. More formally, the "setting-to-zero" method entails the following. In a first step, the filtered parameter is obtained as

$$\hat{f}_{t+1}(\theta) = \omega + \beta \hat{f}_t(\theta) + \alpha I_t s_t, \tag{3}$$

where the filter recursion is initialised at a fixed point $\hat{f}_1(\theta) \in \mathbb{R}$. The corresponding average log-likelihood function is obtained by

$$\hat{L}_T(\theta) = T^{-1} \sum_{t=1}^{T} I_t \log p(y_t | \hat{f}_t(\theta); \theta), \tag{4}$$

where T is the time series sample length, including the missing entries. We refer to (4) as the pseudo log-likelihood function. Finally, the pseudo ML estimator is obtained as

$$\hat{\theta}_T = \arg\sup_{\theta\in\Theta} \hat{L}_T(\theta), \tag{5}$$

where Θ is a compact set that has the true parameter vector θ_0 in its interior.

The "setting-to-zero" approach has been considered by Creal et al. (2014), Koopman et al. (2015), Lucas et al. (2016), Delle Monache et al. (2016) and Buccheri et al. (2017), amongst others. It provides a practical way to treat missing observation in the GAS framework. By considering a multivariate score-driven model, Lucas et al. (2016) present some arguments to justify why this approach could be a reasonable way to handle missing observations. Their arguments are based on the Expectation-Maximization algorithm, however, the asymptotic properties of the resulting pseudo ML estimator are not discussed.

In the next section we argue that the "setting-to-zero" approach does not lead to the consistent estimation of θ_0. The problem is due to the fact that the pseudo likelihood (4) is not the actual likelihood of the observations and this leads to an asymptotic bias in the parameter estimates. In general, it is not clear how the true likelihood function for the observables can be obtained for observation-driven models. We do not know of theoretical results related to parameter estimation within score-driven models or, more generally, within observations-driven models, when we have missing observations. This is the case even for well known models such as the GARCH model. An exception is the very specific case of the least squares estimator of the parameter vector in the autoregressive conditional heteroskedasticity (ARCH) model that is explored by Bondon and Bahamonde (2012).

3 Inconsistency of the pseudo ML estimator with illustrations

We formally discuss the inconsistency of the pseudo ML estimator for a location, or local mean, score-driven model. We present a simulation experiment that provides further evidence of the inconsistency. We illustrate the inconsistency of the pseudo ML estimator as a specific problem for score-driven time series models. Two additional examples feature volatility models: the GARCH model and the conditional variance Student's t model.

3.1 Local mean model

Consider the data generating process for a conditional Gaussian distribution with a time varying mean as given by

$$y_t = \mu_t^o + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, \sigma_0^2), \qquad \mu_{t+1}^o = \omega_0 + \beta_0 \mu_t^o + \alpha_0 (y_t - \mu_t^o), \qquad t \in \mathbb{Z}, \tag{6}$$

where $\{\mu_t^o\}_{t\in\mathbb{Z}}$ is the time-varying mean process, $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is an independent and identically distributed (i.i.d) sequence of Gaussian random variables with mean zero and variance σ_0^2, and ω_0, β_0 and α_0 are static coefficients. Here we assume that the model is for a univariate series y_t. A multivariate version of this model is obtained by considering y_t, μ_t^o and ε_t as (equally sized) vectors; this model is considered in the illustration of Lucas et al. (2016). The local mean model (6) is a special case of the GAS model (1) – (2) with $p(y_t|f_t;\theta) = \mathrm{N}(\mu_t^o, \sigma^2)$ and $\mu_t^o \equiv f_t$. The scaled score function is simply the prediction error $s_t = y_t - \mu_t^o \equiv \varepsilon_t$. Since we can replace μ_τ^o by $y_\tau - \varepsilon_\tau$, for $\tau = t, t+1$, it follows almost immediately that the updating equation for μ_t^o in (6) implies an autoregressive moving average model, an $\mathrm{ARMA}(1,1)$ model, for y_t with autoregressive coefficient β_0 and moving average coefficient $\alpha_0 - \beta_0$. Therefore, $|\beta_0| < 1$ ensures the strict stationarity of the process (6).

For the developments given in this section, we simply assume that $\{I_t\}_{t\in\mathbb{Z}}$ is an i.i.d. sequence of Bernoulli random variables with success probability π. In case of model (6) for an observed sequence

$y_1, \ldots, y_T$, we obtain the filtered parameter $\hat{\mu}_t(\theta)$ recursively by

$$\hat{\mu}_{t+1}(\theta) = \omega + \beta \hat{\mu}_t(\theta) + \alpha I_t \left[y_t - \hat{\mu}_t(\theta) \right], \tag{7}$$

where $\hat{\mu}_1(\theta) \in \mathbb{R}$ is an arbitrary chosen initial condition for the filter. The pseudo log-likelihood function is then given by

$$\hat{L}_T(\theta) = -T^{-1} \frac{\sum_{t=1}^{T} I_t}{2} \log \sigma^2 - \frac{T^{-1}}{2} \sum_{t=1}^{T} I_t \left[y_t - \hat{\mu}_t(\theta) \right]^2 / \sigma^2.$$

Under the assumption that the coefficients ω_0, β_0 and α_0 are known, we can show that the estimator of σ_0^2 is inconsistent as follows. The estimator of σ_0^2 is

$$\hat{\sigma}_T^2 = (\sum_{t=1}^{T} I_t)^{-1} \sum_{t=1}^{T} I_t (y_t - \hat{\mu}_t(\theta_0))^2.$$

We start by noticing that $\hat{\mu}_t(\theta_0)$ does not converge to the true μ_t^o as $t \to \infty$ because μ_t^o depends on the infinite past of $\{y_t\}_{t \in \mathbb{Z}}$ and for any $\pi \in (0,1)$ there are infinitely many missing observations. Let $\{\mu_t(\theta_0)\}_{t \in \mathbb{Z}}$ denote the limit sequence to which $\hat{\mu}_t(\theta_0)$ converges as $t \to \infty$. We further have that $\hat{\sigma}_T^2$ converges in probability to $\mathbb{E}[(y_t - \mu_t(\theta_0))^2] = \sigma_0^2 + \mathbb{E}[(\mu_t^o - \mu_t(\theta_0))^2]$. The expectation $\mathbb{E}[(\mu_t^o - \mu_t(\theta_0))^2]$ is strictly larger than zero and therefore $\hat{\sigma}_T^2$ overestimates the variance σ_0^2. This inconsistency is not limited to the variance estimator. The next result shows the non-trivial fact that also the dependence coefficients β_0 and α_0 cannot be estimated consistently when the "setting-to-zero" method is applied for missing observations. Without loss of generality, we assume for the next result that ω_0 and σ_0^2 are known and equal to zero and one, respectively.

Theorem 3.1. *The pseudo ML estimator $\hat{\theta}_T$ defined in (5) for the local mean GAS model (6) is not consistent for some $\theta_0 := (\alpha_0, \beta_0)$ in the interior of some compact parameter space $\Theta \subset (0,1)^2$. In particular, there exists an $\epsilon > 0$ such that*

$$\mathbb{P}\left(\liminf_{T \to \infty} \|\hat{\theta}_T - \theta_0\| > \epsilon \right) = 1,$$

for some $\theta_0 \in \Theta$ and some $\pi \in (0,1)$.

Theorem 3.1 shows that the pseudo ML estimator of θ_0 in the GAS model (6) is inconsistent. This highlights a general problem for the treatment of missing observations in the context of GAS models.

Remark 3.1. *The proof of Theorem 3.1 is written for some $(\alpha_0, \beta_0) \in (0,1)^2$ but the result is considerably stronger as the same argument seems applicable to any $(\alpha_0, \beta_0) \in (0,1)^2$.*

Figure 1 presents the finite sample behaviour of the pseudo ML estimator for different sample sizes and $\pi = 0.75$. The simulations suggest that the estimator is indeed inconsistent. The sample distribution of the estimator is not collapsing towards the true parameter value. The results reveal the inconsistency for the estimators of α_0 and σ_0^2. In particular, we learn from Figure 1 that σ_0^2 is overestimated. This is coherent with the inconsistency argument presented above. The results in Figure 1 also provide some evidence that

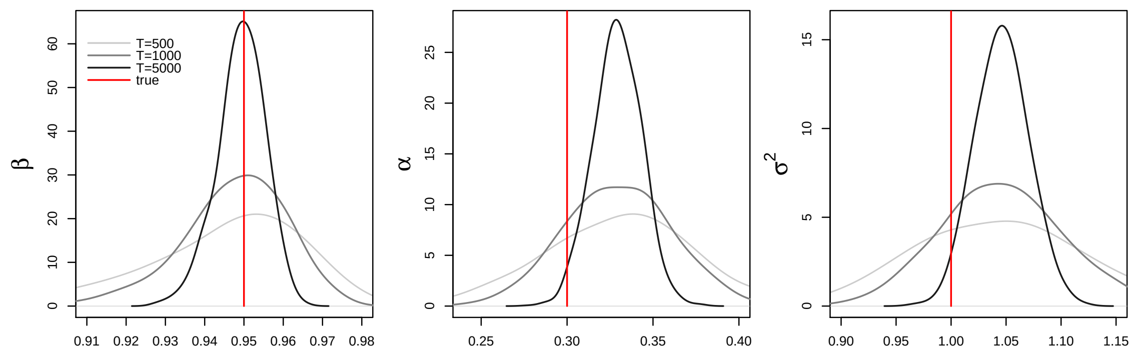

Figure 1: *Kernel distribution of the pseudo ML estimator for the Gaussian local mean model. The results are obtained from* $1,000$ *Monte Carlo replications. Different sample sizes are considered and* $\pi = 0.75$.

α_0 tends to be overestimated which is very intuitive. Assume that we have some sequence of consecutive missing observations, then the first observation after this sequence is highly informative about the current level of μ_t^o. Therefore, in order to approximate the true μ_t^o accurately, the parameter α should be large to give the new observation much weight. This intuition originates from the Kalman filter equations of the "local level model" with missing observations; see (Durbin and Koopman, 2012, section 2.7). After a sequence of missing values the filter is updated faster. In case of the GAS local mean model, the magnitude of the step is constant and therefore we obtain a positive bias.

3.2 GARCH model

The generalised autoregressive conditional heteroscedasticity (GARCH) model is specified for a univariate zero-mean time series y_t and is, in a slightly different fashion than usual, given by

$$y_t = \sqrt{h_t}\varepsilon_t, \qquad h_{t+1} = \omega_0 + \beta_0 h_t + \alpha_0(y_t^2 - h_t), \tag{8}$$

where $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is an i.i.d sequence of normal random variables with zero mean and unit variance, and ω_0, β_0 and α_0 are static coefficients. The GARCH model (8) is a special case of the GAS model (1) – (2) with $p(y_t|f_t;\theta) = \mathrm{N}(0, h_t)$ and $h_t \equiv f_t$. The scaled score function is simply the prediction error $s_t = y_t^2 - h_t$. Maximum likelihood estimation of the parameters in the GARCH model is the default option in most empirical work. However, except for a few special cases such as the ARCH model estimator of Bondon and Bahamonde (2012), parameter estimation with missing observations has not been widely discussed.

Figure 2 is indicative of how the “setting-to-zero” estimation method in Section 2 can be problematic. This becomes particularly clear by observing the sampling distribution of the pseudo ML estimator for the parameter α_0. The parameter α_0 tends to be overestimated. A similar intuitive explanation as for the GAS local mean model as discussed above applies here as well. The simulations strongly suggest that the estimators of the parameters ω_0 and β_0 are biased.

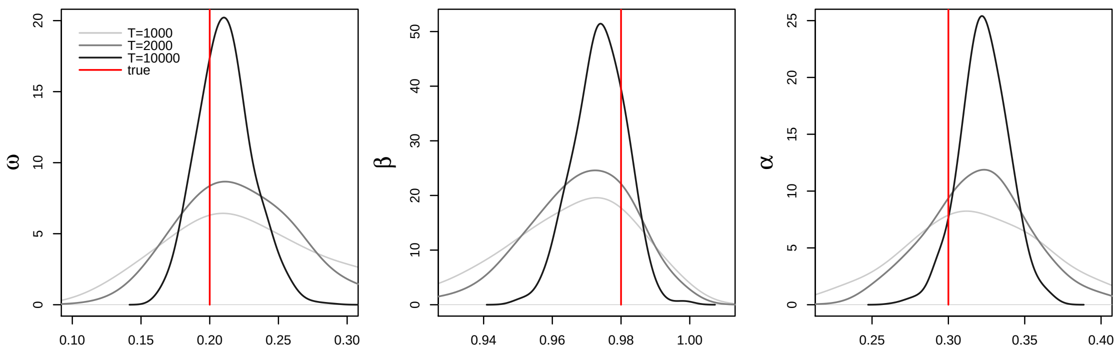

Figure 2: *Kernel distribution of the pseudo ML estimator for the Gaussian GARCH model. The results are obtained from* $1,000$ *Monte Carlo replications. Different sample sizes are considered and* $\pi = 0.75$.

3.3 Conditional volatility Student's t model

For our final illustration we consider the conditional volatility Student's t model of Creal et al. (2013) and Harvey (2013) for a univariate zero-mean time series y_t. The model has rapidly become a widely used framework for extracting volatility from time series of daily financial returns. It accounts for extreme observations by not only considering a fat-tailed distribution for the observations but also through a robust updating function of the conditional variance. The conditional volatility Student's t model is a special case of the GAS model (1) – (2) with $p(y_t|f_t;\theta) = t(0, h_t, \nu)$ and $h_t \equiv f_t$ where $t(0, h_t, \nu)$ is the Student's t density with mean zero, variance h_t and degrees of freedom ν. The resulting model becomes

$$y_t = \sqrt{h_t}\varepsilon_t, \qquad \varepsilon_t \sim t(0, 1, \nu_0), \qquad h_{t+1} = \omega_0 + \beta_0 h_t + \alpha_0 \left[\frac{(\nu_0 + 1)y_t^2}{(\nu_0 - 2) + y_t^2 h_t^{-1}} - h_t \right], \tag{9}$$

where $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is an i.i.d. sequence of Student's t distributed random variables and ω_0, β_0, α_0 and ν_0 are static coefficients.

The same simulation experiment as above has been carried to assess the inconsistency of the pseudo ML estimator in finite samples. Figure 3 presents the Kernel estimates of the distributions of the pseudo ML estimates. The distributions seem to converge towards values that are different from the corresponding true parameter values. This is particularly the case for the parameters α_0 and ν_0. For example, the parameter α_0 is clearly overestimated in the same way as for the Gaussian local mean and GARCH models. The parameter ν_0 appears to be underestimated by the pseudo-ML estimator.

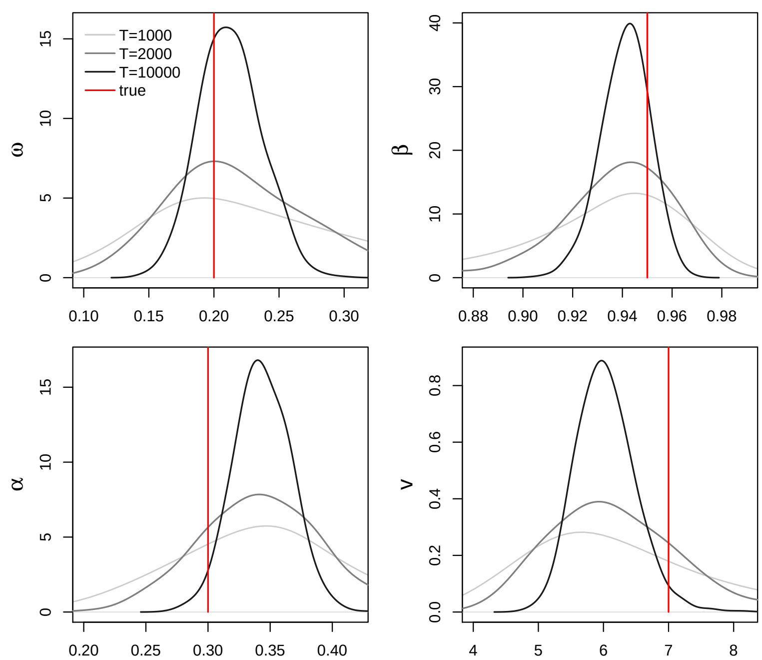

Figure 3: *Kernel distribution for conditional volatility Student's t model as in Figure 2.*

4 The indirect inference estimator and its properties

To overcome the inconsistency problem of the pseudo ML estimator for the GAS models with missing observations, we use a composite indirect inference estimator similar to the one proposed in Varin et al. (2011) and Gourieroux and Monfort (2017). This indirect inference estimator averages the log-likelihoods of the auxiliary models and delivers unbiased estimates of the parameter of interest. The idea is that we can generate data from our GAS model and introduce missing observations for those time periods where the actual observed data is missing. In this way, under the assumption of data missing at random, we can simulate from the true generating process with missing data. We do not require any further assumption on the missing value process $\{I_t\}_{t\in\mathbb{Z}}$ except that it needs to be stationary and ergodic with $\pi = \mathbb{E}I_t > 0$. These assumptions are imposed to rule out the possibility that from a certain time point onwards all data are missing. Once we have our simulated samples, we can proceed with indirect inference and we consider the pseudo ML estimator as auxiliary statistics. This approach provides consistent inference because the bias of the pseudo ML estimator is present both in the simulation and real data estimates.

More formally, we simulate S paths of length T from the GAS model in (1) and (2) for a given parameter value $\bar{\theta} \in \Theta$, which we denote with $\{y_{i,t}(\bar{\theta})\}_{t=1}^{T}$, $i = 1, \dots, S$. We treat $y_{i,t}$, for $i = 1, \dots, S$, as missing data if the corresponding real observation y_t is missing. For each simulated path we obtain the pseudo log-likelihood function as described in (4), which we denote with $\hat{L}_{i,T}(\theta, \bar{\theta})$. We then compute the average of these pseudo log-likelihoods as follows

$$\hat{L}_{S,T}(\theta, \bar{\theta}) = \frac{1}{S} \sum_{i=1}^{S} \hat{L}_{i,T}(\theta, \bar{\theta}),$$

and we obtain the maximiser of $\hat{L}_{S,T}(\theta, \bar{\theta})$ with respect to θ, that is,

$$\hat{\theta}_{S,T}(\bar{\theta}) = \arg\sup_{\theta \in \Theta} \hat{L}_{S,T}(\theta, \bar{\theta}).$$

The estimator $\hat{\theta}_{S,T}(\bar{\theta})$ is not consistent to $\bar{\theta}$ and in general it converges to a pseudo true parameter vector $\theta^*(\bar{\theta}) \neq \bar{\theta}$ as $T \to \infty$. Finally, we define the indirect inference estimator $\tilde{\theta}_{S,T}$ as the parameter value $\bar{\theta}$ the minimises a distance between the average pseudo ML estimator $\hat{\theta}_{S,T}(\bar{\theta})$ obtained from simulations and the point estimate $\hat{\theta}_T$ obtained from the real data, that is,

$$\tilde{\theta}_{S,T} = \arg\inf_{\bar{\theta} \in \Theta} \left\| \hat{\theta}_{S,T}(\bar{\theta}) - \hat{\theta}_T \right\|. \tag{10}$$

In practice, the minimization can be performed using the Newton-Raphson methods that are implemented in standard computer softwares for data analysis. The choice of the distance is irrelevant because we have exact identification and therefore there is a parameter value $\bar{\theta}$ that sets any distance to zero. We propose to average the log-likelihoods instead of the more common approach of averaging parameter estimates because this leads to a more efficient estimator from a computational point of view. Our main results for consistency and asymptotic normality of the indirect inference estimator (10) are presented next.

4.1 Consistency

Next we formulate sufficient conditions for the consistency and the asymptotic normality of the indirect inference estimator. Assumption 4.1 imposes that the GAS model has a compact parameter space and that it is correctly specified. In particular, the sample of observed data $\{y_t\}_{t=1}^{T}$ is generated by the GAS model in (1) and (2) with true parameter vector $\theta_0 \in \Theta$.

Assumption 4.1. *The observed data $\{y_t\}_{t=1}^{T}$ is a realised path from stochastic process $\{y_t\}_{t \in \mathbb{Z}}$ that satisfies the GAS's equations (1) and (2) at $\theta_0 \in \Theta$. Furthermore, the set Θ is compact.*

Assumption 4.2 states that the conditional density function $p(y|f;\theta)$ is continuous in all arguments. This is needed to ensure the continuity of the log-likelihood function.

Assumption 4.2. *The function* $(y, f, \theta) \mapsto p(y|f;\theta)$ *is continuous in* $\mathbb{R} \times \mathbb{R} \times \Theta$.

Assumption 4.3 requires the GAS model to generate stationary and ergodic data for any $\bar{\theta} \in \Theta$; see Blasques et al. (2014b) for primitive conditions that ensure the stationarity and ergodicity of GAS processes. This implies that also the observed data $\{y_t\}_{t=1}^{T}$ is stationary and ergodic since $\theta_0 \in \Theta$. Assumption 4.3 further requires the independence of observed and simulated data from the missing values process $\{I_t\}_{t\in\mathbb{Z}}$. Note that we do not specify a data generating process for the sequence $\{I_t\}_{t\in\mathbb{Z}}$. Instead, we take $\{I_t\}_{t\in\mathbb{Z}}$ as being an exogenous variable; see the discussion in Gourieroux et al. (1993).

Assumption 4.3. *The sequence* $\{y_{i,t}(\bar{\theta})\}_{t\in\mathbb{Z}}$ *is stationary and ergodic for every* $\bar{\theta} \in \Theta$*. Furthermore, the sequences* $\{y_{i,t}(\bar{\theta})\}_{t\in\mathbb{Z}}$*,* $i = 1, \ldots, S$*, and* $\{y_t\}_{t\in\mathbb{Z}}$ *are independent of the missing values process* $\{I_t\}_{t\in\mathbb{Z}}$*.*

Assumption 4.4 imposes conditions on the filtered sequence, as defined in (3), obtained from the simulated data. We let $\hat{f}_{i,t}(\theta, \bar{\theta})$ denote the filter in (3) evaluated at $\theta \in \Theta$ using a sample of data $\{y_{i,t}(\bar{\theta})\}_{t\in\mathbb{Z}}$, which is simulated under $\bar{\theta} \in \Theta$. In particular, the filter is required to be invertible and to converge exponentially fast and almost surely (e.a.s.)[1] to a strictly stationary and ergodic limit sequence, uniformly over $(\theta, \bar{\theta}) \in \Theta \times \Theta$. In practice, this assumption can be checked by means of Theorem 3.1 of Bougerol (1993); see also Straumann and Mikosch (2006) for an application of this theorem to GARCH-type models and Blasques et al. (2016a) for an application to GAS models.

Assumption 4.4. *The function* $(\theta, \bar{\theta}) \mapsto \hat{f}_{i,t}(\theta, \bar{\theta})$ *is a.s. continuous in* $\Theta \times \Theta$*. Furthermore, the filter* $\{\hat{f}_{i,t}(\theta, \bar{\theta})\}_{t\in\mathbb{N}}$ *converges e.a.s. and uniformly to a limit strictly stationary and ergodic sequence* $\{f_{i,t}(\theta, \bar{\theta})\}_{t\in\mathbb{Z}}$*,*

$$\sup_{\theta\in\Theta} \sup_{\bar{\theta}\in\Theta} |\hat{f}_{i,t}(\theta, \bar{\theta}) - f_{i,t}(\theta, \bar{\theta})| \xrightarrow{e.a.s.} 0 \quad as \quad t \to \infty,$$

for every initialization $\hat{f}_{i,1}(\theta, \bar{\theta}) \in \mathbb{R}$*.*

Assumption 4.5 gives moment conditions that are standard in the ML estimation literature of GAS models under misspecification; see Blasques et al. (2016a) for further details. The moment bounds on the pseudo log-likelihood and its derivative with respect to the time-varying parameter $f_{i,t}$, which we denote by ∇_f, allow the application of a uniform law of large numbers for the log-likelihood of the pseudo ML estimator. In particular, the uniform log moment condition on the derivative ∇_f is imposed to ensure that

[1]A sequence of positive random variables $\{x_t\}_{t\in\mathbb{Z}}$ is said to converge e.a.s. to zero if there is an $\gamma > 1$ such that $\gamma^t x_t \xrightarrow{a.s.} 0$ as t diverges, see Straumann and Mikosch (2006).

the log-likelihood evaluated at the filter $\hat{f}_{i,t}(\theta,\bar{\theta})$ converges to the limit log-likelihood evaluated at the limit filter $f_{i,t}(\theta,\bar{\theta})$. This moment condition is typically implied by the contraction condition that is used in GAS models to ensure that the filter is invertible.

Assumption 4.5. *The log-likelihood satisfies the following moment conditions:*

$$\mathbb{E}\sup_{\theta\in\Theta}\Big|\log p\big(y_{i,t}(\bar{\theta})|f_{i,t}(\theta,\bar{\theta});\theta\big)\Big| < \infty \quad \textit{for every} \quad \bar{\theta}\in\Theta,$$

$$\mathbb{E}\log^{+}\sup_{\theta\in\Theta}\sup_{f}\Big|\nabla_f \log p\big(y_{i,t}(\bar{\theta})|f;\theta\big)\Big| < \infty \quad \textit{for every } \bar{\theta}\in\Theta.$$

Assumption 4.6, together with the compactness of Θ and the continuity of the limit pseudo log-likelihood on Θ, ensures the identifiable uniqueness of the pseudo true parameter $\theta^*(\bar{\theta})$ for data obtained from any parameter vector $\bar{\theta}\in\Theta$.

Assumption 4.6. *The pseudo-true parameter $\theta^*(\bar{\theta})$ is the unique maximiser of the limit pseudo log-likelihood $L(\cdot,\bar{\theta})$ in Θ for every $\bar{\theta}\in\Theta$.*

Proposition 4.1 establishes the consistency of the *auxiliary* pseudo ML estimators $\hat{\theta}_{S,T}(\bar{\theta})$ and $\hat{\theta}_T$ as $T\to\infty$ to their respective pseudo true parameters $\theta^*(\bar{\theta})$ and $\theta^*(\theta_0)$ for any $\bar{\theta}\in\Theta$. The proof explores the argument laid down in Blasques et al. (2014a) and it is based on the classical results reviewed in White (1994).

Proposition 4.1. *Let Assumptions 4.1-4.6. hold. Then $\hat{\theta}_{S,T}(\bar{\theta}) \stackrel{a.s.}{\to} \theta^*(\bar{\theta})$ for every $\bar{\theta}\in\Theta$ and $\hat{\theta}_T \stackrel{a.s.}{\to} \theta^*(\theta_0)$ as $T\to\infty$.*

The consistency of our indirect inference estimator requires more than just the pointwise convergence of the auxiliary estimator $\hat{\theta}_{S,T}(\bar{\theta}) \stackrel{a.s.}{\to} \theta^*(\bar{\theta})$ for every every $\bar{\theta}\in\Theta$. Assumptions 4.7–4.10 impose sufficient conditions for the functional estimator $\hat{\theta}_{S,T}(\cdot)$ to converge a.s. and uniformly in Θ to the binding function $\theta^*(\cdot)$. Assumption 4.7 imposes that $p(y|f;\theta)$ is smooth in (y,f,θ) and that the filter $\hat{f}_{i,t}(\theta,\bar{\theta})$ is smooth in θ. These additional differentiability requirements allow us to work with the score and Hessian of the log-likelihood to establish the uniform convergence of our auxiliary estimator.

Assumption 4.7. *The function $(y,f,\theta)\mapsto p(y|f;\theta)$ is 3 times continuously differentiable in $\mathbb{R}\times\mathbb{R}\times\Theta$ and $\theta\mapsto\hat{f}_{i,t}(\theta,\bar{\theta})$ is a.s. 2 times continuously differentiable in Θ for any $\bar{\theta}\in\Theta$.*

Assumption 4.8 ensures that the filter derivative processes are invertible and converge exponentially fast to their respective stationary and ergodic limits. This is a standard regularity condition which is designed to ensure that the score and Hessian satisfy laws of large numbers and central limit theorems; see Potscher and Prucha (1997) for further details. We adopt the following notation: $\nabla_\theta^k \hat{f}_{i,t}(\theta,\bar{\theta})$ is the kth derivative of $\hat{f}_{i,t}(\theta,\bar{\theta})$ w.r.t. θ; $\nabla_\theta^{(0:k)} \hat{f}_{i,t}(\theta,\bar{\theta})$ denotes the vector containing the filter $\hat{f}_{i,t}(\theta,\bar{\theta})$ and its derivatives of up to kth order.

Assumption 4.8. *The derivative filter $\nabla_\theta^k \hat{f}_{i,t}(\theta,\bar{\theta})$ converges e.a.s. and uniformly to a stationary and ergodic sequence $\{\nabla_\theta^k f_{i,t}(\theta,\bar{\theta})\}_{t\in\mathbb{Z}}$ as $t\to\infty$ for $k=0,1,2$, that is,*

$$\sup_{\theta\in\Theta}\sup_{\bar{\theta}\in\Theta}\left\|\nabla_\theta^{(0:2)}\hat{f}_{i,t}(\theta,\bar{\theta})-\nabla_\theta^{(0:2)} f_{i,t}(\theta,\bar{\theta})\right\| \overset{e.a.s.}{\to} 0 \quad \textit{as} \quad t\to\infty.$$

Assumption 4.9 provides additional moment bounds on the score and the Hessian of the log-likelihood functions. Additional logarithmic moments are imposed on the derivative of the score and the Hessian with respect to the filter $f_{i,t}(\theta,\bar{\theta})$. These moment conditions ensure that we can apply uniform laws of large numbers to the score and the Hessian. We let $\nabla_\theta \log p$ denote the score function, $\nabla^2_{\theta f}\log p$ denote the derivative of the score with respect to the filter $f_{i,t}(\theta,\bar{\theta})$, $\nabla^2_{\theta\theta}\log p$ denote the Hessian, and $\nabla^3_{\theta\theta f}\log p$ denote the derivative of the Hessian with respect to the filter $f_{i,t}(\theta,\bar{\theta})$.

Assumption 4.9. *The score and Hessian of the log-likelihood have one bounded moment,*

$$\mathbb{E}\sup_{\bar{\theta}\in\Theta}\sup_{\theta\in\Theta}\left\|\nabla_\theta \log p\big(y_{i,t}(\bar{\theta})|f_{i,t}(\theta,\bar{\theta});\theta\big)\right\|<\infty;$$

$$\mathbb{E}\sup_{\bar{\theta}\in\Theta}\sup_{\theta\in\Theta}\left\|\nabla^2_{\theta\theta} \log p\big(y_{i,t}(\bar{\theta})|f_{i,t}(\theta,\bar{\theta});\theta\big)\right\|<\infty.$$

The derivatives of the score and Hessian have a logarithmic bounded moment,

$$\mathbb{E}\log^+\sup_{\bar{\theta}\in\Theta}\sup_{\theta\in\Theta}\sup_{f^{(0:1)}}\left\|\nabla^2_{\theta f^{(0:1)}} \log p\big(y_{i,t}(\bar{\theta})|f_{i,t}(\theta,\bar{\theta});\theta\big)\right\|<\infty;$$

$$\mathbb{E}\log^+\sup_{\bar{\theta}\in\Theta}\sup_{\theta\in\Theta}\sup_{f^{(0:2)}}\left\|\nabla^3_{\theta\theta f^{(0:2)}} \log p\big(y_{i,t}(\bar{\theta})|f_{i,t}(\theta,\bar{\theta});\theta\big)\right\|<\infty.$$

Assumption 4.10 ensures that the Hessian converges to a non-singular limit.

Assumption 4.10. *The Hessian matrix $\mathbb{E}\nabla^2_{\theta\theta}\log p\big(y_{i,t}(\bar{\theta})|f_{i,t}(\theta,\bar{\theta});\theta\big)$ is non-singular for every $\theta,\bar{\theta}\in\Theta$.*

Assumption 4.7 states the fundamental identification condition for the indirect inference estimator. The assumption that the so-called *binding function* θ^* is injective is standard for indirect inference estimators but often difficult to verify; see Gourieroux et al. (1993) for a discussion.

Assumption 4.11. *The binding function $\bar{\theta}\mapsto\theta^*(\bar{\theta})$ is continuous and injective in Θ.*

Theorem 4.1 delivers the consistency of the indirect inference estimator. The proof is different from that found in Gourieroux et al. (1993) as we average auxiliary log-likelihoods instead of averaging auxiliary estimators. This makes our estimator computationally faster when S is larger than 1. Additionally, our proof handles the convergence of the filter $\{\hat{f}_{i,t}(\theta, \bar{\theta})\}$.

Theorem 4.1. *Let Assumptions 4.1-4.11 hold. Then the indirect inference estimator is strongly consistent:* $\tilde{\theta}_{S,T} \overset{a.s.}{\to} \theta_0$ *as* $T \to \infty$.

4.2 Asymptotic normality

Asymptotic normality of the indirect inference estimator is derived from the asymptotic normality of the auxiliary pseudo ML estimators. The additional assumptions are designed to ensure that the auxiliary pseudo ML estimators $\hat{\theta}_{S,T}(\bar{\theta})$ and $\hat{\theta}_T$ of the GAS model are asymptotically normally distributed. Given that the auxiliary model is misspecified, the score of the log-likelihood will generally fail to be a martingale difference sequence. In any case, the score can still satisfy a central limit theorem for sequences that are *near epoch dependent* (NED) on an α-mixing sequence; see Wooldridge (1986), Gallant and White (1988), and Theorem 10.2 in Potscher and Prucha (1997). Assumptions 4.12 and 4.13 impose the NED property on the data, the filter, and the filter's derivative; see Blasques et al. (2014a) for conditions that ensure the NED property for score models.

Assumption 4.12. *The sequence* $\{y_{i,t}(\theta_0)\}$ *is NED of size* -1 *of an* α*-mixing sequence of size* $-2r/(r-1)$ *for some* $r > 2$.

Assumption 4.13. *The filter* $\{f_{i,t}(\theta^*(\theta_0), \theta_0)\}$ *and its derivative* $\{\nabla_\theta f_{i,t}(\theta^*(\theta_0), \theta_0)\}$ *are both NED of size* -1 *of an* α*-mixing sequence of size* $-2r/(r-1)$ *for some* $r > 2$.

Assumption 4.14 imposes additional smoothness and bounded moments for the score. The Lipschitz smoothness assumption ensures that the score inherits the NED property from the data and the score. The $2+\delta$ bounded moments ensure that the score satisfies a central limit theorem.

Assumption 4.14. *The score* $\nabla_\theta \log p$ *is Lipschitz continuous in* $y_{i,t}$, $f_{i,t}$ *and* $\nabla_\theta f_{i,t}$. *Furthermore, the following moment is finite*

$$\mathbb{E}\Big|\nabla_\theta \log p\Big(y_{i,t}(\theta_0)\big|f_{i,t}(\theta^*(\theta_0), \theta_0), \theta^*(\theta_0)\Big)\Big|^{2+\delta} < \infty \quad \textit{for some } \delta > 0.$$

Proposition 4.2 delivers the asymptotic normality of the auxiliary pseudo ML estimators.

Proposition 4.2. *Let Assumptions 4.1-4.14. hold. Then*

$$\sqrt{T}\Big(\hat{\theta}_T - \theta^*(\theta_0)\Big) \xrightarrow{d} N\Big(0, \Omega^*(\theta_0)^{-1}\Sigma^*(\theta_0)\Omega^*(\theta_0)^{-1}\Big),$$

$$\text{and} \quad \sqrt{T}\Big(\hat{\theta}_{S,T}(\theta_0) - \theta^*(\theta_0)\Big) \xrightarrow{d} N\Big(0, \Omega^*(\theta_0)^{-1}\Sigma^*_S(\theta_0)\Omega^*(\theta_0)^{-1}\Big) \quad \text{as} \quad T \to \infty,$$

where $\Omega^*(\theta_0) = \mathbb{E}\nabla^2_{\theta\theta} \log p\big(y_{i,t}(\theta_0)|f_{i,t}(\theta^*(\theta_0), \theta_0); \theta^*(\theta_0)\big)$ *and* $\Sigma^*_S(\theta_0) = \frac{1}{S}\Sigma^*(\theta_0) + \frac{S-1}{S}K^*(\theta_0)$*, with*

$$\Sigma^*(\theta_0) = \lim_{T\to\infty} \mathbb{V}ar\left(\frac{1}{\sqrt{T}}\sum_{t=2}^{T} \nabla_\theta \log p\big(y_{i,t}(\theta_0)\big|f_{i,t}\big(\theta^*(\theta_0), \theta_0\big); \theta^*(\theta_0)\big)\right) \quad \text{and}$$

$$K^*(\theta_0) = \lim_{T\to\infty} \mathbb{C}ov\left(\frac{1}{\sqrt{T}}\sum_{t=2}^{T} \nabla_\theta \log p\big(y_{i,t}(\theta_0)\big|f_{i,t}\big(\theta^*(\theta_0), \theta_0\big); \theta^*(\theta_0)\big)\ , \right.$$
$$\left. \frac{1}{\sqrt{T}}\sum_{t=2}^{T} \nabla_\theta \log p\big(y_{j,t}(\theta_0)\big|f_{j,t}\big(\theta^*(\theta_0), \theta_0\big); \theta^*(\theta_0)\big)\right) \quad \text{for some } i \neq j.$$

Finally, we obtain the asymptotic normality of the indirect inference estimator $\tilde{\theta}_{S,T}$ as $T \to \infty$. Assumption 4.15 imposes the continuous differentiability of θ^*.

Assumption 4.15. *The binding function* $\bar{\theta} \mapsto \theta^*(\bar{\theta})$ *is continuously differentiable in* Θ.

Theorem 4.2 delivers the desired asymptotic normality of the indirect inference estimator as proven in Gourieroux et al. (1993). As usual, the asymptotic variance is smaller for larger S. The expression for the asymptotic variance is simpler than usual due to the fact that the structural and auxiliary parameter spaces are the same.

Theorem 4.2. *Let Assumptions 4.1-4.15. hold. Then*

$$\sqrt{T}\Big(\tilde{\theta}_{S,T} - \theta_0\Big) \xrightarrow{d} N(0, W_S) \quad \text{as} \quad T \to \infty,$$

where

$$W_S := \left(1 + \frac{1}{S}\right)\left[\frac{\partial\theta^*(\theta_0)}{\partial\theta^\top}\right]^{-1} V(\theta_0) \left[\frac{\partial\theta^*(\theta_0)}{\partial\theta}^\top\right]^{-1}$$

where $V(\theta_0)$ *denotes the asymptotic variance* $V(\theta_0) := \Omega^*(\theta_0)^{-1}(\Sigma^*(\theta_0) - K^*(\theta_0))\Omega^*(\theta_0)^{-1}$.

4.3 Finite sample Monte Carlo study

We evaluate the finite sample behavior of the indirect inference estimator (10) through a simulation study. More specifically, we consider the indirect inference estimator of the GARCH model (8) we employ the same simulation setting as in Section 3.2. Figure 4 displays the distribution of the estimator for different sample sizes. We can see that the distributions are centered around the true parameter values. This suggests that the indirect inference estimator can successfully eliminate the bias caused by the missing data; see Figure 2 for a comparison with the pseudo-ML estimator.

We observe clearly that the distributions are collapsing towards the true parameter values as the sample size increases. Furthermore, the distributions tend to become more symmetric and with a more normal shape for larger sample sizes. These results confirm strongly the reliability of the indirect inference estimator and the validity of its asymptotic properties. Similar findings are obtained for other models but are not reported here for space considerations.

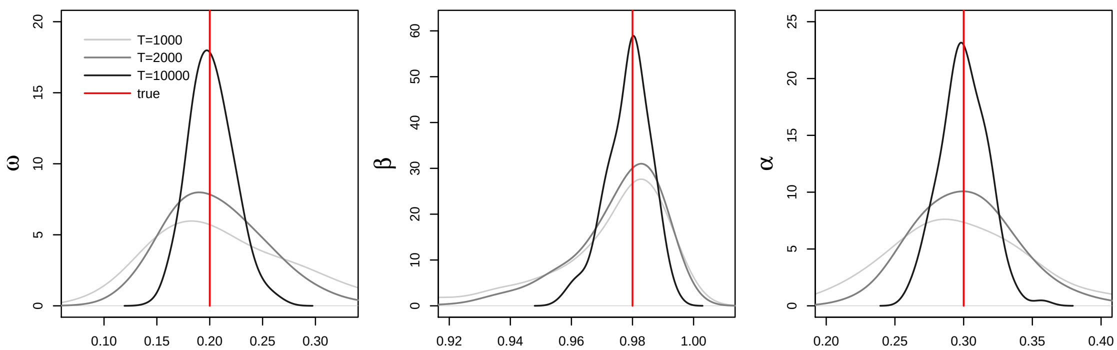

Figure 4: *Distribution of the indirect inference estimator for the GARCH model. The results are obtained from* 500 *Monte Carlo replications and* $S = 10$. *Different sample sizes are considered and* $\pi = 0.75$.

5 Monte Carlo comparison among different estimators

In this section, we present the results of a Monte Carlo experiment to evaluate the finite sample performance of the indirect inference estimator compared to the exact ML estimator and the pseudo ML estimator. We consider the Gaussian GAS local mean model in (6). This choice is due to the fact that only for this model the exact ML estimator is available when we have missing data. The GAS model (6), with $\omega_0 = 0$, can be rewritten as a Gaussian ARMA(1,1) model of the form

$$y_t = \beta_0 y_{t-1} + \phi_0 \varepsilon_{t-1} + \varepsilon_t,$$

where $\phi_0 = \alpha_0 - \beta_0$. Therefore, we can use the Kalman filter to consistently estimate the model. In presence of missing observations, the consistency of the ML estimator based on the Kalman filter has been formally discussed in Jones (1980) and Kohn and Ansley (1986). Note that this comparison is possible only for this specific model because in general there is not a clear way to obtain the exact likelihood function for

GAS models with missing data. However, it is useful to see how our indirect inference estimator performs compared to exact ML in this setting.

Table 1 reports a finite sample comparison among the indirect inference estimator, the exact ML estimator and the inconsistent pseudo ML estimator in terms of relative bias and mean squared error (MSE). The relative bias is the bias relative to the true parameter vale, which is computed as $(\hat{\theta}_T - \theta_0)/\theta_0$. The results are presented for sample sizes of $T = 500$, $1,000$ and $2,000$ observations. The missing observations are generated from independent Bernoulli random variables where π is the probability of observing y_t and $1 - \pi$ is the probability of having a missing observation. We study the behaviour of the estimators for several values of π, namely $\pi = 0.4$, 0.6, 0.8 and 1. The latter case corresponds naturally to a sample without any missing observations. The values reported are for a true parameter θ_0 with $\beta_0 = 0.95$, $\alpha_0 = 0.3$, and $\sigma_0^2 = 1.0$. The parameter ω_0 is assumed to be known and it is set equal to zero.

Table 1 reveals very clearly that the bias of the pseudo ML estimator does not converge to zero when the sample size increases. This is particularly clear for small values of π. The bias is also relevant in relative terms. For instance, the parameter α has a bias of about 30% and σ^2 has a bias of about 15% when $\pi = 0.4$. Instead, the indirect inference estimator and the exact ML estimator have a negligible bias, even in relative terms. In terms of the MSE, we find that the impact of the bias is more relevant for larger sample sizes. These gains for the indirect inference estimator and the exact ML estimator over the pseudo ML are stronger for the larger samples. Finally, the indirect inference estimator shows comparable performances when compared to the exact ML estimator. In particular, the MSE of these two estimators are very close for all the configurations considered in the experiment. This emphasises the accuracy of our proposed indirect inference estimator. Indeed the advantage of the indirect inference estimator is that it can be applied to GAS models in general while the exact ML estimator is only available in this particular setting.

Figure 5 presents the bias of the pseudo ML, exact ML and our proposed indirect inference estimator. The plots show bias with respect to β, α and σ^2 over a range of values of π. The advantage of our new estimator becomes more relevant for small π, that is when the fraction of missing values is large. This seems to be especially true for the estimation of the parameters α_0 and σ_0^2. Furthermore, Figure 5 further confirms how the exact ML estimator and the indirect inference estimator have a very similar performance.

6 An empirical experiment for the S&P500 daily returns time series

To illustrate how the inconsistency problem of the pseudo ML estimator can affect inference in an empirical study and how the use of the indirect inference estimator alleviates the problem, we analyse daily log-differences of the Standard and Poor's 500 stock index (S&P500) from January 2000 to December 2016.

			$\pi = 0.40$			$\pi = 0.60$		
			β	α	σ^2	β	α	σ^2
$T = 500$	Rel. bias	**PML**	-0.009	0.321	0.164	-0.006	0.172	0.078
		ML	-0.012	0.001	-0.006	-0.009	-0.012	-0.009
		II	-0.009	0.016	-0.008	-0.006	0.001	-0.012
	MSE	**PML**	0.028	0.120	0.204	0.023	0.075	0.121
		ML	0.029	0.060	0.109	0.023	0.047	0.087
		II	0.025	0.064	0.118	0.020	0.051	0.091
$T = 1000$	Rel. bias	**PML**	-0.004	0.320	0.166	-0.004	0.174	0.084
		ML	-0.005	-0.007	-0.001	-0.005	-0.011	-0.002
		II	-0.004	0.009	-0.005	-0.004	-0.002	-0.005
	MSE	**PML**	0.018	0.108	0.187	0.016	0.065	0.105
		ML	0.018	0.042	0.078	0.016	0.033	0.058
		II	0.014	0.045	0.084	0.013	0.035	0.062
$T = 2000$	Rel. bias	**PML**	-0.002	0.316	0.168	-0.001	0.173	0.084
		ML	-0.002	-0.016	0.003	-0.002	-0.014	-0.002
		II	-0.001	-0.000	-0.003	-0.001	-0.003	-0.006
	MSE	**PML**	0.011	0.100	0.177	0.010	0.058	0.095
		ML	0.011	0.028	0.052	0.010	0.023	0.042
		II	0.008	0.029	0.055	0.007	0.024	0.045
			$\pi = 0.80$			$\pi = 1.00$		
			β	α	σ^2	β	α	σ^2
$T = 500$	Rel. bias	**PML**	-0.006	0.072	0.029	-0.006	-0.003	-0.005
		ML	-0.009	-0.013	-0.009	-0.009	-0.009	-0.008
		II	-0.007	-0.004	-0.010	-0.006	-0.007	-0.007
	MSE	**PML**	0.022	0.049	0.080	0.020	0.037	0.064
		ML	0.022	0.041	0.073	0.021	0.037	0.065
		II	0.019	0.044	0.076	0.018	0.039	0.067
$T = 1000$	Rel. bias	**PML**	-0.004	0.074	0.034	-0.003	-0.001	-0.002
		ML	-0.005	-0.009	-0.003	-0.005	-0.004	-0.004
		II	-0.004	-0.005	-0.004	-0.003	-0.005	-0.003
	MSE	**PML**	0.015	0.037	0.061	0.014	0.025	0.044
		ML	0.015	0.028	0.049	0.014	0.025	0.044
		II	0.012	0.029	0.051	0.011	0.026	0.046
$T = 2000$	Rel. bias	**PML**	-0.001	0.072	0.033	-0.001	-0.004	-0.003
		ML	-0.002	-0.010	-0.003	-0.002	-0.006	-0.004
		II	-0.001	-0.006	-0.004	-0.001	-0.007	-0.004
	MSE	**PML**	0.010	0.030	0.049	0.009	0.018	0.032
		ML	0.010	0.020	0.036	0.009	0.018	0.032
		II	0.007	0.021	0.037	0.007	0.019	0.033

Table 1: *Simulation results for the pseudo ML (PML), indirect inference (II) and exact maximum likelihood (ML). We report relative bias (Rel. bias) and mean squared error (MSE). The results are obtained from 500 Monte Carlo replications with $S = 10$. The true parameter vector is $\theta_0 = (0.95, 0.3, 1)^\top$.*

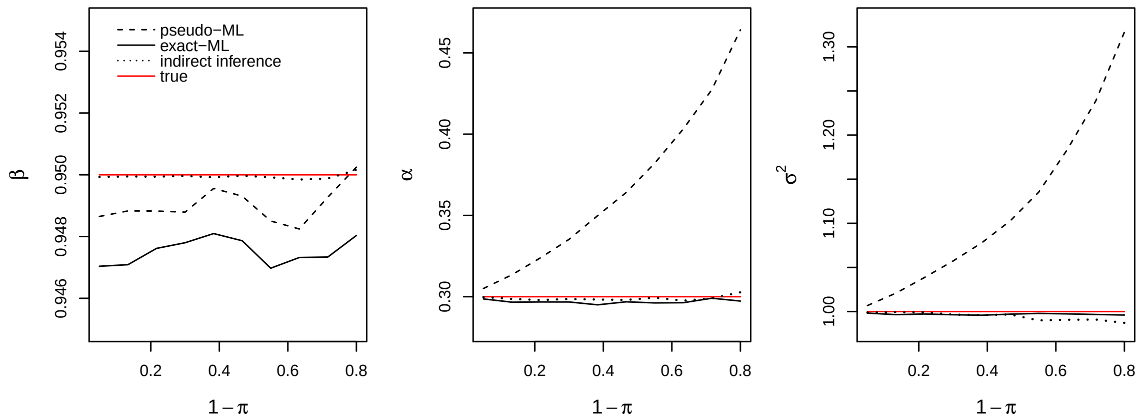

Figure 5: *Median of the sampling distribution of the pseudo ML estimator, the exact ML estimator and the indirect inference estimator for different values of the probability* π. *The results are obtained from 500 Monte Carlo replications and the sample size of the simulated series is* $T = 1,000$. *For, the indirect inference estimator* $S = 10$ *is considered.*

We adopt the conditional volatility Student's t model (9) and carry out the pseudo ML and the indirect inference methods for parameter estimation. The method of exact maximum likelihood is not feasible for this model when there are missing observations. The occurrence of missing observations is widespread in financial returns data because markets are regularly closed during the year. On these closure days there are no financial transactions and hence we do not observe price changes. However, the underlying price of the asset may still be changing during these days; see, for example, the discussions in Bondon and Bahamonde (2012).

In our empirical experiment, we aim to investigate the behaviour of the two estimators when we have a growing number of missing observations in the sample. We first estimate the model using all available data in the sample. Then we artificially remove observations from the sample by drawing a Bernoulli random variable with success probability π for each observation. If the outcome of the draw is zero, then we consider the corresponding observation as missing. For this resulting sample with missing data, we estimate the parameters in the model using the two methods that account for the missing observations. We repeat this procedure 100 times for a given value of π. In this way, for a given value of π, we obtain the distribution of the estimator. We use the full sample estimates as the benchmark to evaluate the performance of the estimates based on the samples with missing data. We consider a range of different π values and repeat the exercise as described. Clearly, this experiment is conditional on the full sample of observed data. The variability of the estimates with missing data only originates from the randomness of the observations that are removed and treated as missing through the Bernoulli draws.

Figures 6 and 7 report the results of this experiment. In particular, the figures show the bias distribution of the estimators compared to the full sample estimators for different values of π. Figure 6 clearly reveals that the pseudo ML estimates have a strong bias for the parameters α and ν. This is coherent with the

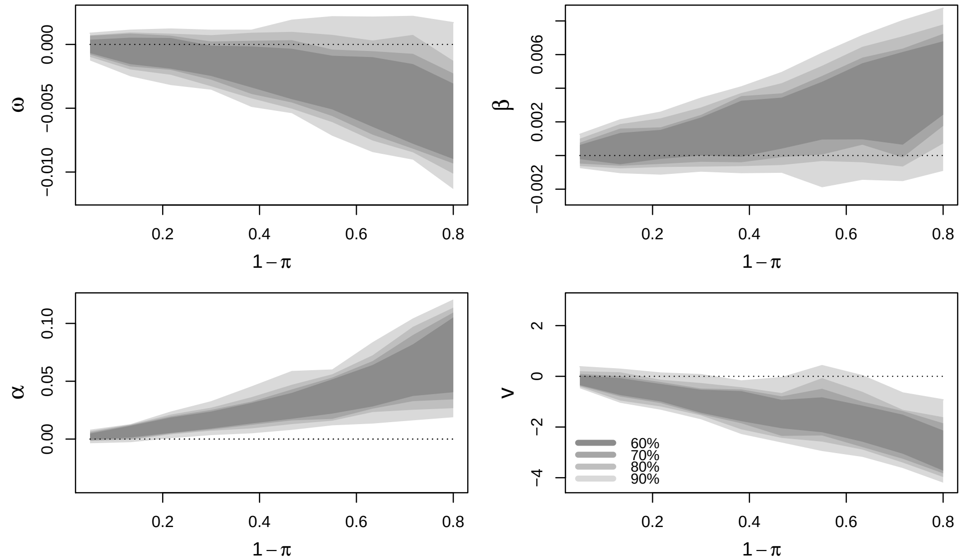

Figure 6: *Bias of the pseudo ML estimator compared to the corresponding full sample estimator for different values of π. The grey areas represent confidence bounds of the bias, which are obtained from 100 random draws of the missing data. The dashed line represents a bias equal to zero.*

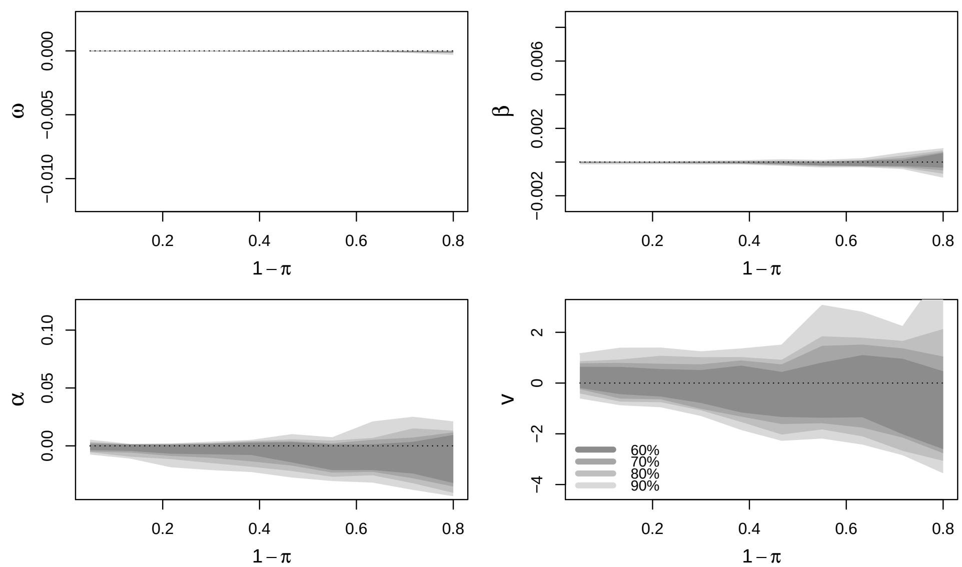

Figure 7: *Bias of the Indirect Inference estimator (with $S = 5$) compared to the corresponding full sample estimator for different values of π. The grey areas represent confidence bounds of the bias, which are obtained from 100 random draws of the missing data. The dashed line represents zero bias. The dashed line represents a bias equal to zero.*

findings provided by the simulation experiment. In particular, the estimator of α gets further away from the corresponding full sample estimator as the probability of missing observations π increases. We observe this divergence clearly in Figure 6 where the zero-line is not within the 90% variability bounds for large values of π. A similar situation occurs for the parameter ν. As we have discussed throughout the paper, this issue can be addressed by the consistent indirect inference estimator as proposed in Section 3 and studied in detail in Section 4.

Figure 7 provides evidence that the indirect inference estimation procedure does not lead to any bias for any parameter, in particular when compared to the pseudo ML estimation results in Figure 6. We have expected this result since the indirect inference estimator is consistent. However, a small bias may be observed in this experiment since we are dealing with real data and the analysis is conditional on an observed time series. Therefore, the model is possibly misspecified and may cause a slight bias. Furthermore, we emphasise that the variability observed in the estimation is not due to the variability of the estimation. Our analyses are based on a single time series and the randomness in the different draws is only due to the Bernoulli missing value generator.

7 Conclusion

We have highlighted the theoretical issues that arise when missing observations are present in observation-driven time series models and in particular in score-driven models. We have argued that the "setting-to-zero" method may lead to the inconsistency of the maximum likelihood estimator. Based on theoretical arguments and simulation experiments, we have confirmed the inconsistency problem. We further have proposed a new estimation procedure based on the method of indirect inference that provides a simple and general approach to obtain consistency and asymptotic normality in the presence of missing observations for observation-driven time series models. Simulation experiments have shown that the proposed estimator has comparable performances to the exact maximum likelihood estimator for a Gaussian score-driven location model. Finally, an experiment with real financial data has illustrated the key importance of our results in a practical context.

A Appendix

A.1 Proofs of Section 3

Proof of Theorem 3.1. Let $\mu_t(\theta)$ denote the limit of the filtered parameter $\hat{\mu}_t(\theta)$ that is given by

$$\mu_t(\theta) = \alpha \sum_{k=1}^{\infty} \left[\prod_{i=1}^{k-1} \tilde{\xi}_{t-i} \right] I_{t-k} y_{t-k}, \tag{11}$$

where $\tilde{\xi}_t = \beta - \alpha I_t$. Furthermore, we denote with L_T the pseudo log-likelihood function evaluated at the limit filter $\mu_t(\theta)$, i.e. $L_T(\theta) = -2^{-1}T^{-1}\sum_{t=1}^{T} I_t(y_t - \mu_t(\theta))^2$. Finally, we define the limit of the pseudo log-likelihood $L(\theta)$ as $L(\theta) = -2^{-1}\pi\mathbb{E}[(y_t - \mu_t(\theta))^2]$.

To prove the theorem, we first show that the pseudo likelihood function $\hat{L}_T(\theta)$ converges a.s. and uniformly to $L(\theta)$, i.e. $\sup_{\theta\in\Theta} |\hat{L}_T(\theta) - L(\theta)| \xrightarrow{a.s.} 0$. Then, we show that this uniform convergence together with Lemma A.2 implies that $\liminf_{T\to\infty} \|\hat{\theta}_T - \theta_0\| > \epsilon$ with probability 1.

As concerns the uniform convergence, an application of the triangle inequality yields

$$\sup_{\theta\in\Theta} \left| \hat{L}_T(\theta) - L(\theta) \right| \leq \sup_{\theta\in\Theta} \left| \hat{L}_T(\theta) - L_T(\theta) \right| + \sup_{\theta\in\Theta} \left| L_T(\theta) - L(\theta) \right|. \tag{12}$$

Therefore, we just need to show that both terms on the right hand side of the inequality in (12) go to zero almost surely. Regarding the first term, we have that $\sup_{\theta\in\Theta} |\hat{\mu}_t(\theta) - \mu_t(\theta)|$ goes to zero exponentially almost surely (e.a.s.) by Lemma A.1. Then we obtain that the following inequality is satisfied for large enough t

$$\sup_{\theta\in\Theta} |(y_t - \mu_t(\theta))^2 - (y_t - \hat{\mu}_t(\theta))^2| \leq \eta_t \sup_{\theta\in\Theta} |\hat{\mu}_t(\theta) - \mu_t(\theta)|, \tag{13}$$

where

$$\eta_t = 2\sup_{\theta\in\Theta} |\mu_t(\theta)| + 2|y_t| + 1 \geq 2\sup_{\theta\in\Theta} |\mu_t^*(\theta)| + 2|y_t|$$

for any μ_t^* between μ_t and $\hat{\mu}_t$. Therefore, since $\{\eta_t\}_{t\in\mathbb{Z}}$ is a stationary and ergodic sequence with bounded moments of any order and $\sup_{\theta\in\Theta} |\hat{\mu}_t(\theta) - \mu_t(\theta)|$ goes to zero e.a.s., we conclude that the left hand side of the inequality in (13) goes to zero almost surely by an application of Lemma 2.1 of Straumann and Mikosch (2006). It is then immediate to see that $\sup_{\theta\in\Theta} |(y_t - \mu_t(\theta))^2 - (y_t - \hat{\mu}_t(\theta))^2| \xrightarrow{a.s.} 0$ implies the desired result, i.e. $\sup_{\theta\in\Theta} |\hat{L}_T(\theta) - L_T(\theta)| \xrightarrow{a.s.} 0$. Finally, the second term on the right hand side of the inequality in (12) goes to zero almost surely by an application of the ergodic theorem of Rao (1962) provided that $\mathbb{E}\sup_{\theta\in\Theta} |y_t - \mu_t(\theta)|^2 < \infty$. We note that $\mathbb{E}\sup_{\theta\in\Theta} |y_t - \mu_t(\theta)|^2 < \infty$ holds true as Θ is a compact set contained in $(0,1)^2$ and moments of any order for $\mu_t(\theta)$ exists for any $\theta \in \Theta$.

From the uniform convergence $\sup_{\theta\in\Theta}|\hat{L}_T(\theta) - L(\theta)| \xrightarrow{a.s.} 0$ and Lemma A.2, we infer that there exists an $\epsilon > 0$ such that the following inequality is satisfied with probability 1

$$\limsup_{n\to\infty}\left(\sup_{\theta\in B_\epsilon(\theta_0)}\hat{L}_T(\theta) - \sup_{\theta\in B^c_\epsilon(\theta_0)}\hat{L}_T(\theta)\right) < 0, \tag{14}$$

where $B_\epsilon(\theta_0) = \{\theta \in \Theta : \|\theta_0 - \theta\| < \epsilon\}$ and $B^c_\epsilon(\theta_0) = \Theta/B_\epsilon(\theta_0)$. From the definition of $\hat{\theta}_T$, we know that $\hat{L}_T(\hat{\theta}_T) = \sup_{\theta\in\Theta}\hat{L}_T(\theta)$ for any $n \in \mathbb{N}$. Therefore, if we assume that $\liminf_{T\to\infty}\|\hat{\theta}_T - \theta_0\| > \epsilon$ with probability smaller than 1, then the inequality in (14) must be satisfied with probability smaller than 1 since,

$$\|\hat{\theta}_T - \theta_0\| > \epsilon \quad \Leftrightarrow \quad \sup_{\theta\in B_\epsilon(\theta_0)}\hat{L}_T(\theta) < \sup_{\theta\in B^c_\epsilon(\theta_0)}\hat{L}_T(\theta)$$

and hence

$$\mathbb{P}(\|\hat{\theta}_T - \theta_0\| > \epsilon) = \mathbb{P}\Big(\sup_{\theta\in B_\epsilon(\theta_0)}\hat{L}_T(\theta) < \sup_{\theta\in B^c_\epsilon(\theta_0)}\hat{L}_T(\theta)\Big) < 1.$$

This is a contradiction with respect to (14). Therefore, we can conclude that $\liminf_{T\to\infty}\|\hat{\theta}_T - \theta_0\| > \epsilon$ with probability 1. This concludes the proof of the theorem. □

Lemma A.1. *For any $(\alpha_0, \beta_0, \pi) \in (0,1)^3$ and any compact set $\Theta \subset (0,1)^2$, we have that*

$$\sup_{\theta\in\Theta}|\hat{\mu}_t(\theta) - \mu_t(\theta)| \xrightarrow{e.a.s.} 0, \quad as \quad t \to \infty,$$

for any initialization $\hat{\mu}_1(\theta) \in \mathbb{R}$.

Proof. The result can be obtained by an application of Theorem 3.1 of Bougerol (1993) to a sequence of random functions $\{x_t(\cdot)\}_{t\in\mathbb{N}}$ defined through a Stochastic Recurrence Equation (SRE) of the form

$$x_{t+1}(\theta) = \phi_t(x_t(\theta), \theta),\ t \in \mathbb{N}, \tag{15}$$

where $x_1(\theta) \in \mathbb{R}$, the map $(x, \theta) \mapsto \phi_t(x, \theta)$ from $\mathbb{R} \times \Theta$ into $\mathbb{R}$ is almost surely continuous and the sequence $\{\phi_t(x, \theta)\}_{t\in\mathbb{Z}}$ is stationary and ergodic for any $(x, \theta) \in \mathbb{R} \times \Theta$. Bougerol's theorem ensures that for any initialization $x_1(\theta)$ the sequence defined by the SRE in (15) converges e.a.s. and uniformly in Θ to a unique stationary and ergodic sequence $\{\tilde{x}_t(\theta)\}$. The conditions required to apply Bougerol's result are:
(i) There is an $x \in \mathbb{R}$ such that $\mathbb{E}\log^+(\sup_{\theta\in\Theta}|\phi_0(x, \theta)|) < \infty$,
(ii) $\mathbb{E}\log^+(\sup_{\theta\in\Theta}\Lambda_0(\theta)) < \infty$,
(iii) $\mathbb{E}\log(\sup_{\theta\in\Theta}\Lambda_0(\theta)) < 0$,
where $\Lambda_t(\theta) = \sup_{x\in\mathbb{R}}|\partial\phi_t(x, \theta)/\partial x|$.

In our case we have that the sequence $\{\hat{\mu}_t(\theta)\}$ is defined through the SRE in (7). As a result, we have that $\phi_t(x,\theta) = \beta x + \alpha I_t(y_t - x)$. Therefore we immediately obtain that $\Lambda_t(\theta) = |\beta - \alpha I_t|$. Furthermore, the limit function $\tilde{x}_t(\theta)$ in our case is given by $\mu_t(\theta)$, which is defined in (11). In the following we show that the conditions of Bougerol's theorem are satisfied.

First we note that there exists an $x \in \mathbb{R}$ such that $\mathbb{E}\log^+(\sup_{\theta\in\Theta}|\phi_0(x,\theta)|) < \infty$ because we can set $x = 0$ and we immediately obtain that

$$\mathbb{E}\log^+\left(\sup_{\theta\in\Theta}|\phi_0(x,\theta)|\right) \leq \mathbb{E}\left(\sup_{\theta\in\Theta}|\alpha I_t y_t|\right) \leq \sup_{\theta\in\Theta}|\alpha|\mathbb{E}|y_t| < \infty,$$

where the last equality is implied by the fact that $\sup_{\theta\in\Theta}|\alpha|$ is finite by compactness of Θ and $\mathbb{E}|y_t|$ is finite because y_t is a stationary ARMA(1,1) process for any $(\alpha_0, \beta_0) \in (0,1)^2$ and thus moments of any order exist. Second, we note that $\mathbb{E}\log^+(\sup_{\theta\in\Theta}\Lambda_0(\theta)) < \infty$ and $\mathbb{E}\log(\sup_{\theta\in\Theta}\Lambda_0(\theta)) < 0$ since $\Lambda_t(\theta) = |\beta - \alpha I_t|$ is smaller than 1 with probability 1 for any $\theta \in \Theta$ and therefore by compactness $\sup_{\theta\in\Theta}\Lambda_t(\theta) < 1$ with probability 1. This concludes the proof of the lemma.

□

Lemma A.2. *For some $(\alpha_0, \beta_0, \pi) \in (0,1)^3$ there exists an $\epsilon > 0$ such that*

$$\sup_{\theta\in B_\epsilon(\theta_0)} L(\theta) < \sup_{\theta\in B^c_\epsilon(\theta_0)} L(\theta).$$

Proof. In the following, we shall show that $\partial L(\theta)/\partial\beta|_{\theta=\theta_0} \neq 0$ for some $(\alpha_0, \beta_0, \pi) \in (0,1)^3$. Then, given the smoothness of the function $L(\theta)$ in Θ and the assumption that θ_0 is an interior point of Θ, we can conclude that the supremum of $L(\theta)$ in Θ is not contained in the closure of the set $B_\epsilon(\theta_0)$ for small enough $\epsilon > 0$. This immediately proves the statement of the Lemma.

We are therefore left with showing that $\partial L(\theta)/\partial\beta|_{\theta=\theta_0} \neq 0$ for some $(\alpha_0, \beta_0, \pi) \in (0,1)^3$. First, we obtain a closed form expression for $L(\theta)$ and $\partial L(\theta)/\partial\beta$. Expanding the square in the expression of $L(\theta)$, we obtain that $L(\theta) = -2^{-1}\pi(1 + \mathbb{E}[\mu^o_t - \mu_t(\theta)]^2)$ as ε_t is independent of the past observations as well as the missing value process $\{I_t\}_{t\in\mathbb{Z}}$. We also note that, expanding the recursion in (6), μ^o_t can be written as

$$\mu^o_t = \alpha_0 \sum_{k=1}^{\infty} \xi_0^{k-1} y_{t-k},$$

where $\xi_0 = \beta_0 - \alpha_0$. Therefore, considering the expression of $\mu_t(\theta)$ in (11), we obtain that

$$(\mu_t^o - \mu_t(\theta))^2 = \sum_{k=1}^{\infty}\sum_{s=1}^{\infty}\Bigg(\alpha_0^2\xi_0^{k+s-2} + \alpha^2\left[\prod_{i=1}^{k-1}\tilde{\xi}_{t-i}\right]\left[\prod_{i=1}^{s-1}\tilde{\xi}_{t-i}\right]I_{t-k}I_{t-s} \\ - \alpha\alpha_0\xi_0^{k-1}\left[\prod_{i=1}^{s-1}\tilde{\xi}_{t-i}\right]I_{t-s} - \alpha\alpha_0\xi_0^{s-1}\left[\prod_{i=1}^{k-1}\tilde{\xi}_{t-i}\right]I_{t-k}\Bigg)y_{t-s}y_{t-k}. \tag{16}$$

For convenience, we split the double sum in (16) in three terms, namely the sum of elements such that $k = s$, $k < s$ and $k > s$.

Taking into account that $\{I_t\}_{t\in\mathbb{Z}}$ is an i.i.d. sequence of Bernoulli random variables and the independence between $\{y_t\}_{t\in\mathbb{Z}}$ and $\{I_t\}_{t\in\mathbb{Z}}$, we obtain that the expectation of the sum of terms in (16) such that $k = s$, which we denote as s_1, is given by

$$s_1 = \sum_{k=1}^{\infty}\left(\alpha_0^2\xi_0^{2(k-1)} + \alpha^2\pi\xi_Z^{k-1} - 2\alpha\alpha_0\pi\xi_0^{k-1}\xi_B^{k-1}\right)\gamma_0 \\ = \left(\frac{\alpha_0^2}{1-\xi_0^2} + \frac{\alpha^2\pi}{1-\xi_Z} - \frac{2\alpha\alpha_0\pi}{1-\xi_0\xi_B}\right)\gamma_0,$$

where $\xi_B = \mathbb{E}(\tilde{\xi}_t) = \beta - \pi\alpha$, $\xi_Z = \mathbb{E}(\tilde{\xi}_t^2) = \pi(\beta-\alpha)^2 + (1-\pi)\beta^2$ and $\gamma_k = \mathbb{E}(y_t y_{t-k})$ is given in Lemma A.3 for $k \in \mathbb{N}$. Similarly, the expectation of the sum of terms in (16) such that $k < s$, which we denote as s_2, is given by

$$s_2 = \sum_{k=1}^{\infty}\sum_{s=1}^{\infty}\Bigg(\alpha_0^2\xi_0^{2(k-1)+s} + \alpha^2\pi\xi_A\xi_B^{s-1}\xi_Z^{k-1} \\ - \alpha\alpha_0\pi\left(\xi_0^{k-1}\xi_B^{s+k-1} + \xi_0^{s+k-1}\xi_B^{k-1}\right)\Bigg)\beta_0^{s-1}\tilde{\gamma} \\ = \frac{\alpha_0^2\xi_0\tilde{\gamma}}{(1-\xi_0^2)(1-\xi_0\beta_0)} + \frac{\alpha^2\pi^2\xi_A\tilde{\gamma}}{(1-\xi_B\beta_0)(1-\xi_Z)} \\ - \frac{\alpha\alpha_0\pi\xi_B\tilde{\gamma}}{(1-\xi_0\xi_B)(1-\xi_B\beta_0)} - \frac{\alpha\alpha_0\pi\xi_0\tilde{\gamma}}{(1-\xi_0\xi_B)(1-\xi_0\beta_0)},$$

where $\xi_A = \beta - \alpha$ and $\tilde{\gamma}$ is given in Lemma A.3. Finally, it can be easily noted that the expectation of the sum of terms in (16) such that $k > s$ is equal to s_2. As a result, we can conclude that $\mathbb{E}(\mu_t^o - \mu_t(\theta))^2 = s_1 + 2s_2$. We can now compute the derivative with respect to β of s_1 and s_2. By elementary calculus, we obtain that the derivative of s_1 evaluated at $\theta = \theta_0$ is given by

$$\dot{s}_1 = \frac{\partial s_1}{\partial\beta}\bigg|_{\theta=\theta_0} = \left(\frac{2\xi_B^o\alpha_0^2\pi}{(1-\xi_Z^o)^2} - \frac{2\alpha_0^2\xi_0\pi}{(1-\xi_0\xi_B^o)^2}\right)\gamma_0.$$

Similarly, the derivative of s_2 evaluated at θ_0 is given by

$$\dot{s}_2 = \left.\frac{\partial s_2}{\partial \beta}\right|_{\theta=\theta_0} = \dot{s}_{22} + \dot{s}_{23} + \dot{s}_{24},$$

where

$$\begin{aligned}
\dot{s}_{22} =& \alpha_0^2\pi^2\tilde{\gamma}\left(\frac{(1-\xi_B^o\beta_0)(1-\xi_Z^o) + \xi_A^o\beta_0(1-\xi_Z^o) + 2\xi_A^o\xi_B^o(1-\xi_B^o\beta_0)}{(1-\xi_B^o\beta_0)^2(1-\xi_Z^o)^2}\right),\\
\dot{s}_{23} =& -\alpha_0^2\pi\tilde{\gamma}\left(\frac{(1-\xi_B^o\beta_0)(1-\xi_B^o\xi_0^o) + \xi_B^o\xi_0(1-\xi_B^o\beta_0) + \xi_B^o\beta_0(1-\xi_B^o\xi_0)}{(1-\xi_B^o\beta_0)^2(1-\xi_0\xi_B^o)^2}\right),\\
\dot{s}_{24} =& -\alpha_0^2\pi\tilde{\gamma}\left(\frac{\xi_0^2(1-\xi_0\beta_0)}{(1-\xi_0\beta_0)^2(1-\xi_0\xi_B^o)^2}\right),
\end{aligned}$$

with ξ_A^o, ξ_B^o and ξ_Z^o denoting ξ_A, ξ_B and ξ_Z evaluated at $(\alpha,\beta) = (\alpha_0,\beta_0)$. The derivative of $L(\theta)$ with respect to β and evaluated at θ_0 is therefore given by $\partial L(\theta)/\partial\beta|_{\theta=\theta_0} = -2^{-1}\pi(\dot{s}_1 + 2\dot{s}_2)$. Finally, we conclude the proof of the theorem by noticing that the derivative is different from zero for some $(\alpha_0,\beta_0,\pi) \in (0,1)^3$. For instance, it is easy to verify that the derivative is different from zero at the point $(\alpha_0,\beta_0,\pi) = (0.2, 0.95, 0.5)$. Other values can be used to obtain the same result.

□

Lemma A.3. *When $\sigma_0^2 = 1$ and $\omega_0 = 0$, the autocovariance function of y_t, namely $\gamma_k = \mathbb{E}(y_t y_{t-k})$, is given by*

$$\gamma_k = \begin{cases} 1 + \frac{\alpha_0^2}{1-\beta_0^2}, & \text{if } k = 0 \\ \beta_0^{k-1}\tilde{\gamma}, & \text{if } k \geq 1 \end{cases},$$

where $\tilde{\gamma} = \alpha_0 + \frac{\alpha_0^2\beta_0}{1-\beta_0^2}$.

Proof. The proof follows immediately by noting that y_t is an ARMA(1,1) that has the following MA(∞) representation

$$y_t = \alpha_0 \sum_{i=1}^{\infty} \beta_0^{i-1}\varepsilon_{t-i} + \varepsilon_t,$$

where $\varepsilon_t \sim N(0,1)$. It is then straightforward to obtain the expression for the autocovariance function γ_k. □

A.2 Proofs of Section 4

Proof of Proposition 4.1. We obtain the consistency of $\hat{\theta}_{S,T}(\bar{\theta})$, for every $\bar{\theta} \in \Theta$, by appealing to Theorem 3.4 in White (1994). In particular, we show that $\hat{L}_{S,T}(\theta, \bar{\theta})$ converges a.s. to a limit deterministic function $L(\theta, \bar{\theta}) = \mathbb{E} \log p\big(y_{i,t}(\bar{\theta}) | f_{i,t}(\theta, \bar{\theta}); \theta\big)$ uniformly in $\theta \in \Theta$, for every $\bar{\theta} \in \Theta$, that is,

$$\sup_{\theta \in \Theta} |\hat{L}_{S,T}(\theta, \bar{\theta}) - L(\theta, \bar{\theta})| \overset{a.s.}{\to} 0 \quad \forall\, \bar{\theta} \in \Theta \text{ as } T \to \infty, \tag{17}$$

and that $\theta = \theta^*(\bar{\theta})$ is the identifiably unique minimiser of the limit criterion $L(\theta, \bar{\theta})$, that is,

$$\sup_{\theta \in \Theta \,:\, \|\theta - \theta^*(\bar{\theta})\| > \delta} L(\theta, \bar{\theta}) < L(\theta^*(\bar{\theta}), \bar{\theta}), \;\; \forall\, \delta > 0, \; \bar{\theta} \in \Theta\,. \tag{18}$$

The identifiable uniqueness of $\theta^*(\bar{\theta})$ in (18) follows by the compactness of Θ (Assumption 4.1), the uniqueness of $\theta^*(\bar{\theta})$ $\forall\, \bar{\theta} \in \Theta$ (Assumption 4.6) and the continuity of $L(\cdot, \bar{\theta})$ for every $\bar{\theta} \in \Theta$, which is ensured by the continuity and uniform convergence of $\hat{L}_{S,T}(\cdot, \bar{\theta})$ shown below.

As concerns the uniform convergence in (17), for every $\bar{\theta} \in \Theta$, the triangle inequality yields

$$\sup_{\theta \in \Theta} |\hat{L}_{S,T}(\theta, \bar{\theta}) - L(\theta, \bar{\theta})| \le \sup_{\theta \in \Theta} |\hat{L}_{S,T}(\theta, \bar{\theta}) - L_{S,T}(\theta, \bar{\theta})| + \sup_{\theta \in \Theta} |L_{S,T}(\theta, \bar{\theta}) - L(\theta, \bar{\theta})|, \tag{19}$$

where $L_{S,T}(\theta, \bar{\theta})$ denotes the log-likelihood function evaluated at the limit filter $\{f_{i,t}(\theta, \bar{\theta})\}$. Therefore, the desired uniform convergence follows if both terms on the right side of the inequality in (19) go to zero almost surely. As concerns the first term, from Assumptions 4.4 and 4.5, we obtain that

$$\begin{aligned}
&\sup_{\theta \in \Theta} |\hat{L}_{S,T}(\theta, \bar{\theta}) - L_{S,T}(\theta, \bar{\theta})| = \\
&\quad = \sup_{\theta \in \Theta} \Big| \frac{1}{ST} \sum_{i=1}^{S} \sum_{t=2}^{T} \Big(\log p\big(y_{i,t}(\bar{\theta}) | \hat{f}_{i,t}(\theta, \bar{\theta}); \theta\big) - \log p\big(y_{i,t}(\bar{\theta}) | f_{i,t}(\theta, \bar{\theta}); \theta\big) \Big) \Big| \\
&\quad \le \frac{1}{ST} \sum_{i=1}^{S} \sum_{t=2}^{T} \sup_{\theta \in \Theta} \Big| \log p\big(y_{i,t}(\bar{\theta}) | \hat{f}_{i,t}(\theta, \bar{\theta}); \theta\big) - \log p\big(y_{i,t}(\bar{\theta}) | f_{i,t}(\theta, \bar{\theta}); \theta\big) \Big| \\
&\quad \le \frac{1}{ST} \sum_{i=1}^{S} \sum_{t=2}^{T} \sup_{\theta \in \Theta} \sup_{f} \Big| \nabla_f \log p\big(y_{i,t}(\bar{\theta}) | f; \theta\big) \Big| \times \sup_{\theta \in \Theta} \Big| \hat{f}_{i,t}(\theta, \bar{\theta}) - f_{i,t}(\theta, \bar{\theta}) \Big| \overset{a.s.}{\to} 0.
\end{aligned}$$

The a.s. convergence to zero follows by an application of Lemma 2.1 of Straumann and Mikosch (2006) since the Lipschitz coefficient $\{\sup_{\theta \in \Theta} \sup_f |\nabla_f \log p(y_{i,t}(\bar{\theta}) | f; \theta)|\}$ is strictly stationary and ergodic with a logarithmic moment (Assumption 4.5) and $\sup_{\theta \in \Theta} |\hat{f}_{i,t}(\theta, \bar{\theta}) - f_{i,t}(\theta, \bar{\theta})| \overset{e.a.s.}{\to} 0$ (Assumption 4.4).[2]

[2] We emphasise that $\{\sup_{\theta \in \Theta} \sup_f |\nabla_f \log p(y_{i,t}(\bar{\theta}) | f; \theta)|\}$ is stationary and ergodic since it is a measurable function of the stationary and ergodic sequence $\{y_{i,t}(\bar{\theta})\}$ (Assumption 4.3).

As concerns the second term on the right hand side of (19), we obtain the uniform convergence

$$\sup_{\theta\in\Theta} |L_{S,T}(\theta,\bar{\theta}) - L(\theta,\bar{\theta})| \xrightarrow{a.s.} 0,$$

by an application of the ergodic theorem of Rao (1962), applied to the sequence $\{\log p\big(y_{i,t}(\bar{\theta})|f_{i,t}(\cdot,\bar{\theta});\cdot\big)\}$ with elements taking values in the Banach space of continuous functions $\mathbb{C}(\Theta)$ equipped with supremum norm. We notice that the sequence $\{\log p(y_{i,t}(\bar{\theta}), f_{i,t}(\cdot,\bar{\theta}),\cdot)\}$ is strictly stationary and ergodic since each element is a measurable function of the strictly stationary data $y_{i,t}(\bar{\theta})$ and the limit filter $f_{i,t}(\cdot,\bar{\theta})$, for every $\bar{\theta}\in\Theta$ (Assumptions 4.3 and 4.4). Additionally, $\log p\big(y_{i,t}(\bar{\theta})|f_{i,t}(\cdot,\bar{\theta});\cdot\big)$ has a uniform bounded moment for every $\bar{\theta}\in\Theta$ by Assumption 4.5. This enables the application of Rao (1962)'s law of large numbers and obtain the desired result.

We can therefore conclude that $\hat{\theta}_{S,T}(\bar{\theta})$ is strongly consistent for $\theta^*(\bar{\theta})$. Furthermore, we note that the strong consistency of $\hat{\theta}_T$ to $\theta^*(\theta_0)$ follows immediately since $\hat{\theta}_T$ has the same stochastic properties of $\hat{\theta}_{S,T}(\bar{\theta})$ with $S=1$ and $\bar{\theta}=\theta_0$. This concludes the proof of the Proposition.

□

Proof of Theorem 4.1. Following Theorem 3.4 in White (1994), we obtain the consistency of our indirect inference estimator by showing that the indirect inference criterion $\|\hat{\theta}_{S,T}(\bar{\theta}) - \hat{\theta}_T\|$ satisfies

$$\sup_{\bar{\theta}\in\Theta}\Big|\, \|\hat{\theta}_{S,T}(\bar{\theta}) - \hat{\theta}_T\| \;-\; \|\theta^*(\bar{\theta}) - \theta^*(\theta_0)\| \,\Big| \xrightarrow{a.s.} 0 \quad \text{as} \quad T\to\infty, \tag{20}$$

and that θ_0 is the identifiably unique minimiser of the limit criterion

$$\inf_{\bar{\theta}\in\Theta\,:\,\|\bar{\theta}-\theta_0\|>\delta} \|\theta^*(\bar{\theta}) - \theta^*(\theta_0)\| > \|\theta^*(\theta_0) - \theta^*(\theta_0)\| = 0\;\forall\,\delta>0.$$

The identifiable uniqueness follows immediately from the compactness of the parameter space and the continuity and injective nature of the binding function $\theta^*(\cdot)$ (Assumption 4.11); see Potscher and Prucha (1997).

As concerns the uniform convergence of the criterion in (20), the reverse triangle inequality and the triangle inequality yield

$$\begin{aligned}
&\sup_{\bar{\theta}\in\Theta}\big|\, \|\hat{\theta}_{S,T}(\bar{\theta}) - \hat{\theta}_T\| \;-\; \|\theta^*(\bar{\theta}) - \theta^*(\theta_0)\| \,\big| \leq \\
&\qquad\qquad \leq \sup_{\bar{\theta}\in\Theta} \|\hat{\theta}_{S,T}(\bar{\theta}) - \hat{\theta}_T - \theta^*(\bar{\theta}) + \theta_0\| \leq \sup_{\bar{\theta}\in\Theta} \|\hat{\theta}_{S,T}(\bar{\theta}) - \theta^*(\bar{\theta})\| + \|\hat{\theta}_T - \theta_0\|.
\end{aligned}$$

Therefore, the desired result follows if both terms on the right hand side of the above inequality go a.s. to zero. We obtain that the convergence of $\hat{\theta}_T$ to $\theta^*(\theta_0)$ follows by an application of Proposition 4.1 and the uniform convergence of $\hat{\theta}_{S,T}(\bar{\theta})$ to $\theta^*(\bar{\theta})$ follows by an application of Lemma A.4.

□

Proof of Proposition 4.2. The asymptotic normality of the auxiliary statistics is obtained by appealing to Theorem 6.2 in White (1994). In particular, we obtain the asymptotic normality of $\hat{\theta}_{ST}(\theta_0)$ by verifying the following conditions:

(i) The strong consistency of the auxiliary estimator $\hat{\theta}_{S,T}(\theta_0) \overset{a.s.}{\to} \theta^*(\theta_0)$;

(ii) Twice continuous differentiability of the pseudo log-likelihood function $\hat{L}_{S,T}(\theta, \theta_0)$ with respect to θ;

(iii) Asymptotic normality of the score evaluated at the pair $(\theta^*(\theta_0), \theta_0)$

$$\sqrt{T}\nabla_\theta \hat{L}_{S,T}(\theta^*(\theta_0), \theta_0) \overset{d}{\to} N(0, \Sigma_S^*(\theta_0));$$

(iv) Uniform convergence of the Hessian

$$\sup_{\theta\in\Theta} \left\| \nabla^2_{\theta\theta}\hat{L}_{S,T}(\theta, \theta_0) - \mathbb{E}\nabla^2_{\theta\theta} L_{S,T}(\theta, \theta_0) \right\| \overset{a.s.}{\to} 0.$$

(v) The Hessian matrix $\Omega^*(\theta_0) = \mathbb{E}\nabla^2_{\theta\theta} L_{S,T}(\theta^*(\theta_0), \theta_0)$ is non-singular.

First we note that Condition (i) is satisfied by an application of Proposition 4.1 and Condition (ii) is satisfied by assumption.

As concerns condition (iii), we can re-write the score of the likelihood as

$$\begin{aligned}\sqrt{T}\nabla_\theta \hat{L}_{S,T}(\theta^*(\theta_0), \theta_0) =& \sqrt{T}\nabla_\theta \hat{L}_{S,T}(\theta^*(\theta_0), \theta_0) - \sqrt{T}\nabla_\theta L_{S,T}(\theta^*(\theta_0), \theta_0) \\ &+ \sqrt{T}\nabla_\theta L_{S,T}(\theta^*(\theta_0), \theta_0).\end{aligned}$$

Therefore, the desired result can be proved by showing that a central limit theorem applies to the limit score

$$\sqrt{T}\nabla_\theta L_{S,T}(\theta^*(\theta_0), \theta_0) \overset{d}{\to} N(0, \Sigma_S^*(\theta_0)),$$

and showing that that the remainder term vanishes almost surely

$$\sqrt{T}\nabla_\theta \hat{L}_{S,T}(\theta^*(\theta_0), \theta_0) - \sqrt{T}\nabla_\theta L_{S,T}(\theta^*(\theta_0), \theta_0) \overset{a.s.}{\to} 0 \quad \text{as} \quad T \to \infty.$$

In particular, we obtain that the limit score is asymptotically Gaussian by an application of a central limit theorem for sequences that are NED on α-mixing sequences; see Theorem 10.2 in Potscher and Prucha (1997). Note that $\{y_{i,t}(\theta_0)\}$, $\{f_{i,t}(\theta^*(\theta_0),\theta_0)\}$ and $\{\nabla_\theta f_{i,t}(\theta^*(\theta_0),\theta_0)\}$ are NED of size -1 on an α-mixing sequence (Assumptions 4.12 and 4.13), the score $\nabla_\theta \log p$ is Lipschitz continuous on the first two arguments (Assumption 4.14) and it has two bounded moments (Assumption 4.14). Therefore, it follows that the score is NED of size -1 on the same α-mixing sequence; see Lemma 1 in Andrews (1991) and Corollary 6.8 in Potscher and Prucha (1997). Furthermore, since the α-mixing sequence has size $-2r/(r-1)$ for some $r > 2$, and the score has mean zero and r moments (Assumption 4.14), we conclude that the score satisfies a central limit theorem; see Wooldridge (1986), Gallant and White (1988) and Theorem 10.2 in Potscher and Prucha (1997). We thus have that

$$\begin{aligned}&\sqrt{T}\nabla_\theta L_{S,T}(\theta^*(\theta_0),\theta_0) := \\ &\quad \frac{1}{\sqrt{T}S}\sum_{i=1}^{S}\sum_{t=2}^{T}\nabla_\theta \log p\big(y_{i,t}(\theta_0)\big|f_{i,t}\big(\theta^*(\theta_0),\theta_0\big);\theta^*(\theta_0)\big) \xrightarrow{d} N(0,\Sigma_S^*(\theta_0)),\end{aligned}$$

where the asymptotic covariance matrix of the score is $\Sigma_S^*(\theta_0) = S^{-1}\Sigma^*(\theta_0) + \frac{S-1}{S}K^*(\theta_0)$ with

$$\Sigma^*(\theta_0) = \lim_{T\to\infty} \mathbb{V}\mathrm{ar}\left(\frac{1}{\sqrt{T}}\sum_{t=2}^{T}\nabla_\theta \log p\big(y_{i,t}(\theta_0)\big|f_{i,t}\big(\theta^*(\theta_0),\theta_0\big);\theta^*(\theta_0)\big)\right)$$

$$\begin{aligned}\text{and} \quad K^*(\theta_0) = \lim_{T\to\infty} \mathbb{C}\mathrm{ov}\Bigg(&\frac{1}{\sqrt{T}}\sum_{t=2}^{T}\nabla_\theta \log p\big(y_{i,t}(\theta_0)\big|f_{i,t}\big(\theta^*(\theta_0),\theta_0\big);\theta^*(\theta_0)\big)\ , \\ &\frac{1}{\sqrt{T}}\sum_{t=2}^{T}\nabla_\theta \log p\big(y_{j,t}(\theta_0)\big|f_{j,t}\big(\theta^*(\theta_0),\theta_0\big);\theta^*(\theta_0)\big)\Bigg) \quad \text{for some } i \neq j.\end{aligned}$$

Note that this expression of the covariance matrix $\Sigma^*(\theta_0)$ is due to the fact that the scores of the pseudo log-likelihood can be correlated. Additionally, we obtain that

$$\begin{aligned}&\big\|\sqrt{T}\nabla_\theta \hat{L}_{S,T}(\theta^*(\theta_0),\theta_0) - \sqrt{T}\nabla_\theta L_T(\theta^*(\theta_0),\theta_0)\big\| \\ &\quad \leq \frac{1}{\sqrt{T}S}\sum_{i=1}^{S}\sum_{t=2}^{T}\sup_{f^{(0:1)}}\Big\|\nabla^2_{\theta f^{(0:1)}}\log p\big(y_{i,t}(\theta_0)\big|f_{i,t}\big(\theta^*(\theta_0),\theta_0\big);\theta^*(\theta_0)\big)\Big\| \\ &\qquad\qquad \times \big\|\nabla_\theta^{0:1}\hat{f}_{i,t}(\theta^*(\theta_0),\theta_0) - \nabla_\theta^{0:1}f_{i,t}(\theta^*(\theta_0),\theta_0)\big\| \xrightarrow{a.s.} 0.\end{aligned}$$

The almost sure convergence to zero follows by Lemma 2.1 in Straumann and Mikosch (2006) since the first factor of the product above is stationary and ergodic with a logarithmic moment (Assumption 4.9) and the second factor vanishes e.a.s. by Assumption 4.8. This implies that the double sum converges almost surely and therefore we obtain the a.s. convergence to zero because of $1/\sqrt{T}$.

As concerns Condition (iv), we obtain that the triangle inequality yields

$$\begin{aligned}&\sup_{\theta\in\Theta}\left\|\nabla_{\theta\theta}\hat{L}_{S,T}(\theta,\theta_0)-\mathbb{E}\nabla_{\theta\theta}L_{S,T}(\theta,\theta_0)\right\|\\&\qquad\le\sup_{\theta\in\Theta}\left\|\nabla_{\theta\theta}\hat{L}_{S,T}(\theta,\theta_0)-\nabla_{\theta\theta}L_{S,T}(\theta,\theta_0)\right\|+\sup_{\theta\in\Theta}\left\|\nabla_{\theta\theta}L_{S,T}(\theta,\theta_0)-\mathbb{E}\nabla_{\theta\theta}L_{S,T}(\theta,\theta_0)\right\|\end{aligned}$$

We are therefore left with showing that both terms on right hand side of the above inequality go to zero almost surely. First, we note that the second therm vanishes by application of Rao (1962)'s ergodic theorem since the required uniform moment condition is provided by Assumption 4.9. Instead, for the second term we obtain that

$$\begin{aligned}&\sup_{\theta\in\Theta}\left\|\nabla_{\theta\theta}\hat{L}_{S,T}(\theta,\theta_0)-\nabla_{\theta\theta}L_{S,T}(\theta,\theta_0)\right\|\\&\qquad\le\frac{1}{\sqrt{T}S}\sum_{i=1}^{S}\sum_{t=2}^{T}\sup_{\theta\in\Theta}\sup_{f^{(0:2)}}\left\|\nabla^3_{\theta\theta f^{(0:2)}}\log p\big(y_{i,t}(\theta_0)\big|f_{i,t}(\theta,\theta_0);\theta\big)\right\|\\&\qquad\qquad\times\sup_{\theta\in\Theta}\left\|\nabla^{0:2}_{\theta\theta}\hat{f}_{i,t}(\theta,\theta_0)-\nabla^{0:2}_{\theta\theta}f_{i,t}(\theta,\theta_0)\right\|\overset{a.s.}{\to}0.\end{aligned}$$

where the almost sure convergence to zero follows by Lemma 2.1 in Straumann and Mikosch (2006) since the required log moment condition for the first factor is given by Assumption 4.9 and the e.a.s. convergence of the second factor is given by Assumption 4.8.

As concerns Condition (v), we have that this condition is immediately satisfied by Assumption 4.10. Therefore, we conclude that

$$\sqrt{T}\Big(\hat{\theta}_{S,T}(\theta_0)-\theta^*(\theta_0)\Big)\overset{d}{\to}N\Big(0,\Omega^*(\theta_0)^{-1}\Sigma^*_S(\theta_0)\Omega^*(\theta_0)^{-1}\Big)\quad\text{as}\quad T\to\infty.$$

Finally, we note that the asymptotic normality of $\hat{\theta}_T$ follows immediately since $\hat{\theta}_T$ has the same stochastic properties of $\hat{\theta}_{S,T}(\theta_0)$ with $S=1$. This concludes the proof of the Proposition. □

Proof of Theorem 4.2. The proof of this theorem is available in Gourieroux et al. (1993). Note that the asymptotic normality of the auxiliary statistics is obtained in Proposition 4.2 and that, asymptotically, we have

$$\begin{aligned}\mathbb{V}\mathrm{ar}\Big(\sqrt{T}\big(\hat{\theta}_T-\hat{\theta}_{S,T}(\theta_0)\big)\Big)&=\Omega^*(\theta_0)^{-1}\Big[\Sigma^*(\theta_0)+\frac{1}{S}\Sigma^*(\theta_0)+\frac{S-1}{S}K^*(\theta_0)-2K^*(\theta_0)\Big]\Omega^*(\theta_0)^{-1}\\&=\Big(1+\frac{1}{S}\Big)\Omega^*(\theta_0)^{-1}\Big(\Sigma^*(\theta_0)-K^*(\theta_0)\Big)\Omega^*(\theta_0)^{-1}.\end{aligned}$$

Finally, we note that the form of the asymptotic covariance matrix simplifies because of exact identification.

$\square$

Lemma A.4. *Let Assumptions 4.1-4.11 hold. Then, the pseudo ML estimator $\hat{\theta}_{S,T}(\bar{\theta})$ converges a.s. and uniformly to $\theta^*(\bar{\theta})$, that is,*

$$\sup_{\bar{\theta}\in\Theta} \|\hat{\theta}_{S,T}(\bar{\theta}) - \theta^*(\bar{\theta})\| \xrightarrow{a.s.} 0 \quad \textit{as} \quad T \to \infty.$$

Proof. First, we note that an application of the mean value theorem yields

$$\sup_{\bar{\theta}\in\Theta} \|\hat{\theta}_{S,T}(\bar{\theta}) - \theta^*(\bar{\theta})\| \leq \sup_{\bar{\theta}\in\Theta} \|\nabla_\theta \hat{L}_{S,T}(\theta^*(\bar{\theta}),\bar{\theta})\| \sup_{\theta\in\Theta}\sup_{\bar{\theta}\in\Theta} \left\| \left(\nabla^2_{\theta\theta}\hat{L}_{S,T}(\theta,\bar{\theta})\right)^{-1} \right\|$$

Therefore, the desired result is obtained if

$$\sup_{\bar{\theta}\in\Theta} \|\nabla_\theta \hat{L}_{S,T}(\theta^*(\bar{\theta}),\bar{\theta})\| \xrightarrow{a.s.} 0 \quad \text{as} \quad T \to \infty, \tag{21}$$

and

$$\sup_{\theta\in\Theta}\sup_{\bar{\theta}\in\Theta} \left\| \left(\nabla^2_{\theta\theta}\hat{L}_{S,T}(\theta,\bar{\theta})\right)^{-1} \right\| \xrightarrow{a.s.} c \neq 0 \quad \text{as} \quad T \to \infty, \tag{22}$$

are satisfied.

As concerns the convergence in (21), we obtain that

$$\begin{aligned} &\sup_{\bar{\theta}\in\Theta} \|\nabla_\theta \hat{L}_{S,T}(\theta^*(\bar{\theta}),\bar{\theta})\| \\ &\qquad \leq \sup_{\bar{\theta}\in\Theta} \|\nabla_\theta \hat{L}_{S,T}(\theta^*(\bar{\theta}),\bar{\theta}) - \nabla_\theta L_{S,T}(\theta^*(\bar{\theta}),\bar{\theta})\| + \sup_{\bar{\theta}\in\Theta} \|\nabla_\theta L_{S,T}(\theta^*(\bar{\theta}),\bar{\theta})\|. \end{aligned} \tag{23}$$

The second term on the right hand side of (23) vanishes a.s. by application of the ergodic theorem of Rao (1962). In particular, $\mathbb{E}\nabla_\theta L_{S,T}(\theta^*(\bar{\theta}),\bar{\theta}) = 0$ for any $\bar{\theta} \in \Theta$ and the uniform moment condition on the score in Assumption 4.9 ensures the a.s. uniform convergence of $\nabla_\theta L_{S,T}(\theta^*(\bar{\theta}),\bar{\theta})$ to zero as T diverges. Instead, for the the first term on the right hand side of (23), we obtain that

$$\begin{aligned} &\sup_{\bar{\theta}\in\Theta} \big\|\nabla_\theta \hat{L}_{S,T}(\theta^*(\bar{\theta}),\bar{\theta}) - \nabla_\theta L_{S,T}(\theta^*(\bar{\theta}),\bar{\theta})\big\| \\ &\quad \leq \frac{1}{ST}\sum_{i=1}^{S}\sum_{t=2}^{T} \sup_{\theta\in\Theta}\sup_{\bar{\theta}\in\Theta} \big\|\nabla_\theta \log p\big(y_{i,t}(\bar{\theta})|\hat{f}_{i,t}(\theta,\bar{\theta});\theta\big) - \nabla_\theta \log p\big(y_{i,t}(\bar{\theta})|f_{i,t}(\theta,\bar{\theta});\theta\big)\big\| \\ &\quad \leq \frac{1}{ST}\sum_{i=1}^{S}\sum_{t=2}^{T} \sup_{\theta\in\Theta}\sup_{\bar{\theta}\in\Theta}\sup_{f^{(0:1)}} \big\|\nabla^2_{\theta f^{(0:1)}} \log p\big(y_{i,t}(\bar{\theta})|\hat{f}_{i,t}(\theta,\bar{\theta});\theta\big)\big\| \\ &\qquad\qquad\qquad \times \sup_{\theta\in\Theta}\sup_{\bar{\theta}\in\Theta} \big\|\nabla_\theta^{(0:1)}\hat{f}_{i,t}(\theta,\bar{\theta}) - \nabla_\theta^{(0:1)} f_{i,t}(\theta,\bar{\theta})\big\| \xrightarrow{a.s.} 0. \end{aligned}$$

This follows by an application of Lemma 2.1 in Straumann and Mikosch (2006) since the first factor on the right hand side of the above inequality is stationary and ergodic with a logarithmic moment (Assumption 4.9) and the second factor vanishes e.a.s. by Assumption 4.8.

The uniform convergence of the inverse Hessian in (22) is obtained by establishing the uniform convergence of the Hessian to a non-singular limit $\mathbb{E}\nabla^2_{\theta\theta}\log p\big(y_{i,t}(\bar{\theta})|f_{i,t}(\theta,\bar{\theta});\theta\big)$ (Assumption 4.10), that is,

$$\sup_{\theta\in\Theta}\sup_{\bar{\theta}\in\Theta}\left\|\Big(\nabla^2_{\theta\theta}\hat{L}_{S,T}(\theta,\bar{\theta})\Big)^{-1}\right\| \overset{a.s.}{\to} c\neq 0$$
$$\Leftarrow \quad \sup_{\theta\in\Theta}\sup_{\bar{\theta}\in\Theta}\left\|\nabla^2_{\theta\theta}\hat{L}_{S,T}(\theta,\bar{\theta})-\mathbb{E}\nabla^2_{\theta\theta}\log p\big(y_{i,t}(\bar{\theta})|f_{i,t}(\theta,\bar{\theta});\theta\big)\right\| \overset{a.s.}{\to} 0.$$

The uniform convergence above is shown as follows. First, we obtain that

$$\begin{aligned}
&\sup_{\theta\in\Theta}\sup_{\bar{\theta}\in\Theta}\left\|\nabla^2_{\theta\theta}\hat{L}_{S,T}(\theta,\bar{\theta})-\mathbb{E}\nabla^2_{\theta\theta}\log p\big(y_{i,t}(\bar{\theta})|f_{i,t}(\theta,\bar{\theta});\theta\big)\right\| \\
&\qquad\leq \sup_{\theta\in\Theta}\sup_{\bar{\theta}\in\Theta}\left\|\nabla^2_{\theta\theta}\hat{L}_{S,T}(\theta,\bar{\theta})-\nabla^2_{\theta\theta}L_{S,T}(\theta,\bar{\theta})\right\| \\
&\qquad\qquad + \sup_{\theta\in\Theta}\sup_{\bar{\theta}\in\Theta}\left\|\nabla^2_{\theta\theta}L_{S,T}(\theta,\bar{\theta})-\mathbb{E}\nabla^2_{\theta\theta}\log p\big(y_{i,t}(\bar{\theta})|f_{i,t}(\theta,\bar{\theta});\theta\big)\right\|.
\end{aligned} \tag{24}$$

The second term on the right hand side of inequality (24) vanishes a.s. to zero by an application of the ergodic theorem of Rao (1962), since $\nabla^2_{\theta\theta}\log p\big(y_{i,t}(\bar{\theta})|f_{i,t}(\theta,\bar{\theta});\theta\big)$ has a uniformly bounded moment by Assumption 4.9. As concerns the first term on the right hand side of inequality (24), we obtain that

$$\begin{aligned}
&\sup_{\theta\in\Theta}\sup_{\bar{\theta}\in\Theta}\|\nabla^2_{\theta\theta}\hat{L}_{S,T}(\theta,\bar{\theta})-\nabla^2_{\theta\theta}L_{S,T}(\theta,\bar{\theta})\| \\
&\quad\leq \frac{1}{ST}\sum_{i=1}^{S}\sum_{t=2}^{T}\sup_{\theta\in\Theta}\sup_{\bar{\theta}\in\Theta}\left\|\nabla^2_{\theta\theta}\log p\big(y_{i,t}(\bar{\theta})|\hat{f}_{i,t}(\theta,\bar{\theta}),\theta\big)-\nabla^2_{\theta\theta}\log p\big(y_{i,t}(\bar{\theta})|f_{i,t}(\theta,\bar{\theta});\theta\big)\right\| \\
&\quad\leq \frac{1}{ST}\sum_{i=1}^{S}\sum_{t=2}^{T}\sup_{\theta\in\Theta}\sup_{\bar{\theta}\in\Theta}\sup_{f^{(0:2)}}\left\|\nabla^3_{\theta\theta f^{(0:2)}}\log p(y_{i,t}(\bar{\theta}),\hat{f}_{i,t}(\theta,\bar{\theta}),\theta)\right\| \\
&\qquad\qquad\times\sup_{\theta\in\Theta}\sup_{\bar{\theta}\in\Theta}\left\|\nabla^{(0:2)}_{\theta}\hat{f}_{i,t}(\theta,\bar{\theta})-\nabla^{(0:2)}_{\theta}f_{i,t}(\theta,\bar{\theta})\right\| \overset{a.s.}{\to} 0,
\end{aligned}$$

where the a.s. convergence to zero is obtained by an application of Lemma 2.1 in Straumann and Mikosch (2006). In particular, the first factor on the right hand side of the above inequality is strictly stationary and ergodic with a bounded logarithmic moment (Assumption 4.9) and the second factor vanishes e.a.s. by Assumption 4.8. This concludes the proof of the lemma. □

References

Andrews, D. W. (1991). An empirical process central limit theorem for dependent non-identically distributed random variables. *Journal of Multivariate Analysis*, 38(2):187 – 203.

Blasques, F., Gorgi, P., Koopman, S. J., and Wintenberger, O. (2016a). Feasible invertibility conditions and maximum likelihood estimation for observation-driven models. Technical report, Tinbergen Institute Discussion Papers 16-082/III.

Blasques, F., Koopman, S. J., and Lucas, A. (2014a). Maximum likelihood estimation for generalized autoregressive score models. Technical report, Tinbergen Institute Discussion Papers 14-029/III.

Blasques, F., Koopman, S. J., and Lucas, A. (2014b). Stationarity and ergodicity of univariate generalized autoregressive score processes. *Electronic Journal of Statistics*, 8(1):1088–1112.

Blasques, F., Koopman, S. J., Lucas, A., and Schaumburg, J. (2016b). Spillover dynamics for systemic risk measurement using spatial financial time series models. *Journal of Econometrics*, 195(2):211 – 223.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Bondon, P. and Bahamonde, N. (2012). Least squares estimation of ARCH models with missing observations. *Journal of Time Series Analysis*, 33(6):880–891.

Bougerol, P. (1993). Kalman filtering with random coefficients and contractions. *SIAM Journal on Control and Optimization*, 31(4):942–959.

Buccheri, G., Bormetti, G., Corsi, F., and Lillo, F. (2017). A score-driven conditional correlation model for noisy and asynchronous data: An application to high-frequency covariance dynamics. *Available at SSRN: https://ssrn.com/abstract=2912438*.

Cox, D. R. (1981). Statistical analysis of time series: some recent developments. *Scandinavian Journal of Statistics*, 8:93–115.

Creal, D., Koopman, S. J., and Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5):777–795.

Creal, D., Schwaab, B., Koopman, S. J., and Lucas, A. (2014). Observation-driven mixed-measurement dynamic factor models with an application to credit risk. *Review of Economics and Statistics*, 96(5):898–915.

Davis, R. A., Dunsmuir, W. T. M., and Streett, S. B. (2003). Observation-driven models for poisson counts. *Biometrika*, 90(4):777–790.

Delle Monache, D., Petrella, I., and Venditti, F. (2016). Adaptive state space models with applications to the business cycle and financial stress. *CEPR Discussion Paper (DP11599)*.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford University Press, 2nd edition.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007.

Gallant, R. and White, H. (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Cambridge University Press.

Gourieroux, C. and Monfort, A. (2017). Composite indirect inference with application to corporate risks. *Econometrics and Statistics*.

Gourieroux, C., Monfort, A., and Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics*, 8(S1):S85–S118.

Harvey, A. C. (2013). *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*. New York: Cambridge University Press.

Harvey, A. C. and Luati, A. (2014). Filtering with heavy tails. *Journal of the American Statistical Association*, 109(507):1112–1122.

Harvey, A. C. and Pierse, R. G. (1984). Estimating missing observations in economic time series. *Journal of the American Statistical Association*, 79(385):125–131.

Jones, R. H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics*, 22(3):389–395.

Kohn, R. and Ansley, C. F. (1986). Estimation, prediction, and interpolation for ARIMA models with missing data. *Journal of the American Statistical Association*, 81(395):751–761.

Koopman, S. J., Lit, R., and Lucas, A. (2015). Intraday stock price dependence using dynamic discrete copula distributions. *Tinbergen Institute Discussion Papers TI 15-037/III/ DSF90*.

Lucas, A., Opschoor, A., and Schaumburg, J. (2016). Accounting for missing values in score-driven time-varying parameter models. *Economics Letters*, 148:96 – 98.

Nelson, D. B. (1991). Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica*, 59(2):347–370.

Pigott, T. D. (2001). A review of methods for missing data. *Educational research and evaluation*, 7(4):353–383.

Potscher, B. M. and Prucha, I. R. (1997). *Dynamic nonlinear econometric models: Asymptotic theory*. Springer-Verlag Berlin Heidelberg.

Rao, R. R. (1962). Relations between weak and uniform convergence of measures with applications. *The Annals of Mathematical Statistics*, 33(2):659–680.

Salvatierra, I. D. L. and Patton, A. J. (2015). Dynamic copula models and high frequency data. *Journal of Empirical Finance*, 30:120–135.

Straumann, D. and Mikosch, T. (2006). Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach. *The Annals of Statistics*, 34(5):2449–2495.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42.

White, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge Books. Cambridge University Press.

Wooldridge, J. (1986). *Asymptotic Properties of Econometric Estimators*. PhD thesis, Department of Economics, University of California, San Diego.