The Annals of Statistics
2022, Vol. 50, No. 5, 3039–3063
https://doi.org/10.1214/22-AOS2220

ESTIMATION OF TIME SERIES MODELS USING RESIDUALS DEPENDENCE MEASURES

BY CARLOS VELASCO[a]

Department of Economics, Universidad Carlos III de Madrid, [a]*carlos.velasco@uc3m.es*

We propose new estimation methods for time series models, possibly noncausal and/or noninvertible, using serial dependence information from the characteristic function of model residuals. This allows to impose the *i.i.d.* or martingale difference assumptions on the model errors to identify the unknown location of the roots of the lag polynomials for ARMA models without resorting to higher order moments or distributional assumptions. We consider generalized spectral density and cumulative distribution functions to measure residuals dependence at an increasing number of lags under both assumptions and discuss robust inference to higher order dependence when only mean independence is imposed on model errors. We study the consistency and asymptotic distribution of parameter estimates and discuss efficiency when different restrictions on error dependence are used simultaneously, including serial uncorrelation. Optimal weighting of continuous moment conditions yields maximum likelihood efficiency under independence for unknown error distribution. We investigate numerical implementation and finite sample properties of the new classes of estimates.

1. Introduction. Most dynamic models are based on sequences of errors or innovations that satisfy some form of unpredictability, such as serial statistical independence or conditional moment independence given past information, which implies independence of a given order imposed on joint moments. In particular, the conditional mean independence or martingale difference condition implies second-order independence, that is, serial uncorrelation or white noise property. These dependence conditions on model errors permit to describe the dynamic properties of the observed data and provide identification of model parameters under general conditions. As a result, time series model specification testing is often based on checking that residuals satisfy these forms of unpredictability or necessary conditions for them, such as uncorrelation, which can also be sufficient for model identification under some further restrictions (e.g., invertibility and causality).

For parameter estimation, the serial independence condition (together with the identical distribution assumption) is exploited to construct the likelihood function. In absence of distributional assumptions on model errors, Gaussian Pseudo Maximum Likelihood (PML) estimates based on least squares are typically prescribed. The Gaussian PML estimates try in fact to match data sample autocovariances with the model implied ones, or equivalently, minimize the magnitude of residuals autocorrelations to match the zero serial correlation white noise assumption, which only under Gaussianity is equivalent to serial independence. Conditional moments based models lead to unconditional moment restrictions using the uncorrelation of errors with past information described by instrumental variables (see, e.g., the survey by Anatolyev (2007)). These instruments are constructed with lags of observations and/or residuals, though these alternative representations of past information are not equivalent in general, for instance, when the true model is noninvertible.

Received January 2022; revised July 2022.
MSC2020 subject classifications. Primary 62M10; secondary 62M15.
Key words and phrases. Characteristic function, martingale difference, generalized method of moments, generalized spectral density, noncausal processes, noninvertible processes.

In this regard, a fundamental drawback of these methods based on correlations or second-order moments is that they cannot discriminate between uncorrelated residuals generated by alternative stationary representations of the model. However, residuals are not serially independent nor martingale difference sequences for the wrong representations, even if the original non-Gaussian errors satisfy either of these conditions, because they become nonlinearly predictable (Rosenblatt (2000), Section 5.4). Identifying the true representation to recover the original errors with the required independence conditions and the true impulse response function is key in many applications, including the design of optimal estimates and predictions, which can be nonlinear. For instance, noncausal processes can display nonlinear or explosive dynamics very different from the linearity of their causal version sharing the same autocorrelation function; see, for example, Gouriéroux and Zakoïan (2017), while noninvertibility questions the approximation of time series by long autoregressions, for example, Lippi and Reichlin (1994).

This identification problem can be alleviated by the use of non-Gaussian likelihoods or approximations as in Lii and Rosenblatt (1992, 1996), Huang and Pawitan (2000) and Lanne and Saikkonen (2011) or nonquadratic loss functions, based on, for example, LAD as in Breidt, Davis and Trindade (2001) and ranks as in Andrews, Davis and Breidt (2007). Alternatively, error independence of finite order can be imposed using higher order cumulants and spectral densities, permitting to identify general dynamic models which encompass nonminimum phase (i.e., noncausal or noninvertible) representations, both in the time domain, for example, Ramsey and Montenegro (1992) and Gospodinov and Ng (2015), and in the frequency domain, for example, Lii and Rosenblatt (1982) and Velasco and Lobato (2018). This last article also showed that this additional information from higher order dynamics can lead to efficiency improvements over Gaussian PML estimates.

While residual uncorrelation tests such as Box and Pierce (1970) are standard in empirical goodness-of-fit analysis, there is an increasing number of proposals to check serial dependence hypotheses on observed time series or on model residuals beyond uncorrelation; see, for example, the survey of Tjøstheim, Otneim and Støve (2022). Some methods check general dependence of a finite vector of observations at all possible lags using characterizing families of transformations of the data that are able to describe any type of functional relationship between different random variables. Pinkse (1998), Hong (1999), Hong and Lee (2005), Escanciano and Velasco (2006a, 2006b) consider the characteristic and cumulative probability distribution functions based on the exponential and indicator transformations, but many others are possible; see Stinchcombe and White (1998).

In this paper, we consider estimation of dynamic models based on independence conditions of the model errors, such as serial independence (or independent and identically distributed, *i.i.d.*, for stationary sequences), conditional mean independence (martingale difference sequences or *mds*) and uncorrelation (white noise sequences or *wns*). The first two conditions guarantee identification of models in situations where second-order moments are not able to discriminate between alternative representations and also could lead to efficiency improvements for non-Gaussian time series over least squares methods when considered alone or together with other restrictions. Further, our estimates are consistent without requiring additional distributional assumptions on the model errors apart from some finite moment and can achieve ML efficiency under an appropriate weighting of continuous moment conditions.

To measure the dependence in model residuals for a given parameter value, instead of using second or higher order moments through usual or higher order autocorrelations and spectral density functions, we employ generalized spectral densities as proposed by Hong (1999) for dependence testing on observed data and Hong and Lee (2005) for model residuals. Minimum distance loss functions compare the integrated empirical generalized spectral density and distribution functions of model residuals with the restricted estimates under an independence

assumption. This approach exploits similar dependence measures to the ones used to test independence or *mds* and other related hypotheses deduced from the relationship between the (joint) characteristic function derivatives and (multivariate) moments, and have closed-form expressions for integrable kernels. Fokianos and Pitsillou (2018) and Yao, Zhang and Shao (2018) make similar testing proposals based on the distance covariance of Székely, Rizzo and Bakirov (2007). Our measures also help to alleviate strong moment conditions associated to dependence descriptions based on higher order moments. Further, cumulative measures also avoid the choice of smoothing parameters necessary for pointwise consistency of spectral density estimates.

We investigate under general forms of higher order dependence, such as conditional heteroskedasticity, asymptotic properties and robust inference for parameter estimates developed under the *mds* restriction. We also consider overidentified generalized method of moments (GMM) estimation exploiting different but compatible dependence characterizations. In this way, there is no information loss in the Gaussian case when serial uncorrelation is imposed directly, while permitting efficiency improvements for non-Gaussian series.

Our methods focus on dependence of model residuals measured through their joint empirical characteristic function, but we do not specify any parametric model for this function as previous proposals trying to resemble ML estimation based on blocks of data; see, for example, Feuerverger (1990) and Knight and Yu (2002). Also, in contrast to Carrasco, Chernov, Florens and Ghysels (2007), we do not specify our moment conditions in terms of a closed-form expression of the conditional *cf* of observed data given past observations (or on a simulated joint *cf* when the dynamic model is not Markov), but focus directly on residuals. By estimating the residuals pairwise dependence at an increasing number of lags, our estimates are easier to compute in the non-Markovian case, do not require any distributional assumption and can be optimally continuously weighted to also achieve the ML efficiency under independence as explored by Gassiat (1993) for noncausal autoregressions using Kreiss (1987) adaptive estimation methods.

The rest of the paper is organized as follows. Section 2 presents the model and the dependence measures based on the characteristic function. Section 3 investigates model identification while Section 4 describes the asymptotic properties of parameter estimates based on the *mds* assumption. Section 5 discusses GMM estimation based on several dependence restrictions, including *i.i.d.* and white noise, and Section 6 discusses optimal continuous GMM estimation. Section 7 contains a simulation study of finite sample properties of our methods and further computational details.

2. Time series models and characteristic function based residuals dependence measures. We assume that the observed time series Y_t is generated by

$$Y_t = \mu_0 + \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{t-j}, \tag{1}$$

where ε_t is a stationary sequence with zero mean, which we will assume to be an independent identically distributed (*i.i.d.*) or martingale difference sequence (*mds*), but might have no finite variance. Summability conditions on ψ_j together with moment conditions on ε_t guarantee stationarity of Y_t, for example, $\sum_{j=-\infty}^{\infty} |\psi_j|^a < \infty$ and $\mathbb{E}|\varepsilon_t|^a < \infty$ for $a = 1, 2$ under *i.i.d.* and *mds*, respectively. Note that the two-sided summation in (1) allows Y_t to be noncausal. We also allow for noninvertible dynamics when inverting the linear filter $\psi(L) = \sum_{j=-\infty}^{\infty} \psi_j L^j$ to recover $\varepsilon_t = \sum_{j=-\infty}^{\infty} \psi_j^{(-1)}(Y_{t-j} - \mu_0)$ with the inverse filter $\psi^{-1}(L) := 1/\psi(L) = \sum_{j=-\infty}^{\infty} \psi_j^{(-1)} L^j$ using possibly future and past values of Y_t.

A model $\psi(\theta; L) = \sum_{j=-\infty}^{\infty} \psi_j(\theta) L^j$ establishes a structure on the coefficients $\psi_j(\theta)$ in terms of some parameter vector $\theta \in \Theta \subset \mathbb{R}^q$, so the target is the estimation of θ. The primary example are ARMA(q_1, q_2) models where

$$\alpha(L)(Y_t - \mu_0) = \beta(L)\varepsilon_t \tag{2}$$

and the polynomials $\alpha(L) = 1 - \sum_{j=1}^{q_1} \alpha_j L$ and $\beta(L) = 1 + \sum_{j=1}^{q_2} \beta_j L$ are of order q_1 and q_2, respectively, have all their roots away from the unit circle, inside or outside it, and do not have any common roots. In this case, the model parameters are $\theta = (\alpha_1, \ldots, \alpha_{q_1}, \beta_1, \ldots, \beta_{q_2})'$ and $\Theta \subset \{\theta \in \mathbb{R}^{q_1+q_2} : \alpha(z)\beta(z) \neq 0 \text{ for } |z| = 1, |\alpha(z)| + |\beta(z)| > 0 \text{ for all } z \in \mathbb{C}, \alpha_{q_1} \neq 0, \beta_{q_2} \neq 0\}$. Other parameterizations of $\psi(\theta)$ include the Bloomfield (1973) exponential model, fractional models, Hosking (1981) and noncausal autoregressions, Lanne and Saikkonen (2011).

In practice, given a parametric model $\psi(\theta; L)$, we compute residuals for any given (θ, μ),

$$\varepsilon_t(\theta; \mu) = \psi^{-1}(\theta; L)(Y_t - \mu),$$

where both $\psi^{-1}(\theta; L)$ and $\psi(\theta; L)$ are at least square summable for all $\theta \in \Theta$ and can include lags and leads. Denoting by θ_0 and μ_0 the true value of the parameters, that is, $\psi(L) = \psi(\theta_0; L)$, we have that $\varepsilon_t(\theta_0; \mu_0) = \psi^{-1}(\theta_0; L)(Y_t - \mu_0) = \varepsilon_t$. In this paper, we use generalized spectral densities based on the characteristic function (*cf*) transformation, Hong (1999), to measure the serial dependence in the residuals $\varepsilon_t(\theta; \mu)$ and check the suitability of candidate values of θ and μ. These dependence measures lead to population loss functions to identify θ, which are invariant to centering by μ_0 and μ, so we do not consider identification of μ_0, whose estimation can be pursued using the sample mean of observations, $\bar{Y}_T = n^{-1} \sum_{t=1}^{T} Y_t$. Then we denote by $\varepsilon_t(\theta) = \varepsilon_t(\theta; \mu_0)$ the residuals obtained with $\mu = \mu_0$ and generated by

$$\varepsilon_t(\theta) = \psi^{-1}(\theta; L)\psi(\theta_0; L)\varepsilon_t = \phi(\theta; L)\varepsilon_t$$

for $\phi(\theta; L) := \psi^{-1}(\theta; L)\psi(\theta_0; L)$.

For a stationary time series of residuals $\varepsilon_t(\theta)$ with marginal *cf* given by $\varphi_\theta(u) = \varphi_{\varepsilon_t(\theta)}(u) = \mathbb{E}[e^{iu\varepsilon_t(\theta)}]$, we define the pairwise *cf* of $(\varepsilon_t(\theta), \varepsilon_{t-j}(\theta))$ by $\varphi_{\theta,j}(u, v) = \mathbb{E}[e^{i(u\varepsilon_t(\theta) + v\varepsilon_{t-j}(\theta))}]$ for $j = 0, \pm 1, \ldots$, $i = \sqrt{-1}$ and $(u, v) \in \mathbb{R}^2$. Let $\sigma_{\theta,j}(u, v)$ be the covariance between $e^{iu\varepsilon_t(\theta)}$ and $e^{iv\varepsilon_{t-|j|}(\theta)}$,

$$\sigma_{\theta,j}(u, v) = \operatorname{Cov}\big(e^{iu\varepsilon_t(\theta)}, e^{iv\varepsilon_{t-|j|}(\theta)}\big) = \varphi_{\theta,|j|}(u, v) - \varphi_\theta(u)\varphi_\theta(v).$$

Then $\sigma_{\theta,j}(u, v) = 0$ for all $(u, v) \in \mathbb{R}^2$ and $j \neq 0$ if and only if $\varepsilon_t(\theta)$ and $\varepsilon_{t-|j|}(\theta)$ are pairwise independent, capturing all types of dependence, including those described by autocorrelations and higher order moments.

Assuming that $\sup_{(u,v)\in\mathbb{R}^2} \sum_{j=-\infty}^{\infty} |\sigma_{\theta,j}(u, v)| < \infty$ if dependence decays fast enough with j, we can define the generalized spectral density of $\varepsilon_t(\theta)$ as the Fourier transform of $\sigma_{\theta,j}(u, v)$, that is,

$$f_\theta(\omega, u, v) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \sigma_{\theta,j}(u, v) e^{-ij\omega}, \quad \omega \in [-\pi, \pi],$$

which describes pairwise dependence at all lags without assuming finite moments of any order or smoothness conditions on the distribution of $\varepsilon_t(\theta)$, so that for all (ω, u, v) under independence

$$f_\theta(\omega, u, v) = f_{\theta,i.i.d.}(\omega, u, v) := \frac{1}{2\pi} \sigma_{\theta,0}(u, v).$$

Sufficient weak dependence and moment assumptions guarantee the existence of the following generalized partial derivatives of the spectral density $f_\theta(\omega, u, v)$ of order (m, ℓ) with respect to (u, v),

$$f_\theta^{(m,\ell)}(\omega, u, v) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \sigma_{\theta,j}^{(m,\ell)}(u, v) e^{-ij\omega},$$

where $\sigma_{\theta,j}^{(m,\ell)}(u, v) = (\partial^{m+\ell}/\partial u^m \partial v^\ell)\sigma_{\theta,j}(u, v)$. In particular, setting $(m, \ell) = (1, 0)$ and $u = 0$, $\sigma_{\theta,j}^{(1,0)}(0, v) = \operatorname{Cov}(i\varepsilon_t(\theta), e^{iv\varepsilon_{t-|j|}(\theta)}) = 0$ for all $v \in \mathbb{R}$ and all $j \neq 0$ if $\varepsilon_t(\theta)$ is a martingale difference sequence, so that for all (ω, v) under *mds*,

$$f_\theta^{(1,0)}(\omega, 0, v) = f_{\theta,mds}^{(1,0)}(\omega, 0, v) := \frac{1}{2\pi}\sigma_{\theta,0}^{(1,0)}(0, v).$$

Similarly, we can define the spectral cumulative distribution function (*cdf*),

$$F_\theta(\omega, u, v) = 2\int_0^\omega f_\theta(\lambda, u, v)\, d\lambda = \sum_{j=-\infty}^{\infty} \sigma_{\theta,j}(u, v) \frac{\sin j\omega}{j\pi}, \quad \omega \in [0, \pi],$$

to describe serial dependence in the sequence $\varepsilon_t(\theta)$ up to a given frequency ω.

To investigate the properties of these residual dependence measures, we introduce the following assumption. Let denote by $C > 0$ a generic bounded constant.

ASSUMPTION 1. Θ is assumed compact and let $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$ be the complex unit circle. Assume $\psi(\theta, z) \in \mathbb{L}^2(\mathbb{T})$ and $\psi^{-1}(\theta, z) \in \mathbb{L}^2(\mathbb{T})$ for every $\theta \in \Theta$ so that for some $\eta_0 > 1/2$,

$$\sup_{\theta\in\Theta}|\psi_j(\theta)| + \sup_{\theta\in\Theta}|\psi_j^{(-1)}(\theta)| \leq C|j|^{-\eta_0}, \quad j = \pm 1, \pm 2, \ldots.$$

Assumption 1 implies the same summability restriction for the coefficients $\phi_j(\theta)$ of the filter $\phi(\theta; z)$ and when $\eta_0 > 1$ together with ε_t *i.i.d.*, zero mean and $\mathbb{E}|\varepsilon_t| < \infty$, guarantees that Y_t and $\varepsilon_t(\theta)$ are (strictly and first-order) stationary and that $\varepsilon_t(\theta)$ is weakly dependent with $|\sigma_{\theta,j}(u, v)| \to 0$ as $j \to \infty$ for each θ, u and v as we now show.

LEMMA 1. *Under Assumption 1, $\eta_0 > 1$, ε_t i.i.d., zero mean, $\mathbb{E}|\varepsilon_t| < \infty$,*

$$\sup_{\theta\in\Theta}|\sigma_{\theta,j}(u, v)| \leq C(|u| + |v|)j^{1-\eta_0} \quad \text{as } j \to \infty.$$

Proofs of results are contained in Appendix A of the Supplementary Material (Velasco (2022)) with technical lemmas compiled in its Appendix B. Note that $\sigma_{\theta,j}(u, v) = \operatorname{Cov}[z_t(\theta; u), z_{t-|j|}(\theta; v)]$ and that, under causality, the sequence $z_t(\theta; u) := e^{iu\varepsilon_t(\theta)} - \varphi_\theta(u)$ is strong mixing for each θ and u under the summability condition on $\psi_j(\theta)$ in Assumption 1 ($\eta_0 > 1$), a δ-moment condition on the *i.i.d.* ε_t, $\delta > 1$, and an integral Lipschitz condition on the probability density function (*pdf*) of ε_t; see Gorodetskii (1977). Similar results are found by Rosenblatt ((2000), Section 4.4) for processes with two-sided representations. Then the results of Hong (1999) on $\sigma_{\theta,j}$ could be easily transposed pointwise (i.e., for each θ). We use instead the weak dependent processes framework of Doukhan and Louhichi (1999, Lemma 9) to bound the covariances of centered finite functions of two-sided linear filters of *i.i.d.* sequences like $z_t(\theta; u)$; see also Lemma 3.1 in Dedecker et al. (2007).

Using a similar approach, we show in next lemma that the *mds* condition on the sequence ε_t together with two finite moments are sufficient to evaluate the rate of decay of $\sigma_{\theta,j}^{(1,0)}(0, v)$ uniformly in θ as j grows without need of mixing conditions. This result also guarantees that the population loss functions exploiting the *mds* property are well defined under our assumptions. Denote by I_{t-1} the past information of the sequence ε_t, $I_{t-1} = \sigma(\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots)$.

LEMMA 2. *Under Assumption* 1, $\eta_0 > 1/2$, ε_t *stationary* mds, $\mathbb{E}[\varepsilon_t | I_{t-1}] = 0$ *and* $\mathbb{E}\varepsilon_t^2 < \infty$,

$$\sup_{\theta \in \Theta} |\sigma_{\theta,j}^{(1,0)}(0, v)| \leq C|v| j^{1/2-\eta_0} \quad \text{as } j \to \infty.$$

The rate of decay of the *mds* dependence measures obtained in this lemma is faster for the same η_0 compared to the one in Lemma 1 for general dependence measures $\sigma_{\theta,j}$ because it exploits two finite moments instead of just one, only restricting the ψ_j to be square summable.

3. Model identification.

3.1. *Identification under serial independence.* Second-order dynamic properties of residuals $\varepsilon_t(\theta)$ are not sufficient to identify the true value of $\theta \in \Theta$ for some parametric models, so stronger assumptions have to be imposed on the error sequence. The criteria we propose to identify θ in this paper are L_2 norms between the generalized spectral density or cumulative distribution functions of the sequence of residuals based on $\sigma_{\theta,j}(u, v)$ and its derivatives, and the restricted versions under the serial independence assumptions.

We now provide conditions to guarantee that for any $\theta \neq \theta_0$ the sequence of residuals $\varepsilon_t(\theta) = \phi(\theta; L)\varepsilon_t$ is not serially pairwise independent if ε_t is non-Gaussian. This holds despite $\varepsilon_t(\theta)$ could be serially uncorrelated when $\psi(\theta; L) = a(L)\psi(\theta_0; L)$ for an all-pass or Blaschke factor a satisfying $a(z) = a^{-1}(z^{-1})$; see Hannan (1970, pp. 65–67). An all-pass filter is a product of ratios of possibly nonminimum phase lag polynomials with denominator roots which are the reciprocals of the numerator ones, implying that residuals have been generated by a different representation $\psi(\theta; L)$ of the true model which shifts the roots location but generates the same spectral density as $|a(e^{i\omega})|^2 = 1$.

ASSUMPTION 2.

1. For all $\theta \in \Theta$, $\theta \neq \theta_0$, $\phi(\theta; z) = \psi^{-1}(\theta; z)\psi(\theta_0; z) \neq c_0 z^{j_0}$ for any $j_0 \in \mathbb{Z}$ and $c_0 \in \mathbb{R}\backslash\{0\}$ in a subset of positive measure of $\mathbb{T}$ and $\psi(\theta_0; z)$ has only simple zeros or poles for $|z| \neq 0$.
2. If $|\phi(\theta; z)|^2$ is constant for some $\theta \neq \theta_0$ a.e. in $\mathbb{T}$, then ε_t is non-Gaussian with some finite and different from zero higher order cumulant $\kappa_a^{\varepsilon} \neq 0$, $a = 3, 4, \ldots$.

Assumption 2.1 ensures that the linear representation is essentially unique by discarding parameterizations that can generate *i.i.d.* residuals by a rescaling and time shift of the original error sequence, that is, $\varepsilon_t(\theta) = c_0\varepsilon_{t-j_0}$; see Cheng (1992). It is satisfied by ARMA models with no unit or multiple roots. We make no other assumptions on the location of the (nonunit) roots of the lag polynomials $\psi(\theta; z)$ and $\psi^{-1}(\theta; z)$ with respect to the unit circle, unless we allow for Gaussian ε_t. Note that it is not possible to identify a unique solution for Gaussian processes when, for example, both causal and noncausal representations are allowed. If these different representations are possible for some $\theta \in \Theta$, then ε_t is presumed non-Gaussian in Assumption 2.2 to avoid that uncorrelated residuals $\varepsilon_t(\theta)$, generated by $\theta \neq \theta_0$ for which $|\phi(\theta; z)|^2$ and its spectral density are constant, become full independent as happens under Gaussianity.

For $\theta \neq \theta_0$, the residuals $\varepsilon_t(\theta)$ necessarily depend on several values of ε_t under Assumption 2.1 and, therefore, are serially correlated except if $\phi(\theta; L)$ is an all-pass filter. However, from the analysis of Section 5.4 of Rosenblatt (2000), concluding that the optimal mean square prediction of nonminimum phase AR and MA processes with non-Gaussian innovations is nonlinear, we can expect that the $\varepsilon_t(\theta)$ cannot form a serially (pairwise) independent sequence even if they are uncorrelated. We now provide a result that confirms this claim.

LEMMA 3. *For an uncorrelated ARMA process X_t generated by a nonconstant all-pass $\psi(z)$ with nonunit simple lag polynomial roots and non-Gaussian i.i.d. innovations ε_t with some finite and different from zero higher order cumulant $\kappa_a^\varepsilon \neq 0$, $a = 3, 4, \ldots$, the pairwise conditional expectations $\mathbb{E}[X_t|X_{t-j}]$, $j = 1, 2, \ldots$, are not all linear functions of X_{t-j} and, therefore, are not all zero.*

This lemma extends Rosenblatt (2000) argument in two important directions. First, we consider explicitly predictions by a single element of the past, that is, pairwise dependence, and not only predictions given all infinite past observations, $\mathbb{E}[X_t|X_{t-1}, X_{t-2}, \ldots]$. Second, we consider serially uncorrelated ARMA processes with all-pass $\psi(z)$, which may contain both noninvertible and noncausal factors, not just serially correlated pure noncausal AR or pure noninvertible MA processes. See Chen, Choi and Escanciano (2017) for a similar result for multivariate all-pass noninvertible but causal processes when predicted by all infinite past.

For our identification purposes, the case when X_t is an uncorrelated residual process $\varepsilon_t(\theta)$ is key, as Lemma 3 shows that residuals generated by imposing any different representation of the true process with $\theta \neq \theta_0$ (flipping some of the lag polynomial roots around the unit circle) is not (full nor pairwise) *i.i.d.* nor *mds*. Note, however, that for such θ the pairwise and infinite-past optimal linear predictions of $\varepsilon_t(\theta)$ are zero and that the implied model representation is able to match the true autocorrelation function of Y_t, but not its complete dynamics described by the impulse response function $\psi(\theta_0; z) \neq \psi(\theta; z)$. We conjecture that this result holds without the restriction of $\psi(\theta_0; z)$ having only simple zeros or poles for $|z| \neq 1$ used also by Rosenblatt (2000).

Therefore, under Assumption 2 and for any continuous and increasing W with unbounded support, $\int |\sigma_{\theta,j}(u,v)|^2 dW(u,v) > 0$ for some $j \neq 0$ because the joint *cf* $\varphi_{\theta,j}(u,v)$ of $(\varepsilon_t(\theta), \varepsilon_{t-j}(\theta))$ does not factorize into the product of the marginals since $\mathbb{E}[\varepsilon_t(\theta)|\varepsilon_{t-j}(\theta)] \neq 0$. Thus, with Lemma 1 showing that $|\sigma_{\theta,j}(u,v)|$ decays fast enough with increasing j we can define the population L_2 distance between f_θ and $f_{\theta,i.i.d.}$,

$$\begin{aligned} Q_0^{i.i.d.}(\theta) &:= \int \int_{-\pi}^{\pi} \left| f_\theta(\omega,u,v) - \frac{1}{2\pi}\sigma_{\theta,0}(u,v) \right|^2 d\omega\, dW(u,v) \\ &= \frac{2}{\pi} \sum_{j=1}^{\infty} \int |\sigma_{\theta,j}(u,v)|^2 dW(u,v), \end{aligned}$$

where the second line follows by Parseval's equality, for a weighting function W satisfying the next assumption.

ASSUMPTION 3. $W(u,v) = W(u)W(v)$ where $W : \mathbb{R} \to \mathbb{R}^+$ is continuous, symmetric and increasing with unbounded support and $\int |u|^3 dW(u) < \infty$.

Assumption 3 on W is similar to the corresponding one used in Hong (1999) to argue for the consistency of serial dependence tests and is stronger than the nondecreasing with bounded total variation condition he used to derive the null asymptotic distribution of test statistics. We also introduce a moment condition on W, to control fluctuations of $|\sigma_{\theta,j}(u,v)|$ in u and v when using derivatives of the *cf*, and a factorization, to simplify numerical calculations and asymptotic analysis.

Then, under Assumptions 1–3, $Q_0^{i.i.d.}(\theta) > 0$ for $\theta \neq \theta_0$, and similarly $\mathcal{L}_0^{i.i.d.}(\theta) := L_2^2(F_\theta, F_{\theta,i.i.d.}) > 0$ for $\theta \neq \theta_0$, where

$$\mathcal{L}_0^{i.i.d.}(\theta) = \frac{2}{\pi} \sum_{j=1}^{\infty} j^{-2} \int |\sigma_{\theta,j}(u,v)|^2 dW(u,v).$$

We also note that for any μ_0 and μ, $\sigma_{\varepsilon(\theta;\mu),j}(u,v) = \exp\{i\psi^{(-1)}(\theta;1)(u+v)(\mu_0-\mu)\}\sigma_{\theta,j}(u,v)$ because centering only affects the *cf* of $\varepsilon_t(\theta;\mu)$ by a complex scaling in the unit circle. Then $|\sigma_{\varepsilon(\theta;\mu),j}(u,v)|^2 = |\sigma_{\theta,j}(u,v)|^2$ for all μ, showing that the identification criteria for θ based on the *i.i.d.* assumption are invariant to location of Y_t or $\varepsilon_t(\theta)$.

3.2. *Identification under a martingale difference assumption.* Lemma 3 shows that residuals $\varepsilon_t(\theta)$ for $\theta \neq \theta_0$ and non-Gaussian ε_t are not pairwise *i.i.d.* nor *mds* even when innovations are *i.i.d.*, so that we can expect that under Assumptions 2 and 3

$$\int \left|\sigma_{\theta,j}^{(1,0)}(0,v)\right|^2 dW(v) > 0 \quad \text{for some } j \neq 0 \tag{3}$$

for a large class of dynamic processes driven by *mds* innovations, including all that impose a particular causality and invertibility representation. Therefore, under (3), $Q_0^{mds}(\theta) > 0$ for $\theta \neq \theta_0$, where

$$Q_0^{mds}(\theta) := L_2^2(f_\theta^{(1,0)}, f_{\theta,mds}^{(1,0)}) = \frac{2}{\pi}\sum_{j=1}^{\infty}\int |\sigma_{\theta,j}^{(1,0)}(0,v)|^2 dW(v),$$

and similarly $\mathcal{L}_0^{mds}(\theta) := L_2^2(F_\theta^{(1,0)}, F_{\theta,mds}^{(1,0)}) = \frac{2}{\pi}\sum_{j=1}^{\infty} j^{-2}\int |\sigma_{\theta,j}^{(1,0)}(0,v)|^2 dW(v) > 0$ for $\theta \neq \theta_0$, which are well defined by Lemma 2.

By a similar argument as for the *i.i.d.* criteria, we can check for any μ_0 and μ that $\sigma_{\varepsilon(\theta;\mu),j}^{(1,0)}(0,v) = \exp\{i\psi^{(-1)}(\theta;1)v(\mu_0-\mu)\}\sigma_{\theta,j}^{(1,0)}(0,v)$ because the mean component is automatically eliminated by centering in the covariance definition, so that $|\sigma_{\varepsilon(\theta;\mu),j}^{(1,0)}(0,v)|^2 = |\sigma_{\theta,j}^{(1,0)}(0,v)|^2$ and *mds* identification criteria for θ are also invariant to location of Y_t or $\varepsilon_t(\theta)$.

3.3. *Interpretation of loss functions and data scaling.* The loss functions based on (generalized) autocovariances are not scale free. Replacing $\varepsilon_t(\theta)$ by $\sigma\varepsilon_t(\theta)$ for some $\sigma > 0$ can be interpreted naturally as a rescaling of the weighting measure W, which leads to a local analysis as $\sigma \to 0$. Focusing on $Q_0^{i.i.d.}(\theta)$ defined for $\sigma\varepsilon_t(\theta)$, we can expand the *cf* of $\sigma\varepsilon_t(\theta)$ around $u = 0$ assuming all moments exist and obtain

$$Q_0^{i.i.d.}(\theta;\sigma) = \frac{2}{\pi}\sum_{a,b,a',b'=0}^{\infty}\sigma^{a+b+a'+b'}\frac{i^{a+b-a'-b'}}{a!b!a'!b'!}w_{a+a'}w_{b+b'}\sum_{j=1}^{\infty}\gamma_{a,b}^{\theta}(j)\gamma_{a',b'}^{\theta}(j)$$

for $\gamma_{a,b}^{\theta}(j) := \mathrm{Cov}[\varepsilon_t^a(\theta), \varepsilon_{t-|j|}^b(\theta)]$ and $w_a := \int u^a dW(u) < \infty$ assuming that W has finite moments of all orders. Then, as $\sigma \to 0$, the first term in the expansion is in the square of $\gamma_{1,1}^{\theta}(j) = \mathrm{Cov}[\varepsilon_t(\theta), \varepsilon_{t-|j|}(\theta)]$, the usual covariance of residuals, while higher order cross-moments appear successively.

Then, for small σ, linear residual correlations dominate $Q_0^{i.i.d.}(\theta;\sigma)$, explaining why our estimates achieve ML efficiency for Gaussian data when $\sigma \to 0$, as will be formally showed in next sections. However, linear correlation does not contain all relevant information on dependence and other choices of σ or of the weighting may lead to more efficient estimates. For the *mds* criterion a similar interpretation is possible, but involving only the correlation of residuals $\varepsilon_t(\theta)$ in levels with powers of its lags, see Escanciano and Velasco (2006a).

4. Asymptotic properties of estimates based on the *mds* criterion. In this section, we explore the asymptotic properties of parameter estimates based on the *mds* assumption on model errors ε_t and how to robustify inference when in absence of serial independence there could be predictability in second moments (i.e., conditional dynamic heteroskedasticity) and in nonlinear functions of future ε_t, for example, $e^{iu\varepsilon_{t+j}}$, $j > 0$.

To impose the *mds* restriction on model residuals, we use empirical versions of the identifying population loss functions Q_0^{mds} and $\mathcal{L}_0^{mds}$. However, given a finite sample $\{Y_t\}_{t=1}^T$, we can only have an approximation to the population model residuals $\varepsilon_t(\theta)$ due to unknown centering and information truncation, that is, for each θ we can only compute

$$\hat{\varepsilon}_t(\theta) := \hat{\varepsilon}_t(\theta; \bar{Y}_T) := \psi^{-1}(\theta; L)(Y_t - \bar{Y}_T)\mathbf{1}\{1 \le t \le T\}$$

after centering by the sample mean. Given $\hat{\varepsilon}_t(\theta)$, we construct estimates

$$\hat{\sigma}_{\theta,j}(u, v) := \hat{\sigma}_{\hat{\varepsilon}_t(\theta),j}(u, v) = \hat{\varphi}_{\theta,|j|}(u, v) - \hat{\varphi}_{\theta,|j|}(u, 0)\hat{\varphi}_{\theta,|j|}(0, v)$$

with

$$\hat{\varphi}_{\theta,j}(u, v) = \frac{1}{T - |j|} \sum_{t=1+|j|}^{T} e^{i(u\hat{\varepsilon}_t(\theta) + v\hat{\varepsilon}_{t-j}(\theta))}$$

to obtain estimation criteria equal to the squared L_2 norm between the (derivatives of the) empirical generalized spectral density or cumulative distribution functions of the residuals $\hat{\varepsilon}_t(\theta)$ and their restricted versions under the different serial dependence hypotheses. After standard calculations, we can check that

$$\hat{\varepsilon}_t(\theta) = \varepsilon_t(\theta) + \upsilon_T(\theta) - \upsilon_{t,T}(\theta),$$

where the unknown centering term $\upsilon_T(\theta) := \psi^{(-1)}(\theta; 1)(\mu_0 - \bar{Y}_T)$ is constant in t, so does not contribute to the modulus of $\hat{\sigma}_{\theta,j}(u, v)$ as in the case of the population $\sigma_{\theta,j}(u, v)$. The information truncation term, $\upsilon_{t,T}(\theta) := (\sum_{j=-\infty}^{t-T-1} + \sum_{j=t}^{\infty})\psi_j^{(-1)}(\theta)(Y_{t-j} - \bar{Y}_T)$, is negligible asymptotically because the coefficients of $\psi^{-1}(\theta; L)$ decay fast uniformly in θ.

The empirical spectral densities and their derivatives are constructed following usual spectral analysis:

$$\hat{f}_{\theta,T}^{(m,\ell)}(\omega, u, v) = \frac{1}{2\pi} \sum_{j=1-T}^{T-1} \left(1 - \frac{|j|}{T}\right)^{1/2} k\left(\frac{j}{p}\right) \hat{\sigma}_{\theta,j}^{(m,\ell)}(u, v) e^{-ij\omega},$$

where $\hat{\sigma}_{\theta,j}^{(m,\ell)}(u, v) = (\partial^{m+\ell}/\partial u^m \partial v^\ell)\hat{\sigma}_{\theta,j}(u, v)$, $j = 0, \pm 1, \ldots, \pm(T-1)$. The kernel function k is a lag window with $k(0) = 1$ and p is a bandwidth or lag order required to increase slower than T for consistency but without impact on asymptotic results, while the factor $(1 - |j|/T)^{1/2}$ is introduced to improve finite sample properties of estimates. Similarly, estimates of the corresponding spectral *cdf* $F_\theta^{(m,\ell)}(\omega, u, v)$ can be devised,

$$\hat{F}_{\theta,T}^{(m,\ell)}(\omega, u, v) = \sum_{j=1-T}^{T-1} \hat{\sigma}_{\theta,j}^{(m,\ell)}(u, v) \frac{\sin j\omega}{j\pi},$$

for which it is not needed to use smoothing to achieve consistency.

When imposing the martingale difference assumption on ε_t but allowing for other type of higher order dependence, we use the derivatives of order $(1, 0)$ of the estimated spectral and covariance functions f and σ, which identify this hypothesis by means of

$$Q_T^{mds}(\theta) := L_2^2(\hat{f}_{\theta,T}^{(1,0)}, \hat{f}_{\theta,mds,T}^{(1,0)}) = \frac{2}{\pi} \sum_{j=1}^{T-1} k^2\left(\frac{j}{p}\right)\left(1 - \frac{|j|}{T}\right) \int |\hat{\sigma}_{\theta,j}^{(1,0)}(0, v)|^2 \, dW(v),$$

while when using the spectral *cdf* we set

$$\mathcal{L}_T^{mds}(\theta) := L_2^2(\hat{F}_{\theta,T}^{(1,0)}, \hat{F}_{\theta,mds,T}^{(1,0)}) = \frac{2}{\pi} \sum_{j=1}^{T-1} j^{-2} \int |\hat{\sigma}_{\theta,j}^{(1,0)}(0, v)|^2 \, dW(v),$$

where

$$\hat{\sigma}_{\theta,j}^{(1,0)}(0,v) := \frac{\partial}{\partial u}\hat{\sigma}_{\theta,j}(u,v)\Big|_{u=0} = \hat{\varphi}_{\theta,j}^{(1,0)}(0,v) - i\overline{\hat{\varepsilon}}_T(\theta)\hat{\varphi}_{\theta,j}(0,v),$$

with $\overline{\hat{\varepsilon}}_T(\theta) = (T-|j|)^{-1}\sum_{t=1+|j|}^{T}\hat{\varepsilon}_t(\theta) = (\partial/\partial u)\hat{\varphi}_{\theta,j}(u,0)|_{u=0}/i$, so the modulus of $\hat{\sigma}_{\theta,j}^{(1,0)}(0,v)$ is also invariant to location of $\hat{\varepsilon}_t(\theta)$, but not to information truncation when replacing $\varepsilon_t(\theta)$ by $\hat{\varepsilon}_t(\theta)$. The loss function Q_T^{mds} with local smoothing is the main ingredient of several test statistics for the martingale difference hypothesis when $\hat{\varepsilon}_t(\theta)$ are replaced by observed time series or by residuals $\hat{\varepsilon}_t = \varepsilon_t(\theta_T)$ evaluated at a particular parameter estimate θ_T (see Hong (1999), Hong and Lee (2005) and Chen et al. (2017)), while $\mathcal{L}_T^{mds}$ is similar to the test statistic used in Escanciano and Velasco (2006a) for the *mds* hypothesis.

The consistency analysis of parameter estimates under a correct parametrization assumption is based on uniform convergence of the empirical objective function to the theoretical counterpart that identifies properly the unique solution. We show the uniform convergence of $\mathcal{L}_T^{mds}(\theta)$ to $\mathcal{L}_0^{mds}(\theta)$ under fairly weak assumptions due to the improved convergence for large lags built in the definition of $\mathcal{L}_T^{mds}(\theta)$, and propose as initial estimation

$$\hat{\theta}_T^{mds} := \arg\min_{\theta\in\Theta}\mathcal{L}_T^{mds}(\theta).$$

To exploit potential efficiency improvements of estimates based on Q_T^{mds} over those based on $\mathcal{L}_T^{mds}$, we compute a Newton–Raphson step from $\hat{\theta}_T^{mds}$ to set

$$\tilde{\theta}_T^{mds} := \hat{\theta}_T^{mds} - \left(\frac{\partial^2}{\partial\theta\,\partial\theta'}Q_T^{mds}(\hat{\theta}_T^{mds})\right)^{-1}\frac{\partial}{\partial\theta}Q_T^{mds}(\hat{\theta}_T^{mds}).$$

We first show consistency of $\hat{\theta}_T^{mds}$ and then compare the asymptotic distributions of both estimates and obtain closed-form expressions for their asymptotic variances for Gaussian W and consistent standard errors. For that, we introduce further assumptions on the smoothness and weak dependence of the model and the innovations.

ASSUMPTION 4. The filter $\phi(\theta;z)$ is differentiable for all $\theta\in\Theta$ with derivative $\delta(\theta;z) := (\partial/\partial\theta)\phi(\theta;z) = \sum_{j=-\infty}^{\infty}\delta_j(\theta)z^j$ satisfying, for $\eta_1 > 1$,

$$\sup_{\theta\in\Theta}\|\delta_j(\theta)\| \le C|j|^{-\eta_1}, \quad j = \pm1, \pm2, \ldots.$$

ASSUMPTION 5. For some $\nu \ge 3$:

1. ε_t is a stationary *mds*, $\mathbb{E}[\varepsilon_t|I_{t-1}] = 0$, and $\mathbb{E}|\varepsilon_t|^\nu < \infty$.
2. ε_t is strong mixing with mixing coefficients satisfying $\sum_{j=1}^{\infty}\alpha(j)^{\frac{\nu-2}{\nu}} < \infty$.

Assumption 4 controls the rate of decay on the model scores, which are key in the asymptotic analysis. It is satisfied at once by ARMA models, possibly noncausal or noninvertible, for any $\eta_1 > 0$ when unit roots are excluded. The mixing condition in Assumption 5 on ε_t is used to control in the asymptotic analysis the nonlinear predictability on top of the absence of conditional mean dependence implied by the *mds* assumption. Together with a finite third absolute moment for ε_t, it further allows us to evaluate the variance of the estimates $\hat{\sigma}_{\theta,j}^{(1,0)}$ used to construct $\mathcal{L}_T^{mds}(\theta)$ and $Q_T^{mds}(\theta)$ and show that $\mathcal{L}_T^{mds}(\theta)$ converges to $\mathcal{L}_0^{mds}(\theta)$ uniformly for $\theta\in\Theta$. The mixing assumption could be replaced by conditions guaranteeing that ε_t could be approximated by a *mds* sequence, which is finite dependent as in Hong and Lee (2005).

THEOREM 1. *Under Assumptions 1–5, $\eta_0 > 2$, $\eta_1 > 1$ and (3), $\hat{\theta}_T^{mds} \to_p \theta_0$ as $T\to\infty$.*

To analyze the asymptotic distribution of estimates under a *mds* condition, we need to reinforce our moment and mixing conditions of Assumption 5 to evaluate higher order moments of estimates of residuals *cf* and its derivatives. We also need a further assumption on the differentiability of the model to analyse the asymptotic distribution of parameter estimates and on the kernel k and lag parameter p for Q_T^{mds} based estimates.

ASSUMPTION 6. For some $\nu \geq 5$:

1. ε_t is stationary *mds*, $\mathbb{E}[\varepsilon_t | I_{t-1}] = 0$, and $\mathbb{E}|\varepsilon_t|^\nu < \infty$.
2. ε_t is strong mixing with mixing coefficients satisfying $\sum_{j=1}^{\infty} j^2 \alpha(j)^{\frac{\nu-4}{\nu}} < \infty$.

ASSUMPTION 7. The filter $\phi(\theta; z)$ has three derivatives with $\phi^{(a,n)}(\theta; z) := (\partial/\partial\theta_n)^a \phi(\theta; z) = \sum_{j=-\infty}^{\infty} \phi_j^{(a,n)}(\theta) z^j$ so that for $\eta_a > 1$, $a = 1, 2, 3$ and $n = 1, \ldots, q$,

$$\sup_{\theta \in \Theta} |\phi_j^{(a,n)}(\theta)| \leq C|j|^{-\eta_a}, \quad j = \pm 1, \pm 2, \ldots.$$

ASSUMPTION 8.

1. $k : \mathbb{R} \to [-1, 1]$ is symmetric and continuous at 0 and all but a finite number of points, with $k(1) = 1$ and $|k(x)| \leq C|x|^{-b}$, $b \geq 1$, for large x, and $1 - k(x) = k_\tau |x|^\tau + o(x)$ as $x \to 0$ for some $\tau \in (0, \infty)$ and $k_\tau > 0$.
2. $1/p + p^2/T \to 0$ as $T \to \infty$.

Assumption 6 also implies the mixing condition used in Andrews (1991) for summability of the fourth-order cumulants and the conditions in Yoshihara (1978) to bound the fourth moment of sums of mixing processes. Assumption 7 imposes further smoothness on the filter ϕ for convergence of higher order derivatives of the objective functions as for ARMA models. Assumption 8.1 was used by Hong (1999) for the analysis of dependence tests and is standard in the related literature of smoothed spectral density estimation. Assumption 8.2 allows to choose p for optimal MSE estimation of the generalized spectral densities for standard kernels, but our theory does not provide a rule for the choice of p because first-order asymptotic properties of parameter estimates $\tilde{\theta}_T^{mds}$ based on $Q_T^{mds}(\theta)$ do not depend on p once Assumption 8.2 holds.

Define $R_t^{(a)}$ for $a = 0, 1$ and $z_t^0 = z_t(\theta_0; u) = e^{iu\varepsilon_t} - \varphi(u)$, $\varphi(u) = \mathbb{E}[e^{iu\varepsilon_t}]$, as

$$R_t^{(a)} := \sum_{j=1}^{\infty} j^{-2a} i \int z_{t-j}^0(v) \zeta_j^0(-v)\, dW(v), \qquad \zeta_j^0(v) := -\sum_{n=j}^{\infty} \delta_n(\theta_0) \varphi_{j-n}^{(1,0)}(0, v),$$

noting that $\varphi_{j-n}^{(1,0)}(0, v) = i\mathbb{E}[\varepsilon_{t-n} e^{iv\varepsilon_{t-j}}]$ is only different from zero under *i.i.d.* when $n = j$, in which case $\zeta_j^0(v) = -\delta_j(\theta_0)\varphi^{(1)}(v)$. Define also for $a = 0, 1$

$$S_t^{(a)} := \sum_{j=1}^{\infty} j^{-2a} i \int z_{t-j}^0(v) \beta_j^0(-v)\, dW(v), \qquad \beta_j^0(v) := -\delta_{-j}(\theta_0) v \varphi_j^{(2,0)}(0, v),$$

which would simplify if ε_t were *i.i.d.* or a conditional homoskedastic *mds*, as in this case $\varphi_j^{(2,0)}(0, v) = -\mathbb{E}[\varepsilon_t^2 e^{iv\varepsilon_{t-j}}] = -\sigma_\varepsilon^2 \varphi(v)$ and $\beta_j^0(v) = \sigma_\varepsilon^2 \delta_{-j}(\theta_0) v \varphi(v)$, $\sigma_\varepsilon^2 = \mathbb{E}[\varepsilon_t^2]$. Note that, under Assumption 6, $\varphi_{j-n}^{(1,0)}(0, v)$ tends to zero with a rate $\alpha(n - j)^{\frac{\nu-1}{\nu}}$ for $\nu \geq 5$ uniformly in v with $\sup_v \sum_{n=j}^{\infty} |\varphi_{j-n}^{(1,0)}(0, v)| < \infty$. Therefore, $\zeta_j^0(v) = O(\|\delta_j(\theta_0)\|)$ as $j \to \infty$ and it is immediate that $\beta_j^0(v) = O(|v|\|\delta_{-j}(\theta_0)\|)$ as $j \to \infty$, both sequences being absolute summable for $\eta_1 > 1$. Also, $\beta_j^0(v) = 0$ for all $j = 1, 2, \ldots$ for causal and invertible processes.

Finally, we set $V_a^{mds} := \mathbb{V}[\varepsilon_t(R_t^{(a)} + S_t^{(a)})] = \mathbb{E}[\varepsilon_t^2(R_t^{(a)} + S_t^{(a)})(\overline{R_t^{(a)} + S_t^{(a)}})]$, $a = 0, 1$, so that

$$V_a^{mds} = \sum_{j=1}^{\infty}\sum_{\ell=1}^{\infty}(j\ell)^{-2a}\int\int \mathbb{E}[\varepsilon_t^2 z_{t-j}^0(v)z_{t-\ell}^0(-u)] \times \{\zeta_j^0(-v) + \beta_j^0(-v)\}\{\zeta_\ell^0(u) + \beta_\ell^0(u)\}' dW(v)\, dW(u),$$

while the also real matrix H_a^{mds}, $a = 0, 1$, is defined as

$$H_a^{mds} := -\sum_{j=1}^{\infty} j^{-2a}\int \{\zeta_j^0(-v) + \beta_j^0(-v)\}\{\zeta_j^0(v) + \beta_j^0(v)\}' dW(v).$$

THEOREM 2. *Under Assumptions 1–8, $\eta_0 > 2$, $\eta_a > 1$, $a = 1, 2, 3$, $\theta_0 \in Int(\Theta)$, (3), and H_0^{mds} and H_1^{mds} positive definite, as $T \to \infty$,*

$$T^{1/2}(\tilde{\theta}_T^{mds} - \theta_0) \to_d N\big(0, (H_0^{mds})^{-1}V_0^{mds}(H_0^{mds})^{-1}\big),$$
$$T^{1/2}(\hat{\theta}_T^{mds} - \theta_0) \to_d N\big(0, (H_1^{mds})^{-1}V_1^{mds}(H_1^{mds})^{-1}\big).$$

The kernel k and the bandwidth p do not play a direct role in the asymptotic distribution of $\tilde{\theta}_T^{mds}$ and its root-T convergence rate can be achieved without loss of asymptotic efficiency from slowly increasing p because coefficients in the score representation converge very fast (given that $k(x)$ is smooth around $x = 0$).

When we impose the *i.i.d.* assumption instead of the mixing *mds* condition, we can find that for $a = 0, 1$,

$$V_a^{mds} = \sigma_\varepsilon^2\{\sigma_x^2\mathbf{\Sigma}_{2a} + \sigma_s^2\mathbf{\Sigma}_{2a}^- + \sigma_{sx}(\mathbf{\Sigma}_{2a}^{\mp} + \mathbf{\Sigma}_{2a}^{\mp\prime})\},$$

where σ_s^2 and σ_x^2 are the variances of the (bounded) zero mean *i.i.d.* random variables (s_t^0, x_t^0),

$$s_t^0 := \frac{\sigma_\varepsilon^2}{i}\int z_t^0(u)u\varphi(-u)\, dW(u),$$
$$x_t^0 := \frac{1}{i}\int z_t^0(u)\varphi^{(1)}(-u)\, dW(u),$$

σ_{sx} is their covariance and, $a = 0, 1, 2$,

$$\mathbf{\Sigma}_a := \sum_{j=1}^{\infty} j^{-2a}\delta_j(\theta_0)\delta_j'(\theta_0),$$

$\mathbf{\Sigma}_a^- := \sum_{j=1}^{\infty} j^{-2a}\delta_{-j}(\theta_0)\delta_{-j}'(\theta_0)$ and $\mathbf{\Sigma}_a^{\mp} := \sum_{j=1}^{\infty} j^{-2a}\delta_{-j}(\theta_0)\delta_j'(\theta_0)$.

The assumptions on H_a^{mds} are local identification conditions and, similarly, we find under *i.i.d.* that $a = 0, 1$,

$$H_a^{mds} = \rho_1\mathbf{\Sigma}_a + \sigma_\varepsilon^4\rho_2\mathbf{\Sigma}_a^- - \rho_0\sigma_\varepsilon^2(\mathbf{\Sigma}_a^{\mp} + \mathbf{\Sigma}_a^{\mp\prime}),$$

where the scalar coefficients ρ_a are defined by $\rho_0 := -\int \varphi^{(1)}(u)u\varphi(-u)\, dW(u)$, $\rho_1 := \int |\varphi^{(1)}(u)|^2\, dW(u)$ and $\rho_2 := \int u^2|\varphi(u)|^2\, dW(u)$. When the model is further causal and invertible, $\mathbf{\Sigma}_a^- = \mathbf{\Sigma}_a^{\mp} = 0$ because $\delta_j(\theta) = 0$ for $j = 0, -1, \ldots$. Then H_a^{mds} are positive definite if $\mathbf{\Sigma}_a$ are positive definite for pure causal-invertible models (or if $\mathbf{\Sigma}_a^-$ is positive definite for pure noncausal and noninvertible ones because in this case $\delta_j(\theta) = 0$ for $j = 1, 2, \ldots$). Note

that $\mathbf{\Sigma}_0 > 0$ is the local identification condition for Gaussian PML estimates of causal and invertible models, while the contribution of filters with forward components is reflected in $\mathbf{\Sigma}_a^-$ and $\mathbf{\Sigma}_a^{\mp}$. In this case, the asymptotic variance of $\tilde{\theta}_T^{mds}$ simplifies to $\kappa^{mds}\mathbf{\Sigma}_0^{-1}$, where the scalar factor $\kappa^{mds} := \sigma_\varepsilon^2\sigma_x^2/\rho_1^{-2}$ measures the asymptotic relative efficiency of the estimates with respect to the Gaussian PMLE. See similar representations in Velasco and Lobato (2018) for estimates based on different sets of spectral densities also under *i.i.d.*

In the causal and invertible case, the asymptotic variance of $\hat{\theta}_T^{mds}$ also simplifies to $\kappa^{mds}\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1}$, where $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are weighted versions of $\mathbf{\Sigma}_0$ with higher discount of higher lags information when using the spectral *cdf*. It is easy to show that $\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_0^{-1}$ is positive semidefinite, so $\tilde{\theta}_T^{mds}$ is asymptotically more efficient than $\hat{\theta}_T^{mds}$ for pure causal and invertible models with independent errors, and also for pure noncausal and noninvertible ones, for which the discussion is the same replacing $\mathbf{\Sigma}_a$ by $\mathbf{\Sigma}_a^-$.

If additionally ε_t is Gaussian with distribution $N(0, \sigma_\varepsilon^2)$ and W is the standard normal *cdf*, we can obtain closed-form expressions for these scalar coefficients (see Appendix C of Velasco (2022)), and the asymptotic relative efficiency of $\tilde{\theta}_T^{mds}$ with respect to the MLE is

$$\kappa^{mds} = \kappa^{mds}(\sigma_\varepsilon^2) := \left(\frac{(2\sigma_\varepsilon^2+1)^2}{(2\sigma_\varepsilon^2+1)^2 - \sigma_\varepsilon^4}\right)^{3/2},$$

where $\kappa^{mds}(\sigma_\varepsilon^2) \to 1$ as $\sigma_\varepsilon \to 0$, while $\kappa^{mds}(\sigma_\varepsilon^2) \to (4/3)^{3/2} = 1.54$ as $\sigma_\varepsilon \to \infty$ and $\kappa^{mds}(1) = (9/8)^{3/2} = 1.19 > 1$, so efficiency improves as σ_ε^2 becomes smaller. This significant role of σ_ε^2 in estimates' efficiency implies that in practice the residuals $\hat{\varepsilon}_t(\theta)$ could be normalized (for a fixed W) to obtain parameter estimates with asymptotic variance approaching the optimal one for normal data, or exploit dependence measures based on the *wns* condition, which can be made scale independent as do not require integration.

4.1. *Robust standard errors.* Under the *mds* assumption we need robust estimates of the asymptotic variances to nonlinear dependence, including conditional heteroskedasticity. Using a consistent $\theta_T \in \{\tilde{\theta}_T^{mds}, \hat{\theta}_T^{mds}\}$, we propose to estimate V_a^{mds}, $a = 0, 1$, with

$$\hat{V}_{a,p}^{mds} := \frac{1}{T-1}\sum_{t=2}^{T} \hat{\varepsilon}_t^2 \hat{Z}_{t,p}^{(a)} \hat{Z}_{t,p}^{(a)\prime},$$

where $\hat{\varepsilon}_t := \hat{\varepsilon}_t(\theta_T)$ and $\hat{Z}_{t,p}^{(a)} := i\sum_{j=1}^{t-1} j^{-2a} k(\frac{j}{p}) \int \hat{z}_{t-j}(v)\{\hat{\zeta}_{j,p}(-v) + \hat{\beta}_j(-v)\}\, dW(v)$ for

$$\hat{\zeta}_{j,p}(v) := -\sum_{n=j}^{T+j-1} k\left(\frac{n-j}{p}\right)\delta_n(\theta_T)\hat{\varphi}_{\theta_T,|j|-n}^{(1,0)}(0,v),$$

and $\hat{\beta}_j(v) := -\delta_{-j}(\theta_T) v \hat{\varphi}_{\theta_T,j}^{(2,0)}(0,v)$, with k and p as in Assumption 8. Similarly, we can estimate H_a^{mds} with

$$\hat{H}_{a,p}^{mds} = -\sum_{j=1}^{T-1} j^{-2a} k\left(\frac{j}{p}\right) \int \{\hat{\zeta}_{j,p}(-v) + \hat{\beta}_j(-v)\}\{\hat{\zeta}_{j,p}(v) + \hat{\beta}_j(v)\}' dW(v),$$

where estimates are real and have closed-form expressions for W with known *cf* like the Gaussian (see Appendix D of Velasco (2022)), and easily showed to be consistent using the methods in the proof of Theorem 2.

4.2. *Standard errors under i.i.d.* Under the *i.i.d.* assumption on ε_t, standard errors of estimates can be obtained by direct estimation of $(\sigma_s^2, \sigma_x^2, \sigma_{sx})$ with the sample covariance matrix of

$$\hat{s}_t := s_t(\theta_T) = \frac{\hat{\sigma}_\varepsilon^2}{i} \int \hat{z}_t(u) u \hat{\varphi}_{\theta_T}(-u)\, dW(u),$$

$$\hat{x}_t := x_t(\theta_T) = \frac{1}{i} \int \hat{z}_t(u) \hat{\varphi}^{(1)}_{\theta_T}(-u)\, dW(u),$$

noting that the integrals are pure imaginary because W is symmetric, where $\hat{z}_t(u) := z_t(\theta_T; u) = e^{iu\hat{\varepsilon}_t} - \hat{\varphi}_{\theta_T}(u)$ and $\hat{\sigma}_\varepsilon^2$ is the sample variance of $\hat{\varepsilon}_t$. In the case of Gaussian or with closed-form characteristic function W, $\hat{s}_t$ and $\hat{x}_t$ have explicit expressions amenable to fast computation without numerical integration. Alternatively, we can approximate numerically the integrals if no explicit expressions exist for a given W. For the estimation of (ρ_0, ρ_1, ρ_2), we could also perform numerical integration using $\hat{\varphi}_{\theta_T}(u)$ and $\hat{\varphi}^{(1)}_{\theta_T}(u)$ in place of their population counterparts, or alternatively, use closed-form expressions; see Appendix D of Velasco (2022) for Gaussian W. For estimation of $\mathbf{\Sigma}_a$, $\mathbf{\Sigma}_a^-$, $\mathbf{\Sigma}_a^{\mp}$, we just plug-in θ_T in the expression for $\delta_j(\theta)$ for a given model.

5. Overidentified GMM estimation. In this section, we explore efficient estimation when more than one dependence condition is imposed on the residuals. Thus, when imposing the *i.i.d.* assumption on estimation it is also possible to use information from the weaker identification *mds* condition and in this case and when estimating under *mds* it is possible to use the even weaker white noise sequence (*wns*) or serial uncorrelation condition that motivates the loss functions

$$Q_T^{wns}(\theta) := \frac{2}{\pi} \sum_{j=1}^{T-1} k^2\left(\frac{j}{p}\right)\left(1 - \frac{|j|}{T}\right)\hat{\sigma}^{(1,1)}_{\theta,j}(0,0)^2$$

and $\mathcal{L}_T^{wns}(\theta) := \frac{2}{\pi} \sum_{j=1}^{T-1} j^{-2} \hat{\sigma}^{(1,1)}_{\theta,j}(0,0)^2$, where

$$\hat{\sigma}^{(1,1)}_{\theta,j}(0,0) = -\frac{1}{T-|j|} \sum_{t=1+|j|}^{T} \hat{\varepsilon}_t(\theta)\hat{\varepsilon}_{t-|j|}(\theta) + \frac{1}{(T-|j|)^2} \sum_{t=1+|j|}^{T} \hat{\varepsilon}_t(\theta) \sum_{t=1+|j|}^{T} \hat{\varepsilon}_{t-|j|}(\theta)$$

are (minus) the usual sample autocovariances of residuals $\hat{\varepsilon}_t(\theta)$.

First, in parallel with *mds* analysis, we can propose estimates based on the *i.i.d.* condition,

$$\hat{\theta}_T^{i.i.d.} := \arg\min_{\theta\in\Theta} \mathcal{L}_T^{i.i.d.}(\theta)$$

and

$$\tilde{\theta}_T^{i.i.d.} := \hat{\theta}_T^{i.i.d.} - \left(\frac{\partial}{\partial\theta\,\partial\theta'} Q_T^{i.i.d.}(\hat{\theta}_T^{i.i.d.})\right)^{-1} \frac{\partial}{\partial\theta} Q_T^{i.i.d.}(\hat{\theta}_T^{i.i.d.}).$$

With serially independent ε_t, consistency of $\hat{\theta}_T^{i.i.d.}$ follows by the identification provided by Lemma 3 under slightly different conditions from those for $\hat{\theta}_T^{mds}$, not requiring mixing or bounded higher order moments on ε_t if the model does not admit $\theta \neq \theta_0$ for which $|\phi(\theta; z)|$ is constant, as identification can rely on the serial correlation of residuals when using the wrong parameter values. However, the summability conditions on the linear filter are stronger as the proof does not exploit higher order moments.

THEOREM 3. *Under Assumptions 1–4, $\eta_0 \geq 3$, $\eta_1 > 1$, ε_t i.i.d., zero mean, $\mathbb{E}|\varepsilon_t| < \infty$, $\theta_0 \in \Theta$, $\hat{\theta}_T^{i.i.d.} \to_p \theta_0$ as $T \to \infty$.*

Define for $a = 0, 1, 2$,

$$V_a := \sigma_e^2\sigma_x^2(\mathbf{\Sigma}_{2a} + \mathbf{\Sigma}_{2a}^-) + \sigma_{xe}^2(\mathbf{\Sigma}_{2a}^{\mp} + \mathbf{\Sigma}_{2a}^{\mp\prime}),$$
$$H_a := \rho_1\rho_2(\mathbf{\Sigma}_a + \mathbf{\Sigma}_a^-) + \rho_0^2(\mathbf{\Sigma}_a^{\mp} + \mathbf{\Sigma}_a^{\mp\prime}),$$

where (σ_x^2, σ_e^2) and σ_{xe} are the variances and covariance of the (bounded and real) zero mean *i.i.d.* random variables (x_t^0, e_t^0) where

$$e_t^0 = e_t(\theta_0) := \frac{1}{i}\int z_t(\theta_0; u)u\varphi(-u)\,dW(u).$$

Then we obtain the asymptotic distribution of *i.i.d.* based estimates also under weaker moment conditions than for *mds* based estimates if $|\phi(\theta; z)|$ is constant only for $\theta = \theta_0$ or if otherwise $\kappa_3^{\varepsilon} \neq 0$ and no higher than third-order cumulants are required for identification.

THEOREM 4. *Under Assumptions* 1–4, 7, 8, $\eta_0 \geq 3$, $\eta_a > 1$, $a = 1, 2, 3$, ε_t i.i.d., *zero mean*, $\mathbb{E}|\varepsilon_t|^3 < \infty$, H_0 *and* H_1 *positive definite and* $\theta_0 \in Int(\Theta)$, *as* $T \to \infty$,

$$T^{1/2}(\tilde{\theta}_T^{i.i.d.} - \theta_0) \to_d N(0, H_0^{-1}V_0H_0^{-1}),$$
$$T^{1/2}(\hat{\theta}_T^{i.i.d.} - \theta_0) \to_d N(0, H_1^{-1}V_1H_1^{-1}).$$

Standard errors for *i.i.d.* estimates can be obtained in the same fashion as for *mds* estimates under serial independence using $\hat{x}_t$ and $\hat{e}_t := e_t(\theta_T) = \frac{1}{i}\int \hat{z}_t(u)u\hat{\varphi}_{\theta_T}(-u)\,dW(u)$. For pure causal-invertible models, the assumptions on H_a are satisfied if $\mathbf{\Sigma}_a$ are positive definite and the asymptotic variance of $\tilde{\theta}_T^{i.i.d.}$ simplifies to $\kappa^{i.i.d.}\mathbf{\Sigma}_0^{-1}$, where the scalar factor $\kappa^{i.i.d.} := \sigma_e^2\sigma_x^2/(\rho_1^2\rho_2^2)^{-1}$ measures the asymptotic relative efficiency of the estimates compared to the Gaussian PMLE, with $\tilde{\theta}_T^{i.i.d.}$ being more efficient than $\hat{\theta}_T^{i.i.d.}$, whose asymptotic variance becomes $\kappa^{i.i.d.}\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1}$.

When ε_t is Gaussian with distribution $N(0, \sigma_{\varepsilon}^2)$ and $W(u)$ is the standard normal *cdf*, we obtain in Appendix C of Velasco (2022) that $\kappa^{i.i.d.} = \kappa^{i.i.d.}(\sigma_{\varepsilon}^2) = \kappa^{mds}(\sigma_{\varepsilon}^2)^2$. Therefore, *mds* based estimation is more efficient than *i.i.d.* based one for Gaussian processes, but this asymptotic relative efficiency might differ for other distributions, for example, for scaled chi-square distributions with degrees of freedom under 4 and large σ_{ε}; see Figures 1 and 2 in Appendix C.

It is easy to check that estimates based on the *wns* criterion have the same asymptotic distribution as the Gaussian PMLE under global identification (e.g., assuming invertibility and causality) and a moment condition under serial independence. Further, we have argued in Section 3 that, as the scaling $\sigma \to 0$, the *i.i.d.* and *mds* criteria are equivalent to the *wns* restriction, so using *i.i.d.* and *mds* estimation with a fixed σ together with the *wns* restriction, can provide robust estimates to the scaling choice, which are never less efficient than the Gaussian PMLE, while achieving identification under non-Gaussianity.

Following Velasco and Lobato (2018), who considered parameter estimation using simultaneously information from different moments, our GMM overidentified estimation considers simultaneously the scores of the three objective functions $Q_T^{i.i.d.}$, Q_T^{mds} and Q_T^{wns},

$$\mathbf{S}_T(\theta) = \begin{pmatrix} (\partial/\partial\theta)Q_T^{i.i.d.}(\theta) \\ (\partial/\partial\theta)Q_T^{mds}(\theta) \\ (\partial/\partial\theta)Q_T^{wns}(\theta) \end{pmatrix},$$

and optimally weight the joint information provided by a Newton–Raphson iteration over the objective function $Q_T(\theta) = \mathbf{S}_T(\theta)'\hat{\boldsymbol{V}}_T^{-1}\mathbf{S}_T(\theta)$, that is,

$$\tilde{\theta}_T^{gmm} := \theta_T - \left(\mathbf{H}_T(\theta_T)'\hat{\boldsymbol{V}}_T^{-1}\mathbf{H}_T(\theta_T)\right)^{-1}\mathbf{H}_T(\theta_T)'\hat{\boldsymbol{V}}_T^{-1}\mathbf{S}_T(\theta_T), \tag{4}$$

where $\hat{\boldsymbol{V}}_T$ is a consistent estimator of $\mathbf{V}$, the asymptotic variance of $T^{1/2}\mathbf{S}_T(\theta_0)$, $\mathbf{H}_T(\theta) = (\partial/\partial\theta')\mathbf{S}_T(\theta)$ and the initial estimate θ_T satisfies

$$\theta_T - \theta_0 = O_p(T^{-1/2}). \tag{5}$$

Further, we can replace $\mathbf{H}_T(\theta_T)$ by the asymptotically equivalent $\widehat{\mathbf{H}}(\theta_T)$ where $\widehat{\mathbf{H}}$ uses the functional form of $\mathbf{H} = \mathbf{H}(\theta_0) := p\lim_{T\to\infty}\mathbf{H}_T(\theta_0)$, but replacing unknown moments by sample averages, errors by residuals and (derivatives of) the *cf* by empirical estimates. For implementing (4), we can employ for θ_T some version of *i.i.d.* or *mds* estimation, which correctly identifies θ_0 in the case that noncausal or noninvertible models are allowed, while a similar estimation strategy based on $\mathcal{L}_T$ distances can be proposed. Given (5), the consistency and asymptotic normality of $\tilde{\theta}_T^{gmm}$ is immediate as described by the next theorem.

THEOREM 5. *Under the assumptions of Theorem 2,* ε_t i.i.d., (5), $\hat{\boldsymbol{V}} \to_p \mathbf{V} > 0$ *as* $T \to \infty$,

$$\sqrt{T}(\tilde{\theta}_T^{gmm} - \theta_0) \to_d N\big(0, (\mathbf{H}'\mathbf{V}^{-1}\mathbf{H})^{-1}\big).$$

For causal and invertible models, $\mathbf{H} = (\rho_1\rho_2, \rho_1, \sigma_\varepsilon^4)' \otimes \boldsymbol{\Sigma}_0$ and $\mathbf{V} = \{\mathbb{V}(w_t) \circ \mathbb{V}(W_t)\} \otimes \boldsymbol{\Sigma}_0$, where $w_t := (e_t^0, \varepsilon_t, \varepsilon_t)'$, $W_t := (x_t^0, x_t^0, \sigma_\varepsilon^2\varepsilon_t)'$ and $\circ$ is the Hadamart element by element product. Then

$$\begin{aligned}\mathbf{V} &= \begin{pmatrix} \sigma_e^2 & -\rho_0 & -\rho_0 \\ -\rho_0 & \sigma_\varepsilon^2 & \sigma_\varepsilon^2 \\ -\rho_0 & \sigma_\varepsilon^2 & \sigma_\varepsilon^2 \end{pmatrix} \circ \begin{pmatrix} \sigma_x^2 & \sigma_x^2 & \sigma_\varepsilon^2\rho_1 \\ \sigma_x^2 & \sigma_x^2 & \sigma_\varepsilon^2\rho_1 \\ \sigma_\varepsilon^2\rho_1 & \sigma_\varepsilon^2\rho_1 & \sigma_\varepsilon^6 \end{pmatrix} \otimes \boldsymbol{\Sigma}_0 \\ &= \begin{pmatrix} \sigma_e^2\sigma_x^2 & -\rho_0\sigma_x^2 & -\sigma_\varepsilon^2\rho_0\rho_1 \\ -\rho_0\sigma_x^2 & \sigma_\varepsilon^2\sigma_x^2 & \sigma_\varepsilon^4\rho_1 \\ -\sigma_\varepsilon^2\rho_0\rho_1 & \sigma_\varepsilon^4\rho_1 & \sigma_\varepsilon^8 \end{pmatrix} \otimes \boldsymbol{\Sigma}_0,\end{aligned}$$

using that $\sigma_{e\varepsilon} = \mathbb{E}[e_t^0\varepsilon_t] = -\rho_0$ and $\sigma_{x\varepsilon} = \mathbb{E}[x_t^0\varepsilon_t] = \rho_1$, so that $(\mathbf{H}'\mathbf{V}^{-1}\mathbf{H})^{-1} = \kappa\boldsymbol{\Sigma}_0^{-1}$ for $\kappa \le 1$, and $\tilde{\theta}_T^{gmm}$ is never less efficient than the Gaussian PMLE.

Similarly, to make more robust the identification and estimation, we could drop the *i.i.d.* assumption and rely only on identification from the *mds* condition by considering estimates based on *mds* and *wns* criteria that do not impose in $\mathbf{V}$ lack of dynamics in higher order moments of ε_t.

6. Efficient continuous GMM. In this section, we propose parameter estimates based on the efficient weighting across (u, v) of the continuous moment conditions that set $\hat{\sigma}_{\theta,j}(u, v)$ and $\hat{\sigma}_{\theta,j}^{(1,0)}(0, v)$ equal to zero for all $(u, v) \in \mathbb{R}^2$ and $j = 1, 2, \ldots$ Since these restrictions use implicitly instruments based on the *cf* (of lagged residuals), the optimal score functions are always included in the span of our specific continuous moment conditions, allowing for optimal estimation in contrast to methods based on a finite set of moment conditions as when using ordinary spectral densities for non-Gaussian data.

We follow the approach of Carrasco et al. (2007) to optimally weight the empirical continuous moments for each (u, v) using the appropriate covariance kernel and obtain the same efficiency as the ML estimate using the likelihood function, which contains the same information as the *cf*. However, by using an increasing number of pairwise dependence conditions based on model residuals, we do not need to specify a conditional model for the *cf* of Y_t given a finite vector of past observations, or simulate their joint *cf* as in Carrasco et al. (2007). Since residuals evaluated at the true parameter value become (mean) independent of all the past under the *i.i.d.* (*mds*) assumption, their conditional *cf* (and its first derivative at zero) is equal

to their marginal *cf* (expectation), which can be estimated using the residuals empirical *cf* (mean) based on a preliminary estimate of θ_0.

The continuous moments averaging the residual-based $\xi_{t,j}(w;\theta) = z_t(\theta;u)z_{t-|j|}(\theta;v)$, $w=(u,v)\in\mathbb{R}^2$, or $\xi_{t,j}(w;\theta)=\varepsilon_t(\theta)z_{t-|j|}(\theta;v)$, $w=v\in\mathbb{R}$, implicit in our *i.i.d.* and *mds* estimation, have covariance function

$$C_{j,n}(w_1,w_2;\theta) = \operatorname{Cov}\big[\xi_{t,j}(w_1;\theta), \overline{\xi_{t,n}(w_2;\theta)}\big], \quad j,n=1,2,\ldots.$$

For the former and under serial independence, $C_{j,n}(w_1,w_2;\theta_0)=0$ for $j\neq n$, while for $j=n$, $C_{j,j}$ is equal to

$$\begin{aligned} C^{i.i.d.}(w_1,w_2;\theta_0) &:= \mathbb{E}\big[z_t(\theta_0;u_1)z_{t-|j|}(\theta_0;v_1)\overline{z_t(\theta_0;u_2)z_{t-|j|}(\theta_0;v_2)}\big] \\ &= \big\{\varphi(u_1-u_2)-\varphi(u_1)\varphi(-u_2)\big\}\big\{\varphi(v_1-v_2)-\varphi(v_1)\varphi(-v_2)\big\}, \end{aligned}$$

which does not depend on j because both $z_t(\theta_0;u)$ and $\varepsilon_t(\theta_0)$ become *i.i.d.*, but for the latter,

$$C^{mds}_{j,n}(v_1,v_2;\theta_0) := \mathbb{E}\big[\varepsilon_t^2\big\{e^{iv_1\varepsilon_{t-|j|}}-\varphi(v_1)\big\}\big\{e^{-iv_2\varepsilon_{t-|n|}}-\varphi(-v_2)\big\}\big]$$

might depend on j and n when imposing only a *mds* condition. However, $C^{mds}_{j,n}(v_1,v_2;\theta_0)=0$ for $j\neq n$ under *i.i.d.*, while for $j=n$, $C^{mds}_{j,j}$ simplifies to

$$\begin{aligned} C^{mds}(v_1,v_2;\theta_0) &:= \sigma_\varepsilon^2\mathbb{E}\big[\big\{e^{iv_1\varepsilon_{t-|j|}}-\varphi(v_1)\big\}\big\{e^{-iv_2\varepsilon_{t-|j|}}-\varphi(-v_2)\big\}\big] \\ &= \sigma_\varepsilon^2\big(\varphi(v_1-v_2)-\varphi(v_1)\varphi(-v_2)\big). \end{aligned}$$

Therefore, both efficient estimation strategies can be pursued as in Carrasco et al. (2007) under *i.i.d.* despite we explicitly consider an increasing number of moment conditions indexed by j. This is feasible because the moment conditions become asymptotically independent at different lags for all values of (w_1,w_2) when evaluated at the true value of the parameter and, by stationarity and independence, all have the same covariance function, which then only needs to be estimated once. However, under only a *mds* condition, estimation of $C^{mds}_{j,n}$ for all pairs (j,n) would be required as moment conditions are in general nonhomoskedastic nor uncorrelated at different lags.

To construct continuous GMM (CGMM) efficient estimates, we need estimates of $(K^{i.i.d.})^{-1/2}$ and $(K^{mds})^{-1/2}$ to standardize each of the empirical moment conditions, where $K^{i.i.d.}$ and K^{mds} are the operators implied by the corresponding covariance functions of the sample moments $\hat\sigma_{\theta_0,j}(w)$ and $\hat\sigma^{(1,0)}_{\theta_0,j}(w)$, respectively, defined by

$$\big(K^\bullet g\big)(w_1;\theta_0) := \int C^\bullet(w_1,w_2;\theta_0)g(w_2)\,dW(w_2).$$

For estimation of the inverse of the operators $K^{i.i.d.}$ and K^{mds}, we pursue the same strategy as in Carrasco et al. (2007) using a Tikhonov regularized inverse of an estimate $K^\bullet_T$ of $K^\bullet$,

$$\big(K^\bullet_{T,\alpha_T}\big)^{-1} := \big(K^{\bullet 2}_T+\alpha_T I\big)^{-1}K^\bullet_T,$$

for the identity operator I and a penalizing term $\alpha_T\to 0$ to avoid the problem of $(K^\bullet_T)^{-1}$ not existing on the whole Hilbert space $\mathbb{L}^2(W)$ defined by the weighting function W, but only on a subset (the reproducing kernel Hilbert space of $K^\bullet$).

Following our pairwise approach, we define

$$\bar Q^{i.i.d.}_T(\theta) := \frac{2}{\pi}\sum_{j=1}^{T-1}k^2\Big(\frac{j}{p}\Big)\Big(1-\frac{|j|}{T}\Big)\int\big|\big(K^{i.i.d.}_{T,\alpha_T}\big)^{-1/2}\hat\sigma_{\theta,j}(u,v)\big|^2\,dW(u,v)$$

for $K_{T,\alpha_T}^{i.i.d.}$ based on a consistent estimator $K_T^{i.i.d.}$ satisfying

$$\|K_T^{i.i.d.} - K^{i.i.d.}\|^2 = \sup_{\|g\|\le 1} \int |(K_T^{i.i.d.} - K^{i.i.d.})g(w)|^2 \, dW(w) = O_p(T^{-\varrho}) \tag{6}$$

for some $\varrho \ge 0$, and the corresponding estimates

$$\bar{\theta}_T^{i.i.d.} = \theta_T - \left(\frac{\partial}{\partial\theta\,\partial\theta'}\bar{Q}_T^{i.i.d.}(\theta_T)\right)^{-1} \frac{\partial}{\partial\theta}\bar{Q}_T^{i.i.d.}(\theta_T) \tag{7}$$

iterating over the preliminary root-T consistent estimate θ_T. A similar approach leads to the definition of $\bar{\theta}_T^{mds}$ based on $\bar{Q}_T^{mds}(\theta)$ built in terms of $(K_{T,\alpha_T}^{mds})^{-1/2}\hat{\sigma}_{\theta,j}^{(1,0)}(0,v)$ for a consistent K_T^{mds}.

We set

$$\hat{C}_T^{\bullet}(w_1, w_2; \theta_T) := \frac{1}{T-1}\sum_{t=2}^{T} \hat{\xi}_{t,1}(w_1;\theta_T)\overline{\hat{\xi}_{t,1}(w_2;\theta_T)},$$

where for $\hat{C}_T^{i.i.d.}$ and $w = (u, v)$,

$$\hat{\xi}_{t,1}(w;\theta) := \big(e^{iu\hat{\varepsilon}_t(\theta)} - \hat{\varphi}_{\theta,1}(u)\big)\big(e^{iv\hat{\varepsilon}_{t-1}(\theta)} - \hat{\varphi}_{\theta,1}(v)\big), \tag{8}$$

and for $\hat{C}_T^{mds}$ and $w = v$,

$$\hat{\xi}_{t,1}(v;\theta) := \big(\hat{\varepsilon}_t(\theta) - \overline{\hat{\varepsilon}}_T(\theta)\big)\big(e^{iv\hat{\varepsilon}_{t-1}(\theta)} - \hat{\varphi}_{\theta,1}(v)\big), \tag{9}$$

using the first lag to exploit the maximum possible number of observations in $\hat{\varphi}_{\theta,1}(v)$ because the moments evaluated at $\theta = \theta_0$ have the same variance for all $j = 1, 2, \ldots$. This leads to consistent estimates of $K_T^{\bullet}$ with $\varrho = 1/2$.

For computation purposes, we write as in Carrasco et al. (2007) for each $j = 1, 2, \ldots$

$$\int |(K_{T,\alpha_T}^{i.i.d.})^{-1/2}\hat{\sigma}_{\theta,j}(u,v)|^2 \, dW(u,v) = \overline{\mathbf{v}_{T,j}(\theta)}'\big(C_T^2 + \alpha_T I_{T-1}\big)^{-1}\mathbf{v}_{T,j}(\theta),$$

where I_{T-1} is the $(T-1)$-identity matrix, $\mathbf{v}_{T,j}(\theta) = (v_{j,2}(\theta), \ldots, v_{j,T}(\theta))'$ with $v_{j,t}(\theta) := \int \overline{\hat{\xi}_{t,1}(u,v;\theta_T)}\hat{\sigma}_{\theta,j}(u,v)\,dW(u,v)$ and C_T is the $(T-1)\times(T-1)$ matrix with (t,r) element $\hat{c}_{tr}/(T-1)$, where $\hat{c}_{tr} := \int \overline{\hat{\xi}_{t,1}(u,v;\theta_T)}\xi_{r,1}(u,v;\theta_T)\,dW(u,v)$ for $\hat{\xi}_{t,1}$ in (8) and a preliminary $\theta_T \to_p \theta_0$. A similar expression holds for $\int |(K_{T,\alpha_T}^{mds})^{-1/2}\hat{\sigma}_{\theta,j}^{(1,0)}(0,v)|^2\,dW(v)$ for $\hat{\xi}_{t,1}$ in (9), where all integrals have closed form for Gaussian W; see Appendix D in Velasco (2022).

Assuming that ε_t has finite variance and *pdf* $f(x)$ with derivative $\dot{f}(x) := (\partial/\partial x)f(x)$ satisfying $\bar{\sigma}_0^2 := \mathbb{E}(\dot{f}(\varepsilon_t)/f(\varepsilon_t))^2 < \infty$, so that $|\bar{\rho}_0| < \infty$, $\bar{\rho}_0 := \mathbb{E}[\varepsilon_t\dot{f}(\varepsilon_t)/f(\varepsilon_t)]$, we define

$$\bar{H}_0 := \sigma_\varepsilon^2\bar{\sigma}_0^2\big(\mathbf{\Sigma}_0 + \mathbf{\Sigma}_0^{-}\big) + \bar{\rho}_0^2\big(\mathbf{\Sigma}_0^{\mp} + \mathbf{\Sigma}_0^{\mp\prime}\big).$$

Note that $\bar{\sigma}_0^2$ and $\bar{\rho}_0$ are the variance of the (location) score of ε_t and its covariance with ε_t itself, respectively, and

$$\frac{\partial}{\partial\theta}\log f\big(\varepsilon_t(\theta)\big) = \dot{\varepsilon}_t(\theta)\frac{\dot{f}(\varepsilon_t(\theta))}{f(\varepsilon_t(\theta))},$$

where $\dot{\varepsilon}_t(\theta) := (\partial/\partial\theta)\varepsilon_t(\theta)$ is linear in leads and lags of ε_t and $\dot{f}(\varepsilon_t(\theta_0)) = \dot{f}(\varepsilon_t)$. Therefore, $\bar{H}_0$ is the limit of the variance of the average score $T^{-1/2}\sum_{t=1}^{T}(\partial/\partial\theta)\log f(\varepsilon_t(\theta_0))$ and $\bar{H}_0^{-1}$ is the asymptotic variance of the ML estimate of θ for known f under usual regularity conditions. We now show that it is also the asymptotic variance of the optimal CGMM estimate (7), which is not using knowledge of f, under the following assumption.

ASSUMPTION 9. The null space of the kernel $K^{\bullet}(\cdot;\theta_0)$ of the covariance function of the sample moment $\xi_{t,1}(\cdot;\theta_0)$ consists of only the null element, and $\mathbb{E}[\xi_{t,1}(\cdot;\theta)]$ and $\mathbb{E}[(\partial/\partial\theta)\xi_{t,1}(\cdot;\theta)]$ belong to its reproducing kernel Hilbert space for all $\theta \in \Theta$.

This condition is equivalent to Assumption A.5 of Carrasco et al. (2007) and we discuss sufficient conditions for it in the proof of next theorem under *i.i.d.*

THEOREM 6. *Under Assumptions 1–4, 7, 8 and 9 for* $K^{i.i.d.}$, $\eta_0 \geq 3$, $\eta_a > 1$, $a = 1, 2, 3$, ε_t i.i.d., *zero mean*, $\mathbb{E}|\varepsilon_t|^3 < \infty$, $\theta_0 \in Int(\Theta)$, *with* $(T^{\varrho}\alpha_T)^{-1} + \alpha_T \to 0$ *as* $T \to \infty$,

$$T^{1/2}(\bar{\theta}_T^{i.i.d.} - \theta_0) \to_d N(0, \bar{H}_0^{-1}).$$

This result extends the MLE efficiency results of Carrasco et al. (2007) of optimal CGMM estimates to the non-Markov (or noncausal) case described by our class of models without assumptions on the conditional *cf* thanks to the use of an increasing number of lags in our objective function, providing an alternative method to the adaptive estimation of Gassiat (1993). Similarly, we now provide the asymptotic distribution of CGMM estimates based on the optimal continuous *mds* criterion under serial independence. Define

$$\bar{H}_0^{mds} := \mathbf{\Sigma}_0 + \sigma_\varepsilon^2 \bar{\sigma}_0^2 \mathbf{\Sigma}_0^{-} - \bar{\rho}_0(\mathbf{\Sigma}_0^{\mp} + \mathbf{\Sigma}_0^{\mp\prime}).$$

THEOREM 7. *Under Assumptions 1–8 and 9 for* K^{mds}, $\eta_0 > 2$, $\eta_a > 1$, $a = 1, 2, 3$, ε_t i.i.d. *zero mean*, $\theta_0 \in Int(\Theta)$, *with* $(T^{\varrho}\alpha_T)^{-1} + \alpha_T \to 0$ *as* $T \to \infty$,

$$T^{1/2}(\bar{\theta}_T^{mds} - \theta_0) \to_d N(0, (\bar{H}_0^{mds})^{-1}).$$

For Gaussian $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$, $\bar{\sigma}_0^2 = \sigma_\varepsilon^{-2}$ and $\bar{\rho}_0 = -1$, so that $\bar{H}_0^{mds} = \bar{H}_0$ and the *mds* based estimation also achieves the efficiency bound as with the optimal continuous *i.i.d.* estimation, but this property does no need to hold for other distributions or in absence of the independence assumption.

The asymptotic variance of CGMM estimates based on continuous optimal weighting could be estimated directly using smoothed estimates of the *pdf* of ε_t with residuals to construct estimates of $(\sigma_\varepsilon^2, \bar{\sigma}_0^2, \bar{\rho}_0)$. Alternatively, we can adapt the approach of Proposition 3.5 in Carrasco et al. (2007) for the asymptotic variance of the generic efficient estimate. For that, we set for some consistent θ_T, for example, $\tilde{\theta}_T^{i.i.d.}$, C_T defined as before and $\hat{\varepsilon}_t = \hat{\varepsilon}_t(\theta_T)$,

$$\begin{aligned}\hat{\bar{\sigma}}_0^2 &:= \frac{1}{\hat{\sigma}_\varepsilon^2} \int \nabla\hat{\sigma}_{1,\theta_T}(u,v)\overline{(K_{T,\alpha_T}^{i.i.d.})^{-1}\nabla\hat{\sigma}_{1,\theta_T}(u,v)}\, dW(u,v) \\ &= \frac{1}{\hat{\sigma}_\varepsilon^2}\overline{\hat{\mathbf{v}}_{T,1}}'(\alpha_T I_{T-1} + C_T^2)^{-1}\hat{\mathbf{v}}_{T,1}\end{aligned}$$

and

$$\begin{aligned}\hat{\bar{\rho}}_0^2 &:= \int \nabla\hat{\sigma}_{1,\theta_T}(v,u)\overline{(K_{T,\alpha_T}^{i.i.d.})^{-1}\nabla\hat{\sigma}_{1,\theta_T}(u,v)}\, dW(u,v) \\ &= \overline{\hat{\mathbf{w}}_{T,1}}'(\alpha_T I_{T-1} + C_T^2)^{-1}\hat{\mathbf{v}}_{T,1}\end{aligned}$$

for $\nabla\hat{\sigma}_{1,\theta_T}(u,v) = (T-1)^{-1}\sum_{t=2}^{T} iu\hat{\varepsilon}_{t-1}e^{iu\hat{\varepsilon}_t + iv\hat{\varepsilon}_{t-1}}$, $\hat{\mathbf{v}}_{T,1} = (\hat{v}_{1,2}, \ldots, \hat{v}_{1,T})'$ and $\hat{\mathbf{w}}_{T,1} = (\hat{w}_{1,2}, \ldots, \hat{w}_{1,T})'$, where the integrals $\hat{v}_{1,t} := \int \overline{\hat{\xi}_{t,1}(u,v;\theta_T)}\nabla\hat{\sigma}_{1,\theta_T}(u,v)\, dW(u,v)$ and $\hat{w}_{1,t} := \int \overline{\hat{\xi}_{t,1}(u,v;\theta_T)}\nabla\hat{\sigma}_{1,\theta_T}(v,u)\, dW(u,v)$ do not require numerical integration for Gaussian W. For the asymptotic variance of *mds* estimates, we can use the same strategy.

7. Simulations. In this section, we explore the finite sample properties of our estimates under different model specifications and innovation distributions. As discussed, the scaling of the data can make a large difference in practice for a given W so in our simulation exercises we standardize the residuals before computing dependence measures and loss functions, replacing $\hat{\varepsilon}_t(\theta)$ by

$$\hat{\varepsilon}_t(\theta;\sigma) := \sigma \frac{\hat{\varepsilon}_t(\theta)}{s_T(\theta)}, \qquad s_T^2(\theta) := \frac{1}{T}\sum_{t=1}^{T}\big(\hat{\varepsilon}_t(\theta) - \overline{\hat{\varepsilon}}_T(\theta)\big)^2,$$

where σ is a user chosen scaling parameter that balances efficiency and identification robustness; see Section 3.3. Normalization of $\hat{\varepsilon}_t(\theta)$ by $s_T(\theta)$ eliminates scaling effects in finite samples, making parameter comparison invariant to the variance of the residuals induced by the different causality and invertibility properties generated by each θ. For instance, polynomial roots close to zero induce very small (large) residuals variance in the noninvertible (noncausal) case, which can distort unscaled dependence comparisons.

This normalization does not affect asymptotic theory apart from shifting the innovations variance σ_ε^2 to the user chosen σ. This property follows because $s_T^2(\theta)$ converges uniformly to $s^2(\theta) := \sigma_\varepsilon^2 \sum_{j=-\infty}^{\infty} \phi_j^2(\theta)$ under the conditions of Theorems 2 or 4. Then $s^2(\theta)$ is bounded and bounded away from zero uniformly for $\theta \in \Theta$, with $s^2(\theta_0) = \sigma_\varepsilon^2$ and derivative $\dot{s}^2(\theta) := (\partial/\partial\theta)s^2(\theta) = 2\sigma_\varepsilon^2 \sum_{j=-\infty}^{\infty} \phi_j(\theta)\delta_j(\theta)$ satisfying $\dot{s}^2(\theta_0) = 2\sigma_\varepsilon^2\delta_0(\theta_0)$. Then $\dot{s}^2(\theta_0)$ is zero for causal and invertible ARMA processes because $\phi_0(\theta) = 1$ and $\delta_0(\theta) = 0$ for all θ, so the standardization has no asymptotic effect in this case. In the general case when $\delta_0(\theta_0) \neq 0$, we can check that the residual derivatives contain an additional term due to normalization,

$$\frac{\partial}{\partial\theta}\hat{\varepsilon}_t(\theta;\sigma) = \sigma\frac{\dot{\hat{\varepsilon}}_t(\theta)}{s_T(\theta)} - \sigma\frac{\dot{s}_T(\theta)}{s_T^2(\theta)}\hat{\varepsilon}_t(\theta),$$

where $\dot{\hat{\varepsilon}}_t(\theta) := (\partial/\partial\theta)\hat{\varepsilon}_t(\theta) = \sum_{j=-\infty}^{\infty} \delta_j(\theta_0)\varepsilon_{t-j}$ and $\dot{s}_T(\theta_0) := (\partial/\partial\theta)s_T(\theta_0) \to_p \sigma_\varepsilon\delta_0(\theta_0)$. Therefore, for $\dot{\varepsilon}_t(\theta) := (\partial/\partial\theta)\varepsilon_t(\theta)$,

$$\frac{\partial}{\partial\theta}\hat{\varepsilon}_t(\theta_0;\sigma) \to_p \frac{\sigma}{\sigma_\varepsilon}\{\dot{\varepsilon}_t(\theta_0) - \delta_0(\theta_0)\varepsilon_t\} = \frac{\sigma}{\sigma_\varepsilon}\sum_{j\neq 0}\delta_j(\theta_0)\varepsilon_{t-j},$$

leaving asymptotic properties of parameter estimates unaltered because the residual scores still do not depend on $\delta_0(\theta_0)$ as without normalization; see Theorems 2 or 4.

For W, we use the standard normal *cdf*; see Appendix D, Velasco (2022). For $Q_T^\bullet$ criteria, we employ the Daniell kernel, $k(x) = \sin(x)/x$, while we set $p = T^{1/5}$, the results being robust to these standard choices is spectral analysis. We use three sample sizes, $T = 100$, 200 and 400 and report results across 10,000, 5000 and 1000 replications for three values of the normalizing parameter $\sigma \in \{0.5, 1.0, 2.0\}$ for both *i.i.d.* and *mds* criteria. We only report results for $\sigma = 1.0$; complete results can be found in Appendix E, Velasco (2022).

In our first experiment, we consider standardized non-Gaussian *i.i.d.* innovations ε_t given by exponential, t_5 and uniform distributions as in Velasco and Lobato (2018), where the first one is highly asymmetric ($\kappa_3 = 2$) with strong positive kurtosis ($\kappa_4 = 6$) and the other two distributions are symmetric, with positive ($\kappa_4 = 6$) and negative kurtosis ($\kappa_4 = -1.2$), respectively. We consider AR(1), MA(1) and ARMA(1, 1) models with parameters 0.5 and 0.9 to generate causal and invertible processes and their reciprocals to generate noncausal and noninvertible models. Noncausal models are simulated using their stationary forward-looking moving average representation.

In Tables 1 to 3, we report the percentage of correct identification of the location of the lag polynomial roots with respect to the unit circle across simulations to investigate the global identification achieved by our estimation criteria. For both objective functions, based on the

TABLE 1
AR(1) *Percentage of correct root identification*, i.i.d. ε_t

			$\mathcal{L}_T^{i.i.d.}$				$\mathcal{L}_T^{mds}$			
ε_t	T	σ	**0.5**	**0.9**	**0.9**$^{-1}$	**0.5**$^{-1}$	**0.5**	**0.9**	**0.9**$^{-1}$	**0.5**$^{-1}$
Exp	**100**	**1.0**	97.43	88.13	89.04	99.82	98.14	89.85	52.31	69.39
	200	**1.0**	97.95	96.35	96.34	100.00	98.00	96.18	55.32	75.57
	400	**1.0**	100.00	99.50	99.50	100.00	100.00	99.20	57.20	97.30
t_5	**100**	**1.0**	67.41	59.28	58.12	70.00	59.04	59.41	49.70	62.35
	200	**1.0**	78.13	62.36	61.55	78.11	65.56	62.91	51.81	64.36
	400	**1.0**	89.20	66.60	64.80	88.20	67.90	67.20	52.00	68.50
Unif	**100**	**1.0**	74.53	53.72	53.53	74.40	50.85	55.59	48.05	55.76
	200	**1.0**	87.68	61.76	61.99	87.48	53.59	69.12	48.23	62.55
	400	**1.0**	96.40	75.20	72.30	95.60	59.70	83.80	47.20	74.10

Note: Percentage of replications in which $\mathcal{L}_T^{i.i.d.}(\theta)$ and $\mathcal{L}_T^{mds}(\theta)$ are minimized for a value θ_T so that $\mathbf{1}\{|\theta_T| < 1\} = \mathbf{1}\{|\theta_0| < 1\}$. $\theta_0 = 0.5, 0.9, 0.5^{-1}, 0.9^{-1}$.

i.i.d. or the *mds* assumptions, we only report results for $\mathcal{L}_T^{\bullet}$ estimates because results for $Q_T^{\bullet}$ are fairly similar, as they use initial values equal to $\hat{\theta}_T^{\bullet}$. The results in Tables 1 and 2 are very similar for simple models, as AR(1) and MA(1) models face symmetric identification problems. Both estimation approaches report similar results for the exponential distribution, but the *i.i.d.* criterion provides symmetric identification results between roots outside or inside the unit circle, while the *mds* criterion can be very asymmetric, especially for roots close to the unit circle, $\theta = 0.9, 0.9^{-1}$, possibly due to the nonreversibility of the *mds* property. For the other distributions, which are symmetric, the *mds* criterion results are more balanced, but typically dominated by the *i.i.d.* criterion, which reports similar results for the uniform distribution to those of the higher order moments-based method of Velasco and Lobato (2018) for the MA(1) model and but slightly worse for the t_5 distribution. For exponential innovations, our estimates improve the results of moment estimation despite not directly focusing of the strong skewness of this distribution. In terms of the choice of σ, there are not systematic differences in many cases, but $\sigma = 0.5$ and 1.0 seem to perform better for the exponential and t_5 distributions and $\sigma = 1.0$ for the uniform, so $\sigma = 1.0$ seems an overall robust choice once residuals are standardized.

In Table 3 reporting the results for the ARMA(1,1) models, we observe a similar pattern, though identification of root location becomes more difficult with the complexity of the model, noting that there are now up to four potential root configurations. The exponential case is still the easier to identify, while the results in the uniform case seem to improve quite slowly with sample size, especially for the *mds* criterion.

Our second experiment investigates model identification with *mds* innovations generated by a GARCH model $\varepsilon_t = \gamma_t e_t$ with $\gamma_t^2 = 1 + 0.8\varepsilon_{t-1}^2 + 0.2\gamma_{t-1}^2$ and centered *i.i.d.* exponentially distributed e_t. The main conclusion from the results in Table 4 is the asymmetry for *i.i.d.* and *mds* estimation criteria for both AR(1) and MA(1) models. The loss function based on *mds* works as expected for models with roots outside the unit circle (with the *i.i.d.* criterion obtaining not surprisingly misleading results as T grows for some σ). However, for models with roots inside the unit circle, the $\mathcal{L}_T^{mds}$ criterion loses much of its identification power, improving very slowly with sample size, while $\mathcal{L}_T^{i.i.d.}$ has a reasonable performance despite simulated series have no linear representation with fully independent innovations. It appears that the wrong linear representation, together with the higher order dependence of the innovations, leads to stronger departures from the *i.i.d.* hypothesis than from the *mds* one, so the *mds* criterion, which for the t distribution was already performing slightly worse for

TABLE 2
MA(1) *Percentage of correct root identification*, i.i.d. ε_t

			$\mathcal{L}_T^{i.i.d.}$				$\mathcal{L}_T^{mds}$			
ε_t	T	σ	**0.5**	**0.9**	$\mathbf{0.9^{-1}}$	$\mathbf{0.5^{-1}}$	**0.5**	**0.9**	$\mathbf{0.9^{-1}}$	$\mathbf{0.5^{-1}}$
Exp	**100**	**1.0**	99.39	89.66	86.06	99.39	96.10	62.66	84.60	69.39
	200	**1.0**	99.72	98.05	96.92	99.98	99.69	70.44	94.16	75.57
	400	**1.0**	100.00	99.90	99.90	100.00	100.00	81.40	99.20	83.60
t_5	**100**	**1.0**	69.19	61.19	55.85	69.32	59.61	53.35	54.33	59.72
	200	**1.0**	78.21	62.72	59.19	77.06	62.99	54.52	55.61	59.67
	400	**1.0**	88.30	69.30	65.80	87.80	69.40	54.80	60.20	61.10
Unif	**100**	**1.0**	73.26	57.76	52.47	71.63	58.24	52.07	56.88	49.61
	200	**1.0**	86.96	63.63	58.59	86.63	65.08	53.09	67.65	52.42
	400	**1.0**	96.40	73.50	72.10	96.70	75.90	55.40	80.70	59.50

Note: See Table 1.

independent innovations, has more difficulties to pick up the right model for moderate sample sizes.

In our third simulation analysis, we investigate the empirical root mean square error (RMSE) of parameter estimates of the moving average parameter of MA(1) and ARMA(1,1) models with *i.i.d.* innovations (see Tables 5* and 6* in Appendix E in Velasco (2022)). We use both (*i.i.d.* and *mds*) identification criteria with both spectral cumulative ($\mathcal{L}_T^{\bullet}$) and density ($Q_T^{\bullet}$) functions for initial estimates ($\hat{\theta}_T$ and $\tilde{\theta}_T$, respectively). We also consider efficient GMM estimates $\tilde{\theta}_T^{gmm}$ using jointly *i.i.d.*, *mds* and *wns* or *mds* and *wns* moments and efficient CGMM estimates $\bar{\theta}_T$ based on $\bar{Q}_T^{i.i.d.}$ and $\bar{Q}_T^{mds}$. For the later estimate, we use $\alpha_T = T^{-2}$, which, despite does not satisfy the assumptions of Theorems 6 and 7, performed well when estimating the asymptotic variance of criterion scores for Gaussian errors. RMSE calculations only use replications in which the location of the lag polynomials were correctly identified.

In terms of efficiency of nonoptimal methods, *i.i.d.* based estimation works better than *mds* for exponential and t_5 distributions, but it is outperformed by *mds* estimation methods for invertible models with uniform innovations (Table 5*). Whittle estimation imposing

TABLE 3
ARMA(1,1) *Percentage of correct root identification*, $\varepsilon_t \sim$i.i.d.

			$\mathcal{L}_T^{i.i.d.}$				$\mathcal{L}_T^{mds}$			
ε_t	T	σ	**C-I**	**NC-I**	**C-NI**	**NC-NI**	**C-I**	**NC-I**	**C-NI**	**NC-NI**
Exp	**100**	**1.0**	95.02	88.95	89.05	93.47	89.38	71.98	68.82	54.85
	200	**1.0**	99.60	98.28	98.16	99.48	98.54	85.56	80.94	66.12
	400	**1.0**	100.00	100.00	100.00	100.00	100.00	94.90	90.20	79.50
t_5	**100**	**1.0**	42.77	41.58	40.13	41.76	34.88	36.29	33.61	34.76
	200	**1.0**	54.04	47.86	48.00	52.76	40.66	38.76	35.20	37.84
	400	**1.0**	68.60	55.90	56.50	67.70	45.00	43.00	39.20	41.20
Unif	**100**	**1.0**	40.74	41.15	41.11	37.81	31.55	31.71	25.64	20.85
	200	**1.0**	53.04	55.38	56.66	50.46	36.14	39.68	30.02	23.50
	400	**1.0**	65.90	76.30	75.10	63.00	41.40	55.50	38.40	28.50

Note: Percentage of replications in which $\mathcal{L}_T^{i.i.d.}(\theta)$ and $\mathcal{L}_T^{mds}(\theta)$ are minimized for a value θ_T, which identifies correctly the location of both the AR and MA roots with respect the complex unit circle. Models: C-I is causal and invertible, $\theta_0 = (0.5, 0.5)'$; NC-I is noncausal and invertible, $\theta_0 = (1/0.5, 0.5)'$; C-NI is causal and noninvertible, $\theta_0 = (0.5, 1/0.5)'$; NC-NI is noncausal and noninvertible, $\theta_0 = (1/0.5, 1/0.5)'$.

TABLE 4
AR(1) & *MA*(1) *Percentage of correct root ident.* $\varepsilon_t \sim$mds.*GARCH.Exp*(1)

			AR(1)				MA(1)			
			0.5		**0.5**$^{-1}$		**0.5**		**0.5**$^{-1}$	
ε_t	T	σ	$\mathcal{L}_T^{i.i.d.}$	$\mathcal{L}_T^{mds}$	$\mathcal{L}_T^{i.i.d.}$	$\mathcal{L}_T^{mds}$	$\mathcal{L}_T^{i.i.d.}$	$\mathcal{L}_T^{mds}$	$\mathcal{L}_T^{i.i.d.}$	$\mathcal{L}_T^{mds}$
Exp	**100**	**1.0**	54.79	61.54	67.35	47.39	41.84	61.04	63.32	44.18
	200	**1.0**	53.70	79.71	72.70	45.86	35.60	72.46	66.33	46.22
	400	**1.0**	53.50	89.74	79.65	45.60	27.35	84.75	69.20	52.05
	800	**1.0**	93.90	99.40	99.20	69.40	80.70	99.20	99.10	68.30

Note: See Table 1.

invertibility and causality (and inverting the roots for noninvertible or noncausal models) performs systematically worse than our estimates, which exploit non-Gaussian information, except again for the uniform distribution

Across parameter configurations and choices of σ, there are no systematic differences between $\hat{\theta}_T$ and $\tilde{\theta}_T$ exploiting spectral distribution and density functions, respectively, but the latter shows better relative efficiency in more persistent set-ups as expected, though the empirical RMSE seems very sensitive to a small number of outlying replications. For exponential and uniform distributions, $\sigma = 1.0$ seems the best choice in terms of robustness and efficiency, but for t_5 innovations $\sigma = 0.5$ provides better results in most model and estimation method configurations.

The two optimal GMM estimates can improve substantially with respect to unweighted methods, specially for *i.i.d.* based estimation. We only report results for $\sigma = 1.0$ as results after optimal weighting of moment conditions are more stable across σ. CGMM provides improved efficiency overall, specially for large sample sizes, with the RMSE of GMM estimates being more sensitive for roots close to the unit circle, perhaps because it relies on asymptotic approximations to the score variances instead of standardizing directly empirical moments as in CGMM. These optimal estimates under *i.i.d.* and *mds* perform noticeably better than the higher order moments method of Velasco and Lobato (2018) for exponential errors. For t and uniform errors and nonminimum phase models our estimates also improve on those, with comparable performance for noninvertible models.

The results in Table 6* for the ARMA(1,1) model and exponential errors lead to analogous conclusions, with weighted estimates displaying a more erratic behavior for the smallest sample size. However, for the larger samples, the efficient CGMM estimate outperforms almost uniformly any other estimate also for this more complex model, being our general recommended procedure for not too short time series, considering the modest additional computation cost from initial estimates like $\hat{\theta}_T$, both based on standardized residuals with $\sigma = 1.0$. With respect to which restriction to apply, *mds* is more robust by construction, but in moderate sample sizes $i.i.d.$ estimates typically have a higher rate of success to identify the right model, even with misspecified innovations dependence, and smaller RMSE. We illustrate this estimation strategy in the empirical application described in Appendix F of Velasco (2022).

Acknowledgments. The author would like to thank an Associate Editor and two referees for very helpful comments on earlier versions of the paper.

Funding. Financial support from Ministerio de Economía y Competitividad (Spain), grants ECO2017-86009-P and PID2020-114664GB-I00 is gratefully acknowledged.

SUPPLEMENTARY MATERIAL

Technical and numerical appendices (DOI: 10.1214/22-AOS2220SUPPA; .pdf). The Supplementary Material contains six Appendices: A. Proofs of results, B. Auxiliary Lemmas, C. Asymptotic variance for Gaussian errors and W, D. Closed-form expressions for Gaussian W, E. Extended Monte Carlo Results, F. Empirical Application.

Matlab Code (DOI: 10.1214/22-AOS2220SUPPB; .zip). MATLAB code is provided for replicating the simulations reported in the paper.

REFERENCES

ANATOLYEV, S. (2007). Optimal instruments in time series: A survey. *J. Econ. Surv.* **21** 143–173.

ANDREWS, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59** 817–858. MR1106513 https://doi.org/10.2307/2938229

ANDREWS, B., DAVIS, R. A. and BREIDT, F. J. (2007). Rank-based estimation for all-pass time series models. *Ann. Statist.* **35** 844–869. MR2336871 https://doi.org/10.1214/009053606000001316

BLOOMFIELD, P. (1973). An exponential model for the spectrum of a scalar time series. *Biometrika* **60** 217–226. MR0323048 https://doi.org/10.1093/biomet/60.2.217

BOX, G. E. P. and PIERCE, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *J. Amer. Statist. Assoc.* **65** 1509–1526. MR0273762

BREIDT, F. J., DAVIS, R. A. and TRINDADE, A. A. (2001). Least absolute deviation estimation for all-pass time series models. *Ann. Statist.* **29** 919–946. MR1869234 https://doi.org/10.1214/aos/1013699986

CARRASCO, M., CHERNOV, M., FLORENS, J.-P. and GHYSELS, E. (2007). Efficient estimation of general dynamic models with a continuum of moment conditions. *J. Econometrics* **140** 529–573. MR2408918 https://doi.org/10.1016/j.jeconom.2006.07.013

CHEN, B., CHOI, J. and ESCANCIANO, J. C. (2017). Testing for fundamental vector moving average representations. *Quant. Econ.* **8** 149–180. MR3638602 https://doi.org/10.3982/QE393

CHENG, Q. S. (1992). On the unique representation of non-Gaussian linear processes. *Ann. Statist.* **20** 1143–1145. MR1165613 https://doi.org/10.1214/aos/1176348677

DEDECKER, J., DOUKHAN, P., LANG, G., LEÓN R., J. R., LOUHICHI, S. and PRIEUR, C. (2007). *Weak Dependence*: *With Examples and Applications*. *Lecture Notes in Statistics* **190**. Springer, New York. MR2338725

DOUKHAN, P. and LOUHICHI, S. (1999). A new weak dependence condition and applications to moment inequalities. *Stochastic Process. Appl.* **84** 313–342. MR1719345 https://doi.org/10.1016/S0304-4149(99)00055-1

ESCANCIANO, J. C. and VELASCO, C. (2006a). Generalized spectral tests for the martingale difference hypothesis. *J. Econometrics* **134** 151–185. MR2328319 https://doi.org/10.1016/j.jeconom.2005.06.019

ESCANCIANO, J. C. and VELASCO, C. (2006b). Testing the martingale difference hypothesis using integrated regression functions. *Comput. Statist. Data Anal.* **51** 2278–2294. MR2307501 https://doi.org/10.1016/j.csda.2006.07.039

FEUERVERGER, A. (1990). An efficiency result for the empirical characteristic function in stationary time-series models. *Canad. J. Statist.* **18** 155–161. MR1067167 https://doi.org/10.2307/3315564

FOKIANOS, K. and PITSILLOU, M. (2018). Testing independence for multivariate time series via the auto-distance correlation matrix. *Biometrika* **105** 337–352. MR3804406 https://doi.org/10.1093/biomet/asx082

GASSIAT, É. (1993). Adaptive estimation in noncausal stationary AR processes. *Ann. Statist.* **21** 2022–2042. MR1245779 https://doi.org/10.1214/aos/1176349408

GORODECKIĬ, V. V. (1977). The strong mixing property for linearly generated sequences. *Theory Probab. Appl.* **22** 411–413. MR0438450

GOSPODINOV, N. and NG, S. (2015). Minimum distance estimation of possibly noninvertible moving average models. *J. Bus. Econom. Statist.* **33** 403–417. MR3372667 https://doi.org/10.1080/07350015.2014.955175

GOURIÉROUX, C. and ZAKOÏAN, J.-M. (2017). Local explosion modelling by non-causal process. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 737–756. MR3641405 https://doi.org/10.1111/rssb.12193

HANNAN, E. J. (1970). *Multiple Time Series*. Wiley, New York. MR0279952

HONG, Y. (1999). Hypothesis testing in time series via the empirical characteristic function: A generalized spectral density approach. *J. Amer. Statist. Assoc.* **94** 1201–1220. MR1731483 https://doi.org/10.2307/2669935

HONG, Y. and LEE, Y.-J. (2005). Generalized spectral tests for conditional mean models in time series with conditional heteroscedasticity of unknown form. *Rev. Econ. Stud.* **72** 499–541. MR2129829 https://doi.org/10.1111/j.1467-937X.2005.00341.x

HOSKING, J. R. M. (1981). Fractional differencing. *Biometrika* **68** 165–176. MR0614953 https://doi.org/10.1093/biomet/68.1.165

HUANG, J. and PAWITAN, Y. (2000). Quasi-likelihood estimation of non-invertible moving average processes. *Scand. J. Stat.* **27** 689–702. MR1804170 https://doi.org/10.1111/1467-9469.00216

KNIGHT, J. L. and YU, J. (2002). Empirical characteristic function in time series estimation. *Econometric Theory* **18** 691–721. MR1906331 https://doi.org/10.1017/S026646660218306X

KREISS, J.-P. (1987). On adaptive estimation in stationary ARMA processes. *Ann. Statist.* **15** 112–133. MR0885727 https://doi.org/10.1214/aos/1176350256

LANNE, M. and SAIKKONEN, P. (2011). Noncausal autoregressions for economic time series. *J. Time Ser. Econom.* **3** Art. 2. MR2928654 https://doi.org/10.2202/1941-1928.1080

LII, K. S. and ROSENBLATT, M. (1982). Deconvolution and estimation of transfer function phase and coefficients for non-Gaussian linear processes. *Ann. Statist.* **10** 1195–1208. MR0673654

LII, K.-S. and ROSENBLATT, M. (1992). An approximate maximum likelihood estimation for non-Gaussian non-minimum phase moving average processes. *J. Multivariate Anal.* **43** 272–299. MR1193615 https://doi.org/10.1016/0047-259X(92)90037-G

LII, K.-S. and ROSENBLATT, M. (1996). Maximum likelihood estimation for non-Gaussian nonminimum phase ARMA sequences. *Statist. Sinica* **6** 1–22. MR1379046

LIPPI, M. and REICHLIN, L. (1994). VAR analysis, nonfundamental representations, Blaschke matrices. *J. Econometrics* **63** 307–325. MR1309986 https://doi.org/10.1016/0304-4076(93)01570-C

PINKSE, J. (1998). A consistent nonparametric test for serial independence. *J. Econometrics* **84** 205–231. MR1630190 https://doi.org/10.1016/S0304-4076(97)00084-5

RAMSEY, J. B. and MONTENEGRO, A. (1992). Identification and estimation of non-invertible non-Gaussian MA(q) processes. *J. Econometrics* **54** 301–320.

ROSENBLATT, M. (2000). *Gaussian and Non-Gaussian Linear Time Series and Random Fields*. *Springer Series in Statistics*. Springer, New York. MR1742357 https://doi.org/10.1007/978-1-4612-1262-1

STINCHCOMBE, M. B. and WHITE, H. (1998). Consistent specification testing with nuisance parameters present only under the alternative. *Econometric Theory* **14** 295–325. MR1628586 https://doi.org/10.1017/S0266466698143013

SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35** 2769–2794. MR2382665 https://doi.org/10.1214/009053607000000505

TJØSTHEIM, D., OTNEIM, H. and STØVE, B. (2022). Statistical dependence: Beyond Pearson's ρ. *Statist. Sci.* **37** 90–109. MR4371097 https://doi.org/10.1214/21-sts823

VELASCO, C. (2022). Supplement to "Estimation of time series models using residuals dependence measures." https://doi.org/10.1214/22-AOS2220SUPPA, https://doi.org/10.1214/22-AOS2220SUPPB

VELASCO, C. and LOBATO, I. N. (2018). Frequency domain minimum distance inference for possibly non-invertible and noncausal ARMA models. *Ann. Statist.* **46** 555–579. MR3782377 https://doi.org/10.1214/17-AOS1560

YAO, S., ZHANG, X. and SHAO, X. (2018). Testing mutual independence in high dimension via distance covariance. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 455–480. MR3798874 https://doi.org/10.1111/rssb.12259

YOSHIHARA, K. (1978). Moment inequalities for mixing sequences. *Kodai Math. J.* **1** 316–328. MR0508749