An Alternative Asymptotic Analysis of Residual-Based Statistics

Elena Andreou* and Bas J.M. Werker‡
University of Cyprus and Tilburg University

May 20, 2004

Abstract

This paper offers an alternative technique to derive the limiting distribution of residual-based statistics or, more general, the limiting distribution of statistics with estimated nuisance parameters. This technique allows us to unify many known results on two-stage estimators and tests and we also derive new results. The technique is especially useful in situations where smoothness of the statistic of interest with respect to the parameters to be estimated does not hold or is difficult to establish, e.g., rank-based statistics. We essentially replace this differentiability condition with a distributional invariance property that is often satisfied in specification tests. Our results on statistics that have not been considered before all use nonparametric statistics. On the technical side, we provide a novel approach to the pre-estimation problem using Le Cam's third lemma. The resulting formula for the correction in the limiting variance as a result of pre-estimation some parameters is a simple expression involving some appropriate covariances. The regularity conditions required fairly minimal. Numerous examples show the strength and wide applicability of our approach.

JEL codes: C32, C51, C52.
Keywords and phrases. Asymptotic size, Discretized estimators, Goodness-of-Fit tests, Local asymptotic normality, Rank statistics, Structural break tests, Temporal dependence tests, Two-stage inference.

---

*Department of Economics, University of Cyprus, P.O.Box 537, CY1678, Nicosia, Cyprus.

†Econometrics and Finance Group, CentER, Tilburg University, P.O.Box 90153, 5000 LE, Tilburg, The Netherlands.

1   Introduction

Residual-based tests represent an important area of research in econometrics, generally used for diagnostic checking of a proposed statistical model. Such specification tests are covered in many econometrics and statistics textbooks and remain of interest in ongoing research[1]. Similarly, residual-based estimators (often referred to as two-stage estimators) are widely applied in econometric work[2]. Usually, the asymptotic distribution of residual-based statistics (tests or estimators) is derived on a case-by-case basis, using a particular model specification and/or some stringent assumptions about the statistic and the first-stage estimators employed. The key assumption for deriving the limiting distribution of a residual-based statistic or estimator is a smoothness condition (e.g., differentiability) of the statistic with respect to the estimator. In this traditional approach, the question whether the limiting distribution of a statistic is affected by the estimator or not, is primarily a question of whether or not it has a nonzero derivative with respect to the estimator (see for instance, Pierce, 1982, and Randles, 1982). We present a new and alternative approach that does not involve such differentiability conditions. Instead, we rely on an invariance condition which is generally trivially satisfied for residual-based statistics. Especially in cases where smoothness conditions do not hold or are non-trivial to establish (as, for instance, is the case for many rank-based statistics), our approach offers a useful and unifying alternative.

Our alternative and novel approach for deriving the asymptotic distribution of residual-based statistics, or, more generally, the limiting distribution of statistics that involve estimated nuisance parameters, is based on Le Cam's third lemma applied to Locally Asymptotically Normal (LAN) models (see the appendix for a very brief summary of Le Cam's third lemma). The proposed method owns the following advantages: (i) It is based on the unifying and coherent framework of Hájek and Le Cam's theory of asymptotic statistics (see, e.g., Bickel et al., 1993, Le Cam and Yang, 1990, and Pollard, 2004). As such it applies to general model specifications, as long as they satisfy the LAN condition, and the limiting distribution of statistics need not be derived on a case-by-case model specific basis. (ii) It offers a simple, yet general, method for deriving the asymptotic distribution of various statistics using arbitrary initial consistent estimators, subject to only weak regularity conditions. (iii) It does not require asymptotic smoothness conditions. (iv) It generalizes the

---

[1]A non-exhaustive list of references is Andrews (1988a, 1988b), Bera and Jarque (1982), Bierens (1990), Chesher and Irish (1997), Godfrey (1988), Hallin and Puri (1990), Koul and Stute (1999), Lundberg and Terasvirta (2002), McKean, Sheather, and Hettmansperger (1990), MacKinnon (1992), Newey (1985), Newey and McFadden (1994), Pagan and Vella (1989), Ploberger and Kramer (1992), Spanos (1986), Stute (1997), Tauchen (1985), White (1982), and the Journal of Econometrics special issues edited by Blundell (1987), Hillier and King (1991), and Keuzenkamp and Magnus (1995).

[2]For example, Anderson and Takamitsu (1979), Amemiya (1982), Pagan (1986), and Powell (1983).

traditional asymptotic methods based on the pioneering work of Pierce (1982) and Randles (1982). More precisely, it covers most existing residual-based tests (both classical and recent) which can now also be studied for more general dynamic models (think of, e.g., residual-based dependence tests). (v) It allows for several new results, e.g., in the area of rank-based tests for temporal dependence and heterogeneity.

The main Theorem 2.1 of the paper shows, under a LAN condition and an appropriate asymptotic invariance condition for the statistic of interest, that the residual-based statistic is asymptotically normally distributed with a variance that is a simple function of the limiting variance of the innovation-based statistic[3] and the estimator, as well as the covariances of the innovation-based statistic with the central sequence and with the estimator. Besides the wide applicability of this theorem, it also provides insights into the general structure of the problem by indicating precisely when the asymptotic variance of the residual-based static equals that of the innovation-based one, exceeds it, or is smaller (which, contrary to widespread believe, often happens as well). Moreover, the theorem can be used directly to assess the local power of residual-based tests. Therefore the technical analysis of the paper contributes to the pre-estimation problem of statistics using Le Cam's theory of convergence of experiments.

The traditional approach derives the limiting distribution of statistics that involve an estimated nuisance parameter making either model-specific assumptions and/or assuming appropriate differentiability conditions. Within Le Cam's framework we state some high-level assumptions regarding local asymptotic normality of the model and the joint asymptotic distribution of the central sequence, the test statistic, and the estimator. More precisely, our asymptotic analysis replaces the differentiability assumption by a distributional invariance condition which is trivially satisfied in most residual-based statistics. The central sequence is the new element in our asymptotic analysis. In the traditional approach the result that the asymptotic variance of the statistic is not affected by estimation is due to a zero derivative of the statistic with respect to the estimator. In our framework it follows when the statistic and the central sequence are orthogonal. The proposed method also yields the limiting distribution when distributional invariance does not hold, subject to somewhat more cumbersome calculus. In addition, our approach can be considered for semi- and nonparametric models for which the central sequence will not enter the limiting distribution of the estimator or statistic. In this context, the results of Theorem 2.1 yield an auxiliary model that can be considered as the least-favorable parametric submodel for semiparametric efficiency with respect to the statistic of interest.

Based on Theorem 2.1, we address a vast number of applications to residual-based statistics that cover tests from all three categories of assumptions on model innovations – dependence, distributional assumptions, and heterogeneity. It is shown

---

[3]Throughout the paper, we use the term innovation-based statistic for the statistic applied to the true innovations in the model, i.e., the residuals obtained if the true value of the parameter of interest were known. The precise meaning of this term will be clear in all the examples of Section 4.

how our method can be used as an alternative, simple technique to reach the same asymptotic distribution of many residual-based tests in the (classical and recent) literature and as a method that provides new results on the asymptotic distribution of residual-based tests such as rank-statistics for serial dependence or structural breaks. The paper presents the three categories of examples as applications of Theorem 2.1 for a large family of both location and scale time series models that satisfy the LAN condition (such as ARMA and GARCH models). Our focus on time-series models in the applications we discuss is only for illustrative reasons. The first category deals with temporal dependence tests such as linear correlation residual tests (e.g., the Ljung and Box, 1978, test), second-order correlation tests (e.g., the McLeod and Li, 1983, test), and rank-based residual correlation tests (see Hallin and Werker, 1999, for an overview). For a general scale model we obtain the result on the squared-residual correlation tests that has recently been derived in Berkes, Horvath, and Kokoszka (2003) for the more specific GARCH model. The asymptotic results for rank-based serial dependence tests are new in the econometric literature. The second category of tests revisits goodness-of-fit statistics based on the empirical distribution function and tests of homogeneity. Applying our main result, we find for the residual-based tests in location and scale time series models an asymptotic distribution which is that originally due to Durbin (1973) and recently extended, for instance, in Koul and Stute (1999) and Bai (2003). Our third category of tests considers structural break CUSUM tests based on the ranks of the residuals and shows that these tests are asymptotically distribution free for both location and scale time series models. This is a new and complementary result to the case of regression models in Sen (1984) and to other empirical distribution function statistics (Horvath, Kokoszka and Teyssiere, 2001, and Koul, 2002) in the literature. In addition we present one application of the limiting distribution of two-stage estimators in AR-GARCH specifications. Finally, we also consider discrete choice model and tests, based on traditional conditional moments, for omitted variables. This illustrates that our analysis is also valid in LAN limited dependent variable models. Moreover, for this application we present and study a new nonparametric test for omitted variables based on Kendall's tau.

The rest of the paper is organized as follows. In Section 2 we formulate our basic idea for deriving the limiting distribution of a statistic when applied to some model's residuals. To illustrate our results, we consider in that section the classical problem of autocorrelation tests applied to residuals of an estimated AR(1) model as a running example. Moreover, we show how our results relate to the more standard approach based on smoothness conditions of the statistics with respect to the estimated parameters. Finally, we indicate that a technical complication arises when making our ideas rigorous. Section 3 deals with this complication by discretizing the estimator of the model's parameters and we provide a formal proof of the limiting distribution of the residual-based statistic, for a discretization that becomes finer and finer with the sample size. Section 4 gives many applications to illustrate that our approach is a general and simple technique for deriving the limiting distribution of residual-based

statistics in a wide range of models, statistics, and first-stage estimators. Section 5 concludes.

2   Main results: Intuition

Our goal is to give a new and widely applicable method to derive the asymptotic distribution of statistics applied to residuals of some statistical or econometric model. As mentioned in the introduction, this classical problem occurs in many applications, several of which are discussed in Section 4. Traditionally, the appropriate size-adjustments for residual-based tests have been calculated using some form of differentiability of the test-statistic as a function of the underlying parameters. While this approach works for many interesting situations, it is more difficult to apply in situations such as rank and/or sign-based statistics, due to the inherent non-differentiability of these kinds of statistics. We propose an alternative approach that does not require any analytical (e.g., differentiablility) smoothness or some form of asymptotic linearity of the statistics. Our proposed method requires some high-level assumptions. In particular, we impose an asymptotic distributional invariance property on the statistic which is generally satisfied in a specification testing setting. We also impose regularity on the underlying distributions in an appropriate Local Asymptotic Normality sense and we resort to discretized estimators (see Section 3). The relation of our method with the traditional approach is outlined in more detail in Section 2.3.

Our results are derived within the Hájek and Le Cam framework of Locally Asymptotically Normal (LAN) models. "Most" of the common models in econometrics and statistics are LAN. The LAN property has been considered in regression models by, e.g., Bickel (1982) and Fabian and Hannan (1982). Autoregressive models are LAN as shown by Kreiss (1987a), as well as ARMA models that are discussed in Kreiss (1987b). Non-linear regression and autoregression models are LAN for regression functions that are smooth in the parameters as is shown in Drost, Klaassen, and Werker (1997). Panel data models are discussed in Hahn and Kuersteiner (2002) and limited dependent variables in Bickel et al. (1993). ARCH-type models were shown to satisfy the LAN condition by Linton (1993) and GARCH-type models were treated in detail by Drost and Klaassen (1997) and Sun and Tsengos (2004). Duration models like the Autoregressive Conditional Duration model of Engle and Russell (1998) are discussed in Drost and Werker (2003). Two final references are Bickel et al. (1993) that discusses other classes of LAN models with iid observations and Taniguchi and Kakizawa (2000) that discusses several more general time-series models. Applications of our results to the above time series models are given in Section 4. It is illustrative to mention also some models where the LAN condition is not satisfied. Two common phenomena generally lead to non-LAN behavior: non-stationary data and non-smooth functional dependence. To start with the first, models for non-

stationary (and possibly cointegrated) processes are discussed in a series of papers, Jegannathan (1995, 1997, 1999), where quadratic likelihood approximations as in the LAN condition are derived. However, the limiting distribution of the central sequence in these models is no longer normal, but a scale-mixture of normals. The situation of non-smooth functional dependence of a regression function, for instance, occurs in Threshold AutoRegressive models. In such models, the regression function is of the form $m(x) = m_1(x)I\{x \leq x_0\} + m_2(x)I\{x > x_0\}$. The fact that this regression function is not-differentiable with respect to the threshold parameter x_0 leads to a situation where the limiting experiment is not Gaussian as in the LAN case, but of a compound Poisson type. Inference for the threshold parameter is discussed in, e.g., Qian (1998) and Hansen (2000). While these latter two cases present models that are not LAN, the idea of our approach is likely to carry over to these situations since Le Cam's third lemma, on which our results are based, is not restricted to the LAN situation. However, the details are beyond the scope of the present paper. Some concluding remarks suggest possible ways to extend this method.

2.1 Conditions

Before we introduce the LAN condition, let us formalize the statistical model we are interested in. Let $\mathcal{E}^{(n)}$ denote a sequence of experiments defined on a common parameter set $\Theta \subset \mathbb{R}^k$:

$$\mathcal{E}^{(n)} = \left\{ X^{(n)}, \mathcal{A}^{(n)}, \mathcal{P}^{(n)} = \left\{ \mathbb{P}_\theta^{(n)} \; : \; \theta \in \Theta \right\} \right\},$$

where $\left(X^{(n)}, \mathcal{A}^{(n)} \right)$ is a sequence of measurable spaces and, for each n and $\theta \in \Theta$, $\mathbb{P}_\theta^{(n)}$ a probability measure on $\left(X^{(n)}, \mathcal{A}^{(n)} \right)$. We assume throughout this paper that Θ is a subset of $\mathbb{R}^k$ so that we consider the effect of pre-estimating a Euclidean parameter. While we focus our discussion on the situation where the underlying model is parametric, we do not exclude semi- or nonparametric models. These are discussed in Section 2.4. Also, we assume throughout that pertinent asymptotics in the sequence of experiments takes place at the usual $\sqrt{n}$ rate. Other rates can be easily adopted at the cost of adapted notation only. Let θ_0 denote a fixed value of the parameter of interest θ and let θ_n and θ'_n denote sequences local (sometimes called contiguous) to θ_0, i.e., $\delta_n = \sqrt{n}(\theta_n - \theta_0)$ and $\delta'_n = \sqrt{n}(\theta'_n - \theta_0)$ are bounded in $\mathbb{R}^k$. Write $\Lambda^{(n)}(\theta'_n|\theta_n) = \log \left(d\mathbb{P}_{\theta'_n}^{(n)} / d\mathbb{P}_{\theta_n}^{(n)} \right)$ for the log-likelihood of $\mathbb{P}_{\theta'_n}^{(n)}$ with respect to $\mathbb{P}_{\theta_n}^{(n)}$. In case $\mathbb{P}_{\theta'_n}^{(n)}$ is not dominated by $\mathbb{P}_{\theta_n}^{(n)}$, we mean the Radon-Nikodym derivative of the absolute continuous part in the Lebesgue decomposition of $\mathbb{P}_{\theta'_n}^{(n)}$ with respect to $\mathbb{P}_{\theta_n}^{(n)}$ (see Strasser, 1985, Definition 1.3). We impose a condition which is somewhat stronger than LAN. This condition is usually referred to as Uniform Local Asymptotic Normality (ULAN) and is generally indispensable for the construction of efficient inference procedures. Although not all papers cited above discuss this uniform version, it is satisfied in all these models. As a matter of fact, we are unaware

of any non-trivial statistical model which is LAN but not ULAN.

Condition (ULAN): The sequence of experiments $\mathcal{E}^{(n)}$ is Uniformly Locally Asymptotically Normal (ULAN) in the sense that there exists a sequence of random variables $\Delta^{(n)}(\theta)$ (the *central sequence*) such that for all sequences θ_n and θ'_n local to θ_0, we have

$$\Lambda^{(n)}(\theta'_n|\theta_n) = (\delta'_n - \delta_n)^T \Delta^{(n)}(\theta_0 + \delta_n/\sqrt{n}) - \frac{1}{2}(\delta'_n - \delta_n)^T I_F(\delta'_n - \delta_n) + o_{\mathrm{P}}(1)$$

$$= (\delta'_n - \delta_n)^T \Delta^{(n)}(\theta_0 + \delta'_n/\sqrt{n}) + \frac{1}{2}(\delta'_n - \delta_n)^T I_F(\delta'_n - \delta_n) + o_{\mathrm{P}}(1),$$

where, as before, $\theta_n = \theta_0 + \delta_n/\sqrt{n}$ and $\theta'_n = \theta_0 + \delta'_n/\sqrt{n}$. Moreover, the central sequence $\Delta^{(n)}(\theta_n)$ is asymptotically normally distributed with zero mean and variance I_F, i.e., $\Delta^{(n)}(\theta_n) \xrightarrow{\mathcal{L}} N(0, I_F)$, as $n \to \infty$, under $\mathbb{P}^{(n)}_{\theta_n}$. I_F is called the Fisher information matrix. $\square$

REMARK 2.1 The (U)LAN condition presents a prime example in the theory of convergence of statistical experiments. The quadratic expansion of the log-likelihood ratio in the local parameter $\delta'_n - \delta_n$ is equal to the log-likelihood ratio in the Gaussian shift model $\{N(I_F^{-1}\delta, I_F^{-1}) : \delta \in \mathbb{R}^k\}$. This can be shown to imply that the sequence of localized experiments $\{\mathbb{P}^{(n)}_{\theta_n+\delta} : \delta \in \mathbb{R}^k\}$ converges, in an appropriate sense, to the Gaussian shift experiment. This in turn implies that asymptotic analysis in the original experiments can be based on properties of the limiting Gaussian shift model. It also implies that the sequences $\mathbb{P}^{(n)}_{\theta'_n}$ and $\mathbb{P}^{(n)}_{\theta'_n}$ are contiguous (see, e.g., Le Cam and Yang, 1990). As a result, the $o_{\mathrm{P}}(1)$-terms in the above definition converge in probability to zero both under $\mathbb{P}^{(n)}_{\theta_n}$ and $\mathbb{P}^{(n)}_{\theta'_n}$. Moreover, the central sequence is the equivalent of the score function (the derivative of the log-likelihood with respect to the parameter) in likelihood analysis in the Cramér sense. $\square$

In order to illustrate our results, we consider the well-known example of testing for residual autocorrelation in ARMA models. For expository simplicity, we restrict attention to the AR(1) model, whereas the widely applied residual serial autocorrelation type test of, e.g., Ljung and Box (1978) is examined in Section 4 for general ARMA models. The final result is, of course, well-known and can be found in, e.g., Brockwell and Davis (1991). However, our derivation is novel and easily extended to many other models and statistics as shown in Section 4.

Example 2.1 Let the time-series (Y_t) follow a stationary and invertible AR(1) model, i.e.,

$$Y_t = \theta Y_{t-1} + \varepsilon_t,$$

where $\theta \in (-1, 1)$ and (ε_t) is a sequence of i.i.d. random variables from a distribution with density f with expectation zero and finite variance σ_ε^2. Kreiss's (1987b) Theorem 3.1 shows that the ARMA model satisfies the LAN condition if the innovation

density f is absolutely continuous (with respect to Lebesgue measure) with finite Fisher information for location, i.e., $I_l := \int (f'/f)^2 f < \infty$. Some weak conditions on the starting values are needed as well, but that need not concern us here. For the AR(1) model the central sequence is given by

$$\Delta^{(n)}(\theta) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} -\frac{f'}{f}(\varepsilon_t) Y_{t-1},$$

with Fisher information

$$I_F = I_l \text{Var}\{Y_{t-1}\} = I_l \frac{\sigma_\varepsilon^2}{1 - \theta^2}.$$

For notational convenience we consider the stationary solution to the AR(1) equation. The repercussions of this for the LAN condition are detailed in Koul and Schick (1997). □

The interest of the present paper lies in the asymptotic behavior of (test) statistics applied to residuals calculated using a given estimator $\hat{\theta}_n$ for the parameter θ. We define the localized version of this estimator as $\hat{\delta}_n = \sqrt{n}(\hat{\theta}_n - \theta_0)$. We are interested in a test statistic that depends on the unknown parameter θ, say $T_n(\theta)$. In the above example $T_n(\theta)$ could be the l-th order autocorrelation of residuals $\varepsilon_t(\theta) = Y_t - \theta Y_{t-1}$. We assume that we know the asymptotic behavior of the test statistic $T_n(\theta)$, under $\mathbb{P}_\theta^{(n)}$, for all $\theta \in \Theta$. Our goal is to derive the limiting distribution of the statistic when θ_n is replaced by the estimator $\hat{\theta}_n$, i.e., the limiting distribution of $T_n(\hat{\theta}_n)$ under $\mathbb{P}_\theta^{(n)}$. In order to achieve this goal, we impose a second condition.

Condition (AN): Consider a sequence θ_n local to θ_0. The test statistic $T_n(\theta_n)$, the central sequence $\Delta^{(n)}(\theta_n)$, and the estimation error $\sqrt{n}(\hat{\theta}_n - \theta_n) = \hat{\delta}_n - \delta_n$ are jointly asymptotically normally distributed, under $\mathbb{P}_{\theta_n}^{(n)}$, as $n \to \infty$ and as $\delta_n \to \delta$, more precisely,

$$\begin{bmatrix} T_n(\theta_n) \\ \Lambda^{(n)}(\theta_n|\theta_0) \\ \hat{\delta}_n - \delta_n \end{bmatrix} \xrightarrow{\mathcal{L}} \begin{bmatrix} T \\ \frac{1}{2}\delta^T I_F \delta + \delta^T \Delta \\ Z \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \frac{1}{2}\delta^T I_F \delta \\ 0 \end{bmatrix} ; \begin{bmatrix} \tau^2 & c^T \delta & \alpha^T \\ \delta^T c & \delta^T I_F \delta & \delta^T \\ \alpha & \delta & \Gamma \end{bmatrix} \right).$$

□

In the above condition, the distribution of Z does not depend on the sequence (δ_n). This is to say that the estimator being used is regular in the sense of Bickel et al. (1993), page 18. This regularity also implies that the asymptotic covariance between the estimator and the central sequence is the $k \times k$ identity matrix I_k. Moreover, we impose that the asymptotic mean of the test statistic $T_n(\theta_n)$ does not depend on δ, and, hence, is put to zero. In this we follow Pierce (1982), but we do

not need his differentiability condition (1.2). Pierce's (1982) differentiability condition essentially states $c = \mathrm{Cov}\{T, \Delta\} = \lim_{n\to\infty} \mathrm{E}\{\partial T_n(\theta)/\partial\theta\}$. Moreover, Pierce (1982) restricts attention to the case of efficient estimators $\hat{\theta}_n$, which implies $\alpha = \Gamma c$ and $\Gamma = I_F^{-1}$. We discuss the relation of our approach with previous results in the literature in more detail in Section 2.3.

Example 2.2 In our AR(1) example, we may estimate θ using the standard least squares estimator $\hat{\theta}_n$. It is well-known that this estimator satisfies the asymptotic linear representation

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \frac{1-\theta^2}{\sigma_\varepsilon^2}\frac{1}{\sqrt{n}}\sum_{t=1}^n \varepsilon_t Y_{t-1} + o_\mathrm{P}(1) \xrightarrow{\mathcal{L}} N(0, 1-\theta^2),$$

as $n \to \infty$, under the imposed conditions on the AR(1) model, i.e., $\Gamma = 1 - \theta^2$.

In this example, we are interested in testing for serial correlation in the residuals of the AR(1) model. Based on the true innovations $\varepsilon_t(\theta)$, the standard l-th order autocorrelation test statistic satisfies

$$
\begin{aligned}
T_n(\theta) &= \hat{\rho}_n(\theta; l) \\
&:= \sqrt{n}\frac{(n-l+1)^{-1}\sum_{t=l+1}^n \varepsilon_t(\theta)\varepsilon_{t-l}(\theta) - \left(n^{-1}\sum_{t=1}^n \varepsilon_t(\theta)\right)^2}{n^{-1}\sum_{t=1}^n \varepsilon_t(\theta)^2 - \left(n^{-1}\sum_{t=1}^n \varepsilon_t(\theta)\right)^2} \\
&= \frac{1}{\sqrt{n}}\sum_{t=l+1}^n \frac{\varepsilon_t(\theta)\varepsilon_{t-l}(\theta)}{\sigma_\varepsilon^2} + o_\mathrm{P}(1).
\end{aligned}
$$

In order to verify Condition (AN), the following moment results for the stationary distribution of the AR(1) process are needed:

$$
\begin{aligned}
\mathrm{E}\frac{-f'(\varepsilon_t)}{f(\varepsilon_t)}Y_{t-1}\frac{\varepsilon_t\varepsilon_{t-l}}{\sigma_\varepsilon^2} &= \mathrm{E}\frac{Y_{t-1}\varepsilon_{t-l}}{\sigma_\varepsilon^2} = \theta^{l-1}, \\
\mathrm{E}\frac{1-\theta^2}{\sigma_\varepsilon^2}\varepsilon_t Y_{t-1}\frac{\varepsilon_t\varepsilon_{t-l}}{\sigma_\varepsilon^2} &= (1-\theta^2)\mathrm{E}\frac{Y_{t-1}\varepsilon_{t-l}}{\sigma_\varepsilon^2} = (1-\theta^2)\theta^{l-1}, \\
\mathrm{E}\frac{1-\theta^2}{\sigma_\varepsilon^2}\varepsilon_t Y_{t-1}\frac{-f'(\varepsilon_t)}{f(\varepsilon_t)}Y_{t-1} &= \frac{1-\theta^2}{\sigma_\varepsilon^2}\mathrm{E}Y_{t-1}^2 = 1.
\end{aligned}
$$

These moments are easily derived upon noting $\mathrm{E}[-f'(\varepsilon_t)/f(\varepsilon_t)]\varepsilon_t = -\int xf'(x)\mathrm{d}x = \int f(x)\mathrm{d}x = 1$, using the finiteness of the Fisher information for location. A standard application of the martingale central limit theorem now shows that Condition (AN) is satisfied with

$$
\begin{aligned}
\begin{bmatrix} T \\ \Delta \\ Z \end{bmatrix} &\sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}; \begin{bmatrix} \tau^2 & c^T & \alpha^T \\ c & I_F & 1 \\ \alpha & 1 & \Gamma \end{bmatrix}\right) \\
&= N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}; \begin{bmatrix} 1 & \theta^{l-1} & \theta^{l-1}(1-\theta^2) \\ \theta^{l-1} & I_l\sigma_\varepsilon^2/(1-\theta^2) & 1 \\ \theta^{l-1}(1-\theta^2) & 1 & 1-\theta^2 \end{bmatrix}\right),
\end{aligned}
$$

Observe that none of the covariance terms in this limiting distribution depends on the actual innovation density f. $\qquad\square$

2.2   Statement of main result

We may now state the main result of the paper in an informal way. The statement will be made precise in the next section, that also presents a formal proof. For a better understanding of the result, we provide here an intuitive "proof". Note that we study the behavior of the residual statistic $T_n(\hat{\theta}_n)$ under local alternatives θ_n of the parameter value θ_0.

Theorem 2.1 *Under the Conditions (ULAN) and (AN) and in a way that will be made precise in the next section, we have for the residual statistic $T_n(\hat{\theta}_n)$, under $\mathbf{P}_{\theta_n}^{(n)}$, approximately*

$$
\begin{aligned}
T_n(\hat{\theta}_n) \;&\sim\; N\left(0, \tau^2 + (\alpha - \Gamma c)^T \Gamma^{-1}(\alpha - \Gamma c) - \alpha^T \Gamma^{-1}\alpha\right) \qquad (2.1)\\
&=\; N\left(0, \tau^2 + c^T \Gamma c - 2\alpha^T c\right).
\end{aligned}
$$

Proof (intuition): Introduce the distribution

$$
\begin{bmatrix} T \\ \Delta \\ Z \end{bmatrix} \sim N\left(0, \begin{bmatrix} \tau^2 & c^T & \alpha^T \\ c & I_F & I_k \\ \alpha & I_k & \Gamma \end{bmatrix}\right). \qquad (2.2)
$$

where I_k denotes the $k \times k$ identity matrix. From Condition (AN), we have for all $\delta \in \mathbb{R}^k$, under $\mathbf{P}_{\theta_n + \delta/\sqrt{n}}^{(n)}$ and as $n \to \infty$,

$$
\begin{bmatrix} T_n(\theta_n + \delta/\sqrt{n}) \\ \Lambda^{(n)}(\theta_n | \theta_n + \delta/\sqrt{n}) \\ \sqrt{n}(\hat{\theta}_n - \theta_n - \delta/\sqrt{n}) \end{bmatrix} \xrightarrow{\mathcal{L}} \begin{bmatrix} T \\ -\frac{1}{2}\delta^T I_F \delta - \delta^T \Delta \\ Z \end{bmatrix},
$$

while, as a consequence of Le Cam's third lemma (see, e.g., Le Cam and Yang, 1990, Proposition 3.1.1), the same vector converges under $\mathbf{P}_{\theta_n}^{(n)}$ in distribution to

$$
\begin{bmatrix} T - c^T \delta \\ +\frac{1}{2}\delta^T I_F \delta - \delta^T \Delta \\ Z - \delta \end{bmatrix}.
$$

The quantity of interest now can be written as, for $t \in \mathbb{R}$,

$$
\begin{aligned}
&\mathbf{P}_{\theta_n}^{(n)}\left\{T_n(\hat{\theta}_n) \le t\right\} \\
&= \int_{\delta \in \mathbf{R}^k} \mathbf{P}_{\theta_n}^{(n)}\left\{T_n(\hat{\theta}_n) \le t | \hat{\theta}_n = \theta_n + \delta/\sqrt{n}\right\} d\mathbf{P}_{\theta_n}^{(n)}\left\{\sqrt{n}(\hat{\theta}_n - \theta_n) \le \delta\right\}
\end{aligned}
$$

$$
= \int_{\delta \in \mathbf{R}^k} \mathbb{P}_{\theta_n}^{(n)} \left\{ T_n(\theta_n + \delta/\sqrt{n}) \le t | \sqrt{n}(\hat{\theta}_n - \theta_n) = \delta \right\} \mathrm{d}\mathbb{P}_{\theta_n}^{(n)} \left\{ \sqrt{n}(\hat{\theta}_n - \theta_n) \le \delta \right\}
$$

$$
\rightarrow \int_{\delta \in \mathbf{R}^k} \mathbb{P} \left\{ T - c^T \delta \le t | Z = \delta \right\} \mathrm{d}\mathbb{P} \left\{ Z \le \delta \right\}
$$

$$
= \int_{\delta \in \mathbf{R}^k} \Phi \left(\frac{t + (c - \Gamma^{-1}\alpha)^T \delta}{\sqrt{\tau^2 - \alpha^T \Gamma^{-1}\alpha}} \right) \mathrm{d}\mathbb{P} \left\{ Z \le \delta \right\},
$$

where Φ denotes the cumulative distribution function of the standard normal distribution and we used the result that, conditionally on $Z = z$, T has a $N(\alpha^T \Gamma^{-1} z, \tau^2 - \alpha^T \Gamma^{-1}\alpha)$ distribution. Observe that, if we introduce the distribution

$$
\begin{bmatrix} X \\ Z \end{bmatrix} \sim N \left(0, \begin{bmatrix} \tau^2 - \alpha^T \Gamma^{-1}\alpha + (\alpha - \Gamma c)^T \Gamma^{-1}(\alpha - \Gamma c) & (\alpha - \Gamma c)^T \\ (\alpha - \Gamma c) & \Gamma \end{bmatrix} \right),
$$

the distribution of X conditionally on $Z = \delta$ is $N(-(c - \Gamma^{-1}\alpha)^T \delta, \tau^2 - \alpha^T \Gamma^{-1}\alpha)$. Consequently, the limit of $\mathbb{P}_{\theta_n}^{(n)} \left\{ T_n(\hat{\theta}_n) \le t \right\}$ can be written as

$$
\int_{\delta \in \mathbf{R}^k} \mathbb{P} \left\{ X \le t | Z = \delta \right\} \mathrm{d}\mathbb{P} \left\{ Z \le \delta \right\} = \mathbb{P} \left\{ X \le t \right\},
$$

from which (2.1) follows. □

REMARK 2.2 In the above derivation, the convergence of the conditional distribution $\mathbb{P}_{\theta_n}^{(n)} \left\{ T_n(\theta_n + \delta/\sqrt{n}) \le t | \sqrt{n}(\hat{\theta}_n - \theta_n) = \delta \right\}$ to the limit $\mathbb{P} \left\{ T - c^T \delta \le t | Z = \delta \right\}$ is the most delicate part, since the convergence takes place in the conditioning event as well. A formalization of such a convergence would require conditions under which a conditional probability or expectation is continuous with respect to the conditioning event. This question has been studied in the literature, by introducing various topologies on the space of conditioning σ-fields. A good reference is the paper by Cotter (1986) that compares some topologies. From our point of interest, Cotter (1986) essentially shows that the required continuity property only holds for discrete probability distributions. Indeed, we solve the problem by discretizing the estimator $\hat{\theta}_n$ appropriately. See Section 3 for details. □

If we think of the canonical examples given in the introduction, $T_n(\theta)$ represents a test statistic for distributional or dynamic properties of some innovations in the model, while $T_n(\hat{\theta})$ denotes the same statistic applied to estimated residuals in the model. Theorem 2.1 shows that replacing innovations by residuals may leave the asymptotic variance of the test-statistic unchanged, increase it, or decrease it, depending on the value of $(\alpha - \Gamma c)^T \Gamma^{-1}(\alpha - \Gamma c)$ as compared to $\alpha^T \Gamma^{-1}\alpha$. Several special cases may occur.

First, if $c = 0$, the residual-based statistic has the same asymptotic variance as the statistic based on the true innovations. In particular, no adaptation is necessary in critical values in order to guarantee the appropriate asymptotic size of the test

when applied to estimated residuals. However, the power of the test may change which is detailed in Section 2.5. Recall that under $c = 0$ the test statistic and the central sequence are asymptotically independent. As a result, the distribution of the test statistic is invariant to local changes in the parameter θ. In particular, the asymptotic distribution of $T_n(\theta_0)$ is the same under all probability distributions $\mathbf{P}_{\theta_n}^{(n)}$, whatever the local parameter sequence θ_n. As estimated parameter values $\hat{\theta}$ also differ from θ_0 in the order of magnitude of $\sqrt{n}$, this property consequently carries over to the residual-based statistic. As we will see, this situation occurs, for example, when applying the McLeod and Li (1983) test for correlation in squared residuals from least-squares estimation of ARMA or regression models (Example 4.2) or when estimating a general scale model on such residuals (Example 4.4). Also, rank-based change point tests in GARCH models turn out to fall under this scheme (Example 4.9).

A second special case occurs if $\alpha = \Gamma c$. For instance, if the estimator used is efficient we have $\Gamma = I_f^{-1}$, $\alpha = I_F^{-1}c$, and, consequently, $\alpha = \Gamma c$. However, we will see below that this situation also occurs, for instance, when applying the Ljung and Box (1978) test to least-squares residuals in an ARMA or regression model, also when the actual underlying distribution of the innovations is not Gaussian and the least-squares estimator consequently is not parametrically efficient. In case $\alpha = \Gamma c$, the limiting variance of the residual-based statistic is smaller than the limiting variance of the statistic applied to the true innovations, with strict inequality if $\alpha \neq 0$. Pierce (1982) restricts attention to this efficient initial estimator case and, imposing a differentiability condition on $T_n(\theta)$ finds the same reduction in the limiting variance. This occurs, for instance, in our discrete choice example (Example 4.8).

Finally, it might be that $\alpha = 0$. In that case the limiting variance of the residual statistic becomes $\tau^2 + c^T \Gamma c \geq \tau^2$. This is the case where the test statistic $T_n(\theta)$ is asymptotically independent from the estimator $\hat{\theta}_n$ and a test based on estimated residuals always has a larger asymptotic variance than the same test applied to the actual innovations, unless $c = 0$. The asymptotic independence of the statistic $T_n(\theta)$ and the estimator $\hat{\theta}_n$ implies that the residual-based statistic $T_n(\hat{\theta}_n)$ essentially behaves as a mixture over various values of θ. Such a mixture distribution clearly has a larger variance than the distribution of $T_n(\theta)$ with θ fixed. This situation arises, among others, in linear regression models where the regressors have zero expectation (see Examples 4.6 and 4.7).

Example 2.3 In our AR(1) running example, we can immediately apply the result (2.1). From the calculations above, we find that the asymptotic variance of the l-th order autocorrelation of the residuals equals

$$
\tau^2 + \frac{(\alpha - \Gamma c)^2}{\Gamma} - \frac{\alpha^2}{\Gamma} = 1 + 0 - \frac{[\theta^{l-1}(1-\theta^2)]^2}{1-\theta^2}
$$
$$
= 1 - \theta^{2(l-1)}(1-\theta^2).
$$

This result is, of course, well-known and can be found, e.g., in Example 9.4.1 in

Brockwell and Davis (1991). Observe that this result does not depend on the actual underlying distribution of the innovations f. □

Theorem 2.1 has been stated for univariate statistics $T_n(\theta)$, but can easily be extended to the multivariate case using the Cramér-Wold device. For multivariate T_n, τ^2, c, and α in Condition (AN) become matrices. By taking arbitrary linear combinations of the components of T_n and applying the univariate version of Theorem 2.1, we find that the same limiting distribution 2.1 holds with τ^2 replaced by the limiting variance matrix of T_n, c the limiting covariance matrix between the statistic and the central sequence, and α the limiting covariance matrix between the statistic and the estimator used. This result can be applied when deriving, for instance, the limiting distribution of a two-stage estimator, i.e., where a model is estimated on residuals from a first-stage estimation, as shown in Example 4.4.

2.3   Relation with traditional approach

As we propose an alternative way to study the asymptotic behavior of residual-based statistics, it is important to relate our approach to the traditional one which is based on smoothness arguments of the statistics under consideration. More precisely, following Pierce (1982) and Randles (1982), several authors impose essentially the condition

$$T_n(\theta_0 + \delta_n/\sqrt{n}) = T_n(\theta_0) + b^T\delta_n + o_{\mathbf{P}}(1), \tag{2.3}$$

under $\mathbf{P}_{\theta_0}^{(n)}$, for bounded sequences δ_n as $n \to \infty$. Condition (2.3) is sometimes reinforced to hold for random sequences $\hat{\delta}_n = O_{\mathbf{P}}(1)$. However, this reinforcement is not required for our purposes if one resorts to discretized estimators as in Section 3.

Under (2.3), Condition (AN) implies $c = -b$. This can be seen as follows. Under $\mathbf{P}_{\theta_0}^{(n)}$ the limiting distribution of $T_n(\theta_n)$ follows from (2.3) as $N(b^T\delta, \tau^2)$. However, as in the proof of Theorem 2.1, it also follows from Le Cam's third lemma as $N(-c^T\delta, \tau^2)$, hence the result. Pierce's (1982) formula (1.3), which is derived for the case the estimator $\hat{\theta}_n$ is efficient, now follows from (2.1) since $\alpha = \Gamma c$ and $\alpha^T\Gamma^{-1}\alpha = c^T\Gamma c = b^T\Gamma b$. In this respect, our alternative approach replaces the differentiability condition (2.3) by a distributional invariance condition (AN), i.e., the assumption that the limiting distribution of $T_n(\theta_n)$ under $\mathbf{P}_{\theta_n}^{(n)}$ does not depend on the sequence θ_n other than through its limit θ_0. This latter assumption is often trivially satisfied for residual-based statistics as the distribution of the residuals calculated using θ_n under $\mathbf{P}_{\theta_n}^{(n)}$ is identical to the distribution of residuals calculated using θ_0 under $\mathbf{P}_{\theta_0}^{(n)}$. See Section 4 for the applications.

The special case $c = 0$ which is discussed in the previous section, squares in the traditional approach with the situation that the expectation of the derivative of the statistic with respect to the parameter θ equals zero, i.e., $b = 0$ in (2.3). It is well-known that in that situation the traditional approach shows that no adaptation is

needed in the limiting distribution of the statistic $T_n(\hat{\theta}_n)$, as follows from an application of Theorem 2.1. Our contribution in this special situation is that we replace the zero derivative condition by a orthogonality condition of the statistic and the central sequence.

Summarizing, our approach essentially replaces the smoothness condition (2.3) by a distributional invariance condition. However, the ideas underlying Theorem 2.1 can also be applied in case this distributional invariance does not hold, i.e., in case the derivative b in (2.3) does not equal $-c$. The simplest way to proceed in such case is to introduce an auxiliary statistic $\tilde{T}_n(\theta)$ as

$$\tilde{T}_n(\theta) := T_n(\theta) - (b+c)^T \sqrt{n}(\theta - \theta_0).$$

This auxiliary statistic does satisfy the asymptotic distributional invariance property, as the limiting distribution of $\tilde{T}_n(\theta_0 + \delta/\sqrt{n}) = T_n(\theta_0) - c^T\delta + o_P(1)$, under $\mathbf{P}_{\theta_n}^{(n)}$, is $N(0, \tau^2)$ again from Le Cam's third lemma. Consequently, Condition (AN) is satisfied and Theorem 2.1 can be applied to the statistic $\tilde{T}_n(\theta)$. The limiting behavior of $T_n(\hat{\theta}_n)$ now follows from that of $\tilde{T}_n(\hat{\theta}_n)$ jointly with $\sqrt{n}(\hat{\theta}_n - \theta_0)$. This latter joint limiting behavior is that of $[X, Z^T]^T$ in the proof of Theorem 2.1. Elementary algebra finds the resulting limiting distribution of $T_n(\hat{\theta}_n)$ as $N(0, \tau^2 + 2b^T\alpha + b^T\Gamma b)$. Observe that this latter expression no longer involves the joint behavior of the statistic or estimator with the central sequence.

2.4   Application to semi- and nonparametric models

As we have seen, under the differentiability condition (2.3), the limiting distribution of the residual-based statistic no longer involves properties of the central sequence. Consequently, the underlying parametric model does not play a role in that case. In general one may consider situations where no parametric model has been specified. As we will see below, in such situations Theorem 2.1 may still be applicable by constructing an appropriate auxiliary parametric model. This is much in the same spirit as the construction of least-favorable parametric submodels in the analysis of semiparametric efficiency. In this vein, note that the running AR(1) example is in fact semiparametric as the innovation density f remains unspecified. Clearly, the limiting distribution of residual-based statistics may very well depend on the underlying value of such innovation density f. The fact that it does not in the AR(1) residual autocorrelation example is an exception, rather than a rule.

Although an in depth analysis of the applicability of our approach to residual-based statistics in semi- and nonparametric models is beyond the scope of the present paper, we will provide one example where the ideas can be applied within a standard GMM setting. Consider a setting with i.i.d. observations $Z_1, \ldots, Z_n$ where we have a Euclidean parameter of interest that is defined by the moment condition $\mathrm{E}m(\theta; Z_i) = 0$. We assume that this condition is exactly identifying, i.e., also the sample equivalent of this moment condition (2.4) below has a unique solution. As

an example we consider the regression model with conditional moment restriction $\mathrm{E}\{Y_i - X_i^T\theta | X_i\} = 0$ with $Z_i = (Y_i, X_i)$ and use $m(\theta; Z_i) = (Y_i - X_i^T\theta)X$. We assume that our test statistic of interest is based on an alternative one-dimensional moment condition $\mathrm{E}h(\theta; Z_i) = 0$. where, to facilitate the analysis, $\mathrm{E}h(\theta; Z_i)m(\theta; Z_i) = 0$. In the regression model this condition may be about the unconditional second moment of the innovations, i.e., $h(\theta; Z_i) = (Y_i - X_i^T\theta)^2 - \sigma^2$ where $\sigma^2 := \mathrm{E}(Y_i - X_i^T\theta)^2$. The imposed orthogonality between $m(\theta; Z_i)$ and $h(\theta; Z_i)$ in this regression setting amounts to imposing zero conditional skewness for the innovations, i.e., $\mathrm{E}\{(Y_i - X_i^T\theta)^3 | X_i\} = 0$. Summarizing, our estimator of interest is $\hat{\theta}_n$ solving

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_i(\hat{\theta}_n; Z_i) = 0, \tag{2.4}$$

and our statistic of interest is given by

$$T_n(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} h_i(\theta; Z_i). \tag{2.5}$$

In this situation, Condition (AN) becomes, under θ_n,

$$\left[\begin{array}{c} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} h(\theta_n; Z_i) \\ \sqrt{n}(\hat{\theta}_n - \theta_n) \end{array}\right] \xrightarrow{\mathcal{L}} N\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right]; \left[\begin{array}{cc} \tau^2 & 0^T \\ 0 & \Gamma \end{array}\right]\right), \tag{2.6}$$

with $\tau^2 = \mathrm{E}h(\theta_0; Z_i)^2$. The delicate part in checking (2.6) is the limiting mean of zero for the statistic of interest $T_n(\theta_n)$. However, the point is again that this is often easily verified without resorting to smoothness arguments, but by using a distribution invariance argument. For instance, in the regression model example we are considering, one readily obtains $T_n(\theta_n) = n^{-1/2}\sum_{i=1}^{n}(Y_i - X_i^T\theta_n) - \sigma^2$ which under θ_n has exactly the same distribution as $T_n(\theta_0)$ under θ_0. This distributional invariance carries over to the limit which verifies (2.6) and thus Condition (AN). Applying Theorem 2.1 with $\alpha = c = 0$ leads to the conclusion that the residual-based statistic $T_n(\hat{\theta}_n)$ has the same limiting distribution as the statistic $T_n(\theta_0)$.

Theorem 2.1 assumes the existence of a central sequence. The actual form is irrelevant in the present situation, since the limiting distribution of the residual-based statistic does not depend on it. However, we should verify that at least one appropriate central sequence can be defined. We do so by introducing a parametric model that is defined locally around a fixed value of $\theta = \theta_0$. More precisely, we define the central sequence $\Delta^{(n)}(\theta_0) = \Gamma^{-1}\sqrt{n}(\hat{\theta}_n - \theta_0)$ and the Fisher information $I_F = \Gamma^{-1}$. As before, Γ denotes the limiting variance of the estimator $\hat{\theta}_n$ which, in the present GMM case under standard regularity conditions equals

$$\Gamma = \left[\mathrm{E}\frac{\partial}{\partial\theta}m(\theta)\Big|_{\theta=\theta_0}\right]^{-1}\mathrm{E}m(\theta_0)m(\theta_0)^T.$$

The auxiliary parametric model is now defined via

$$\frac{\mathrm{d}\mathbb{P}^{(n)}_{\theta_0+\delta/\sqrt{n}}}{\mathrm{d}\mathbb{P}^{(n)}_{\theta_0}} = \frac{\exp\left(\delta^T \Delta^{(n)}(\theta_0) - \frac{1}{2}\delta^T I_F \delta\right)}{\mathrm{E}\exp\left(\delta^T \Delta^{(n)}(\theta_0) - \frac{1}{2}\delta^T I_F \delta\right)}, \tag{2.7}$$

In other words, we construct a parametric model such that the given estimator $\hat{\theta}_n$ is efficient in this model. It is easy to see that the model satisfies the Condition (ULAN) as long as $\exp\left(\delta^T \Gamma^{-1}\sqrt{n}(\hat{\theta}_n - \theta_0)\right)$ is uniformly integrable.

In the analysis above we imposed the orthogonality condition $\mathrm{E}h(\theta; Z_i)m(\theta; Z_i) = 0$. Without this condition, Condition (AN) may be more difficult to verify or may even not hold as the limiting mean of $T_n(\theta_n)$ is not necessarily constant zero in δ. If the appropriate limiting behavior would be known, one may proceed as in Section 2.3 by introducing the auxiliary statistic $\tilde{T}_n$ that does satisfy the limiting invariance property.

At this point it may be important to stress that we do not exclude simulation based estimators as SMM, EMM, or indirect inference. These estimators often do not provide efficient inference in a given model, but our results are not restricted to efficient first-stage estimators (although the resulting analysis for the residual-based statistic does simplify in the efficient case). It is important that the joint limiting behavior of the first-stage estimator and the innovation-based statistic is known.

2.5 Power considerations

A question that arises naturally at this point is the effect on the power of tests that are applied to residuals instead of actual innovations. First of all, note that the limiting distribution of the residual test statistic in (2.1) does not depend on the local parameter sequence θ_n. This implies that the statistic's distribution is invariant with respect to local changes in the underlying parameter θ. The test statistic T_n, consequently, has no local power against alternatives of this type.

Consider, however, the case where there is an additional parameter ψ in the model and we are interested in the (local) power of the residual-based statistic $T_n(\hat{\theta}_n)$ with respect to this parameter. The model now consists of a set of probability measures $\{\mathbb{P}^{(n)}_{\theta,\psi} : \theta \in \Theta, \ \psi \in \Psi\}$. For ease of notation we assume that the original model is obtained by setting $\psi = 0$, i.e., $\mathbb{P}^{(n)}_{\theta,0} = \mathbb{P}^{(n)}_\theta$. As before, fix $\theta_0 \in \Theta$ and consider the local parametrization $(\theta_n, \psi_n) = (\theta_0 + \delta/\sqrt{n}, 0 + \eta/\sqrt{n})$. Introduce the log-likelihood

$$\tilde{\Lambda}(\psi_n|0) = \log \frac{\mathrm{d}\mathbb{P}^{(n)}_{\theta_0,\psi_n}}{\mathrm{d}\mathbb{P}^{(n)}_{\theta_0,0}},$$

with respect to the parameter ψ. We are interested in the behavior of our test-statistic $T_n(\hat{\theta}_n)$ under $\mathbb{P}^{(n)}_{\theta_0,\psi_n}$. Assume that Condition (ULAN) is satisfied jointly in θ and ψ.

Moreover, assume the equivalent of Condition (AN) under $\psi = 0$, i.e., under $\mathbf{P}_{\theta_n,0}^{(n)}$ and as $n \to \infty$,

$$
\begin{bmatrix}
T_n(\theta_n) \\
\Lambda^{(n)}(\theta_n|\theta_0) \\
\hat{\delta}_n - \delta_n \\
\tilde{\Lambda}^{(n)}(\psi_n|0)
\end{bmatrix}
\xrightarrow{\mathcal{L}}
\begin{bmatrix}
T \\
\frac{1}{2}\delta^T I_F \delta + \delta^T \Delta \\
Z \\
-\frac{1}{2}\eta^T I_P \eta + \delta^T I_{FP}\eta + \eta^T \tilde{\Delta}
\end{bmatrix}
\tag{2.8}
$$

$$
\sim \quad N\left(
\begin{bmatrix}
0 \\
\frac{1}{2}\delta^T I_F \delta \\
0 \\
-\frac{1}{2}\eta^T I_P \eta + \delta^T I_{FP}\eta
\end{bmatrix}
;
\begin{bmatrix}
\tau^2 & c^T\delta & \alpha^T & d^T\eta \\
\delta^T c & \delta^T I_F \delta & \delta^T & \delta^T I_{FP}\eta \\
\alpha & \delta & \Gamma & B^T\eta \\
\eta^T d & \eta^T I_{FP}\delta & \eta^T B & \eta^T I_P \eta
\end{bmatrix}
\right).
$$

Here I_P denotes the Fisher information for the parameter ψ, while I_{FP} denotes the cross Fisher information between θ and ψ. The matrix B measures the covariance between the log-likelihood ratio with respect to ψ and the estimator for θ_n. Consequently, this matrix measures the bias in $\hat{\theta}_n$ that occurs due to possible local changes in ψ. The special case $B = 0$ refers to the situation where $\hat{\theta}_n$ is insensitive to local changes in ψ. This occurs, e.g., if $\hat{\theta}_n$ is an efficient estimator for θ in a model where ψ is considered a nuisance parameter. The asymptotic mean of $\tilde{\Lambda}^{(n)}(\psi_n|0)$ in (2.8) is a direct consequence of the fact that the limiting distribution is studied under $(\theta, \psi) = (\theta_n, 0)$. The derivations leading to Theorem 2.1 remain valid and can be carried out while taking into account the joint behavior of $T_n(\hat{\theta}_n)$ and $\tilde{\Lambda}^{(n)}(\psi_n|0)$. Under $\mathbf{P}_{\theta_n,0}^{(n)}$, one easily verifies for $(T_n(\hat{\theta}_n), \tilde{\Lambda}^{(n)}(\psi_n|0))$ the following limiting distribution:

$$
N\left(
\begin{bmatrix}
0 \\
-\frac{1}{2}\eta^T I_P \eta + \delta^T I_{FP}\eta
\end{bmatrix}
,
\begin{bmatrix}
\tau^2 + (\alpha - \Gamma c)^T \Gamma^{-1}(\alpha - \Gamma c) - \alpha^T \Gamma^{-1}\alpha & \eta^T(d - Bc) \\
(d - Bc)^T \eta & \eta^T I_P \eta
\end{bmatrix}
\right).
$$

Applying Le Cam's third lemma once more, we see that the shift in the innovation-based statistic $T_n(\theta)$ due to local changes in ψ is given by $d^T\eta$, while the same local change in ψ induces a shift of size $(d - Bc)^T\eta$ in the residual-based statistic $T_n(\hat{\theta})$. In the special case that $B = 0$, we thus find that the power against local changes in ψ in the residual-based statistic decreases, remains unchanged, or increases as the limiting variance under $\psi = 0$ increases, remains unchanged, or decreases, respectively. It may thus very well be the case that residual-based statistics have larger power against certain local alternatives than the same statistic applied to actual innovations.

Alternatively, the results in this section can be interpreted in terms of specification testing with locally misspecified alternatives much in the same spirit as Bera and Yoon (1993). That paper derives a correction to standard LM tests which makes them insensitive to local misspecification. Not surprisingly, this correction exactly contains the covariance term B, which is $J_{\psi\phi}$ in their Formula (3.2).

2.6 Non-Gaussian limiting distributions

The results in the present paper have been written with multivariate normally distributed test statistics, central sequences, and estimators in mind. While this covers the vast majority of everyday econometric practice, it is worthwhile mentioning that ideas underlying Theorem 2.1 carry over to non-Gaussian situations. In particular, Le Cam's third lemma is not restricted to Gaussian distributions, but it takes a particularly simple form in that case, see the appendix. As illustrated by the applications to dependence tests, for instance the Ljung-Box and McLeod-Li tests (Examples 4.1 and 4.2), traditional χ^2-tests can be handled by our Theorem 2.1 applied to the autocorrelation function of the (squared) residuals instead of the autocorrelation's quadratic transformation. The same idea of quadratic representation of normal statistics can be applied to most other χ^2-tests.

To handle still different limiting distribution, one would need to specify the joint density of the limiting distribution of the statistic, the log-likelihood ratio, and the estimation error in Condition (AN). This may, however, be very cumbersome. Also note that Condition (AN) is in fact sufficient to derive Theorem 2.1 as the Condition (ULAN) only serves to establish part of Condition (AN). Given this joint limiting density, the proof of Theorem 2.1, or the formal version Theorem 3.2, could be carried out by carefully calculating the appropriate conditional densities, applying Le Cam's third Lemma, and calculating the given integrals. These calculations are likely to be tedious and whether they result in simple expressions as in Theorem 2.1 remains for future research.

3 Main result: Formalization

The problem with studying the asymptotic behavior of $T_n(\hat{\theta}_n)$ is that arbitrary estimators (even if they are regular) $\hat{\theta}_n$ can pick out very special points of the parameter space. Without strong uniformity conditions on the behavior of $T_n(\theta)$ as a function of θ (such as, continuous differentiability in some way), the residual statistic $T_n(\hat{\theta}_n)$ can behave in an erratic way. We solve this problem by discretizing the estimator $\hat{\theta}_n$. This is a well-known trick due to Le Cam, however, usually applied to the construction of optimal tests and estimators in ULAN models. We introduce this approach now and study the behavior of the statistic based on the discretized estimated parameter.

The discretized estimator $\overline{\hat{\theta}}_n$ is obtained by rounding the original estimator $\hat{\theta}_n$ to the nearest midpoint of a regular grid of cubes. To be precise, consider a grid of cubes in $\mathbb{R}^k$ with sides of length $d/\sqrt{n}$. We call d the discretization constant. Then $\overline{\hat{\theta}}_n$ is the estimator obtained by taking the midpoint of the cube to which $\hat{\theta}_n$ belongs. To formalize the above even further, introduce the function $d : \mathbb{R}^k \to \mathbb{Z}^k$ which arithmetically rounds each of the components of the input vector to the nearest integer. Then, we may write, with $\hat{\theta}_n$ our initial non-discretized estimator,

$\overline{\overline{\theta}}_n = d(\sqrt{n}\widehat{\theta}_n)/\sqrt{n}$. Our ultimate interest lies in the asymptotic behavior of $T_n(\overline{\overline{\theta}}_n)$. We first study the behavior of $\overline{\overline{\theta}}_n$ in the following lemma.

Lemma 3.1 *Let the discretization constant $d > 0$ be given. Define the "discretized truth" $\overline{\theta}_n = d(\sqrt{n}\theta_0)/\sqrt{n}$. Then, the localized version $\overline{\overline{\delta}}_n = \sqrt{n}(\overline{\overline{\theta}}_n - \overline{\theta}_n)$ of the discretized estimator $\overline{\overline{\theta}}_n$ is degenerated on $\{dj : j \in \mathbb{Z}^k\}$. Moreover, for $\delta_n \to \delta$ as $n \to \infty$, we have*

$$\mathbb{P}_{\overline{\theta}_n + \delta_n/\sqrt{n}}^{(n)}\left\{\overline{\overline{\delta}}_n = dj\right\} \to \mathbb{P}\left\{N(\delta - dj, \Gamma) \in (-\frac{d}{2}\iota, \frac{d}{2}\iota]\right\}, \tag{3.1}$$

where $\iota = (1, 1, \ldots, 1)^T \in \mathbb{Z}^k$.

PROOF: The fact that $\overline{\overline{\delta}}_n$ is degenerated on $\{dj : j \in \mathbb{Z}^k\}$ follows easily from $\overline{\overline{\delta}}_n = \sqrt{n}(\overline{\overline{\theta}}_n - \overline{\theta}_n) = d(\sqrt{n}\widehat{\theta}_n) - d(\sqrt{n}\theta_0)$. To deduce its limiting distribution, observe the following equalities of events, for fixed $j \in \mathbb{Z}^k$,

$$
\begin{aligned}
\left\{\overline{\overline{\delta}}_n = dj\right\} &= \left\{d(\sqrt{n}\widehat{\theta}_n) - d(\sqrt{n}\theta_0) = dj\right\} \\
&= \left\{d(\sqrt{n}\theta_0 + \widehat{\delta}_n) = d(\sqrt{n}\theta_0) + dj\right\} \\
&= \left\{d(\sqrt{n}\theta_0) + dj - \frac{d}{2}\iota < \sqrt{n}\theta_0 + \widehat{\delta}_n \le d(\sqrt{n}\theta_0) + dj + \frac{d}{2}\iota\right\} \\
&= \left\{dj - \frac{d}{2}\iota < \widehat{\delta}_n + \sqrt{n}\theta_0 - d(\sqrt{n}\theta_0) \le dj + \frac{d}{2}\iota\right\} \\
&= \left\{dj - \frac{d}{2}\iota < \sqrt{n}(\widehat{\theta}_n - \overline{\theta}_n) \le dj + \frac{d}{2}\iota\right\}.
\end{aligned}
$$

From the Conditions (ULAN) and (AN), we find, under $\mathbb{P}_{\overline{\theta}_n + dj/\sqrt{n}}^{(n)}$, as $\delta_n \to \delta$, and as $n \to \infty$,

$$
\begin{bmatrix} \Lambda(\overline{\theta}_n + \delta_n/\sqrt{n}|\overline{\theta}_n + dj/\sqrt{n}) \\ \sqrt{n}(\widehat{\theta}_n - \overline{\theta}_0 - dj/\sqrt{n}) \end{bmatrix} \xrightarrow{\mathcal{L}} \begin{bmatrix} -\frac{1}{2}(\delta - dj)^T I_F(\delta - dj) + (\delta - dj)^T \Delta \\ Z \end{bmatrix},
$$

with $[Z, \Delta^T]^T$ as in (2.2). From Le Cam's third lemma, this implies, under $\mathbb{P}_{\overline{\theta}_n + \delta_n/\sqrt{n}}^{(n)}$ and as $n \to \infty$, $\sqrt{n}(\widehat{\theta}_n - \overline{\theta}_n - dj/\sqrt{n}) \xrightarrow{\mathcal{L}} N(\delta - dj, \Gamma)$. Together with the above result on the event $\left\{\overline{\overline{\delta}}_n = dj\right\}$, the lemma now follows. $\qquad\square$

The above lemma is basic to our formal main result that now can be stated.

Theorem 3.2 *With the notation introduced above and under Conditions (ULAN) and (AN), we have for $\delta_n \to \delta$ and as $n \to \infty$,*

$$\lim_{d \downarrow 0} \lim_{n \to \infty} \mathbb{P}^{(n)}_{\bar{\theta}_n + \delta_n / \sqrt{n}} \left\{ T_n(\bar{\bar{\theta}}_n) \le t \right\} = \mathbb{P}\{X \le t\}, \tag{3.2}$$

where

$$X \sim N\left(0, \tau^2 + (\alpha - \Gamma c)^T \Gamma^{-1} (\alpha - \Gamma c) - \alpha^T \Gamma^{-1} \alpha\right).$$

PROOF: From the proof of Lemma 3.1, we know

$$\left\{ \bar{\bar{\delta}}_n = dj \right\} = \left\{ -\frac{d}{2}\iota < \sqrt{n}(\hat{\theta}_n - \bar{\theta}_n) - dj \le \frac{d}{2}\iota \right\}.$$

Moreover, applying Le Cam's third lemma as in the proof of Lemma 3.1, we find under $\mathbb{P}^{(n)}_{\bar{\theta}_n + \delta_n / \sqrt{n}}$ and as $n \to \infty$,

$$\left[\begin{array}{c} T_n(\bar{\theta}_n + dj/\sqrt{n}) \\ \sqrt{n}(\hat{\theta}_n - \bar{\theta}_n) - dj \end{array} \right] \xrightarrow{\mathcal{L}} N\left(\left[\begin{array}{c} (\delta - dj)^T c \\ \delta - dj \end{array} \right], \left[\begin{array}{cc} \tau^2 & \alpha^T \\ \alpha & \Gamma \end{array} \right] \right).$$

Taking these two results together, we get, for all $j \in \mathbb{Z}^k$ and with the distribution (2.2),

$$\mathbb{P}^{(n)}_{\bar{\theta}_n + \delta_n / \sqrt{n}} \left\{ T_n(\bar{\theta}_n + dj/\sqrt{n}) \le t \text{ and } \bar{\bar{\delta}}_n = dj \right\}$$

$$\to \mathbb{P}\left\{ T + (\delta - dj)^T c \le t \text{ and } -\frac{d}{2}\iota < Z + (\delta - dj) \le \frac{d}{2}\iota \right\}.$$

The number of values that $\bar{\bar{\delta}}_n$ takes in a bounded set, is finite. Consequently, we may write for each $M > 0$,

$$\mathbb{P}^{(n)}_{\bar{\theta}_n + \delta_n / \sqrt{n}} \left\{ T_n(\bar{\bar{\theta}}_n) \le t \text{ and } \left|\bar{\bar{\delta}}_n\right| \le M \right\}$$

$$= \sum_{j \in \mathbb{Z}^k, \ d|j| \le M} \mathbb{P}^{(n)}_{\bar{\theta}_n + \delta_n / \sqrt{n}} \left\{ T_n(\bar{\theta}_n + dj) \le t \text{ and } \bar{\bar{\delta}}_n = dj \right\}$$

$$\to \sum_{j \in \mathbb{Z}^k, \ d|j| \le M} \mathbb{P}\left\{ T + (\delta - dj)^T c \le t \text{ and } -\frac{d}{2}\iota < Z + (\delta - dj) \le \frac{d}{2}\iota \right\},$$

as $n \to \infty$. Since $\limsup_{n \to \infty} \mathbb{P}^{(n)}_{\bar{\theta}_n + \delta_n / \sqrt{n}} \left\{ \left|\bar{\bar{\delta}}_n\right| > M \right\} \to 0$ as $M \to \infty$, we obtain

$$\mathbb{P}^{(n)}_{\bar{\theta}_n + \delta_n / \sqrt{n}} \left\{ T_n(\bar{\bar{\theta}}_n) \le t \right\}$$

$$\to \sum_{j \in \mathbb{Z}^k} \mathbb{P}\left\{ T \le t - (\delta - dj)^T c \text{ and } -\frac{d}{2}\iota < Z + (\delta - dj) \le \frac{d}{2}\iota \right\},$$

as $n \to \infty$. Let φ_{TZ} denote the probability density function of $[T, Z^T]^T$ and φ_Z that of Z. Observe again that, conditionally on $Z = z$, T has a $N(\alpha^T \Gamma^{-1} z, \tau^2 - \alpha^T \Gamma^{-1} \alpha)$ distribution. Consequently,

$$
\sum_{j \in \mathbb{Z}^k} \mathbf{P} \left\{ T \leq t - (\delta - dj)^T c \text{ and } -\frac{d}{2}\iota < Z + (\delta - dj) \leq \frac{d}{2}\iota \right\}
$$

$$
= \sum_{j \in \mathbb{Z}^k} \int_{x=-\infty}^{t-(\delta-dj)^T c} \int_{z=-(\delta-dj)-\frac{d}{2}\iota}^{-(\delta-dj)+\frac{d}{2}\iota} \varphi_{TZ}(x, z) \mathrm{d}x \mathrm{d}z
$$

$$
= \sum_{j \in \mathbb{Z}^k} \int_{z=-(\delta-dj)-\frac{d}{2}\iota}^{-(\delta-dj)+\frac{d}{2}\iota} \Phi\left(\frac{t - (\delta - dj)^T c - \alpha^T \Gamma^{-1} z}{\sqrt{\tau^2 - \alpha^T \Gamma^{-1} \alpha}} \right) \varphi_Z(z) \mathrm{d}z
$$

$$
= \sum_{j \in \mathbb{Z}^k} \int_{z=-(\delta-dj)-\frac{d}{2}\iota}^{-(\delta-dj)+\frac{d}{2}\iota} \Phi\left(\frac{t - (\alpha - \Gamma c)^T \Gamma^{-1} z}{\sqrt{\tau^2 - \alpha^T \Gamma^{-1} \alpha}} \right) \varphi_Z(z) \mathrm{d}z + O(d)
$$

$$
= \int_{z \in \mathbf{R}^k} \Phi\left(\frac{t - (\alpha - \Gamma c)^T \Gamma^{-1} z}{\sqrt{\tau^2 - \alpha^T \Gamma^{-1} \alpha}} \right) \varphi_Z(z) \mathrm{d}z + O(d)
$$

$$
= \int_{z \in \mathbf{R}^k} \mathbf{P}\left\{ X \leq t | Z = z \right\} \varphi_Z(z) \mathrm{d}z + O(d)
$$

$$
ra \quad \mathbf{P}\left\{ X \leq t \right\},
$$

as $d \downarrow 0$, with

$$
\begin{bmatrix} X \\ Z \end{bmatrix} \sim N\left(0, \begin{bmatrix} \tau^2 + (\alpha - \Gamma c)^T \Gamma^{-1}(\alpha - \Gamma c) - \alpha^T \Gamma^{-1}\alpha & \alpha - \Gamma c \\ (\alpha - \Gamma c)^T & \Gamma \end{bmatrix} \right).
$$

This completes the proof. $\qquad \square$

REMARK 3.1 As the informal derivations in Section 3, the above proof is strongly based on a conditioning argument with respect to the value of the estimator $\hat{\theta}_n$, or, more precisely, that of the local estimation error $\hat{\delta}_n$. This leads one to believe that it is meaningfully possible to derive LAN conditions for conditional distributions, where the conditioning event is the value of the estimation error. The authors of the present paper have, however, not seen any results in this direction. $\qquad \square$

Theorem 3.2 utilizes the technique of discretization to avoid initial estimators $\hat{\theta}_n$ that pick out points of the likelihood which do not behave properly. The same technique is applied usually in the construction of efficient estimators in parametric and semiparametric models. In practise, one would, of course, rarely implement it. We take the result (3.2) as an approximate limiting distribution for the residual-based statistic for large number of observations n and small discretization constant d. The accuracy of this approximation for finite samples is case specific. Many of the papers referenced in Section 4 provide simulation studies to assess this accuracy in particular situations.

4 Applications

Our results are applicable to cases where the underlying model satisfies the LAN condition. In the present section we work through several examples, showing the scope of our results. We concentrate on time series models to gain coverage of tests and conciseness of exposition while noting that cross-sectional models can be handled as well (see, e.g., Example 4.8). The first category of statistics addresses residual-based tests for temporal dependence, both linear and non-linear (quadratic). In particular, we consider testing for linear and second order serial correlation in ARMA and regression models in Examples 4.1 and 4.2, respectively. After this, we extend these results to testing for non-linear dependence in the form of autocorrelation in squared residuals and linear dependence tests applied to residuals of an estimated scale model. Example 4.4 considers the case of estimating a GARCH model based on ARMA residuals. For reasons of robustness, applied work often uses rank-based statistics to test for serial correlation in residuals or in their squares. The fifth example focuses on this situation. The second category of examples considers goodness-of-fit tests for evaluating the innovations' distributional assumptions. There is a large class of empirical distribution function (EDF) goodness-of-fit tests (e.g., D'Agostino and Stephens, 1986) recently revisited for residuals of regression and time series models (see, e.g., Andrews, 1997, Koul and Stute, 1999, Koul, 2002, and Horváth et al., 2001). Examples 4.6 and 4.7 revisit the asymptotic distribution of such residual EDF-based tests in location and scale time series, when the null distribution is completely specified and when it contains nuisance parameters. The approach followed in the paper yields asymptotic results that match those of Durbin (1973) and Bai (2003), among others. In Example 4.8 we consider tests for omitted variables in a probit framework. As an aside we introduce a new nonparametric test for this problem based on Kendall's tau. The last category (Example 4.9) considers homogeneity test statistics. The CUSUM test for structural breaks in the innovations' distribution based on the ranks of residuals (compare Sen, 1984, for the linear model) is studied. We present a new result on the asymptotic distribution of these statistics for scale models.

4.1 Tests for temporal dependence

Example 4.1 Ljung-Box in ARMA/regression models with OLS residuals
Consider a location model of the form

$$Y_t = \mu_{t-1}(\theta) + \varepsilon_t, \ t = 1, \ldots, n, \tag{4.1}$$

where $\mu_{t-1}(\theta)$ depends on past values $Y_{t-1}, Y_{t-2}, \ldots$ and ε_t is a sequence of i.i.d. mean zero innovations with finite variance σ_ε^2 and finite Fisher information for location $I_l := \int (f'/f)^2 f < \infty$. As mentioned in the introduction, such a model satisfies

the LAN property under the condition that $\mu_{t-1}(\theta)$ depends smoothly on θ and the process satisfies some regularity conditions (see Drost et al., 1997, for details). In particular, stationary and invertible ARMA models and linear regression models are allowed. In the latter case, $\mu_{t-1}(\theta)$ depends on some observable exogenous variables $X_1, \ldots, X_n$. The central sequence can generally be written as

$$\Delta^{(n)}(\theta) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} -\frac{f'}{f}(\varepsilon_t) \frac{\partial \mu_{t-1}(\theta)}{\partial \theta},$$

where $\varepsilon_t(\theta) := Y_t - \mu_{t-1}(\theta)$ and where the Fisher information becomes

$$I_f = I_l \mathrm{E}\left[\frac{\partial \mu_{t-1}(\theta)}{\partial \theta} \frac{\partial \mu_{t-1}(\theta)}{\partial \theta^T} \right].$$

For notational convenience we assume stationarity here .

We study in this example the Ljung and Box (1978) statistic which is based on l-th order residual autocorrelations $\hat{\rho}(\theta; l)$ that satisfy, as we have seen,

$$\hat{\rho}(\theta; l) = \frac{1}{\sqrt{n}} \sum_{t=l+1}^{n} \frac{\varepsilon_t(\theta)\varepsilon_{t-l}(\theta)}{\sigma_\varepsilon^2} + o_{\mathrm{P}}(1), \tag{4.2}$$

as $n \to \infty$, where $\varepsilon_t(\theta) = Y_t - \mu_{t-1}(\theta)$.

The third and last ingredient that determines the behavior of the residual-based Ljung and Box (1978) statistic is the actual estimator used. Consider, for example, the standard least-squares estimator that satisfies

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \left(\mathrm{E}\left[\frac{\partial \mu_{t-1}(\theta)}{\partial \theta} \frac{\partial \mu_{t-1}(\theta)}{\partial \theta^T} \right] \right)^{-1} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \varepsilon_t \frac{\partial \mu_{t-1}(\theta)}{\partial \theta} + o_{\mathrm{P}}(1),$$

as $n \to \infty$.

In order to apply Theorem 2.1, we calculate

$$\begin{aligned}
\tau^2 &= 1, \\
\alpha &= \left(\mathrm{E}\left[\frac{\partial \mu_{t-1}(\theta)}{\partial \theta} \frac{\partial \mu_{t-1}(\theta)}{\partial \theta^T} \right] \right)^{-1} \mathrm{E}\left[\varepsilon_{t-l} \frac{\partial \mu_{t-1}(\theta)}{\partial \theta} \right], \\
\Gamma &= \sigma_\varepsilon^2 \left(\mathrm{E}\left[\frac{\partial \mu_{t-1}(\theta)}{\partial \theta} \frac{\partial \mu_{t-1}(\theta)}{\partial \theta^T} \right] \right)^{-1}, \\
c &= \sigma_\varepsilon^{-2} \mathrm{E}\left[\varepsilon_{t-l} \frac{\partial \mu_{t-1}(\theta)}{\partial \theta} \right].
\end{aligned} \tag{4.3}$$

Note that in the present example we have $\alpha = \Gamma c$. This implies that the asymptotic distribution of the l-th order autocorrelation calculated on least-squares residuals is

$$N\left(0, 1 - \sigma_\varepsilon^{-2} \mathrm{E}\left[\varepsilon_{t-l} \frac{\partial \mu_{t-1}(\theta)}{\partial \theta^T} \right] \left(\mathrm{E}\left[\frac{\partial \mu_{t-1}(\theta)}{\partial \theta} \frac{\partial \mu_{t-1}(\theta)}{\partial \theta^T} \right] \right)^{-1} \mathrm{E}\left[\frac{\partial \mu_{t-1}(\theta)}{\partial \theta} \varepsilon_{t-l}^T \right] \right). \tag{4.4}$$

Note, once more, that decreased limiting variance does not depend on the density f of the underlying innovations, but through some standard moments.

The Ljung and Box (1978) or the Box and Pierce (1970) statistic is based on a simultaneous comparison of the empirical autocorrelation at various lags. In order to derive the joint behavior of $\hat{\rho}(\hat{\theta}_n; l)$ for $l = 1, \ldots, L$, Theorem 2.1 can be utilized in its multivariate extension as discussed in Section 2. This leads to the limiting distribution (4.4) above, with ε_{t-l} replaced by the L-dimensional vector $(\varepsilon_{t-1}, \ldots, \varepsilon_{t-L})^T$. Restricting attention further to ARMA(p,q) model, we have $\partial \mu_{t-1}(\theta)/\partial \theta = (Y_{t-1}, \ldots, Y_{t-p}, \varepsilon_{t-1}$ so that the limiting variance only depends on the autocorrelation structure. Completing the calculation and taking $L \to \infty$ as $n \to \infty$, one verifies readily that the limiting variance is approximately a projection matrix with trace $L - p - q$, which leads to the classical result as in, e.g., Brockwell and Davis (1991). $\qquad\square$

Example 4.2 McLeod-Li in ARMA/regression models with OLS residuals

Following up on the previous example, we consider the situation where we want to test for serial correlation in the squared innovations. The McLeod and Li (1983) statistic is based on the empirical autocorrelation of squared innovations which are given by

$$
\begin{aligned}
\hat{\rho}_2(\theta; l) &= \sqrt{n} \frac{(n-l+1)^{-1} \sum_{t=l+1}^{n} \left[\varepsilon_t^2(\theta) - n^{-1} \sum_{t=1}^{n} \varepsilon_t^2(\theta)\right] \left[\varepsilon_{t-l}^2(\theta) - n^{-1} \sum_{t=1}^{n} \varepsilon_t^2(\theta)\right]}{n^{-1} \sum_{t=1}^{n} \varepsilon_t^4(\theta) - \left(n^{-1} \sum_{t=1}^{n} \varepsilon_t^2(\theta)\right)^2} \\
&= \frac{1}{\sqrt{n}} \sum_{t=l+1}^{n} \frac{(\varepsilon_t^2(\theta)/\sigma_\varepsilon^2 - 1)(\varepsilon_{t-l}^2(\theta)/\sigma_\varepsilon^2 - 1)}{\kappa_\varepsilon - 1} + o_{\mathbf{P}}(1),
\end{aligned}
\tag{4.5}
$$

as $n \to \infty$, assuming that the innovations have finite fourth moments $\kappa_\varepsilon \sigma_\varepsilon^4$.

Compared to the previous example, Γ doesn't change as it depends on the model and the estimator only. One easily verifies $\tau^2 = 1$ and finds $c = 0$, since, using integration by parts and $E\varepsilon_t = 0$, $E[-f'(\varepsilon_t)/f(\varepsilon_t)][\varepsilon_t^2/\sigma_\varepsilon^2 - 1] = 2 \int x f(x)\mathrm{d}x/\sigma_\varepsilon^2 = 0$. The actual form of α is easily obtained as well, but that need not concern us here as $c = 0$ implies that the limiting distribution of the residual-based statistic equals that of the innovation-based statistic, i.e., $N(0, \tau^2) = N(0, 1)$. The McLeod and Li (1983) statistic is based on $\sum_{l=1}^{L} \hat{\rho}_2^2(\hat{\theta}_n; l)$, compare the theorem of McLeod and Li (1983), page 271. Apparently, when applied to the residuals of a regression of ARMA model, the limiting distribution remains χ_L^2 with no correction for pre-estimated parameters. $\square$

Example 4.3 Ljung-Box/McLeod-Li in scale models with QMLE residuals

Tests for residual autocorrelation or squared residual autocorrelation are also often applied to the residuals of scale models. This situation occurs in financial modelling using ARCH-type processes or ACD-type models (Engle and Russell, 1998). In general terms, the scale model can be written as

$$
Y_t = \sigma_{t-1}(\theta)\varepsilon_t, \ t = 1, \ldots, n,
\tag{4.6}
$$

where $\sigma_{t-1}(\theta)$ depends on past values $Y_{t-1}, Y_{t-2}, \ldots$ and ε_t is a sequence of i.i.d. mean zero, unit variance innovations with finite Fisher information for scale $I_s := \int (1 + xf'(x)/f(x))^2 f(x) \mathrm{d}x < \infty$. As mentioned in the introduction, such a model satisfies the LAN property under sufficient regularity conditions. The central sequence for θ in these models reads

$$\Delta^{(n)}(\theta) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} -\frac{1}{2} \left(1 + \varepsilon_t(\theta) \frac{f'(\varepsilon_t(\theta))}{f(\varepsilon_t(\theta))} \right) \frac{\partial}{\partial \theta} \log \sigma_{t-1}^2(\theta),$$

where $\varepsilon_t(\theta) := Y_t/\sigma_{t-1}(\theta)$, while the Fisher information is given by

$$I_F = \frac{1}{4} I_s \mathrm{E} \left[\frac{\partial}{\partial \theta} \log \sigma_{t-1}^2(\theta) \frac{\partial}{\partial \theta^T} \log \sigma_{t-1}^2(\theta) \right].$$

The most often applied estimator in these models is the QMLE estimator $\hat{\theta}_n$ based on a imposed Gaussian distribution for the innovations ε_t. In various more specific cases, this QMLE estimator has been shown to satisfy the asymptotically linear representation:

$$\sqrt{n}(\hat{\theta}_n - \theta_n) = -\left(\mathrm{E} \left[\frac{\partial}{\partial \theta} \log \sigma_{t-1}^2(\theta) \frac{\partial}{\partial \theta^T} \log \sigma_{t-1}^2(\theta) \right] \right)^{-1} \times \qquad (4.7)$$
$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left(1 - \varepsilon_t^2 \right) \frac{\partial}{\partial \theta} \log \sigma_{t-1}^2(\theta) + o_{\mathrm{P}}(1),$$

under $\mathbb{P}_{\theta_n}^{(n)}$ and as $n \to \infty$. From this representation one immediately finds the asymptotic variance of the QMLE estimator as

$$\Gamma = (\kappa_\varepsilon - 1) \left(\mathrm{E} \left[\frac{\partial}{\partial \theta} \log \sigma_{t-1}^2(\theta) \frac{\partial}{\partial \theta^T} \log \sigma_{t-1}^2(\theta) \right] \right)^{-1}, \qquad (4.8)$$

with, as before, $\kappa_\varepsilon = \mathrm{E}\varepsilon_t^4$ (recall that in this scale model we normalized $\mathrm{E}\varepsilon_t^2 = 1$).

In order to find out the limiting distribution of the empirical l-th order autocorrelation of the residuals, or the squared residuals, we may use (4.2) and (4.5), to get the appropriate covariances in Condition (AN) for the Ljung and Box (1978) type test (c_{LB} and α_{LB}) and the McLeod and Li (1983) type test (c_{ML} and α_{ML}). This leads to

$$c_{LB} = 0,$$
$$\alpha_{LB} = \frac{\Gamma}{\kappa_\varepsilon - 1} \mathrm{E}\varepsilon_t^3 \mathrm{E} \left[\varepsilon_{t-l} \frac{\partial}{\partial \theta} \log \sigma_{t-1}^2(\theta) \right],$$
$$c_{ML} = \frac{1}{\kappa_\varepsilon - 1} \mathrm{E} \left[(\varepsilon_{t-l}^2 - 1) \frac{\partial}{\partial \theta} \log \sigma_{t-1}^2(\theta) \right],$$
$$\alpha_{ML} = \frac{\Gamma}{\kappa_\varepsilon - 1} \mathrm{E} \left[(\varepsilon_{t-l}^2 - 1) \frac{\partial}{\partial \theta} \log \sigma_{t-1}^2(\theta) \right],$$

since $E[1 + \varepsilon_t f'(\varepsilon_t)/f(\varepsilon_t)]\varepsilon_t = -2 \int x f(x)\mathrm{d}x = 0$ and $E[1 + \varepsilon_t f'(\varepsilon_t)/f(\varepsilon_t)][\varepsilon_t^2 - 1] = \int(x^3 - x)f'(x)\mathrm{d}x = -\int(3x^2 - 1)f(x)\mathrm{d}x = -2$, as $E\varepsilon_t = 0$ and $E\varepsilon_t^2 = 1$.

From $c_{LB} = 0$, we find that applying the Ljung and Box (1978) statistic to residuals of a scale model estimated by Gaussian QMLE, does not lead to an adaptation in the limiting distribution, in particular not to a reduction of the number of degrees of freedom in the χ^2 distribution as in the classical ARMA case. For the McLeod and Li (1983) statistic the situation is quite different. No further simplification occurs and the limiting distribution of the individual squared autocorrelations is given by (2.1) as

$$N\left(0, 1 - \frac{q}{\kappa_\varepsilon - 1}\right), \tag{4.9}$$

with

$$
\begin{aligned}
q = {} & E\left[(\varepsilon_{t-l}^2 - 1)\frac{\partial}{\partial\theta^T}\log\sigma_{t-1}^2(\theta)\right] \times \\
& \left(E\left[\frac{\partial}{\partial\theta}\log\sigma_{t-1}^2(\theta)\frac{\partial}{\partial\theta^T}\log\sigma_{t-1}^2(\theta)\right]\right)^{-1} \times \\
& E\left[(\varepsilon_{t-l}^2 - 1)\frac{\partial}{\partial\theta}\log\sigma_{t-1}^2(\theta)\right].
\end{aligned}
$$

The result (4.9) is also derived in Berkes et al. (2003) for residuals of the GARCH(p,q) model (compare also Horváth and Kokoszka, 2001). They, however, pay much more attention to the specific conditions needed so that, in our terminology, Condition (AN) is satisfied. Their Theorem 2.2 is the counterpart of (4.9) with the notation $d_0^2 = \kappa_\varepsilon - 1$, $i_k = l$, $c_{i_k} = E\left[(\varepsilon_{t-l}^2 - 1)\frac{\partial}{\partial\theta^T}\log\sigma_{t-1}^2(\theta)\right]$, $A_0 = \frac{1}{4}(\kappa_\varepsilon - 1)^2\Gamma^{-1}$, and $B_0 = -\frac{1}{2}(\kappa_\varepsilon - 1)\Gamma^{-1}$. Note that their Theorem 2.2 gives the limiting distribution of $(\kappa_\varepsilon - 1)\hat{\rho}_2(\hat{\theta}; l)$. $\qquad\square$

Example 4.4 Estimating GARCH on ARMA residuals
As we have seen in Section 2, our results can also be used to derive the limiting distribution of a two-step estimator. To illustrate this, we assume that the Gaussian QMLE as in (4.7) for the scale model (4.6) is calculated on residuals of an ARMA model that has been estimated at the first stage using least-squares as in Example 4.1. In order to apply Theorem 2.1, note that the statistic of interest now is the Gaussian QMLE for the scale model, while the underlying model and estimator are as in Example 4.1. Consequently, we have that Γ is as in (4.3), while

$$\tau^2 = (\kappa_\varepsilon - 1)\left(E\left[\frac{\partial}{\partial\theta}\log\sigma_{t-1}^2(\theta)\frac{\partial}{\partial\theta^T}\log\sigma_{t-1}^2(\theta)\right]\right)^{-1},$$

and $c = 0$ since $E[-f'(\varepsilon_t)/f(\varepsilon_t)][1 - \varepsilon_t^2] = -\int 2x f(x)\mathrm{d}x = 0$. Note that τ^2 and c are matrices in this case. Once more, the actual form of α, although it can be easily

derived, is irrelevant as $c = 0$ implies that the limiting variance of the Gaussian QMLE applied to the residuals is the same as that applied to the innovations, namely τ^2 above. The asymptotic distribution of Lagrange multiplier (LM) tests that are based on such a two-stage approach, and examine, for instance, dependence and nonlinearity in the residuals of ARMA models, can also be considered in the above context. The two-stage and related estimator's properties in Pagan (1986) can also be treated in the context of Theorem 2.1. □

Example 4.5 Rank test for residual autocorrelation

One of the advantages of our approach is that we do not require differentiability of our test-statistic with respect to the parameter θ. This is particularly helpful when considering rank-based statistics since they are, by definition, not smooth in the parameter θ for given observations. To introduce the statistic, write $R_t(\theta)$ for the rank of the t-th innovation $\varepsilon_t(\theta)$ among all innovations $\varepsilon_1(\theta), \ldots, \varepsilon_n(\theta)$. Consider a rank-based test for l-th order autocorrelation

$$r_n(\theta; l; g) = \frac{1}{n-l} \sum_{t=l+1}^{n} \frac{-g'}{g} \left(G^{-1} \left(\frac{R_t(\theta)}{n+1} \right) \right) G^{-1} \left(\frac{R_{t-l}(\theta)}{n+1} \right) / \sqrt{I_g}, \qquad (4.10)$$

where g denotes some zero mean and unit variance reference density, with corresponding cumulative distribution function G and Fisher information for location $I_g = \int (g'/g)^2 g < \infty$. The so-called van der Waerden autocorrelations are obtained by taking g the standard normal density, while the logistic density leads to the Wilcoxon autocorrelations. Many more examples can be found in the overview of Hallin and Werker (1999), which also gives the relevant asymptotically linear representations used below. The prime advantage of using rank-based autocorrelations is that they are insensitive to misspecification of the innovation distribution (since they are distribution-free), while they still may lead to semiparametrically efficient inference procedures (Hallin and Werker, 2003).

Our interest lies in the behavior of the rank-based autocorrelation (4.10), when applied to residuals of some model estimated during a first-stage analysis. Let's consider the situation mentioned in the abstract of residuals of a scale model (like GARCH(2,2)) estimated using Gaussian QMLE. The relevant model is thus described in Example 4.3 and Γ is given by (4.8). In order to verify Condition (AN), an asymptotically linear representation is needed for the rank-based autocorrelation $r_n(\theta; l; g)$. These results are well-known in the statistics literature and, assuming that the density g is strongly unimodal (i.e., $-g'/g$ is monotone increasing), we find

$$r_n(\theta; l; g) = \frac{1}{n-l} \sum_{t=l+1}^{n} \frac{-g'}{g} \left(G^{-1} \left(F(\varepsilon_t) \right) \right) G^{-1} \left(F(\varepsilon_{t-l}) \right) / \sqrt{I_g} + o_{\mathrm{P}}(n^{-1/2}), \quad (4.11)$$

as $n \to \infty$. Note that F denotes the true (unknown) distribution of the innovations, while G is a reference distribution that need not equal F. The rank-based autocorrelations are asymptotically normally distributed with unit variance even if $G \neq F$.

Their power for detecting l-th order autocorrelation, however, is maximal if G is close to F. Using the asymptotically linear representation (4.11), we find

$$c = -\frac{\tilde{c}}{2\sqrt{I_g}} \mathrm{E}\left[G^{-1}(F(\varepsilon_t))\frac{\partial}{\partial\theta}\log\sigma_{t-1}^2(\theta)\right],$$

$$\alpha = \frac{\tilde{\alpha}}{(\kappa_\varepsilon - 1)\sqrt{I_g}}\Gamma\mathrm{E}\left[G^{-1}(F(\varepsilon_t))\frac{\partial}{\partial\theta}\log\sigma_{t-1}^2(\theta)\right],$$

with

$$\tilde{c} := \mathrm{E}[1 + \varepsilon_t f'(\varepsilon_t)/f(\varepsilon_t)][g'(G^{-1}(F(\varepsilon_t)))/g(G^{-1}(F(\varepsilon_t)))]$$

$$= \int_{u=0}^1 F^{-1}(u)(f'/f)(F^{-1}(u))(g'/g)(G^{-1}(u))\mathrm{d}u,$$

and $\tilde{\alpha} := \int_{u=0}^1 [1 - F^{-1}(u)^2][(g'/g)(G^{-1}(u))]$.

With the above expressions, Theorem 2.1 can be applied directly. Note that the limiting distribution of the rank-based autocorrelations are not distribution free, i.e., depend on the underlying distribution F of the innovations. However, if both the true distribution F and the reference distribution G are symmetric about zero, one finds $\tilde{c} = \tilde{\alpha} = 0$. In that case, c is zero and also the rank-based autocorrelation calculated on the residuals is asymptotically standard normally distributed. $\qquad\square$

4.2 Goodness-of-fit tests

Example 4.6 Goodness-of-Fit tests
Next to testing for linear or non-linear dependence, one is often also interested in testing a particular distribution for the innovations ε_t. Having standard Goodness-of-Fit tests in mind, we are, therefore, interested in the limiting distribution of the empirical distribution function of residuals. We consider the empirical distribution at a fixed point $z \in \mathbb{R}$ first, i.e., the statistic of interest can be written as

$$T_n(\theta) = \sqrt{n}(F_n(z) - F(z)) = \frac{1}{\sqrt{n}}\sum_{t=1}^n \left(I\{\varepsilon_t(\theta) \leq z\} - F(z)\right).$$

For expository reasons, we consider residuals of an ARMA or regression model as in Example 4.1 only. Once more, the calculations to verify Condition (AN) are straightforward:

$$c = -f(z)\mathrm{E}\left[\frac{\partial\mu_{t-1}(\theta)}{\partial\theta}\right],$$

$$\alpha = m(z)\left(\mathrm{E}\left[\frac{\partial\mu_{t-1}(\theta)}{\partial\theta}\frac{\partial\mu_{t-1}(\theta)}{\partial\theta^T}\right]\right)^{-1}\mathrm{E}\left[\frac{\partial\mu_{t-1}(\theta)}{\partial\theta}\right],$$

since $\mathrm{E}[f'(\varepsilon_t)/f(\varepsilon_t)]I\{\varepsilon_t \le z\} = \int f'(x)I\{x \le z\}\mathrm{d}x = f(z)$ and with $m(z) := \mathrm{E}\varepsilon_t I\{\varepsilon_t \le z\} = \int xf(x)I\{x \le z\}\mathrm{d}x$. Since $\tau^2 = F(z)[1 - F(z)]$, the residual-based empirical distribution function at z has, according to Theorem 2.1, limiting variance

$$F(z)[1 - F(z)]+$$
$$\left(f(z)^2\sigma_\varepsilon^2 + 2f(z)m(z)\right)\mathrm{E}\left[\frac{\partial\mu_{t-1}(\theta)}{\partial\theta^T}\right]\left(\mathrm{E}\left[\frac{\partial\mu_{t-1}(\theta)}{\partial\theta}\frac{\partial\mu_{t-1}(\theta)}{\partial\theta^T}\right]\right)^{-1}\mathrm{E}\left[\frac{\partial\mu_{t-1}(\theta)}{\partial\theta}\right].$$

The above analysis is restricted in the sense that the empirical distribution function is evaluated at a fixed point z only. An extension to the multivariate situation of the empirical distribution function evaluated in the points $(z_1, \ldots, z_m)$ is straightforward. More difficult, and beyond the scope of the present paper, would be to find a functional limit theorem for the residual-based empirical distribution. The first to study such a problem is Durbin (1973). His Theorem 1 is comparable to our Theorem 2.1 with the notation $t(t-1) = F(z)[1 - F(z)] = \tau^2$, $h = \alpha$, $g_2 = c$, and $L = \Gamma$, under the null-hypothesis $\gamma = 0$. $\qquad\square$

Example 4.7 Goodness-of-Fit tests with nuisance parameters
The previous example considers the case where residuals are tested against a completely specified distribution F. Clearly, one often encounters the situation where this distribution is not completely specified. For instance, consider the same setup as in Example 4.6. Now, however, we want to test whether the residuals belong to the normal scale family $\{N(0, \sigma^2) : \sigma^2 > 0\}$. To this extent, we use the test statistic

$$
\begin{aligned}
T_n(\theta) &= \sqrt{n}\left(F_n(z) - \Phi\left(\frac{z}{s_n}\right)\right) \\
&= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left(I\{\varepsilon_t \le z\} - \Phi\left(\frac{z}{\sigma_\varepsilon}\right) + \frac{1}{2}\varphi\left(\frac{z}{\sigma_\varepsilon}\right)\frac{z}{\sigma_\varepsilon}\left(\frac{s_n^2}{\sigma_\varepsilon^2} - 1\right)\right) + o_\mathrm{P}(1) \\
&= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left(I\{\varepsilon_t \le z\} - \Phi\left(\frac{z}{\sigma_\varepsilon}\right) + \frac{1}{2}\varphi\left(\frac{z}{\sigma_\varepsilon}\right)\frac{z}{\sigma_\varepsilon}\left(\frac{\varepsilon_t^2}{\sigma_\varepsilon^2} - 1\right)\right) + o_\mathrm{P}(1),
\end{aligned}
$$

as $n \to \infty$, and where the estimated variance of the innovations is $s_n^2 = \sum_{t=1}^{n}\varepsilon_t^2/n$ and Φ denotes the standard normal distribution function; φ its density. Once more, the derivations are straightforward and lead to

$$
\begin{aligned}
c &= -f(z)\mathrm{E}\left[\frac{\partial\mu_{t-1}(\theta)}{\partial\theta}\right], \\
\alpha &= \left(m(z) + \frac{1}{2}\varphi\left(\frac{z}{\sigma_\varepsilon}\right)\frac{z}{\sigma_\varepsilon}\mathrm{E}\varepsilon_t^3\right)\left(\mathrm{E}\left[\frac{\partial\mu_{t-1}(\theta)}{\partial\theta}\frac{\partial\mu_{t-1}(\theta)}{\partial\theta^T}\right]\right)^{-1}\mathrm{E}\left[\frac{\partial\mu_{t-1}(\theta)}{\partial\theta}\right].
\end{aligned}
$$

The limiting distribution of the test-statistic applied to residual of the ARMA or regression model, follows again immediately. Note that, in case $\mathrm{E}\varepsilon_t^3 = 0$, the formulae

for c and α are the same as in Example 4.6. Consequently, the change in variance due to applying the statistic on residuals instead of actual innovations is the same, although the limiting distribution of the statistic applied to innovations clearly differs in both cases. $\square$

Example 4.8 Misspecification testing in discrete choice models
Consider the binary choice model

$$\mathbb{P}\{Y_i = 1 | X_i\} = F(X_i^T \theta),$$

where, for instance, $F = \Phi$ denotes the probit model. The log-likelihood function is

$$\log L(\theta) = \sum_{i=1}^{n} Y_i \log F(X_i^T \theta) + (1 - Y_i) \log \left(1 - F(X_i^T \theta)\right),$$

which leads, under some standard regularity conditions, to the central sequence

$$\Delta^{(n)}(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{Y_i - F(X_i^T \theta)}{F(X_i^T \theta) \left(1 - F(X_i^T \theta)\right)} f(X_i^T \theta) X_i = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i^G X_i. \qquad (4.12)$$

Here f denotes the density corresponding to F and $\varepsilon_i^G = \varepsilon_i^G(\theta)$ is sometimes called the generalized residual. The corresponding Fisher information matrix is given by

$$I_F = \mathrm{E} \left(\varepsilon_i^G\right)^2 X_i X_i^T = \mathrm{E} \frac{f(X_i^T \theta)^2}{F(X_i^T \theta) \left(1 - F(X_i^T \theta)\right)} X_i X_i^T.$$

We consider the parametric case, where θ is estimated using maximum likelihood, so that $\sqrt{n}(\hat{\theta}_n - \theta_0) = I_F^{-1} \Delta^{(n)}(\theta_0) + o_{\mathrm{P}}(1)$ and $\Gamma = I_F^{-1}$.

The omitted variables test based on exogenous variables Z_i uses the test-statistic

$$T_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i^G(\theta) Z_i. \qquad (4.13)$$

In order to derive the limiting distribution of the statistic $T_n(\hat{\theta}_n)$, note that the use of the efficient maximum likelihood estimator implies using Theorem 2.1,

$$T_n(\hat{\theta}_n) \approx N \left(0; \tau^2 - \alpha^T \Gamma^{-1} \alpha\right).$$

Here the limiting variance τ^2 of $T_n(\theta)$ is easily found to be

$$\tau^2 = \mathrm{E} \left(\varepsilon_i^G\right)^2 Z_i Z_i^T = \mathrm{E} \frac{f(X_i^T \theta)^2}{F(X_i^T \theta) \left(1 - F(X_i^T \theta)\right)} Z_i Z_i^T.$$

The asymptotic covariance between the statistic T_n and the estimator $\sqrt{n}(\hat{\theta}_n - \theta)$ is easily seen to be given by

$$\alpha = I_F^{-1} \mathrm{E} \left(\varepsilon_i^G\right)^2 X_i Z_i^T = I_F^{-1} \mathrm{E} \frac{f(X_i^T \theta)^2}{F(X_i^T \theta) \left(1 - F(X_i^T \theta)\right)} X_i Z_i^T.$$

Consequently, the variance which appropriately corrects for pre-estimating θ in the omitted variables test is

$$\mathrm{E}w_i Z_i Z_i^T - \mathrm{E}w_i X_i Z_i^T \left(\mathrm{E}w_i X_i X_i^T\right)^{-1} \mathrm{E}w_i X_i Z_i^T,$$

with

$$w_i = \frac{f(X_i^T\theta)^2}{F(X_i^T\theta)\left(1 - F(X_i^T\theta)\right)}.$$

This result has been derived using the traditional approach with conditional moment or lagrange multiplier test (see for instance Pagan and Vella, 1989). Tests for the presence of heteroskedasticity in discrete choice models use test-statistics similar to (4.13) and can be analyzed in the same framework.

The test statistic (4.13) checks for linear correlation between the generalized residuals and the possibly omitted variables Z_i. In case one prefers a test with power against non-linear forms of correlation, one could consider the statistic Kendall's tau applied to the pairs $(\varepsilon_i^G(\theta), Z_i)$. For simplicity we consider the case where the possibly omitted variables are univariate, i.e., $Z_i \in \mathbb{R}$. Recall that Kendall's tau is defined by

$$4\mathbb{P}\{\varepsilon_i^G(\theta) < \varepsilon_j^G(\theta),\ Z_i < Z_j\} - 1,\ i \neq j.$$

An appropriately rescaled version of Kendall's tau is the U-statistic

$$T_n^\tau(\theta) = n^{1/2} \binom{n}{2}^{-1} \sum_{i=1}^n \sum_{j=1}^{i-1} \left[4I\{\varepsilon_i^G(\theta) < \varepsilon_j^G(\theta),\ Z_i < Z_j\} - 1\right].$$

The limiting distribution of $T_n^\tau(\theta)$ is known to be $\tau^2 = 4/9$ under the null hypothesis of independent ε_i^G and Z_i. In order to derive the appropriate variance correction when calculating Kendall's tau using generalized residuals on the basis of the maximum likelihood estimator $\hat{\theta}_n$, we need to calculate

$$
\begin{aligned}
\alpha &= \lim_{n \to \infty} \mathrm{Cov}_{\theta_0}\left\{\sqrt{n}(\hat{\theta}_n - \theta_0), T_n^\tau(\theta_0)\right\} \\
&= I_F^{-1} \lim_{n \to \infty} \binom{n}{2}^{-1} \sum_{l=1}^n \sum_{i=1}^n \sum_{j=1}^{i-1} \mathrm{Cov}\left\{\varepsilon_l^G X_l, 4I\{\varepsilon_i^G(\theta) < \varepsilon_j^G(\theta),\ Z_i < Z_j\} - 1\right\} \\
&= 4I_F^{-1} \lim_{n \to \infty} \binom{n}{2}^{-1} \sum_{i=1}^n \sum_{j=1}^{i-1} \mathrm{E}\left(\varepsilon_i^G X_i + \varepsilon_j^G X_j\right) I\{\varepsilon_i^G(\theta) < \varepsilon_j^G(\theta),\ Z_i < Z_j\}.
\end{aligned}
$$

Further simplification of α is not necessary as the above expression is easily consistently estimated. Applying Theorem 2.1 we immediately obtain, under the null hypothesis of no omitted variables,

$$T_n^\tau(\theta) \sim N\left(0; \frac{4}{9} - \alpha^T I_F \alpha\right).$$

As far as we know, this nonparametric test for omitted variables in the binary choice model has not been considered before in the literature. We provide this example to show that its limiting distribution is easily derived in the framework we propagate. $\square$

4.3 Homogeneity tests

Example 4.9 Rank-based tests for structural breaks
As a final example we consider the problem of testing for a structural break in the innovation's distribution, using a rank-based CUSUM type test (see, for instance, Sen, 1984, for the linear regression model). We focus here on the case where the possible break-point is known. The case with unknown break-point leads to non-normally distributed test statistics (see, for instance, the sup of Brownian bridge asymptotic results of historical or sequential rank- and EDF-based tests in Bhatacharyya and Frierson, 1981, Picard, 1981, and Horváth et al., 2001). These non-normal limiting distributions cannot be handled directly by our approach. For illustrative purposes, we consider in this example the scale model as described in Example 4.3. Other models can be handled in exactly the same way with adapted expressions for the relevant variances and covariance c, α, and Γ. For a known change-point at the s-th quantile of the sample, the test statistic of interest is

$$T_n(\theta) = \frac{1}{\sqrt{n}} \sum_{t=1}^{\langle ns \rangle} \left(\frac{R_t(\theta)}{n+1} - \frac{1}{2} \right),$$

where $R_t(\theta)$ denotes the rank of the t-th innovation $\varepsilon_t(\theta)$ among all n innovations $\varepsilon_1(\theta), \ldots, \varepsilon_n(\theta)$ and $\langle \cdot \rangle$ denotes the entier function. A standard theorem on the asymptotically linear representation of rank-statistics (e.g., Hájek, Šidák, and Sen, 1999, Chapter 6) shows

$$T_n(\theta) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left(I\{t \le \langle ns \rangle\} - s \right) \left(F(\varepsilon_t) - \frac{1}{2} \right) + o_P(1),$$

as $n \to \infty$. One immediately verifies

$$
\begin{aligned}
c &= \frac{\int x f^2(x) \mathrm{d}x}{2} \frac{1}{n} \sum_{t=1}^{n} [I\{t \le \langle ns \rangle\} - s] \frac{\partial}{\partial \theta} \log \sigma_{t-1}^2(\theta) + o_P(1) \\
&\to 0,
\end{aligned}
$$

since $\mathrm{E}[1 + \varepsilon_t f'(\varepsilon_t)/f(\varepsilon_t)] F(\varepsilon_t) = \int (x f'(x) + f(x)) F(x) \mathrm{d}x = -\int x f^2(x) \mathrm{d}x$, assuming that the process $\frac{\partial}{\partial \theta} \log \sigma_{t-1}^2(\theta)$ satisfies a law-of-large numbers, so that

$$
\begin{aligned}
\frac{1}{n} \sum_{t=1}^{n} [I\{t \le \langle ns \rangle\} - s] \frac{\partial}{\partial \theta} \log \sigma_{t-1}^2(\theta) &\to \{s(1-s) - (1-s)s\} \mathrm{E} \frac{\partial}{\partial \theta} \log \sigma_{t-1}^2(\theta) \\
&= 0.
\end{aligned}
$$

Such an assumption is satisfied in the standard models described in the introduction. From $c = 0$, we deduce that the rank-based CUSUM statistic when applied to residuals of a GARCH-type model, does not require any size correction. $\qquad\square$

5 Final remarks and future work

The present paper considers the asymptotic analysis of residuals-based statistics in a Gaussian limiting framework: The models under consideration are assumed to be asymptotically Gaussian shift experiments (through the LAN condition), the statistics being studied have limiting Gaussian distributions, and the estimators under consideration are limiting Gaussian as well. For this situation we introduce a novel approach to the derivation of residual-based statistics and two-stage estimators. We illustrate the power of this approach by numerous examples.

We envisage that our approach can be extended in several interesting directions. First of all, while the Gaussian context has many applications for residual-based tests in econometric models as discussed in the paper, it also represents the foundations for an alternative and simple approach of deriving the asymptotic distribution of certain other statistics in non-Gaussian limiting statistical experiments (like for non-stationary time series). One could also think of extensions to statistics that have sup-of-Gaussian processes as their limiting distribution (like Kolmogorov-Smirnov type goodness-of-fit tests). Also, while we have shown how our results can be applied in semi- and nonparametric models where a Euclidean parameter is pre-estimated, it is of interest to see how the results can be extended to pre-estimation of infinite dimensional parameters, like densities or regression functions. A deeper analysis of the applicability to simulation based estimators (Indirect inference, EMM, SMM) would largely enhance the scope of the analysis. Finally, Le Cam's third lemma is not restricted to LAN models. Therefore, non-stationary and non-smooth functional models can be considered. For instance, the extension of the proposed method to non-stationary processes with locally and asymptotically mixed normal behavior remains for future work.

A Le Cam's third Lemma

This appendix recalls Le Cam's third lemma in its best-known form, i.e., for asymptotically normal distribution. Consider two sequences of probability measures $\left(\mathbb{Q}^{(n)}\right)_{n=1}^{\infty}$ and $\left(\mathbb{P}^{(n)}\right)_{n=1}^{\infty}$. Assume that the corresponding likelihood ratios $d\mathbb{Q}^{(n)}/d\mathbb{P}^{(n)}$ satisfy, jointly with some statistic T_n, under $\mathbb{P}^{(n)}$ and as $n \to \infty$,

$$
\begin{pmatrix} T_n \\ \log \frac{d\mathbb{Q}^{(n)}}{d\mathbb{P}^{(n)}} \end{pmatrix} \xrightarrow{\mathcal{L}} \begin{pmatrix} T \\ \Lambda \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ -\frac{\sigma^2}{2} \end{bmatrix}, \begin{bmatrix} \tau^2 & c \\ c & \sigma^2 \end{bmatrix} \right). \tag{A.1}
$$

Le Cam's third lemma now gives the limiting behavior of the statistic T_n also under $\mathbb{Q}^{(n)}$. More precisely,

$$
T_n \xrightarrow{\mathcal{L}} N(c, \tau^2),
$$

under $\mathbb{Q}^{(n)}$ and as $n \to \infty$. The intuition for this result is based on the fact that a statistic T which is jointly normally distributed with some log-likelihood ratio Λ as in (A.1), has $N(c, \tau^2)$ distribution under the alternative measure. This non-asymptotic version follows trivially from writing down the appropriate densities and likelihood ratios. Le Cam's third lemma provides in fact a continuity result in the sense that this transformation remains valid in the limit. It is not restricted to limiting Gaussian distributions, but it takes an especially simple form in that case.

Le Cam's third lemma is discussed in any modern book on asymptotic statistics, e.g., Hájek and Šidák (1967), Le Cam and Yang (1990), Bickel et al. (1993), or van der Vaart (1998).

References

Amemiya, T. (1982). Two Stage Least Absolute Deviations Estimators. *Econometrica* **50**, 689–712.

Anderson, T.W., and S. Takamitsu (1979). Evaluation of the Distribution Function of the Two-Stage Least Squares Estimate. *Econometrica* **47**, 163–182.

Andrews, D.W.K. (1988a). Chi-Square Diagnostic Tests for Econometric Models: Introduction and Applications. *Journal of Econometrics* **37**, 135–156.

Andrews, D.W.K. (1988b). Chi-Square Diagnostic Tests for Econometric Models: Theory. *Econometrica* **56**, 1419–1453.

Andrews, D.W.K. (1997). A conditional Kolmogorov test. *Econometrica* **65**, 1097–1128.

Bai, J. (2003). Testing parametric conditional distributions of dynamic models. *Review of Economics and Statistics*, forthcoming.

Bera, A.K., and C.M. Jarque (1982). Model specification tests: A simultaneous approach. *Journal of Econometrics* **20**, 59–82.

Bera, A.K., and M.J. Yoon (1993). Specification testing with misspecified local alternatives. *Econometric Theory* **9**, 649–658.

Berkes, I., L. Horváth, and P. Kokoszka (2003). Asymptotics for GARCH squared residual correlations. *Econometric Theory* **19**, 515–540.

Bhatacharyya, P.K., and D. Frierson (1981). A nonparametric control chart for detecting small disorders. *The Annals of Statistics* **9**, 544–554.

Bickel, P.J. (1982). On adaptive estimation. *Annals of Statistics* **10**, 647–671.

BICKEL, P.J., C.A.J. KLAASSEN, Y. RITOV, and J.A. WELLNER (1993). *Efficient and Adaptive Statistical Inference for Semiparametric Models.* John Hopkins University Press, Baltimore.

BIERENS, H.J. (1990). A consistent conditional moment test of functional form. *Econometrica* **58**, 1443–1458.

BLUNDELL, R.W. (eds.) (1987). Specification testing in limited and discrete variable models. *Journal of Econometrics* **34**.

BOX, G.E.P., and D.A. PIERCE (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of American Statistical Association* **65**, 1509–1526.

BROCKWELL, P.J., and R.A. DAVIS (1991). *Time series: Theory and methods.* Springer, Berlin.

CHESHER, A., and R. SMITH (1997). Likelihood ratio specification tests. *Econometrica* **65**, 627–646.

COTTER, K.D., (1986). Similarity of information and behavior with a pointwise convergence topology. *Journal of Mathematical Economics* **15**, 25–38.

D'AGOSTINO, R.B., and M.A. STEPHENS (eds.) (1986). *Goodness-of-Fit Techniques.* Statistics: textbooks and monographs, **68**, Marcel Dekker.

DROST, F.C. and C.A.J. KLAASSEN (1997). Efficient estimation in semiparametric GARCH models. *Journal of Econometrics* **81**, 193–221.

DROST, F.C., C.A.J. KLAASSEN, and B.J.M. WERKER (1997). Adaptive estimation in time-series models. *Annals of Statistics* **25**, 786–818.

DROST, F.C., and B.J.M. WERKER (2003). Semiparametric duration models. *Journal of Business and Economic Statistics* forthcoming.

DURBIN, J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *Annals of Statistics* **1**, 279–290.

ENGLE, R.F., and J.R. RUSSELL (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica* **66**, 1127–1162.

FABIAN, V. and J. HANNAN (1982). On estimation and adaptive estimation for locally asymptotically normal families. *Zeitschrift für Wahrscheinlichkeitsrechnung und verwandte Gebiete* **59**, 459–478.

GODFREY, L.G. (1988) *Misspecification tests in econometrics.* Cambridge University Press, Cambridge.

HAHN, J., and G. KUERSTEINER (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large. *Econometrica* **70**, 1639-1657.

HÁJEK, J. and Z. ŠIDÁK (1967). *Theory of Rank Tests.* Academic Press, New York.

HÁJEK, J., Z. ŠIDÁK, and P.K. SEN (1999). *Theory of Rank Tests.* Academic Press, New York.

HALLIN, M., and M.L. PURI (1990). Rank tests for time series analysis: A Survey. *New Directions in Time Series Analysis: Part I.* D. Brillinger, P. Caines, J. Geweke, E. Parzen, M. Rosenblatt, and M.S. Taqqu (eds), 111–153.

HALLIN, M. and B.J.M. WERKER (1999). Optimal testing for semiparametric AR models: from Lagrange multipliers to autoregression rank scores and adaptive tests. In *Asymptotics, Nonparametrics and Time Series*, (S. Ghosh, Ed.), 295-350. M. Dekker, New York.

HALLIN, M. and B.J.M. WERKER (2003). Semiparametric Efficiency, Distribution-Freeness, and Invariance. *Bernoulli* **9**, 137–165.

HANSEN, B.E. (2000). Sample splitting and threshold estimation. *Econometrica* **68**, 575–603.

HILLIER, G.H., and M.L. KING (eds.) (1991). 40 years of diagnostic testing. *Journal of Econometrics* **47**.

HORVÁTH, L. and P. KOKOSZKA (2001). Large sample distribution of weighted sums of ARCH(p) squared residual correlations. *Econometric Theory* **17**, 283–295.

HORVÁTH, L., P. KOKOSZKA, and G. TEYSSIERE (2001). Empirical process of the squared residuals of an ARCH sequence. *Annals of Statistics* **29**, 445–469.

JEGANATHAN, P. (1995). Some aspects of asymptotic theory with applications to time series models. *Econometric Theory* **11**, 818–887.

JEGANATHAN, P. (1997). On asymptotic inference in linear cointegrated time series systems. *Econometric Theory* **13**, 692–745.

JEGANATHAN, P. (1999). On asymptotic inference in linear cointegrated time series with fractionally integrated errors. *Econometric Theory* **15**, 583–621.

KEUZENKAMP, H.A., and J.R. MAGNUS (1995) (eds). The significance of testing in econometrics. *Journal of Econometrics* **67**.

KHMALADZE, E.V. (1981). An innovation approach to goodness-of-test in $\mathbb{R}^n$. *Annals of Statistics* **16**, 1503–1516.

KOUL, H.L. and A. SCHICK (1997). Efficient estimation in nonlinear autoregressive time series models. *Bernoulli* **3**, 247–277.

KOUL, H.L. and W. STUTE (1999). Nonparametric model checks for time series. *Annals of Statistics* **27**, 204–236.

KOUL, H.L. (2002). Weighted Empirical Processes in Dynamic Nonlinear Models. *Lecture Notes in Statistics* **166**, Springer, Berlin.

KREISS, J.P. (1987a). On adaptive estimation in autoregressive models when there are nuisance functions, *Statistics and Decisions* **5**, 59–76.

KREISS, J.P. (1987b). On adaptive estimation in stationary ARMA processes. *Annals of Statistics* **15**, 112–133.

LE CAM, L. and G.L. YANG (1990). *Asymptotics in Statistics; Some Basic Concepts*. Springer, Berlin.

LINTON, O. (1993). Adaptive estimation in ARCH models. *Econometric Theory* **9**, 539–569.

LJUNG, G.M., and G.E.P. BOX (1978). On a measure of lack of fit in time series models. *Biometrika* **65**, 297–303.

LUNDBERG, S., and T. TERASVIRTA (2002). Evaluating GARCH models. *Journal of Econometrics* **110**, 417–435.

MCKEAN, J.W., S.J. SHEATHER, and T.P. HETTMANSPERGER (1990). Regression diagnostics for rank-based methods. *Journal of American Statistical Association* **85**, 1018–1028.

MACKINNON, J.G. (1992). Model misspecification tests and artificial regression. *Journal of Economic Literature* **30**, 102–146.

MCLEOD, A.I., and W.K. LI (1983). Diagnostic checking ARMA time series models using squared-residual autocorrelations. *Journal of Time Series Analysis* **4**, 269–273.

NEWEY, W. (1985). Maximum likelihood specification testing and conditional moment tests. *Econometrica* **53**, 1047–1070.

NEWEY, W., and D. MCFADDEN (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, Ch. 36, Vol. IV, R. Engle and D. McFadden (eds.).

PAGAN, A. (1986). Two Stage and Related Estimators and Their Applications. *The Review of Economic Studies* **53**, 517–538.

PAGAN, A., and F. VELLA (1989). Diagnostic tests for models based on individual data: A Survey. *Journal of Applied Econometrics* **4**, S29–S59.

PICARD, D. (1985). Testing and estimating change-points in time series. *Advances in Applied Probability* **17**, 841–867.

PIERCE, D.A. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *The Annals of Statistics* **10**, 475–478.

PLOBERGER, W., and W. KRAMER (1992). The CUSUM test for OLS residuals. *Econometrica* **60**, 271–285.

POLLARD, D. (2004). *Asymptopia*, manuscript, Yale.

POWELL, J.L. (1983). The Asymptotic Normality of Two-Stage Least Absolute Deviations Estimators. *Econometrica* **51**, 1569–1576.

QIAN, L. (1998). On maximum likelihood estimation for a threshold autoregression. *Journal of Statistical Planning and Inference* **75**, 21–46.

RANDLES, R.H. (1982). On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics* **10**, 462–474.

SEN, P.K. (1984). Nonparametric procedures for some miscellaneous problems. *Handbook of Statistics* **4**, Chapter 29.

SPANOS, A. (1986). *Statistical Foundations of Econometric Modelling.* Cambridge University Press, Cambridge.

STRASSER, H. (1985). *Mathematical Theory of Statistics.* Walter de Gruyter, New York.

STUTE, W. (1997). Nonparametric model checks for regression. *Annals of Statistics* **25**, 613–641.

SUN, Y., and T. TSENGOS (2004). Semiparametric efficient adaptive estimation for asymmetric GARCH models. Working paper, Department of Economics, University of Guelph.

TANIGUCHI, M., and Y. KAKIZAWA (2000). *Asymptotic Theory of Statistical Inference for Time Series.* Springer, Berlin.

TAUCHEN, G. (1985). Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics* **30**, 415–443.

WHITE, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica* **50**, 1–26.