Robust Non-Gaussian Inference for Linear Simultaneous Equations Models*

Adam Lee[(a)] *and Geert Mesters*[(a,b)]

[(a)] Universitat Pompeu Fabra and Barcelona GSE

[(b)] Vrije Universiteit Amsterdam

October 21, 2021

Abstract

All parameters in linear simultaneous equations models can be identified (up to permutation and scale) if the underlying structural shocks are independent and if at most one of them is Gaussian. Unfortunately, existing inference methods that exploit such a non-Gaussian identifying assumption suffer from size distortions when the true shocks are close to Gaussian. To address this *weak non-Gaussian* problem, we develop a robust semi-parametric inference method that yields valid confidence intervals for the structural parameters of interest regardless of the *distance to Gaussianity*. We treat the densities of the structural shocks non-parametrically and construct identification robust tests based on the efficient score function. The approach is shown to be applicable for a broad class of linear simultaneous equations models in cross-sectional and panel data settings. A simulation study and an empirical study for production function estimation highlight the practical relevance of the methodology.

JEL classification: C12, C14, C30

Keywords: Weak identification, semiparametric modeling, independent component analysis, simultaneous equations.

*Email: adam.lee@upf.edu, geert.mesters@upf.edu. Address: Jaume 1, Ramon Trias Fargas 25-27, 08005, Barcelona, Spain. We thank numerous seminar participants for helpful comments. Mesters acknowledge support from the Spanish Ministry of Economy and Competitiveness through the Ramon y Cajal fellowship (RYC2019-028287-I), the Spanish Ministry of Economy and Competitiveness through the Severo Ochoa Programme for Centres of Excellence in R&D (CEX2019-000915-S), and the Netherlands Organization for Scientific Research (NWO) through the VENI research grant (016.Veni.195.036).

1 Introduction

The linear simultaneous equations model (LSEM) is a benchmark model used to analyse general equilibrium relationships in economics. It was formalized in its modern form by Haavelmo (1943, 1944), building on Frisch (1933) and Tinbergen (1939) among others. As is well known, without further restrictions, not all parameters of the LSEM can be uniquely recovered from the observed data series, see Dhrymes (1994) for an in-depth discussion.

Interestingly, this identification problem vanishes (up to permutation and scale) when the underlying structural shocks are independent and at most one of them follows a Gaussian distribution (e.g. Comon, 1994). This non-Gaussian identification approach has a long history in the statistics literature where it is often referred to as independent components analysis (ICA), see Hyvärinen, Karhunen and Oja (2001) for a textbook treatment. More recently, the economics literature has started investigating this approach and developing the corresponding methodology for conducting inference on the parameters of various LSEMs based on non-Gaussian identification.[1]

Unfortunately, existing inference methods suffer from size distortions when the true distributions are close to Gaussian. To understand the source of this *weak non-Gaussianity* problem consider the simple ICA model,

$$Y = A^{-1}\epsilon \,, \tag{1}$$

where Y is a $K \times 1$ vector, A is a $K \times K$ invertible matrix and ϵ is a $K \times 1$ vector that has independent components. The general approach for conducting inference on A is as follows: (i) *assume* that sufficiently many components of ϵ follow a non-Gaussian distribution, (ii) estimate A using likelihood-based methods or (generalized) method of moments, and (iii) construct confidence bands for some function of A based on the sampling variation of the estimator. Both parametric and semi-parametric estimators can be considered, see Chen and Bickel (2006) and Gouriéroux, Monfort and Renne (2017) for different examples.

The problem with this general approach occurs when the true densities are close to the Gaussian density. In such *weakly non-Gaussian* cases local identification deteriorates and coverage distortions occur. The root of the problem lies in the fact that the aforementioned inference approach is based on a binary treatment of non-Gaussianity, ignoring that what matters for correctly sized inference is the distance to the Gaussian distribution.[2]

[1] See for instance: Lanne and Lütkepohl (2010), Moneta et al. (2013), Lanne, Meitz and Saikkonen (2017), Maxand (2018), Lanne and Luoto (2019), Gouriéroux, Monfort and Renne (2017, 2019), Tank, Fox and Shojaie (2019), Herwartz (2019), Bekaert, Engstrom and Ermolov (2019, 2020), Fiorentini and Sentana (2020), Velasco (2020), Guay (2020) and Sims (2021).

[2] We note that several recent works have highlighted coverage distortions for weak non-Gaussian scenarios

To this extent, this paper develops a robust approach for conducting inference in LSEMs that is inspired by the identification robust methods developed in econometrics (e.g. Stock and Wright, 2000; Kleibergen, 2005; Andrews and Mikusheva, 2015) and the general semiparametric statistical theory that is discussed in Bickel et al. (1998) and van der Vaart (2002). In brief, we cast the LSEM as a semiparametric model, where the densities of the errors are treated non-parametrically, and we construct confidence bands for the possibly unidentified (Euclidean) parameters of interest by inverting semiparametric score tests. The approach efficiently exploits non-Gaussianity when it is present in the data and yields correct coverage regardless of the distance to the Gaussian distribution.

To structure our approach we start by providing a general and quite high level framework for conducting identification robust hypothesis tests in semiparametric models where the null hypothesis concerns a Euclidean parameter and there exists an infinite dimensional, but well identified, nuisance parameter. For this general framework our testing approach is characterized by two steps. In the first step an estimate for the efficient score function of the Euclidean parameter of interest is constructed and in the second step this estimate is used to construct a robust score statistic. This test statistic can be viewed as the semiparametric version of the Neyman-Rao score statistic and under appropriate assumptions it has a standard χ^2 limiting distribution regardless of whether the Euclidean parameter of interest is (well) identified.

Within our general framework we allow for singular efficient information matrices — the variance matrix of the efficient score function — as the semi-parametric models that we consider often have singularities at the points where identification fails.[3] For instance, we show that in model (1) the efficient information matrix becomes singular when more than one component of ϵ follows an exact Gaussian distribution.

Further, in our general framework we allow for both Euclidean and infinite dimensional nuisance parameters, and treat them separately. This allows the general theory to be stated with discretized versions of $\sqrt{n}$-consistently estimated Euclidean nuisance parameters directly plugged into the score statistic. This often greatly simplifies the application of the general theory to specific models as the high-level convergence requirements only need to be shown to hold along certain deterministic sequences.[4] This is demonstrated in our leading example of LSEMs with independent components.

The semi-parametric score tests that we propose control the size of the test regardless of whether the parameters of interest are identified. Moreover, under regularity conditions

in simulation exercises (e.g. Gouriéroux, Monfort and Renne, 2017; Lanne and Luoto, 2019).

[3]See Andrews and Guggenberger (2019) for examples of a similar phenomenon in a class of moment condition models.

[4]The discretization trick, due to Le Cam, is discussed in, among others, Le Cam and Yang (2000).

which include non-singularity of the efficient information matrix, the test belongs to the class of asymptotically uniformly most powerful invariant (AUMPI) tests (e.g. Choi, Hall and Schick, 1996).

With our general framework in hand, we turn to the class of LSEMs. We first provide a complete implementation for the simple ICA model (1). We start by casting the ICA model as a semiparametric model in which a Euclidean parameter α determines A and the densities of the components of ϵ form the non-parametric part. We make no functional form assumptions on the densities of the components of ϵ, requiring only certain moment conditions to hold. We analytically derive the efficient score function following Amari and Cardoso (1997) and show that it can be consistently estimated using the B-spline based log density score estimator of Jin (1992) and Chen and Bickel (2006). Based on the estimate of the efficient score function we can directly compute the score statistic which is shown to have a chi-squared limiting distribution. Importantly, this result does not assume any form of non-Gaussianity. In practice, computing the score statistic is simple and fast as it essentially only requires K regressions to estimate the log density scores, thus avoiding the usage of numerical optimization routines.

Next, we turn to the broader class of LSEMs which includes models with additional exogenous explanatory variables. Prominent members included in this class are the classical simultaneous equations model and several short T panel data models. The main restriction we impose is that the observations are independent across entities. Conceptually, our approach to inference is the same as for the baseline ICA model: (i) we cast the LSEM as a semi-parametric model, (ii) determine and estimate the efficient score functions and (iii) use these to compute the score test.

We evaluate the finite sample performance of the semiparametric score test in a large simulation study. We show that regardless of how close ϵ is to the Gaussian distribution our test is correctly sized. In contrast, tests that are based on the sampling variation of (pseudo)-maximum likelihood or GMM estimators have large size distortions in weakly non-Gaussian settings. Further, for moderate sample sizes the power of the semiparametric test is comparable to the parametric score test that relies on knowing the functional form of the density. When the parametric density of the (pseudo)-maximum likelihood score test is misspecified the semi-parametric test is always found preferable. This performance demonstrates that our asymptotic theory is a useful guide to finite sample performance.

To showcase the empirical value of our methodology we consider the estimation of the coefficients in a production function (e.g. Marschak and Andrews, 1944; Hoch, 1958; Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg, Caves and Frazer, 2015). In contrast to the more recent literature, we explicitly model the correlation between the error

term and the production function inputs (e.g. Hoch, 1958), and we exploit non-Gaussianity to identify the product function coefficients. We adopt this strategy for a large sample of manufacturing firms using both cross-sectional and panel data designs.

We find that this approach is able to accurately pin down the production function coefficients. We estimate the coefficient for labor between 0.4 and 0.8 and the coefficient for capital is between 0.2 and 0.5. These estimates are (i) robust across a variety of model specifications and (ii) vastly different from standard OLS estimates, potentially indicating a strongly endogenous relationship. Inspection of the model residuals is suggestive of non-Gaussianity, which explains the precision of the estimates we obtain.

The remainder of this paper is organized as follows. We complete the introduction by carefully relating our approach to the existing literature. In the next section we discuss a general framework for conducting identification robust tests in semiparametric models. Section 3 gives the implementation details and primitive assumptions for the LSEMs. Sections 4 and 5 summarize the results from the simulation and empirical studies. Section 6 concludes. Unless otherwise mentioned all proofs are provided in Appendix A. Any references to sections, equations, lemmas etc. which start with "S" refer to the supplementary material.

Relation to the literature

Our approach builds on three strands of literature: identification robust testing, semiparametric inference and the LSEM with independent non-Gaussian errors.

Regarding the weak identification robust literature, a useful analogy is obtained when comparing the non-Gaussian identification approach to an instrumental variable (IV) based identification approach. In textbook IV, identification is established theoretically by assuming that the covariance matrix between the instruments and the endogenous variables has full rank. In practice however, what matters for reliable standard inference is that the first stage (effective) F-statistic is larger then some threshold value, informally put, the correlation between the instruments and the endogenous variables should be sufficiently strong (e.g. Staiger and Stock, 1997; Stock and Yogo, 2005; Olea and Pflueger, 2013). In a similar way, in the LSEM non-Gaussianity can be viewed as a theoretical identification assumption (e.g. Comon, 1994; Hyvärinen, Karhunen and Oja, 2001), but what matters in practice is the distance to the Gaussian distribution. To avoid relying on the strict non-Gaussian identification assumption we consider test statistics whose asymptotic size does not depend on this assumption, similar in spirit to the identification robust tests that have been constructed for the IV problem which avoid explicitly relying on the covariance between instruments and the endogenous variables for inference (e.g. Anderson and Rubin, 1949; Staiger and Stock, 1997; Kleibergen, 2002).

More generally, the score testing approach of this paper is the semi-parametric equivalent of the Neyman-Rao test for parametric models (e.g. Hall and Mathiason, 1990). The latter have been shown to be robust to identification failures in, for instance, Andrews and Mikusheva (2015). Similar identification robust approaches have been developed for generalized moment models in Stock and Wright (2000); Kleibergen (2005); Andrews and Mikusheva (2016), among others. In the GMM context Andrews and Guggenberger (2019) provide an important extension that allows the variance matrix of the moments to be near singular or singular, see also Andrews (1987). We adopt a similar approach for constructing singularity robust tests in our setting.

The semiparametric literature in statistics has mainly focused on efficient estimation in well identified models Bickel et al. (1998) and van der Vaart (2002). A few papers focus on testing in well-identified semiparametric models (e.g. Choi, Hall and Schick, 1996; Bickel, Ritov and Stoker, 2006).

Finally, there exists a rich literature on ICA and LSEM models, and applications thereof (e.g. Dhrymes, 1994; Hyvärinen, Karhunen and Oja, 2001). This paper relates most closely to papers that treat the density functions of ϵ non-parametrically, see Bach and Jordan (2002), Samarov and Tsybakov (2004) and Chen and Bickel (2006). While the majority of the ICA literature has focused on efficient estimation under non-Gaussianity, recent works in econometrics have considered inference in such models (e.g. Gouriéroux, Monfort and Renne, 2017) and we contribute to this literature by developing inference methods robust to the failure of the assumption of non-Gaussianity.

As an alternative to our semi-parametric score approach one could imagine combining a higher order moment based approach as in Lanne and Luoto (2019) with robust GMM inference methodology as developed in Kleibergen (2005) for instance. The downsides of this approach are (i) typically a a large number of higher order moments need to be used, which is known to cause size distortions in existing weak identification robust methods (e.g. Andrews and Stock, 2007) and (ii) such an approach would not in general share the power optimality properties of the approach we consider.

2 Robust testing in semiparametric models

In this section we present a general approach for conducting identification and singularity robust hypothesis tests in semiparametric models. Our treatment is high-level and can be applied to a variety of models.

To outline the setting, consider the random vector $Y \in \mathcal{Y} \subset \mathbb{R}^K$ defined on some underlying probability space $(\Omega, \mathcal{F}, \mathrm{P})$ with its distribution on $\mathcal{Y}$ specified by the law P_{θ_0} that

depends on parameters $\theta_0 \in \Theta$. The parameter space Θ has the form $\Theta = \mathcal{A} \times \mathcal{B} \times \mathcal{H}$, where $\mathcal{A} \subset \mathbb{R}^{L_\alpha}$, $\mathcal{B} \subset \mathbb{R}^{L_\beta}$ and $\mathcal{H}$ a metric space. We write a typical element of Θ as $\theta = (\alpha, \beta, \eta)$, where it is understood that $\alpha \in \mathcal{A}$, $\beta \in \mathcal{B}$ and $\eta \in \mathcal{H}$.

The model that the researcher considers is the collection

$$\mathcal{P}_\Theta = \{P_\theta : \theta \in \Theta\} \ , \tag{2}$$

where each $P_\theta \ll \mu$ for some σ-finite measure μ on $\mathcal{Y}$. Typically, when $\mathcal{H}$ is finite dimensional we think of model (2) as parametric, whereas if $\mathcal{H}$ is infinite dimensional the model is classified as either non- or semi-parametric, see Bickel et al. (1998) and van der Vaart (2002) for textbook treatments. The model includes two types of Euclidean parameters: the parameters of interest α and the (Euclidean) nuisance parameters β. For future reference we define $\gamma = (\alpha, \beta)$ and $\Gamma = \mathcal{A} \times \mathcal{B}$, which implies that $\Gamma \subset \mathbb{R}^L$ with $L = L_\alpha + L_\beta$, and $P_\theta = P_{(\gamma,\eta)}$.

In general, we assume that the nuisance parameters β and η do not suffer from identification problems, but α may. In particular, for different points $\beta \in \mathcal{B}$ and $\eta \in \mathcal{H}$ the vector α may be strongly identified, weakly identified or completely unidentified. To conduct inference on α without making a priori assumptions on the identification of α we consider hypothesis tests of the form

$$H_0 : \alpha = \alpha_0 \ , \ \beta \in \mathcal{B} \ , \ \eta \in \mathcal{H} \qquad \text{against} \qquad H_1 : \alpha \neq \alpha_0 \ , \ \beta \in \mathcal{B} \ , \ \eta \in \mathcal{H} \ . \tag{3}$$

The main idea is to develop test statistics whose asymptotic size control is invariant to the identification strength of α. Such test statistics can then be inverted to yield confidence intervals for α with correct coverage.[5]

To derive our tests, we first define the scores of model (2) to be the quadratic mean derivatives of root-density paths.

Definition 1 (Cf. Definition 1.6 in van der Vaart, 2002)**.** *A* differentiable path *is a map $t \mapsto P_t$ from a neighbourhood $\mathscr{U}$ of $0 \in [0, \infty)$ to $\mathcal{P}_\Theta$ such that for some measurable function $s : \mathcal{Y} \to \mathbb{R}$, as $t \downarrow 0$,*

$$\int \left[\frac{\sqrt{p_t} - \sqrt{p}}{t} - \frac{1}{2} s \sqrt{p} \right]^2 \mathrm{d}\mu \to 0 \ , \tag{4}$$

where p_t and p respectively denote the densities of P_t and P relative to μ. Here s is the score function *of the submodel $\{P_t : t \in \mathscr{U}\}$ at $t = 0$.*

If we let $t \mapsto P_t$ range over a collection of submodels, indexed by $\mathcal{I}$, we will obtain a

[5] In parametric settings this approach is considered in Andrews and Mikusheva (2015) among others.

collection of score functions, say s_i for $i \in \mathcal{I}$. This collection, $\{s_i : i \in \mathcal{I}\}$, will be denoted by $\mathcal{T}_{P,\mathcal{I}}$ and as we only consider models with linear spaces we refer to it as a *tangent space*. For the semiparametric model (2) we define tangent spaces along restricted paths concerning the two parts of the parameter $\theta = (\gamma, \eta)$ separately.

Assumption 0. *The map $t \mapsto P_{\gamma+tg,\eta_t(\gamma,\eta,h)}$ is a differentiable path for each $(g,h) \in \mathbb{R}^L \times H =: \mathcal{J}$. The tangent space $\mathcal{T}_{P_\theta,\mathcal{J}}$ has the form*

$$\mathcal{T}_{P_\theta,\mathcal{J}} = \mathcal{T}^{\gamma|\eta}_{P_\theta,\mathbb{R}^L} + \mathcal{T}^{\eta|\gamma}_{P_\theta,H} \ , \tag{5}$$

where $\mathcal{T}^{\gamma|\eta}_{P_\theta,\mathbb{R}^L} = \{g'\dot{\ell}_\theta : g \in \mathbb{R}^L\}$, for $\dot{\ell}_\theta$ a L-vector of measurable functions from $\mathcal{Y} \to \mathbb{R}$, is the tangent space for γ and $\mathcal{T}^{\eta|\gamma}_{P_\theta,H}$ is the tangent space for η.

The assumption defines the tangent spaces for the semiparametric model (2) and imposes that the tangent space of the complete model is the sum of the tangent spaces of the parametric and non-parametric parts of the model. The assumption is mild and can typically be satisfied by imposing that the square root of the density function is continuously differentiable almost everywhere with respect to the parameters θ.[6]

For the parametric part of the model we note that $\dot{\ell}_\theta$ is simply the $L \times 1$ vector of scores of γ evaluated at $\theta = (\gamma, \eta)$, and the tangent space of γ is simply the span of $\dot{\ell}_\theta$. The tangent space of the non-parametric part, i.e. $\mathcal{T}^{\eta|\gamma}_{P_\theta,H}$, is formed by scores corresponding to paths of the form $t \mapsto P_{(\gamma,\eta_t(\gamma,\eta,h))}$ for $h \in H$, where the choice for $\eta_t(\gamma,\eta,h)$ depends on η such that $\eta_t(\gamma,\eta,h)|_{t=0} = \eta$.

Having defined the tangent spaces of γ and η, let Π_θ be the orthogonal projection from $L_2(P_\theta)$ onto the closure of $\mathcal{T}^{\eta|\gamma}_{P_\theta,H}$, i.e. $\operatorname{cl}\mathcal{T}^{\eta|\gamma}_{P_\theta,H}$. The *efficient score function* for γ is defined as (e.g. Definition 2.15 in van der Vaart, 2002)

$$\tilde{\ell}_\theta \coloneqq \dot{\ell}_\theta - \Pi_\theta \dot{\ell}_\theta \ , \tag{6}$$

where the projection is understood to apply componentwise. The accompanying *efficient information matrix* for γ is given by

$$\tilde{I}_\theta \coloneqq \mathbb{E}_\theta \tilde{\ell}_\theta \tilde{\ell}'_\theta \ . \tag{7}$$

When η is finite dimensional the efficient score is equivalent to the population residual of the regression of $\dot{\ell}_\theta$ on the scores of η and the efficient information matrix is the variance of this residual (e.g. Neyman, 1979; Choi, Hall and Schick, 1996).

[6] See e.g. Lemma 7.6 in van der Vaart (1998), Lemma 1.8 in van der Vaart (2002) or Proposition 2.1.1 in Bickel et al. (1998).

To obtain the efficient score function for α which is the part of $\gamma = (\alpha, \beta)$ that is of interest, note that the previous two displays imply the partitioning

$$\tilde{\ell}_\theta = \left(\tilde{\ell}'_{\theta,\alpha}, \tilde{\ell}'_{\theta,\beta}\right)' \quad \text{and} \quad \tilde{I}_\theta = \begin{bmatrix} \tilde{I}_{\theta,\alpha\alpha} & \tilde{I}_{\theta,\alpha\beta} \\ \tilde{I}_{\theta,\beta\alpha} & \tilde{I}_{\theta,\beta\beta} \end{bmatrix} . \tag{8}$$

If $\tilde{I}_{\theta,\beta\beta}$ is nonsingular,[7] we can (orthogonally) project once more to obtain the efficient score function for α:

$$\tilde{\kappa}_\theta \coloneqq \tilde{\ell}_{\theta,\alpha} - \tilde{I}_{\theta,\alpha\beta}\tilde{I}^{-1}_{\theta,\beta\beta}\tilde{\ell}_{\theta,\beta} , \tag{9}$$

which has corresponding efficient information matrix

$$\tilde{\mathcal{I}}_\theta \coloneqq \tilde{I}_{\theta,\alpha\alpha} - \tilde{I}_{\theta,\alpha\beta}\tilde{I}^{-1}_{\theta,\beta\beta}\tilde{I}_{\theta,\beta\alpha} . \tag{10}$$

Building tests or estimators based on the efficient score function $\tilde{\kappa}_\theta$ is attractive as efficiency results are well established, see Choi, Hall and Schick (1996), Bickel et al. (1998) and van der Vaart (2002). Note that an identical efficient score function $\tilde{\kappa}_\theta$ is obtained if one would project $\dot{\ell}_{\theta,\alpha}$ on the orthogonal complement of the tangent space of (β, η) in one step. We have purposely separated the steps (i.e. first projecting off η and then off β) to facilitate the analytical derivation of the efficient score.[8]

2.1 Semiparametric identification robust score test

Our interest lies in testing the null hypothesis (3) in a robust way that does not impose restrictions on the identification strength of α. From the previous section it follows that at $\theta_0 = (\alpha_0, \beta, \eta)$, where $\beta \in \mathcal{B}$ and $\eta \in \mathcal{H}$, we have

$$\mathbb{E}_{\theta_0}\tilde{\kappa}_{\theta_0} = 0 . \tag{11}$$

This implies that (11) defines a set of L_α moment conditions based on which we can construct hypothesis tests. See for instance Stock and Wright (2000) or Kleibergen (2005) for related approaches with finite dimensional nuisance parameters. Unlike these papers, the nuisance parameter in our model includes a Euclidean parameter and an infinite dimensional object.[9]

[7] If $\tilde{I}_{\theta,\beta\beta}$ is singular, we may drop components from $\tilde{\ell}_{\theta,\beta}$ until the remaining components form a linearly independent collection which span the same subspace of $L_2(P_\theta)$ as $\tilde{\ell}_{\theta,\beta}$. The corresponding variance matrix of this smaller vector will be non-singular and $\tilde{\ell}_{\theta,\beta}$ can be replaced throughout by this smaller vector.

[8] Cf. the discussion on p. 74 of Bickel et al. (1998).

[9] Andrews and Mikusheva (2016) also consider robust testing in models with an infinite dimensional nuisance parameter, however their approach is not directly applicable here as they assume the moment functions are known.

To construct test statistics we assume that we observe n independent and identically distributed copies of the vector Y that are denoted by $\{Y_i\}_{i=1}^n$. These observations satisfy the following high level assumption.

Assumption 1. *Let $\gamma_0 = (\alpha_0, \beta)$ and $\theta_0 = (\alpha_0, \beta, \eta)$ for any $(\beta, \eta) \in \mathcal{B} \times \mathcal{H}$. Additionally, let $\gamma_n = \{(\alpha_0, \beta_n)\}_{n\geq 1}$ be a deterministic sequence such that $\sqrt{n}(\gamma_n - \gamma_0) = O(1)$ and define $\theta_n = (\gamma_n, \eta)$ for each $n \in \mathbb{N}$. We have that*

1. *$\frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\ell}_{\theta_0}(Y_i) \rightsquigarrow Z \sim \mathcal{N}(0, \tilde{I}_{\theta_0})$ under P_{θ_0} where $\tilde{I}_{\theta_0}$ is nonsingular*
2. *We have an array of estimates $\{\hat{\ell}_{\theta_n}(Y_i)\}_{n\geq 1, i\leq n}$ such that:*

$$\frac{1}{n} \sum_{i=1}^n \left(\hat{\ell}_{\theta_n}(Y_i) - \tilde{\ell}_{\theta_n}(Y_i) \right) = o_{P_{\theta_n}}(n^{-1/2})$$

3. *$\hat{I}_{\theta_n} \xrightarrow{P_{\theta_n}} \tilde{I}_{\theta_0}$ for some sequence of estimates $\{\hat{I}_{\theta_n}\}_{n\geq 1}$*
4. *We have that*

$$\int \left\| \tilde{\ell}_{\theta_n} p_{\theta_n}^{1/2} - \tilde{\ell}_{\theta_0} p_{\theta_0}^{1/2} \right\|^2 \mathrm{d}\mu \to 0.$$

Clearly, Assumption 1 is high level and should be verified for any specific model of the form (2). Nevertheless, the strategy for verifying the different parts of the assumption is similar. In particular, part 1 amounts to verifying a central limit theorem for the efficient score function, which given the i.i.d. assumption requires only the existence of second moments.[10] Part 2 imposes that we should be able to construct a sequence of estimates for the efficient score functions, which in practice amounts to being able to estimate η or a function thereof sufficiently accurately. The third part imposes that the efficient information matrix can be consistently estimated. The final part is a continuity condition which is used (along with assumption 0) in the proof of the theorem below.[11]

Parts 2 and 3 concern convergence of functions which depend on the deterministic sequence $\theta_n = (\theta_0, \beta_n, \eta)$ under the corresponding sequence of measures P_{θ_n}. Permitting this convergence to be demonstrated under this sequence of measures typically simplifies verification of these conditions. The requirement that $\tilde{I}_{\theta_0}$ has full rank is relaxed below.

[10] In fact efficient score functions have finite second moments by construction and therefore automatically satisfy the required moment condition. We leave the weak convergence condition in the assumption as some of the results based on it do not rely on any other properties of efficient score functions and apply to any function satisfying these conditions. Additionally, extensions that allow for dependent observations can equally well be accommodated.

[11] See for instance Lemma 7.3 in van der Vaart (2002) or the proof of Theorem 25.57 in van der Vaart (1998).

Two important observations follow from Assumption 1. First, we do not model the identification strength for α. This is not required as we impose that $\alpha = \alpha_0$ under H_0 in the construction of our test statistic. Second, we effectively do require that η — or at least some aspect of it — is strongly identified as typically $\tilde{\ell}_{\theta_0}$ depends on this parameter and we impose that $\tilde{\ell}_{\theta_0}$ can be $\sqrt{n}$-consistently estimated.

Based on Assumption 1-part 2 we define the following estimators for the efficient score and information matrix for α:

$$\hat{\kappa}_\theta \coloneqq \hat{\ell}_{\theta,\alpha} - \hat{I}_{\theta,\alpha\beta}\hat{I}^{-1}_{\theta,\beta\beta}\hat{\ell}_{\theta,\beta} \ , \qquad \text{and} \qquad \hat{\mathcal{I}}_\theta \coloneqq \hat{I}_{\theta,\alpha\alpha} - \hat{I}_{\theta,\alpha\beta}\hat{I}^{-1}_{\theta,\beta\beta}\hat{I}_{\theta,\beta\alpha} \ . \tag{12}$$

With these estimates we can test the null hypothesis (3) using the efficient score statistic

$$\hat{S}_\theta \coloneqq \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\kappa}_\theta(Y_i)\right)' \hat{\mathcal{I}}_\theta^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\kappa}_\theta(Y_i)\right) . \tag{13}$$

For parametric models this score statistic reduces to Neyman's $C(\alpha)$ statistic (e.g. Neyman, 1979), which is asymptotically equivalent to Rao's score statistic when certain regularity conditions hold and maximum likelihood estimates are used for the nuisance parameters (e.g. Kocherlakota and Kocherlakota, 1991).

The limiting distribution of the efficient score statistic is summarized in the following theorem.

Theorem 1. *Let $\theta_0 = (\alpha_0, \beta, \eta)$ for any $(\beta, \eta) \in \mathcal{B} \times \mathcal{H}$. Suppose that $\hat{\beta}_n$ is a $\sqrt{n}$-consistent estimator of β under H_0. Let $B_n = n^{-1/2}C\mathbb{Z}^{L_\beta}$ for some $C > 0$ and let $\bar{\beta}_n$ be a discretised version of $\hat{\beta}_n$ which replaces its value with the closest point in B_n. Suppose assumptions 0 and 1 hold, and let $\bar{\theta}_n = (\alpha_0, \bar{\beta}_n, \eta)$. Then, under H_0 we have*

$$\hat{S}_{\bar{\theta}_n} \rightsquigarrow \chi^2_{L_\alpha} \ .$$

The theorem implies that, regardless of whether α is well identified, the score static $\hat{S}_{\bar{\theta}_n}$ has a standard χ^2 limiting distribution under the null. Confidence regions for α can be obtained by inverting $\hat{S}_{\bar{\theta}_n}$ over a grid of values for α. By construction such confidence regions will have correct coverage.

This theorem demonstrates that under the one can effectively "plug-in" discretised $\sqrt{n}$-consistent estimators of the well-identified Euclidean nuisance parameters in the efficient score statistic and obtain the usual χ^2 limiting distribution under the null hypothesis. The discretisation is a technical device which permits us to require only convergence along non-random sequences in assumption 1. This construction often dramatically simplifies the ap-

plication of Theorem 1 as the the high-level conditions in Assumption 1 only need to be verified for deterministic sequences. See, for example, the discussion in Le Cam and Yang (2000, p. 125) or van der Vaart (1998, p. 72-73).

It follows from Choi, Hall and Schick (1996) that tests based on $\hat{S}_{\bar{\theta}_n}$ are asymptotically uniformly most powerful within the class of rotation invariant tests (when $L = 1$ the rotational invariance can be dropped for one-sided tests and replaced with unbiasedness for two-sided tests). This implies that asymptotically when testing the hypothesis (3), the power of the test is the greatest possible in the class of rotationally invariant tests. This makes tests based on $\hat{S}_{\bar{\theta}_n}$ attractive for scenarios where there is no explicit direction in which one want to maximize power. When such directions are given alternative test statistics, also based on the efficient score function, can be considered (e.g. Bickel, Ritov and Stoker, 2006).

The identification robust test statistic $\hat{S}_{\bar{\theta}_n}$ is broadly applicable for the class of semiparametric models we consider. The key difficulty for its application lies in the construction and estimation of the efficient score function $\tilde{\ell}_\theta$. For this no general recipe exists but guidance and examples are given in Bickel et al. (1998), van der Vaart (1998) and Rabinowitz (2000).

Besides the LSEM model with non-Gaussian distributions a variety of other models can be cast within our general framework. Prominent examples include, instrumental variable models (e.g. Cattaneo, Crump and Jansson, 2012), mixed-proportional hazard models (e.g. Hahn, 1994) and single index models (e.g. Horowitz, 2009). In each of these models the parameter of interest could become weakly/not-identified depending on the value of (infinite dimensional) nuisance parameters, and pending the verification of Assumption 1, the robust score test can be used to conduct inference.

2.2 Semiparametric identification and singularity robust score test

In this section we extend the main result from the previous section to allow for singular information matrices. This is an important extension motivated by the fact that many models that suffer from identification problems will have a singular information matrices for certain values of the nuisance parameters, see also Andrews and Guggenberger (2019) for examples in a class of moment condition models. A leading example in our setting is the ICA model, where if more then one component of ϵ follows an exact Gaussian distribution the efficient information matrix may be singular.[12]

To allow for singular efficient information matrices in our general theory we modify assumption 1 as follows.

[12]See Lemma S1 in the supplementary material for a proof of this fact.

Assumption 2. *Let $\gamma_0 = (\alpha_0, \beta)$ and $\theta_0 = (\alpha_0, \beta, \eta)$ for any $(\beta, \eta) \in \mathcal{B} \times \mathcal{H}$. Additionally, let $\gamma_n = \{(\alpha_0, \beta_n)\}_{n\in\mathbb{N}}$ be a deterministic sequence such that $\sqrt{n}(\gamma_n - \gamma_0) = O(1)$ and define $\theta_n = (\gamma_n, \eta)$ for each $n \in \mathbb{N}$. Suppose that*

1. *$\frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\ell}_{\theta_0}(Y_i) \rightsquigarrow Z \sim \mathcal{N}(0, \tilde{I}_{\theta_0})$ under P_{θ_0} where $\tilde{I}_{\theta_0,\beta\beta}$ is nonsingular*
2. *We have an array of estimates $\{\hat{\ell}_{\theta_n}(Y_i)\}_{n\geq 1, i\leq n}$ such that:*

$$\frac{1}{n}\sum_{i=1}^n \left(\hat{\ell}_{\theta_n}(Y_i) - \tilde{\ell}_{\theta_n}(Y_i)\right) = o_{P_{\theta_n}}(n^{-1/2})$$

3. *For some sequence of estimates $\{\hat{I}_{\theta_0}\}_{n\geq 1}$ and some sequence $\{\nu_n\}_{n\geq 1}$ with $0 \leq \nu_n \to 0$*

$$\|\hat{I}_{\theta_n} - \tilde{I}_{\theta_0}\|_2 = o_{P_{\theta_n}}(\nu_n)$$

4. *We have that*

$$\int \left\| \tilde{\ell}_{\theta_n} p_{\theta_n}^{1/2} - \tilde{\ell}_{\theta_0} p_{\theta_0}^{1/2} \right\|^2 \mathrm{d}\mu \to 0.$$

This modified assumption allows the limiting distribution of the re-scaled sum of efficient scores to have a singular variance matrix. However, note that we continue to impose that β and η are well identified, as for instance the efficient information matrix corresponding to β is nonsingular as part 1 requires. Part 3 imposes that we can determine a convergence rate for our estimate of the efficient information matrix for γ.

We take $\hat{\kappa}_\theta$ and $\hat{\mathcal{I}}_\theta$ as in equation (12) and, given ν_n, we define a truncated eigenvalue version of the information matrix estimate as

$$\hat{\mathcal{I}}_\theta^t = \hat{U}_n \hat{\Lambda}_n(\nu_n) \hat{U}_n' \,, \tag{14}$$

where $\hat{\Lambda}_n(\nu_n)$ is a diagonal matrix with the ν_n-truncated eigenvalues of $\hat{\mathcal{I}}_\theta$ on the main diagonal and $\hat{U}_n$ is the matrix of corresponding orthonormal eigenvectors. To be specific, let $\{\hat{\lambda}_{n,i}\}_{i=1}^L$ denote the non-increasing eigenvalues of $\hat{\mathcal{I}}_\theta$, then the (i,i)th element of $\hat{\Lambda}_n(\nu_n)$ is given by $\hat{\lambda}_{n,i}\mathbf{1}(\hat{\lambda}_{n,i} \geq \nu_n)$.

Based on this we define the singularity and identification robust score statistic as follows.

$$\hat{S}_\theta^{SR} \coloneqq \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\kappa}_\theta(Y_i)\right)' \hat{\mathcal{I}}_\theta^{t,\dagger} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\kappa}_\theta(Y_i)\right). \tag{15}$$

where $\hat{\mathcal{I}}_\theta^{t,\dagger}$ is the Moore-Penrose psuedo-inverse of $\hat{\mathcal{I}}_\theta^t$. The limiting distribution of $\hat{S}_\theta^{SR}$ is

characterized in the following theorem, which implies that we can use the estimated rank of $\hat{\mathcal{I}}_\theta^t$ to compute the critical value for $\hat{S}_\theta^{SR}$.

Theorem 2. *Let $\theta_0 = (\alpha_0, \beta, \eta)$ for any $(\beta, \eta) \in \mathcal{B} \times \mathcal{H}$. Suppose that $\hat{\beta}_n$ is a $\sqrt{n}$-consistent estimator of β under P_{θ_0}. Let $B_n = n^{-1/2} C \mathbb{Z}^{L_\beta}$ for some $C > 0$ and let $\bar{\beta}_n$ be a discretised version of $\hat{\beta}_n$ which replaces its value with the closest point in B_n. Suppose assumptions 0 and 2 hold and let $\bar{\theta}_n = (\alpha_0, \bar{\beta}_n, \eta)$. Let $r_n = \operatorname{rank}(\hat{\mathcal{I}}_{\bar{\theta}_n}^t)$ and denote by c_n the $1-a$ quantile of the $\chi^2_{r_n}$ distribution for any $a \in (0,1)$.*[13] *Then*

$$\lim_{n\to\infty} P_{\theta_0}\left(\hat{S}_{\bar{\theta}_n}^{SR} > c_n\right) \leq a,$$

with inequality only if $\operatorname{rank}(\tilde{\mathcal{I}}_{\theta_0}) = 0$.

If we use the singular robust test statistic while the true information matrix is non-singular the asymptotic consequences are negligible as following lemma shows.

Lemma 1. *Suppose assumptions 0 and 1 hold. Then*

$$\hat{S}_{\bar{\theta}_n}^{SR} = \hat{S}_{\bar{\theta}_n} + o_{P_{\theta_0}}(1) \ .$$

Therefore the singularity robust score statistic $\hat{S}_{\bar{\theta}_n}^{SR}$ can be adopted for both singular and non-singular information matrices. Moreover, for the case where $r = L$ the optimality properties of $\hat{S}_{\bar{\theta}_n}$ carry over to $\hat{S}_{\bar{\theta}_n}^{SR}$.

3 Robust non-Gaussian inference

In this section we provide details on how to use the high level framework from the previous section to conduct inference on parameters in LSEMs. We start with the details for the baseline ICA model and then consider a more general class of LSEMs. We present the results only for the case where we allow for singular information matrices as Lemma 1 shows that the non-singular score test is asymptotically equivalent under non-singularity.

3.1 Semi-parametric ICA model

For convenience we restate the ICA model

$$Y = A^{-1}\epsilon \ . \tag{16}$$

[13] If $r_n = 0$ we take $c_n = 0$.

We start by casting model (16) as a semiparametric model as defined in general in equation (2), see also Amari and Cardoso (1997) and Chen and Bickel (2006). We consider A to be determined by a Euclidean vector, whereas the infinite dimensional nuisance parameters are the unknown density functions of the components of ϵ. That is $\gamma = (\alpha, \beta)$ controls $A = A(\gamma)$ and $\eta = (\eta_1, \ldots, \eta_K)$, where η_k denotes the density of ϵ_k. The following two examples illustrate possible parametrizations for $A(\gamma)$ that are commonly used in practice.

Example 1 (Rotation matrix)**.** *Let $A^{-1} = \Sigma^{1/2}R$, where $\Sigma^{1/2}$ is lower triangular and R is a rotation matrix. In this setting we can take $\beta = \text{vech}(\Sigma^{1/2})$ and α parametrizes R using the trigonometric transformation or the Cayley or exponential transformation of a skew-symmetric matrix (e.g. Gouriéroux, Monfort and Renne, 2017; Magnus, Pijls and Sentana, 2020). We note that β can be consistently recovered from the variance of Y and a confidence region for α can be obtained by inverting the semi-parametric score test.*

Example 2 (Supply and demand)**.** *Let Y_1 denote the quantity of some good and Y_2 its price. A simple model is given by*

$$\begin{aligned} Y_1^d &= aY_2 + \sigma_1\epsilon_1 \qquad \textit{(demand)} \\ Y_1^s &= bY_2 + \sigma_2\epsilon_2 \qquad \textit{(supply)} \end{aligned}$$

where ϵ_1 and ϵ_2 are independent demand and supply shocks, and in equilibrium we have $Y_1^d = Y_1^s$. We can accommodate this set up by letting $\alpha = (a, b)$, $\beta = (\sigma_1, \sigma_2)$ and defining the mapping $A(\gamma)$ according to

$$A(\gamma) = \begin{bmatrix} \sigma_1^{-1} & 0 \\ 0 & \sigma_2^{-1} \end{bmatrix} \begin{bmatrix} 1 & -a \\ 1 & -b \end{bmatrix}$$

We may compute the confidence region for the demand and supply elasticities $\alpha = (a, b)$ by inverting the score test over the region where a is negative and b positive as economic theory would suggest.

These examples illustrate different possible ways that $A(\gamma)$ may be parametrized. The key restriction is that β should be consistently estimable, which can be ensured by verifying that β can be recovered from the variance of Y. In the remainder of this section we leave the precise parameter mapping $A(\gamma)$ unspecified, up to some smoothness conditions imposed later on.

The nuisance parameters $\eta = (\eta_1, \ldots, \eta_k)$ correspond to the density functions of $\epsilon = (\epsilon_1, \ldots, \epsilon_k)'$ and while we do not impose any parametric form for the density functions, we will place a number of restrictions on the moments of (functions of) ϵ.

Assumption 3. *For $\epsilon = (\epsilon_1, \ldots, \epsilon_K)'$ in model (16), each component ϵ_k has a continuously differentiable root density (where the density is with respect to Lebesgue measure on $\mathbb{R}$). We write the density as η_k with log density score $\phi_k(x) = \partial \log \eta_k(x)/\partial x$. We assume that for all $k = 1, \ldots, K$ and some $\delta > 0$*

1. *$\mathbb{E}\epsilon_k = 0$, $\mathbb{E}\epsilon_k^2 = 1$, $\mathbb{E}\epsilon_k^{4+\delta} < \infty$, $\mathbb{E}(\epsilon_k^4) - 1 > \mathbb{E}(\epsilon_k^3)^2$, and $\mathbb{E}\phi_k^{4+\delta}(\epsilon_k) < \infty$;*
2. *$\mathbb{E}\phi_k(\epsilon_k) = 0$, $\mathbb{E}\phi_k(\epsilon_k)\epsilon_k = -1$, $\mathbb{E}\phi_k(\epsilon_k)\epsilon_k^2 = 0$ and $\mathbb{E}\phi_k(\epsilon_k)\epsilon_k^3 = -3$;*
3. *ϵ_k is independent of ϵ_j for all $k \neq j$.*

The first part normalizes the errors to have mean zero, variance one and finite four+δ moments.[14] Additionally, we require the log density scores $\phi_k(x) = \partial \log \eta_k(x)/\partial x$ evaluated at the errors to have finite four+δ moments. The second part simplifies the construction of the efficient score functions. Whilst this may at first glance appear a strong condition, lemma S8 shows that if the first part holds, then a simple sufficient condition is that the tails of the densities η_k converge to zero at a polynomial rate.[15]

Most important is what is *not* in Assumption 3: there is no condition that imposes that a certain number of components of ϵ have a (sufficiently) non-Gaussian distribution. As a result, our testing approach retains correct size regardless of the true distributions of ϵ, i.e. regardless of the distance to the Gaussian distribution.

To define the parameter space for our semi-parametric model, let $\mathscr{H}$ be given by

$$\mathscr{H} := \Bigg\{ g \in L_1(\lambda) \cap \mathcal{C}^1(\lambda) : g(z) \geq 0, \int g(z)\,\mathrm{d}z = 1, \int z g(z) dz = 0, \int \kappa(z) g(z)\,\mathrm{d}z = 0, \int |z|^{4+\delta} g(z)\,\mathrm{d}z < \infty, \int \left|(g'(z)/g(z))\right|^{4+\delta} g(z)\,\mathrm{d}z < \infty, \int z^4 g(z)\,\mathrm{d}z > 1 + \left[\int z^3 g(z)\,\mathrm{d}z\right]^2 \Bigg\},$$

where λ denotes Lebesgue measure on $\mathbb{R}$, $\mathcal{C}^1(\lambda)$ is the class of real functions on $\mathbb{R}$ which are continuously differentiable λ-a.e. and $\kappa(z) = z^2 - 1$. Let $\mathcal{H} := \prod_{k=1}^K \mathscr{H}$. The semiparametric ICA model we consider is given by $\mathcal{P}_\Theta := \{P_\theta : \theta \in \Theta\}$ with $\Theta := \Gamma \times \mathcal{H}$ and P_θ being the

[14] $\mathbb{E}(\epsilon_k^4) - 1 \geq \mathbb{E}(\epsilon_k^3)^2$ always holds; this is known as Pearson's inequality. See e.g. result 1 in Sen (2012). Assuming that $\mathbb{E}(\epsilon_k^4) - 1 > \mathbb{E}(\epsilon_k^3)^2$ rules out (only) cases where $1, \epsilon_k$ and ϵ_k^2 are linearly dependent when considered as elements of L_2. See e.g. Theorem 7.2.10 in Horn and Johnson (2013).

[15] See example S1 in the supplementary material for an explicit example of a density which satisfies the first part of the assumption but not the second.

law on $\mathbb{R}^K$ defined by the density

$$p_\theta(y) := |\det A(\gamma)| \prod_{k=1}^{K} \eta_k(A_{k\bullet} y) \ , \tag{17}$$

where $A_{k\bullet}$ denotes the kth row of $A(\gamma)$.

Let $\mathcal{H}_0 \subset \mathcal{H}$ denote the set with elements $\eta = (\eta_1, \ldots, \eta_K)$ such that each η_k satisfies the requirements imposed by assumption 3. To implement the score test we first characterize the efficient score function for γ, i.e. equation (6), in terms of estimable quantities. The following lemma provides the key result.[16]

Lemma 2. *Given Assumption 3, if $\gamma \mapsto A(\gamma)$ is continuously differentiable, the components of the efficient score function for γ in the semiparametric ICA model $\mathcal{P}_\Theta$ at any $\theta = (\gamma, \eta)$ with $\eta \in \mathcal{H}_0$ are given by, for $l = 1, \ldots, L$,*

$$\tilde{\ell}_{\theta,l}(y) = \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \zeta_{l,k,j} \phi_k(A_{k\bullet} y) A_{j\bullet} y + \sum_{k=1}^{K} \zeta_{l,k,k} \left[\tau_{k,1} A_{k\bullet} y + \tau_{k,2} \kappa(A_{k\bullet} y)\right],$$

where $\zeta_{l,k,j} := [D_l(\gamma)]_{k\bullet} A^{-1}_{\bullet j}$ with $D_l(\gamma) = \partial A(\gamma)/\partial \gamma_l$, $A^{-1}_{\bullet j}$ is the j-th column of $A(\gamma)^{-1}$ and

$$\tau_k := M_k^{-1} \begin{pmatrix} 0 \\ -2 \end{pmatrix}, \quad \text{where } M_k := \begin{pmatrix} 1 & \mathbb{E}_\theta(A_{k\bullet} y)^3 \\ \mathbb{E}_\theta(A_{k\bullet} y)^3 & \mathbb{E}_\theta(A_{k\bullet} y)^4 - 1 \end{pmatrix}.$$

Lemma 2 is a essentially a special case of Lemma 3 given below for which the proof can be found in the supplementary material. It requires first defining the tangent spaces for γ and η, and then computing the orthogonal projection of the scores for γ on the tangent space for η, see equation (6).

3.2 Non-Gaussian robust score test

Next, to conduct inference on A we consider testing $H_0 : \alpha = \alpha_0$, $(\beta, \eta) \in \mathcal{B} \times \mathcal{H}$ using the identification and singularity robust score statistic given in (15). To compute this test statistic we require an estimate for the efficient score function $\tilde{\ell}_{\theta_0}$ as defined in Lemma 2. This can be done by estimating τ_k and the log density scores ϕ_k for each $k = 1, \ldots, K$. Note that the remaining elements of the efficient score are fixed under H_0.

The estimation of τ_k follows easily by replacing the population moments in its definition

[16] Strictly speaking, the efficient score function is defined relative to a specific tangent set, denoted here by $\mathcal{T}^{\eta|\gamma}_{P_\theta, H}$; see e.g. the discussion in sections 1.2 and 2.2 of van der Vaart (2002).

by their sample counterparts. In particular, we have

$$\hat{\tau}_{k,n} \coloneqq \hat{M}_{k,n}^{-1} \begin{pmatrix} 0 \\ -2 \end{pmatrix}, \quad \text{where } \hat{M}_{k,n} \coloneqq \begin{pmatrix} 1 & \frac{1}{n}\sum_{i=1}^{n}(A_{k\bullet}Y_i)^3 \\ \frac{1}{n}\sum_{i=1}^{n}(A_{k\bullet}Y_i)^3 & \frac{1}{n}\sum_{i=1}^{n}(A_{k\bullet}Y_i)^4 - 1 \end{pmatrix}, \tag{18}$$

where $A = A(\alpha_0, \beta)$ for a given $\beta \in \mathcal{B}$.

The estimation of the log density scores is typically more involved and a variety of options exist. We proceed by stating the requirements that must hold for any density score estimator and we show in Appendix B that the method of Chen and Bickel (2006), who build on Jin (1992), satisfies the requirements under mild conditions. This approach is convenient for two reasons: first the method of Chen and Bickel (2006) is based on B-spline approximations and while easy to implement it is notationally somewhat cumbersome, second different researchers might prefer to use a different density score estimator.

Assumption 4. *Let $\{\beta_n\}_{n\geq 1}$ be a deterministic sequence in $\mathcal{B}$ with $\sqrt{n}(\beta_n - \beta) = O(1)$ and let $\theta_n = (\alpha_0, \beta_n, \eta)$ for some $\eta \in \mathcal{H}_0$ and suppose we have an array of estimates $\{\hat{\phi}_{k,n}(A_{n,k\bullet}Y_i)\}_{n\geq 1, i\leq n}$ for $k = 1, \ldots, K$ where $A_n = A(\alpha_0, \beta_n)$ such that*

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(A_{n,k\bullet}Y_i) - \phi_k(A_{n,k\bullet}Y_i)\right] W_{i,n} = o_{P_{\theta_n}}(n^{-1/2}), \tag{19}$$

and for $\nu_{n,p}^2 = o(\nu_n)$ with $p \coloneqq \min\{1 + \delta/4, 2\}$ and $\nu_{n,p} = n^{(1-p)/p}$ if $p \in (1,2)$ or $\nu_{n,p} = n^{-1/2}\log(n)^{1/2+\rho}$, for some $\rho > 0$, if $p = 2$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\left(\left[\hat{\phi}_{k,n}(A_{n,k\bullet}Y_i) - \phi_k(A_{n,k\bullet}Y_i)\right] W_{i,n}\right)^2 = o_{P_{\theta_n}}(\nu_n). \tag{20}$$

where $\{W_{i,n}\}_{n\geq 1, i\leq n}$ is such that for each $n \in \mathbb{N}$, under P_{θ_n}, the $W_{i,n}$ are i.i.d. with marginal distribution given by G_w, with zero-mean, finite second moments and independent of each $A_{n,k}Y_j$.

The assumption effectively requires a specific function, i.e. ϕ_k, of the nuisance parameters η_k to be estimable sufficiently accurately. For the results in this section we only require the assumption to hold for the special case where $W_{i,n} = A_{n,j\bullet}Y_i$, but in the next section some other choices for $W_{i,n}$ are required. We note that an "upper bound" $\nu_{n,p}^2$ for the rate ν_n is now made explicit and it is split into two parts. The "slow" rate $n^{(1-p)/p}$ (for $p \in (1,2)$) is always sufficient given assumption 3, but if ϵ_k has finite eighth moments the faster rate applies. In appendix B we provide the conditions under which the density score estimator of Jin (1992) and Chen and Bickel (2006) satisfies this assumption and further conditions the

rate ν_n must satisfy; see Proposition 3.

Given any $\beta_n \in \mathcal{B}$, the estimates for τ_k and the density scores ϕ_k (both based on $A_n = A(\alpha_0, \beta_n)$) we can estimate the efficient score function under H_0 by

$$\hat{\ell}_{\theta_n,l}(y) = \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \zeta_{l,k,j} \hat{\phi}_{k,n}(A_{n,k\bullet} y) A_{n,j\bullet} y + \sum_{k=1}^{K} \zeta_{l,k,k} \left[\hat{\tau}_{k,1,n} A_{n,k\bullet} y + \hat{\tau}_{k,2,n} \kappa(A_{n,k\bullet} y)\right] , \quad (21)$$

where, compared to Lemma 2, τ_k and ϕ_k have been replaced by their estimates and $\theta_n = (\alpha_0, \beta_n, \eta)$. We also define an estimate of the corresponding efficient information matrix:

$$\hat{I}_{\theta_n} = \frac{1}{n} \sum_{i=1}^{n} \hat{\ell}_{\theta_n}(Y_i) \hat{\ell}_{\theta_n}(Y_i)' , \quad (22)$$

We now state our main result.

Proposition 1. *Let $\hat{\beta}_n$ be a $\sqrt{n}-$consistent estimator of β and let $\bar{\beta}_n$ denote a discretised version as in Theorem 2. Define $\bar{\theta}_n = (\alpha_0, \bar{\beta}_n, \eta)$ and consider the statistic*

$$\hat{S}^{SR}_{\bar{\theta}_n} = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\kappa}_{\bar{\theta}_n}(Y_i) \right)' \hat{\mathcal{I}}^{t,\dagger}_{\bar{\theta}_n} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\kappa}_{\bar{\theta}_n}(Y_i) \right) ,$$

with $\hat{\kappa}_{\bar{\theta}_n}$, $\hat{\mathcal{I}}_{\bar{\theta}_n}$ as in equation (12) based on $\hat{\ell}_{\bar{\theta}_n}(Y_i)$, $\hat{I}_{\bar{\theta}_n}$ as defined according to equations (21) and (22) and $\hat{\mathcal{I}}^{t,\dagger}_{\bar{\theta}_n}$ the Moore-Penrose inverse of the truncated version of $\hat{\mathcal{I}}^{t}_{\bar{\theta}_n}$ defined analogously to equation (14) with truncation level $\nu_n^{1/2}$. Suppose that assumptions 3 and 4 hold, that $\tilde{I}_{\theta_0,\beta\beta}$ is nonsingular, $\gamma \mapsto A(\gamma)$ is continuously differentiable and the maps $\gamma \to [D_l(\gamma)]_{k\bullet} A(\gamma)^{-1}_{\bullet j}$ are Lipschitz continuous. Let $r_n = \operatorname{rank}(\hat{I}^{t}_{\bar{\theta}_n})$ and denote by c_n the $1-a$ quantile of the $\chi^2_{r_n}$ distribution, for any $a \in (0,1)$.[17] *Then*

$$\lim_{n\to\infty} P_{\theta_0}(\check{S}^{SR}_{\bar{\theta}_n} > c_n) \leq a,$$

with inequality only if $\operatorname{rank}(\tilde{\mathcal{I}}_{\theta_0}) = 0$.

The proof of proposition 1 amounts to verifying the high level conditions stated in assumptions 0 and 2 so that we can apply Theorem 2.

Some comments are in order. First, depending on the parametrization $A(\gamma)$ a suitable estimator for β needs to be selected. This is often easy as β should be recoverable from the variance of Y which can be consistently estimated given assumption 3. Second, tests based on the efficient score statistic (as here) are shown to have various power optimality

[17] If $r_n = 0$ we take $c_n = 0$.

properties under non-singularity (e.g. Choi, Hall and Schick, 1996). These results carry over by Lemma 1. Third, given an $\sqrt{n}$-consistent estimator for β, $\hat{S}_{\bar{\theta}_n}$ is almost trivial to compute as it requires only K regressions to obtain the density score estimates using B-splines, thus avoiding numerical optimization routines entirely. Fourth, the Lipschitz continuity is satisfied for the mappings $A(\gamma)$ discussed in the examples 1 and 2 above and is typically not difficult to establish.

3.3 Extensions for including covariates

This section extends the semiparametric robust score test for a general class of LSEMs that includes covariates. In particular, let $Y = (Z, \tilde{X})$ denote the observable variables that are related by

$$Z = BX + V \ , \qquad V = A^{-1}\epsilon \ , \tag{23}$$

where Z is the dependent variable in $\mathbb{R}^K$, $X = (1, \tilde{X}')'$ is a random vector of explanatory variables in $\mathbb{R}^d$ and V plays the role of Y from the previous section. The additional parameters in this model are the entries of the $K \times d$ coefficient matrix B. The LSEM can be cast as a semi-parametric model when we take $A = A(\alpha, \beta_1)$, $\beta = (\beta_1, b)$, with $b = \text{vec}(B)$, and $\eta = (\eta_0, \eta_1, \ldots, \eta_K)$ includes the densities of X and ϵ. We continue to impose that $A(\alpha, \beta_1)$ should be parametrized such that $A(\alpha, \beta_1)$ is invertible and continuously differentiable with respect to (α, β_1).

The semiparametric LSEM is given by $\mathcal{P}_\Theta := \{P_\theta : \theta \in \Theta\}$ with $\Theta = \mathcal{A} \times \mathcal{B} \times \mathcal{H}$. $\mathcal{A}$ is identical to the previous section. $\mathcal{B}$ now additionally includes the nuisance parameters b and $\mathcal{H} = \mathscr{Z} \times \prod_{k=1}^K \mathscr{H}$, where $\mathscr{Z}$ is the space of density functions η_0 with $X \sim \eta_0$. We emphasize that alternative ways of parametrizing the semiparametric LSEM are also possible.

Our main interest is in testing $H_0 : \alpha = \alpha_0, \ (\beta, \eta) \in \mathcal{B} \times \mathcal{H}$. The following assumption allows us to derive the efficient score function for $\gamma = (\alpha, \beta)$.

Assumption 5. *For model (23) we have (for some $\delta > 0$)*

1. *ϵ satisfies Assumption 3,*
2. *Each $\eta_0 \in \mathscr{Z}$ is a density function (with respect to Lebesgue measure on $\mathbb{R}^{d-1}$) such that if $\tilde{X} \sim \eta_0$, then $\mathbb{E}\tilde{X}\tilde{X}'$ is positive definite and $\mathbb{E}[|\tilde{X}_l|^{4+\delta}] < \infty$ for all $l = 1, \ldots, d-1$,*
3. *ϵ and $\tilde{X}$ are independent,*

Part 1 imposes that ϵ satisfies the same moment assumptions as considered in the baseline ICA model. Part 2 imposes some structure on Y that allows us to identify B. Part 3 requires

the explanatory variables and errors to be independent.[18]

The following lemma provides the efficient score function for γ.

Lemma 3. *Given Assumption 5, if $(\alpha, \beta_1) \mapsto A(\alpha, \beta_1)$ is continuously differentiable, the components of $\tilde{\ell}_\theta$ in the semiparametric linear simultaneous equations model $\mathcal{P}_\Theta$ at any $\theta = (\gamma, \eta)$ with $\gamma = (\alpha, \beta)$, $\beta = (\beta_1, b) \in \mathcal{B}$ and $\eta \in \mathcal{H}_0$ are given by*

$$\tilde{\ell}_{\theta,(\alpha,\beta_1),l}(y) = \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \zeta_{l,k,j} \phi_k(A_{k\bullet} v) A_{j\bullet} v + \sum_{k=1}^{K} \zeta_{l,k,k} \left[\tau_{k,1} A_{k\bullet} v + \tau_{k,2} \kappa(A_{k\bullet} v)\right]$$

$$\tilde{\ell}_{\theta,b,m}(y) = \sum_{k=1}^{K} [-A_{k\bullet} D_{b,m}] \left[(x - \mathbb{E}x)\phi_k(A_{k\bullet} v) - \mathbb{E}x \left(\varsigma_{k,1} A_{k\bullet} v + \varsigma_{k,2} \kappa(A_{k\bullet} v)\right)\right]$$

for $l = 1, \ldots, L_\alpha + \dim(\beta_1)$ and $m = 1, \ldots, \dim(b)$, with $v = z - Bx$, $\zeta_{l,k,j} := [D_l(\alpha, \beta_1)]_{k\bullet} A^{-1}_{\bullet j}$ with $D_l(\alpha, \beta_1) = \partial A(\alpha, \beta_1)/\partial(\alpha, \beta_1)_l$, $D_{b,l} = \partial B / \partial b_l$ and

$$\tau_k := M_k^{-1} \begin{pmatrix} 0 \\ -2 \end{pmatrix}, \quad \varsigma_k := M_k^{-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \text{where } M_k := \begin{pmatrix} 1 & \mathbb{E}_\theta(A_{k\bullet} v)^3 \\ \mathbb{E}_\theta(A_{k\bullet} v)^3 & \mathbb{E}_\theta(A_{k\bullet} v)^4 - 1 \end{pmatrix}.$$

The lemma shows that the efficient scores with respect to the parameters that govern A, i.e. $\tilde{\ell}_{\theta,(\alpha,\beta_1),l}(y)$, do not change apart from being defined in terms of $v = z - Bx$. The efficient scores with respect to the components of the parameters that govern B are denoted by $\tilde{\ell}_{\theta,b,l}(y)$. The proof of Lemma 3 is given in the supplementary material and follows from Amari and Cardoso (1997) for $\tilde{\ell}_{\theta,(\alpha,\beta_1),l}(y)$ and for $\tilde{\ell}_{\theta,b,l}(y)$ the derivations are similar to those found in, for example, Bickel et al. (1998) or Newey (1990).

The following assumption imposes that the density score estimates satisfy Assumption 4 for the linear simultaneous equations model (23).

Assumption 6. *Assumption 4 holds when we replace Y_i by $(Z_i - B_n X_i)$ and take $\beta_n = (\beta_{1,n}, b_n)$.*

Proposition 3 in Appendix B shows that the density score estimator of Chen and Bickel (2006) satisfies this assumption under mild assumptions on η.

Having derived the efficient score function for γ, and with Assumptions 5 and 6 in place, we proceed by proposing estimators for the efficient score and information matrix. In particular, for a given $\beta_n \in \mathcal{B}$ and $\theta_n = (\alpha_0, \beta_n, \eta)$ we define the following estimates for the

[18] The independence assumption could be relaxed by requiring the moment assumptions in 3 to hold conditional on $\tilde{X}$. In this setup, our general approach as outlined in section 2 would continue to be valid though the resulting efficient score function would take a different form.

components of the efficient score for γ.

$$\hat{\ell}_{\theta_n,(\alpha,\beta_1),l}(y) = \sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K}\zeta_{l,k,j}\hat{\phi}_{k,n}(A_{n,k\bullet}v_n)A_{n,j\bullet}v_n + \sum_{k=1}^{K}\zeta_{l,k,k}\left[\hat{\tau}_{k,1,n}A_{n,k\bullet}v_n + \hat{\tau}_{k,2,n}\kappa(A_{n,k\bullet}v_n)\right]$$
$$\hat{\ell}_{\theta_n,b,m}(y) = \sum_{k=1}^{K}[-A_{n,k\bullet}D_{b,m}][(x-\bar{X}_n)\hat{\phi}_{k,n}(A_{n,k\bullet}v_n) - \bar{X}_n(\hat{\varsigma}_{k,1}A_{n,k\bullet}v_n + \hat{\varsigma}_{k,2}\kappa(A_{n,k\bullet}v_n))] \tag{24}$$

where $v_n = z - B_n x$ and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. The estimates $\hat{\tau}_{k,1,n}$, $\hat{\tau}_{k,2,n}$, $\hat{\varsigma}_{k,1,n}$ and $\hat{\varsigma}_{k,2,n}$ are defined analogously to those in equation (18) with Y_i replaced by $(Z_i - B_n X_i)$. The estimate for the corresponding efficient information matrix is given by

$$\hat{I}_{\theta_n} = \frac{1}{n}\sum_{i=1}^{n}\hat{\ell}_{\theta_n}(Y_i)\hat{\ell}_{\theta_n}(Y_i)' \ , \tag{25}$$

where $\hat{\ell}_{\theta_n}(Y_i)$ has components given by (24). Having defined our estimators we state our main result for the linear simultaneous equations model.

Proposition 2. *Let $\hat{\beta}_n$ be a $\sqrt{n}-$consistent estimator of β and let $\bar{\beta}_n$ denote a discretised version as in Theorem 2. Define $\bar{\theta}_n = (\alpha_0, \bar{\beta}_n, \eta)$ and consider the statistic*

$$\hat{S}^{SR}_{\bar{\theta}_n} = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\kappa}_{\bar{\theta}_n}(Y_i)\right)' \hat{\mathcal{I}}^{t,\dagger}_{\bar{\theta}_n}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\kappa}_{\bar{\theta}_n}(Y_i)\right) \ ,$$

with $\hat{\kappa}_{\bar{\theta}_n}$, $\hat{\mathcal{I}}_{\bar{\theta}_n}$ as in equation (12) based on $\hat{\ell}_{\bar{\theta}_n}(Y_i)$, $\hat{I}_{\bar{\theta}_n}$ as defined according to equations (24) and (25) and $\hat{\mathcal{I}}^{t,\dagger}_{\bar{\theta}_n}$ the Moore-Penrose inverse of the truncated version of $\hat{\mathcal{I}}^{t}_{\bar{\theta}_n}$ defined analogously to equation (14) with truncation level $\nu_n^{1/2}$. Suppose that assumptions 5 and 6 hold, that $\tilde{I}_{\theta_0,\beta\beta}$ is nonsingular, $(\alpha,\beta_1) \mapsto A(\alpha,\beta_1)$ is continuously differentiable and the map $(\alpha,\beta_1) \to [D_l(\alpha,\beta_1)]_{k\bullet}A(\alpha,\beta_1)^{-1}_{\bullet j}$ is Lipschitz continuous. Let $r_n = \operatorname{rank}(\hat{\mathcal{I}}^{t}_{\bar{\theta}_n})$ and denote by c_n the $1-a$ quantile of the $\chi^2_{r_n}$ distribution, for any $a \in (0,1)$. Then, under H_0

$$\lim_{n\to\infty} P_{\theta_0}(\hat{S}^{SR}_{\bar{\theta}_n} > c_n) \leq a,$$

with inequality only if $\operatorname{rank}(\tilde{\mathcal{I}}_{\theta_0}) = 0$.

The proposition shows that identification and singularity robust score test $\hat{S}^{SR}_{\bar{\theta}_n}$ can be used to conduct hypothesis tests in the linear simultaneous equations model. Further extensions, for instance nonlinear models that include $A^{-1}\epsilon$ as a component and possibly dynamic models can be handled using a similar approach. The choice for the estimator $\hat{\beta}_n$ is left

open to the researcher. Possible choices include using OLS estimates or one-step efficient estimators (e.g. van der Vaart, 2002, Section 7.2).

4 Simulation results

In this section we study the finite sample properties of the singularity and identification robust score test. We study the size and power of the tests under different data generating processes and compare its performance to several alternatives that have been proposed in the literature. We first study the baseline ICA model (16) after which we consider the linear simultaneous equations model and a panel data model which constitutes a special case of the linear simultaneous equations model.

4.1 Baseline ICA model

We start by drawing independent samples from the ICA model (16) for dimensions $K = 2$ and $K = 3$ and sample sizes $n = 200, 500$. We fix ϵ_1 to have a standard Gaussian density and consider different densities for ϵ_k, with $k = 2, \ldots, K$, that range from standard Gaussian to skewed bi-modal distributions. The non-Gaussian densities are either Student's t or mixtures of normals taken from Marron and Wand (1992). Table 1 provides an overview.

The matrix of interest $A(\gamma) = A(\alpha)$ is taken as a rotation matrix using the trigonometric transformation.[19] In this setting there are no additional nuisance parameters which allows us to concentrate on the consequences of weak non-Gaussianity on the score test and some alternative tests that have been proposed in the literature. In the simulation designs below we include nuisance parameters to show that their inclusion does not alter the size of the test.

For each specification we simulate $S = 5,000$ datasets and for each we compute the singularity robust score statistic as defined in Proposition 1 using the log density score estimator of Jin (1992) and Chen and Bickel (2006) as discussed in Appendix B using $B = 4, 6$ or 8 cubic splines, with the upper and lower endpoints taken to be the 95th and 5th percentile of the samples adjusted respectively up and down by $\log(\log n)$.[20] We threshold the information matrix estimate at machine precision for ν_n for all simulations.

[19] For instance, when $K = 2$ we have that

$$A(\alpha) = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix},$$

with a scalar parameter α.

[20] If this adjustment lead to the endpoint being lower (resp. higher) than the minimum (resp. maximum) of the sample, the minimum (resp. maximum) was used instead.

Size results

In Table 2 we show the empirical rejection frequencies corresponding to the $\hat{S}^{SR}_{\bar{\theta}_n}$ test with nominal size 0.05. The columns correspond to the different choices for the densities ϵ_k for $k \geq 2$.

The first column corresponds to the case where all densities are Gaussian and the expected likelihood takes the same value for all $\alpha \in \mathbb{R}^{L_\alpha}$, e.g. α is unidentified. Nonetheless, we find that the empirical rejection frequency of the score test is always close to the nominal size. This holds regardless of the sample size n, the dimension of the ICA model K and the number of cubic splines B.

Second, when the second (or the second and third) density is non-Gaussian the size remains correct, regardless of the true density and the distance to Gaussianity of this density. Even for complicated skewed bi-modal and outlier densities (e.g. columns 7 and 10) the $\hat{S}^{SR}_n$ test has size close to nominal regardless of the sample size.

Third, overall the number of cubic splines used has little influence on the results. A close inspection reveals that when the number of cubic splines is equal to four the test becomes mildly conservative for some densities, therefore we use $B = 6$ cubic splines in the remaining exercises.

In sum, the size of the semiparametric score test is well controlled for the distributions listed in Table 1.

Comparison to alternative approaches

Next, we compare our semiparametric testing approach to different parametric approaches based on (psuedo) maximum likelihood and the generalized method of moments. Importantly, none of these alternatives are designed to be robust against cases where the true densities are close to Gaussian and previous simulation studies in the literature have highlighted size distortions in such cases for these methods (e.g. Gouriéroux, Monfort and Renne, 2017; Lanne and Luoto, 2019). We merely confirm these findings.

First, we consider the standard maximum likelihood Wald, score and likelihood ratio tests that are based on the students t density for ϵ_k. For densities 1-4 in Table 1 these tests correspond to exact maximum likelihood tests, with the caveat that when the degrees of freedom increases the parameters α become weakly identified, or not-identified when the degrees of freedom tends to infinity as for the Gaussian density. For all other densities the standard maximum likelihood tests are mis-specified.

Second, we consider the psuedo-maximum likelihood tests developed by Gouriéroux, Monfort and Renne (2017). These tests are asymptotically valid for a broader range of true

distribution functions and amount to fixing the functional form of the likelihood. We follow their implementation and choose the Students t density with five degrees of freedom as the pseudo-likelihood and compute the Wald statistic based on this density.

Third, we compare our method to the recently developed GMM method of Lanne and Luoto (2019), which relies on higher order moments to identify the parameter vector α. We follow their implementation and use $\mathbb{E}\epsilon_{i,k}^2 = 1$, $\mathbb{E}\epsilon_{i,k}\epsilon_{i,j} = 0$, $\mathbb{E}\epsilon_{i,k}^3\epsilon_{i,j} = 0$ and $\mathbb{E}\epsilon_{i,k}^2\epsilon_{i,j}^2 = 1$ as moment conditions for all $j \neq k$ and $j, k = 1, \ldots, K$. The GMM likelihood ratio test is then computed as the rescaled difference between the unrestricted and restricted J-statistics, based on the 2-step GMM estimator, see Lanne and Luoto (2019) for more details.

The empirical rejection frequencies are shown in top panel of Table 3 for the case where $K = 2$ and $n = 500$. We find, perhaps not surprisingly, that the MLE Wald test is severely over-sized when the degrees of freedom of the Students t distribution becomes large or the density is mis-specified. In contrast, the likelihood ratio test is under-sized for most of the specifications considered. The parametric score test, e.g. the LM test, performs well when the density is correctly specified (e.g. cases 2-4), which is understandable as α is fixed under the null and no identification problems arise, see Andrews and Mikusheva (2015) for more elaborate examples. When the density is misspecified the parametric score test typically performs less well.

The psuedo-maximum likelihood Wald test of Gouriéroux, Monfort and Renne (2017) is correctly sized when the psuedo-likelihood is close to the true density, but the method performs poorly in all other scenarios. The GMM-based likelihood ratio test of Lanne and Luoto (2019) over-rejects quite severely when the true densities approach the Gaussian, which corresponds to the results in Lanne and Luoto (2019), see their Table 1.

In sum, none of the alternative methods appear to control size under both (a) weakly non-Gaussian densities and (b) mis-specification of the likelihood.

Power results

Finally, we study the power of the semiparametric score test in the baseline ICA model. We consider the case where $K = 2$ and $n = 500$, such that α becomes a scalar parameter. To compare our power we consider the parametric score test, or LM test, based on the Students t density. This approach controls the size of the test reasonably well, see Table 2, and is the natural parametric counterpart for the first four densities considered.

Figure 1 shows the empirical rejection frequencies when we vary α around the, arbitrarily chosen, true value $\alpha = \pi/4$. Each point on the curve is based on $S = 5,000$ simulations and for clarity of the figure we adjusted the power of the parametric score test such that it is size correct, e.g. exactly 0.05 for $\alpha = \alpha_0$, in all specifications.

We find that the power of the parametric score test is larger when compared to the semi-parametric test when the density is correctly specified. This is the top row of Figure 1 where we consider the (normalised) student t density as the truth. Nonetheless the $S^{SR}_{\bar{\theta}_n}$ test comes quite close in terms of power.

For all other density choices the $S^{SR}_{\bar{\theta}_n}$ test convincingly outperforms its parametric counterpart. Especially for bi-modal densities the difference in power is large. We note that α is only identified up to scale and permutation of the columns hence for $\alpha \in [0, 2\pi]$ there are multiple optimal points and the power starts to decrease when it gets close to the next permutation. Based on these results we concludes that the semi-parametric score test has adequate power even when compared to correctly specified parametric tests.

4.2 Linear simultaneous equations model

Next, we discuss the simulation results for the LSEM (23). The dimensions of the design are similar as above with the addition that we consider $d = 2, 3$ for the number of covariates. We now parametrize $A(\gamma)^{-1} = \Sigma^{1/2}(\beta_1)R(\alpha)$ as in example 1, where $\Sigma^{1/2}$ is lower triangular, with the non-zero entries collected in β_1, and the rotation matrix R is specified using the trigonometric transformation. The explanatory variables are drawn from the standard normal distribution. The Euclidean nuisance parameters now include β_1 and the elements of B. The first error term follows a Gaussian distribution and the different distributions from Table 1 are considered for the second and third error terms. For each specification we simulate $S = 5,000$ datasets and for each we compute the singularity robust score statistic as defined in Proposition 2.

The empirical rejection frequencies are shown in Table 4. We find that the rejection frequencies of the $S^{SR}_{\bar{\theta}_n}$ test are generally close to the nominal size. Only for the outlier density the test over-rejects when $n = 200$ and $k = 3$. The reason is that the sample size is too small to estimate the log density scores sufficiently accurately due to the very heavy tails of this density. The rejection frequency improves when n is increased to 500.

The power curves are shown in Figure 2 for two different model specifications. The red curves correspond to the specification discussed above. In this scenario where the first density is always exactly Gaussian the parameter α is always weakly identified. To investigate the effect that the first density can have, the blue curves show the power when we vary the density of ϵ_1 along with the others. This change increases the power of the test substantially indicating that deviations from Gaussianity need to be present in all shocks to have high power for small sample sizes.

4.3 Short T panel data models

In this final simulation design we consider a class of short T panel data models and show that such models can be considered as a special case of the LSEM model above. We explicitly investigate this special case as it is a specification that we also use in our empirical study below. Moreover, given that fixed effects are omnipresent in economic data this extension seems of first order importance.

Let $Z_{i,t}$ be a $K_z \times 1$ vector of observations for individual i at time t. We consider the model

$$Z_{i,t} = c_i + BX_{i,t} + \Sigma^{1/2} R e_{i,t} \ , \tag{26}$$

where c_i is a vector of individual fixed effects, $X_{i,t}$ is an $d \times 1$ vector of explanatory variables whose effect is captured by the $K_z \times d$ matrix B, $\Sigma^{1/2}$ is a $K_z \times K_z$ lower triangular matrix, R a $K_z \times K_z$ rotation matrix and $e_{i,t}$ is the vector of independent components with mean zero and identity variance.

To write this model as an LSEM model we first subtract the time series averages from both sides to obtain[21]

$$Z_{i,t} - \bar{Z}_i = B(X_{i,t} - \bar{X}_i) + \Sigma^{1/2} R(e_{i,t} - \bar{e}_i) \ .$$

We stack the differences $\tilde{Z}_{i,t} = Z_{i,t} - \bar{Z}_i$ and $\tilde{X}_{i,t} = X_{i,t} - \bar{X}_i$ across t and consider the extended LSEM model

$$Y_i = (I_T \otimes B) X_i + A^{-1} \epsilon_i \ , \tag{27}$$

where $Y_i = (\tilde{Z}'_{i,1}, \ldots, \tilde{Z}'_{i,T})'$, $X_i = (\tilde{X}'_{i,1}, \ldots, \tilde{X}'_{i,T})'$, $\epsilon_i = (\epsilon'_{i,1}, \ldots, \epsilon'_{i,T})'$ and

$$A^{-1} = \begin{bmatrix} \Sigma^{1/2}R & 0 & & 0 \\ 0 & \ddots & & \\ & & \ddots & 0 \\ 0 & & 0 & \Sigma^{1/2}R \end{bmatrix} \begin{bmatrix} I_{K_z} & -\frac{1}{T}I_{K_z} & & -\frac{1}{T}I_{K_z} \\ -\frac{1}{T}I_{K_z} & \ddots & & \\ & & \ddots & -\frac{1}{T}I_{K_z} \\ -\frac{1}{T}I_{K_z} & & -\frac{1}{T}I_{K_z} & I_{K_z} \end{bmatrix} .$$

The correction matrix on the right ensures that the original independent errors $e_{i,t}$ are used as errors in the LSEM representation of the short T panel data model. The coefficients in B and $\Sigma^{-1/2}$ can be consistently recovered using OLS after removing the sample means and we will conduct tests for the coefficients in α that determine R similar to the previous two examples.

In this setting we vary $K = 2, 3$, $T = 10, 20$, $n = 500, 1000$ and $d = 2, 3$ to reflect the

[21] With an additional assumption on the initial shock we can also consider taking first differences.

dimensions of the empirical study considered below. The errors are again simulated from the different distributions listed in Table 1. The empirical rejection frequencies are reported in Table 5. We find that for all specifications the rejection frequencies are very close to the nominal size.

To summarize, the simulation results show that for several LSEMs the semi-parametric score test can be used to conduct robust inference on the possibly weakly identified parameters. We note that other specific models can also be cast in the LSEM framework, provided that the observations are independent across i.

5 Testing production function coefficients

In this section we explore whether non-Gaussian distributions can help to identify the coefficients in the production function of a firm. Interestingly, the very first contributions in this literature highlighted the identification problem in this setting using simultaneous equations (e.g. Marschak and Andrews, 1944; Hoch, 1958). This generated a large number of works that aim to address the simultaneity problem in different ways. Prominent examples include using panel data methods (e.g. Arellano and Bond, 1991; Blundell and Bond, 1998) or proxy variable methods (e.g. Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg, Caves and Frazer, 2015).

To study how non-Gaussian distributions may assist in the quest for identification we consider the baseline Cobb-Douglas production function

$$O = e^{c_1} L^{\alpha_1} K^{\alpha_2} e^{\epsilon_1} \ ,$$

where O, L, K denote output, labor and capital, respectively, and ϵ_1 captures unobserved factors that determine output. Our interest is in the coefficients α_1 and α_2 that determine the contributions of labor and capital to output. The, well known, difficulty for learning about α_1 and α_2 is that the inputs L, K are typically choice variables of the firm. Allocations are made to maximize profits and hence will generally depend on unobservables ϵ_1.

To address this simultaneity problem we consider a simultaneous equations approach that allows for correlation among L, K, ϵ_1, and exploits possible non-Gaussianity in the errors to identify the parameters α_1 and α_2. We consider this approach for a single cross-section of firms as well as in a panel data setting. The latter has the benefit that some forms of unobserved heterogeneity can incorporated in the model.

To be specific, the models that we consider are defined for $Y = (\log O, \log L, \log K)'$, and

are of the form

$$S(\alpha, \beta_1)Y = c + BX + D(\beta_1)\epsilon \ , \tag{28}$$

where c can be firm specific in a panel data setting and X may include other exogenous variables. We parametrize the matrices as follows

$$S(\alpha, \beta_1) = \begin{bmatrix} 1 & -\alpha_1 & -\alpha_2 \\ \beta_{1,1} & 1 & -\alpha_3 \\ \beta_{2,1} & \beta_{3,1} & 1 \end{bmatrix} \quad \text{and} \quad D(\beta_1) = \begin{bmatrix} \beta_{4,1} & 0 & 0 \\ 0 & \beta_{5,1} & 0 \\ 0 & 0 & \beta_{6,1} \end{bmatrix} .$$

We note that parameters in β_1 can be recovered from the variance of Y and we simultaneously test $\alpha = \alpha_0$, where $\alpha = (\alpha_1, \alpha_2, \alpha_3)'$, for different choices of α_0 to obtain the confidence sets. The positioning of α_3 is arbitrary in our setting as it is not a parameter of interest, but it can also not be identified from the variance alone. The confidence sets for α_1 and α_2 that we report are obtained by taking the minimum and maximum values for α_1 and α_2 that are not rejected by the score test.[22] Finally, to pin down the desired rotation we impose that α_1 and α_2 are positive and the correlations between L, K and ϵ_1 are non-negative. In other words, positive shocks to output do not decrease labor and capital, a mild sign restriction that corresponds with most economic models (e.g. Hoch, 1958).

We use a sample of 115,000 manufacturing firms that are observed from 2000 until 2017.[23] We perform three exercises. First, to illustrate our methodology we consider the cross section of firms that exist in 2017 and investigate in detail the output of the methodology. Second, we repeat the exercise for different years and assess the changes in α_1 and α_2 over time. Finally, we consider the model for the 2000-2017 panel which allows us to investigate the influence of fixed effects on the location of (α_1, α_2).

5.1 Cross-sectional results

We first illustrate the methodology using the manufacturing firms that existed in 2017. We have $n = 1247$ firms with observations for output, labor and capital. We consider model (28) with a constant and possibly the age of the firm as a control variable (e.g. Olley and Pakes, 1996).

The 95% confidence bounds for the production function coefficients α_1 (labor) and α_2 (capital) are shown in Table 6. We find that these coefficients are generally well identified empirically. In particular, with 95% confidence, α_1 lies between 0.41 and 0.68, while α_2 lies

[22] We note that this projection approach is conservative and refinements along the lines of Kaido, Molinari and Stoye (2019) may improve the current findings.

[23] The data are obtained from CompuStat.

between 0.27 and 0.50, for all choices of the control variables. The joint confidence region for (α_1, α_2) is shown in the top left panel of Figure 3. It shows that we cannot reject that $\alpha_1 + \alpha_2 = 1$ as the confidence region exactly lies on this line.

To understand where the identification in the LSEM is coming from, the other panels in Figure 3 show the empirical densities of the residuals $\hat{\epsilon}_i = \hat{A}(Z_i - \hat{B}X_i)$, where $\hat{A}$ corresponds to the choice for α that minimizes the score statistic. We find that the empirical densities are indeed different from the normal density, notably for the first density. Overall, we can reject the null hypothesis that the errors are normally distributed but the visual inspection shows that the deviations are mild. Indeed the alternative methods that we discussed in the simulation study, which are not robust to weak deviations from Gaussianity, give much smaller confidence bands. This shows that non-Gaussianity can be a useful tool for identification, but robust methods need to be adopted for the approach to be used reliably. We emphasize that besides the sign restrictions that ensure that the correlations between L, K and ϵ_1 are non-negative no further structural assumptions are needed.

Table 6 also shows the baseline OLS estimates as obtained by regressing log output on the controls and log labor and log capital. We find that these estimates are very different and the confidence intervals do not overlap with those of the LSEM.

Next, to highlight that the year 2017 was in no way exceptional we repeat the previous exercise for the years 2000-2017. The results for the model that includes age as a control variable are shown in Figure 4. Overall, the findings are very stable. We do notice a modest decline in the labor input coefficient and an increase of the coefficient on capital towards the end of the sample.

5.2 Panel data results

In the previous section we explored the estimation of the production function coefficients using the classical LSEM. Clearly, such approach does not allow for heterogeneity across firms and in this section we extend our approach to allow for firm fixed effects by using panel data over 2000-2017. We consider the panel data specification given in (26), with $\Sigma^{1/2}R$ replaced by $S^{-1}(\alpha, \beta_1)D(\beta_1)$ and taking $Z_{i,t} = (\log O_{i,t}, \log L_{i,t}, \log K_{i,t})'$. This model can then be cast in the LSEM form similarly as in (27). We include additional time-fixed effects for each equation as controls variables.[24]

Similar as before we adopt the semi-parametric score test to construct confidence bands for the production function coefficients. The results are shown in Table 7. We find that

[24]With this modification tests for serial correlation in the errors could be passed. More specifically, for each equation we applied the portmanteau test developed in Jochmans (2020) and when we included time fixed effects we could not reject the null of no serial correlation.

the confidence bands for the labor coefficients are very similar when compared to the cross-sectional results. In contrast, the bands imply that the coefficient for capital is notably smaller. A possible explanation is that we considered a balanced sample including only firms that are observed over the 2000-2017 period.

When we compare the panel data LSEM confidence intervals with those implied by a standard fixed-effect regression of output on labor and capital, we find that the intervals for labor are noticeably different. In contrast, those for capital are quite similar, providing some evidence that capital could perhaps be treated as pre-determined after including fixed effects.

6 Conclusion

In this paper we highlighted a weak identification problem that arises when non-Gaussian distributions are used to identify coefficients in LSEMs. In particular, existing inference methods suffer from size distortions when the true distributions are close to Gaussian.

To remedy this problem we proposed a class of identification robust score statistics for testing hypotheses in semi-parametric likelihood models. Using high-level assumptions we outlined a general approach for testing finite dimensional parameters in the presence of infinite dimensional, but well identified, nuisance parameters.

The general framework was worked out in detail for a class of LSEMs where the interest was in the contemporaneous effects matrix A and the densities of the errors were treated non-parametrically. We show both theoretically and in simulation that the semi-parametric score statistic is robust to the weak Gaussian problem and controls size over a large class of densities that satisfy mild moment conditions.

While we have restricted our treatment to models where the observations were independently distributed across entities, we note that a similar approach can be considered for dynamic models, but this will require extending our main Theorem 2 to allow for non-i.i.d. data. Similarly, dynamic panel data models can be considered pending a novel strategy for handling the initial conditions. These extensions are left for future work.

References

Ackerberg, Daniel A., Kevin Caves, and Garth Frazer. 2015. "Identification Properties of Recent Production Function Estimators." *Econometrica*, 83(6): 2411–2451.

Amari, S., and J-F. Cardoso. 1997. "Blind Source Separation - Semiparametric Statistical Approach." *IEEE Transactions On Signal Processing*, 45(11).

Anderson, Theodore W., and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *Ann. Math. Statist.*, 20(1): 46–63.

Andrews, Donald W. K. 1987. "Asymptotic Results for Generalized Wald Tests." *Econometric Theory*, 3(3): 348–358.

Andrews, Donald W.K., and James H. Stock. 2007. "Testing with many weak instruments." *Journal of Econometrics*, 138(1): 24–46.

Andrews, Donald W. K., and Patrik Guggenberger. 2019. "Identification- and singularity-robust inference for moment condition models." *Quantitative Economics*, 10(4): 1703–1746.

Andrews, I., and A. Mikusheva. 2015. "Maximum likelihood inference in weakly identified dynamic stochastic general equilibrium models." *Quantitative Economics*, 6.

Andrews, I., and A. Mikusheva. 2016. "Conditional inference with a functional nuisance parameter." *Econometrica*, 84(4).

Arellano, Manuel, and Stephen Bond. 1991. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *The Review of Economic Studies*, 58(2): 277–297.

Bach, Francis R., and Michael I. Jordan. 2002. "Kernel Independent Component Analysis." *Journal of Machine Learning Research*, 3: 1–48.

Bekaert, Geert, Eric Engstrom, and Andrey Ermolov. 2019. "Macro Risks and the Term Structure of Interest Rates." *Working paper.*

Bekaert, Geert, Eric Engstrom, and Andrey Ermolov. 2020. "Aggregate Demand and Aggregate Supply Effects of COVID-19: A Real-time Analysis." *Working paper.*

Bhatia, R. 1997. *Matrix Analysis.* New York, NY, USA:Springer.

Bickel, Peter J., Ya'acov Ritov, and Thomas M. Stoker. 2006. "Tailor-made tests for goodness of fit to semiparametric hypotheses." *Ann. Statist.*, 34(2): 721–741.

Bickel, P. J., C. A. J. Klaasen, Y. Ritov, and J. A. Wellner. 1998. *Efficient and Adaptive Estimation for Semiparametric Models.* New York, NY, USA:Springer.

Blundell, Richard, and Stephen Bond. 1998. "Initial conditions and moment restrictions in dynamic panel data models." *Journal of Econometrics*, 87(1): 115–143.

Cattaneo, Matias D., Richard K. Crump, and Michael Jansson. 2012. "Optimal inference for instrumental variables regression with non-Gaussian errors." *Journal of Econometrics*, 167(1): 1 – 15.

Chen, A., and P. J. Bickel. 2006. "Efficient Independent Component Analysis." *Annals of Statistics*, 34(6).

Choi, Sungsub, W. J. Hall, and Anton Schick. 1996. "Asymptotically uniformly most powerful tests in parametric and semiparametric models." *Ann. Statist.*, 24(2): 841–861.

Comon, P. 1994. "Independent component analysis, A new concept?" *Signal Processing*, 36.

de Boor, C. 2001. *A Practical Guide to Splines.* New York, NY, USA:Springer.

Dhrymes, Phoebus J. 1994. *Topics in Advanced Econometrics, Volume II Linear and Nonlinear Simultaneous Equations.* Springer-Verlag New York.

Durrett, Rick. 2019. *Probability Theory and Examples.* . 5th ed., Cambridge, UK:Cambridge University Press.

Fiorentini, Gabriele, and Enrique Sentana. 2020. "Discrete Mixtures of Normals Pseudo Maximum Likelihood Estimators of Structural Vector Autoregressions." working paper.

Frisch, R. 1933. "Propagation Problems and Impulse Problems In Dynamic Economics." In *Economic Essays in Honor of Gustav Cassel*. George Allen and Unwin.

Garoni, C., and S. Serra-Capizzano. 2017. *Generalized Locally Toeplitz Sequences: Theory and Applications.* Vol. 1, Cham, Switzerland:Springer.

Gouriéroux, C., A. Monfort, and J-P. Renne. 2017. "Statistical inference for independent component analysis: Application to structural VAR models." *Journal of Econometrics*, 196.

Gouriéroux, Christian, Alain Monfort, and Jean-Paul Renne. 2019. "Identification and Estimation in Non-Fundamental Structural VARMA Models." *The Review of Economic Studies*, 87(4): 1915–1953.

Guay, Alain. 2020. "Identification of Structural Vector Autoregressions Through Higher Unconditional Moments." *Journal of Econometrics*. forthcoming.

Haavelmo, T. 1943. "The Statistical Implications of a System of Simultaneous Equations." *Econometrica*, 11: 1–12.

Haavelmo, T. 1944. "The Probability Approach in Econometrics." *Econometrica*, 12. Supplement.

Hahn, Jinyong. 1994. "The Efficiency Bound of the Mixed Proportional Hazard Model." *The Review of Economic Studies*, 61(4): 607–629.

Hall, W. J., and David J. Mathiason. 1990. "On Large-Sample Estimation and Testing in Parametric Models." *International Statistical Review*, 58(1): 77–97.

Herwartz, Helmut. 2019. "Long-run neutrality of demand shocks: Revisiting Blanchard and Quah (1989) with independent structural shocks." *Journal of Applied Econometrics*, 34(5): 811–819.

Hoch, Irving. 1958. "Simultaneous Equation Bias in the Context of the Cobb-Douglas Production Function." *Econometrica*, 26(4): 566–578.

Horn, R. A., and C. R. Johnson. 2013. *Matrix Analysis.* . 2 ed., Cambridge University Press.

Horowitz, Joel L. 2009. *Semiparametric and Nonparametric Methods in Econometrics.* Springer-Verlag New York.

Hyvärinen, A., J. Karhunen, and E. Oja. 2001. *Independent Component Analysis.* John Wiley & Sons, Inc.

Jin, K. 1992. "Empirical Smoothing Parameter Selection In Adaptive Estimation." *Annals of Statistics*, 20(4).

Jochmans, Koen. 2020. "Testing for correlation in error-component models." *Journal of Applied Econometrics*, 35(7): 860–878.

Kaido, Hiroaki, Francesca Molinari, and Jörg Stoye. 2019. "Confidence Intervals for Projections of Partially Identified Parameters." *Econometrica*, 87(4): 1397–1432.

Kleibergen, F. 2005. "Testing parameters in GMM without assuming that they are identified." *Econometrica*, 73(4).

Kleibergen, Frank. 2002. "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression." *Econometrica*, 70(5): 1781–1803.

Kocherlakota, S., and K. Kocherlakota. 1991. "Neyman's C(α) test and Rao's efficient score test for composite hypotheses." *Statistics & Probability Letters*, 11(6): 491 – 493.

Lanne, Markku, and Helmut Lütkepohl. 2010. "Structural Vector Autoregressions With Nonnormal Residuals." *Journal of Business & Economic Statistics*, 28(1): 159–168.

Lanne, Markku, and Jani Luoto. 2019. "GMM Estimation of Non-Gaussian Structural Vector Autoregression." *Journal of Business & Economic Statistics*, 0(0): 1–13.

Lanne, M., M. Meitz, and P. Saikkonen. 2017. "Identification and estimation of non-Gaussian structual vector autoregressions." *Journal of Econometrics*, 196.

Le Cam, Lucien M., and Grace L. Yang. 2000. *Asypmtotics in Statistics: Some Basic Concepts.* . 2 ed., New York, NY, USA:Springer.

Levinsohn, James, and Amil Petrin. 2003. "Estimating Production Functions Using Inputs to Control for Unobservables." *The Review of Economic Studies*, 70(2): 317–341.

Magnus, Jan R., Henk G. J. Pijls, and Enrique Sentana. 2020. "The Jacobian ofthe Exponential Function." Working Paper.

Magnus, J. R., and H. Neudecker. 2019. *Matrix Differential Calculus with Applications in Statistics and Econometrics.* John Wiley & Sons.

Marron, J. S., and M. P. Wand. 1992. "Exact Mean Integrated Squared Error." *Annals of Statistics*, 20(2).

Marschak, Jacob, and William H. Andrews. 1944. "Random Simultaneous Equations and the Theory of Production." *Econometrica*, 12(3/4): 143–205.

Maxand, Simone. 2018. "Identification of independent structural shocks in the presence of multiple Gaussian components." *Econometrics and Statistics*.

Moneta, Alessio, Doris Entner, Patrik O. Hoyer, and Alex Coad. 2013. "Causal Inference by Independent Component Analysis: Theory and Applications*." *Oxford Bulletin of Economics and Statistics*, 75(5): 705–730.

Newey, Whitney K. 1990. "Semiparametric efficiency bounds." *Journal of Applied Econometrics*, 5(2): 99–135.

Neyman, Jerzy. 1979. "C(α) Tests and Their Use." *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 41(1/2): 1–21.

Olea, José Luis Montiel, and Carolin Pflueger. 2013. "A Robust Test for Weak Instruments." *Journal of Business & Economic Statistics*, 31(3): 358–369.

Olley, G. Steven, and Ariel Pakes. 1996. "The Dynamics of Productivity in the Telecommunications Equipment Industry." *Econometrica*, 64(6): 1263–1297.

Powell, M. J. D. 1981. *Approximation Theory and Methods.* Cambridge, UK:Cambridge University Press.

Rabinowitz, Daniel. 2000. "Computing the Efficient Score in Semi-Parametric Problems." *Statistica Sinica*, 10(1): 265–280.

Rao, C. R., and S. K. Mitra. 1971. *Generalized Inverse of Matrices and its Applications.* New York, NY, USA:John Wiley & Sons, Inc.

Samarov, Alexander, and Alexandre Tsybakov. 2004. "Nonparametric independent component analysis." *Bernoulli*, 10(4): 565–582.

Sen, A. 2012. "On the Interrelation Between the Sample Mean and the Sample Variance." *The American Statistician*, 66(2).

Sims, Christopher A. 2021. "SVAR Identification through Heteroskedasticity with Misspecified Regimes." working paper.

Staiger, D., and J. H. Stock. 1997. "Instrumental variables regression with weak instruments." *Econometrica*, 65(3).

Stock, James H., and Motohiro Yogo. 2005. "Testing for Weak Instruments in Linear IV Regression." *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, , ed. Donald W. K. Andrews and James H.Editors Stock, 80–108. Cambridge University Press.

Stock, J. H., and J. H. Wright. 2000. "GMM with weak identification." *Econometrica*, 68(5).

Tank, A, E B Fox, and A Shojaie. 2019. "Identifiability and estimation of structural vector autoregressive models for subsampled and mixed-frequency time series." *Biometrika*, 106(2): 433–452.

Tinbergen, Jan. 1939. *Statistical Testing of Business Cycle Theories: Part I: A Method and Its Application to Investment Activity.*

van der Vaart, A. W. 1988. *Statistical Estimation in Large Parameter Spaces. CWI Tracts*, Amsterdam:Centrum voor Wiskunde en Informatica.

van der Vaart, A. W. 1998. *Asymptotic Statistics.* . 1st ed., New York, NY, USA:Cambridge University Press.

van der Vaart, A. W. 2002. "Semiparametric Statistics." In *Lectures on Probability Theory and Statistics: Ecole d'Eté de Probabilités de Saint-Flour XXIX - 1999.* , ed. P. Bernard. Berlin, Germany:Springer.

van der Vaart, A. W., and J. A. Wellner. 1996. *Weak Convergence and Empirical Processes.* . 1st ed., New York, NY, USA:Springer-Verlag New York, Inc.

Velasco, Carlos. 2020. "Identification and estimation of Structural VARMA models using higher order dynamics." working paper.

Appendix A: Main proofs

In this appendix we provide the main proofs of Theorems 1 and 2 as well as Lemma 1. The proofs of Lemma 2 and Proposition 1 are special cases of Lemma 3 and Proposition 2. The proof of Lemma 3 in included in the supplementary material as it follows along the lines of Amari and Cardoso (1997). Proposition 2 is proven below. In Appendix B below we provide the details for the log density score estimation.

Throughout the appendix we often use the empirical process notation: $Pf = \mathbb{E}f(X_i)$, $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(Y_i)$ and $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - P)f$. Further, G_k denotes the law on $\mathbb{R}$ corresponding to η_k and ϵ_k is distributed according to G_k. Similarly G_0 denotes the law on $\mathbb{R}^{d-1}$ corresponding to η_0 and $\tilde{X}$ is distributed according to G_0.

Proof of Theorem 1. Let $P_0 := P_{\theta_0}$, where θ_0 is defined in Assumption 1. First, we show that under the conditions imposed by assumption 1 we have

$$\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\theta_n} - \tilde{\ell}_{\theta_n}\right] \xrightarrow{P_0} 0, \quad \sqrt{n}\mathbb{P}_n\left[\tilde{\ell}_{\theta_n} - \tilde{\ell}_{\theta_0}\right] + \sqrt{n}\tilde{I}_{\theta_0}(0, (\beta_n - \beta)')' \xrightarrow{P_0} 0, \quad \hat{I}_{\theta_n} \xrightarrow{P_0} \tilde{I}_{\theta_0}. \quad (29)$$

Define $b_n := \sqrt{n}(\beta_n - \beta)$ and let $(n_m)_{m\geq 1}$ be an arbitrary subsequence of $(n)_{n\geq 1}$. It is sufficient for (29) that we can demonstrate that there is a further subsequence $(n_{m(k)})_{k\geq 1}$ along which the claimed convergence holds. There exists a sub-subsequence such that $b_{n_{m(k)}} \to b$ for some $b \in \mathbb{R}^{L_\beta}$.[25] Taking such a subsequence will suffice as we will now demonstrate that the claimed convergence holds for an arbitrary convergent sequence $b_n \to b$.

Let Q_n^n denote the law of $(Y_i)_{i=1}^n$ corresponding to θ_n and P_0^n that corresponding to θ_0. Let $\Lambda_n(Q_n, P_0) = n\mathbb{P}_n \log q_n - \log p_0$ be the corresponding log-likelihood ratio. In view of the differentiability in quadratic mean of the model (e.g. Definition 1) we have by van der Vaart and Wellner, 1996, lemma 3.10.11:

$$\Lambda_n(Q_n, P) = \sqrt{n}\mathbb{P}_n b'\dot{\ell}_{\theta_0,\beta} - \frac{1}{2}b'\dot{I}_{\theta_0,\beta\beta}b + R_n,$$

where $R_n \to 0$ in probability under both P_0^n and Q_n^n and $\dot{I}_{\theta_0} = \mathbb{V}(\dot{\ell}_{\theta_0})$. Noting that $\dot{\ell}_{\theta_0}$ is a score by assumption 0 and hence in $L_2(P_0)$ (e.g. van der Vaart, 2002, Lemma 1.7) it follows by the CLT that

$$\Lambda_n(Q_n, P) \rightsquigarrow \mathcal{N}\left(-\frac{1}{2}b'\dot{I}_{\theta_0,\beta\beta}b, b'\dot{I}_{\theta_0,\beta\beta}b\right),$$

under P_0, from which we can conclude that $P_0^n \triangleleft\triangleright Q_n^n$ (e.g. van der Vaart and Wellner, 1996, example 3.10.6). This mutual contiguity and Le Cam's first lemma (e.g. van der Vaart, 1998, Lemma 6.4) ensure that leftmost and rightmost claims in (29) hold given parts 2 & 3 of assumption 1. Noting that $P_0[\tilde{\ell}_{\theta_0}\dot{\ell}'_{\theta_0,\beta}]b = \tilde{I}_{\theta_0}(0, b')'$, the middle claim of equation (29) follows by proposition A.10 in van der Vaart (1988), which requires Assumption 1-part 4.[26]

Next we show that (29) continues to hold if θ_n is replaced by $\bar{\theta}_n$ as defined in the theorem.[27] Since $\bar{\beta}_n$ remains $\sqrt{n}$-consistent there is an $M > 0$ such that $P_0\left(\sqrt{n}\|\bar{\beta}_n - \beta\| > M\right) <$

[25] Such a subsequence and b exist by the Bolzano-Weierstrass theorem.

[26] Cf. lemma 7.3 in van der Vaart (2002); the proof of theorem 25.57 in van der Vaart (1998).

[27] The proof that (29) continues to hold is adapted from the proof of Theorem 5.48 in van der Vaart (1998).

ε. If $\sqrt{n}\|\bar{\beta}_n - \beta\| \leq M$ then $\bar{\beta}_n$ is equal to one of the values in the finite set $B_n = \{\beta' \in n^{-1/2}C\mathbb{Z}^{L_\beta} : \|\beta' - \beta\| \leq n^{-1/2}M\}$. For each M this set has finite number of elements bounded independently of n, call this upper bound $\overline{B}$. Let

$$R'_n(\beta') \coloneqq \sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\theta'} - \tilde{\ell}_{\theta'}\right],\ R''_n(\beta') \coloneqq \sqrt{n}\mathbb{P}_n\left[\tilde{\ell}_{\theta'} - \tilde{\ell}_{\theta_0}\right] + \sqrt{n}\tilde{I}_{\theta_0}(0, (\beta' - \beta)')',\ R'''_n(\beta') \coloneqq \hat{I}_{\theta'} - \tilde{I}_{\theta_0},$$

where $\theta' = (\alpha_0, \beta', \eta)$. Letting R_n denote either R'_n, R''_n or R'''_n we have that for any $\upsilon > 0$

$$\begin{aligned} P_0\left(\|R_n(\bar{\beta}_n)\| > \upsilon\right) &\leq \varepsilon + \sum_{\beta_n \in B_n} P_0\left(\{\|R_n(\beta_n)\| > \upsilon\} \cap \{\bar{\beta}_n = \beta_n\}\right) \\ &\leq \varepsilon + \sum_{\beta_n \in B_n} P_0\left(\|R_n(\beta_n)\| > \upsilon\right) \\ &\leq \varepsilon + \overline{B} P_0(\|R_n(\beta_n^*)\| > \upsilon), \end{aligned}$$

where $\beta_n^* \in B_n$ maximises $\beta \mapsto P_0\left(\|R_n(\beta_n)\| > \upsilon\right)$. As $(\beta_n^*)_{n\in\mathbb{N}}$ is a deterministic $\sqrt{n}$-consistent sequence for β we have that $P_0(\|R_n(\beta_n^*)\| > \upsilon) \to 0$ by equation (29).

By the version of (29) with θ_n replaced by $\bar{\theta}_n$ we have

$$\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\bar{\theta}_n} - \tilde{\ell}_{\theta_0}\right] = \sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\bar{\theta}_n} - \tilde{\ell}_{\bar{\theta}_n}\right] + \sqrt{n}\mathbb{P}_n\left[\tilde{\ell}_{\bar{\theta}_n} - \tilde{\ell}_{\theta_0}\right] = -\tilde{I}_{\theta_0}(0, \sqrt{n}(\bar{\beta}_n - \beta)')' + o_{P_0}(1).$$

and $\hat{\mathcal{K}}_{\bar{\theta}_n} \xrightarrow{P_0} \tilde{\mathcal{K}}_{\theta_0}$ for

$$\tilde{\mathcal{K}}_\theta \coloneqq \begin{bmatrix} I & -\tilde{I}_{\theta,\alpha\beta}\tilde{I}_{\theta,\beta\beta}^{-1} \end{bmatrix}, \quad \hat{\mathcal{K}}_\theta \coloneqq \begin{bmatrix} I & -\hat{I}_{\theta,\alpha\beta}\hat{I}_{\theta,\beta\beta}^{-1} \end{bmatrix}.$$

Combine these to obtain

$$\begin{aligned} &\sqrt{n}\mathbb{P}_n\left[\hat{\kappa}_{\bar{\theta}_n} - \tilde{\kappa}_{\theta_0}\right] \\ &\quad= \left(\hat{\mathcal{K}}_{\bar{\theta}_n} - \tilde{\mathcal{K}}_{\theta_0}\right)\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\bar{\theta}_n} - \tilde{\ell}_{\theta_0}\right] + \tilde{\mathcal{K}}_{\theta_0}\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\bar{\theta}_n} - \tilde{\ell}_{\bar{\theta}_n}\right] + \left(\hat{\mathcal{K}}_{\bar{\theta}_n} - \tilde{\mathcal{K}}_{\theta_0}\right)\sqrt{n}\mathbb{P}_n\tilde{\ell}_{\theta_0} \\ &\quad= -\tilde{\mathcal{K}}_{\theta_0}\tilde{I}_{\theta_0}(0, \sqrt{n}(\bar{\beta}_n - \beta)')' + o_{P_0}(1) \\ &\quad= -\begin{bmatrix} I & -\tilde{I}_{\theta_0,\alpha\beta}\tilde{I}_{\theta_0,\beta\beta}^{-1} \end{bmatrix}\begin{bmatrix} \tilde{I}_{\theta_0,\alpha\alpha} & \tilde{I}_{\theta_0,\alpha\beta} \\ \tilde{I}_{\theta_0,\beta\alpha} & \tilde{I}_{\theta_0,\beta\beta} \end{bmatrix}\begin{bmatrix} 0 \\ \sqrt{n}(\bar{\beta}_n - \beta) \end{bmatrix} + o_{P_0}(1) \\ &\quad= o_{P_0}(1). \end{aligned}$$

Then, by assumption 1-part 1, under P_0,

$$Z_n \coloneqq \sqrt{n}\mathbb{P}_n\hat{\kappa}_{\bar{\theta}_n} = \sqrt{n}\mathbb{P}_n\left[\hat{\kappa}_{\bar{\theta}_n} - \tilde{\kappa}_{\theta_0}\right] + \sqrt{n}\mathbb{P}_n\tilde{\kappa}_{\theta_0} \rightsquigarrow Z \sim \mathcal{N}(0, \tilde{\mathcal{I}}_{\theta_0})$$

Since $\hat{I}_{\bar{\theta}_n} \xrightarrow{P_0} \tilde{I}_{\theta_0} \succ 0$ an application of the continuous mapping theorem gives that $\hat{\mathcal{I}}_{\bar{\theta}_n}^{-1/2} \xrightarrow{P_0} \tilde{\mathcal{I}}_{\theta_0}^{-1/2}$. Combining this with Slutsky's lemma and the continuous mapping theorem once more, we conclude that $\hat{\mathcal{I}}_{\bar{\theta}_n}^{-1/2} Z_n \rightsquigarrow \tilde{\mathcal{I}}_{\theta_0}^{-1/2} Z$ which has a L_α dimensional standard normal distribution. Hence

$$\hat{S}_{\bar{\theta}_n} = (\hat{\mathcal{I}}_{\bar{\theta}_n}^{-1/2} Z_n)'(\hat{\mathcal{I}}_{\bar{\theta}_n}^{-1/2} Z_n) \rightsquigarrow \chi^2_{L_\alpha}.$$

□

Proof of Theorem 2. Let $P_0 \coloneqq P_{\theta_0}$, where θ_0 is defined in Assumption 2. The first step is to note that assumption 2 implies that

$$\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\theta_n} - \tilde{\ell}_{\theta_n}\right] \xrightarrow{P_0} 0, \quad \sqrt{n}\mathbb{P}_n\left[\tilde{\ell}_{\theta_n} - \tilde{\ell}_{\theta_0}\right] + \sqrt{n}\tilde{I}_{\theta_0}(0, (\beta_n - \beta)')' \xrightarrow{P_0} 0 \tag{30}$$

and

$$\nu_n^{-1}\left\|\hat{I}_{\theta_n} - \tilde{I}_{\theta_0}\right\| = o_{P_0}(1). \tag{31}$$

which together replace equation (29) in this setting.[28]

The next step is to show that (30) and (31) continue to hold with θ_n replaced by $\bar{\theta}_n$. The argument follows analogously to that for the corresponding terms in the proof of Theorem 1, with the definition of $R'''(\beta')$ changed to $R_n'''(\beta') \coloneqq \nu_n^{-1}[\hat{I}_{\theta'} - \tilde{I}_{\theta_0}]$.

By the version of (30) with θ_n replaced by $\bar{\theta}_n$ we have

$$\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\bar{\theta}_n} - \tilde{\ell}_{\theta_0}\right] = \sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\bar{\theta}_n} - \tilde{\ell}_{\bar{\theta}_n}\right] + \sqrt{n}\mathbb{P}_n\left[\tilde{\ell}_{\bar{\theta}_n} - \tilde{\ell}_{\theta_0}\right] = -\tilde{I}_{\theta_0}(0, \sqrt{n}(\bar{\beta}_n - \beta)')' + o_{P_0}(1).$$

and by the version of (31) with θ_n replaced by $\bar{\theta}_n$, $\hat{I}_{\bar{\theta}_n} \xrightarrow{P_0} \tilde{I}_{\theta_0}$ and so $\hat{\mathcal{K}}_{\bar{\theta}_n} \xrightarrow{P_0} \tilde{\mathcal{K}}_{\theta_0}$ for

$$\tilde{\mathcal{K}}_\theta \coloneqq \begin{bmatrix} I & -\tilde{I}_{\theta,\alpha\beta}\tilde{I}_{\theta,\beta\beta}^{-1} \end{bmatrix}, \quad \hat{\mathcal{K}}_\theta \coloneqq \begin{bmatrix} I & -\hat{I}_{\theta,\alpha\beta}\hat{I}_{\theta,\beta\beta}^{-1} \end{bmatrix}.$$

As in the proof of theorem 1 combine these to obtain

$$\begin{aligned}
&\sqrt{n}\mathbb{P}_n\left[\hat{\kappa}_{\bar{\theta}_n} - \tilde{\kappa}_{\theta_0}\right] \\
&\quad = \left(\hat{\mathcal{K}}_{\bar{\theta}_n} - \tilde{\mathcal{K}}_{\theta_0}\right)\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\bar{\theta}_n} - \tilde{\ell}_{\theta_0}\right] + \tilde{\mathcal{K}}_{\theta_0}\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\bar{\theta}_n} - \tilde{\ell}_{\theta_0}\right] + \left(\hat{\mathcal{K}}_{\bar{\theta}_n} - \tilde{\mathcal{K}}_{\theta_0}\right)\sqrt{n}\mathbb{P}_n\tilde{\ell}_{\theta_0} \\
&\quad = -\tilde{\mathcal{K}}_{\theta_0}\tilde{I}_{\theta_0}(0, \sqrt{n}(\bar{\beta}_n - \beta)')' + o_{P_0}(1) \\
&\quad = -\begin{bmatrix} I & -\tilde{I}_{\theta_0,\alpha\beta}\tilde{I}_{\theta_0,\beta\beta}^{-1} \end{bmatrix}\begin{bmatrix} \tilde{I}_{\theta_0,\alpha\alpha} & \tilde{I}_{\theta_0,\alpha\beta} \\ \tilde{I}_{\theta_0,\beta\alpha} & \tilde{I}_{\theta_0,\beta\beta} \end{bmatrix}\begin{bmatrix} 0 \\ \sqrt{n}(\bar{\beta}_n - \beta) \end{bmatrix} + o_{P_0}(1) \\
&\quad = o_{P_0}(1).
\end{aligned}$$

Then, by assumption 2-part 1, under P_0,

$$Z_n \coloneqq \sqrt{n}\mathbb{P}_n\hat{\kappa}_{\bar{\theta}_n} = \sqrt{n}\mathbb{P}_n\left[\hat{\kappa}_{\bar{\theta}_n} - \tilde{\kappa}_{\theta_0}\right] + \sqrt{n}\mathbb{P}_n\tilde{\kappa}_{\theta_0} \rightsquigarrow Z \sim \mathcal{N}(0, \tilde{\mathcal{I}}_{\theta_0}).$$

For the next step, observe that

$$\left\|\hat{\mathcal{I}}_{\bar{\theta}_n} - \tilde{\mathcal{I}}_{\theta_0}\right\|_2 \leq \left\|\hat{I}_{\bar{\theta}_n,\alpha\alpha} - \tilde{I}_{\theta_0,\alpha\alpha}\right\|_2 + \left\|\hat{I}_{\bar{\theta}_n,\alpha\beta}\hat{I}_{\bar{\theta}_n,\beta\beta}^{-1}\hat{I}_{\bar{\theta}_n,\beta\alpha} - \tilde{I}_{\theta_0,\alpha\beta}\tilde{I}_{\theta_0,\beta\beta}^{-1}\tilde{I}_{\theta_0,\beta\alpha}\right\|_2.$$

By repeated addition and subtraction along with the observations that any submatrix has a smaller operator norm than the original matrix and the matrix inverse is Lipschitz continuous

[28] That these equations hold can be demonstrated by arguing entirely analogously to in the proof of Theorem 1.

at a non-singular matrix we obtain

$$\left\|\hat{\mathcal{I}}_{\bar{\theta}_n} - \tilde{\mathcal{I}}_{\theta_0}\right\|_2 \lesssim \left\|\hat{I}_{\bar{\theta}_n} - \tilde{I}_{\theta_0}\right\|_2.$$

Hence by equation (31) with $\bar{\theta}_n$ replacing θ_n we have $P_0\left(\left\|\hat{\mathcal{I}}_{\bar{\theta}_n} - \tilde{\mathcal{I}}_{\theta_0}\right\|_2 < \nu_n\right) \to 1$.

The remainder of the proof is split into two cases. First consider the case where $\text{rank}(\tilde{\mathcal{I}}_{\theta_0}) = r > 0$. We first show that $\hat{\mathcal{I}}^t_{\bar{\theta}_n} \xrightarrow{P_0} \tilde{\mathcal{I}}_{\theta_0}$ and the rank estimate $r_n = \text{rank}(\hat{\mathcal{I}}^t_{\bar{\theta}_n})$ satisfies $P_0(\{r_n = r\}) \to 1$.

Let λ_l denote the lth largest eigenvalue of $\tilde{\mathcal{I}}_{\theta_0}$, similarly define $\hat{\lambda}_{l,n}$ for $\hat{\mathcal{I}}_{\bar{\theta}_n}$ and $\hat{\lambda}^t_{l,n}$ for $\hat{\mathcal{I}}^t_{\bar{\theta}_n}$. Define the set $R_n \coloneqq \{r_n = r\}$, let $\underline{\nu} \coloneqq \lambda_r/2 > 0$ and note that $\|\hat{\mathcal{I}}_{\bar{\theta}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 = o_{P_0}(\nu_n)$ implies that $\|\hat{\mathcal{I}}_{\bar{\theta}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 = o_{P_0}(1)$.

By Weyl's perturbation theorem[29] we have $\max_{l=1,\dots,L_\alpha} |\hat{\lambda}_{l,n} - \lambda_l| \leq \|\hat{\mathcal{I}}_{\bar{\theta}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 = o_{P_0}(1)$. Hence, if we define $E_n \coloneqq \{\hat{\lambda}_{r,n} \geq \nu_n\}$, for n large enough such that $\nu_n < \underline{\nu}$, we have

$$P_0(E_n) = P_0\left(\hat{\lambda}_{r,n} \geq \nu_n\right) \geq P_0\left(\hat{\lambda}_{r,n} \geq \underline{\nu}\right) \geq P_0\left(|\hat{\lambda}_{r,n} - \lambda_r| < \underline{\nu}\right) \to 1.$$

If $r = L_\alpha$ we have that $R_n \supset E_n$ and therefore $P_0(R_n) \to 1$. Additionally, if $\hat{\lambda}_{L_\alpha,n} \geq \nu_n$ then $\hat{\lambda}^t_{l,n} = \hat{\lambda}_{l,n}$ for each $l \in [L_\alpha]$ and hence $\hat{\mathcal{I}}^t_{\bar{\theta}_n} = \hat{\mathcal{I}}_{\bar{\theta}_n}$. Thus, $E_n \cap \{\|\hat{\mathcal{I}}_{\bar{\theta}_n} - \tilde{\mathcal{I}}_{\theta_0}\| \leq \upsilon\} \subset \{\|\hat{\mathcal{I}}^t_{\bar{\theta}_n} - \tilde{\mathcal{I}}_{\theta_0}\| \leq \upsilon\}$, from which it follows that $\hat{\mathcal{I}}^t_{\bar{\theta}_n} \xrightarrow{P_0} \tilde{\mathcal{I}}_{\theta_0}$.

Now suppose instead that $r < L_\alpha$ and define $F_n \coloneqq \{\hat{\lambda}_{r+1,n} < \nu_n\}$. It follows by Weyl's perturbation theorem and the fact that $\lambda_l = 0$ for $l > r$ that as $n \to \infty$

$$P(F_n) = P(\hat{\lambda}_{r+1,n} < \nu_n) \geq P(\|\hat{\mathcal{I}}_{\bar{\theta}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 < \nu_n) \to 1.$$

Since $R_n \supset E_n \cap F_n$, this implies that $P(R_n) \to 1$ as $n \to \infty$. Additionally, if $\hat{\lambda}_{r,n} \geq \nu_n$, $\hat{\lambda}_{r+1,n} < \nu_n$ and $\|\hat{\mathcal{I}}_{\bar{\theta}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 \leq \upsilon$, we have that $\hat{\lambda}^t_{k,n} = \hat{\lambda}_{k,n}$ for $k \leq r$ and $\hat{\lambda}^t_{l,n} = 0 = \lambda_l$ for $l > r$ and so

$$\|\hat{\Lambda}_n(\nu_n) - \Lambda\|_2 = \max_{l=1,\dots,r} |\hat{\lambda}^t_{l,n} - \lambda_l| = \max_{l=1,\dots,r} |\hat{\lambda}_{l,n} - \lambda_l| \leq \|\hat{\Lambda}_n - \Lambda\|_2 \leq \|\hat{\mathcal{I}}_{\bar{\theta}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 \leq \upsilon,$$

and hence $\{\|\hat{\mathcal{I}}_{\bar{\theta}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 \leq \upsilon\} \cap E_n \cap F_n \subset \{\|\hat{\Lambda}_n(\nu_n) - \Lambda\|_2 \leq \upsilon\}$, from which it follows that $\hat{\Lambda}_n(\nu_n) \xrightarrow{P_0} \Lambda$.

To complete this part of the proof, suppose that $(\lambda_1, \dots, \lambda_r)$ consists of s distinct eigenvalues with values $\lambda^1 > \lambda^2 > \cdots > \lambda^s$ and multiplicities $\mathfrak{m}_1, \dots, \mathfrak{m}_s$ (each at least one), where the superscripts on the λs are indices, not exponents. $\lambda^{s+1} = 0$ is an eigenvalue with multiplicity $\mathfrak{m}_{s+1} = L_\alpha - r$. Let l^k_i for $k = 1, \dots, s+1$ and $i = 1, \dots, \mathfrak{m}_k$ denote the column indices of the eigenvectors in U corresponding to each λ^k. For each λ^k, the total

[29] E.g. Corollary III.2.6 in Bhatia (1997).

eigenprojection is $\Pi_k := \sum_{i=1}^{\mathfrak{m}_k} u_{l_i^k} u'_{l_i^k}$.[30] Total eigenprojections are continuous.[31] Therefore, if we construct $\hat{\Pi}_{k,n}$ in in an analogous fashion to Π_k but replace columns of U with columns of $\hat{U}_n$, we have $\hat{\Pi}_{k,n} \xrightarrow{P_0} \Pi_k$ for each $k \in [s+1]$ since $\hat{\mathcal{I}}_{\bar{\theta}_n} \xrightarrow{P_0} \tilde{\mathcal{I}}_{\theta_0}$. Spectrally decompose $\tilde{\mathcal{I}}_{\theta_0}$ as $\tilde{\mathcal{I}}_{\theta_0} = \sum_{k=1}^{s} \lambda^k \Pi_k$, where the sum runs to s rather than $s+1$ since $\lambda^{s+1} = 0$. Then,

$$\hat{\mathcal{I}}^t_{\bar{\theta}_n} = \sum_{k=1}^{s+1} \sum_{i=1}^{\mathfrak{m}_k} \hat{\lambda}^t_{l_i^k,n} \hat{u}_{l_i^k,n} \hat{u}'_{l_i^k,n} = \sum_{k=1}^{s+1} \sum_{i=1}^{\mathfrak{m}_k} (\hat{\lambda}^t_{l_i^k,n} - \lambda^k) \hat{u}_{l_i^k,n} \hat{u}'_{l_i^k,n} + \sum_{k=1}^{s} \lambda^k \hat{\Pi}_{k,n},$$

and so

$$\|\hat{\mathcal{I}}^t_{\bar{\theta}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 \leq \sum_{k=1}^{s+1} \sum_{i=1}^{\mathfrak{m}_k} |\hat{\lambda}^t_{l_i^k,n} - \lambda^k| \|\hat{u}_{l_i^k,n} \hat{u}'_{l_i^k,n}\|_2 + \sum_{k=1}^{s} |\lambda^k| \|\hat{\Pi}_{k,n} - \Pi_k\|_2 \xrightarrow{P_0} 0,$$

by $\hat{\Pi}_{k,n} \xrightarrow{P} \Pi_k$, $\hat{\Lambda}_n(\nu_n) \xrightarrow{P_0} \Lambda$ and since we have $\|u_{l_i^k,n} u'_{l_i^k,n}\|_2 = 1$ for any i, k, n.

Hence, we have that $\hat{\mathcal{I}}^t_{\bar{\theta}_n} \xrightarrow{P_0} \tilde{\mathcal{I}}_{\theta_0}$ and $P_0(\{r_n = r\}) \to 1$. This implies that $\hat{\mathcal{I}}^{t,\dagger}_{\bar{\theta}_n} \xrightarrow{P_0} \tilde{\mathcal{I}}^{\dagger}_{\theta_0}$ where $\tilde{\mathcal{I}}^{\dagger}_{\theta_0}$ is the Moore-Penrose inverse of $\tilde{\mathcal{I}}_{\theta_0}$.[32]

Now consider the score statistic $\hat{S}^{SR}_{\bar{\theta}_n}$, by Slutsky's lemma and the continuous mapping theorem we have that

$$\hat{S}^{SR}_{\bar{\theta}_n} = Z'_n \hat{\mathcal{I}}^{t,\dagger}_{\bar{\theta}_n} Z_n \rightsquigarrow Z' \tilde{\mathcal{I}}^{\dagger}_{\theta_0} Z \sim \chi^2_r$$

where the distributional result $X := Z' \tilde{\mathcal{I}}^{\dagger}_{\theta_0} Z \sim \chi^2_r$, follows from e.g. Theorem 9.2.2 in Rao and Mitra (1971).

Finally, recall that $R_n = \{r_n = r\}$. On these sets c_n is the $1-a$ quantile of the χ^2_r distribution, which we will call c. Hence, we have $c_n \xrightarrow{P_0} c$ as $P_0(R_n) \to 1$. As a result, we obtain $\hat{S}^{SR}_{\bar{\theta}_n} - c_n \rightsquigarrow X - c$ where $X \sim \chi^2_r$. Since the χ^2_r distribution is continuous, we have by the Portmanteau theorem

$$P_0\left(\hat{S}^{SR}_{\bar{\theta}_n} > c_n\right) = 1 - P_0\left(\hat{S}^{SR}_{\bar{\theta}_n} - c_n \leq 0\right) \to 1 - P_0\left(X - c \leq 0\right) = 1 - P_0\left(X \leq c\right) = a \ ,$$

which completes the proof in the case that $r > 0$.

It remains to handle the case with $r = 0$. We first note that $Z_n \rightsquigarrow Z \sim \mathcal{N}(0, \tilde{\mathcal{I}}_{\theta_0})$ continues to hold by our assumptions, though in this case $\tilde{\mathcal{I}}_{\theta_0}$ is the zero matrix and hence the limiting distribution is degenerate: $Z = 0$ a.s.. Let $E_n = \{r_n = 0\}$. Part 3 of assumption 2 and Weyl's perturbation theorem imply that

$$P_0(E_n) = P_0\left(r_n = 0\right) = P_0\left(\max_{l=1,\ldots,L_\alpha} |\hat{\lambda}_{n,l}| < \nu_n\right) \geq P_0\left(\|\hat{\mathcal{I}}_{\bar{\theta}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 < \nu_n\right) \to 1.$$

On the sets E_n we have that $\hat{\mathcal{I}}^t_{\bar{\theta}_n}$ is the zero matrix, whose Moore-Penrose inverse is also the zero matrix. Hence on the sets E_n we have $\hat{S}^{SR}_{\bar{\theta}_n} = 0$ and $c_n = 0$ and therefore do not reject,

[30] See e.g Chapter 8.8 of Magnus and Neudecker (2019).
[31] E.g. Theorem 8.7 of Magnus and Neudecker (2019).
[32] See e.g. Theorem 2 of Andrews (1987).

implying

$$P_0(\hat{S}^{SR}_{\bar{\theta}_n} > c_n) \leq 1 - P_0(E_n) \to 0.$$

It follows that $P_0(\hat{S}^{SR}_{\bar{\theta}_n} > c_n) \to 0$. □

Proof of Lemma 1. We have that $\tilde{\mathcal{I}}^{\dagger}_{\theta_0} = \tilde{\mathcal{I}}^{-1}_{\theta_0}$ and can write

$$\hat{S}^{SR}_n - \hat{S}_n = Z'_n \left[\hat{\mathcal{I}}^{t,\dagger}_{\bar{\theta}_n} - \hat{\mathcal{I}}^{-1}_{\bar{\theta}_n}\right] Z_n,$$

where $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\kappa}_{\bar{\theta}_n}(Y_i) = O_{P_0}(1)$ as shown in the proof of Theorem 1. We have that each $\hat{\lambda}_{n,i} \xrightarrow{P_0} \lambda_i > 0$ where $\{\lambda_i\}_{i=1}^{L_\alpha}$ are the eigenvalues of $\tilde{\mathcal{I}}_{\theta_0}$ (ordered non-increasingly) and $\{\hat{\lambda}_{n,i}\}_{i=1}^{L}$ are the non-increasing eigenvalues of $\hat{\mathcal{I}}_{\bar{\theta}_n}$. Since $\nu_n \to 0$, it follows that with probability approaching one, $\hat{\mathcal{I}}^{t}_{\bar{\theta}_n} = \hat{\mathcal{I}}_{\bar{\theta}_n}$ and this matrix is of full rank. Hence, with probability approaching one $\hat{\mathcal{I}}^{t,\dagger}_{\bar{\theta}_n} = \hat{\mathcal{I}}^{-1}_{\bar{\theta}_n}$, implying that $\hat{\mathcal{I}}^{t,\dagger}_{\bar{\theta}_n} - \hat{\mathcal{I}}^{-1}_{\bar{\theta}_n} = o_{P_0}(1)$. □

Proof of Proposition 2. The proof amounts to verifying assumptions 0 and 2 for the LSEM under Assumption 5 given a suitable log score estimator as defined in Assumption 6. The proposition then will follow by Theorem 2. The regularity assumption 0 follows by lemma S4 in the supplementary material.

First, we note that assumption 2-part 1 follows by the CLT since our data is iid and the efficient score $\tilde{\ell}_{\theta_0}$ lies in $L_2(P_0)$ by construction. Next, let $\theta_n = (\alpha_0, \beta_n, \eta)$ and note that under P_{θ_n}, each $A_{n,k}(Z_i - B_n X_i) \simeq \epsilon_{i,k} \sim \eta_k$. Hence we can compute certain properties of the efficient score using the equality in distribution:

$$\tilde{\ell}_{\theta_n,(\alpha,\beta_1),l}(Y_i) \simeq \sum_{k=1}^K \sum_{j=1,j\neq k}^K \zeta_{l,k,j,n} \phi_k(\epsilon_{i,k})\epsilon_{i,j} + \sum_{k=1}^K \zeta_{l,k,k,n} \left[\tau_{k,1}\epsilon_{i,k} + \tau_{k,2}\kappa(\epsilon_{i,k})\right] \tag{32}$$

$$\tilde{\ell}_{\theta_n,b,l}(Y_i) \simeq \sum_{k=1}^K [-A_{n,k\bullet} D_{b,l}] \left[(X - \mathbb{E}X)\phi_k(\epsilon_{i,k}) - \mathbb{E}X \left(\varsigma_{k,1}\epsilon_{i,k} + \varsigma_{k,2}\kappa(\epsilon_{i,k})\right)\right] \tag{33}$$

where we note that the same observation implies that $\tau_{k,n} = \tau_k$ and $\varsigma_{k,n} = \varsigma_k$ for each n.[33] By our assumptions on the map $(\alpha, \beta_1) \mapsto A(\alpha, \beta_1)$, we have $\zeta_{l,k,j,n} \to \zeta_{l,k,j,\infty} := [D_l(\gamma_0)]_{k\bullet} A(\gamma_0)^{-1}_{\bullet j}$ for $\gamma = (\alpha_0, \beta)$. Note that the entries of $D_{b,l}$ are all zero except for entry l (corresponding to b_l) which is equal to one.

We verify assumption 1-part 2 for each component of the efficient score (32) & (33). For (32) and $y = (z, x)$ we define

$$\varphi_{1,n}(y) := \sum_{k=1}^K \sum_{j=1,j\neq k}^K \zeta_{l,k,j,n} \phi_k(A_{n,k\bullet} v_n) A_{n,j\bullet} v_n \ ,$$

[33] In the preceding display we have written $\zeta_{l,k,j,n}$ rather than $\zeta_{l,k,j}$ to indicate their dependence on β_n. $\zeta_{l,k,j,\infty}$ corresponds to evaluation at the point (α_0, β).

and

$$\hat{\varphi}_{1,n}(y) \coloneqq \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \zeta_{l,k,j,n} \hat{\phi}_{k,n}(A_{n,k\bullet} v_n) A_{n,j\bullet} v_n \ ,$$

with $v_n = z - B_n x$, and let $\overline{\zeta}_n \coloneqq \max_{l\in[L], j\in[K], k\in[K]} |\zeta_{l,j,k,n}|$ which converges to $\overline{\zeta} \coloneqq \max_{l\in[L], j\in[K], k\in[K]} |\zeta_{l,j,k,\infty}| < \infty$. We have that

$$\sqrt{n}\mathbb{P}_n(\hat{\varphi}_{1,n} - \varphi_{1,n}) \leq \sqrt{n} \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \overline{\zeta}_n \left| \frac{1}{n} \sum_{i=1}^{n} \hat{\phi}_{k,n}(V_{i,k,n}) V_{i,j,n} - \phi_k(V_{i,k,n}) V_{i,j,n} \right| \ ,$$

with $V_{i,j,n} = A_{n,j\bullet}(Z_i - B_n X_i)$. Since each $\left| \frac{1}{n} \sum_{i=1}^{n} \hat{\phi}_{k,n}(V_{i,k,n}) V_{i,j,n} - \phi_k(V_{i,k,n}) V_{i,j,n} \right| = o_{P_{\theta_n}}(n^{-1/2})$ by applying Assumption 6 with $W_{i,n} = V_{i,j,n}$ (noting that under P_{θ_n}, $V_{i,k,n} \simeq \epsilon_{k,i}$ and $V_{i,j,n} \simeq \epsilon_{j,i}$ are independent with $\mathbb{E}_{\theta_n} V_{i,j,n}^2 = 1$ by Assumption 3) and the outside summations are finite, it follows that

$$\sqrt{n}\mathbb{P}_n(\hat{\varphi}_{1,n} - \varphi_{1,n}) = o_{P_{\theta_n}}(1). \tag{34}$$

Next, we note that $\hat{\tau}_{k,n} - \tau_k \to 0$ and $\hat{\varsigma}_{k,n} - \varsigma_k \to 0$ in P_{θ_n}-probability by Lemma 7 where $\hat{\tau}_{k,n}$ is defined in (18) with Y_i replaced by $Z_i - B_n X_i$ and $\hat{\varsigma}_{k,n}$ is defined analogously with $(1, 0)'$ replacing $(0, -2)'$.

Now, consider $\varphi_{2,\tau,n}(y)$ defined by

$$\varphi_{2,\tau,n}(y) \coloneqq \sum_{k=1}^{K} \zeta_{l,k,k,n} \left[\tau_{k,1} A_{n,k\bullet} v_n + \tau_{k,2} \kappa(A_{n,k\bullet} v_n) \right].$$

Since sum is finite and each $|\zeta_{l,k,k,n}| \to |\zeta_{l,k,k,\infty}| < \infty$ it is sufficient to consider the convergence of the summands. In particular we have that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\hat{\tau}_{k,n,1} - \tau_{k,1} \right] V_{i,k,n} = \left[\hat{\tau}_{k,n,1} - \tau_{k,1} \right] \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_{i,k,n} = o_{P_{\theta_n}}(1) \times O_{P_{\theta_n}}(1) = o_{P_{\theta_n}}(1),$$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\hat{\tau}_{k,n,2} - \tau_{k,2} \right] \kappa(V_{i,k,n}) = \left[\hat{\tau}_{k,n,2} - \tau_{k,2} \right] \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \kappa(V_{i,k,n}) = o_{P_{\theta_n}}(1) \times O_{P_{\theta_n}}(1) = o_{P_{\theta_n}}(1).$$

since $V_{i,k,n} \simeq \epsilon_{k,i} \sim \eta_k$ under P_{θ_n} and $(\epsilon_{i,k})_{i\geq 1}$ and $(\kappa(\epsilon_{i,k}))_{i\geq 1}$ are i.i.d. mean-zero sequences with finite second moments such that the CLT holds. Together these yield that

$$\sqrt{n}\mathbb{P}_n(\varphi_{2,\hat{\tau}_n,n} - \varphi_{2,\tau,n}) = o_{P_{\theta_n}}(1). \tag{35}$$

Putting (34) and (35) together yields the required convergence for components of the type (32), since $\tilde{\ell}_{\theta_n,(\alpha,\beta_1),l} = \varphi_{1,n} + \varphi_{2,\tau,n}$ and $\hat{\ell}_{\theta_n,(\alpha,\beta_1),l} = \hat{\varphi}_{1,n} + \varphi_{2,\hat{\tau}_n,n}$.

Next, we consider components (33). Let $a_{n,k,l} \coloneqq -A_{n,k\bullet}D_{b,l}$ and write

$$\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\theta_n,b,l} - \tilde{\ell}_{\theta_n,b,l}\right] = \sum_{k=1}^{K} a_{n,k,l}\sqrt{n}\mathbb{P}_n\left[(X_i - \mathbb{E}X_i)[\hat{\phi}_k(V_{i,k,n}) - \phi_k(V_{i,k,n})] + (\mathbb{E}X_i - \bar{X}_n)\phi_k(V_{i,k,n})\right]$$
$$+ \sum_{k=1}^{K} a_{n,k,l}\sqrt{n}\mathbb{P}_n\left[(\mathbb{E}X_i - \bar{X}_n)[\hat{\varsigma}_{k,n,1}V_{i,k,n} + \hat{\varsigma}_{k,n,2}\kappa(V_{i,k,n})]\right]$$
$$- \sum_{k=1}^{K} a_{n,k,l}\sqrt{n}\mathbb{P}_n\left[\mathbb{E}X_i[(\hat{\varsigma}_{k,n,1} - \varsigma_{k,1})V_{i,k,n} + (\hat{\varsigma}_{k,n,2} - \varsigma_{k,2})\kappa(V_{i,k,n})]\right]$$

Taking the right hand side terms (inside the outer summation) in order, we have that $\sqrt{n}\mathbb{P}_n(X_i - \mathbb{E}X_i)[\hat{\phi}_k(V_{i,k,n}) - \phi_k(V_{i,k,n})] = o_{P_{\theta_n}}(1)$ by assumption 6 applied with $W_{i,n} = X_i - \mathbb{E}X_i$. For the second, $\sqrt{n}\mathbb{P}_n(\mathbb{E}X_i - \bar{X}_n)\phi_k(V_{i,k,n}) = (\mathbb{E}X_i - \bar{X}_n)\sqrt{n}\mathbb{P}_n\phi_k(V_{i,k,n}) = o_{P_{\theta_n}}(1) \times O_{P_{\theta_n}}(1) = o_{P_{\theta_n}}(1)$ by the WLLN & CLT, noting for the latter that $V_{i,k,n} \simeq \epsilon_{i,k}$. We know from Lemma 7 that $\varsigma_{k,n} \xrightarrow{P_{\theta_n}} \varsigma_k$ and hence adding & subtracting and using the WLLN & CLT again yields that $\sqrt{n}\mathbb{P}_n(\mathbb{E}X_i - \bar{X}_n)[\hat{\varsigma}_{k,n,1}V_{i,k,n} + \hat{\varsigma}_{k,n,2}\kappa(V_{i,k,n})] = o_{P_{\theta_n}}(1)$. The CLT & $\varsigma_{k,n} \xrightarrow{P_{\theta_n}} \varsigma_k$ ensure that $\sqrt{n}\mathbb{P}_n[(\hat{\varsigma}_{k,n,1} - \varsigma_{k,1})V_{i,k,n} + (\hat{\varsigma}_{k,n,2} - \varsigma_{k,2})\kappa(V_{i,k,n})] = o_{P_{\theta_n}}(1)$. Together these observations and that $a_{n,k,l} \to a_{\infty,n,l} \coloneqq A_{k\bullet}D_{b,l}$ imply that the required condition, $\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\theta_n,b,l} - \tilde{\ell}_{\theta_n,b,l}\right] = o_{P_{\theta_n}}(1)$, is satisfied.

To verify part 3 we will show that

$$\left\|\hat{I}_{\theta_n} - \tilde{I}_{\theta_0}\right\|_2 \leq \left\|\hat{I}_{\theta_n} - \tilde{I}_{\theta_n}\right\|_2 + \left\|\tilde{I}_{\theta_n} - \tilde{I}_{\theta_0}\right\|_2 = o_{P_{\theta_n}}(\nu_n^{1/2}). \tag{36}$$

where $\tilde{I}_{\theta_n} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \tilde{\ell}_{\theta_n}(Y_i)\tilde{\ell}_{\theta_n}(Y_i)'$. To obtain the rates we start with $\|\tilde{I}_{\theta_n} - \tilde{I}_{\theta_0}\|_2$, for which we show that each component satisfies the required rate. To set this up, let $Q^{r,s}_{l,m,i,n} = \tilde{\ell}_{\theta_n,r,l}(Y_i)\tilde{\ell}_{\theta_n,s,m}(Y_i) - \tilde{\ell}_{\theta_0,r,l}(Y_i)\tilde{\ell}_{\theta_0,s,m}(Y_i)$, where $r, s \in \{(\alpha,\beta_1), b\}$ and l, m denote the indices of the components of the efficient scores. Let $\breve{Q}^{r,s}_{l,m,i,n}$ be defined analogously with $V_{i,k,n}$ replaced by $\epsilon_{i,k}$. Under P_{θ_n} we have that $Q^{r,s}_{l,m,i,n} \simeq \breve{Q}^{r,s}_{l,m,i,n}$. Therefore to show $[\tilde{I}_{\theta_n} - \tilde{I}_{\theta_0}]_{l,m} = o_{P_{\theta_n}}(\nu_n^{1/2})$ it suffices to show that for any r, s and l, m

$$\frac{1}{n}\sum_{i=1}^{n} \breve{Q}^{r,s}_{l,m,i,n} - G\breve{Q}^{r,s}_{l,m,i,n} + \frac{1}{n}\sum_{i=1}^{n} G[\breve{Q}^{r,s}_{l,m,i,n} - \breve{Q}^{r,s}_{l,m,i,\infty}] = o_G(\nu_n^{1/2}),$$

where G is the product measure $\prod_{k=0}^{K} G_k$ and each $\breve{Q}^{r,s}_{l,m,i,n}$ is shown to satisfy $\|\breve{Q}^{r,s}_{l,m,i,n}\|_{G,p} < \infty$ in Lemma 6 given below. The convergence of the second term follows from the assumed Lipschitz continuity of the map defining the ζ's and the $\sqrt{n}$-consistency of β_n for β, since $n^{-1/2} = o(\nu_n^{1/2})$.[34] For the first term, if $p = 2$ in lemma 6, by Theorem 2.5.11 in Durrett

[34]Note that for large enough $n \in \mathbb{N}$ β_n is in a ball of radius, say, $\delta > 0$ around β. The (continuous) differentiability of $(\alpha,\beta_1) \mapsto A(\alpha,\beta_1)$ and the fact that $D_{b,l}$ is a constant matrix implies that the map $(\alpha,\beta_1) \mapsto [-A(\alpha,\beta_1)_{k\bullet}D_{b,l}]$ is Lipschitz on this set.

(2019), we have that for all $\iota > 0$

$$\frac{1}{n}\sum_{i=1}^{n} \breve{Q}^{r,s}_{l,m,i,n} - G\breve{Q}^{r,s}_{l,m,i,n} = o_G\left(n^{-1/2}\log(n)^{1/2+\iota}\right).$$

It follows that

$$\|\tilde{I}_{\theta_n} - \tilde{I}_{\theta_0}\|_2 \leq \|\tilde{I}_{\theta_n} - \tilde{I}_{\theta_0}\|_F = o_{P_{\theta_n}}\left(n^{-1/2}\log(n)^{1/2+\iota}\right).$$

If, instead, $p = 1 + \nu/4 < 2$ in Lemma 6, then by the Marcinkiewicz & Zygmund SLLN (e.g. Theorem 2.5.12 in Durrett, 2019)

$$\frac{1}{n}\sum_{i=1}^{n} \breve{Q}^{r,s}_{l,m,i,n} - G\breve{Q}^{r,s}_{l,m,i,n} = o_G\left(n^{\frac{1-p}{p}}\right),$$

and similarly

$$\|\tilde{I}_{\theta_n} - \tilde{I}_{\theta_0}\|_2 \leq \|\tilde{I}_{\theta_n,n} - \tilde{I}_{\theta_0}\|_F = o_{P_{\theta_n}}\left(n^{\frac{1-p}{p}}\right).$$

That is, for any $p \in (1,2]$ we have $\|\tilde{I}_{\theta_n} - \tilde{I}_{\theta_0}\|_2 = o_{P_{\theta_n}}(\nu_{n,p}) = o_{P_{\theta_n}}(\nu_n^{1/2})$.

For the other component of the sum, let $r \in \{(\alpha,\beta_1), b\}$ and let l denote an index, we write $\hat{U}_{n,i,r,l} \coloneqq \hat{\ell}_{\theta_n,r,l}(Y_i)$, $\tilde{U}_{i,r,l} \coloneqq \tilde{\ell}_{\theta_n,r,l}(Y_i)$ and $D_{n,i,r,l} \coloneqq \hat{\ell}_{\theta_n,r,l}(Y_i) - \tilde{\ell}_{\theta_n,r,l}(Y_i)$.

Since it is the absolute value of the $(r,l)-(s,m)$ component of $\hat{I}_{\theta_n,n} - \tilde{I}_{\theta_0,n}$, it is sufficient to show that $\left|\frac{1}{n}\sum_{i=1}^n \hat{U}_{n,i,r,l}D_{n,i,s,m} + \frac{1}{n}\sum_{i=1}^n D_{n,i,r,l}\tilde{U}_{i,s,m}\right| = o_{P_{\theta_n}}(\nu_n^{1/2})$ as $n \to \infty$ for any $r,s \in \{(\alpha,\beta_1), b\}$ and l,m. By Cauchy-Schwarz and lemma 8

$$\left|\frac{1}{n}\sum_{i=1}^{n} D_{n,i,r,l}\tilde{U}_{i,s,m}\right| \leq \left(\frac{1}{n}\sum_{i=1}^{n}\tilde{U}^2_{i,s,m}\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n} D^2_{n,i,r,l}\right)^{1/2} = O_{P_{\theta_n}}(1)\times o_{P_{\theta_n}}(\nu_n^{1/2}) = o_{P_{\theta_n}}(\nu_n^{1/2}),$$

$$\left|\frac{1}{n}\sum_{i=1}^{n} \hat{U}_{n,i,r,l}D_{n,i,s,m}\right| \leq \left(\frac{1}{n}\sum_{i=1}^{n}\hat{U}^2_{n,i,r,l}\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n} D^2_{n,i,s,m}\right)^{1/2} = O_{P_{\theta_n}}(1)\times o_{P_{\theta_n}}(\nu_n^{1/2}) = o_{P_{\theta_n}}(\nu_n^{1/2}),$$

for any $(r,l)-(s,m)$. It follows that

$$\left[\frac{1}{n}\sum_{i=1}^{n} \hat{U}_{n,i,r,l}D_{n,i,s,m} + D_{n,i,r,l}\tilde{U}_{i,s,m}\right]^2 \leq 2\left[\frac{1}{n}\sum_{i=1}^{n} \hat{U}_{n,i,r,l}D_{n,i,s,m}\right]^2 + 2\left[\frac{1}{n}\sum_{i=1}^{n} D_{n,i,r,l}\tilde{U}_{i,s,m}\right]^2 = o_{P_{\theta_n}}(\nu_n)$$

and hence $\|\hat{I}_{\theta_n,n} - \tilde{I}_{\theta_0,n}\|_2 \leq \|\hat{I}_{\theta_n,n} - \tilde{I}_{\theta_0,n}\|_F = o_{P_{\theta_n}}(\nu_n^{1/2})$. We can combine these results to obtain:

$$\|\hat{I}_{\theta_n,n} - \tilde{I}_{\theta_0}\|_2 \leq \|\hat{I}_{\theta_n,n} - \tilde{I}_{\theta_n,n}\|_2 + \|\tilde{I}_{\theta_n,n} - \tilde{I}_{\theta_0}\|_2 = o_{P_{\theta_n}}(\nu_n^{1/2}) + o_{P_{\theta_n}}(\nu_n^{1/2}) = o_{P_{\theta_n}}(\nu_n^{1/2}).$$

It remains to show that part 4 holds. Recall that the dominating measure here is λ and

re-write the integral in question as

$$\int \left\| \tilde{\ell}_{\theta_n} p_{\theta_n}^{1/2} - \tilde{\ell}_{\theta_0} p_{\theta_0}^{1/2} \right\|^2 \mathrm{d}\lambda = \sum_{l=1}^{L} \int \left[\tilde{\ell}_{\theta_n, l} p_{\theta_n}^{1/2} - \tilde{\ell}_{\theta_0, l} p_{\theta_0}^{1/2} \right]^2 \mathrm{d}\lambda. \tag{37}$$

It is evidently sufficient to show that each of the integrals in the sum on the rhs converges to zero. To this end, let $f_{r,n} \coloneqq \tilde{\ell}_{\theta_n, r, l} p_{\theta_n}^{1/2}$ and $f_r \coloneqq \tilde{\ell}_{\theta_0, r, l} p_{\theta_0}^{1/2}$ for $r \in \{(\alpha, \beta_1), b\}$ corresponding to (32) & (33) for some arbitrary l. By the expressions for $\tilde{\ell}_{\theta_n}$ and p_{θ_n} given in lemma 3 and equation (17) respectively along with the continuity of A, D_l and each η_k and ϕ_k (each of which follows from our assumptions), we have that $f_{r,n} \to f_r$ λ-a.e. for all r. Moreover, using the representation in (32) we have

$$\begin{aligned}
\int f_{(\alpha,\beta_1),n}^2 \, \mathrm{d}\lambda &= \int \left(\sum_{k=1}^{K} \left[\zeta_{l,k,k,n} \left[\tau_{k,1} \epsilon_{k,i} + \tau_{k,2} \kappa(\epsilon_{k,i}) \right] + \sum_{j=1, j \neq k}^{K} \zeta_{l,k,j,n} \phi_k(\epsilon_{k,i}) \epsilon_{j,i} \right] \right)^2 \mathrm{d}G \\
&= \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \sum_{b=1}^{K} \sum_{m=1, m\neq b}^{K} \zeta_{l,k,j,n} \zeta_{l,b,m,n} \int \phi_k(\epsilon_{k,i}) \epsilon_{j,i} \phi_b(\epsilon_{b,i}) \epsilon_{m,i} \, \mathrm{d}G \\
&\quad + 2 \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \sum_{b=1}^{K} \zeta_{l,k,j,n} \zeta_{l,b,b,n} \int \phi_k(\epsilon_{k,i}) \epsilon_{j,i} \left[\tau_{b,1} \epsilon_{b,i} + \tau_{b,2} \kappa(\epsilon_{b,i}) \right] \mathrm{d}G \\
&\quad + \sum_{k=1}^{K} \sum_{b=1}^{K} \zeta_{l,k,k,n} \zeta_{l,b,b,n} \int \left[\tau_{b,1} \epsilon_{b,i} + \tau_{b,2} \kappa(\epsilon_{b,i}) \right] \left[\tau_{k,1} \epsilon_{k,i} + \tau_{k,2} \kappa(\epsilon_{k,i}) \right] \mathrm{d}G
\end{aligned}$$

where G is the law of ϵ and each of the integrals are finite by assumption 3. By the continuity of A and D_l, this converges to

$$\begin{aligned}
\int f_{(\alpha,\beta_1)}^2 \, \mathrm{d}\lambda &= \int \left(\sum_{k=1}^{K} \left[\zeta_{l,k,k,\infty} \left[\tau_{k,1} \epsilon_{k,i} + \tau_{k,2} \kappa(\epsilon_{k,i}) \right] + \sum_{j=1, j \neq k}^{K} \zeta_{l,k,j,\infty} \phi_k(\epsilon_{k,i}) \epsilon_{j,i} \right] \right)^2 \mathrm{d}G \\
&= \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \sum_{b=1}^{K} \sum_{m=1, m\neq b}^{K} \zeta_{l,k,j,\infty} \zeta_{l,b,m,\infty} \int \phi_k(\epsilon_{k,i}) \epsilon_{j,i} \phi_b(\epsilon_{b,i}) \epsilon_{m,i} \, \mathrm{d}G \\
&\quad + 2 \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \sum_{b=1}^{K} \zeta_{l,k,j,\infty} \zeta_{l,b,b,\infty} \int \phi_k(\epsilon_{k,i}) \epsilon_{j,i} \left[\tau_{b,1} \epsilon_{b,i} + \tau_{b,2} \kappa(\epsilon_{b,i}) \right] \mathrm{d}G \\
&\quad + \sum_{k=1}^{K} \sum_{b=1}^{K} \zeta_{l,k,k,\infty} \zeta_{l,b,b,\infty} \int \left[\tau_{b,1} \epsilon_{b,i} + \tau_{b,2} \kappa(\epsilon_{b,i}) \right] \left[\tau_{k,1} \epsilon_{k,i} + \tau_{k,2} \kappa(\epsilon_{k,i}) \right] \mathrm{d}G,
\end{aligned}$$

which is finite by assumption 3. By Proposition 2.29 in van der Vaart (1998) we conclude that $\int (f_{(\alpha,\beta_1),n} - f_{(\alpha,\beta_1)})^2 \, \mathrm{d}\lambda \to 0$. Analogous arguments hold for $r = b$; we omit the details. The convergence of each $\int (f_{r,n} - f_r)^2 \, \mathrm{d}\lambda \to 0$ in conjunction with equation (37) is sufficient for part 4. □

Lemma 4. *Suppose that assumption 3 holds and let $k, j, s, b \in [K]$ with $j \neq k$ and $s \neq b$.*

Then, for G the law of ϵ and any $p \in [1,2]$ we have that

(i) $\|\phi_k(\epsilon_k)\epsilon_j\phi_s(\epsilon_s)\epsilon_b\|_{G,p} < \infty$,

(ii) $\|\phi_k(\epsilon_k)\epsilon_j\epsilon_s\|_{G,p} < \infty$,

(iii) $\|\epsilon_k\epsilon_s\|_{G,p} < \infty$.

Proof. By Cauchy-Schwarz, independence and our moment conditions we have

$$\begin{aligned}
\|\phi_k(\epsilon_k)\epsilon_j\phi_s(\epsilon_s)\epsilon_b\|_{G,p} &\leq \left[G[\phi_k(\epsilon_k)]^{2p}G[\epsilon_j]^{2p}G[\phi_s(\epsilon_s)]^{2p}G[\epsilon_b]^{2p}\right]^{\frac{1}{2p}} < \infty, \\
\|\phi_k(\epsilon_k)\epsilon_j\epsilon_s\|_{G,p} &\leq \left[G[\phi_k(\epsilon_k)]^{2p}G[\epsilon_j]^{2p}G[\epsilon_s]^{2p}\right]^{1/(2p)} < \infty, \\
\|\epsilon_k\epsilon_s\|_{G,p} = \|(\epsilon_k)^p(\epsilon_s)^p\|_{G,1}^{1/p} &\leq \|(\epsilon_k)^p\|_{G,2}^{1/p}\|(\epsilon_s)^p\|_{G,2}^{1/p} < \infty.
\end{aligned}$$

□

Lemma 5. *Suppose that assumption 3 holds and let $k,j,s \in [K]$ with $j \neq k$. Then, for G the law of ϵ and $1 \leq p \leq \min(1+\delta/4, 2)$, we have*

(i) $\|\phi_k(\epsilon_k)\epsilon_j\kappa(\epsilon_s)\|_{G,p} < \infty$,

(ii) $\|\epsilon_k\kappa(\epsilon_s)\|_{G,p} < \infty$,

(iii) $\|\kappa(\epsilon_k)\kappa(\epsilon_s)\|_{G,p} < \infty$.

Proof. By Cauchy-Schwarz, independence and our assumed moment conditions we have

$$\begin{aligned}
\|\phi_k(\epsilon_k)\epsilon_j\kappa(\epsilon_s)\|_{G,p} &\leq \left[\left[G[\phi_k(\epsilon_k)]^{2p}G[\epsilon_s]^{4p}\right]^{1/(2p)} + \|\phi_k(\epsilon_k)\|_{G,p}\right]\|\epsilon_j\|_{G,p} < \infty, \\
\|\epsilon_k\kappa(\epsilon_s)\|_{G,p} &\leq \|(\epsilon_k)^p\|_{G,2}^{1/p}\|(\epsilon_s)^{2p}\|_{G,2}^{1/p} + \|\epsilon_k\|_{G,p} < \infty, \\
\|\kappa(\epsilon_k)\kappa(\epsilon_s)\|_{G,p} &\leq \|(\epsilon_k)^{2p}\|_{G,2}^{1/p}\|(\epsilon_s)^{2p}\|_{G,2}^{1/p} + 2\|(\epsilon_k)^2\|_{G,p} + 2\|(\epsilon_s)^2\|_{G,p} + 1 < \infty.
\end{aligned}$$

□

Lemma 6. *Define*

$$\begin{aligned}
q_{l,i,n}^{(\alpha,\beta_1)} &:= \sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K}\zeta_{l,k,j,n}\phi_k(\epsilon_{k,i})\epsilon_{j,i} + \sum_{k=1}^{K}\zeta_{l,k,k,n}\left[\tau_{k,1}\epsilon_{k,i} + \tau_{k,2}\kappa(\epsilon_{k,i})\right] \\
q_{l,i,n}^{b} &:= -\sum_{k=1}^{K}[A_{n,k\bullet}D_{b,l}]\left[(X_i - \mathbb{E}X_i)\phi_k(\epsilon_{k,i}) - \mathbb{E}X_i(\varsigma_{k,1}\epsilon_{k,i} + \varsigma_{k,2}\kappa(\epsilon_{k,i}))\right]
\end{aligned}$$

where the dependence of e.g. $\zeta_{l,k,j,n}$ on n is as in the proof of proposition 2.[35] *Let $\breve{Q}_{l,m,i,n}^{r,s} := q_{l,i,n}^{r}q_{m,i,n}^{s}$. Suppose that assumption 5 holds. Then, for $1 \leq p \leq \min(1+\delta/4, 2)$ we have $\|\breve{Q}_{l,m,i,n}^{r,s}\|_{G,p} < \infty$ for G the law of $(\tilde{X}, \epsilon)$.*

[35] See footnote 33.

Proof. By definition we have

$$\begin{aligned}
\breve{Q}^{(\alpha,\beta_1),(\alpha,\beta_1)}_{l,m,i,n} &= \sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K}\sum_{s=1}^{K}\sum_{b=1,b\neq s}^{K}\zeta_{l,k,j,n}\zeta_{m,s,b,n}\phi_k(\epsilon_{k,i})\epsilon_{j,i}\phi_s(\epsilon_{s,i})\epsilon_{b,i} \\
&\quad + 2\sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K}\sum_{s=1}^{K}\zeta_{l,k,j,n}\zeta_{m,s,s,n}\phi_k(\epsilon_{k,i})\epsilon_{j,i}[\tau_{s,1}\epsilon_{s,i}+\tau_{s,2}\kappa(\epsilon_{s,i})] \\
&\quad + \sum_{k=1}^{K}\sum_{s=1}^{K}\zeta_{l,k,k,n}\zeta_{m,s,s,n}[\tau_{k,1}\epsilon_{k,i}+\tau_{k,2}\kappa(\epsilon_{k,i})][\tau_{s,1}\epsilon_{s,i}+\tau_{s,2}\kappa(\epsilon_{s,i})]. \\
\breve{Q}^{(\alpha,\beta_1),b}_{l,m,i,n} &= -\sum_{s=1}^{K}\sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K}\zeta_{l,k,j,n}\phi_k(\epsilon_{k,i})\epsilon_{j,i}[A_{n,s\bullet}D_{b,l}](X_i-\mathbb{E}X_i)\phi_s(\epsilon_{s,i}) \\
&\quad + \sum_{s=1}^{K}\sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K}\zeta_{l,k,j,n}\phi_k(\epsilon_{k,i})\epsilon_{j,i}[A_{n,s\bullet}D_{b,l}]\mathbb{E}X_i(\varsigma_{s,1}\epsilon_{s,i}+\varsigma_{s,2}\kappa(\epsilon_{s,i})) \\
&\quad - \sum_{s=1}^{K}\sum_{k=1}^{K}\zeta_{l,k,k,n}[\tau_{k,1}\epsilon_{k,i}+\tau_{k,2}\kappa(\epsilon_{k,i})][A_{n,s\bullet}D_{b,l}](X_i-\mathbb{E}X_i)\phi_s(\epsilon_{s,i}) \\
&\quad + \sum_{s=1}^{K}\sum_{k=1}^{K}\zeta_{l,k,k,n}[\tau_{k,1}\epsilon_{k,i}+\tau_{k,2}\kappa(\epsilon_{k,i})][A_{n,s\bullet}D_{b,l}]\mathbb{E}X_i(\varsigma_{s,1}\epsilon_{s,i}+\varsigma_{s,2}\kappa(\epsilon_{s,i})) \\
\breve{Q}^{b,b}_{l,m,i,n} &= \sum_{s=1}^{K}\sum_{k=1}^{K}[A_{n,s\bullet}D_{b,l}](X_i-\mathbb{E}X_i)\phi_s(\epsilon_{s,i})[A_{n,k\bullet}D_{b,l}](X_i-\mathbb{E}X_i)\phi_k(\epsilon_{k,i}) \\
&\quad + 2\sum_{s=1}^{K}\sum_{k=1}^{K}[A_{n,s\bullet}D_{b,l}]\mathbb{E}X_i(\varsigma_{s,1}\epsilon_{s,i}+\varsigma_{s,2}\kappa(\epsilon_{s,i}))[A_{n,k\bullet}D_{b,l}](X_i-\mathbb{E}X_i)\phi_k(\epsilon_{k,i}) \\
&\quad + \sum_{s=1}^{K}\sum_{k=1}^{K}[A_{n,s\bullet}D_{b,l}]\mathbb{E}X_i(\varsigma_{s,1}\epsilon_{s,i}+\varsigma_{s,2}\kappa(\epsilon_{s,i}))[A_{n,k\bullet}D_{b,l}]\mathbb{E}X_i(\varsigma_{k,1}\epsilon_{k,i}+\varsigma_{k,2}\kappa(\epsilon_{k,i}))
\end{aligned}$$

Hence, by Minkowski's inequality, the independence of ϵ from $\tilde{X}$ (with finite second moments) and lemmas 4 & 5, $\|\breve{Q}^{r,s}_{l,m,i,n}\|_{G,p} < \infty$. □

Lemma 7. *Suppose assumption 5 holds and θ_n, $\nu_{n,p}$ and ν_n are as in assumption 4. Then $\|\hat{\varkappa}_{k,n} - \varkappa_{k,n}\|_2 = o_{P_{\theta_n}}(\nu_{n,p}) = o_{P_{\theta_n}}(\nu_n^{1/2})$ for $\varkappa \in \{\tau, \varsigma\}$.*

Proof. Under P_{θ_n}, $A_{n,k\bullet}(Z_i - B_n X_i) \simeq \epsilon_{k,i} \sim \eta_k$, hence the claim will follow if we show that $\breve{\varkappa}_{k,n} - \breve{\varkappa}_k = o_{G_k}(\nu_n^{1/2})$, where

$$\breve{\varkappa}_{k,n} \coloneqq \breve{M}_{k,n}^{-1} w, \quad \text{where } \breve{M}_{k,n} \coloneqq \begin{pmatrix} 1 & \frac{1}{n}\sum_{i=1}^{n}(\epsilon_{k,i})^3 \\ \frac{1}{n}\sum_{i=1}^{n}(\epsilon_{k,i})^3 & \frac{1}{n}\sum_{i=1}^{n}(\epsilon_{k,i})^4 - 1 \end{pmatrix},$$

$$\breve{\varkappa}_{k,n} \coloneqq \breve{M}_{k,n}^{-1} w, \quad \text{where } \breve{M}_{k,n} \coloneqq \begin{pmatrix} 1 & G_k(\epsilon_{k,i})^3 \\ G_k(\epsilon_{k,i})^3 & G_k(\epsilon_{k,i})^4 - 1 \end{pmatrix},$$

and $w \in \mathbb{R}^2$. By the preceding definitions and the fact that the map $M \mapsto M^{-1}$ is Lipschitz at a positive definite matrix M_0 we have that for a positive constant C then for large enough n, with probability approaching one

$$\|\breve{\varkappa}_{k,n} - \breve{\varkappa}_{k,n}\|_2 = \|(\breve{M}_{k,n}^{-1} - \breve{M}_k^{-1})w\|_2 \leq \|w\|_2 \|\breve{M}_{k,n}^{-1} - \breve{M}_k^{-1}\|_2 \lesssim C\|\breve{M}_{k,n} - \breve{M}_k\|_2. \tag{38}$$

If $\upsilon \coloneqq \delta/4 \geq 1$, we have that by Theorem 2.5.11 in Durrett (2019)

$$\frac{1}{n}\sum_{i=1}^{n}[(\epsilon_{k,i})^3 - G_k(\epsilon_{k,i})^3] = o_{G_k}\left(n^{-1/2}\log(n)^{1/2+\iota}\right)$$
$$\frac{1}{n}\sum_{i=1}^{n}[(\epsilon_{k,i})^4 - G_k(\epsilon_{k,i})^4] = o_{G_k}\left(n^{-1/2}\log(n)^{1/2+\iota}\right)$$

for $\iota > 0$, which implies that

$$\|\breve{M}_{k,n} - \breve{M}_k\|_2 \leq \|\breve{M}_{k,n} - \breve{M}_k\|_F = o_{G_k}\left(n^{-1/2}\log(n)^{1/2+\iota}\right).$$

If $0 < \upsilon < 1$, we have by Theorems 2.5.11 & 2.5.12 in Durrett (2019) that for $\iota > 0$,

$$\frac{1}{n}\sum_{i=1}^{n}[(\epsilon_{k,i})^3 - G_k(\epsilon_{k,i})^3] = \begin{cases} o_{G_k}\left(n^{-1/2}\log(n)^{1/2+\iota}\right) & \text{if } \upsilon \in [1/2, 1) \\ o_{G_k}\left(n^{\frac{1-p}{p}}\right) & \text{if } \upsilon \in (0, 1/2) \end{cases},$$
$$\frac{1}{n}\sum_{i=1}^{n}[(\epsilon_{k,i})^4 - G_k(\epsilon_{k,i})^4] = o_{G_k}\left(n^{\frac{1-p}{p}}\right).$$

which together imply that

$$\|\breve{M}_{k,n} - \breve{M}_k\|_2 \leq \|\breve{M}_{k,n} - \breve{M}_k\|_F = o_{G_k}\left(n^{\frac{1-p}{p}}\right).$$

Combining these convergence rates with equation (38) yields the result in light of the observations made at the beginning of the proof. □

Lemma 8. *Suppose assumptions 5 and 6 hold and $\theta_n = (\alpha_0, \beta_n, \eta)$ where $\sqrt{n}(\beta_n - \beta) = O(1)$ is a deterministic sequence. Then for each $r \in \{(\alpha, \beta_1), b\}$ and l*

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\ell}_{\theta_n,r,l}(Y_i) - \tilde{\ell}_{\theta_n,r,l}(Y_i)\right)^2 = o_{P_{\theta_n}}(\nu_n).$$

Proof. In this proof we let $M_k \coloneqq M_{k\bullet}$ for any matrix M. We start by considering elements in $\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\ell}_{\theta_n,(\alpha,\beta_1),l}(Y_i) - \tilde{\ell}_{\theta_n,(\alpha,\beta_1),l}(Y_i)\right)^2$. We define $\tilde{\tau}_{k,n,q} \coloneqq \hat{\tau}_{k,n,q} - \tau_{k,q}$ and $V_{i,n} = Z_i - B_n X_i$. Since each $|\zeta_{l,k,j,n}| < \infty$ and the sums over k, j are finite, it is sufficient to demonstrate that

for every $k, j, m, s \in [K]$, with $k \neq j$ and $s \neq m$,

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(A_{n,k}V_{i,n}) - \phi_k(A_{n,k}V_{i,n})\right]\left[\hat{\phi}_{s,n}(A_{n,s}V_{i,n}) - \phi_s(A_{n,s}V_{i,n})\right]A_{n,j}V_{i,n}A_{n,m}V_{i,n} = o_{P_{\theta_n}}(\nu_n), \tag{39}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(A_{n,k}V_{i,n}) - \phi_k(A_{n,k}V_{i,n})\right]A_{n,j}V_{i,n}\left[\tilde{\tau}_{s,n,1}A_{n,s}V_{i,n} + \tilde{\tau}_{s,n,2}\kappa(A_{n,s}V_{i,n})\right] = o_{P_{\theta_n}}(\nu_n), \tag{40}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left[\tilde{\tau}_{s,n,1}A_{n,s}V_{i,n} + \tilde{\tau}_{s,n,2}\kappa(A_{n,s}V_{i,n})\right]\left[\tilde{\tau}_{k,n,1}A_{n,k}V_{i,n} + \tilde{\tau}_{k,n,2}\kappa(A_{n,k}V_{i,n})\right] = o_{P_{\theta_n}}(\nu_n). \tag{41}$$

For (41), let $\xi_1(x) = x$ and $\xi_2(x) = \kappa(x)$. Then, we can split the sum into 4 parts, each of which has the following form for some $q, w \in \{1, 2\}$

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{\tau}_{s,n,q}\tilde{\tau}_{k,n,w}\xi_q(A_{n,s}V_{i,n})\xi_w(A_{n,k}V_{i,n}) = \tilde{\tau}_{s,n,q}\tilde{\tau}_{k,n,w}\frac{1}{n}\sum_{i=1}^{n}\xi_q(A_{n,s}V_{i,n})\xi_w(A_{n,k}V_{i,n}) = o_{P_{\theta_n}}(\nu_n),$$

since we have that each $\tilde{\tau}_{s,n,q}\tilde{\tau}_{k,n,w} = o_{P_{\theta_n}}(\nu_n)$ by lemma 7.[36] For (40) we can argue similarly. Again let $\xi_1(x) = x$ and $\xi_2(x) = \kappa(x)$. Then, we can split the sum into 2 parts, each of which has the following form for some $q \in \{1, 2\}$

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(A_{n,k}V_{i,n}) - \phi_k(A_{n,k}V_{i,n})\right]A_{n,j}V_{i,n}\tilde{\tau}_{s,n,q}\xi_q(A_{n,s}V_{i,n}) \\
&\quad\leq \tilde{\tau}_{s,n,q}\left(\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(A_{n,k}V_{i,n}) - \phi_k(A_{n,k}V_{i,n})\right]^2(A_{n,j}V_{i,n})^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_q(A_{n,s}V_{i,n})^2\right)^{1/2} \\
&\quad= o_{P_{\theta_n}}(\nu_n).
\end{aligned}$$

by assumption 6 applied with $W_{i,n} = A_{n,j}V_{i,n}$ and $\tilde{\tau}_{s,n,q} = o_{P_{\theta_n}}(\nu_n^{1/2})$.[37] For (39) use Cauchy-

[36] The fact that $\frac{1}{n}\sum_{i=1}^{n}\xi_q(A_{n,s}V_{i,n})\xi_w(A_{n,k}V_{i,n}) = O_{P_{\theta_n}}(1)$ can be seem to hold using the moment and i.i.d. assumptions from assumption 3 and Markov's inequality, noting once more that $A_{n,k}V_{i,n} \simeq \epsilon_{k,i}$ under P_{θ_n}.

[37] See footnote 36.

Schwarz with assumption 6:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(A_{n,k}V_{i,n})-\phi_k(A_{n,k}V_{i,n})\right]\left[\hat{\phi}_{s,n}(A_{n,s}V_{i,n})-\phi_s(A_{n,s}V_{i,n})\right]A_{n,j}V_{i,n}A_{n,m}V_{i,n}\\
&\quad\leq\left(\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(A_{n,k}V_{i,n})-\phi_k(A_{n,k}V_{i,n})\right]^2(A_{n,j}V_{i,n})^2\right)^{1/2}\\
&\qquad\qquad\times\left(\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{s,n}(A_{n,s}V_{i,n})-\phi_s(A_{n,s}V_{i,n})\right]^2(A_{n,m}V_{i,n})^2\right)^{1/2}\\
&\quad=o_{P_{\theta_n}}(\nu_n).
\end{aligned}$$

Finally, we consider the elements in $\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\ell}_{\theta_n,b,l}(Y_i)-\tilde{\ell}_{\theta_n,b,l}(Y_i)\right)^2$, where we let $a_{n,k,l}\coloneqq -A_{n,k}D_{b,l}$ and note that

$$\begin{aligned}
&\hat{\ell}_{\theta_n,b,l}(Y_i)-\tilde{\ell}_{\theta_n,b,l}(Y_i)\\
&\quad=\sum_{k=1}^{K}a_{n,k,l}\left[(X_i-\mathbb{E}X_i)[\hat{\phi}_k(V_{i,k,n})-\phi_k(V_{i,k,n})]+(\mathbb{E}X_i-\bar{X}_n)\phi_k(V_{i,k,n})\right]\\
&\qquad+\sum_{k=1}^{K}a_{n,k,l}\left[(\mathbb{E}X_i-\bar{X}_n)[\hat{\varsigma}_{k,n,1}V_{i,k,n}+\hat{\varsigma}_{k,n,2}\kappa(V_{i,k,n})]\right]\\
&\qquad-\sum_{k=1}^{K}a_{n,k,l}\left[\mathbb{E}X_i[(\hat{\varsigma}_{k,n,1}-\varsigma_{k,1})V_{i,k,n}+(\hat{\varsigma}_{k,n,2}-\varsigma_{k,2})\kappa(V_{i,k,n})]\right]
\end{aligned}$$

We have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\ell}_{\theta_n,b,l}(Y_i)-\tilde{\ell}_{\theta_n,b,l}(Y_i)\right)^2\\
&\quad\lesssim\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n}[a_{n,k,l}(X_i-\mathbb{E}X_i)]^2[\hat{\phi}_k(V_{i,k,n})-\phi_k(V_{i,k,n})]^2+[a_{n,k,l}(\mathbb{E}X_i-\bar{X}_n)]^2\phi_k(V_{i,k,n})^2\\
&\qquad+\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n}[a_{n,k,l}(\mathbb{E}X_i-\bar{X}_n)]^2[\hat{\varsigma}_{k,n,1}V_{i,k,n}+\hat{\varsigma}_{k,n,2}\kappa(V_{i,k,n})]^2\\
&\qquad+\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n}[a_{n,k,l}\mathbb{E}X_i]^2[(\hat{\varsigma}_{k,n,1}-\varsigma_{k,1})V_{i,k,n}+(\hat{\varsigma}_{k,n,2}-\varsigma_{k,2})\kappa(V_{i,k,n})]^2
\end{aligned}$$

The first term is $o_{P_{\theta_n}}(\nu_n)$ by Cauchy-Schwarz and applying assumption 6, the second and third terms follows from $(a_{n,k,l}(\bar{X}_n-\mathbb{E}X_i))^2=O_{P_{\theta_n}}(n^{-1})=o_{P_{\theta_n}}(\nu_n)$ and the fourth term follows from Lemma 7. □

Appendix B: Log density score estimation

In this section we discuss the details for the estimation of the log density scores ϕ_k. In particular, following Chen and Bickel (2006) and Jin (1992) we define a convenient B-spline based estimator $\hat{\phi}_{n,k}$ and show that this estimate satisfies Assumption 6.

B-spline based density score estimation

Let $\xi_1 < \cdots < \xi_N$ be a knot sequence, the first order B-splines are defined according to $b_i^{(1)}(x) \coloneqq \mathbf{1}_{[\xi_i, \xi_{i+1})}(x)$. Subsequent order B-splines can be computed according to the recurrence relation

$$b_i^{(\kappa)}(x) = \frac{x - \xi_i}{\xi_{i+\kappa-1} - \xi_i} b_i^{(\kappa-1)}(x) + \frac{\xi_{i+\kappa} - x}{\xi_{i+\kappa} - \xi_{i+1}} b_{i+1}^{(\kappa-1)}(x), \tag{42}$$

for $\kappa > 1$ and $i = 1, \ldots, N - \kappa$. A κ-th order B-spline is $\kappa - 2$ times differentiable in x with first derivative

$$c_i^{(\kappa)}(x) = \frac{\kappa - 1}{\xi_{i+\kappa-1} - \xi_i} b_i^{(\kappa-1)}(x) - \frac{\kappa - 1}{\xi_{i+\kappa} - \xi_{i+1}} b_{i+1}^{(\kappa-1)}(x). \tag{43}$$

See de Boor (2001) for more details on B-splines.

Let $b_{k,n} = (b_{k,n,1}, \ldots, b_{k,n,B_{k,n}})'$ be a collection of $B_{k,n}$ cubic B-splines and let $c_{k,n} = (c_{k,n,1}, \ldots, c_{k,n,B_{k,n}})'$ be their derivatives: $c_{k,n,i}(x) \coloneqq \frac{\mathrm{d} b_{k,n,i}(x)}{\mathrm{d}x}$ for each $i \in [B_{k,n}]$. Let $\gamma_k \in \mathbb{R}^{B_{k,n}}$. The knots of the splines, $\xi_{k,n} = (\xi_{k,n,i})_{i=1}^{K_{k,n}}$ are equally spaced in $[\Xi_{k,n}^L, \Xi_{k,n}^U]$ with $\delta_{k,n} \coloneqq \xi_{k,n,i+1} - \xi_{k,n,i} > 0$.[38] For each (k, n) pair the relationships between the number of knots ($K_{k,n}$), the number of spline functions ($B_{k,n}$) and $\delta_{k,n}$ are given by $B_{k,n} = K_{k,n} - 4$ and $K_{k,n} = 1 + (\Xi_{k,n}^U - \Xi_{k,n}^L)/\delta_{k,n}$.[39]

Since the B-splines vanish at infinity for any $n \in \mathbb{N}$, integration by parts gives that

$$\begin{aligned} \int (\phi_k(z) - \gamma_k' b_{k,n}(z))^2 \eta_k(z) \, \mathrm{d}z &= \int \phi_k^2 \, \mathrm{d}G_k + \int (\gamma_k' b_{k,n})^2 \, \mathrm{d}G_k + 2 \int \gamma_k' c_{k,n}(z) \eta_k(z) \, \mathrm{d}z \\ &= G_k \phi_k^2 + \gamma_k' G_k [b_{k,n} b_{k,n}'] \gamma_k + 2 \gamma_k' G_k c_{k,n}. \end{aligned} \tag{44}$$

The solution to minimising this mean-squared error is given by:[40]

$$\gamma_{k,n} = -G_k [b_{k,n} b_{k,n}']^{-1} G_k c_{k,n}. \tag{45}$$

Replacing the population expectations with sample counterparts we arrive at our estimate of γ_k

$$\hat{\gamma}_{k,n} \coloneqq -\left[\frac{1}{n} \sum_{i=1}^n b_{k,n}(\epsilon_{k,i}) b_{k,n}(\epsilon_{k,i})' \right]^{-1} \frac{1}{n} \sum_{i=1}^n c_{k,n}(\epsilon_{k,i}), \tag{46}$$

where $\epsilon_{k,i}$ is set equal to $A_{n,k\bullet}(Y_i - B_n X_i)$, which under P_{θ_n} has the same distribution. Our

[38] For each $k \in [K]$ the sequences $(\Xi_{k,n}^L)_{n\in\mathbb{N}}$, $(\Xi_{k,n}^U)_{n\in\mathbb{N}}$, $(B_{k,n})_{n\in\mathbb{N}}$ and $(\delta_{k,n})_{n\in\mathbb{N}}$ are deterministic.

[39] Implicitly we choose $K_{k,n}$ and the endpoints and $\delta_{k,n}$ adjusts such that these formulae hold; this way we do not need to adjust anything to ensure these are integers.

[40] This differs from the expression in Chen and Bickel (2006) by a factor of -1 as they estimate $-\phi_k$.

estimate for ϕ_k is given by

$$\hat{\phi}_{k,n}(z) \coloneqq \hat{\gamma}'_{k,n} b_{k,n}(z) \ . \tag{47}$$

We note that computing (47) effective only requires computing the B-spline regression coefficients $\hat{\gamma}_{k,n}$ in (46). To implement the score test we need to estimate K density scores, hence the computational costs is quite modest.

Theoretical properties density score estimator

We will now show that the estimates $\hat{\phi}_{k,n}$ satisfy assumption 6 under regularity conditions on η_k and the choice of knot points. We first state the main proposition.

Proposition 3. *Let $\phi_{k,n} \coloneqq \phi_k \mathbf{1}_{[\Xi^L_{k,n}, \Xi^U_{k,n}]}$ and $\Delta_{k,n} \coloneqq \Xi^U_{k,n} - \Xi^L_{k,n}$ and suppose that for ν_n as in assumption 6, $[\Xi^L_{k,n}, \Xi^U_{k,n}] \uparrow \tilde{\Xi} \supset \operatorname{supp}(\eta_k)$ and $\delta_{k,n} \downarrow 0$ such that*

(i) $G_k(\epsilon_k \notin [\Xi^L_{k,n}, \Xi^U_{k,n}]) = o(\nu_n^2)$;

(ii) *For some $\iota > 0$,* $n^{-1}\Delta_{k,n}^{2+2\iota}\delta_{k,n}^{-(8+2\iota)} = o(\nu_n)$;

(iii) η_k *is bounded ($\|\eta_k\|_\infty < \infty$) and differentiable, with a bounded derivative:* $\|\eta'_k\|_\infty < \infty$;

(iv) *For each n, $\phi_{k,n}$ is three-times continuously differentiable on $[\Xi^L_{k,n}, \Xi^U_{k,n}]$ and $\|\phi^{(3)}_{k,n}\|^2_\infty \delta^6_{k,n} = o(\nu_n)$;*[41]

(v) *There are $c > 0$ and $N \in \mathbb{N}$ such that for $n \geq N$ we have $\inf_{t \in [\Xi^L_{k,n}, \Xi^U_{k,n}]} |\eta_k(t)| \geq c\delta_{k,n}$.*

Then, under assumption 5, the estimates $\hat{\phi}_{k,n}$ satisfy assumption 6.

Proof. We start by showing that $\hat{\phi}_{k,n}$ satisfies equation (19) when we replace Y_i by $(Z_i - B_n X_i)$ and take $\beta_n = (\beta_{1,n}, b_n)$. Under P_{θ_n}, we have that $A_{n,k\bullet}(Z_i - B_n X_i) \simeq \epsilon_{k,i} \sim \eta_k$. Additionally, we can write

$$\begin{aligned}
\left| \frac{1}{n} \sum_{i=1}^n \hat{\phi}_{k,n}(\epsilon_{k,i}) W_{i,n} - \phi_k(\epsilon_{k,i}) W_{i,n} \right| \leq & \left| \frac{1}{n} \sum_{i=1}^n \left[\hat{\phi}_{k,n}(\epsilon_{k,i}) - \tilde{\phi}_{k,n}(\epsilon_{k,i}) \right] W_{i,n} \right| \\
& + \left| \frac{1}{n} \sum_{i=1}^n \left[\tilde{\phi}_{k,n}(\epsilon_{k,i}) - \phi_{k,n}(\epsilon_{k,i}) \right] W_{i,n} \right| \\
& + \left| \frac{1}{n} \sum_{i=1}^n \left[\phi_{k,n}(\epsilon_{k,i}) - \phi_k(\epsilon_{k,i}) \right] W_{i,n} \right|,
\end{aligned} \tag{48}$$

where $\hat{\phi}_{k,n}(z) = \hat{\gamma}'_{k,n} b_{k,n}(z)$, $\tilde{\phi}_{k,n}(z) \coloneqq \gamma'_{k,n} b_{k,n}(z)$ and $\phi_{k,n} \coloneqq \phi_k \mathbf{1}_{[\Xi^L_{k,n}, \Xi^U_{k,n}]}$. We will show that each of these three terms on the right hand side are $o_G(n^{-1/2})$, where G is the product of G_k and G_w, which implies that

$$\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\phi}_{k,n}(A_{n,k} Y_i) W_{i,n} - \phi_k(A_{n,k} Y_i) W_{i,n} \right| \xrightarrow{P_{\theta_n}} 0.$$

[41] The differentiability and continuity requirements at the end-points are one-sided.

For the last term in (48), by assumption $G_k\{\epsilon_k \notin [\Xi^L_{k,n}, \Xi^U_{k,n}]\} \downarrow 0$ and hence by independence and Cauchy-Schwarz

$$\begin{aligned} G\left([\phi_{k,n}(\epsilon_k) - \phi_k(\epsilon_k)]^2 W_{i,n}^2\right) &= G_k\left[\phi_k(\epsilon_k)^2 \mathbf{1}\{\epsilon_k \notin [\Xi^L_{k,n}, \Xi^U_{k,n}]\}\right] G_w W_{i,n}^2 \\ &\leq \left[G_k \phi_k(\epsilon_k)^4\right]^{1/2} \left[G_k \mathbf{1}\{\epsilon_k \notin [\Xi^L_{k,n}, \Xi^U_{k,n}]\}\right]^{1/2} G_w W_{i,n}^2 \\ &\to 0. \end{aligned} \tag{49}$$

By Markov's inequality it follows that for any $\upsilon > 0$,

$$G\left(\left|\frac{1}{\sqrt{n}} \sum_{i=1}^n [\phi_{k,n}(\epsilon_{k,i}) - \phi_k(\epsilon_{k,i})] W_{i,n}\right| > \upsilon\right) \leq \frac{nG\left([\phi_{k,n}(\epsilon_k) - \phi_k(\epsilon_k)]^2 W_{i,n}^2\right)}{n\upsilon} \to 0.$$

For the second term, we note that by our hypotheses and lemma 11 we have

$$G\left([\tilde{\phi}_{k,n}(\epsilon_k) - \phi_{k,n}(\epsilon_k)]^2 W_{i,n}^2\right) = G_k\left([\tilde{\phi}_{k,n}(\epsilon_k) - \phi_{k,n}(\epsilon_k)]^2\right) G_w W_{i,n}^2 \leq C^2 \delta_{k,n}^6 \|\phi_k^{(3)}\|_\infty^2 G_w W_{i,n}^2 \to 0, \tag{50}$$

as $n \to \infty$, and hence again by Markov's inequality for any $\upsilon > 0$,

$$G\left(\left|\frac{1}{\sqrt{n}} \sum_{i=1}^n [\tilde{\phi}_{k,n}(\epsilon_{k,i}) - \phi_{k,n}(\epsilon_{k,i})] W_{i,n}\right| > \upsilon\right) \leq \frac{nG\left([\tilde{\phi}_{k,n}(\epsilon_k) - \phi_{k,n}(\epsilon_k)]^2 W_{i,n}^2\right)}{n\upsilon} \to 0.$$

For the first term, by Cauchy-Schwarz

$$\left|\frac{1}{n} \sum_{i=1}^n \left[\hat{\phi}_{k,n}(\epsilon_{k,i}) - \tilde{\phi}_{k,n}(\epsilon_{k,i})\right] W_{i,n}\right| \leq \|\hat{\gamma}_{k,n} - \gamma_{k,n}\|_2 \left\|\frac{1}{n} \sum_{i=1}^n b_{k,n}(\epsilon_{k,i}) W_{i,n}\right\|_2 = o_G(n^{-1/2}),$$

by lemmas 12 and 13.

Next, we show that $\hat{\phi}_{k,n}$ satisfies equation (20) when we replace Y_i by $(Z_i - B_n X_i)$ and take $\beta_n = (\beta_{1,n}, b_n)$. We write:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \left(\left[\hat{\phi}_{k,n}(\epsilon_{k,i}) - \phi_k(\epsilon_{k,i})\right] W_{i,n}\right)^2 \leq{}& \frac{4}{n} \sum_{i=1}^n \left[\hat{\phi}_{k,n}(\epsilon_{k,i}) - \tilde{\phi}_{k,n}(\epsilon_{k,i})\right]^2 W_{i,n}^2 \\ &+ \frac{4}{n} \sum_{i=1}^n \left[\tilde{\phi}_{k,n}(\epsilon_{k,i}) - \phi_{k,n}(\epsilon_{k,i})\right]^2 W_{i,n}^2 \\ &+ \frac{4}{n} \sum_{i=1}^n \left[\phi_{k,n}(\epsilon_{k,i}) - \phi_k(\epsilon_{k,i})\right]^2 W_{i,n}^2. \end{aligned} \tag{51}$$

We will show that (1/4 of) each of the right hand side terms is $o_G(\nu_n)$ under our assumptions, which is sufficient for equation (20) since $A_{k,n}(Z_i - B_n X_i) \simeq \epsilon_{k,i} \sim \eta_k$ under P_{θ_n}. For the last

term, for any $\upsilon > 0$, by Markov's inequality, independence and Cauchy-Schwarz we have

$$G\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left[\phi_{k,n}(\epsilon_{k,i})-\phi_k(\epsilon_{k,i})\right]^2 W_{i,n}^2\right| > \upsilon\nu_n\right) \lesssim \frac{G_k\mathbf{1}\{\epsilon_k \notin [\Xi_{k,n}^L, \Xi_{k,n}^U]\}G_w W_{i,n}^2}{\upsilon\nu_n} = o(1).$$

For the second term, for any $\upsilon > 0$, by Markov's inequality, independence and lemma 11:

$$\begin{aligned} G\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left[\tilde{\phi}_{k,n}(\epsilon_{k,i})-\phi_{k,n}(\epsilon_{k,i})\right]^2 W_{i,n}^2\right| > \upsilon\nu_n\right) &\leq \frac{G_k\left(\left[\tilde{\phi}_{k,n}(\epsilon_k)-\phi_{k,n}(\epsilon_k)\right]^2\right)G_w W_{i,n}^2}{\upsilon\nu_n} \\ &\leq \frac{C\delta_{k,n}^6\|\phi_k^{(3)}\|_\infty^2 G_w W_{i,n}^2}{\upsilon\nu_n} \\ &= o(1). \end{aligned}$$

Finally, for the first term in the decomposition, by lemma 13 and condition (ii) we have

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(\epsilon_{k,i})-\tilde{\phi}_{k,n}(\epsilon_{k,i})\right]^2 W_{i,n}^2 \leq \|\hat{\gamma}_{k,n}-\gamma_{k,n}\|_2^2\left[\frac{1}{n}\sum_{i=1}^{n}\|b_{k,n}(\epsilon_{k,i})\|_2^2 W_{i,n}^2\right] = o_G(\nu_n).$$

□

Auxiliary results

Lemma 9. *The smallest eigenvalue of the $B_{k,n}\times B_{k,n}$ Gram matrix $\tilde{\Gamma}_{k,n} \coloneqq \int b_{k,n}b_{k,n}'\,\mathrm{d}\lambda$ satisfies*

$$\lambda_{\min}(\tilde{\Gamma}_{k,n}) \geq \upsilon\delta_{k,n} > 0,$$

for a $\upsilon > 0$.

Proof. Since $b_{k,n,m}(x)b_{k,n,s}(x)$ is non-zero only for $|m-s| \leq 3$ and each $b_{k,n,m}$ is non-zero only on $[\xi_{k,n,m}, \xi_{k,n,m+4})]$ (e.g. (20) p. 91 of de Boor, 2001), $\tilde{\Gamma}_{k,n}$ is a symmetric banded Toeplitz matrix.[42] Its entries can be computed by direct integration:

$$[\tilde{\Gamma}_{k,n}]_{m,s} = \delta_{k,n}\times\begin{cases}\frac{151}{315} & \text{if } m=s\\ \frac{397}{1680} & \text{if } |m-s|=1\\ \frac{1}{42} & \text{if } |m-s|=2\\ \frac{1}{5040} & \text{if } |m-s|=3\\ 0 & \text{if } |m-s|>3\end{cases}.$$

For simplicity of notation let $f_0 \coloneqq \frac{151}{315}$, $f_1 \coloneqq f_{-1} \coloneqq \frac{397}{1680}$, $f_2 \coloneqq f_{-2} \coloneqq \frac{1}{42}$ and $f_3 \coloneqq f_{-3} \coloneqq \frac{1}{5040}$ and let $f_s \coloneqq 0$ for $|s| > 3$. Now, let $f(\theta) \coloneqq \sum_{s=-3}^{3} f_s e^{i(s\theta)}$. Then, $\tilde{\Gamma}_{k,n}/\delta_{k,n}$ is then the matrix

[42] As can be easily verified, unlike in the case of linear ($\kappa = 2$) or quadratic splines ($\kappa = 3$), this matrix is *not* diagonally dominant. In the case of $\kappa \in \{2, 3\}$ this argument could be completed in a simpler fashion by using the Gershgorin circle theorem.

generated by f in the sense that $\tilde{\Gamma}_{k,n}/\delta_{k,n} = \mathscr{T}_n(f) \coloneqq \sum_{s=-\min(B_{k,n}-1,3)}^{\min(B_{k,n}-1,3)} f_k J_n^s$ where each J_n^s is the $B_{k,n} \times B_{k,n}$ matrix which is zero everywhere except for the (i,j)-th entries where $i-j=s$, where it has a value of 1.[43] Since $f \in L_1([-\pi,\pi])$ and is real on $[-\pi,\pi]$ by Theorem 6.1 in Garoni and Serra-Capizzano (2017) we have that $\lambda_{\min}(\tilde{\Gamma}_{k,n}) = \delta_{k,n}\lambda_{\min}(\tilde{\Gamma}_{k,n}/\delta_{k,n}) \geq \delta_{k,n} \inf_{\theta\in[-\pi,\pi]} f(\theta) = \delta_{k,n}\upsilon$, where $\upsilon \coloneqq \inf_{\theta\in[-\pi,\pi]} f(\theta) > 0$. □

Lemma 10. *Suppose $\xi \in \mathbb{R}^{N+1}$ such that $a = \xi_0 < \xi_1 < \cdots < \xi_N = b$, $h \coloneqq \max_{i\in[N]} \xi_i - \xi_{i-1}$, and let $\mathscr{G}_k(\xi)$ be the linear space formed by degree k splines with knots ξ. Then, if $f \in C^{k-1}[a,b]$ we have that*

$$\inf_{g\in\mathscr{G}_k(\xi)} \|g-f\|_\infty \leq \frac{(k+1)!}{2^k} h^{k-1} \|f^{(k-1)}\|_\infty = c_k h^{k-1} \|f^{(k-1)}\|_\infty,$$

where c_k depends only on k.

Proof. This follows as a special case of Theorem 20.3 in Powell (1981). □

Lemma 11 (Cf. Lemma A.5, Chen and Bickel, 2006)**.** *Let $\tilde{\phi}_{k,n}(z)$ and $\phi_{k,n}$ be defined as in Proposition 3. If (iv) of the hypotheses of proposition 3 holds, we have*

$$G_k\left(\tilde{\phi}_{k,n}(\epsilon_k) - \phi_{k,n}(\epsilon_k)\right)^2 \leq C^2 \delta_{k,n}^6 \|\phi_{k,n}^{(3)}\|_\infty^2.$$

Proof. By the definition of $\tilde{\phi}_{k,n}$ and lemma 10 we have

$$G_k\left(\tilde{\phi}_{k,n}(\epsilon_k) - \phi_{k,n}(\epsilon_k)\right)^2 = \inf_{g\in\mathscr{G}_k(\xi_{k,n})} G_k\left(g(\epsilon_k) - \phi_{k,n}(\epsilon_k)\right)^2 \leq C^2\delta_{k,n}^6 \|\phi_{k,n}^{(3)}\|_\infty^2.$$

The first inequality comes from the fact that we can equivalently see $\gamma_{k,n} = -G_k[b_{k,n}b_{k,n}']^{-1}G_k c_{k,n}$ as the solution to a version of the mean-squared error problem based on equation (44) where we only integrate over the support of $\phi_{k,n}$ since this is also the support of $b_{k,n}$ and $c_{k,n}$. □

Lemma 12 (Cf. Lemma A.3, Chen and Bickel, 2006)**.** *Under assumption 5 and that $W_{i,n}$ is independent of $\epsilon_{k,i}$ we have*

$$\left\| \frac{1}{n}\sum_{i=1}^n b_{k,n}(\epsilon_{k,i}) W_{i,n} \right\|_2 = O_G(n^{-1/2}).$$

Proof. By the fact that $\sum_{m=1}^{B_{k,n}} b_{m,k,n}(x)^2 \leq 1$ (see e.g. (36) on p. 96 of de Boor, 2001) and the given assumptions we have that

$$G\left(\left\| \frac{1}{n}\sum_{i=1}^n b_{k,n}(\epsilon_{k,i}) W_{i,n} \right\|_2^2\right) = \frac{1}{n} G_k\left(\sum_{m=1}^{B_{k,n}} b_{k,n,m}(\epsilon_k)^2\right) G_w W_{i,n}^2 \leq \frac{G_w W_{i,n}^2}{n}$$

[43] See section 6.1 in Garoni and Serra-Capizzano (2017), noting that it is clear that $f \in L_1([-\pi,\pi])$.

Fix $\epsilon > 0$ and take $M > 0$ large enough such that $G_w W_{i,n}^2 / M^2 < \epsilon$. Markov's inequality yields

$$G\left(\sqrt{n}\left\|\frac{1}{n}\sum_{i=1}^{n} b_{k,n}(\epsilon_{k,i})W_{i,n}\right\|_2 > M\right) \leq \frac{G\left(n\left\|\frac{1}{n}\sum_{i=1}^{n} b_{k,n}(\epsilon_{k,i})W_{i,n}\right\|_2^2\right)}{M^2} \leq \frac{G_w W_{i,n}^2}{M^2} < \epsilon.$$

□

Lemma 13 (Cf. Lemma A.2, Chen and Bickel, 2006)**.** *Let $\hat{\gamma}_{k,n}$ and $\gamma_{k,n}$ be defined as in equations (46) and (45) respectively. Suppose that conditions (ii), (iii) and (v) of proposition 3 and assumption 5 hold. Then, if we define*

$$\hat{\Gamma}_{k,n} \coloneqq \frac{1}{n}\sum_{i=1}^{n} b_{k,n}(\epsilon_{k,i}) b_{k,n}(\epsilon_{k,i})', \quad \Gamma_{k,n} \coloneqq G_k b_{k,n} b_{k,n}',$$

and

$$\hat{C}_{k,n} \coloneqq \frac{1}{n}\sum_{i=1}^{n} c_{k,n}(\epsilon_{k,i}), \quad C_{k,n} \coloneqq G_k c_{k,n},$$

we have that

(i) $\|C_{k,n}\|_2 = O(\delta_{k,n} B_{k,n}^{1/2})$,

(ii) $\|\hat{C}_{k,n} - C_{k,n}\|_2 = O_G\left(\sqrt{\frac{B_{k,n}\log B_{k,n}}{n\delta_{k,n}^2}}\right)$,

(iii) $\|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_2 = O_G\left(\sqrt{\frac{B_{k,n}\log B_{k,n}}{n}}\right)$,

(iv) $\|\Gamma_{k,n}\|_2 = O(\delta_{n,k})$

(v) $\|\Gamma_{k,n}^{-1}\|_2 = O(\delta_{k,n}^{-2})$.

In particular, $\|\hat{\gamma}_{k,n} - \gamma_{k,n}\|_2 = O_G(n^{-1/2}\Delta_{k,n}\delta_{k,n}^{-4}(\Delta_{k,n}\delta_{k,n}^{-1})^{\iota}) = o_G(1)$ *and* $\|\hat{\Gamma}_{k,n}\|_2 = o_G(1)$.

Proof. The proof follows the relevant parts of the proof of lemma A.2 in Chen and Bickel (2006). Firstly, from the representation of the derivative of the cubic spline in (42) we can write $c_{k,n,i} = \left(b_{k,n,i}^{(3)} - b_{k,n,i+1}^{(3)}\right)/\delta_{k,n}$. We have, for large enough $n \in \mathbb{N}$,

$$\begin{aligned}
|C_{k,n,i}| = |G_k c_{k,n,i}| &= \delta_{k,n}^{-1}\left|\int b_{k,n,i}^{(3)}(t)\eta_k(t)\,\mathrm{d}t - \int b_{k,n,i+1}^{(3)}(t)\eta_k(t)\,\mathrm{d}t\right| \\
&= \delta_{k,n}^{-1}\left|\int b_{k,n,i}^{(3)}(t)\eta_k(t)\,\mathrm{d}t - \int b_{k,n,i}^{(3)}(t)\eta_k(t+\delta_{k,n})\,\mathrm{d}t\right| \\
&\leq \left|\int b_{k,n,i}^{(3)}(t)\frac{\eta_k(t+\delta_{k,n}) - \eta_k(t)}{\delta_{k,n}}\,\mathrm{d}t\right| \\
&\leq 2\|\eta_k'\|_\infty \int b_{k,n,i}^{(3)}(t)\,\mathrm{d}t \\
&\leq 6\|\eta_k'\|_\infty \delta_{k,n},
\end{aligned}$$

where the last inequality is due to (20) on p. 91 in de Boor (2001) and the fact that splines (of any order) take values in $[0,1]$.[44] It follows immediately that for large enough $n \in \mathbb{N}$,

$$\sum_{i=1}^{B_{k,n}} C_{k,n,i}^2 \leq \sum_{i=1}^{B_{k,n}} 6^2 \|\eta_k'\|_\infty^2 \delta_{k,n}^2 = B_{k,n} 6^2 \|\eta_k'\|_\infty^2 \delta_{k,n}^2,$$

from which (i) follows.

We have that $c_{k,n,i} = \left(b_{k,n,i}^{(3)} - b_{k,n,i+1}^{(3)}\right)/\delta_{k,n}$ and since splines (of any order) take values in $[0,1]$ (both as noted above), we have that $c_{k,n,i} \in [-\delta_{k,n}^{-1}, \delta_{k,n}^{-1}]$. Hence, by Hoeffdings's inequality for $t \geq 0$ we have

$$G\left(\left|\frac{1}{n}\sum_{i=1}^{n} c_{k,n,m}(\epsilon_{k,i}) - G_k c_{k,n,m}\right| \geq t\right) \leq 2\exp\left(\frac{-n^2t^2}{2n\delta_{k,n}^{-2}}\right) = 2\exp(-nt^2\delta_{k,n}^2/2).$$

Therefore,

$$\begin{aligned} G\left(\|\hat{C}_{k,n} - C_{k,n}\|_2 \geq t\right) &\leq \sum_{m=1}^{B_{k,n}} G\left(\left|\frac{1}{n}\sum_{i=1}^{n} c_{k,n,m}(\epsilon_{k,i}) - G_k c_{k,n,m}\right| \geq \frac{t}{\sqrt{B_{k,n}}}\right) \\ &\leq 2B_{k,n}\exp(-nt^2 B_{k,n}^{-1}\delta_{k,n}^2/2), \end{aligned}$$

and so for any fixed $\epsilon > 0$ we can take $t = \sqrt{\frac{4B_{k,n}\log B_{k,n}}{n\delta_{k,n}^2}}$ to obtain

$$G\left(\|\hat{C}_{k,n} - C_{k,n}\|_2 \geq t\right) \leq 2B_{k,n}^{-1} \to 0,$$

yielding (ii).

Since for any $m, s \in [B_{k,n}]$ we have $b_{k,n,m}b_{k,n,s} \in [0,1]$ by Hoeffding's inequality it follows that for any $t \geq 0$

$$G\left(\left|\frac{1}{n}\sum_{i=1}^{n} b_{k,n,m}(\epsilon_{k,i})b_{k,n,s}(\epsilon_{k,i}) - G_k b_{k,n,m}b_{k,n,s}\right| \geq t\right) \leq 2\exp\left(\frac{-2n^2t^2}{n}\right) = 2\exp(-2nt^2).$$

Therefore, since $\|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_2 \leq \|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_F$ and both $\hat{\Gamma}_{k,n}$ and $\Gamma_{k,n}$ are zero for all (m,s)

[44] This is evident from their definition in (42). See also property (36) (p. 96) of de Boor (2001).

entries where $|m-s|>3$ (de Boor, 2001, (20), p. 91) we have that

$$\begin{aligned}
&G\left(\|\hat{\Gamma}_{k,n}-\Gamma_{k,n}\|_2 \geq t\right) \\
&\quad \leq G\left(\|\hat{\Gamma}_{k,n}-\Gamma_{k,n}\|_F \geq t\right) \\
&\quad \leq \sum_{m=1}^{B_{k,n}} \sum_{s=\max(m-3,1)}^{\min(B_{k,n},m+3)} G\left(\left|\frac{1}{n}\sum_{i=1}^{n} b_{k,n,m}(\epsilon_{k,i})b_{k,n,s}(\epsilon_{k,i}) - G_k b_{k,n,m}b_{k,n,s}\right| \geq \frac{t}{\sqrt{7B_{k,n}}}\right) \\
&\quad \leq 14B_{k,n}\exp\left(\frac{-2nt^2}{7B_{k,n}}\right).
\end{aligned}$$

Putting $t=\sqrt{\frac{7B_{k,n}\log B_{k,n}}{n}}$ we obtain

$$G\left(\|\hat{\Gamma}_{k,n}-\Gamma_{k,n}\|_2 \geq t\right) \leq 14B_{k,n}^{-1} \to 0,$$

yielding (iii).

Since $\Gamma_{k,n}$ is symmetric and positive (semi-)definite we have that $\|\Gamma_{k,n}\|_2 \leq \|\Gamma_{k,n}\|_\infty = \max_{m=1,\ldots,B_{k,n}} \sum_{s=1}^{B_{k,n}} G_k b_{k,n,m}b_{k,n,s}$.[45] Then, since for any $z\in\mathbb{R}$, each row of $b_{k,n}(z)b_{k,n}(z)'$ has at most 7 non-zero entries,[46] all of which are bounded above by 1 we have

$$\begin{aligned}
\|\Gamma_{k,n}\|_2 &\leq \max_{m=1,\ldots,B_{k,n}} \sum_{s=1}^{B_{k,n}} G_k b_{k,n,m}b_{k,n,s} \\
&= \max_{m=1,\ldots,B_{k,n}} \sum_{s=1}^{B_{k,n}} \int_{\xi_{k,n,m}}^{\xi_{k,n,m+4}} b_{k,n,m}(z)b_{k,n,s}(z)\eta_k(z)\,\mathrm{d}z \\
&\leq \max_{m=1,\ldots,B_{k,n}} 7\|\eta_k\|_\infty 4\delta_{k,n} \\
&= 28\|\eta_k\|_\infty \delta_{k,n},
\end{aligned}$$

which yields (iv) in conjunction with requirement (iii) of proposition 3.

By (v) of proposition 3, on $[\Xi_{k,n}^L, \Xi_{k,n}^U]$ we have $\eta(x) \geq c\delta_{k,n}$. Hence $\eta(x)-c\delta_{k,n}\geq 0$ and so $\int b_{k,n}b_{k,n}'(\eta-c\delta_{k,n})\lambda = \int (b_{k,n}\sqrt{\eta-c\delta_{k,n}})(b_{k,n}\sqrt{\eta-c\delta_{k,n}})'\lambda$. Note that the functions $b_{k,n,i}\sqrt{\eta-c\delta_{k,n}}$ satisfy $\int (b_{k,n,i}\sqrt{\eta-c\delta_{k,n}})^2\,\mathrm{d}\lambda < \infty$ and hence belong to $L_2(\lambda)$. It follows that the matrix $\int b_{k,n}b_{k,n}'(\eta-c\delta_{k,n})\lambda$ is a Gram matrix and hence positive semi-definite. This implies that $\Gamma_{k,n} \succeq c\delta_{k,n}\tilde{\Gamma}_{k,n}$ where $\tilde{\Gamma}_{k,n}$ is defined as in lemma 9. Hence, by the Rayleigh quotient theorem (see e.g. Theorem 4.2.2 in Horn and Johnson, 2013) and lemma 9

$$\lambda_{\min}(\Gamma_{k,n}) \geq \lambda_{\min}(c\delta_{k,n}\tilde{\Gamma}_{k,n}) = c\delta_{k,n}\lambda_{\min}(\tilde{\Gamma}_{k,n}) \geq c\upsilon\delta_{k,n}^2,$$

[45] See e.g. Theorem 5.6.9 in Horn and Johnson (2013).

[46] $b_{k,n,m}(z)=0$ outside $[\xi_{k,n,m},\xi_{k,n,m+4})$. See (20) on p. 91 in de Boor (2001).

for a $\upsilon > 0$, from which we may conclude that

$$\|\Gamma_{k,n}^{-1}\|_2 = \frac{1}{\lambda_{\min}(\Gamma_{k,n})} \leq (c\upsilon)^{-1}\delta_{k,n}^{-2},$$

which yields (v).

To demonstrate the last claim, note that with the results just derived, under our assumptions we have,

$$\|\hat{C}_{k,n}\|_2 \leq \|\hat{C}_{k,n} - C_{k,n}\|_2 + \|C_{k,n}\|_2 = O_G\left(\sqrt{\frac{B_{k,n}\log B_{k,n}}{n\delta_{k,n}^2}}\right) + O\left(\delta_{k,n}\sqrt{B_{k,n}}\right) = O_G\left(\delta_{k,n}\sqrt{B_{k,n}}\right),$$

and, using inequality (5.8.2) from Horn and Johnson (2013),

$$\begin{aligned}
\|\hat{\Gamma}_{k,n}^{-1}\|_2 &\leq \|\Gamma_{k,n}^{-1}(I + [\hat{\Gamma}_{k,n} - \Gamma_{k,n}]\Gamma_{k,n}^{-1})^{-1}\|_2 \\
&\leq \|\Gamma_{k,n}^{-1}\|_2 \|(I + [\hat{\Gamma}_{k,n} - \Gamma_{k,n}]\Gamma_{k,n}^{-1})^{-1}\|_2 \\
&\leq \|\Gamma_{k,n}^{-1}\|_2 \left(1 - \|[\hat{\Gamma}_{k,n} - \Gamma_{k,n}]\Gamma_{k,n}^{-1}\|_2\right)^{-1} \\
&\leq \|\Gamma_{k,n}^{-1}\|_2 \left(1 - \|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_2 \|\Gamma_{k,n}^{-1}\|_2\right)^{-1} \\
&= O_G(\delta_{k,n}^{-2}).
\end{aligned} \tag{52}$$

Using these intermediate results along with (ii) - (v) and our hypotheses we obtain that

$$\begin{aligned}
\|\hat{\gamma}_{k,n} - \gamma_{k,n}\|_2 &= \|\hat{\Gamma}_{k,n}^{-1}\hat{C}_{k,n} - \Gamma_{k,n}^{-1}C_{k,n}\|_2 \\
&\leq \|(\hat{\Gamma}_{k,n}^{-1} - \Gamma_{k,n}^{-1})\hat{C}_{k,n}\|_2 + \|\Gamma_{k,n}^{-1}(\hat{C}_{k,n} - C_{k,n})\|_2 \\
&\leq \|\Gamma_{k,n}^{-1}\|_2 \|\Gamma_{k,n} - \hat{\Gamma}_{k,n}\|_2 \|\hat{\Gamma}_{k,n}^{-1}\|_2 \|\hat{C}_{k,n}\|_2 + \|\Gamma_{k,n}^{-1}\|_2 \|\hat{C}_{k,n} - C_{k,n}\|_2 \\
&= O_G\left(\sqrt{\frac{B_{k,n}^2 \log B_{k,n}}{\delta_{k,n}^6 n}}\right) + O_G\left(\sqrt{\frac{B_{k,n}\log B_{k,n}}{\delta_{k,n}^6 n}}\right) \\
&= o_G(1),
\end{aligned}$$

by condition (ii) of proposition 3, since we have $B_{k,n} \leq \Delta_{k,n}\delta_{k,n}^{-1}$ and hence the dominant term above vanishes since for all large enough n,

$$\sqrt{\frac{B_{k,n}^2 \log B_{k,n}}{\delta_{k,n}^6 n}} \leq n^{-1/2}\Delta_{k,n}\delta_{k,n}^{-4}\log(\Delta_{k,n}\delta_{k,n}^{-1}) \leq n^{-1/2}\Delta_{k,n}\delta_{k,n}^{-4}(\Delta_{k,n}\delta_{k,n}^{-1})^{\iota} = o(1).$$

Finally, by (iii) and (iv) and condition (ii) of proposition 3 we have

$$\|\hat{\Gamma}_{k,n}\|_2 \leq \|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_2 + \|\Gamma_{k,n}\|_2 = O_G\left(\sqrt{\frac{B_{k,n}\log B_{k,n}}{n}}\right) + O(\delta_{k,n}) = o_G(1),$$

since $\delta_{k,n} \to 0$ and for large enough n,

$$\sqrt{\frac{B_{k,n} \log B_{k,n}}{n}} \leq n^{-1/2} \Delta_{k,n} \delta_{k,n}^{-1} \log(\Delta_{k,n} \delta_{k,n}^{-1}) \leq \delta_{k,n}^{3} n^{-1/2} \Delta_{k,n} \delta_{k,n}^{-4} (\Delta_{k,n} \delta_{k,n}^{-1})^{\iota} = o(1).$$

□

Appendix C: Tables and figures

Table 1: True Error Distributions

	Distribution
1	$\mathcal{N}(0,1)$
2	$t'(15)$
3	$t'(10)$
4	$t'(5)$
5	"skewed unimodal"
6	"kurtotic unimodal"
7	"outlier"
8	"bimodal"
9	"separate bimodal"
10	"skewed bimodal"

Notes: Distributions 2-4 are t-distributions normalised to have unit variance. Distributions 5 - 10 (and their names) are taken from Marron and Wand (1992); see their table 1 for the definitions and the plots on p. 717.

Table 2: Empirical Rejection Frequencies $\hat{S}^{SR}_{\hat{\theta}_n}$ test for Baseline ICA

n	K	B	1	2	3	4	5	6	7	8	9	10
200	2	4	0.041	0.047	0.038	0.043	0.047	0.051	0.047	0.052	0.047	0.044
200	2	6	0.045	0.043	0.042	0.044	0.045	0.054	0.047	0.053	0.051	0.047
200	2	8	0.046	0.047	0.047	0.046	0.043	0.051	0.046	0.050	0.053	0.047
200	3	4	0.031	0.040	0.037	0.037	0.043	0.047	0.041	0.047	0.046	0.042
200	3	6	0.038	0.042	0.038	0.037	0.045	0.046	0.044	0.042	0.049	0.044
200	3	8	0.041	0.046	0.040	0.042	0.048	0.047	0.043	0.044	0.045	0.042
500	2	4	0.047	0.041	0.041	0.045	0.045	0.048	0.048	0.051	0.048	0.050
500	2	6	0.043	0.044	0.046	0.041	0.048	0.052	0.049	0.050	0.050	0.048
500	2	8	0.047	0.048	0.043	0.044	0.049	0.046	0.051	0.053	0.049	0.050
500	3	4	0.041	0.043	0.040	0.042	0.047	0.041	0.045	0.052	0.048	0.050
500	3	6	0.039	0.044	0.043	0.043	0.045	0.047	0.047	0.046	0.048	0.046
500	3	8	0.041	0.043	0.045	0.046	0.045	0.045	0.051	0.046	0.050	0.047

Notes: The table shows the empirical rejection frequencies for the $S^{SR}_{\hat{\theta}_n}$ test based on $S = 5,000$ Monte Carlo replications for the baseline ICA model. The test has nominal size $a = 0.05$. The columns denote the sample size n, the dimension of the ICA model K, the number of B-splines B and the choice for densities ϵ_k, for $k > 1$, where the numbers correspond to the different densities listed in Table 1.

Table 3: Empirical Rejection Frequencies Alternative Tests for Baseline ICA

Test	1	2	3	4	5	6	7	8	9	10
$\hat{S}^{SR}_{\bar{\theta}_n}$	0.043	0.044	0.046	0.041	0.048	0.052	0.049	0.050	0.050	0.048
W	0.257	0.231	0.187	0.092	0.282	0.076	0.022	0.176	0.188	0.252
LM	0.072	0.090	0.075	0.065	0.109	0.063	0.069	0.065	0.066	0.087
LR	0.011	0.035	0.045	0.050	0.045	0.035	0.021	0.000	0.001	0.026
W^G	0.231	0.093	0.050	0.014	0.037	0.024	0.035	0.982	1.000	0.849
LR^L	0.164	0.141	0.106	0.092	0.149	0.168	0.345	0.117	0.112	0.161

Notes: The table shows the empirical rejection frequencies based on $S = 5,000$ Monte Carlo replications for the baseline ICA model with $n = 500$ and $K = 2$. All tests have nominal size $a = 0.05$. The first column indicates the test. In particular, $\hat{S}^{SR}_{\bar{\theta}_n}$ denotes the robust semi-parametric score test, W denotes the MLE-based Wald test, LM denotes the MLE-based Lagrange multiplier test, LR denotes the MLE-based likelihood ratio test, W^G denotes the Wald test based on the psuedo-maximum likelihood estimator of Gouriéroux, Monfort and Renne (2017), LR^L denotes the likelihood ratio test based on the GMM estimator of Lanne and Luoto (2019). The remaining columns denote the choice for densities ϵ_k, for $k \geq 2$, where the numbers correspond to the different densities listed in Table 1.

Table 4: Empirical Rejection Frequencies $\hat{S}^{SR}_{\bar{\theta}_n}$ test for LSEM

n	K	d	1	2	3	4	5	6	7	8	9	10
200	2	2	0.050	0.053	0.057	0.061	0.057	0.064	0.064	0.053	0.054	0.059
200	2	3	0.054	0.058	0.058	0.064	0.061	0.060	0.058	0.055	0.058	0.049
200	3	2	0.061	0.068	0.066	0.086	0.070	0.049	0.127	0.049	0.050	0.056
200	3	3	0.065	0.074	0.069	0.085	0.064	0.051	0.111	0.059	0.059	0.058
500	2	2	0.049	0.050	0.046	0.056	0.049	0.058	0.055	0.051	0.049	0.050
500	2	3	0.051	0.059	0.052	0.057	0.055	0.056	0.058	0.048	0.046	0.045
500	3	2	0.049	0.050	0.051	0.070	0.056	0.043	0.081	0.042	0.043	0.038
500	3	3	0.058	0.057	0.055	0.062	0.049	0.045	0.077	0.043	0.039	0.045

Notes: The table shows the empirical rejection frequencies for the $S^{SR}_{\bar{\theta}_n}$ test based on $S = 5,000$ Monte Carlo replications for the linear simultaneous equations model. The test has nominal size $a = 0.05$. The columns denote the sample size n, the dimension of the ICA model K, the number of covariates d and the choice for densities ϵ_k, for $k \geq 2$, where the numbers correspond to the different densities listed in Table 1. The $S^{SR}_{\bar{\theta}_n}$ test was implemented using $B = 6$ B-splines.

Table 5: Empirical Rejection Frequencies $\hat{S}^{SR}_{\bar{\theta}_n}$ test for Panel

T	n	K_z	d	1	2	3	4	5	6	7	8	9	10
5	200	2	2	0.042	0.044	0.040	0.048	0.036	0.033	0.052	0.031	0.034	0.030
5	200	2	3	0.039	0.042	0.048	0.056	0.031	0.031	0.064	0.027	0.031	0.033
5	200	3	2	0.047	0.043	0.044	0.043	0.040	0.032	0.053	0.033	0.038	0.034
5	200	3	3	0.047	0.043	0.046	0.059	0.036	0.031	0.047	0.031	0.036	0.036
5	500	2	2	0.045	0.042	0.045	0.041	0.036	0.034	0.044	0.031	0.034	0.032
5	500	2	3	0.044	0.037	0.037	0.044	0.032	0.034	0.051	0.033	0.033	0.029
5	500	3	2	0.046	0.042	0.044	0.041	0.035	0.032	0.043	0.035	0.033	0.030
5	500	3	3	0.040	0.046	0.040	0.047	0.031	0.028	0.044	0.032	0.032	0.028
10	200	2	2	0.047	0.043	0.044	0.049	0.035	0.029	0.051	0.032	0.035	0.034
10	200	2	3	0.041	0.042	0.047	0.050	0.034	0.033	0.055	0.028	0.034	0.032
10	200	3	2	0.045	0.040	0.042	0.043	0.037	0.032	0.054	0.034	0.035	0.032
10	200	3	3	0.050	0.049	0.042	0.052	0.038	0.030	0.044	0.031	0.037	0.031
10	500	2	2	0.046	0.038	0.040	0.045	0.038	0.032	0.040	0.032	0.031	0.030
10	500	2	3	0.043	0.041	0.039	0.042	0.036	0.029	0.045	0.034	0.030	0.030
10	500	3	2	0.046	0.042	0.042	0.046	0.034	0.028	0.046	0.032	0.037	0.034
10	500	3	3	0.042	0.037	0.042	0.045	0.038	0.029	0.038	0.033	0.033	0.029

Notes: The table shows the empirical rejection frequencies for the $S^{SR}_{\bar{\theta}_n}$ test based on $S = 5,000$ Monte Carlo replications for the short T panel data model. The test has nominal size $a = 0.05$. The columns denote the time series dimension T, the cross section dimension n, the dimension of the individual ICA model K_z, the number of covariates d and the choice for densities ϵ_k, for $k \geq 2$, where the numbers correspond to the different densities listed in Table 1. The $S^{SR}_{\bar{\theta}_n}$ test was implemented using $B = 6$ B-splines.

Table 6: Production function estimates 2017

	LSEM		OLS
Labor	[0.41, 0.64]	[0.44,0.68]	[0.89, 0.99]
Capital	[0.27, 0.50]	[0.32,0.50]	[0.18, 0.26]
Age		✓	✓
n	1247	1247	1247

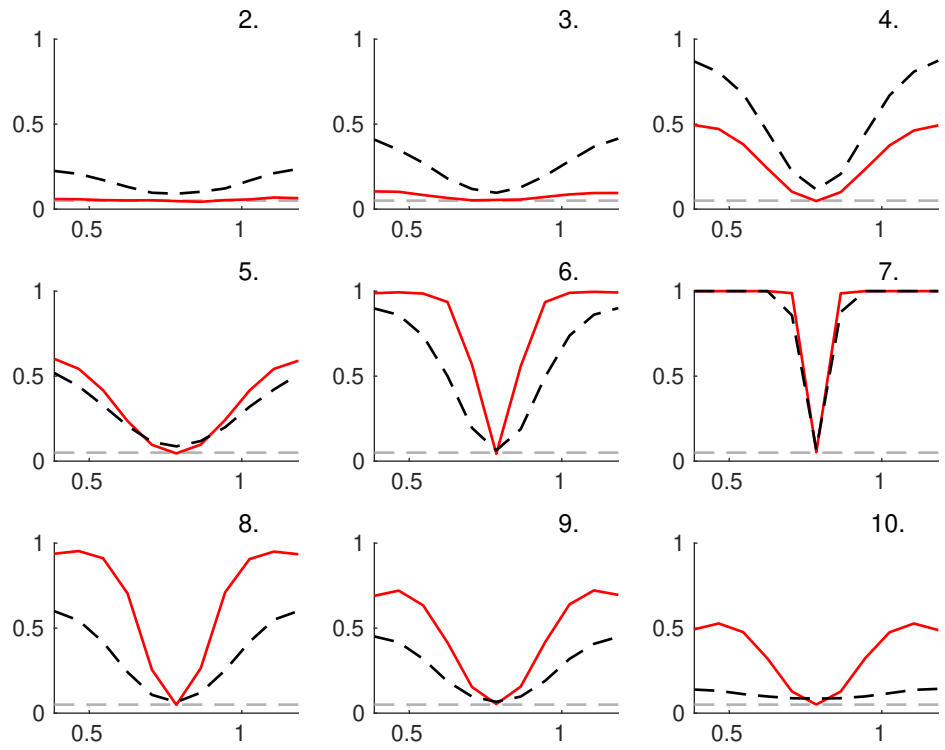
Notes: We report the 95% confidence bands for the production function coefficients for labor and capital. The first three columns consider the bounds obtained by considering the three-variable LSEM (i.e. $Y_i = (\log O_i, \log L_i, \log K_i)'$) with different explanatory variables as indicated in the rows. The right-most column displays the baseline OLS estimates for comparison.

Table 7: Production function estimates 2000-2017

	LSEM		FE
Labor	[0.53, 0.71]	[0.51,0.69]	[0.75, 0.92]
Capital	[0.09, 0.22]	[0.10,0.24]	[0.11, 0.30]
Time dummies		✓	✓
n	638	638	638

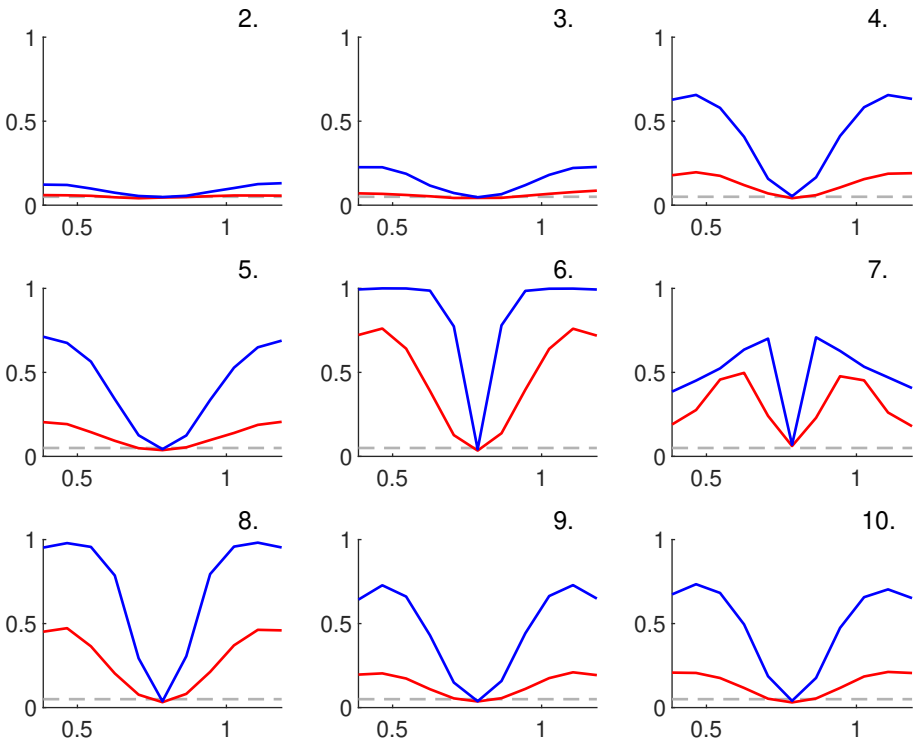
Notes: We report the 95% confidence bands for the production function coefficients for labor and capital. The first three columns consider the bounds obtained by considering the three-variable LSEM (i.e. $Y_i = (\log O_i, \log L_i, \log K_i)'$) with different explanatory variables as indicated in the rows. The right-most column displays the baseline OLS estimates for comparison.

Figure 1: Power Baseline ICA model



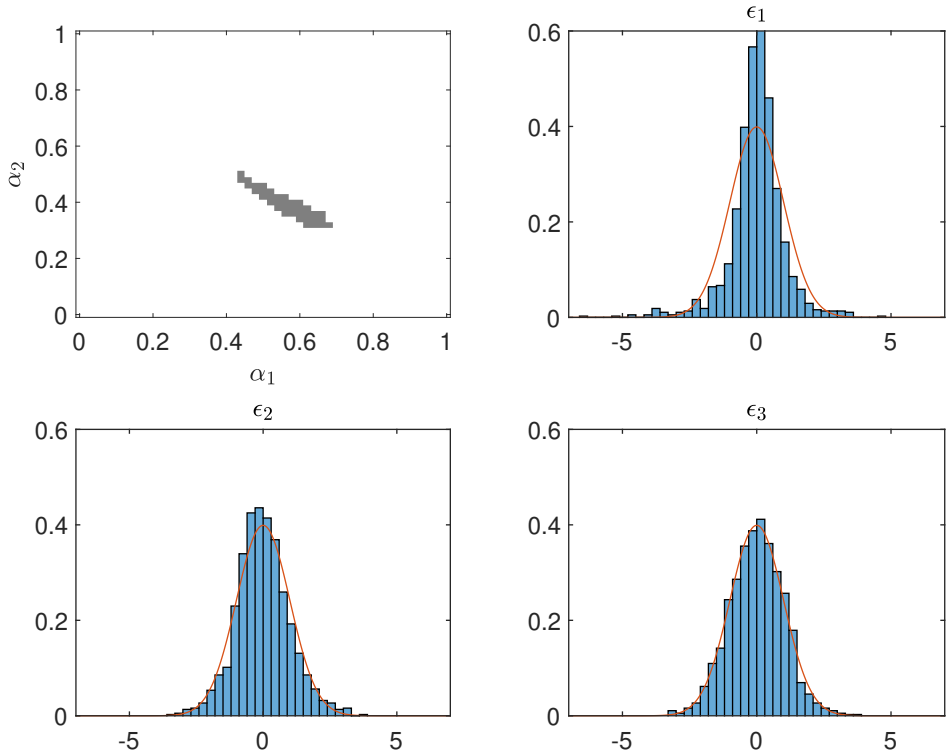
Notes: Empirical power curves for the baseline ICA model with $k = 2$ and $n = 500$. Each plot corresponds to the choice for densities ϵ_k, for $k \geq 2$, where the numbers correspond to the different densities listed in Table 1. The solid red line shows the empirical rejection frequency for the S_n^{SR} test whereas the black dashed line corresponds to the parametric LM test which is size-adjusted. Note that the parametric LM test is size adjusted.

Figure 2: Power LSEM



Notes: Empirical power curves for the LSEM model with $k = 2$, $d = 2$ and $n = 500$. Each plot corresponds to the choice for densities ϵ_k, for $k \geq 2$, where the numbers correspond to the different densities listed in Table 1. The red curve corresponds to the empirical rejection frequency of the $S^{SR}_{\bar{\theta}_n}$ test when the first component ϵ_1 is standard normal. The blue curve is when ϵ_1 varies across the different densities listed in Table 1.

Figure 3: LSEM Production Function Output 2017



Notes: The top left panel shows the confidence region for the labor α_1 and capital α_2. The other three panels show the empirical densities of the residuals together with the standard normal distribution.

Figure 4: Confidence intervals labor and capital 2000-2017

Notes: The vertical lines describe the confidence bands for labor and capital for each year between 2000 and 2017. Each pair of bands is based on firms observed in the corresponding year and estimated using the LSEM .