Robust Inference for Non-Gaussian Linear Simultaneous Equations Models*

Adam Lee and Geert Mesters

Universitat Pompeu Fabra and Barcelona School of Economics

July 6, 2022

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 identifying assumptions suffer from size distortions when the true distributions of the 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 finite sample properties of the methodology are illustrated in a large simulation study and an empirical study for production function estimation.

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 analyze 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 identified from the first and second moments of 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 identification approach has a long history in the statistics and signal processing literatures where it is often referred to as independent components analysis, see Hyvärinen, Karhunen and Oja (2001) for a textbook treatment. More recently, the econometrics 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, if in the true data generating process multiple structural shocks follow a Gaussian distribution some structural parameters may be under- or un-identified and standard inference methods that aim to exploit non-Gaussian distributions may fail to control size. Moreover, as is typical in models with points of identification failure, such behavior is also observed if the true distributions of the shocks are sufficiently to close to Gaussianity, relative to the sampling variation. Intuitively, in such *weakly non-Gaussian* settings local identification deteriorates leading to coverage distortions when using standard inference methods, such as maximum likelihood and moment methods.

Similar (weak) identification problems occur in many other econometric models, e.g. instrumental variable models, nonlinear regression models and many others, see Andrews and Cheng (2012, 2013) for numerous examples. The key difference between this existing literature and the non-Gaussian LSEM is that, in the latter, the parameters responsible for the possible identification failure are density functions, i.e. infinite dimensional parameters. Therefore, whilst conceptually the identification problem is the same, providing robust inferential methods requires a new approach which is capable of handling identification failure caused by infinite dimensional nuisance parameters.

[1] See for instance: Lanne and Lütkepohl (2010), Moneta et al. (2013), Lanne, Meitz and Saikkonen (2017), Maxand (2018), Lanne and Luoto (2021), Gouriéroux, Monfort and Renne (2017, 2019), Tank, Fox and Shojaie (2019), Herwartz (2019), Herwartz, Lange and Maxand (2019), Bekaert, Engstrom and Ermolov (2019, 2020), Fiorentini and Sentana (2022), Velasco (2022), Guay (2020), Moneta and Pallante (2020), Drautzburg and Wright (2021), Sims (2021) and Davis and Ng (2022).

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 semi-parametric statistical theory that is discussed in Bickel et al. (1998) and van der Vaart (2002). In brief, we treat the LSEM as a semiparametric model, where the densities of the independent structural shocks are treated non-parametrically, and we construct confidence bands for the possibly unidentified structural 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 true distribution of the shocks.

Intuitively, the efficient score test that we propose is the semi-parametric analog of Neyman's $C(\alpha)$ test (e.g. Neyman, 1979; Hall and Mathiason, 1990). In the conventional $C(\alpha)$ test the scores of the parameter of interest are orthogonalized with respect to the scores of the *finite dimensional* nuisance parameters. In our setting the nuisance parameter includes the densities of the shocks, i.e. an *infinite dimensional* parameter. While such nuisance functions result in the orthogonal projection being more technically demanding to derive, the main idea of Neyman (1979) continues to apply.

We evaluate the finite sample performance of the semiparametric score test in a large simulation study. This shows that regardless of how close the errors are 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 to be preferable.

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; capital and labor (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.

Overall, 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.

Throughout this paper we retain the assumption that the structural shocks are independent which may not be the case in practice, see the discussions in Matteson and Tsay (2017), Davis and Ng (2022) and Montiel Olea, Plagborg-Møller and Qian (2022). Therefore, in our empirical study we test the independence of the structural shocks following the approach of Matteson and Tsay (2017) and find that for our empirical application we cannot reject the independence assumption.

The remainder of this paper is organized as follows. In the next section we provide a simple example that illustrates the identification problem and intuitively discusses our solution. Section 3 presents the main LSEM model and provides the implementation details for the efficient score test. Section 4 discusses the main theoretical results including the required assumptions. Sections 5 and 6 summarize the results from the simulation and empirical studies. Section 7 concludes. Unless otherwise mentioned all proofs are provided in the Appendix. Any references to sections, equations, lemmas etc. which start with "S" refer to the supplementary material.

2 Illustrative example

In this section we use a simple example to illustrate: (i) the identification problem in LSEMs, (ii) why conventional inference methods suffer from size distortions when the structural shocks have densities close to Gaussian and (iii) how our proposed approach circumvents such distortions.

The identification problem

Consider the simple bi-variate model

$$Y_i = R'\epsilon_i \ , \qquad i = 1, \ldots, n \ , \tag{1}$$

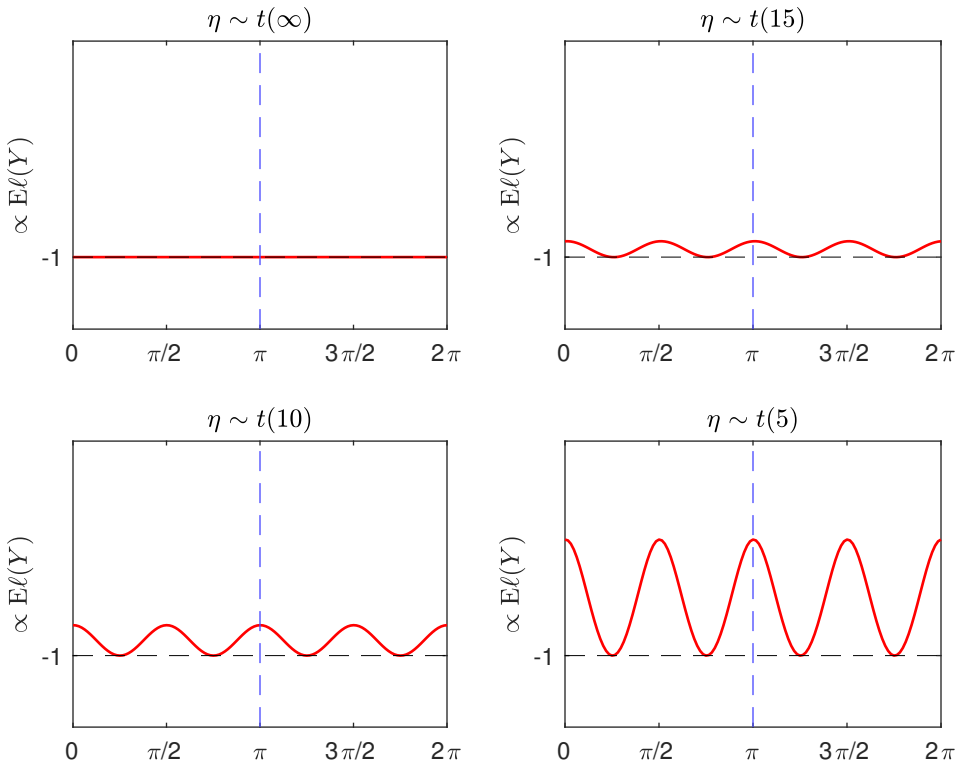
where Y_i is a vector of observable variables, R is rotation matrix (i.e. $R'R = I_2$) and ϵ_i is a vector with independent structural shocks $\epsilon_{i,k}$, for $k = 1, 2$, that have mean zero, unit variance and common density η. For concreteness, we will parameterize the rotation matrix as follows

$$R = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix} \ , \tag{2}$$

where $\alpha \in [0, 2\pi]$ and we let α_0 denote the true parameter.[2]

[2]Note that in general a researcher may consider $Y_i = \Sigma^{1/2} R'\epsilon_i$, where $\Sigma^{1/2}$ is lower triangular. However, the elements of $\Sigma^{1/2}$ can be identified from the variance of Y_i and pose no difficulty. Therefore we set the

Figure 1: (Weak) Non-Gaussian Identification



Notes: In the figure we show the expected log likelihood (red line) as a function of α (the true value is $\alpha_0 = \pi$).

Model (1) has two parameters: the parameter of interest α and the infinite dimensional nuisance parameter η. Suppose for now that η is known and let the log likelihood function for Y_i be denoted by $\ell_\alpha(\cdot)$. α is locally identified if the expected score of $\ell_\alpha(Y_i)$ with respect to α is non-zero for all $\alpha \neq \alpha_0$ in a neighborhood of α_0.

Whether local identification occurs turns out to depend crucially on η. To illustrate, consider the case where η is equal to the Gaussian density. Since ϵ_i is normalized we have

$$\mathbb{E}\ell_\alpha(Y_i) \propto -\frac{1}{2}\mathbb{E}(RY_i)'(RY_i) = -1$$

and hence the expected loglikelihood takes the same value irrespective of α. This is plotted in the top left panel of Figure 1, where we show the expected likelihood $\mathbb{E}\ell_\alpha(Y_i)$ as a function of α with $\alpha_0 = \pi$ as the true parameter (an arbitrary choice). This illustrates the

variance of Y_i to unity and exclude $\Sigma^{1/2}$ for simplicity.

standard identification problem in linear simultaneous equations models: without additional identifying restrictions, the impact effects of the structural shocks are not identifiable when the structural shocks follow a Gaussian distribution.

The other plots in Figure 1 show that this is no longer the case when we move away from the Gaussian distribution. In each case the expected gradient becomes non-zero at values $\alpha \neq \alpha_0$ in the vicinity of α_0, i.e. local identification occurs. While for the Student's t distribution with five degrees of freedom (i.e. $t(5)$) the change in the value of the expected likelihood is substantial it is easy to see that for more modest deviations from Gaussianity (e.g. $t(15)$) the difference is less pronounced. Further, note that non-Gaussian densities do not imply that α is globally identified, instead identification is only up to permutation and sign of the shocks.

Finite sample size distortions

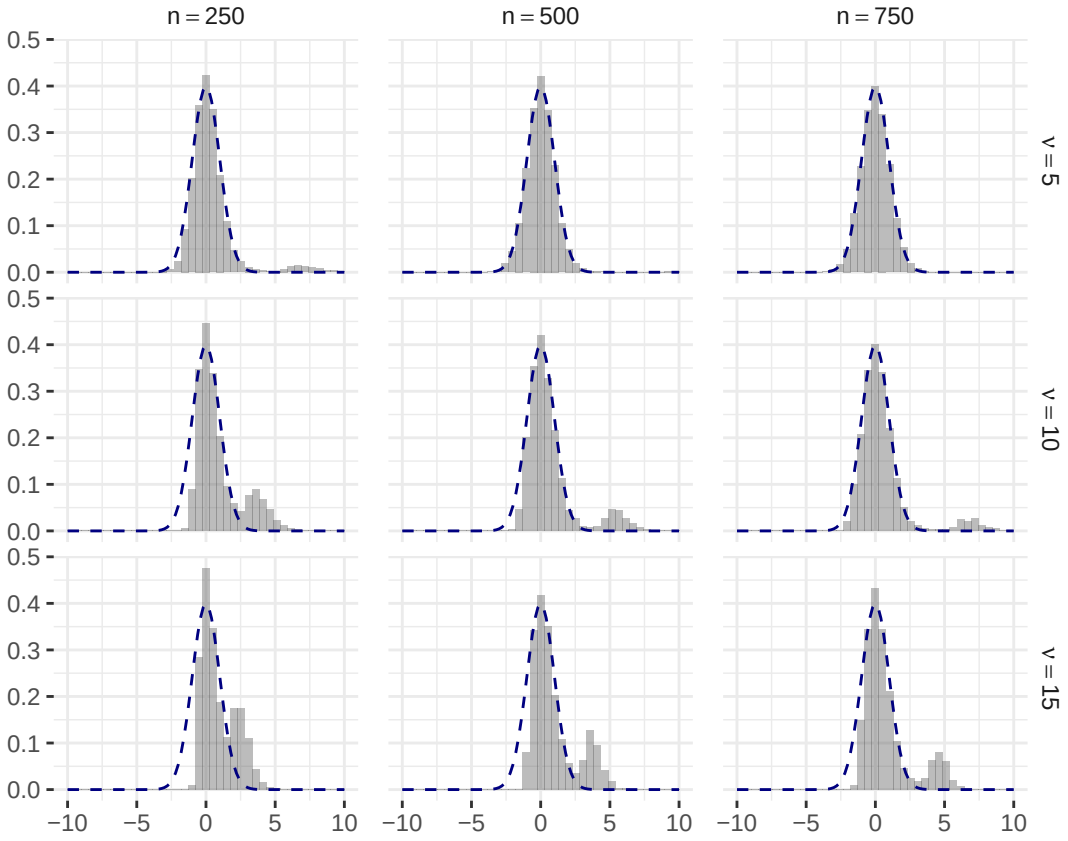
In population α is always locally identified when all but one component of η is non-Gaussian (Comon, 1994), but this is not sufficient for good performance of standard testing procedures in finite samples. In particular, if the structural errors are too close to Gaussian, the available identifying information may be small relative to the sampling variability. Standard asymptotic approximations are not reliable in this setting and, as a result, testing procedures based on these approximations may fail to provide reliable inference.

To illustrate how the density η affects standard inference methods in finite sample consider Figure 2 which depicts the finite sample distribution of the t-statistic for the hypothesis $H_0 : \alpha = \alpha_0$, based on the maximum likelihood estimator under the assumption that η is known. The blue dashed lines show the $\mathcal{N}(0, 1)$ density. As can clearly be seen in this figure, the quality of the approximation provided by the standard Normal depends crucially on the underlying density, η. For a given sample size, the approximation deteriorates substantially the closer η is to a standard Gaussian density.

This deterioration results in poor size control of standard tests. Table 1 shows the empirical rejection frequencies for three standard tests in the same setting: Wald (W), likelihood ratio (LR) and Lagrange multiplier (LM) (or score) tests, all computed under the assumption that η is known. Specifically we drew 5000 samples $\{Y_i\}_{i=1}^n$ from model (1) for different η's using different sample sizes $n = 250, 500, 750$. The empirical rejection frequencies correspond to the test for $H_0 : \alpha = \alpha_0$ with nominal size $a = 0.05$, where the critical values are based on the standard $\chi^2(1)$ asymptotic approximation.

We find that the Wald test is severely size distorted for η close to Gaussian; in view of the poor quality of asymptotic approximation depicted in Figure 2 this is not surprising. As η gets closer to Gaussianity, the likelihood ratio test starts to under-reject as when α is poorly

Figure 2: Poor asymptotic approximation close to Gaussianity



Notes: In the figure we show the finite sample distribution of the t-statistic based on the maximum likelihood estimator of α (the true value is $\alpha_0 = \pi$) for different sample sizes (n) and different degrees of freedom (ν) in the (standardised) t distribution, all based on 5000 replications.

identified the likelihood values are very similar. Both of these tests are based on estimates of α and, in weakly identified settings, such estimates will be inaccurate. In contrast, the score test (LM) shows correct size as it fixes $\alpha = \alpha_0$ under the null and α does not need to be (well) identified for this test to be correctly sized.

Towards a semi-parametric score test

Now in practice, η will be unknown and needs to be estimated. To build up to our semi-parametric approach, consider first the case where η is known up to a finite dimensional parameter vector, say β (for example β may include the degrees of freedom of the Student's t distribution). For this case Neyman (1979) proposed a convenient extension of the standard

Table 1: Rejection Frequencies for ML tests close to Gaussianity

	t(15)			t(10)			t(5)		
n	W	LM	LR	W	LM	LR	W	LM	LR
250	25.26	4.42	3.74	20.56	4.24	4.04	8.88	4.84	4.08
500	21.76	4.54	4.52	13.10	4.38	3.60	6.38	4.42	4.92
750	17.12	4.96	3.94	9.90	4.88	3.42	6.12	5.28	5.64

Notes: The table shows the empirical rejection frequencies for the three maximum likelihood tests, under the assumption that η is known and based on 5000 Monte Carlo replications for the baseline model $Y_i = R'\epsilon_i$. The test has nominal size $a = 0.05$.

score test, that amounts to first orthogonalizing the scores for α with respect to the scores for β and then computing a quadratic form of the score statistic. To illustrate let $\dot{\ell}(Y_i) = (\dot{\ell}_\alpha(Y_i), \dot{\ell}_\beta(Y_i))'$, $\dot{\ell}_\alpha(Y_i) = \nabla_\alpha \ell(Y_i)$, $\dot{\ell}_\beta(Y_i) = \nabla_\beta \ell(Y_i)$ and $\hat{I} = \frac{1}{n}\sum_{i=1}^n \dot{\ell}(Y_i)\dot{\ell}(Y_i)'$, denote the score and information matrix for α and β. Neyman's $C(\alpha)$ test statistic is given by

$$C(\alpha) = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\kappa}(Y_i)\right)' \hat{\mathcal{I}}^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\kappa}(Y_i)\right) ,$$

with

$$\hat{\kappa}(Y_i) = \dot{\ell}_\alpha - \hat{I}_{\alpha\beta}\hat{I}_{\beta\beta}^{-1}\dot{\ell}_\beta \qquad \text{and} \qquad \hat{\mathcal{I}} = \hat{I}_{\alpha\alpha} - \hat{I}_{\alpha\beta}\hat{I}_{\beta\beta}^{-1}\hat{I}_{\beta\alpha} ,$$

where $\hat{I}_{..}$ denote the corresponding blocks of $\hat{I}$.[3] The (estimated) orthogonalized scores $\hat{\kappa}(\cdot)$ are often referred to as the (estimates of the) efficient scores and $\hat{\mathcal{I}}$ is the corresponding (estimate of the) efficient information matrix. When evaluating $C(\alpha)$ at $\alpha = \alpha_0$ and $\hat{\beta}$, some $\sqrt{n}$ consistent estimate for β, this statistic will converge to a standard χ^2 limit under the null provided that $\hat{\mathcal{I}}$ is invertible.[4] Tests based on $C(\alpha)$ retain correct size regardless of whether α is well identified as α is fixed under H_0, making them attractive for settings where identification failure due to finite dimensional nuisance parameters is a concern (e.g. Andrews and Mikusheva, 2015).

In the present paper, we will not impose that the parametric form of η is known up to finite dimensional parameters but instead treat η non-parametrically. Despite this change,

[3]This is numerically equivalent to the "usual" score test provided the nuisance parameter β is estimated by (restricted) maximum likelihood under the null hypothesis (Kocherlakota and Kocherlakota, 1991).

[4]In our general framework below we explicitly allow $\hat{\mathcal{I}}$ to be singular and rely on an eigenvalue truncated generalized inverse, see also Andrews (1987), Lütkepohl and Burda (1997) and Andrews and Guggenberger (2019).

our approach is similar to that sketched above. We will first orthogonalize the score for α with respect to the scores for η and obtain a semi-parametric analog of the conventional Neyman $C(\alpha)$ test. This requires technical adjustments as the scores with respect to η need to be defined differently and the projection with respect to η scores requires more care. For this we follow the semi-parametric literature as outlined in the textbooks of Bickel et al. (1998) and van der Vaart (2002).

3 Robust inference for LSEMs

In this section we discuss the implementation of the semi-parametric score test for a general class of linear simultaneous equations models.

3.1 General model and objectives

We consider the linear simultaneous equations model for a random sample of the $K \times 1$ endogenous variables Y_i, the $d \times 1$ exogenous variables $X_i = (1, \tilde{X}_i')'$ and the $K \times 1$ structural shocks ϵ_i. Specifically,

$$Y_i = BX_i + A^{-1}\epsilon_i \ , \qquad i = 1, \ldots, n \ , \tag{3}$$

where the matrices B and A^{-1} map the explanatory variables and the structural shocks to the endogenous variables. The density functions of the components of $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iK})'$ are denoted by $(\eta_1, \ldots, \eta_K)$ and the density of $\tilde{X}_i$ is given by η_0. We set $\eta = (\eta_0, \eta_1, \ldots, \eta_K)$.

As illustrated in the previous section, depending on the shapes of $\eta_1, \ldots, \eta_K$ we may not be able to identify all parameters in A. To model this we let $A = A(\alpha, \sigma)$, where $A(\alpha, \sigma)$ is a function of the possibly unidentified parameters α and parameters σ which can be always identified from the variance of $Y_i - BX_i$. We let $\alpha \in \mathcal{A} \subset \mathbb{R}^{L_\alpha}$ and set $\beta = (\sigma, b) \in \mathcal{B} \subset \mathbb{R}^{L_\sigma} \times \mathbb{R}^{L_b} = \mathbb{R}^{L_\beta}$, with $b = \text{vec}(B)$. The following two examples illustrate possible parametrizations for $A(\alpha, \sigma)$ that are of practical interest.

Example 1 (Rotation matrix)**.** *Let $A(\alpha, \sigma)^{-1} = \Sigma^{1/2}R'$, where $\Sigma^{1/2}$ is lower triangular and R is a rotation matrix. In this setting we can take $\sigma = \text{vech}(\Sigma^{1/2})$ and α parametrizes R using the trigonometric transformation (as in Section 2) or the Cayley or exponential transformation of a skew-symmetric matrix (e.g. Gouriéroux, Monfort and Renne, 2017; Magnus, Pijls and Sentana, 2021).*

Example 2 (Supply and demand)**.** *For $K = 2$ let Y_{i1} denote the quantity of some good and*

Y_{i2} its price. A simple model (omitting covariates for convenience) is given by

$$Y_{i1}^d = aY_{i2} + \sigma_1\epsilon_{i1} \qquad \textit{(demand)}$$
$$Y_{i1}^s = bY_{i2} + \sigma_2\epsilon_{i2} \qquad \textit{(supply)}$$

where ϵ_{i1} and ϵ_{i2} are independent demand and supply shocks, and in equilibrium we have $Y_{i1}^d = Y_{i2}^s$. We can accommodate this set up by letting $\alpha = (a, b)$, $\beta = (\sigma_1, \sigma_2)$ and defining the mapping $A(\alpha, \sigma)$ according to

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

In the remainder we leave the precise mapping $A(\alpha, \sigma)$ unspecified, but we will require that it satisfies certain smoothness conditions.

The general LSEM (3) depends on the triplet of parameters $\theta = (\alpha, \beta, \eta)$, which includes the possibly unidentified parameters α, the finite dimensional nuisance parameters $\beta = (\sigma, b)$ and the infinite dimensional nuisance parameters η. We will refer to β as nuisance parameters as our main interest is in conducting inference on α, but clearly β could also be an object of interest. To conduct inference on α without making a priori assumptions on the identification strength of α, i.e. without assuming that sufficiently many η_k's are non-Gaussian, we consider hypothesis tests of the form

$$H_0 : \alpha = \alpha_0 \qquad \text{against} \qquad H_1 : \alpha \neq \alpha_0 \ . \tag{4}$$

Such test statistics can then be inverted to yield confidence intervals for α with correct coverage.

The problem formulation reflects that we aim for a procedure that is valid for all densities η_k, for $k = 1, \ldots, K$, Gaussian or not. A related set-up is found in Risk, Matteson and Ruppert (2019) and Jin, Risk and Matteson (2019) who assume that the structural shocks can be separated into *exactly* Gaussian and non-Gaussian shocks. We do not impose such structure, but we note that if indeed shocks can be separated in this way our approach will remain valid, but likely less efficient when compared to Risk, Matteson and Ruppert (2019).

3.2 Efficient score test for LSEMs

Next, we provide a step by step implementation guide for the semi-parametric score test, with the theoretical justification postponed to the next section.

Efficient score and information matrix estimates

As a first step, let $\hat{\ell}_\gamma(V_i)$ denote the estimates for efficient scores of the finite dimensional parameters $\gamma = (\alpha, \beta)$ of the LSEM (3) evaluated at $V_i = Y_i - BX_i$ and γ. Intuitively, these are the estimates for the scores of the parameters γ that are obtained after projecting out the infinite dimensional nuisance parameter η. As we show in the appendix, consistent estimates for the components of $\hat{\ell}_\gamma(V_i)$ are given by

$$\hat{\ell}_\gamma(V_i) = \begin{bmatrix} \hat{\ell}_{\gamma,\alpha}(V_i) \\ \hat{\ell}_{\gamma,\beta}(V_i) \end{bmatrix} = \begin{bmatrix} \{\hat{\ell}_{\gamma,\alpha_l}(V_i)\}_{l=1}^{L_\alpha} \\ \{\hat{\ell}_{\gamma,\beta_l}(V_i)\}_{l=1}^{L_\beta} \end{bmatrix} \quad \text{with} \quad \hat{\ell}_{\gamma,\beta}(V_i) = \begin{bmatrix} \hat{\ell}_{\gamma,\sigma}(V_i) \\ \hat{\ell}_{\gamma,b}(V_i) \end{bmatrix} = \begin{bmatrix} \{\hat{\ell}_{\gamma,\sigma_l}(V_i)\}_{l=1}^{L_\sigma} \\ \{\hat{\ell}_{\gamma,b_l}(V_i)\}_{l=1}^{L_b} \end{bmatrix},$$

and

$$\begin{aligned}
\hat{\ell}_{\gamma,\alpha_l}(V_i) &= \sum_{j,k=1,j\neq k}^{K} \zeta_{l,k,j}^{\alpha} \hat{\phi}_k(A_{k\bullet}V_i) A_{j\bullet}V_i + \sum_{k=1}^{K} \zeta_{l,k,k}^{\alpha} \left[\hat{\tau}_{k,1} A_{k\bullet}V_i + \hat{\tau}_{k,2}\kappa(A_{k\bullet}V_i)\right] \\
\hat{\ell}_{\gamma,\sigma_l}(V_i) &= \sum_{j,k=1,j\neq k}^{K} \zeta_{l,k,j}^{\sigma} \hat{\phi}_k(A_{k\bullet}V_i) A_{j\bullet}V_i + \sum_{k=1}^{K} \zeta_{l,k,k}^{\sigma} \left[\hat{\tau}_{k,1} A_{k\bullet}V_i + \hat{\tau}_{k,2}\kappa(A_{k\bullet}V_i)\right] \qquad (5) \\
\hat{\ell}_{\gamma,b_l}(V_i) &= \sum_{k=1}^{K} [-A_{k\bullet}D_{b,l}][(X_i - \bar{X}_n)\hat{\phi}_k(A_{k\bullet}V_i) - \bar{X}_n(\hat{\varsigma}_{k,1}A_{k\bullet}V_i + \hat{\varsigma}_{k,2}\kappa(A_{k\bullet}V_i))]
\end{aligned}$$

where $A_{k\bullet}$ denotes the kth row of A, $\kappa(z) = 1 - z^2$, $\zeta_{l,k,j}^{\alpha} := [D_{\alpha,l}]_{k\bullet} A_{\bullet j}^{-1}$, $\zeta_{l,k,j}^{\sigma} := [D_{\sigma,l}]_{k\bullet} A_{\bullet j}^{-1}$, $D_{\alpha,l} = \partial A(\alpha,\sigma)/\partial\alpha_l$, $D_{\sigma,l} = \partial A(\alpha,\sigma)/\partial\sigma_l$, $D_{b_l} = \partial B/\partial b_l$ and $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$. The coefficients $\hat{\tau}_k = (\hat{\tau}_{k,1}, \hat{\tau}_{k,2})'$ and $\hat{\varsigma}_k = (\hat{\varsigma}_{k,1}, \hat{\varsigma}_{k,2})'$ are given, for $k = 1, \ldots, K$, by

$$\hat{\tau}_k = \hat{M}_k^{-1} \begin{pmatrix} 0 \\ -2 \end{pmatrix}, \quad \hat{\varsigma}_k = \hat{M}_k^{-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \hat{M}_k = \begin{pmatrix} 1 & \frac{1}{n}\sum_{i=1}^n (A_{k\bullet}V_i)^3 \\ \frac{1}{n}\sum_{i=1}^n (A_{k\bullet}V_i)^3 & \frac{1}{n}\sum_{i=1}^n (A_{k\bullet}V_i)^4 - 1 \end{pmatrix}. \quad (6)$$

Finally, the efficient score estimates (5) depend on $\hat{\phi}_k(\cdot)$: the estimate for the log density score $\phi_k(x) = \partial\eta_k(x)/\partial x$. Such estimates can be obtained in different ways and our preferred approach is based on using B-splines as in Jin (1992) and Chen and Bickel (2006). We can define such estimates as

$$\hat{\phi}_k(x) = \hat{\gamma}_k' b_k(x) \qquad \text{with} \quad \hat{\gamma}_k = -\left[\sum_{i=1}^n b_k(A_{k\bullet}V_{k,i}) b_k(A_{k\bullet}V_{k,i})'\right]^{-1} \sum_{i=1}^n c_k(A_{k\bullet}V_{k,i}) \,, \qquad (7)$$

where $b_k(x) = (b_{k,1}(x), \ldots, b_{k,B_k}(x))'$ is a collection of B_k cubic B-splines and $c_k(x) = (c_{k,1}(x), \ldots, c_{k,B_k}(x))'$ are their derivatives: $c_{k,i}(x) = \frac{\mathrm{d}b_{k,i}(x)}{\mathrm{d}x}$ for each $i = 1, \ldots, B_k$, see de Boor (2001) for more details on B-splines. In practice we rely on equally spaced knots

with upper and lower end points taken to be the 95th and 5th percentile of the samples $\{A_{k\bullet}V_{i,k}\}_{i=1}^n$ adjusted by $\log(\log(n))$. We use $B_k = 6$ splines in our main simulations below and investigate the sensitivity of this choice.

Given the estimates of the efficient scores we estimate the efficient information matrix, which is the variance matrix of the efficient score function, as

$$\hat{I}_\gamma = \frac{1}{n}\sum_{i=1}^n \hat{\ell}_\gamma(V_i)\hat{\ell}_\gamma(V_i)' \qquad \text{with partitioning} \quad \hat{I}_\gamma = \begin{bmatrix} \hat{I}_{\gamma,\alpha\alpha} & \hat{I}_{\gamma,\alpha\beta} \\ \hat{I}_{\gamma,\beta\alpha} & \hat{I}_{\gamma,\beta\beta} \end{bmatrix} . \tag{8}$$

Efficient score statistic

To compute the efficient semi-parametric score statistic for testing $H_0 : \alpha = \alpha_0$ we first orthogonalize the efficient scores for α with respect to those for $\beta = (\sigma, b)$. Since, β is finite dimensional the estimates of the resulting orthogonalized scores and information for α are given by

$$\hat{\kappa}_\gamma(V_i) = \hat{\ell}_{\gamma,\alpha}(V_i) - \hat{I}_{\gamma,\alpha\beta}\hat{I}^{-1}_{\gamma,\beta\beta}\hat{\ell}_{\gamma,\beta}(V_i) \qquad \text{and} \qquad \hat{\mathcal{I}}_\gamma = \hat{I}_{\gamma,\alpha\alpha} - \hat{I}_{\gamma,\alpha\beta}\hat{I}^{-1}_{\gamma,\beta\beta}\hat{I}_{\gamma,\beta\alpha} . \tag{9}$$

These are estimates of the population efficient score and efficient information matrix. Importantly, the latter may not be positive definite in our setting. For instance, when the densities η_k correspond to the Gaussian density, $\mathcal{I}_\gamma$ is singular, see Lemma S11 in the supplementary material.

With $\hat{\kappa}_\gamma(V_i)$ and $\hat{\mathcal{I}}_\gamma$ we can define the efficient score statistic for the LSEM model as function of $\gamma = (\alpha, \beta)$ and $V_i = Y_i - BX_i$ by

$$\hat{S}_\gamma = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\kappa}_\gamma(V_i)\right)' \hat{\mathcal{I}}^{t,\dagger}_\gamma \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\kappa}_\gamma(V_i)\right) , \tag{10}$$

where $\hat{\mathcal{I}}^{t,\dagger}_\gamma$ denotes the generalized inverse of the eigenvalue truncated efficient information matrix $\hat{\mathcal{I}}_\gamma$ (e.g. Lütkepohl and Burda, 1997). Formally,

$$\hat{\mathcal{I}}^t_\gamma = \hat{U}_n\hat{\Lambda}_n(\nu_n)\hat{U}_n' , \tag{11}$$

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 $\hat{\lambda}_{n,i}\mathbf{1}(\hat{\lambda}_{n,i} \geq \nu_n)$.

Equations (5)-(11) define the semi-parametric score statistic for the LSEM model (3) for

a given parameter vector $\gamma = (\alpha, \beta)$. To test the null hypothesis (4) we will evaluate this test statistic at $\alpha = \alpha_0$, i.e. fixing the possibly unidentified parameters under the null, and at $\hat{\beta}$, which can be any $\sqrt{n}$ consistent estimate for β. In our simulations, we use ordinary least squares estimates for σ and $b = \text{vec}(B)$, or one-step efficient estimates following van der Vaart (2002, Section 7.2). Let $\hat{\gamma} = (\alpha_0, \hat{\beta})$, in our theoretical section below we show that under suitable assumptions the score statistic will converge to a χ^2 limit. Specifically, we prove that under H_0 for any $a \in (0, 1)$ we have

$$\lim_{n\to\infty} P(\hat{S}_{\hat{\gamma}} > c_n) \leq a \ , \tag{12}$$

where c_n is the $1 - a$ quantile of the $\chi^2_{r_n}$ distribution with $r_n = \text{rank}(\hat{\mathcal{I}}^t_{\hat{\gamma}})$. Importantly, as we show in section 4 this result does not rely on any assumptions regarding the shape of the densities η, i.e. we do not need to assume that η is non-Gaussian. Only conventional moment assumptions and some regularity conditions on the densities are required. The following algorithm summarizes the complete implementation.

Algorithm: Efficient score test for LSEM

1. Obtain $\sqrt{n}$-consistent estimates $\hat{\beta} = (\hat{\sigma}, \hat{b})$ and residuals $\hat{V}_i = Y_i - \hat{B}X_i$;
2. For $k = 1, \ldots, K$, compute $\hat{\phi}_k(\hat{A}_{k\bullet}\hat{V}_i)$ from (7) with $\hat{A} = A(\alpha_0, \hat{\sigma})$;
3. Compute the efficient scores $\hat{\ell}_{\hat{\gamma}}(\hat{V}_i)$ from (5) and the information matrix $\hat{I}_{\hat{\gamma}}$ from (8) using $\hat{\gamma} = (\alpha_0, \hat{\beta})$;
4. Compute $\hat{\kappa}_{\hat{\gamma}}(\hat{V}_i)$ and $\hat{\mathcal{I}}_{\hat{\gamma}}$ from (9).
5. Compute the score statistic $\hat{S}_{\hat{\gamma}}$ from (10) and reject $H_0 : \alpha = \alpha_0$ if $\hat{S}_{\hat{\gamma}} > c_n$, where c_n is the $1 - a$ quantile of the $\chi^2_{r_n}$ distribution with $r_n = \text{rank}(\hat{\mathcal{I}}^t_{\hat{\gamma}})$.

The algorithm highlights that the computational cost for evaluating the semi-parametric score statistic $\hat{S}_{\hat{\gamma}}$ is modest; effectively one only needs to compute K B-spline regressions to obtain the log density scores. Importantly, this implies that the algorithm can be implemented without relying on numerical optimization routines. Confidence sets for α can be constructed by inverting the score statistic over a range of values for α_0.

4 Asymptotic theory

In this section we present our main theoretical results and discuss the required underlying assumptions.

4.1 Assumptions

We assume that we observe a random sample $\{(Y_i, \tilde{X}_i)\}_{i=1}^n$ from model (3) where the underlying components satisfy the following.

Assumption 1. *For $\epsilon_i = (\epsilon_{i,1}, \ldots, \epsilon_{i,K})'$ in model (3), each component $\epsilon_{i,k}$ has a continuously differentiable root density (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. *$\mathbb{E}\epsilon_{i,k} = 0$, $\mathbb{E}\epsilon_{i,k}^2 = 1$, $\mathbb{E}\epsilon_{i,k}^{4+\delta} < \infty$, $\mathbb{E}(\epsilon_{i,k}^4) - 1 > \mathbb{E}(\epsilon_{i,k}^3)^2$, and $\mathbb{E}\phi_k^{4+\delta}(\epsilon_{i,k}) < \infty$;*
2. *$\mathbb{E}\phi_k(\epsilon_{i,k}) = 0$, $\mathbb{E}\phi_k(\epsilon_{i,k})\epsilon_{i,k} = -1$, $\mathbb{E}\phi_k(\epsilon_{i,k})\epsilon_{i,k}^2 = 0$ and $\mathbb{E}\phi_k(\epsilon_{i,k})\epsilon_{i,k}^3 = -3$;*
3. *$\epsilon_{i,k}$ is independent of $\epsilon_{i,l}$ for all $k \neq l$;*
4. *$\eta_0 \in \mathscr{Z}$ is a density function (with respect to Lebesgue measure on $\mathbb{R}^{d-1}$) such that if $\tilde{X}_i \sim \eta_0$, then $\mathbb{E}\tilde{X}_i\tilde{X}_i'$ is positive definite and $\mathbb{E}[|\tilde{X}_{i,l}|^{4+\delta}] < \infty$ for all $l = 1, \ldots, d-1$;*
5. *ϵ_i and $\tilde{X}_i$ are independent.*

The first part normalizes the errors to have mean zero, variance one and finite four+δ moments,[5] hence ruling out heavy tailed errors.[6] 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 S12 in the supplementary material 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.[7] The third part imposes that the components of ϵ_i are independent. Part four imposes some structure on $\tilde{X}_i$ that allows us to identify B; notably positive definite second moments and four+δ finite moments are required. Part five requires

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

[6] Heavy tailed errors in ICA and SVAR models have recently been considered in Davis and Ng (2022) and Davis and Fernandes (2022), but an inferential theory remains to be developed.

[7] 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.

the explanatory variables and errors to be independent. This can be relaxed by requiring the moment assumptions in 1 to hold conditional on $\tilde{X}_i$. In this setup, our general theory as outlined in this section would continue to be valid though the resulting efficient score function would take a different form.

Most important is what is *not* in Assumption 1: there is no condition that imposes that a certain number of components of ϵ_i have a (sufficiently) non-Gaussian distribution.

The second assumption that we impose is only required for the estimation of the log density scores $\phi(x) = \partial\eta(x)/\partial x$ using B-spline regressions and can be appropriately replaced when a different density score estimator is used. For notation purposes, let $\Xi^L_{k,n}$ and $\Xi^U_{k,n}$ denote the lower and upper endpoints of the cubic B-splines for $\phi_k(x)$ for $k = 1, \ldots, K$.[8]

Assumption 2. *Define ν_n according to $\nu^2_{n,p} = o(\nu_n)$ with $p := \min\{1 + \delta/4, 2\}$ 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 $p = 2$. Let $\phi_{k,n} := \phi_k \mathbf{1}_{[\Xi^L_{k,n}, \Xi^U_{k,n}]}$ and $\Delta_{k,n} := \Xi^U_{k,n} - \Xi^L_{k,n}$ and suppose that for , $[\Xi^L_{k,n}, \Xi^U_{k,n}] \uparrow \tilde{\Xi} \supset \text{supp}(\eta_k)$ and $\delta_{k,n} \downarrow 0$ such that*

(i) $P(\epsilon_{i,k} \notin [\Xi^L_{k,n}, \Xi^U_{k,n}]) = o(\nu^2_n)$;

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

(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 $\|\phi^{(3)}_{k,n}\|^2_\infty \delta^6_{k,n} = o(\nu_n)$;*[9]

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

First, the assumption makes explicit the truncation rate ν_n that is needed for the truncation of the eigenvalues in (11). This rate is split into two parts. The "slow" rate $n^{(1-p)/p}$ (for $p \in (1,2)$) is always sufficient given assumption 1, but if ϵ_k has finite eighth moments the faster rate applies.

Part (i) imposes that the tails of $\epsilon_{i,k}$ decay to zero sufficiently fast.[10] Part (ii) ensures that the number of knots does not grow to fast relative to the sample size (and the truncation rate). Part (iii) requires the density and its derivative to be bounded. Part (iv) requires the existence of the third derivatives of ϕ_k and that the rate of increase of the third derivative is not too great. Part (v) ensures that the density is bounded away from zero

[8] In practice, we select these points as the lower 5th and upper 95th percentiles of the samples $\{V_{i,k}\}^n_{i=1}$ adjusted by $\log\log n$, see the implementation section 3.

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

[10] The required speed of decay is linked to the truncation rate.

on $[\Xi^L_{k,n}, \Xi^U_{k,n}]$. Overall, these assumptions are similar as in Chen and Bickel (2006), with two key differences.[11] Firstly, Chen and Bickel (2006) require the conditions to hold for the functions $v \mapsto \phi_k(A_{k\bullet}v)$ (rather than ϕ_k), uniformly over shrinking balls (at rate $n^{-1/2}$) around A. In our setting we are only interested in testing as consistent estimation is ruled out by the possible lack of identification, hence we only require the conditions to hold for the functions ϕ_k. Secondly, unlike Chen and Bickel (2006), we require convergence at a rate ν_n which satisfies certain decay conditions. This is due to the fact that we may have a singular efficient information matrix and in order to obtain a consistent estimate of the Moore – Penrose inverse of this matrix, we require knowledge of the rate of convergence of our estimate.

4.2 Main result

In this section we formally state our main result for the efficient score test $\hat{S}_{\hat{\gamma}}$. To do so, instead of evaluating the efficient score test at the $\sqrt{n}$-consistent estimates $\hat{\gamma} = (\alpha_0, \hat{\beta})$ we will evaluate the score test at its discretized version $\bar{\gamma} = (\alpha_0, \bar{\beta}_n)$. Formally, let $B_n = n^{-1/2} C \mathbb{Z}^{L_\beta}$ for some $C > 0$ and define $\bar{\beta}_n$ as a new version of $\hat{\beta}$ that replaces its value with the closest point in B_n. Note that this changes each coordinate of $\hat{\beta}$ by a quantity which is at most $O(n^{-1/2})$, hence the $\sqrt{n}$-consistency is retained by discretization. Since the constant C can be chosen arbitrarily small this change has no practical relevance for the implementation of the test.

The advantage of relying on discretized estimates is that it simplifies the proof of the main result. Specifically, it removes the need to show uniform convergence between the efficient scores evaluated at $\hat{\beta}$ and β. The discretization trick is due to Le Cam (1960) and is widely used in statistics, see the detailed discussion in Le Cam and Yang (2000, Section 6.3), or van der Vaart (1998, page 72). It has also been adopted in econometrics, see Cattaneo, Crump and Jansson (2012) for instance.

With this modification we have the following result.

Theorem 1. *Suppose that Assumptions 1 and 2 hold, that $(\alpha, \sigma) \mapsto A(\alpha, \sigma)$ is continuously differentiable and the maps $(\alpha, \sigma) \mapsto \zeta^\alpha_{l,k,j}$ and $(\alpha, \sigma) \mapsto \zeta^\sigma_{l,k,j}$ are Lipschitz continuous. Let $r_n = \operatorname{rank}(\hat{\mathcal{I}}^t_{\bar{\gamma}})$ 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*

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

with inequality only if $\operatorname{rank}(\tilde{\mathcal{I}}_{\gamma_0}) = 0$ where $\gamma_0 = (\alpha_0, \beta)$.

[11] Cf. their conditions C3, C5 – C7, p. 2834.

The proposition shows that semi-parametric score test $\hat{S}_{\bar{\gamma}}$ has correct asymptotic size for all densities η that satisfy the requirements in Assumptions 1 and 2. The requirements that $(\alpha, \sigma) \mapsto A(\alpha, \sigma)$ is continuously differentiable and $(\alpha, \sigma) \to \zeta^{\alpha}_{l,k,j}$, $(\alpha, \sigma) \to \zeta^{\sigma}_{l,k,j}$ are Lipschitz continuous are easily verified for Examples 1 and 2. The choice for the estimator $\hat{\beta}$ 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). Our simulation study explores the finite sample differences between these two estimators.

It follows from Choi, Hall and Schick (1996) that for non-singular information matrices tests based on $\hat{S}_{\bar{\gamma}}$ are asymptotically uniformly most powerful within the class of rotation invariant tests. This implies that asymptotically when testing the hypothesis $H_0 : \alpha = \alpha_0$, the power of the test is the greatest possible in the class of rotationally invariant tests. This makes tests based on $\hat{S}_{\bar{\gamma}}$ 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). Uniformity results and minimax optimality results which permit singular information matrices can be found in Lee (2022) for efficient score tests in general semi-parametric models.

5 Simulation results

In this section we study the finite sample properties of the singularity and identification robust score test $\hat{S}_{\hat{\gamma}}$. We study the size and power of the test under different data generating processes and compare its performance to several alternatives that have been proposed in the literature. We first study the simple model of section (2) after which we consider the general linear simultaneous equations model (3). The supplementary material provides additional results.

5.1 Baseline model

We start by drawing independent samples from model (1), which we restate for convenience

$$Y_i = R'\epsilon_i \ , \qquad i = 1, \ldots, n \ .$$

We take Y_i to be $K \times 1$ and consider $K = 2, 3$ and $K = 5$. The sample size is taken as $n = 200, 500$ or $n = 1000$. We fix $\epsilon_{i,1}$ to have a standard Gaussian density and consider different densities for $\epsilon_{i,k}$, with $k = 2, \ldots, K$. The non-Gaussian densities are either Student's t or mixtures of normals taken from Marron and Wand (1992). Figure 3 provides an overview.

The matrix of interest $R = R(\alpha)$ is orthogonal and parametrized by the Cayley transformation of a skew-symmetric matrix (e.g. Gouriéroux, Monfort and Renne, 2017):

$$R(\alpha) = (I - \Omega(\alpha))(I + \Omega(\alpha))^{-1} ,$$

where $\Omega(\alpha)$ is a skew-symmetric matrix (i.e. $\Omega(\alpha)' = -\Omega(\alpha)$) parameterized by α which we sample at random from $\alpha \sim N(0, I_{L_\alpha})$.

In this setting there are no additional nuisance parameters which allows us to concentrate on the consequences of weak non-Gaussianity on the efficient score test and some alternative tests that have been proposed in the literature. In the simulation designs below we include additional finite dimensional nuisance parameters (i.e. $\beta = (\sigma, b)$) and investigate whether their inclusion alters the size and power of the test.

For each specification we simulate $S = 5,000$ datasets and for each we compute the efficient score statistic $\hat{S}_{\hat{\gamma}}$ as defined in equation (10) following the Algorithm given in Section 3.[12] We implement the log density score estimator (7) using $B = 4, 6$ or 8 cubic splines.

In Table 2 we show the empirical rejection frequencies corresponding to the $S_{\hat{\gamma}}$ 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 $\alpha \in \mathbb{R}^{L_\alpha}$, i.e. 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 model K and the number of cubic splines B.

Second, when the densities for $k \geq 2$ are non-Gaussian the size remains correct. Specifically, columns 2-4 show the results for the case where $\epsilon_{i,k}$ follows a Student's t distribution with decreasing degrees of freedom ($\nu = 15, 10, 5$). No matter how close we get to the Gaussian density the size remains correct. Columns 5-10 show similarly correct size for a variety of mixture distributions. Even for complicated skewed bi-modal densities (e.g. columns 8-10) the $S_{\hat{\gamma}}$ 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.

Overall, the asymptotic approximation in Theorem 1 seems to provide a good approximation for the finite sample behavior of the semiparametric score test, at least for the

[12] To be specific, since the model does not contain any finite dimensional nuisance parameters step 1 in the algorithm can be skipped and the score statistic is simply evaluated at α_0.

distributions shown in Figure 3.

5.2 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. We concentrate on evaluating different tests based on size and power in the vicinity of Gaussianity.[13]

Alternative tests

Conceptually, there are two types of alternative tests that we consider: (i) tests that rely on estimates for α and (ii) tests that fix $\alpha = \alpha_0$ under the null. Clearly, from our intuitive discussion in Section 2 it follows that we expect tests that fix α under the null to perform relatively well.

In category (i) we consider the standard maximum likelihood Wald (W^{mle}) and likelihood ratio (LR^{mle}) tests based on the Student's t density for ϵ_k. For densities 2-4 in Figure 3 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. For all other densities these tests are mis-specified.

In addition, we consider the psuedo-maximum likelihood Wald test (W^{pmle}) from Gouriéroux, Monfort and Renne (2017). This test is asymptotically valid for a broader range of true distribution functions and amount to fixing the functional form of the densities $\eta_1, \ldots, \eta_K$. We follow the implementation of Gouriéroux, Monfort and Renne (2017) and choose the Students t density with five degrees of freedom as the pseudo-likelihood and compute the Wald statistic based on this density.

Finally, we consider the recently developed GMM method of Lanne and Luoto (2021), which relies on higher order moments to identify the parameters α. We use $\mathbb{E}\epsilon_{i,k}^2\epsilon_{i,j} = 0$, $\mathbb{E}\epsilon_{i,k}^3\epsilon_{i,j} = 0$ 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 (LR^{gmm}), see Lanne and Luoto (2021) for details.[14]

In category (ii) we consider tests which fix $\alpha = \alpha_0$ under the null. Specifically, we include the standard LM test (LM^{mle}) based on the Student's t density where the degrees of freedom parameter is estimated from the data. Second, we consider the pseudo-maximum likelihood

[13] The recent simulation studies of Herwartz, Lange and Maxand (2019) and Moneta and Pallante (2020) provide further simulation evidence for existing methods, also focusing on estimation accuracy.

[14] Note that lower order moments are not required as the baseline model $Y_i = R'\epsilon_i$ implies that the observations have mean zero and unit variance.

version of the LM test (LM^{pmle}) based on Gouriéroux, Monfort and Renne (2017), which fixes the degrees of freedom at five. Finally, we consider the GMM-based identification robust S-statistic (S^{gmm}) of Stock and Wright (2000), which was recently considered in Drautzburg and Wright (2021) in the context of structural VAR models with non-Gaussian errors. We use the same moment conditions as considered in Drautzburg and Wright (2021) for the LM^{gmm} test.

Size comparison

We compare the size of the different tests for the simulation designs described in Section 5.1. The empirical rejection frequencies are shown in Table 3 for the case where $K = 2$ and $n = 200, 500, 1000$. Overall we find, perhaps not surprisingly, that all tests in category (i) do not have correct size when the true density is close to Gaussian nor when the corresponding method is based on a mis-specified model. This shows that tests based on estimates for α are generally unreliable. Second, tests in category (ii) overall control the size of the test well.

More specifically, we find that the Wald tests (W^{mle} and W^{pmle}) tend to over-reject quite severely whilst the standard likelihood ratio test (LR^{mle}) tends to be undersized for most densities, especially in the vicinity of the Gaussian density, as ought to be expected given the earlier evidence in shown in Figure 1. Finally, the GMM likelihood ratio test (LR^{gmm}) is also over-sized, which confirms findings in Lanne and Luoto (2021) where the LR^{gmm} also over-rejects when the densities of the structural shocks are close to Gaussian.

In the second category the semi-parametric score test $\hat{S}_{\hat{\gamma}}$ (as proposed in this paper) and the pseudo maximum likelihood LM test (LM^{pmle}), inspired by Gouriéroux, Monfort and Renne (2017), both have near perfect size across all densities. The standard LM test (LM^{mle}) also performs reasonably well, but when the functional form of the true densities is very different from the Student's t density (e.g. separate bi-modal, column 9) the test tends to under-reject.[15] Finally, the GMM based robust S test (S^{gmm}) tends to be over-sized for small samples, but for large samples it generally shows correct size except for densities with moderately heavy tails such as the $t(5)$ density (column 4). In these cases the S^{gmm} is over-sized which can be understood when realizing that the GMM approach requires eight finite moments for inference when based on fourth-order moment restrictions. The $t(5)$ density does not have eight finite moments.

In sum, we recommend avoiding statistics that are based on estimates for α as these are overall unreliable when the shock distributions are close to Gaussian. All tests that fix α under the null perform at least reasonably well. In the next section we compare these tests based on their finite sample power.

[15] Recall here that this test is based on a misspecified density.

Power comparison

We compare the power of all tests that fix α under the null, that is $\hat{S}_\gamma$, LM^{mle}, LM^{pmle} and S^{gmm}.

We consider the case where $K = 2$ and $n = 1000$.[16] In this setting α is a scalar parameter and we fixed the true value at 0 (an arbitrary choice). Figure 4 shows the empirical rejection frequencies when we vary α around $\alpha = 0$. Each point on the curve is based on $S = 5,000$ simulations.

Two main findings stand out. First, for the Student's t densities $t(15)$, $t(10)$ and $t(5)$ (panels 2-4) the standard LM test (LM^{mle}) shows the highest power. This is not surprising as for these data generating processes the LM^{mle} test is correctly specified and hence takes advantage of fitting the true densities using only a scalar parameter. That said, the semi-parametric score test ($\hat{S}_{\hat{\gamma}}$) and the pseudo maximum likelihood LM test (LM^{pmle}) come reasonably close in terms of power.

Second, for all other densities, i.e. different mixtures of normals in panels 5 – 10, the semi-parametric score test ($\hat{S}_{\hat{\gamma}}$) shows the highest power. Sometimes the difference with the other tests is not very large, but for instance for bi-modal densities (panels 8-10) the differences are substantial. Overall, the good power of the $\hat{S}_{\hat{\gamma}}$ test corresponds to the theoretical finding that for non-singular information matrices the test is asymptotically uniformly most powerful in the class of unbiased tests.[17]

Besides the $\hat{S}_{\hat{\gamma}}$ test, we note that the pseudo maximum likelihood LM test and the GMM based S test shows quite promising power for most of the densities considered. None of these dominates the other. The caveat for the GMM test is that it is size-distorted for moderately heavy tails (panel 4).

5.3 Linear simultaneous equations model

Next, we discuss the simulation results for the general linear simultaneous equations model (3). 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 $A(\alpha, \sigma)^{-1} = \Sigma^{1/2}(\beta_1)R(\alpha)$ as in example 1, where $\Sigma^{1/2}$ is lower triangular and the rotation matrix R remains to be specified by the Cayley transform. The explanatory variables are drawn from the standard normal distribution.

The vector of finite dimensional nuisance parameters β now includes $\sigma = \text{vech}(\Sigma^{1/2})$ and $b = \text{vec}(B)$. Our main theoretical result in Theorem 1 shows that β can be approximated by

[16] Power comparisons for different n can be found in the supplementary material.

[17] Cf. Choi, Hall and Schick (1996).

any $\sqrt{n}$-consistent estimate. Obviously, ordinary least squares estimates are attractive for their simplicity, but given the non-normality of the structural shocks these estimates may be improved. Therefore we also consider estimating β by one-step-efficient estimates (e.g. van der Vaart, 2002, Section 7.2), which are easy to compute here since the efficient score of β is computed anyway to construct the score test.

Similar, as before the first error $\epsilon_{i,1}$ follows a Gaussian distribution and the different densities from Figure 3 are assigned to the other error terms. For each specification we simulate $S = 5,000$ datasets and for each sample we compute the semi-parametric score statistic using the Algorithm in Section 3.

Size results

The empirical rejection frequencies are shown in Tables 4 and 5 for the OLS and one-step efficient estimates for β, respectively.

We find that for all the rejection frequencies of the $\hat{S}_{\hat{\gamma}}$ test are generally close to the nominal size. That said, there is more variation in the empirical rejection frequencies compared to Table 2, indicating that the estimation of the finite dimensional nuisance parameters does have consequences.

Starting with Table 4 where $\hat{\beta}$ is estimated by OLS. We find that the size of $\hat{S}_{\hat{\gamma}}$ is the same regardless of how close the densities of $\epsilon_{i,k}$ are to the Gaussian density. Specifically, moving from columns 1-4 (i.e. from Gaussian to $t(5)$) we see virtually no changes in the rejection frequencies. This holds for all specifications considered and highlights the main point of this paper: the semi-parametric score test yields reliable inference even when α is not, or poorly, identified.

Depending on the dimension of β we do find size distortions for small sample sizes, most notably when $K = 5$ and $n = 200$. In this setting β is of dimension 20 or 25 depending on $d = 2, 3$, and we see that the test is often over-sized. This does not hold for all densities considered, but for Gaussian, Student's t and kurtotic unimodal densities the test over-rejects. When n increases the over-rejection vanishes and the test appears correctly sized.

For the one-step efficient estimator for β the results are shown in Table 5. We find that on average the empirical rejection frequencies are larger when compared to the OLS estimator. Notably, when n is small over-rejection becomes more severe. Again, we find that this holds uniformly across densities, i.e. distortions do no depend on being close to Gaussian, and the sizes improve when n increases.

Power results

Next, we investigate the power of the $\hat{S}_{\hat{\gamma}}$ test for the LSEM model. We again consider the case where $K = 2$, $d = 2$ and $n = 1000$, which allows us to compare the results with those for the baseline model. The power curves are shown in Figure 5 for both OLS and one-step estimates for β.

First, when comparing Figure 5 to the case without nuisance parameters (i.e. Figure 4) we find that the power of the test is reduced when we include nuisance parameters. Second, the power of the test using the one-step efficient estimates (dotted blue line) is higher when compared to the same test evaluated at OLS estimates. This holds for all densities considered.

Based on these results we recommend using OLS estimates for β when the sample size is small (e.g. $n = 200, 500$), but for larger sample sizes the one-step efficient estimates are preferable.

6 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. Fittingly, 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

$$O_i = e^{c_1} L_i^{\alpha_1} K_i^{\alpha_2} e^{\epsilon_{i,1}} \ ,$$

where O_i, L_i, K_i denote output, labor and capital, respectively, and $\epsilon_{i,1}$ 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_i, K_i are typically choice variables of the firm. Allocations are made to maximize profits and hence will generally depend on unobservables $\epsilon_{i,1}$.

To address this simultaneity problem we consider a simultaneous equations approach that allows for correlation among L_i, K_i, ϵ_1, and exploits possible non-Gaussianity in the errors to identify the parameters α_1 and α_2.

To be specific, the models that we consider are defined for $Y_i = (\log O_i, \log L_i, \log K_i)'$, and are of the form

$$S(\alpha, \sigma)Y_i = BX_i + D(\sigma)\epsilon_i \ , \tag{13}$$

where X_i includes a constant and any other additional exogenous variables such as the age of the firm. We adopt the following specification for the matrices S and D.

$$S(\alpha, \sigma) = \begin{bmatrix} 1 & -\alpha_1 & -\alpha_2 \\ -\sigma_1 & 1 & -\alpha_3 \\ -\sigma_2 & -\sigma_3 & 1 \end{bmatrix} \qquad \text{and} \qquad D(\sigma) = \begin{bmatrix} \sigma_4 & 0 & 0 \\ 0 & \sigma_5 & 0 \\ 0 & 0 & \sigma_6 \end{bmatrix} .$$

We note that parameters in σ can be recovered from the variance of $Y_i - BX_i$ and we will 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.[18] Finally, to pin down the desired rotation we impose that α_1 and α_2 are positive and the correlations between L_i, K_i and $\epsilon_{i,1}$ 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.[19] We perform two 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.

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 (13) 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 between 0.27 and 0.50, for all choices of the control variables. The joint confidence region

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

[19] The data are obtained from CompuStat.

for (α_1, α_2) is shown in the top left panel of Figure 6. 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 6 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 for the first and second errors using a Jarque-Bera test. For the third error we cannot reject normality.

Given our simulation results such mild deviations from Gaussianity may cause problems for standard inference methods. This is true for the alternative methods that which were found not robust to weak deviations from Gaussianity; they tend to give much smaller confidence bands. This suggests that whilst non-Gaussianity may be a useful tool for identification, 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 or instruments are needed.

Table 6 also reports 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. This highlights that there may indeed be endogeneity in the form of correlation between labor, capital and the error term $\epsilon_{i,1}$.

To verify whether this conclusion is justified we need to test whether the underlying assumption regarding the independence of the underlying structural shocks is indeed true (e.g. Montiel Olea, Plagborg-Møller and Qian, 2022). To do so, we adopt the permutation test for independent components as proposed in Matteson and Tsay (2017). We implement their test on the sample $\{\hat{\epsilon}_i\}$ as defined above.[20] The results are shown in the bottom row of Table 6. Depending on whether age is included as a control variable, the p-values are 0.12 and 0.16 indicating that there is not substantial evidence against independence.

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 7. 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.

[20] The test was implemented using the R package steadyICA using the function permTest.

7 Conclusion

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

To remedy this problem we proposed an identification robust semi-parametric score statistic for testing hypotheses in LSEMs. Under mild regularity conditions we showed that the score test retains correct asymptotic size regardless of the shape of the true density functions. A simulation study shows that our asymptotic theory provides an accurate approximation to the finite sample performance of our test.

While we have restricted our treatment to models where the observations were independently distributed across entities, we note that a similar approach may be considered for dynamic models, but this will require extending our results to allow for non-i.i.d. data. Similarly, dynamic panel data models could 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).

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

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

Andrews, D. W. K., and X. Cheng. 2012. "Estimation and inference with weak, semi-strong and strong identification." *Econometrica*, 80(5).

Andrews, D. W. K., and X. Cheng. 2013. "Maximum likelihood estimation and uniform inference with sporadic identification failure." *Journal of Econometrics*, 173.

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

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.

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., Yaacov 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.

Davis, Richard, and Leon Fernandes. 2022. "Independent Component Analysis with Heavy Tails using Distance Covariance." Working paper.

Davis, Richard, and Serena Ng. 2022. "Time Series Estimation of the Dynamic Effects of Disaster-Type Shocks." Working paper.

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.

Drautzburg, Thorsten, and Jonathan H Wright. 2021. "Refining Set-Identification in VARs through Independence." National Bureau of Economic Research Working Paper 29316.

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

Fiorentini, Gabriele, and Enrique Sentana. 2022. "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.

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.

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.

Herwartz, Helmut, Alexander Lange, and Simone Maxand. 2019. "Statistical Identification in Svars - Monte Carlo Experiments and a Comparative Assessment of the Role of Economic Uncertainties for the US Business Cycle." CEGE Discussion Paper 375.

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.

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).

Jin, Ze, Benjamin B. Risk, and David S. Matteson. 2019. "Optimization and testing in linear non-Gaussian component analysis." *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(3): 141–156.

Kaido, Hiroaki, Francesca Molinari, and Jrg 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).

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. 2021. "GMM Estimation of Non-Gaussian Structural Vector Autoregression." *Journal of Business & Economic Statistics*, 39(1): 69–81.

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. 1960. *Locally Asymptotically Normal Families of Distributions: Certain Approximations to Families of Distributions and Their Use in the Theory of Estimation and Testing Hypotheses. University of California Berkeley, Calif: University of California publications in statistics*, University of California Press.

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

Lee, Adam. 2022. "Robust and Efficient Inference for Non-Regular Semiparametric Models." Working paper.

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

Lütkepohl, Helmut, and Maike M. Burda. 1997. "Modified Wald tests under nonregular conditions." *Journal of Econometrics*, 78(2): 315–332.

Magnus, Jan R., Henk G.J. Pijls, and Enrique Sentana. 2021. "The Jacobian of the exponential function." *Journal of Economic Dynamics and Control*, 127: 104122.

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.

Matteson, David S., and Ruey S. Tsay. 2017. "Independent Component Analysis via Distance Covariance." *Journal of the American Statistical Association*, 112(518): 623–637.

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

Moneta, Alessio, and Gianluca Pallante. 2020. "Identification of Structural VAR Models via Independent Component Analysis: A Performance Evaluation Study." Laboratory of Economics and Management (LEM), Sant'Anna School of Advanced Studies, Pisa, Italy LEM Papers Series 2020/24.

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.

Montiel Olea, José Luis, Mikkel Plagborg-Møller, and Eric Qian. 2022. "SVAR Identification from Higher Moments: Has the Simultaneous Causality Problem Been Solved?" *AEA Papers and Proceedings*, 112: 481–85.

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

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

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

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

Risk, Benjamin B., David S. Matteson, and David Ruppert. 2019. "Linear Non-Gaussian Component Analysis Via Maximum Likelihood." *Journal of the American Statistical Association*, 114(525): 332–343.

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.

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. 2022. "Identification and Estimation of Structural VARMA Models Using Higher Order Dynamics." *Journal of Business & Economic Statistics.* forthcoming.

Appendix

In this appendix we provide the proof for Theorem 1. The proof is structured as follows. We first provide a general approach for conducting identification and singularity robust hypothesis tests in semiparametric models. This general theory is subsequently applied to prove Theorem 1.

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.

A: General theory

We expound a general approach for conducting identification robust hypothesis tests in semiparametric models. The LSEM model of Section 3 constitutes as a special case of the model considered in this section.

Let $Y \in \mathcal{Y} \subset \mathbb{R}^K$ by a random vector 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

$$\mathcal{P}_\Theta = \{P_\theta : \theta \in \Theta\} \ , \tag{14}$$

where each $P_\theta \ll \mu$ for some σ-finite measure μ on $\mathcal{Y}$. 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

$$H_0 : \alpha = \alpha_0 \ , \ \beta \in \mathcal{B} \ , \ \eta \in \mathcal{H} \qquad \text{against} \qquad H_1 : \alpha \neq \alpha_0 \ , \ \beta \in \mathcal{B} \ , \ \eta \in \mathcal{H} \ . \tag{15}$$

To derive our tests, we first define the scores of model (14) following the definition in van der Vaart (2002).

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

$$\int \left[\frac{\sqrt{p_t} - \sqrt{p}}{t} - \frac{1}{2} s \sqrt{p} \right]^2 \mathrm{d}\mu \to 0 \ , \tag{16}$$

where p_t and p respectively denote the densities of P_t and P relative to μ. The map $t \to \sqrt{p_t}$ is the root density path and s is the ***score function*** *of the submodel $\{P_t : t \geq 0\}$ at $t = 0$.*

In words we say that a differentiable path is a parametric submodel $\{P_t : 0 \leq t < \epsilon\}$ that is differentiable in quadratic mean at $t = 0$ with score function s. If we let $t \mapsto P_t$ range over a collection of submodels, indexed by $\mathcal{V}$, we will obtain a collection of score functions, say s_i for $i \in \mathcal{V}$. This collection, $\{s_i : i \in \mathcal{V}\}$, will be denoted by $\mathcal{T}_{P,\mathcal{V}}$ and as we only consider models with linear spaces we refer to it as a *tangent space*. For the semiparametric model (14) we define tangent spaces along restricted paths concerning the two parts of the parameter $\theta = (\gamma, \eta)$ separately.

Assumption 3. *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*

$$\mathcal{T}_{P_\theta,\mathcal{J}} = \mathcal{T}^{\gamma|\eta}_{P_\theta,\mathbb{R}^L} + \mathcal{T}^{\eta|\gamma}_{P_\theta,H} \ , \tag{17}$$

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 (14) 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.[21]

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 $\dot{\ell}_\theta$, i.e. $\mathcal{T}^{\gamma|\eta}_{P_\theta,\mathbb{R}^L} = \{g'\dot{\ell}_\theta : g \in \mathbb{R}^L\}$. 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)

$$\tilde{\ell}_\theta \coloneqq \dot{\ell}_\theta - \Pi_\theta \dot{\ell}_\theta \ , \tag{18}$$

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

$$\tilde{I}_\theta \coloneqq \mathbb{E}_\theta \tilde{\ell}_\theta \tilde{\ell}'_\theta \ . \tag{19}$$

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).

To obtain the efficient score function for α which is the part of $\gamma = (\alpha, \beta)$ that is of

[21] 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).

interest, note that the previous two displays imply the partitioning

$$\tilde{\ell}_\theta = \left(\tilde{\ell}'_{\theta,\alpha}, \tilde{\ell}'_{\theta,\beta}\right)' \quad \text{and} \quad \tilde{I}_\theta = \begin{bmatrix} \tilde{I}_{\theta,\alpha\alpha} & \tilde{I}_{\theta,\alpha\beta} \\ \tilde{I}_{\theta,\beta\alpha} & \tilde{I}_{\theta,\beta\beta} \end{bmatrix} . \tag{20}$$

If $\tilde{I}_{\theta,\beta\beta}$ is nonsingular,[22] we can (orthogonally) project once more to obtain the efficient score function for α:

$$\tilde{\kappa}_\theta \coloneqq \tilde{\ell}_{\theta,\alpha} - \tilde{I}_{\theta,\alpha\beta}\tilde{I}^{-1}_{\theta,\beta\beta}\tilde{\ell}_{\theta,\beta} , \tag{21}$$

which has corresponding efficient information matrix

$$\tilde{\mathcal{I}}_\theta \coloneqq \tilde{I}_{\theta,\alpha\alpha} - \tilde{I}_{\theta,\alpha\beta}\tilde{I}^{-1}_{\theta,\beta\beta}\tilde{I}_{\theta,\beta\alpha} . \tag{22}$$

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).

It follows from (18) and Lemma 1.7 in van der Vaart (2002) that at $\theta_0 = (\alpha_0, \beta, \eta)$, where $\beta \in \mathcal{B}$ and $\eta \in \mathcal{H}$, we have

$$\mathbb{E}_{\theta_0}\tilde{\kappa}_{\theta_0} = 0 . \tag{23}$$

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 4. *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 $\{\hat{\ell}_{\gamma_n}(Y_i)\}_{n\geq 1, i\leq n}$ such that:*

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

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

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

4. *We have that*

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

[22] 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.

We note that the estimates for the efficient scores $\hat{\ell}_{\gamma_n}(Y_i)$ and information matrix $\hat{I}_{\gamma_n}$ no longer depend on η, hence they are only indexed by γ_n. Based on Assumption 4-parts 2 and 3 we define the following estimators for the efficient score and information matrix for α:

$$\hat{\kappa}_\gamma \coloneqq \hat{\ell}_{\gamma,\alpha} - \hat{I}_{\gamma,\alpha\beta}\hat{I}^{-1}_{\gamma,\beta\beta}\hat{\ell}_{\gamma,\beta} \ , \qquad \text{and} \qquad \hat{\mathcal{I}}_\gamma \coloneqq \hat{I}_{\gamma,\alpha\alpha} - \hat{I}_{\gamma,\alpha\beta}\hat{I}^{-1}_{\gamma,\beta\beta}\hat{I}_{\gamma,\beta\alpha} \ . \tag{24}$$

Given ν_n from Assumption 4-part 3, we define a truncated eigenvalue version of the information matrix estimate as

$$\hat{\mathcal{I}}^t_\gamma = \hat{U}_n\hat{\Lambda}_n(\nu_n)\hat{U}'_n \ , \tag{25}$$

where $\hat{\Lambda}_n(\nu_n)$ is a diagonal matrix with the ν_n-truncated eigenvalues of $\hat{\mathcal{I}}_\gamma$ 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 $\hat{\mathcal{I}}_\gamma$, 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 a function of $\gamma = (\alpha, \beta)$ as follows.

$$\hat{S}_\gamma \coloneqq \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\kappa}_\gamma(Y_i)\right)' \hat{\mathcal{I}}^{t,\dagger}_\gamma \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\kappa}_\gamma(Y_i)\right). \tag{26}$$

where $\hat{\mathcal{I}}^{t,\dagger}_\gamma$ is the Moore-Penrose psuedo-inverse of $\hat{\mathcal{I}}^t_\gamma$. The following theorem implies that we can use the estimated rank of $\hat{\mathcal{I}}^t_\gamma$ to compute the critical value for $\hat{S}_\gamma$.

Theorem 2. *Let $\gamma_0 = (\alpha_0, \beta)$ for any $\beta \in \mathcal{B}$. 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 3 and 4 hold and let $\bar{\gamma}_n = (\alpha_0, \bar{\beta}_n)$. Let $r_n = \text{rank}(\hat{\mathcal{I}}^t_{\bar{\gamma}_n})$ and denote by c_n the $1-a$ quantile of the $\chi^2_{r_n}$ distribution for any $a \in (0,1)$.*[23] *Then*

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

with inequality only if $\text{rank}(\tilde{\mathcal{I}}_{\gamma_0}) = 0$.

The proof for Theorem 2 is given below. This theorem provides the main building block for the proof of Theorem 1 for the LSEM model.

B: Proof of Theorem 1

We note that the LSEM model (3) can be viewed as a semi-parametric model defined by

$$\mathcal{P}_\Theta \coloneqq \{P_\theta : \theta \in \Theta\} \tag{27}$$

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

where $\Theta = \mathcal{A} \times \mathcal{B} \times \mathcal{H}$, with $\mathcal{A} \subset \mathbb{R}^{L_\alpha}$, $\mathcal{B} \subset \mathbb{R}^{L_\beta}$ and $\mathcal{H} = \mathscr{Z} \times \prod_{k=1}^{K} \mathscr{H}$, where $\mathscr{Z}$ is the space of density functions η_0 with $\tilde{X}_i \sim \eta_0$ and $\mathscr{H}$ is the space of density functions η_k, i.e.

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

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 $\kappa(z) = z^2 - 1$. We denote by $\mathcal{H}_0 \subset \mathcal{H}$ the set with elements $\eta = (\eta_0, \ldots, \eta_K)$ such that each η_k satisfies the requirements imposed by assumption 1. Finally, P_θ is the law on $\mathcal{Y} \times \mathcal{X}$, with $Y_i \in \mathcal{Y} \subset \mathbb{R}^K$ and $\tilde{X}_i \in \mathcal{X} \subset \mathbb{R}^{d-1}$, defined by the density

$$p_\theta(y, \tilde{x}) := |\det A| \prod_{k=1}^{K} \eta_k(A_{k\bullet} y) \times \eta_0(\tilde{x}) \ , \tag{28}$$

where $A_{k\bullet}$ denotes the kth row of $A = A(\alpha, \sigma)$.

With these formalities established we give three useful lemmas whose proofs are deferred to the web-appendix. The first lemma defines the tangent spaces for the LSEM and effectively ensures that the LSEM model satisfies the high-level assumption 3 in the general theory.

Lemma 1. *Given Assumption 1, if $(\alpha, \sigma) \mapsto A(\alpha, \sigma)$ is continuously differentiable, we have that for any $\theta \in \Theta$ there is a $\delta > 0$ small enough such that the path $t \mapsto P_{\theta_t(\theta,g,h)}$ from $[0, \delta)$ to (a subset of) $\mathcal{P}_\Theta$ is a differentiable path with score function $y \mapsto g' \dot{\ell}_\theta(y, \tilde{x}) + h_0(\tilde{x}) + \sum_{k=1}^{K} h_k(A_{k\bullet} v)$, where $v = y - Bx$. In particular,*

$$\mathcal{T}_{P_\theta, \mathcal{J}} = \left\{ y \mapsto g' \dot{\ell}_\theta(y, \tilde{x}) + h_0(\tilde{x}) + \sum_{k=1}^{K} h_k(A_{k\bullet} v) : g \in \mathbb{R}^L, h \in H \right\} = \mathcal{T}^{\gamma|\eta}_{P_\theta, \mathbb{R}^L} + \mathcal{T}^{\eta|\gamma}_{P_\theta, H},$$

and $\mathcal{T}_{P_\theta, \mathcal{J}}$ is a tangent space to the model at P_θ.

The next lemma presents the efficient score functions (18) for the LSEM model.

Lemma 2. *Given Assumption 1, if $(\alpha, \sigma) \mapsto A(\alpha, \sigma)$ is continuously differentiable, the components of the efficient score function $\tilde{\ell}_\theta$ for the semiparametric linear simultaneous equations model $\mathcal{P}_\Theta$ in (27) at any $\theta = (\gamma, \eta)$ with $\gamma = (\alpha, \beta)$, $\alpha \in \mathcal{A}$, $\beta = (\sigma, b) \in \mathcal{B}$ and*

$\eta \in \mathcal{H}_0$ *are given by*

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

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

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

with $v = y - Bx$, $x = (1,\tilde{x}')'$, $\zeta^{\alpha}_{l,k,j} := [D_{\alpha,l}]_{k\bullet}A^{-1}_{\bullet j}$, $\zeta^{\sigma}_{l,k,j} := [D_{\sigma,l}]_{k\bullet}A^{-1}_{\bullet j}$, $D_{\alpha,l} = \partial A(\alpha,\sigma)/\partial\alpha_l$, $D_{\sigma,l} = \partial A(\alpha,\sigma)/\partial\sigma_l$ *and* $D_{b,l} = \partial B/\partial b_l$. *Further,*

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

The proof of this lemma follows from Amari and Cardoso (1997) for $\tilde{\ell}_{\theta,\alpha_l}(y,\tilde{x})$ and $\tilde{\ell}_{\theta,\sigma_l}(y,\tilde{x})$, and for $\tilde{\ell}_{\theta,b_l}(y,\tilde{x})$ the derivations are similar to those found in, for example, Bickel et al. (1998) or Newey (1990).

The final lemma summarizes which conditions a log density score estimator should satisfy. We will apply this lemma for different choices of $W_{i,n}$ to verify our main result.

Lemma 3. *Given assumptions 1 and 2, let* $\{\beta_n\}_{n\geq 1}$ *be any deterministic sequence in* $\mathcal{B}$ *with* $\sqrt{n}(\beta_n - \beta) = O(1)$ *and let* $\theta_n = (\alpha_0, \beta_n, \eta)$ *for some* $\eta \in \mathcal{H}_0$. *The log density score estimates* $\hat{\phi}_k$ *defined in* (7) *satisfy*

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_k(A_{n,k\bullet}(Y_i - B_nX_i)) - \phi_k(A_{n,k\bullet}(Y_i - B_nX_i))\right]W_{i,n} = o_{P_{\theta_n}}(n^{-1/2}), \tag{29}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\left(\left[\hat{\phi}_k(A_{n,k\bullet}(Y_i - B_nX_i)) - \phi_k(A_{n,k\bullet}(Y_i - B_nX_i))\right]W_{i,n}\right)^2 = o_{P_{\theta_n}}(\nu_n). \tag{30}$$

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* $A_{n,k}Y_j$.

Proof of Theorem 1. We verify assumptions 3 and 4 for the LSEM under Assumptions 1 and 2.

First, the technical assumption 3 is verified in Lemma 1, as given above. Next, we verify each part of Assumption 4 separately. First, we note that assumption 4-part 1 follows by the CLT since our data is iid and the efficient score $\tilde{\ell}_{\theta_0}$ as derived in Lemma 2 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 $A_{n,k}(Y_i - B_nX_i) \simeq$

$\epsilon_{i,k} \sim \eta_k$ where $A_n = A(\alpha_0, \sigma_n)$ and $A_{n,k}$ denotes the kth row of A_n. Hence we can compute certain properties of the efficient score using the equality in distribution:

$$\tilde{\ell}_{\theta_n,\alpha_l}(Y_i, \tilde{X}_i) \simeq \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \zeta^{\alpha}_{l,k,j,n} \phi_k(\epsilon_{i,k})\epsilon_{i,j} + \sum_{k=1}^{K} \zeta^{\alpha}_{l,k,k,n} \left[\tau_{k,1}\epsilon_{i,k} + \tau_{k,2}\kappa(\epsilon_{i,k})\right] \tag{31}$$

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

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

where we note that the same observation implies that $\tau_{k,n} = \tau_k$ and $\varsigma_{k,n} = \varsigma_k$ for each n.[24] By our assumptions on the map $(\alpha, \sigma) \mapsto A(\alpha, \sigma)$, we have $\zeta^{\alpha}_{l,k,j,n} \to \zeta^{\alpha}_{l,k,j,\infty} := [D_{\alpha,l}(\gamma_0)]_{k\bullet} A(\gamma_0)^{-1}_{\bullet j}$ and $\zeta^{\sigma}_{l,k,j,n} \to \zeta^{\alpha}_{l,k,j,\infty} := [D_{\sigma,l}(\gamma_0)]_{k\bullet} A(\gamma_0)^{-1}_{\bullet j}$ 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 4-part 2 for each component of the efficient score (31)-(33), but we note that (31) and (32) are identical hence we concentrate on (31). For (31) and $v_n = y - B_n x$, we define

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

and

$$\hat{\varphi}_{1,n}(v_n) := \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \zeta^{\alpha}_{l,k,j,n} \hat{\phi}_k(A_{n,k\bullet} v_n) A_{n,j\bullet} v_n \ ,$$

Let $\overline{\zeta}^{\alpha}_n := \max_{l\in[L], j\in[K], k\in[K]} |\zeta^{\alpha}_{l,j,k,n}|$ which converges to $\overline{\zeta}^{\alpha} := \max_{l\in[L], j\in[K], k\in[K]} |\zeta^{\alpha}_{l,j,k,\infty}| < \infty$. We have that

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

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 Lemma 3-part (29) 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 $\mathbb{E}_{\theta_n} V^2_{i,j,n} = 1$ by Assumption 1) 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 $\hat{\tau}_{k,n}$ and $\hat{\varsigma}_{k,n}$ are defined in (6).

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

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

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

Since sum is finite and each $|\zeta^{\alpha}_{l,k,k,n}| \to |\zeta^{\alpha}_{l,k,k,\infty}| < \infty$ 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 (31) , since $\tilde{\ell}_{\theta_n,\alpha_l} = \varphi_{1,n} + \varphi_{2,\tau,n}$ and $\hat{\ell}_{\gamma_n,\alpha_l} = \hat{\varphi}_{1,n} + \varphi_{2,\hat{\tau}_n,n}$. The same holds for (32); the only difference is that we replace $\zeta^{\alpha}_{l,k,k,n}$ by $\zeta^{\sigma}_{l,k,k,n}$

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

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

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 Lemma 3-part (29) 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, $\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\gamma_n,b,l} - \tilde{\ell}_{\theta_n,b,l}\right] = o_{P_{\theta_n}}(1)$, is satisfied.

To verify part 3 we will show that

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

where $\tilde{I}_{\theta_n} := \frac{1}{n}\sum_{i=1}^n \tilde{\ell}_{\theta_n}(Y_i)\tilde{\ell}_{\theta_n}(Y_i)'$. 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 $r, s \in \{\alpha, \sigma, b\}$ 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 $n^{-1/2} = o(\nu_n^{1/2})$.[25] For the first term, if $p = 2$ in lemma 6, by Theorem 2.5.11 in Durrett (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 $r \in \{\alpha, \sigma, b\}$ and let l denote an index, we write $\hat{U}_{n,i,r_l} := \hat{\ell}_{\gamma_n,r_l}(Y_i)$, $\tilde{U}_{i,r_l} := \tilde{\ell}_{\theta_n,r_l}(Y_i)$ and $D_{n,i,r_l} := \hat{\ell}_{\gamma_n,r_l}(Y_i) - \tilde{\ell}_{\theta_n,r_l}(Y_i)$.

Since it is the absolute value of the $(r, l) - (s, m)$ component of $\hat{I}_{\gamma_n,n} - \tilde{I}_{\theta_0,n}$, 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

[25] 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.

$r, s \in \{(\alpha, \sigma), b\}$ and l, m. By Cauchy-Schwarz and lemma 8

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

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

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

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

It remains to show that part 4 of Assumption 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 $r \in \{\alpha, \sigma, b\}$ corresponding to (31)-(33) for some arbitrary l. By the expressions for $\tilde{\ell}_{\theta_n}$ and p_{θ_n} given in lemma 2 and equation (28) 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 (31) we have

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

where G is the law of ϵ and each of the integrals are finite by assumption 14. By the

continuity of A and D_l, this converges to

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

which is finite by assumption 1. By Proposition 2.29 in van der Vaart (1998) we conclude that $\int (f_{\alpha,n} - f_\alpha)^2 \,\mathrm{d}\lambda \to 0$. Analogous arguments hold for $r = \sigma, b$; 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. □

B1: Supporting Lemmas

Lemma 4. *Suppose that assumption 1 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 1 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*

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

where the dependence of e.g. $\zeta_{l,k,j,n}^{\alpha}$ on n is as in the proof of Theorem 1.[26] *Let $\breve{Q}_{l,m,i,n}^{r,s} \coloneqq q_{l,i,n}^{r}q_{m,i,n}^{s}$. Suppose that assumption 1 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 $(\tilde{X}, \epsilon)$.*

[26] See footnote 24.

Proof. By definition we have

$$\begin{aligned}
\breve{Q}^{\alpha,\alpha}_{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^{\alpha}_{l,k,j,n}\zeta^{\alpha}_{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^{\alpha}_{l,k,j,n}\zeta^{\alpha}_{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^{\alpha}_{l,k,k,n}\zeta^{\alpha}_{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,b}_{l,m,i,n} &= -\sum_{s=1}^{K}\sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K}\zeta^{\alpha}_{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^{\alpha}_{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^{\alpha}_{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^{\alpha}_{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, $\|\breve{Q}^{r,s}_{l,m,i,n}\|_{G,p} < \infty$, noting that for σ instead of α we have the same expressions. □

Lemma 7. *Suppose assumption 1 holds and $\nu_{n,p}$ and ν_n are as in assumption 2. 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

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

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

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 1 and 2 hold and $\theta_n = (\alpha_0, \beta_n, \eta)$ where $\sqrt{n}(\beta_n - \beta) = O(1)$ is a deterministic sequence. Then for each $r \in \{\alpha, \sigma, b\}$ and l*

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

Proof. In this proof we let $M_k \coloneqq M_{k\bullet}$ for any matrix M. We start by considering elements in $\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\ell}_{\gamma_n, \alpha_l}(Y_i) - \tilde{\ell}_{\theta_n, \alpha_l}(Y_i)\right)^2$ (noting that the result for σ will be the same). 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 $|\zeta^{\alpha}_{l,k,j,n}| < \infty$ 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.[27] 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 Lemma 3 applied with $W_{i,n} = A_{n,j}V_{i,n}$ and $\tilde{\tau}_{s,n,q} = o_{P_{\theta_n}}(\nu_n^{1/2})$.[28] For (39) use Cauchy-

[27] 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 1 and Markov's inequality, noting once more that $A_{n,k}V_{i,n} \simeq \epsilon_{k,i}$ under P_{θ_n}.

[28] See footnote 27.

Schwarz with lemma 3:

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

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

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

We have

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

The first term is $o_{P_{\theta_n}}(\nu_n)$ by Cauchy-Schwarz and applying lemma 3, 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. □

C: Proof of Theorem 2

Proof of Theorem 2. Let $P_0 \coloneqq P_{\theta_0}$, where θ_0 is defined in Assumption 4. The first step is to show that assumption 4 implies that

$$\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\gamma_n} - \tilde{\ell}_{\theta_n}\right] \xrightarrow{P_0} 0, \quad \sqrt{n}\mathbb{P}_n\left[\tilde{\ell}_{\theta_n} - \tilde{\ell}_{\theta_0}\right] + \sqrt{n}\tilde{I}_{\theta_0}(0, (\beta_n - \beta)')' \xrightarrow{P_0} 0 \tag{42}$$

and

$$\nu_n^{-1}\left\|\hat{I}_{\gamma_n} - \tilde{I}_{\theta_0}\right\| = o_{P_0}(1). \tag{43}$$

To do so, define $b_n \coloneqq \sqrt{n}(\beta_n - \beta)$ and let $(n_m)_{m\geq 1}$ be an arbitrary subsequence of $(n)_{n\geq 1}$. It is sufficient for (42)-(43) 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 $b \in \mathbb{R}^{L_\beta}$.[29] 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 3 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 left claim in (42) and (43) hold given parts 2 & 3 of assumption 4. Noting that $P_0[\tilde{\ell}_{\theta_0}\dot{\ell}'_{\theta_0,\beta}]b = \tilde{I}_{\theta_0}(0, b')'$, the right claim of equation (42) follows by proposition A.10 in van der Vaart (1988), which requires Assumption 4-part 4.[30]

Next we show that (42) and (43) continue to hold if γ_n (and $\theta_n = (\gamma_n, \eta)$) is replaced by $\bar{\gamma}_n$ (and $\bar{\theta}_n = (\bar{\gamma}_n, \eta)$) as defined in the theorem.[31] Since $\bar{\beta}_n$ remains $\sqrt{n}$-consistent there is an $M > 0$ such that $P_0\left(\sqrt{n}\|\bar{\beta}_n - \beta\| > M\right) < \varepsilon$. If $\sqrt{n}\|\bar{\beta}_n - \beta\| \leq M$ then the discretized estimator $\bar{\beta}_n$ 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

$$R'_n(\beta') \coloneqq \sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\gamma'} - \tilde{\ell}_{\theta'}\right], \; R''_n(\beta') \coloneqq \sqrt{n}\mathbb{P}_n\left[\tilde{\ell}_{\theta'} - \tilde{\ell}_{\theta_0}\right] + \sqrt{n}\tilde{I}_{\theta_0}(0, (\beta' - \beta)')', \; R'''_n(\beta') \coloneqq \nu_n^{-1}[\hat{I}_{\gamma'} - \tilde{I}_{\theta_0}],$$

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

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

[31] The proof is adapted from the proof of Theorem 5.48 in van der Vaart (1998).

where $\gamma' = (\alpha_0, \beta')$ and $\theta' = (\gamma', \eta)$. 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 equations (42) and (43).

By the version of (42) with γ_n, θ_n replaced by $\bar{\gamma}_n, \bar{\theta}_n$ we have

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

and by the version of (43) with γ_n, θ_n replaced by $\bar{\gamma}_n, \bar{\theta}_n$, $\hat{I}_{\bar{\gamma}_n} \xrightarrow{P_0} \tilde{I}_{\theta_0}$ and so $\hat{\mathcal{K}}_{\bar{\gamma}_n} \xrightarrow{P_0} \tilde{\mathcal{K}}_{\theta_0}$ for

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

We combine these to obtain

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

Then, by assumption 4-part 1, under P_0,

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

For the next step, observe that

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

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 at a non-singular matrix we obtain

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

Hence by equation (43) with $\bar{\gamma}_n$ replacing γ_n we have $P_0\left(\left\|\hat{\mathcal{I}}_{\bar{\gamma}_n} - \tilde{\mathcal{I}}_{\theta_0}\right\|_2 < \nu_n\right) \to 1$.

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 $\hat{\mathcal{I}}^t_{\bar{\gamma}_n} \xrightarrow{P_0} \tilde{\mathcal{I}}_{\theta_0}$ and the rank estimate $r_n = \text{rank}(\hat{\mathcal{I}}^t_{\bar{\gamma}_n})$ 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 $\hat{\mathcal{I}}_{\bar{\gamma}_n}$ and $\hat{\lambda}^t_{l,n}$ for $\hat{\mathcal{I}}^t_{\bar{\gamma}_n}$. Define the set $R_n \coloneqq \{r_n = r\}$, let $\underline{\nu} \coloneqq \lambda_r/2 > 0$ and note that $\|\hat{\mathcal{I}}_{\bar{\gamma}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 = o_{P_0}(\nu_n)$ implies that $\|\hat{\mathcal{I}}_{\bar{\gamma}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 = o_{P_0}(1)$.

By Weyl's perturbation theorem[32] we have $\max_{l=1,\dots,L_\alpha} |\hat{\lambda}_{l,n} - \lambda_l| \leq \|\hat{\mathcal{I}}_{\bar{\gamma}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 = o_{P_0}(1)$. 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 $\hat{\mathcal{I}}^t_{\bar{\gamma}_n} = \hat{\mathcal{I}}_{\bar{\gamma}_n}$. Thus, $E_n \cap \{\|\hat{\mathcal{I}}_{\bar{\gamma}_n} - \tilde{\mathcal{I}}_{\theta_0}\| \leq \upsilon\} \subset \{\|\hat{\mathcal{I}}^t_{\bar{\gamma}_n} - \tilde{\mathcal{I}}_{\theta_0}\| \leq \upsilon\}$, from which it follows that $\hat{\mathcal{I}}^t_{\bar{\gamma}_n} \xrightarrow{P_0} \tilde{\mathcal{I}}_{\theta_0}$.

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 $n \to \infty$

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

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 $\|\hat{\mathcal{I}}_{\bar{\gamma}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 \leq \upsilon$, 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

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

and hence $\{\|\hat{\mathcal{I}}_{\bar{\gamma}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 \leq \upsilon\} \cap E_n \cap F_n \subset \{\|\hat{\Lambda}_n(\nu_n) - \Lambda\|_2 \leq \upsilon\}$, 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 eigenprojection is $\Pi_k \coloneqq \sum_{i=1}^{\mathfrak{m}_k} u_{l^k_i} u'_{l^k_i}$.[33] Total eigenprojections are continuous.[34] 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 $\hat{\mathcal{I}}_{\bar{\gamma}_n} \xrightarrow{P_0} \tilde{\mathcal{I}}_{\theta_0}$. Spectrally decompose $\tilde{\mathcal{I}}_{\theta_0}$

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

[33] See e.g Chapter 8.8 of Magnus and Neudecker (2019).

[34] E.g. Theorem 8.7 of Magnus and Neudecker (2019).

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,

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

and so

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

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 $\hat{\mathcal{I}}^t_{\bar{\gamma}_n} \xrightarrow{P_0} \tilde{\mathcal{I}}_{\theta_0}$ and $P_0(\{r_n = r\}) \to 1$. This implies that $\hat{\mathcal{I}}^{t,\dagger}_{\bar{\gamma}_n} \xrightarrow{P_0} \tilde{\mathcal{I}}^{\dagger}_{\theta_0}$ where $\tilde{\mathcal{I}}^{\dagger}_{\theta_0}$ is the Moore-Penrose inverse of $\tilde{\mathcal{I}}_{\theta_0}$.[35]

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

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

where the distributional result $X \coloneqq Z' \tilde{\mathcal{I}}^{\dagger}_{\theta_0} Z \sim \chi^2_r$, 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 $\hat{S}_{\bar{\gamma}_n} - c_n \rightsquigarrow X - c$ where $X \sim \chi^2_r$. Since the χ^2_r distribution is continuous, we have by the Portmanteau theorem

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

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 4 and Weyl's perturbation theorem imply that

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

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

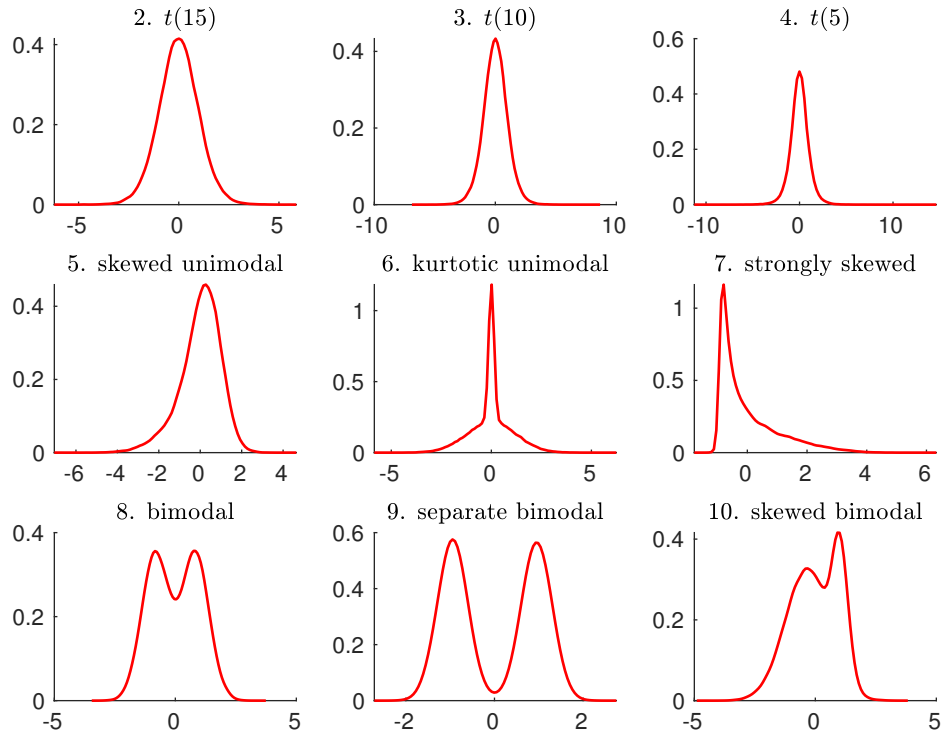
$$P_0(\hat{S}_{\bar{\gamma}_n} > c_n) \leq 1 - P_0(E_n) \to 0.$$

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

[35] See e.g. Theorem 2 of Andrews (1987).

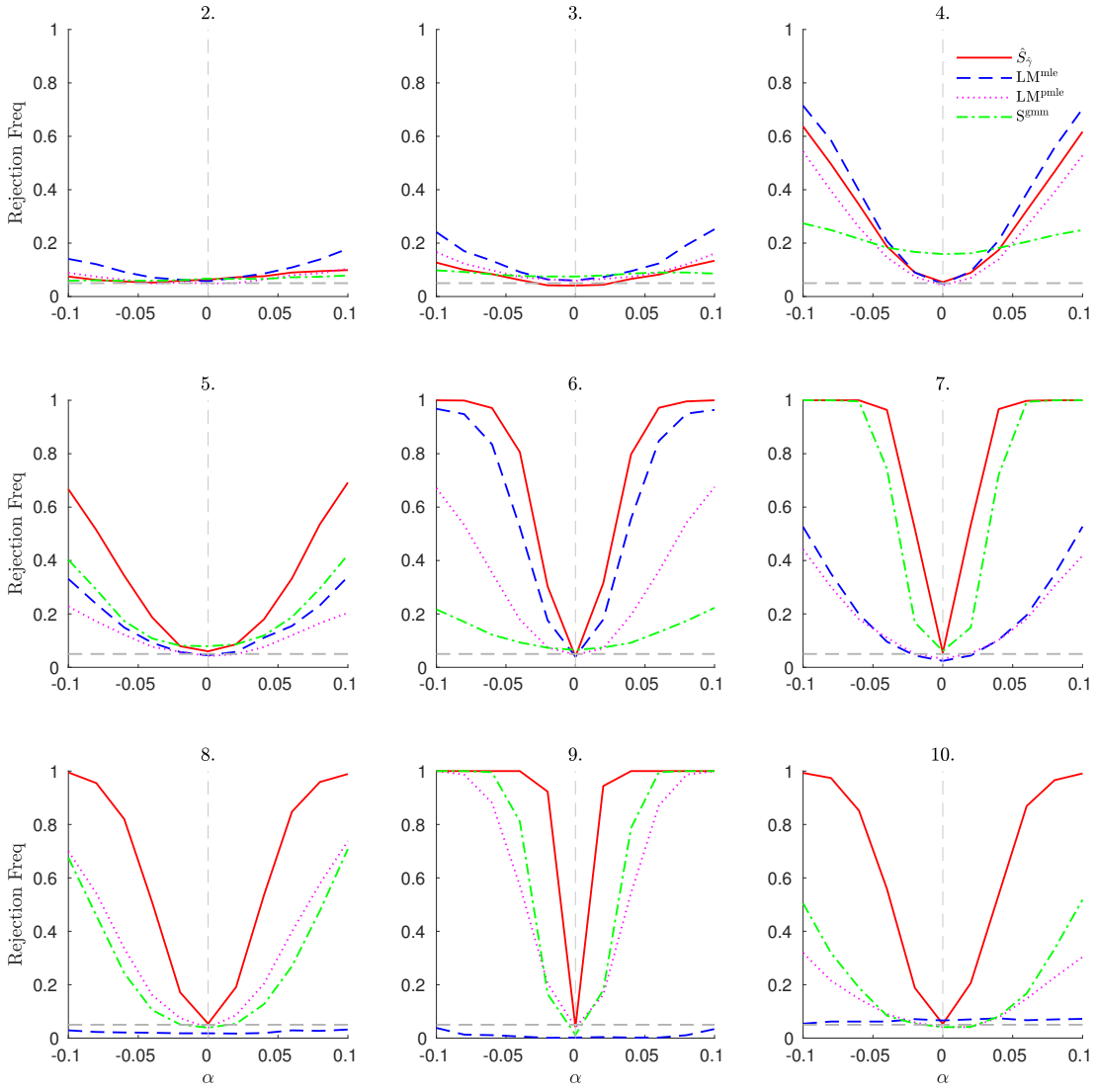
D: Figures and tables

Figure 3: Structural Shock Densities



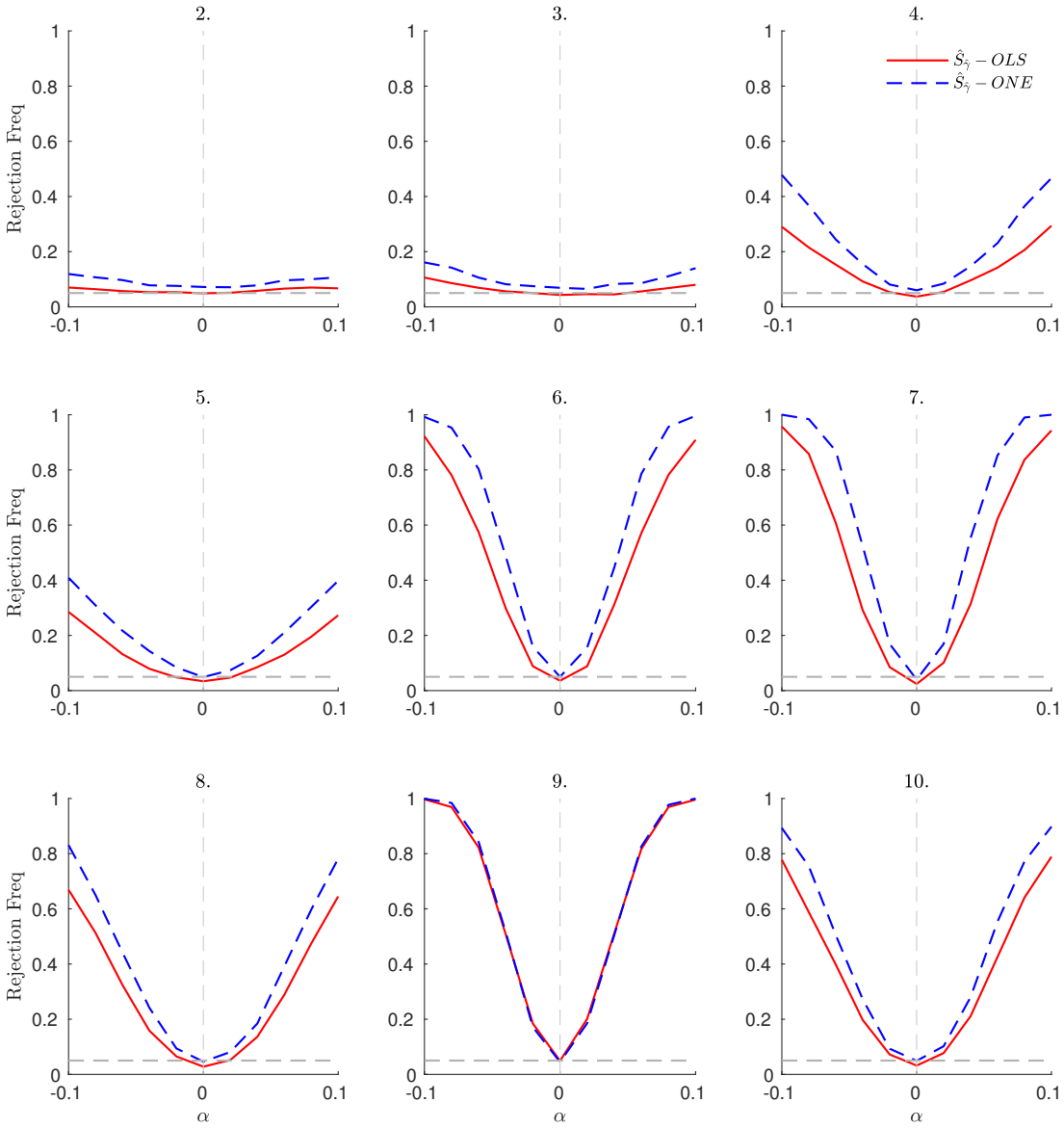
Notes: The plots show the different densities considered for simulating the structural shocks. Densities 2-4 are t-distributions normalised to have unit variance. Densities 5 - 10 (and their names) are mixtures of normals taken from Marron and Wand (1992); see their table 1 for the definitions. Density 1 is the standard Gaussian and omitted from the figure.

Figure 4: Power Comparison Baseline model



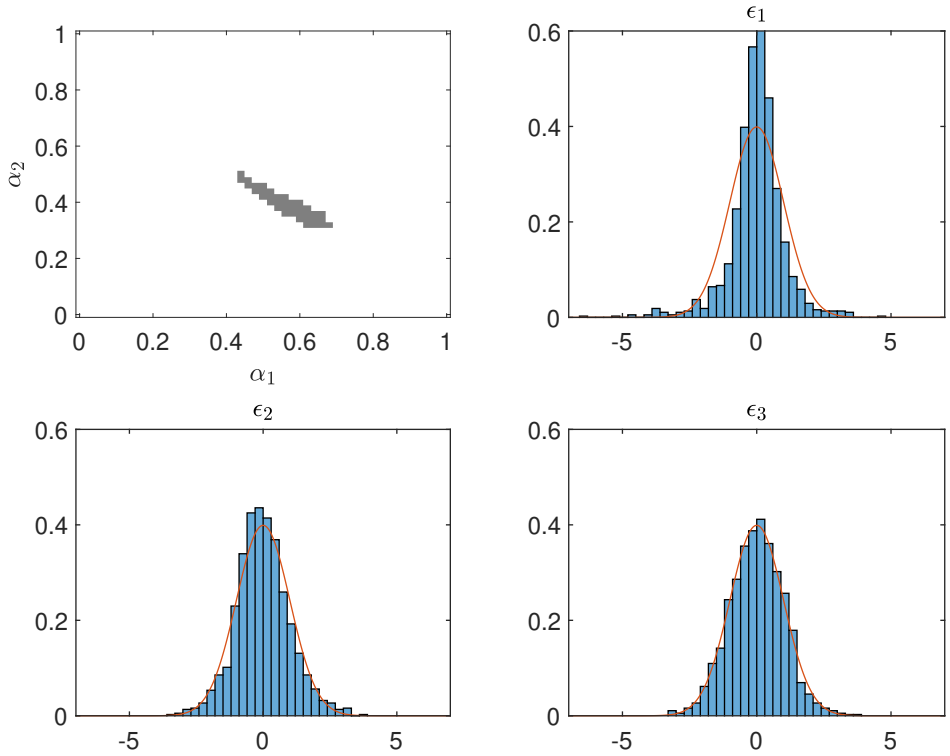
Notes: Empirical power curves for the baseline model with $k = 2$ and $n = 1000$. Each plot corresponds to the choice for densities $\epsilon_{i,k}$, for $k \geq 2$, where the numbers correspond to the different densities shown in Figure 3. The solid red line corresponds to $S_{\hat{\gamma}}$, the dashed blue line to LM^{mle}, the dotted pink line to LM^{pmle} and the dot-dashed green line to S^{gmm}.

Figure 5: Power LSEM



Notes: Empirical power curves for the LSEM model with $k = 2$, $d = 2$ and $n = 1000$. Each plot corresponds to the choice for densities $\epsilon_{i,k}$, for $k \geq 2$, where the numbers correspond to the different densities shown in Figure 3. The solid red line corresponds to the empirical rejection frequency of the $\hat{S}_{\hat{\gamma}}$ test where $\hat{\gamma} = (\alpha_0, \hat{\beta})$, with $\hat{\beta}$ the OLS estimator. The dashed blue line corresponds to the empirical rejection frequency of the $\hat{S}_{\hat{\gamma}}$ test where $\hat{\gamma} = (\alpha_0, \hat{\beta})$, with $\hat{\beta}$ the one-step efficient MLE estimator.

Figure 6: 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 7: 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 .

Table 2: Rejection Frequencies $\hat{S}_{\hat{\gamma}}$ test for Baseline model

n	K	B	1	2	3	4	5	6	7	8	9	10
200	2	4	0.049	0.049	0.048	0.040	0.047	0.049	0.034	0.049	0.048	0.048
200	2	6	0.048	0.045	0.049	0.044	0.048	0.053	0.047	0.045	0.058	0.051
200	2	8	0.050	0.049	0.047	0.044	0.048	0.048	0.053	0.050	0.051	0.047
200	3	4	0.043	0.039	0.039	0.039	0.044	0.048	0.026	0.049	0.052	0.050
200	3	6	0.045	0.038	0.040	0.044	0.041	0.048	0.044	0.047	0.052	0.043
200	3	8	0.047	0.046	0.040	0.040	0.044	0.048	0.042	0.049	0.044	0.051
200	5	4	0.032	0.034	0.033	0.034	0.035	0.039	0.015	0.041	0.045	0.043
200	5	6	0.037	0.033	0.036	0.032	0.032	0.040	0.043	0.045	0.043	0.044
200	5	8	0.039	0.038	0.038	0.030	0.035	0.043	0.045	0.040	0.041	0.038
500	2	4	0.053	0.046	0.053	0.045	0.047	0.052	0.031	0.049	0.045	0.046
500	2	6	0.048	0.049	0.048	0.048	0.049	0.052	0.057	0.047	0.047	0.049
500	2	8	0.048	0.048	0.045	0.049	0.047	0.045	0.051	0.052	0.048	0.045
500	3	4	0.042	0.039	0.040	0.046	0.048	0.048	0.021	0.042	0.046	0.047
500	3	6	0.043	0.045	0.042	0.042	0.045	0.047	0.047	0.051	0.044	0.045
500	3	8	0.046	0.045	0.040	0.035	0.042	0.047	0.044	0.045	0.050	0.047
500	5	4	0.040	0.036	0.039	0.036	0.041	0.046	0.016	0.048	0.047	0.046
500	5	6	0.041	0.039	0.039	0.039	0.040	0.049	0.046	0.045	0.044	0.044
500	5	8	0.039	0.040	0.036	0.041	0.043	0.050	0.050	0.044	0.046	0.047
1000	2	4	0.042	0.052	0.040	0.055	0.047	0.052	0.046	0.052	0.046	0.048
1000	2	6	0.054	0.052	0.045	0.050	0.045	0.049	0.049	0.054	0.045	0.057
1000	2	8	0.047	0.048	0.048	0.047	0.048	0.052	0.050	0.048	0.055	0.052
1000	3	4	0.049	0.041	0.043	0.045	0.048	0.050	0.054	0.051	0.051	0.047
1000	3	6	0.048	0.044	0.038	0.040	0.050	0.047	0.046	0.049	0.051	0.045
1000	3	8	0.046	0.047	0.047	0.042	0.049	0.045	0.050	0.052	0.043	0.047
1000	5	4	0.038	0.035	0.038	0.047	0.041	0.044	0.050	0.046	0.047	0.048
1000	5	6	0.041	0.043	0.039	0.042	0.043	0.049	0.044	0.048	0.048	0.049
1000	5	8	0.042	0.042	0.038	0.039	0.048	0.050	0.049	0.047	0.045	0.049

Notes: The table shows the empirical rejection frequencies for the $S_{\hat{\gamma}}$ test based on $S = 5,000$ Monte Carlo replications for the baseline model $Y_i = R'\epsilon_i$. The test has nominal size $a = 0.05$. The columns denote the sample size n, the dimension of the model K, the number of B-splines B and the choice for densities $\epsilon_{i,k}$, for $k \geq 2$, where the numbers correspond to the different densities shown in Figure 3.

Table 3: Rejection Frequencies Alternative Tests for Baseline model

Cat (i)	n	1	2	3	4	5	6	7	8	9	10
W^{mle}	200	0.179	0.149	0.139	0.127	0.113	0.059	0.097	0.152	0.125	0.171
	500	0.180	0.133	0.114	0.115	0.095	0.167	0.073	0.114	0.097	0.150
	1000	0.188	0.101	0.079	0.074	0.061	0.405	0.058	0.124	0.103	0.170
LR^{mle}	200	0.028	0.054	0.060	0.046	0.054	0.026	0.048	0.017	0.018	0.024
	500	0.043	0.056	0.068	0.054	0.065	0.023	0.053	0.016	0.017	0.024
	1000	0.049	0.065	0.063	0.061	0.053	0.031	0.051	0.022	0.018	0.025
W^{pmle}	200	0.375	0.211	0.198	0.086	0.141	0.058	0.105	0.495	0.998	0.467
	500	0.485	0.264	0.204	0.073	0.163	0.030	0.079	0.973	0.999	0.870
	1000	0.570	0.230	0.180	0.051	0.131	0.023	0.068	0.428	1.000	0.947
LR^{gmm}	200	0.413	0.411	0.425	0.441	0.290	0.379	0.120	0.216	0.086	0.232
	500	0.292	0.246	0.246	0.286	0.141	0.171	0.025	0.109	0.066	0.106
	1000	0.232	0.181	0.155	0.176	0.074	0.115	0.014	0.068	0.059	0.049
Cat (ii)	n	1	2	3	4	5	6	7	8	9	10
$\hat{S}_{\hat{\gamma}}$	200	0.051	0.047	0.048	0.040	0.049	0.049	0.047	0.048	0.050	0.044
	500	0.047	0.047	0.054	0.047	0.044	0.043	0.047	0.048	0.051	0.054
	1000	0.047	0.043	0.046	0.049	0.048	0.047	0.050	0.044	0.049	0.043
LM^{mle}	200	0.052	0.058	0.054	0.043	0.040	0.043	0.023	0.018	0.002	0.059
	500	0.056	0.052	0.052	0.042	0.046	0.047	0.028	0.017	0.001	0.062
	1000	0.062	0.052	0.050	0.049	0.039	0.040	0.029	0.016	0.002	0.052
LM^{plme}	200	0.049	0.045	0.049	0.035	0.038	0.046	0.030	0.041	0.042	0.042
	500	0.049	0.047	0.050	0.039	0.047	0.046	0.034	0.046	0.044	0.051
	1000	0.046	0.048	0.053	0.044	0.041	0.046	0.034	0.042	0.052	0.047
S^{gmm}	200	0.188	0.209	0.248	0.326	0.236	0.264	0.195	0.108	0.059	0.130
	500	0.094	0.105	0.123	0.223	0.116	0.133	0.103	0.057	0.028	0.064
	1000	0.061	0.070	0.081	0.162	0.069	0.078	0.054	0.031	0.019	0.035

Notes: The table shows the empirical rejection frequencies based on $S = 5,000$ Monte Carlo replications for the baseline model $Y_i = R'\epsilon_i$, with $n = 500$ and $K = 2$. All tests have nominal size $a = 0.05$. The first column indicates the test. The remaining columns denote the choice for densities $\epsilon_{i,k}$, for $k \geq 2$, where the numbers correspond to the different densities shown in Figure 3.

Table 4: Rejection Frequencies $\hat{S}_{\hat{\gamma}}$ test for LSEM - OLS $\hat{\beta}$

n	K	d	1	2	3	4	5	6	7	8	9	10
200	2	2	0.050	0.054	0.049	0.049	0.038	0.030	0.038	0.043	0.057	0.046
200	2	3	0.049	0.054	0.054	0.048	0.046	0.059	0.042	0.035	0.029	0.052
200	3	2	0.056	0.058	0.050	0.062	0.059	0.031	0.018	0.038	0.047	0.050
200	3	3	0.063	0.054	0.057	0.065	0.060	0.025	0.023	0.051	0.058	0.049
200	5	2	0.098	0.104	0.109	0.142	0.094	0.051	0.064	0.054	0.023	0.057
200	5	3	0.116	0.116	0.131	0.155	0.103	0.039	0.029	0.061	0.026	0.072
500	2	2	0.049	0.050	0.039	0.042	0.041	0.027	0.029	0.036	0.026	0.029
500	2	3	0.048	0.041	0.047	0.047	0.037	0.029	0.024	0.034	0.050	0.051
500	3	2	0.051	0.051	0.048	0.040	0.037	0.028	0.029	0.038	0.022	0.039
500	3	3	0.048	0.050	0.047	0.051	0.053	0.028	0.048	0.041	0.037	0.036
500	5	2	0.071	0.078	0.068	0.081	0.049	0.023	0.060	0.042	0.039	0.038
500	5	3	0.067	0.068	0.080	0.085	0.063	0.022	0.045	0.049	0.027	0.051
1000	2	2	0.040	0.051	0.049	0.029	0.043	0.032	0.033	0.045	0.049	0.041
1000	2	3	0.048	0.044	0.040	0.040	0.040	0.030	0.038	0.046	0.030	0.044
1000	3	2	0.045	0.038	0.043	0.034	0.033	0.032	0.034	0.040	0.039	0.042
1000	3	3	0.044	0.045	0.043	0.036	0.030	0.032	0.035	0.040	0.024	0.034
1000	5	2	0.059	0.051	0.057	0.051	0.039	0.024	0.063	0.030	0.028	0.036
1000	5	3	0.057	0.058	0.056	0.050	0.035	0.018	0.046	0.036	0.029	0.040

Notes: The table shows the empirical rejection frequencies for the $S_{\hat{\gamma}}$ 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 model K, the number of covariates d and the choice for densities $\epsilon_{i,k}$, for $k \geq 2$, where the numbers correspond to the different densities shown in Figure 3. The $S_{\hat{\gamma}}$ test was implemented using $B = 6$ B-splines.

Table 5: Rejection Frequencies $\hat{S}_{\hat{\gamma}}$ test for LSEM - One-step $\hat{\beta}$

n	K	d	1	2	3	4	5	6	7	8	9	10
200	2	2	0.067	0.080	0.068	0.081	0.070	0.031	0.054	0.056	0.061	0.051
200	2	3	0.068	0.074	0.076	0.072	0.066	0.071	0.057	0.047	0.026	0.061
200	3	2	0.095	0.106	0.104	0.120	0.090	0.041	0.026	0.059	0.036	0.061
200	3	3	0.099	0.103	0.105	0.114	0.098	0.037	0.028	0.071	0.035	0.064
200	5	2	0.187	0.226	0.247	0.264	0.178	0.063	0.040	0.072	0.020	0.068
200	5	3	0.212	0.238	0.262	0.289	0.193	0.064	0.049	0.089	0.036	0.088
500	2	2	0.062	0.062	0.068	0.067	0.057	0.034	0.049	0.041	0.021	0.037
500	2	3	0.059	0.064	0.071	0.069	0.056	0.031	0.019	0.046	0.031	0.051
500	3	2	0.078	0.078	0.081	0.079	0.066	0.026	0.024	0.047	0.021	0.045
500	3	3	0.076	0.081	0.091	0.088	0.068	0.025	0.029	0.050	0.042	0.042
500	5	2	0.112	0.149	0.158	0.181	0.097	0.036	0.035	0.060	0.030	0.044
500	5	3	0.129	0.151	0.168	0.180	0.101	0.033	0.023	0.069	0.031	0.058
1000	2	2	0.059	0.059	0.065	0.048	0.049	0.025	0.021	0.055	0.050	0.038
1000	2	3	0.060	0.060	0.060	0.068	0.057	0.038	0.052	0.050	0.027	0.051
1000	3	2	0.061	0.067	0.068	0.065	0.053	0.023	0.048	0.047	0.023	0.045
1000	3	3	0.064	0.066	0.072	0.070	0.054	0.040	0.016	0.047	0.022	0.041
1000	5	2	0.091	0.105	0.108	0.111	0.069	0.032	0.026	0.042	0.029	0.043
1000	5	3	0.085	0.102	0.120	0.103	0.065	0.026	0.020	0.047	0.026	0.050

Notes: The table shows the empirical rejection frequencies for the $\hat{S}_{\hat{\gamma}}$ test based on $S = 5,000$ Monte Carlo replications for the linear simultaneous equations model (3). The test has nominal size $a = 0.05$. The columns denote the sample size n, the dimension of the observations K, the number of covariates d and the choice for densities $\epsilon_{i,k}$, for $k \geq 2$, where the numbers correspond to the different densities shown in Figure 3. The $S_{\hat{\gamma}}$ test was implemented using $B = 6$ B-splines and using OLS estimates for β.

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
p^{ind}	0.12	0.16	

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. The bottom row shows the p-value for the independence test proposed by Matteson and Tsay (2017) as performed on $\{\hat{A}(Y_i - \hat{B}X_i)\}_{i=1}^n$, where $\hat{A} = D(\hat{\sigma})^{-1}S(\hat{\alpha}, \hat{\sigma})$, with $\hat{\alpha}$ denoting the minimizer of $\hat{S}_{\hat{\gamma}}$ and $\hat{\sigma}$ and $\hat{B}$ the OLS estimates for σ and B.