Robust Non-Gaussian Inference[*]

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

[(a)] Universitat Pompeu Fabra and Barcelona GSE

[(b)] Vrije Universiteit Amsterdam

December 10, 2020

Abstract

Systems of equations become identified (up to permutation and sign) if the underlying structural shocks are independent and if at most one of them is Gaussian. Unfortunately, existing inference methods that exploit such non-Gaussian identifying assumption suffer from size distortions when the true shocks are close to Gaussian. To address this *weak non-Gaussian* problem, we develop a robust semi-parametric testing approach 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-parametric and construct asymptotically efficient tests based on the efficient score function. The approach is applicable for a broad class of simultaneous equations models including structural VAR models. We evaluate the method in a large simulation study and an empirical study.

Keywords: Weak identification, semiparametric modeling, independent component analysis, structural VAR models.

[*] First draft: December 2020. Email: `adam.lee@upf.edu`, `geert.mesters@upf.edu`. We thank numerous seminar participants for helpful comments. Mesters acknowledge support from the Spanish Ministry of Economy and Competitiveness, Grant ECO2015-68136-P and FEDER, EU, the Spanish Ministry of Economy and Competitiveness through the Severo Ochoa Programme for Centres of Excellence in R&D (SEV-2015-0563), and Fundacion BBVA scientific research grant (PR16-DAT-0043).

1 Introduction

Non-Gaussian distributions can help to identify structural parameters in various structural models, see Hyvärinen, Karhunen and Oja (2001) and Gouriéroux, Monfort and Renne (2019) for prominent examples in statistics and econometrics, respectively. Unfortunately, existing inference methods suffer from size distortions when the true distributions are close to Gaussian. To remedy this problem we propose a robust inference method that exploits non-Gaussianity, but does not assume it.

To outline the problem consider the simple model

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

where Y is a $K \times 1$ vector, A is a $K \times K$ invertible matrix and ϵ is a $K \times 1$ vector that has independent components. The goal is to recover A^{-1}, or, perhaps more usefully, A in $AY = \epsilon$, from a sample of independent realizations of Y. This procedure is known in the statistic literature as *independent components analysis* (ICA), see Hyvärinen, Karhunen and Oja (2001).

When the components of ϵ follow Gaussian distributions A can only be identified up to orthogonal transformations. In contrast, when at least $K-1$ components of ϵ follow non-Gaussian distributions A can be recovered up to sign and permutation of its columns (see Common, 1994). In other words, non-Gaussianity shrinks the identified set and can provide useful information for pinning down the location of A.

With this in mind, a common approach for conducting inference on A is as follows: (i) *assume* that sufficiently many components of ϵ follow a non-Gaussian distribution, (ii) estimate A using maximum likelihood methods or (generalized) method of moments, and (iii) construct confidence bands for some function of A based on the sampling variation of the estimator. Both parametric and semi-parametric estimators can be considered, see Hyvärinen, Karhunen and Oja (2001) and Gouriéroux, Monfort and Renne (2017) for important examples.

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

From an economics perspective, model (1) is more usefully viewed a building block in a larger simultaneous equations model possibly including covariates and a dynamic structure. For such models it follows that whenever the object of interest depends on A – for which inference is conducted in the aforementioned approach – the size-distortions carry over. Prominent examples where this problem arises include (i) inference for impulse responses in structural vector autoregressive (SVAR) models (e.g. Hyvärinen et al., 2010; Lanne and Ltkepohl, 2010; Moneta et al., 2013; Lanne, Meitz and Saikkonen, 2017; Maxand, 2018; Lanne and Luoto, 2019*a*; Gouriéroux, Monfort and Renne, 2017, 2019; Tank, Fox and Shojaie, 2019; Herwartz, 2019; Bekaert, Engstrom and Ermolov, 2019, 2020; Fiorentini and Sentana, 2020), (ii) tests for invertibility and fundamentalness (e.g. Sahneh, 2015; Chen, Choi and Escanciano, 2017) and (iii) potentially inference for common components in factor models (e.g. Bonhomme and Robin, 2009). In any of these examples, it holds that size distortions occur whenever the true densities of ϵ are close to the Gaussian distribution.

To this extent, this paper develops a robust approach for conducting inference on A that is inspired by the identification robust methods developed in econometrics (e.g. Stock and Wright, 2000; Kleibergen, 2005; Andrews and Mikusheva, 2015) and the general semiparametric statistical theory that is discussed in Bickel et al. (1998) and van der Vaart (2002). In particular, we construct confidence bands for the elements of A by inverting singularity and identification robust semiparametric score test statistics. The score test is shown to be correctly sized regardless of the *distance-to-Gaussianity* of ϵ and – under non-singularity – it is included in the class of asymptotically uniformly most powerful invariant (AUMPI) tests.

We start by providing a general, and quite high level, framework for conducting singularity and identification robust hypothesis tests in semiparametric likelihood models where the null hypothesis concerns a finite dimensional parameter vector and there exists an infinite dimensional, but well identified, nuisance parameter. The testing approach is characterized by two steps. In the first step an estimate for the efficient score function of the finite dimensional parameter of interest is constructed and in the second step this estimate is used to construct a singularity robust score statistic. The test statistic can be viewed as the semiparametric version of a Neyman-Rao score statistic, with an adjustment for a possibly singular variance matrix, see also Choi, Hall and Schick (1996) for the non-singular case and Andrews and Guggenberger (2019) for singularity adjustments in parametric models.

With our general framework in hand, we turn to the ICA model (1) and its extensions. We start by casting the ICA model as a semiparametric model in which A determines the parametric part and the densities of the components of ϵ form the non-parametric part. Given a set of mild regularity conditions we analytically derive the efficient score function following Amari and Cardoso (1997) and show that it can be consistently estimated using the B-spline based log density score estimator of Jin (1992) and Chen and Bickel (2006). Based on the estimate of the efficient score function we can directly compute the score statistic which is shown to have a standard chi-squared limiting distribution. Importantly, this result does not assume any form of non-Gaussianity, and the score statistic has the same limiting distribution regardless of the distance-to-Gaussianity. Moreover, computing the score statistic is trivial as it essentially only requires K regressions to estimate the log density scores, thus avoiding the usage of numerical optimization routines.

To extend the applicability of our approach to a broad class of simultaneous equations models we consider situations where we do not observe realizations of Y, but instead are only able to estimate Y based on some observable data sample. Prominent examples included in this class are linear simultaneous equations models with predetermined explanatory variables and structural VAR models. In general, the estimation noise from the initial estimation step, required to estimate Y, is non-negligible and we show how to adjust the variance of the efficient score function to account for this. With this adjustment the main results for the baseline model (1) carry over.

We evaluate the finite sample performance of the semiparametric score tests in a large simulation study. We show that regardless of how close ϵ is to the Gaussian distribution (i.e. how well A is identified) our test is correctly sized. In contrast, tests that are based on the sampling variation of (pseudo)-maximum likelihood or GMM estimators for A have large size distortions. Importantly, we find that pre-testing for non-Gaussianity does not fix this problem.[1] Further, for moderate sample sizes the power of the semiparametric test is similar when compared to the parametric score test that relies on knowing the functional form of the density. These findings show that our asymptotic theory is useful for obtaining finite sample approximations.

In an empirical study we consider estimating supply and demand elasticities in the US labor market (e.g. Baumeister and Hamilton, 2015; Lanne and Luoto, 2019*b*). We show that allowing for non-Gaussian distributions creates some identifying power that eliminates the need for some of the assumptions imposed by Baumeister and Hamilton (2015) and can be done in a robust way without actually assuming non-Gaussian densities. However, the resulting confidence sets are larger when compared to the non-robust methods, implying that the weak non-Gaussian problem is likely to be relevant in this setting.

Our approach builds on three strands of literature: identification robust testing, semiparametric inference and the ICA model and its extensions.

Regarding the weak identification robust literature, a useful analogy is obtained when we compare the non-Gaussian identification approach to an instrumental variable based identification approach. In textbook IV, identification is established theoretically by assuming

[1] Pre-testing here is defined as a two-step procedure, where in the first step the elements of Y are tested for normality and if normality is rejected step two proceeds by maximum likelihood or moment based inference for A.

that the covariance matrix between the instruments and the endogenous variables has full rank. In practice however, what matters for reliable standard inference is that the first stage F-statistic is larger then some threshold value, informally put, the correlation between the instruments and the endogenous variables should be sufficiently strong (e.g. Staiger and Stock, 1997; Stock and Yogo, 2005). In a similar way, in the ICA model non-Gaussianity can be viewed as a theoretical identification assumption (e.g. Hyvärinen, Karhunen and Oja, 2001), but what matters in practice is the distance to the Gaussian distribution. To avoid relying on the strict non-Gaussian identification assumption we consider test statistics whose asymptotic size does not depend on this assumption, similar in spirit to the identification robust tests that have been constructed for the IV problem which avoid explicitly relying on the covariance between instruments and the endogenous variables for inference (e.g. Anderson and Rubin, 1949; Staiger and Stock, 1997; Stock and Wright, 2000; Kleibergen, 2005; Andrews and Mikusheva, 2016).

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

The semiparametric literature in statistics has mainly focused on efficient estimation in well identified models Bickel et al. (1998) and van der Vaart (2002). A few papers focus on testing in well-identified semiparametric models (e.g. Choi, Hall and Schick, 1996; Bickel, Ritov and Stoker, 2006). The approach of Choi, Hall and Schick (1996) is most closely related to our general framework but does not deal with identification failures and the associated singularity of the efficient information matrix.

Finally, there exists a rich literature on ICA models and applications thereof (e.g. Hyvärinen, Karhunen and Oja, 2001). This paper relates most closely to papers that treat the density functions of ϵ non-parametrically, see Bach and Jordan (2002) and Chen and Bickel (2006). Empirically our motivation stems from an increasing number of papers in econometrics that rely on a non-Gaussianity assumption for identification in extensions of the ICA model, notably structural vector autoregressive models (e.g. Chen, Choi and Escanciano, 2017; Lanne, Meitz and Saikkonen, 2017; Lanne and Luoto, 2019a; Gouriéroux, Monfort and Renne, 2017, 2019; Bekaert, Engstrom and Ermolov, 2019).

The remainder of this paper is organized as follows. In the next section we discuss a general framework for conducting singularity and identification robust tests in semiparametric models. Section 3 gives the implementation details and primitive assumptions for the ICA model and some of its extensions. Section 4 summarizes the results from the simulation study. Section 6 concludes. Any references to sections, equations, lemmas etc. which start with “S” refer to the supplementary material.

2 Robust testing in semiparametric models

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

To outline the setting, consider the random vector $Y \in \mathcal{Y} \subset \mathbb{R}^K$ that is defined on some underlying probability space $(\Omega, \mathcal{F}, P)$ with distribution specified by the law P_{θ_0} that depends on parameters $\theta_0 \in \Theta$. The parameter space Θ has the form $\Theta = \mathcal{A} \times \mathcal{H}$, where $\mathcal{A} \subset \mathbb{R}^L$ and $\mathcal{H} \subset \mathcal{M}$, with $\mathcal{M}$ a Banach space. We write a typical element of Θ as $\theta = (\alpha, \eta)$, where it is understood that $\alpha \in \mathcal{A}$ and $\eta \in \mathcal{H}$.

The model that the researcher considers is the collection

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

Typically, when $\mathcal{H}$ is finite dimensional we think of model (2) as parametric, whereas if $\mathcal{H}$ is infinite dimensional the model is classified as semiparametric, see Bickel et al. (1998) and van der Vaart (2002) for textbook treatments.

In general, we assume that η does not suffer from identification problems, but α may. In particular, for different points $\eta \in \mathcal{H}$ 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

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

The main idea is to develop test statistics whose limiting distribution is invariant to the identification strength of α. Such test statistics can then be inverted to yield confidence intervals for α with correct coverage.[2] Following Choi, Hall and Schick (1996) and Bickel, Ritov and Stoker (2006) we concentrate our efforts on test statistics that are based on the

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

efficient score function for α.

Formally, we define scores of the model (2) to be the quadratic mean derivatives of root-density paths.

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

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

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

If we let $t \mapsto P_t$ range over a collection of submodels, indexed by $\mathcal{I}$, we will obtain a collection of score functions, say s_i for $i \in \mathcal{I}$. This collection, $\{s_i : i \in \mathcal{I}\}$, will be denoted by $\mathcal{T}_{P,\mathcal{I}}$ and if it is a linear space we refer to it as a *tangent space*. For the semiparametric model (2) we define tangent spaces along restricted paths concerning the two parts of the parameter $\theta = (\alpha, \eta)$ separately.

First, let $\mathcal{T}^{\alpha|\eta}_{P_\theta,\mathbb{R}^L} = \{a'\dot{\ell}_\theta : a \in \mathbb{R}^L\}$, where $\dot{\ell}_\theta$ is the $L \times 1$ vector of scores of α evaluated at $\theta = (\alpha, \eta)$. Formally, this is the space of scores corresponding to paths of the form $t \mapsto P_{(\alpha+ta,\eta)}$ for $a \in \mathbb{R}^L$; these are scores corresponding to the parametric model $\{P_{(\alpha,\eta)} : \alpha \in \mathbb{R}^L\}$. Second, let $\mathcal{T}^{\eta|\alpha}_{P_\theta,H}$ be the tangent space at P_θ formed of scores corresponding to paths of the form $t \mapsto P_{(\alpha,\eta_t(\alpha,\eta,h))}$ for $h \in H$; these are scores corresponding to the nonparametric part of the model. Finally, let $\mathcal{J} = \mathbb{R}^L \times H$. We postulate that $\mathcal{T}_{P_\theta,\mathcal{J}} = \mathcal{T}^{\alpha|\eta}_{P_\theta,\mathbb{R}^L} + \mathcal{T}^{\eta|\alpha}_{P_\theta,H}$. We take the tangent spaces as given in this section; see section S1 in the supplementary material for a formal statement of what we require.

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

$$\tilde{\ell}_\theta := \dot{\ell}_\theta - \Pi_\theta \dot{\ell}_\theta \ , \tag{5}$$

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

$$\tilde{I}_\theta := \mathbb{E}_\theta \tilde{\ell}_\theta \tilde{\ell}_\theta' \ . \tag{6}$$

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). Building tests or estimators based on the efficient score function is attractive as efficiency results are well established, see Bickel et al. (1998).

2.1 Semiparametric identification robust score test

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

$$\mathbb{E}_0 \tilde{\ell}_{\theta_0} = \left[\mathbb{E}_0 \dot{\ell}_{\theta_0} - \mathbb{E}_0 \Pi_{\theta_0} \dot{\ell}_{\theta_0} \right] = 0 \ , \tag{7}$$

since $\mathbb{E}_0 \dot{\ell}_{\theta_0} = 0$ as the l-th element of $\dot{\ell}_{\theta_0}$ is a score function for the submodel $P_{\alpha + t e_l}$ with e_l the l-th canonical basis vector in $\mathbb{R}^L$. Moreover, each component of the vector $\Pi_{\theta_0} \dot{\ell}_{\theta_0}$ is an element of $\operatorname{cl} \mathcal{T}^{\eta|\alpha_0}_{P_{\theta_0}, H}$ and hence $\mathbb{E}_0 \Pi_{\theta_0} \dot{\ell}_{\theta_0} = 0$. This implies that (7) defines a set of L moment conditions based on which we can construct hypothesis tests. See e.g. Stock and Wright (2000), Kleibergen (2005) for related approaches with finite dimensional nuisance parameters. Unlike these papers, the nuisance parameter in our model is not a Euclidean parameter but rather an infinite dimensional object.[3]

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 are such that they satisfy the following high level assumption.

Assumption 1 (Non-Singular). *Under the null hypothesis* (3) *we have that*

1. $\frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\ell}_{\theta_0}(Y_i) \rightsquigarrow Z$, *where* $Z \sim \mathcal{N}(0, \tilde{I}_{\theta_0})$ *and* $\operatorname{rank}(\tilde{I}_{\theta_0}) = L$;
2. *We have an array of estimates* $\{\hat{\ell}_{\theta_0,n}(Y_i)\}_{n \geq 1, i \leq n}$ *such that:*
$$\frac{1}{n} \sum_{i=1}^n \left(\hat{\ell}_{\theta_0,n}(Y_i) - \tilde{\ell}_{\theta_0}(Y_i) \right) = o_P(n^{-1/2}) \ ;$$
3. $\hat{I}_{\theta_0,n} \xrightarrow{P} \tilde{I}_{\theta_0}$ *for some sequence of estimates* $\{\hat{I}_{\theta_0,n}\}$.

Clearly, Assumption 1 is high level and should be verified for any specific model of the form (2). Nevertheless, the strategy for verifying the different parts of the assumption is similar. In particular, part 1 amounts to verifying a central limit theorem for the efficient

[3] Andrews and Mikusheva (2016) also consider robust testing with an infinite dimensional nuisance parameter. Their approach is quite different to ours, focussing on models defined by moment conditions rather than with a full specification of the probability law.

score function, which given the iid assumption requires the existence of sufficient moments.[4] Part 2 imposes that we should be able to construct a sequence of estimates for the efficient score functions, which in practice amounts to being able to estimate η or a functional thereof sufficiently accurately. The third part imposes that the efficient information matrix can be consistently estimated. The assumption $\text{rank}(\tilde{I}_{\theta_0}) = L$ is often too strong for models in which α is not identified and we will relax this assumption in the next section.

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

To test the null hypothesis (3) we consider the efficient score statistic given by

$$\hat{S}_n = \left(\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \hat{\ell}_{\theta_0,n}(Y_i) \right)' \hat{I}_{\theta_0,n}^{-1} \left(\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \hat{\ell}_{\theta_0,n}(Y_i) \right) \ . \tag{8}$$

For parametric models this score statistic reduces to Neyman's $C(\alpha)$ statistic, which is asymptotically equivalent to Rao's score statistic (e.g. Kocherlakota and Kocherlakota, 1991).

The limiting distribution of $\hat{S}_n$ is summarized in the following proposition.

Proposition 1. *Given assumption 1-(Non-Singular), under the null hypothesis* (3) *we have that*

$$\hat{S}_n \rightsquigarrow \chi^2_L \ .$$

All proofs are provided in Appendix A. The proposition implies that, regardless of whether α is well identified, the score static $\hat{S}_n$ has a standard χ^2 limiting distribution. Confidence regions for α can be obtained by inverting $\hat{S}_n$ over a grid of values for α. By construction such confidence regions will have correct coverage.

Choi, Hall and Schick (1996) show that tests based on $\hat{S}_n$ are asymptotically uniformly most powerful within the class of rotation invariant tests (when $L = 1$ the rotational invariance can be dropped). This implies that asymptotically when testing $H_0 : \alpha = \alpha_0\ ,\ \eta \in \mathcal{H}$ against $H_1 : \alpha \neq \alpha_0\ ,\ \eta \in \mathcal{H}$, the power of the test is as good as if η would be known. This makes tests based on $\hat{S}_n$ attractive for scenarios where there is no explicit direction in which one want to maximize power. When such directions are given alternative test statistics, also based on the efficient score function, can be considered (e.g. Bickel, Ritov and Stoker, 2006).

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

2.2 Semiparametric identification and singularity robust score test

In semiparametric models, the efficient information matrix $\tilde{I}_{\theta_0}$ will often be singular at points of (local-)identification failure, see also Andrews and Guggenberger (2019) for many examples in parametric models. This occurs for instance, as we show in the supplementary material, in the ICA model (1) when more than one of the densities of ϵ are exactly Gaussian. In this section we modify the test statistic $\hat{S}_n$ to accommodate singular information matrices.

We first modify Assumption 1 accordingly.

Assumption 1 (Singular)**.** *Under the null hypothesis* (3) *we have that*

1. $\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\ell}_{\theta_0}(Y_i) \rightsquigarrow Z$*, where* $Z \sim \mathcal{N}(0, \tilde{I}_{\theta_0})$ *and* $0 < \operatorname{rank}(\tilde{I}_{\theta_0}) = r \leq L$*;*

2. *We have an array of estimates* $\{\hat{\ell}_{\theta_0,n}(Y_i)\}_{n\geq 1, i\leq n}$ *such that:*

$$\frac{1}{n}\sum_{i=1}^n \left(\hat{\ell}_{\theta_0,n}(Y_i) - \tilde{\ell}_{\theta_0}(Y_i)\right) = o_P(n^{-1/2}) \ ;$$

3. $P\left(\|\hat{I}_{\theta_0,n} - \tilde{I}_{\theta_0}\|_2 < \nu_n\right) \to 1$ *for some sequence of estimates* $\{\hat{I}_{\theta_0,n}\}$ *and some positive sequence* $\nu_n \to 0$ *.*

This modified assumption allows the limiting distribution of the re-scaled sum of efficient scores to have a singular variance matrix. Part 3 imposes that there exists a decreasing sequence ν_n such that the distance between $\hat{I}_{\theta_0,n}$ and $\tilde{I}_{\theta_0}$ is upper bounded with probability tending to one by ν_n as $n \to \infty$. For example, if we have that $\|\hat{I}_{\theta_0,n} - \tilde{I}_{\theta_0}\|_2 = O_P(n^{-1/2})$ we could take $\nu_n = \log(n)/\sqrt{n}$.

Given ν_n we define a truncated eigenvalue version of the variance matrix estimate as

$$\hat{I}^t_{\theta_0,n} = \hat{U}_n \hat{\Lambda}_n(\nu_n)\hat{U}_n' \ , \tag{9}$$

where $\hat{\Lambda}_n(\nu_n)$ is a diagonal matrix with the ν_n-truncated eigenvalues of $\hat{I}_{\theta_0,n}$ 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 $\hat{I}_{\theta_0,n}$, then the (i,i)th element of $\hat{\Lambda}_n(\nu_n)$ is given by $\hat{\lambda}_{n,i}\mathbf{1}(\hat{\lambda}_{n,i} \geq \nu_n)$.

Given the truncated variance matrix estimate, we define the singularity robust score test statistic as

$$\hat{S}_n^{SR} = \left(\frac{1}{\sqrt{n}}\sum_{t=1}^n \hat{\ell}_{\theta_0,n}(Y_i)\right)' \hat{I}^{t,\dagger}_{\theta_0,n}\left(\frac{1}{\sqrt{n}}\sum_{t=1}^n \hat{\ell}_{\theta_0,n}(Y_i)\right) \ , \tag{10}$$

where $\hat{I}^{t,\dagger}_{\theta_0,n}$ is the Moore-Penrose psuedo-inverse of $\hat{I}^t_{\theta_0,n}$. The limiting distribution of $\hat{S}_n^{SR}$ is characterized in the following proposition.

Proposition 2. *Given assumption 1-(Singular), let $r_n = \text{rank}(\hat{I}^t_{\theta_0,n})$ where $\hat{I}^t_{\theta_0,n}$ is defined in (9) 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 we have that*

$$P(\hat{S}^{SR}_n > c_n) \to a\ .$$

The proposition implies that we can use the estimated rank of $\hat{I}^t_{\theta_0,n}$ to compute the critical value for $\hat{S}^{SR}_n$. In practice, we can set ν_n to an arbitrarily small number greater then zero. In this case the Moore-Penrose psuedo-inverse of $\hat{I}^t_{\theta_0,n}$ becomes equal to that of $\hat{I}_{\theta_0,n}$. Hence, the role of ν_n is merely to facilitate the construction of the proof.

A reassuring result is that under the given assumptions we have that if $r = L$, then $\hat{S}^{SR}_n = \hat{S}_n + o_p(1)$; see lemma 2 in the Appendix. Therefore the singularity robust score statistic $\hat{S}^{SR}_n$ can be adopted for all cases: singular or non-singular variance matrix. Moreover, for the case where $r = L$ the optimality properties of $\hat{S}_n$ carry over to $\hat{S}^{SR}_n$.

The singularity and identification robust test statistic $\hat{S}^{SR}_n$ is broadly applicable for the class of semiparametric models we consider. The key difficulty for its application lies in the construction and estimation of the efficient score function. For this no general recipe exists but guidance and examples are given in Rabinowitz (2000).

3 Robust non-Gaussian inference

In this section we provide the details for implementing the high level framework from the previous section for conducting inference on A in the ICA model and extensions thereof. For convenience we restate the ICA model

$$Y = A^{-1}\epsilon\ . \tag{11}$$

We start by casting model (11) as a semiparametric model as defined in general in equation (2), see also Amari and Cardoso (1997) and Chen and Bickel (2006). Then, we provide the details for the estimation of the efficient score function and the singularity robust score test. Finally, we provide extensions of model (11) that allow for the inclusion of covariates and dynamics, thus covering a broader class of simultaneous equations models and vector autoregressive models.

3.1 Semi-parametric ICA model

In the ICA model the finite dimensional parameters of interest are the parameters that determine A, whereas the nuisance parameters are the unknown density functions of the components of ϵ.

Let $\alpha \in \mathcal{A} \subset \mathbb{R}^L$ be the parameter controlling $A = A(\alpha)$. For example, if A lies in the set of invertible matrices and is otherwise unconstrained, we can take $\alpha = \text{vec}(A)$. If A is an orthogonal matrix, it can be parameterized by $\alpha \in \mathbb{R}^L$ with $L = K(K-1)/2$ using the trigonometric transform or the Cayley transform of a skew-symmetric matrix (e.g. Gouriéroux, Monfort and Renne, 2017). We leave the precise parameter mapping unspecified in our theoretical work and simply assume that $A(\alpha)$ is continuously differentiable with respect to α.

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

Assumption 2. *For $\epsilon = (\epsilon_1, \ldots, \epsilon_K)'$ in model (11), each component ϵ_k has a continuously differentiable root density (where the density is with respect to Lebesgue measure on $\mathbb{R}$). We write the density as η_k with log density score $\phi_k(x) = \partial \log \eta_k(x)/\partial x$. We assume that for all $k = 1, \ldots, K$*

1. *$\mathbb{E}\epsilon_k = 0$, $\mathbb{E}\epsilon_k^2 = 1$, $\mathbb{E}\epsilon_k^{4+\delta} < \infty$, $\mathbb{E}(\epsilon_k^4) - 1 > \mathbb{E}(\epsilon_k^3)^2$, and $\mathbb{E}\phi_k^{4+\delta}(\epsilon_k) < \infty$ (for some $\delta > 0$);*

2. *$\mathbb{E}\phi_k(\epsilon_{k,i}) = 0$, $\mathbb{E}\phi_k(\epsilon_{k,i})\epsilon_{k,i} = -1$, $\mathbb{E}\phi_k(\epsilon_{k,i})\epsilon_{k,i}^2 = 0$ and $\mathbb{E}\phi_k(\epsilon_{k,i})\epsilon_{k,i}^3 = -3$;*

3. *ϵ_k is independent of ϵ_j for all $k \neq j$.*

The first part normalizes the errors to have mean zero, variance one and finite four+δ moments.[5] The second part simplifies the construction of the efficient score functions. Whilst this may at first glance appear a strong condition, lemma S11 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.[6] When A is restricted to be an orthogonal matrix the second part can be dropped entirely.

Most important is what is *not* in Assumption 2: there is no condition that imposes that a certain number of components of ϵ have a non-Gaussian distribution. This is precisely the way in which we deviate from the ICA literature and its extensions where such assumptions are commonly imposed. The benefit, as we will see below, is that our testing approach

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

[6] 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.

retains correct size regardless of the true distributions of ϵ, e.g. regardless of the distance to the Gaussian distribution.

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

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

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

$$p_\theta(y) := |\det A(\alpha)| \prod_{k=1}^K \eta_k(A_{k\bullet} y) , \qquad (12)$$

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

Let $\mathcal{H}_0 \subset \mathcal{H}$ denote the set with elements $\eta = (\eta_1, \ldots, \eta_K)$ such that each η_k satisfies the requirements imposed by assumption 2. To implement the score test we first characterize the efficient score function (5) in terms of estimable quantities. The following lemma provides the key result.[7]

Lemma 1. *The components of the efficient score function* (5) *for the semiparametric ICA model $\mathcal{P}_\Theta$ at any $\theta = (\alpha, \eta)$ with $\eta \in \mathcal{H}_0$ are given by, for $l = 1, \ldots, L$,*

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

where $\zeta_{l,k,j} := [D_l(\alpha)]_{k\bullet} A^{-1}_{\bullet j}$ with $D_l(\alpha) = \partial A(\alpha)/\partial \alpha_l$ and $\tau_k := (\tau_{1,k}, \tau_{2,k})'$ is defined as

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

[7] Strictly speaking, the efficient score function is defined relative to a specific tangent set, denoted here by $\mathcal{T}^{\eta|\alpha}_{P_\theta,H}$; see e.g. the discussion in sections 1.2 and 2.2 of van der Vaart (2002). The proof of lemma 1 in the supplementary material (see section S2) provides details on the specific tangent set we consider.

The proof of Lemma 1 follows similarly as in Amari and Cardoso (1997) and can be found in the supplementary material. It requires first defining the tangent spaces for α and η, and then computing the orthogonal projection of the scores for α on the tangent space for η, see equation (5). The main difference with respect to Amari and Cardoso (1997) is that we allow $A(\alpha)$ to be parametrized in an arbitrary (smooth) way. This can be useful from both a theoretical and computational perspective if the researcher has prior information relating to restrictions on A. Additionally, it permits more flexibility in specifying hypotheses about A through α.

3.2 Non-Gaussianity robust score test

Next, to conduct inference on A we consider testing $H_0 : \alpha = \alpha_0\ ,\ \eta \in \mathcal{H}_0$ using the singularity robust score statistic $\hat{S}_n^{SR}$ given in (10). To compute $\hat{S}_n^{SR}$ we require an estimate for the efficient score function $\tilde{\ell}_{\theta_0}$ as defined in Lemma 1. This implies that for each $k = 1, \ldots, K$, we need to estimate τ_k and the log density scores ϕ_k. Note that the remaining elements of the efficient score are fixed under H_0. To estimate $\tilde{\ell}_{\theta_0}$ we assume that we have available a sample of n independent and identically distributed copies of Y that are denoted by $\{Y_i, i = 1, \ldots, n\}$.

The estimation of τ_k follows easily by replacing the population moments in its definition by their sample counterparts. In particular, we have

$$\hat{\tau}_{k,n} := \hat{M}_{k,n}^{-1}\begin{pmatrix} 0 \\ -2 \end{pmatrix}, \quad \text{where } \hat{M}_{k,n} := \begin{pmatrix} 1 & \frac{1}{n}\sum_{i=1}^n (A_{k\bullet}Y_i)^3 \\ \frac{1}{n}\sum_{i=1}^n (A_{k\bullet}Y_i)^3 & \frac{1}{n}\sum_{i=1}^n (A_{k\bullet}Y_i)^4 - 1 \end{pmatrix}. \tag{13}$$

The estimation of the log density scores is typically more involved and a variety of options exist. We proceed by stating the requirements that must hold for any density score estimator and we show in Appendix B (see Proposition 3) that the method of Chen and Bickel (2006), who build on Jin (1992), satisfies the requirements under mild conditions. This approach is convenient for two reasons: first the method of Chen and Bickel (2006) is based on B-spline approximations and while easy to implement it is notationally somewhat cumbersome, second different researchers might prefer to use a different density score estimator.

Assumption 3. *We have an array of estimates $\{\hat{\phi}_{k,n}(A_{k\bullet}Y_i)\}_{n\geq 1, i\leq n}$ for $k = 1, \ldots, K$ such that, under the null $H_0 : \alpha = \alpha_0\ ,\ \eta \in \mathcal{H}_0$, for each $k \neq j$*

$$\frac{1}{n}\sum_{i=1}^n \left[\hat{\phi}_{k,n}(A_{k\bullet}Y_i) - \phi_k(A_{k\bullet}Y_i)\right] A_{j\bullet}Y_i = o_P(n^{-1/2}), \tag{14}$$

and for $\nu_n = \nu_{n,p}^2$ *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* $p = 2$*, we have*

$$\frac{1}{n}\sum_{i=1}^{n}\left(\left[\hat{\phi}_{k,n}(A_{k\bullet}Y_i) - \phi_k(A_{k\bullet}Y_i)\right] A_{j\bullet}Y_i\right)^2 = o_P(\nu_n). \tag{15}$$

The assumption effectively requires a specific functional, e.g. ϕ_k, of the nuisance parameters η_k to be estimable sufficiently accurately. The rate ν_n is now made explicit and it is split into two parts. The "slow" rate $n^{(1-p)/p}$ (for $p \in (1,2)$) is always sufficient given assumption 2, but if ϵ_k has finite eighth moments the faster rate applies. The method of Chen and Bickel (2006) satisfies assumption 3 under assumption 2 and mild assumptions on the densities η_k, see proposition 3 in Appendix B.

Given the estimates for τ_k and the density scores ϕ_k we can estimate the efficient score function under H_0 by

$$\hat{\ell}_{\theta_0,n,l}(y) = \sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K}\zeta_{l,k,j}\hat{\phi}_k(A_{k\bullet}y)A_{j\bullet}y + \sum_{k=1}^{K}\zeta_{l,k,k}\left[\hat{\tau}_{k,1}A_{k\bullet}y + \hat{\tau}_{k,2}\kappa(A_{k\bullet}y)\right] , \tag{16}$$

where, compared to Lemma 1, τ_k and ϕ_k have been replaced by their estimates. Similarly to above we can define the efficient information matrix estimate and its eigenvalue truncated version

$$\hat{I}_{\theta_0,n} = \frac{1}{n}\sum_{i=1}^{n}\hat{\ell}_{\theta_0,n}(Y_i)\hat{\ell}_{\theta_0,n}(Y_i)' , \qquad \hat{I}^t_{\theta_0,n} = \hat{U}_n\hat{\Lambda}(\nu_n)\hat{U}_n' , \tag{17}$$

where $\hat{U}_n$ and $\hat{\Lambda}(\nu_n)$ are defined similarly as in equation (9) with ν_n as in assumption 3.

We now state our main result.

Theorem 1. *Consider the statistic*

$$\hat{S}_n^{SR} = \left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\hat{\ell}_{\theta_0,n}(Y_i)\right)' \hat{I}^{t,\dagger}_{\theta_0,n}\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\hat{\ell}_{\theta_0,n}(Y_i)\right) ,$$

with $\hat{\ell}_{\theta_0,n}(Y_i)$ *defined according to* (16) *and* $\hat{I}^{t,\dagger}_{\theta_0,n}$ *is the Moore-Penrose inverse of* $\hat{I}^t_{\theta_0,n}$ *defined in* (17)*. Given assumptions 2 and 3, let* $r_n = \operatorname{rank}(\hat{I}^t_{\theta_0,n})$ *and denote by* c_n *the* $1-a$ *quantile of the* $\chi^2_{r_n}$ *distribution, for any* $a \in (0,1)$*. Then, under* H_0 *we have that*

$$\mathrm{P}(\hat{S}_n^{SR} > c_n) \to a\ .$$

The proof of Theorem 1 amounts to verifying the three high level conditions stated in Assumption 1-(Singular) so that we can apply proposition 2.

Some comments are in order. First, if we would impose that $\operatorname{rank}(\tilde{I}_{\theta_0}) = L$, e.g. the efficient information matrix is non-singular, then we could use the standard score statistic $\hat{S}_n$ as defined in equation (8). Here it holds again that $\hat{S}_n^{SR} = \hat{S}_n + o_P(1)$. Second, if $\operatorname{rank}(\tilde{I}_{\theta_0}) = L$ the score statistic $\hat{S}_n^{SR}$ is uniformly most powerful in the class of invariant tests (e.g. Choi, Hall and Schick, 1996). Third, the threshold ν_n is important for deriving the distribution of the test statistic, but in practice the size and power seem not to be affected as long as ν_n is chosen to be small. Fourth, $\hat{S}_n^{SR}$ is almost trivial to compute as it requires only K regressions to obtain the density score estimates using B-splines, thus avoiding numerical optimization routines entirely.

3.3 Extensions for covariates and dynamics

As argued in the introduction, from an economics perspective the ICA model (11) is best viewed as a building block in a larger simultaneous equations model (e.g. Gouriéroux, Monfort and Renne, 2017, 2019). Motivated by such examples this section extends the semiparametric robust score test for a general class of models where realizations of Y are not observable, but a sequence of estimates of those realizations can be constructed. The testing approach is similar as for the baseline ICA model, but the estimation noise – stemming from estimating Y – is generally non-negligible and will require an adjustment to the variance of the estimate of the efficient score function. Here we restrict ourselves to linear simultaneous equations models, including SVAR models, but in the supplementary material we illustrate how a broader class of potentially nonlinear models, which include $Y = A\epsilon$ as a component, can be handled.

In this setting, we observe realizations from $W_i = (Z_i, M_i)$ that are generated by the model

$$Z_i = BM_i + Y_i \ , \qquad Y_i = A^{-1}\epsilon_i \ , \tag{18}$$

where M_i a random vector of explanatory variables in $\mathbb{R}^d$, B a $K \times d$ coefficient matrix. Let $\hat{B}_n$ the OLS estimator for B and we define $\hat{Y}_{i,n} \coloneqq Z_i - \hat{B}_n M_i = Y_i + U_n M_i$ for $U_n = B - \hat{B}_n$, as the residuals.[8]

To formalize the approach we impose the following assumptions.

Assumption 4. *Let $Y = A^{-1}\epsilon$, where ϵ satisfies assumption 2, and we have estimates $\hat{Y}_{i,n} = Y_i + U_n M_i$, for $i = 1, \dots, n$, where $\{Y_i, i = 1, \dots, n\}$ are independent copies of Y, $U_n = B - \hat{B}_n$ and the process $\{M_i\}$ in model (18) is such that*

[8] The OLS estimator is chosen for convenience, but any $\sqrt{n}$-consistent estimator can be used.

1. *$\{M_i\}$ is a weakly stationary process with $\mathbb{E}[M_i M_i'] \succ 0$ and absolutely summable autocovariances,*

2. *$\sup_{i\in\mathbb{N}} \mathbb{E}|M_{s,i}|^r < \infty$ for all $s \in [d]$, for some $r \geq 4$ and the strong mixing coefficients of $\{M_i\}$, $\{\alpha_h : h \geq 0\}$, satisfy $\sum_{h=0}^{\infty} \alpha_h^{\frac{r-4}{2r}} < \infty$.*

3. *ϵ_i is independent of $(M_i, \mathcal{F}_{i-1})$, where $\mathcal{F}_i \coloneqq \sigma(\{M_j, \epsilon_j : j \leq i\})$;*

The assumption allows for weak dependence in the process $\{M_i\}$. We require that $\{M_i\}$ is strong mixing with strong mixing coefficients that decay sufficiently fast.[9] In the independent case, $\alpha_h = 0$ for all $h \geq 1$ and hence we can take $r = 4$, in which case we require finite fourth moments. In the dependent case, we require $r > 4$ and there is a (non-degenerate) trade off between the existence of moments and the level of dependence permitted. Part 3 imposes that the structural shocks ϵ_i are independent of the "current" M_i and all "past" M_i and ϵ_i. This is stronger than necessary and could be replaced by assuming that a number of specific sequences are martingale differences with respect to the filtration $\mathcal{F}_i$ and some conditional homoskedasticity type assumptions; for ease of presentation we maintain the stronger condition.

With these assumptions in place we construct the score test statistic. As we no longer observe Y_i, we need to adjust our estimator of the efficient score function under H_0. In place of (16), we will use

$$\hat{\ell}_{\theta_0,n,l}(W_i) = \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \zeta_{l,k,j} \hat{\phi}_k(A_{k\bullet}\hat{Y}_{i,n}) A_{j\bullet}\hat{Y}_{i,n} + \sum_{k=1}^{K} \zeta_{l,k,k} \left[\hat{\tau}_{k,1} A_{k\bullet}\hat{Y}_{i,n} + \hat{\tau}_{k,2} \kappa(A_{k\bullet}\hat{Y}_{i,n}) \right] , \tag{19}$$

where $\hat{\tau}_k$ and $\hat{\phi}_k$ are estimates based on the estimated $\hat{Y}_{i,n}$. Both are defined analogously to the previous section, with Y_i replaced by $\hat{Y}_{i,n}$.

To ensure that the log density scores can be consistently estimated we impose the following assumption, mirroring Assumption 3 for the baseline ICA model.

Assumption 5. *We have an array of estimates $\left\{\hat{\phi}_{k,n}\left(A_{k\bullet}Y_i\right)\right\}_{n\geq 1, i\leq n}$ for $k = 1, \ldots, K$ such that, under the null $H_0 : \alpha = \alpha_0$, $\eta \in \mathcal{H}_0$, for each $k \neq j$*

$$\frac{1}{n}\sum_{i=1}^{n} \left[\hat{\phi}_{k,n}\left(A_{k\bullet}\hat{Y}_{i,n}\right) A_{j\bullet}\hat{Y}_{i,n} - \phi_k(A_{k\bullet}Y_i) A_{j\bullet}Y_i \right] = o_P(n^{-1/2}), \tag{20}$$

and for $\nu_n = \nu_{n,p}^2$ with $p \coloneqq \min\{1 + \delta/4, 2\}$ and $\nu_{n,p} = n^{(1-p)/p}$ if $p \in (1, 2)$ or $\nu_{n,p} =$

[9] What constitutes "sufficiently" fast depends on the (non-)existence of higher moments.

$n^{-1/2}\log(n)^{1/2+\rho}$, *for some* $\rho > 0$, *if* $p = 2$, *we have*

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}\left(A_{k\bullet}\hat{Y}_{i,n}\right)A_{j\bullet}\hat{Y}_{i,n} - \phi_k(A_{k\bullet}Y_i)A_{j\bullet}Y_i\right]^2 = o_P(\nu_n). \tag{21}$$

Proposition 4 in Appendix B demonstrates that this assumption holds for the log density score estimator that is considered in Chen and Bickel (2006), provided assumption 4 and some regularity conditions hold.

To express the form of the variance of our efficient score estimator in this model we require some additional notation. Let $\psi_{\theta_0} = (\psi'_{\theta_0,1}, \ldots, \psi'_{\theta_0,L})'$ and its estimator $\hat{\psi}_{\theta_0,n} = (\hat{\psi}'_{\theta_0,n,1}, \ldots, \hat{\psi}'_{\theta_0,n,L})'$, have components defined by (for $l \in [L]$)

$$\psi_{\theta_0,l}(W_i) \coloneqq M_i\left(\sum_{k=1}^{K}\zeta_{l,k,k}\tau_{k,1}A_kY_i\right), \quad \hat{\psi}_{\theta_0,n,l}(W_i) \coloneqq M_i\left(\sum_{k=1}^{K}\zeta_{l,k,k}\hat{\tau}_{k,n,1}A_k\hat{Y}_{i,n}\right). \tag{22}$$

Now, let $\varphi \coloneqq (\tilde{\ell}'_{\theta_0}, \psi'_{\theta_0})'$ and $\hat{\varphi}_n \coloneqq (\hat{\ell}'_{\theta_0,n}, \hat{\psi}'_{\theta_0,n})'$ and define $V_{\theta_0} \coloneqq \mathbb{E}\varphi\varphi'$ and its sample analogue $\hat{V}_n \coloneqq \frac{1}{n}\sum_{i=1}^{n}\hat{\varphi}_n(W_i)\hat{\varphi}_n(W_i)'$. Finally, let $Q \coloneqq \mathbb{E}\left[M_iM_i'\right]^{-1}\mathbb{E}M_i$, $\hat{Q}_n \coloneqq \left[\frac{1}{n}\sum_{i=1}^{n}M_iM_i'\right]^{-1}\frac{1}{n}\sum_{i=1}^{n}M_i$, $\hat{R}_n' \coloneqq (I_L, -(I_L \otimes \hat{Q}_n)')$, $R' \coloneqq (I_L, -(I_L \otimes Q)')$ and define

$$J_{\theta_0} \coloneqq R'V_{\theta_0}R, \quad \text{and } \hat{J}_n \coloneqq \hat{R}_n'\hat{V}_n\hat{R}_n\ . \tag{23}$$

The eigenvalue truncated version of $\hat{J}_n$ is given by

$$\hat{J}_n^t = \hat{U}_n\hat{\Lambda}(\nu_n)\hat{U}_n'\ , \tag{24}$$

where $\hat{U}_n$ and $\hat{\Lambda}(\nu_n)$ are defined similarly to as in equation (9) with ν_n as in assumption 5.

Theorem 2. *Suppose model (18) holds and consider the statistic*

$$\hat{S}_n^{SR} = \left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\hat{\ell}_{\theta_0,n}(W_i)\right)'\hat{J}_n^{t,\dagger}\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\hat{\ell}_{\theta_0,n}(W_i)\right)\ ,$$

with $\hat{\ell}_{\theta_0,n}(W_i)$ *defined in (19) and* $\hat{J}_n^{t,\dagger}$ *is the Moore-Penrose inverse of* $\hat{J}_n^t$ *defined as in (24) where* $\hat{J}_n$ *is defined as in (23). Given assumptions 4 and 5, let* $r_n = \operatorname{rank}(\hat{J}_n^t)$ *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 *we have that*

$$\mathrm{P}(\hat{S}_n^{SR} > c_n) \to a\ .$$

The theorem shows that we can continue to use the singularity adjusted score statistic in cases where Y_i is not observable. The only consequence of estimating Y_i is that the variance of our estimate of the efficient score function needs to be adjusted.

For concreteness we provide two empirically relevant examples of models to which the preceding proposition can be applied.[10] We first consider a static model with i.i.d. observations which can be considered as a simultaneous equations model with additional (exogenous) explanatory covariates. Second we introduce dynamics and consider a SVAR model.

Example 1. *Suppose that model (18) holds and ϵ_i satisfies the requirements of assumption 2 and is independent of $(M_i, \epsilon_{i-1}, M_{i-1}, \ldots, \epsilon_1, M_1)$. Additionally suppose that $\{M_i\}_{i\in\mathbb{N}}$ is i.i.d., has finite fourth moments and $\mathbb{E}[M_i M_i']$ is full-rank. If assumption 5 holds, then proposition 2 applies.*

Example 2. *Suppose that*

$$Z_i = C + \Phi_1 Z_{i-1} + \cdots + \Phi_q Z_{i-q} + Y_i, \quad Y_i = A^{-1}\epsilon_i, \tag{25}$$

where $C \in \mathbb{R}^K$ and each Φ_m is a $K \times K$ matrix such that $\det \Phi(z) \neq 0$ for all $z \in \mathbb{C}$ with $|z| \leq 1$.[11] This model can be put into the form required by equation (18) as

$$Z_i = BM_i + Y_i, \text{ with } B \coloneqq \begin{bmatrix} C & \Phi_1 & \cdots & \Phi_q \end{bmatrix}, \quad M_i \coloneqq \begin{bmatrix} 1 & Z_{i-1}' & \cdots & Z_{i-q}' \end{bmatrix}'. \tag{26}$$

Additionally suppose that the covariance matrix of $(Z_i' - \mu', \ldots, Z_{i+q-1}' - \mu')'$ is positive definite where $\mu \coloneqq \mathbb{E}[Z_i]$ and $\{\epsilon_i\}_{i\in\mathbb{N}}$ are i.i.d. and satisfy assumption 2. If assumption 5 holds, then proposition 2 applies.

4 Simulation results

In this section we study the finite sample properties of the singularity and identification robust score test. We study the size and power of the test under different data generating processes and compare its performance to several alternatives that have been proposed in the literature. We first study the baseline ICA model (11) after which we consider the structural VAR model discussed in example 2.

4.1 Baseline ICA model

We start by drawing independent samples from the ICA model (11) for dimensions $K = 2$ and $K = 3$ and sample sizes $n = 200, 500$. We fix ϵ_1 to have a standard Gaussian density and

[10] The appendix contains proofs of the application of proposition 2 to both examples.

[11] Here $\Phi(z)$ is defined as the matrix-valued polynomial $\Phi(z) \coloneqq I - \Phi_1 z - \cdots - \Phi_q z^q$.

consider different densities for ϵ_k, with $k = 2, \ldots, K$, that range from standard Gaussian to skewed bi-modal distributions. The non-Gaussian densities are either Student's t or mixtures of normals taken from Marron and Wand (1992). Table 1 provides an overview.

The matrix of interest $A(\alpha)$ is parameterized as a rotation matrix using the trigonometric transformation, with the $L \times 1$ vector α denoting the coefficients.[12] Similar results are obtained for $\alpha = \text{vec}(A)$ which we consider for the SVAR simulations below, but the trigonometric transformation conveniently reduces the dimension of α, which is useful for studying power comparisons among different approaches for the case where $K = 2$ where L becomes equal to one.

For each specification we simulate $S = 5.000$ datasets and for each we compute the singularity robust score statistic as defined in Theorem 1 using the log density score estimator of Jin (1992) and Chen and Bickel (2006) as discussed in Appendix B using $B = 4, 6$ or 8 cubic splines, with the upper and lower endpoints taken to be the 95th and 5th percentile of the samples adjusted respectively up and down by $\log(\log n)$.[13] We threshold the information matrix estimate at machine precision for ν_n for all simulations.

Size results

In Table 2 we show the empirical rejection frequencies corresponding to the $\hat{S}_n^{SR}$ 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 $\alpha \in \mathbb{R}^L$, e.g. α is unidentified. Nonetheless, we find that the empirical rejection frequency of the score test is always close to the nominal size. This holds regardless of the sample size n, the dimension of the ICA model K and the number of cubic splines B.

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

Third, overall the number of cubic splines used has little influence on the results. A close

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

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

with a scalar parameter α.

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

inspection reveals that when the number of cubic splines is equal to four the test becomes mildly conservative for some densities, therefore we use $B = 6$ cubic splines in the remaining exercises.

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

Comparison to alternative approaches

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

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

Second, we consider the psuedo-maximum likelihood tests developed by Gouriéroux, Monfort and Renne (2017). These tests are asymptotically valid for a broader range of true distribution functions and amount to fixing the functional form of the likelihood. We follow their implementation and choose the Students t density with five degrees of freedom as the pseudo-likelihood and compute the likelihood ratio statistic based on this density.[14]

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

Finally, several papers suggest using pre-testing procedures to determine whether the shocks are indeed non-Gaussian. To this extent, we implemented all tests conditional on rejecting the Jarque-Berra test for Gaussianity. For the Gaussian specification (e.g. density 1 in Table 1) this was not feasible as the test was almost never rejected and reaching $S = 5.000$

[14] The Wald statistic suggested in Gouriéroux, Monfort and Renne (2017) was found to be over-sized for all specifications.

accepted samples became computationally too expensive, hence this case was omitted.

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

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

In sum, non of the alternative methods appear to control size under either (i) weakly non-Gaussian densities, or (b) mis-specification of the likelihood.

Power results

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

Figure 1 shows the empirical rejection frequencies when we vary α around the, arbitrarily chosen, true value $\alpha = \pi/4$. Each point on the curve is based on $S = 5.000$ simulations and for clarity of the figure we adjusted the power of the parametric score test such that it is size correct, e.g. exactly 0.05 for $\alpha = \alpha_0$, in all specifications.

We find that the power of the parametric score test is larger when compared to the semi-parametric test when the density is correctly specified. This is the top row of Figure 1 where we consider the students t density as the truth. Nonetheless the S_n^{SR} test comes quite close in terms of power.

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

4.2 Structural VAR model

Next, we evaluate the finite sample size and power of the score test in the extended ICA model following the methodology developed in Section 3.3. We focus on the structural VAR model of example 2 as it corresponds to our empirical study below.

We simulate data from the SVAR model (25) with $q = 1, 2, 4$ lags, $K = 2, 3$, $\epsilon_{i,1} \sim N(0, 1)$, $\epsilon_{i,k}$, with $k \geq 2$, selected from Table 1 and $n = 200, 500, 1000$. The vector of interest is specified as $\alpha = \text{vec}(A)$ and hence $L = 4, 9$ in our setting. Experiments with alternative choices led to similar results.

The empirical rejection frequencies for the robust score test $\hat{S}_n^{SR}$ of Theorem 2 are shown in Table 4. We find that for small samples there is some over-rejection, notably for the heavy-tailed densities ($t(5)$ and the outlier density). When the sample size increases the rejection frequencies become close to the nominal size of the test.

5 Short run labor elasticities

In this section we present the results from an empirical study that we conducted to study supply and demand elasticities in the US labor market. The specification was taken from Baumeister and Hamilton (2015) and was recently revisited by Lanne and Luoto (2019*b*) who criticize some of the restrictions considered in Baumeister and Hamilton (2015) and suggest to remove them by using an identification approach that exploits non-Gaussianity. However, as shown in their and our simulations the adopted GMM approach is not robust to weakly non-Gaussian densities.

To this extent, our objective is similar as in Lanne and Luoto (2019*b*) as we aim to relax the prior specifications in Baumeister and Hamilton (2015) by relying on non-Gaussian identification, but in contrast to Lanne and Luoto (2019*b*) we use the semi-parametric score statistic to conduct inference, which is robust to weakly non-Gaussian distributions.

The bi-variate SVAR model is defined for $Z_i = (\Delta w_i, \Delta n_i)$, where Δn_i is the growth rate of total U.S. employment and Δw_i is the growth rate of real compensation per hour. The quarterly data sample is from 1970:Q1 until 2014-Q2 and the model specification of Baumeister and Hamilton (2015) is given by

$$D_0 Z_i = \tilde{C} + D_1 Z_{i-1} + \ldots + D_q Z_{i-q} + \Sigma^{1/2}\epsilon_i \ , \qquad D_0 = \begin{bmatrix} -\beta^d & 1 \\ -\beta^s & 1 \end{bmatrix} \ , \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

where β^d is the short-run wage elasticity of demand, and β^s is the short-run wage elasticity of supply.

This model can be rewritten in the notation of example 2 as follows

$$Z_i = C + \Phi_1 Z_{i-1} + \ldots + \Phi_q Z_{i-q} + A^{-1}\epsilon_i, \qquad A^{-1} = \begin{bmatrix} \frac{\sigma_1}{\beta^s - \beta^d} & -\frac{\sigma_2}{\beta^s - \beta^d} \\ \frac{\sigma_1 \beta^s}{\beta^s - \beta^d} & -\frac{\sigma_2 \beta^d}{\beta^s - \beta^d} \end{bmatrix}.$$

We compute the semi-parametric score test for different values of $\beta^s, \beta^d, \sigma_1, \sigma_2$ that satisfy the sign restrictions $\sigma_1 > 0$, $\sigma_2 > 0$, $\beta^s > 0$ and $\beta^d < 0$. We evaluate the score test using Theorem 2 and report the confidence region for β^s and β^d for each combination of parameters that satisfies $S_n^{SR} \leq c_n$, where c_n denotes the critical value of the $\chi^2_{r_n}$ distribution for $a = 0.05$ and $a = 0.33$.

The implied 67% and 95% confidence intervals for the demand and supply elasticities are shown in Figure 2. We find that the short run supply elasticity is reasonably well identified to the extent that the confidence regions exclude $\beta^s = 0$. In contrast, the demand elasticity is poorly identified using the non-Gaussian distribution, the confidence bands are wide and we cannot exclude zero. These findings contrast with results reported in Lanne and Luoto (2019*b*) who report considerably smaller confidence intervals.

6 Conclusion

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

The general framework was worked out in detail for a class of simultaneous equations models where the interest was in the mixing matrix A and the densities of the errors were treated as nuisance parameters. The mixing matrix A in this model class is identified (up to sign and permutation of its columns) if and only if at most one component is Gaussian. Existing approaches that exploit non-Gaussianity do not control size when the true densities are close to Gaussian. In contrast, we show both theoretically and in simulation that the semi-parametric score statistic is robust to this type of identification failure and controls size uniformly over a space of densities that satisfy mild moment conditions.

References

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

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

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

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

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

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

Baumeister, Christiane, and James D. Hamilton. 2015. "Sign Restrictions, Structural Vector Autoregressions, and Useful Prior Information." *Econometrica*, 83(5): 1963–1999.

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.*

Ben-Israel, A., and T. N. E. Greville. 2003. *Generalized Inverses: Theory and Applications.* New York, NY, USA:Springer.

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

Bonhomme, S., and J-M. Robin. 2009. "Consistent noisy independent component analysis." *Journal of Econometrics*, 149.

Brockwell, P. J., and R. A. Davis. 1991. *Time Series: Theory and Methods.* . 2 ed., New York, NY, USA:Springer.

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

Chen, Bin, Jinho Choi, and Juan Carlos Escanciano. 2017. “Testing for fundamental vector moving average representations.” *Quantitative Economics*, 8(1): 149–180.

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.

Common, P. 1994. “Independent component analysis, A new concept?” *Signal Processing*, 36.

Davidson, J. 1994. *Stochastic limit theory.* Oxford University Press.

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

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

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

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

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

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

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

Hamilton, J. D. 1994. *Time Series Analysis.* Princeton, NJ, USA:Princeton University Press.

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.

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

Hyvärinen, Aapo, Kun Zhang, Shohei Shimizu, and Patrik O. Hoyer. 2010. “Estimation of a Structural Vector Autoregression Model Using Non-Gaussianity.” *Journal of Machine Learning Research*, 11(56): 1709–1731.

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).

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

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

Lanne, Markku, and Jani Luoto. 2019*b*. "Useful Prior Information in Sign-Identified Structural Vector Autoregression: Replication of Baumeister and Hamilton (2015)." working paper.

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

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).

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

Mokkadem, A. 1988. "Mixing properties of ARMA processes." *Stochastic Processes and their Applications*, 29: 309–315.

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.

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

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

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

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

Sahneh, M. H. 2015. “Are the Shocks Obtained from SVAR Fundamental?” *Working paper.*

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

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

Stock, James H., and Motohiro Yogo. 2005. “Testing for Weak Instruments in Linear IV Regression.” *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, , ed. Donald W. K. Andrews and James H.Editors Stock, 80108. Cambridge University Press.

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

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

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.

Appendix A: Main proofs

Proof of Proposition 1. Let $Z_n \coloneqq \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\ell}_{\theta_0,n}(Y_i)$. This can be rewritten as

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\ell}_{\theta_0}(Y_i) + \frac{1}{\sqrt{n}} \sum_{i=1}^n (\hat{\ell}_{\theta_0,n}(Y_i) - \tilde{\ell}_{\theta_0}(Y_i)).$$

By assumption 1-(Non-Singular) the first term on the RHS converges weakly to a random variable $Z \sim \mathcal{N}(0, \tilde{I}_{\theta_0})$ and the second term on the RHS is $o_P(1)$. We conclude that $Z_n \rightsquigarrow Z$. Since $\hat{I}_{\theta_0,n} \xrightarrow{P} \tilde{I}_{\theta_0} \succ 0$ an application of the continuous mapping theorem gives that $\hat{I}_{\theta_0,n}^{-1/2} \xrightarrow{P} \tilde{I}_{\theta_0}^{-1/2}$. Combining this with Slutsky's lemma and the continuous mapping theorem once more, we conclude that $\hat{I}_{\theta_0,n}^{-1/2} Z_n \rightsquigarrow I_{\theta_0}^{-1/2} Z \sim \mathcal{N}(0, I)$ and hence

$$\hat{S}_n = (\hat{I}_{\theta_0,n}^{-1/2} Z_n)'(\hat{I}_{\theta_0,n}^{-1/2} Z_n) \rightsquigarrow \chi^2_L.$$

□

Proof of Proposition 2. We first show that $\hat{I}^t_{\theta_0} \xrightarrow{P} \tilde{I}_{\theta_0}$ and the rank estimate $r_n = \operatorname{rank}(\hat{I}^t_{\theta_0,n})$ satisfies $P(\{r_n = r\}) \to 1$ where $r = \operatorname{rank}(\tilde{I}_{\theta_0})$.

Let λ_l denote the lth largest eigenvalue of $\tilde{I}_{\theta_0}$, similarly define $\hat{\lambda}_{l,n}$ for $\hat{I}_{\theta_0,n}$ and $\hat{\lambda}^t_{l,n}$ for $\hat{I}^t_{\theta_0,n}$. Define the set $R_n \coloneqq \{r_n = r\}$, let $\underline{\nu} \coloneqq \lambda_r/2 > 0$ and note that Assumption 1-(Singular) part 3 — $P(\|\hat{I}_{\theta_0,n} - \tilde{I}_{\theta_0}\|_2 < \nu_n) \to 1$ — implies that $\|\hat{I}_{\theta_0,n} - \tilde{I}_{\theta_0}\|_2 = o_P(1)$.

By Weyl's perturbation theorem[15] we have $\max_{l=1,\ldots,L} |\hat{\lambda}_{l,n} - \lambda_l| \leq \|\hat{I}_{\theta_0,n} - \tilde{I}_{\theta_0}\|_2 = o_P(1)$. 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

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

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

Now suppose instead that $r < L$ 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 $n \to \infty$

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

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 $\|\hat{I}_{\theta_0,n} - \tilde{I}_{\theta_0}\|_2 \leq \upsilon$, 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

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

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

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

To complete the first 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 $\mathfrak{m}_{s+1} = L - r$. Let l_i^k 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 $\Pi_k \coloneqq \sum_{i=1}^{\mathfrak{m}_k} u_{l_i^k} u'_{l_i^k}$.[16] Total eigenprojections are continuous.[17] 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 $\hat{\Pi}_{k,n} \xrightarrow{P} \Pi_k$ for each $k \in [s+1]$ since $\hat{I}_{\theta_0,n} \xrightarrow{P} \tilde{I}_{\theta_0}$. Spectrally decompose $\tilde{I}_{\theta_0}$ as $\tilde{I}_{\theta_0} = \sum_{k=1}^{s} \lambda^k \Pi_k$, where the sum runs to s rather than $s+1$ since $\lambda^{s+1} = 0$. Then,

$$\hat{I}^t_{\theta_0,n} = \sum_{k=1}^{s+1}\sum_{i=1}^{\mathfrak{m}_k} \hat{\lambda}^t_{l_i^k,n} \hat{u}_{l_i^k,n}\hat{u}'_{l_i^k,n} = \sum_{k=1}^{s+1}\sum_{i=1}^{\mathfrak{m}_k} (\hat{\lambda}^t_{l_i^k,n} - \lambda^k)\hat{u}_{l_i^k,n}\hat{u}'_{l_i^k,n} + \sum_{k=1}^{s} \lambda^k \hat{\Pi}_{k,n},$$

and so

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

by $\hat{\Pi}_{k,n} \xrightarrow{P} \Pi_k$, $\hat{\Lambda}_n(\nu_n) \xrightarrow{P} \Lambda$ 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 $\hat{I}^t_{\theta_0} \xrightarrow{P} \tilde{I}_{\theta_0}$ and $P(\{r_n = r\}) \to 1$. This implies that $\hat{I}^{t,\dagger}_{\theta_0} \xrightarrow{P} \tilde{I}^\dagger_{\theta_0}$ where $\tilde{I}^\dagger_{\theta_0}$ is the Moore-Penrose inverse of $\tilde{I}_{\theta_0}$.[18]

Now consider the score statistic $\hat{S}^{SR}_n$. Similarly to in proposition 1 let $Z_n \coloneqq \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\ell}_{\theta_0,n}(Y_i)$. We have

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\ell}_{\theta_0}(Y_i) + \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\hat{\ell}_{\theta_0,n}(Y_i) - \tilde{\ell}_{\theta_0}(Y_i)\right)$$

By Assumption 1-(Singular) parts 1 and 2, we have that $Z_n \rightsquigarrow Z \sim N(0, \tilde{I}_{\theta_0})$. Slutsky's lemma and the continuous mapping theorem imply

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

where the distributional result $X \coloneqq Z'\tilde{I}^\dagger_{\theta_0} Z \sim \chi^2_r$, 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 $c_n \xrightarrow{P} c$ as $P(R_n) \to 1$. As a result, we obtain $\hat{S}^{SR}_n - c_n \rightsquigarrow X - c$ where $X \sim \chi^2_r$. Since the χ^2_r distribution is continuous, we have

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

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

[18] A necessary and sufficient condition for $(M + E_n)^\dagger \to M^\dagger$ as $E_n \to 0$ is that for all sufficiently large n, $\text{rank}(M + E_n) = \text{rank}(M)$; see, for example, chapter 6.6 of Ben-Israel and Greville (2003).

by the Portmanteau theorem

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

which completes the proof. □

Proof of Theorem 1. The proof amounts to verifying Assumptions 1-(Singular) for the ICA model under assumption 2 given a suitable log score estimator as defined in Assumption 3.

First, to verify Assumption 1-part 1 note that the data $(Y_i)_{i\geq 1}$ is an i.i.d. sequence. Moreover, by its definition, $\tilde{\ell}_\theta \in L_2(P_{\theta_0})$ (componentwise) and hence $\tilde{I}_{\theta_0}$ exists and is finite. Therefore the central limit theorem yields that $\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\ell}_{\theta_0}(Y_i) \rightsquigarrow Z \sim \mathcal{N}(0, \tilde{I}_{\theta_0})$.

Second, to verify Assumption 1-part 2, define

$$\varphi_1(y) \coloneqq \sum_{k=1}^K \sum_{j=1, j\neq k}^K \zeta_{l,k,j}\phi_k(A_k y)A_j y,$$

and

$$\hat{\varphi}_1(y) \coloneqq \sum_{k=1}^K \sum_{j=1, j\neq k}^K \zeta_{l,k,j}\hat{\phi}_{k,n}(A_k y)A_j y,$$

and let $\overline{\zeta} \coloneqq \max_{l\in[L], j\in[K], k\in[K]} |\zeta_{l,j,k}| < \infty$. We have that

$$\sqrt{n}\mathbb{P}_n(\hat{\varphi}_1 - \varphi_1) \leq \sqrt{n}\sum_{k=1}^K \sum_{j=1, j\neq k}^K \overline{\zeta}\left|\frac{1}{n}\sum_{i=1}^n \hat{\phi}_{k,n}(A_k y)A_j y - \phi_k(A_k y)A_j y\right|.$$

Since each $\left|\frac{1}{n}\sum_{i=1}^n \hat{\phi}_{k,n}(A_k y)A_j y - \phi_k(A_k y)A_j y\right| = o_{P_{\theta_0}}(n^{-1/2})$ by assumption 3 and the outside summations are finite, it follows that

$$\sqrt{n}\mathbb{P}_n(\hat{\varphi}_1 - \varphi_1) = o_{P_{\theta_0}}(1). \tag{27}$$

Next, we show that $\hat{\tau}_{k,n} - \tau_k \to 0$ almost surely where $\hat{\tau}_{k,n}$ is defined in (13). The sequences $((A_k Y_i)^3)_{i\geq 1}$ and $((A_k Y_i)^4)_{i\geq 1}$ are i.i.d. and have finite mean by assumption 2. Hence by the strong law of large numbers we have that $\hat{M}_k$ converges to M_k, P_{θ_0}-a.s.. Since M_k is nonsingular by assumption 2, the continuous mapping theorem then yields $\hat{\tau}_{k,n} \to \tau_k$, P_{θ_0}-a.s..

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

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

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

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \left[\hat{\tau}_{k,n,1} - \tau_{k,1}\right] A_k Y_i = \left[\hat{\tau}_{k,n,1} - \tau_{k,1}\right]\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_{i,k} = o_{P_{\theta_0}}(1) \times O_{P_{\theta_0}}(1) = o_{P_{\theta_0}}(1),$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\hat{\tau}_{k,n,2}-\tau_{k,2}\right]\kappa(A_kY_i)=\left[\hat{\tau}_{k,n,2}-\tau_{k,2}\right]\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\epsilon_{i,k}^2-1]=o_{P_{\theta_0}}(1)\times O_{P_{\theta_0}}(1)=o_{P_{\theta_0}}(1).$$

since $(\epsilon_{i,k})_{i\geq 1}$ and $(\epsilon_{i,k}^2-1)_{i\geq 1}$ are i.i.d. mean-zero sequences with finite second moments such that the CLT holds. Together these yield that

$$\sqrt{n}\mathbb{P}_n(\varphi_{2,\hat{\tau}_n}-\varphi_{2,\tau})=o_{P_{\theta_0}}(1). \tag{28}$$

Putting (27) and (28) together yields the claim, since $\tilde{\ell}_{\theta_0}=\varphi_1+\varphi_{2,\tau}$ and $\hat{\ell}_{\theta_0}=\hat{\varphi}_1+\varphi_{2,\hat{\tau}_n}$.

Finally, we verify Assumption 1-part 3. Define $\hat{I}_{\theta_0}:=\frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}_{\theta_0}(Y_i)\tilde{\ell}_{\theta_0}(Y_i)'$. We have

$$\|\hat{I}_n-\tilde{I}_{\theta_0}\|_2\leq\|\hat{I}_n-\hat{I}_{\theta_0}\|_2+\|\hat{I}_{\theta_0}-\tilde{I}_{\theta_0}\|_2.$$

We will obtain rates for the right hand side terms, starting with $\|\hat{I}_{\theta_0}-\tilde{I}_{\theta_0}\|_2$. Note that we have for any $l,m\in[L]$

$$[\hat{I}_{\theta_0}-\tilde{I}_{\theta_0}]_{l,m}=\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\ell}_{\theta_0,l}(Y_i)\tilde{\ell}_{\theta_0,m}(Y_i)-P_{\theta_0}[\tilde{\ell}_{\theta_0,l}(Y_i)\tilde{\ell}_{\theta_0,m}(Y_i)]\right)=\frac{1}{n}\sum_{i=1}^{n}Q_{l,m,i},$$

where $Q_{l,m,i}:=\left(\tilde{\ell}_{\theta_0,l}(Y_i)\tilde{\ell}_{\theta_0,m}(Y_i)-P_{\theta_0}[\tilde{\ell}_{\theta_0,l}(Y_i)\tilde{\ell}_{\theta_0,m}(Y_i)]\right)$ and $(Q_{l,m,i})_{i\geq 1}$ are i.i.d. mean zero random variables with the form $\tilde{Q}_{l,m,i}-P_{\theta_0}\tilde{Q}_{l,m,i}$, where $\tilde{Q}_{l,m,i}$ is defined as in lemma S14, which demonstrates that $\|\tilde{Q}_{l,m,i}\|_{P_{\theta_0},p}<\infty$ and hence $\|Q_{l,m,i}\|_{P_{\theta_0},p}<\infty$, where $p=\min\{1+\nu/4,2\}$.

If $p=2$, then by e.g. Theorem 2.5.11 in Durrett (2019) we have that for $\iota>0$

$$[\hat{I}_{\theta_0}-\tilde{I}_{\theta_0}]_{l,m}=\frac{1}{n}\sum_{i=1}^{n}Q_{l,m,i}=o_{P_{\theta_0}}\left(n^{-1/2}\log(n)^{1/2+\iota}\right).$$

It follows that

$$\|\hat{I}_{\theta_0}-\tilde{I}_{\theta_0}\|_2\leq\|\hat{I}_{\theta_0}-\tilde{I}_{\theta_0}\|_F=\sqrt{\sum_{l=1}^{L}\sum_{m=1}^{L}[\hat{I}_{\theta_0}-\tilde{I}_{\theta_0}]_{l,m}^2}=o_{P_{\theta_0}}\left(n^{-1/2}\log(n)^{1/2+\iota}\right).$$

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

$$[\hat{I}_{\theta_0}-\tilde{I}_{\theta_0}]_{l,m}=\frac{1}{n}\sum_{i=1}^{n}Q_{l,m,i}=o_{P_{\theta_0}}\left(n^{\frac{1-p}{p}}\right).$$

It follows that

$$\|\hat{I}_{\theta_0} - \tilde{I}_{\theta_0}\|_2 \le \|\hat{I}_{\theta_0} - \tilde{I}_{\theta_0}\|_F = \sqrt{\sum_{l=1}^{L}\sum_{m=1}^{L}[\hat{I}_{\theta_0} - \tilde{I}_{\theta_0}]_{l,m}^2} = o_{P_{\theta_0}}\left(n^{\frac{1-p}{p}}\right).$$

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

For the other component of the sum, define for any $l \in [L]$, write $\hat{U}_{n,i,l} := \hat{\ell}_{\theta_0,n,l}(Y_i)$, $\tilde{U}_{i,l} := \tilde{\ell}_{\theta_0,l}(Y_i)$ and $W_{n,i,l} := \hat{\ell}_{\theta_0,n,l}(Y_i) - \tilde{\ell}_{\theta_0,l}(Y_i)$. Since it is the absolute value of the (k,j)-th element of $\hat{I}_n - \hat{I}_{\theta_0}$, it is sufficient to show that $\left|\frac{1}{n}\sum_{i=1}^n \hat{U}_{n,i,k}W_{n,i,j} + \frac{1}{n}\sum_{i=1}^n W_{n,i,k}\tilde{U}_{i,j}\right| = o_{P_{\theta_0}}(1)$ as $n \to \infty$. By Cauchy-Schwarz and lemma S16

$$\left|\frac{1}{n}\sum_{i=1}^n W_{n,i,k}\tilde{U}_{i,j}\right| \le \left(\frac{1}{n}\sum_{i=1}^n \tilde{U}_{i,j}^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^n W_{n,i,k}^2\right)^{1/2} = O_{P_{\theta_0}}(1) \times o_{P_{\theta_0}}(\nu_n^{1/2}) = o_{P_{\theta_0}}(\nu_n^{1/2}),$$

$$\left|\frac{1}{n}\sum_{i=1}^n \hat{U}_{n,i,k}W_{n,i,j}\right| \le \left(\frac{1}{n}\sum_{i=1}^n \hat{U}_{n,i,k}^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^n W_{n,i,j}^2\right)^{1/2} = O_{P_{\theta_0}}(1) \times o_{P_{\theta_0}}(\nu_n^{1/2}) = o_{P_{\theta_0}}(\nu_n^{1/2}),$$

for any $(k,j) \in [L] \times [L]$. It follows that the square of the (k,j)-th element of $\hat{I}_n - \hat{I}_{\theta_0}$ is

$$\left[\frac{1}{n}\sum_{i=1}^n \hat{U}_{n,i,k}W_{n,i,j} + W_{n,i,k}\tilde{U}_{i,j}\right]^2 \le 2\left[\frac{1}{n}\sum_{i=1}^n \hat{U}_{n,i,k}W_{n,i,k}\right]^2 + 2\left[\frac{1}{n}\sum_{i=1}^n W_{n,i,k}\tilde{U}_{i,j}\right]^2 = o_{P_{\theta_0}}(\nu_n)$$

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

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

□

Proof of Theorem 2. The proof of this theorem is a special case of the more general theorem S1. In particular, we will show that assumption 4 (which is a strengthening of assumption S1; see lemma S10) implies that assumption S2 holds with $\hat{J}_n$ and J_{θ_0} defined as in equation (23). The claim of the proposition will then follow directly from theorem S1.

We start by showing the weak convergence result. Define $Q_n := \left[\frac{1}{n}\sum_{i=1}^n M_i M_i'\right]^{-1}\mathbb{E}M_i$. Since assumption 4 ensures that $\mathbb{E}[M_i M_i']$ is positive definite and lemma S18 ensures a WLLN for its sample analogue holds, the continuous mapping theorem ensures that $Q_n \xrightarrow{\text{P}} Q$.

We can write

$$\mathbb{G}_n \begin{pmatrix} \tilde{\ell}_{\theta_0} \\ \hat{q}_n \end{pmatrix} = \begin{pmatrix} \sqrt{n}\mathbb{P}_n \tilde{\ell}_{\theta_0} \\ \sum_{k=1}^K \zeta_{1,k,k}\tau_{k,1} A_k \sqrt{n}(B - \hat{B}_n)\mathbb{E}M_i \\ \vdots \\ \sum_{k=1}^K \zeta_{L,k,k}\tau_{k,1} A_k \sqrt{n}(B - \hat{B}_n)\mathbb{E}M_i \end{pmatrix}$$
$$= \begin{pmatrix} I_L & 0 \\ 0 & -(I_L \otimes Q_n)' \end{pmatrix} \times \frac{1}{\sqrt{n}} \sum_{i=1}^n \begin{pmatrix} \tilde{\ell}_{\theta_0}(Y_i) \\ \psi_{\theta_0}(W_i) \end{pmatrix}.$$

To establish weak convergence of $\mathbb{G}_n \varphi = n^{-1/2} \sum_{i=1}^n (\tilde{\ell}_{\theta_0}(Y_i)', \varphi_{\theta_0}(W_i)')'$ we will demonstrate that $(\varphi(W_i), \mathcal{F}_i)_{i \in \mathbb{N}}$ is a martingale difference sequence which satisfies the conditions of the CLT given in proposition S1. Under our assumptions $\tilde{\ell}_{\theta_0}$ and φ_{θ_0} are integrable and $\mathcal{F}_i$ measurable. Moreover, for each $l \in [L]$

$$\mathbb{E}[\tilde{\ell}_{\theta_0}(Y_i)|\mathcal{F}_{i-i}] = \mathbb{E}[\tilde{\ell}_{\theta_0}(W_i)] = 0, \quad \mathbb{E}[\psi_{\theta_0,l}(W_i)|\mathcal{F}_{i-i}] = \sum_{k=1}^K \zeta_{l,k,k}\tau_{k,1}\mathbb{E}\left[M_i \epsilon_{k,i}|\mathcal{F}_{i-1}\right] = 0.$$

Since $\sigma(Y_i)$ is independent of $\mathcal{F}_{i-1}$ we have $\mathbb{E}[\tilde{\ell}_{\theta_0}\tilde{\ell}_{\theta_0}'|\mathcal{F}_{i-1}] = \mathbb{E}[\tilde{\ell}_{\theta_0}\tilde{\ell}_{\theta_0}] = \tilde{I}_{\theta_0}$. For any $s, l \in [L]$,

$$\begin{aligned} \mathbb{E}_{\mathcal{F}_{i-1}}[\tilde{\ell}_{\theta_0,l}\psi_{\theta_0,s}] &= \sum_{b=1}^K \sum_{k=1}^K \sum_{j=1, j\neq k}^K \zeta_{l,k,j}\zeta_{s,b,b}\mathbb{E}_{\mathcal{F}_{i-1}}[\phi_k(\epsilon_{k,i})\epsilon_{j,i}\epsilon_{b,i}M_i] \\ &\quad + \sum_{b=1}^K \sum_{k=1}^K \zeta_{l,k,k}\zeta_{s,b,b}\tau_{b,1}\mathbb{E}_{\mathcal{F}_{i-1}}[\tau_{k,1}\epsilon_{k,i}\epsilon_{b,i}M_i + \tau_{k,2}\kappa(\epsilon_{k,i})\epsilon_{b,i}M_i] \\ &= \sum_{k=1}^K \zeta_{l,k,k}\zeta_{s,k,k}\tau_{k,1}\left[\tau_{k,1} + \tau_{k,2}\mathbb{E}\epsilon_{k,i}^3\right]\mathbb{E}_{\mathcal{F}_{i-1}}(M_i), \end{aligned}$$

since $\mathbb{E}_{\mathcal{F}_{i-1}}[\phi_k(\epsilon_{k,i})\epsilon_{j,i}\epsilon_{b,i}M_i] = \mathbb{E}_{\mathcal{F}_{i-1}}[\mathbb{E}(\phi_k(\epsilon_{k,i})\epsilon_{j,i}\epsilon_{b,i}|\sigma(\mathcal{F}_{i-1}, \sigma(M_i)))M_i] = 0$ since $j \neq k$ for the first right hand side term; a similar argument holds for the second. Additionally

$$\begin{aligned} \mathbb{E}_{\mathcal{F}_{i-1}}[\psi_{\theta_0,s}\psi_{\theta_0,l}'] &= \mathbb{E}_{\mathcal{F}_{i-1}}\left[\left(\sum_{k=1}^K \zeta_{s,k,k}\tau_{k,1}\epsilon_{k,i}\right)\left(\sum_{k=1}^K \zeta_{l,k,k}\tau_{k,1}\epsilon_{k,i}\right) M_i M_i'\right] \\ &= \sum_{k=1}^K \zeta_{l,k,k}\zeta_{s,k,k}\tau_{k,1}^2 \mathbb{E}_{\mathcal{F}_{i-1}}[M_i M_i']. \end{aligned}$$

Therefore, by the law of iterated expectations along with lemma S22 we have that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathcal{F}_{i-1}}\varphi\varphi' \xrightarrow{\text{P}} \mathbb{E}[\varphi\varphi'] = \tilde{V}_{\theta_0}. \tag{29}$$

To verify the conditional Lindeberg condition required by our CLT (see proposition S1) we will verify an unconditional Lyapunov condition.[19] For this, by the triangle inequality, it is sufficient to show that we have for each $k \in [K]$ (a) $\sup_{i\in\mathbb{N}} \|\phi_k(\epsilon_{k,i})\epsilon_{j,i}\|_{\mathrm{P},2+\delta} < \infty$, (b) $\sup_{i\in\mathbb{N}} \|\epsilon_{k,i}\|_{\mathrm{P},2+\delta} < \infty$, (c) $\sup_{i\in\mathbb{N}} \|\kappa(\epsilon_{k,i})\|_{\mathrm{P},2+\delta} < \infty$, (d) $\sup_{i\in\mathbb{N}} \|M_i\epsilon_{k,i}\|_{\mathrm{P},2+\delta} < \infty$ for some $\delta > 0$. Since the $\epsilon_{k,i}$ are i.i.d., the suprema in (a) - (c) are redundant. (b) and (c) follow directly from the finite $4+\upsilon$ moments of $\epsilon_{k,i}$. For (a) and (d) use the finite 4-th moments of $\phi_k(\epsilon_{k,i})$ and the uniformly bounded 4-th moments of M_i, the finite $4+\upsilon$ moments of $\epsilon_{k,i}$ and the Hölder inequality. With this in hand, proposition S1 applies and we can conclude that

$$\mathbb{G}_n\varphi = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}\tilde{\ell}_{\theta_0}(W_i)\\ \psi_{\theta_0}(W_i)\end{pmatrix} \rightsquigarrow \mathcal{N}(0, \tilde{V}_{\theta_0}).$$

Therefore, by Slutsky's theorem we have that

$$\mathbb{G}_n[\tilde{\ell}_{\theta_0}+\hat{q}_n] = \begin{pmatrix}I_L & I_{dL}\end{pmatrix}\mathbb{G}_n\begin{pmatrix}\tilde{\ell}_{\theta_0}\\ \hat{q}_n\end{pmatrix} = \begin{pmatrix}I_L & I_{dL}\end{pmatrix}\begin{pmatrix}I & 0\\ 0 & -(I_L\otimes Q_n)'\end{pmatrix}\mathbb{G}_n\begin{pmatrix}\tilde{\ell}_{\theta_0}\\ \psi_{\theta_0}\end{pmatrix} \rightsquigarrow \mathcal{Z} \sim \mathcal{N}(0, J_{\theta_0}).$$

For the second part of assumption S2, we start by showing that for any $l \in [L]$ and $s \in [d_M]$ we have

$$\mathbb{P}_n\left(\hat{\psi}_{\theta_0,n,l,s} - \psi_{\theta_0,l,s}\right)^2 = o_{\mathrm{P}}(\nu_n). \tag{30}$$

To demonstrate this, we write

$$\begin{aligned}\mathbb{P}_n\left(\hat{\psi}_{\theta_0,n,l,s} - \psi_{\theta_0,l,s}\right)^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(M_{s,i}\zeta_{l,k,k}\sum_{k=1}^{K}[\hat{\tau}_{k,1,n}\hat{\epsilon}_{k,i,n} - \tau_{k,1}\epsilon_{k,i}]\right)^2\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(M_{s,i}\zeta_{l,k,k}\sum_{k=1}^{K}[\tilde{\tau}_{k,1,n}\epsilon_{k,i} + \hat{\tau}_{k,1,n}A_k(B-\hat{B}_n)M_i]\right)^2\\ &\lesssim \sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n}\left(M_{s,i}^2[\tilde{\tau}_{k,1,n}\epsilon_{k,i}]^2 + M_{s,i}^2[\hat{\tau}_{k,1,n}A_k(B-\hat{B}_n)M_i]^2\right)\\ &\le \sum_{k=1}^{K}\tilde{\tau}_{k,1,n}^2\frac{1}{n}\sum_{i=1}^{n}M_{s,i}^2\epsilon_{k,i}^2 + \sum_{k=1}^{K}\hat{\tau}_{k,1,n}^2A_kU_n\left[\frac{1}{n}\sum_{i=1}^{n}M_{s,i}^2M_iM_i'\right]U_n'A_k'\\ &= o_{\mathrm{P}}(\nu_n)\end{aligned} \tag{31}$$

by lemmas S17, S25 and our moment assumptions. We can upper bound the error in the estimation of $\tilde{V}_{\theta_0}$ by

$$\|\hat{V}_n - \tilde{V}_{\theta_0}\|_2 \le \|\mathbb{P}_n\hat{\varphi}_n\hat{\varphi}_n' - \mathbb{P}_n\varphi\varphi'\|_2 + \|\mathbb{P}_n\varphi\varphi' - \tilde{V}_{\theta_0}\|_2.$$

For the first term, let $l \in [L(d_M+1)]$ and write $\hat{U}_{n,i,l} := \hat{\varphi}_{n,l}(W_i)$, $\tilde{U}_{n,i,l} := \varphi_l(W_i)$ and $R_{n,i,l} := \hat{\varphi}_{n,l}(W_i) - \varphi_l(W_i)$. It is then sufficient to show that $\left|\frac{1}{n}\sum_{i=1}^{n}\hat{U}_{n,i,k}R_{n,i,j} + \frac{1}{n}\sum_{i=1}^{n}R_{n,i,k}\tilde{U}_{n,i,j}\right| =$

[19] This implication is standard; see lemma S23 for a statement & proof.

$o_{\mathrm{P}}(\nu_n)$ as $n \to \infty$.[20] By Cauchy-Schwarz, lemma S19 and equation (30), we have

$$\left| \frac{1}{n} \sum_{i=1}^{n} R_{n,i,k} \tilde{U}_{n,i,j} \right| \leq \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{U}_{n,i,j}^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^{n} R_{n,i,k}^2 \right)^{1/2} = O_{\mathrm{P}}(1) \times o_{\mathrm{P}}(\nu_n^{1/2}) = o_{\mathrm{P}}(\nu_n^{1/2}),$$

$$\left| \frac{1}{n} \sum_{i=1}^{n} \hat{U}_{n,i,k} R_{n,i,j} \right| \leq \left(\frac{1}{n} \sum_{i=1}^{n} \hat{U}_{n,i,k}^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^{n} R_{n,i,j}^2 \right)^{1/2} = O_{\mathrm{P}}(1) \times o_{\mathrm{P}}(\nu_n^{1/2}) = o_{\mathrm{P}}(\nu_n^{1/2}).$$

For the right hand side term, it is sufficient to show that (a) $\|\mathbb{P}_n \tilde{\ell}_{\theta_0} \tilde{\ell}'_{\theta_0} - \tilde{I}_{\theta_0}\|_F = o_{\mathrm{P}}(\nu_{n,p})$, (b) $\|\mathbb{P}_n \tilde{\ell}_{\theta_0} \psi'_{\theta_0} - V_{\ell,\psi}\|_F = o_{\mathrm{P}}(\nu_{n,p})$ and (c) $\|\mathbb{P}_n \psi_{\theta_0} \psi'_{\theta_0} - V_{\psi}\|_F = o_{\mathrm{P}}(\nu_{n,p})$. (a) follows by the same argument as in the proof of theorem 1. For (b) we have for $l, m \in [L]$ and some $s \in [M_d]$,

$$\begin{aligned} \tilde{\ell}_{\theta_0,l}(W_i)\psi_{\theta_0,m,s}(W_i) = & \sum_{b=1}^{K} \sum_{k=1}^{K} \sum_{j=1, j \neq k}^{K} \zeta_{l,k,j} \zeta_{m,b,b} \phi_k(\epsilon_{k,i}) \epsilon_{j,i} \epsilon_{b,i} M_{s,i} \\ & + \sum_{b=1}^{K} \sum_{k=1}^{K} \zeta_{l,k,k} \zeta_{m,b,b} \tau_{b,1} [\tau_{k,1} \epsilon_{k,i} \epsilon_{b,i} M_{s,i} + \tau_{k,2} \kappa(\epsilon_{k,i}) \epsilon_{b,i} M_{s,i}], \end{aligned}$$

and for (c) we have for indices $l, m \in [L]$ and $b, s \in [d_M]$ that

$$\psi_{\theta_0,l,b}(W_i)\psi_{\theta_0,m,s}(W_i) = \left[\left(\sum_{k=1}^{K} \zeta_{l,k,k} \tau_{k,1} \epsilon_{k,i} \right) \left(\sum_{h=1}^{K} \zeta_{m,h,h} \tau_{h,1} \epsilon_{h,i} \right) M_{b,i} M_{s,i} \right].$$

These expressions in conjunction with the rate results in lemma S20 yield that (b) and (c) hold. We then have that $\|\hat{V}_n - \tilde{V}_{\theta_0}\|_2 = o_{\mathrm{P}}(\nu_n^{1/2})$. Decompose the estimation error for J_{θ_0} as

$$\|\hat{J}_n - J_{\theta_0}\|_2 \leq \|R\|_2 \|\tilde{V}_{\theta_0}\|_2 \|R - \hat{R}_n\|_2 + \|R\|_2 \|\tilde{V}_{\theta_0} - \hat{V}_n\|_2 \|\hat{R}_n\|_2 + \|R - \hat{R}_n\|_2 \|\hat{V}_n\|_2 \|\hat{R}_n\|_2,$$

to see that the proof can be completed by demonstrating that $\|\hat{R}_n - R\|_2 = o_{\mathrm{P}}(\nu_n^{1/2}) = o_{\mathrm{P}}(\nu_{n,p})$. For this, note that $\|\mathbb{E}[M_i M_i']^{-1}\|_2 = O(1)$ and $\left\| \frac{1}{n} \sum_{i=1}^{n} M_i \right\|_2 = O_{\mathrm{P}}(1)$ by assumption 4. Additionally, by lemma S18 we have $\left\| \frac{1}{n} \sum_{i=1}^{n} M_i - \mathbb{E} M_i \right\|_2 = O_{\mathrm{P}}(n^{-1/2}) = o_{\mathrm{P}}(\nu_{n,2})$ and $\left\| \frac{1}{n} \sum_{i=1}^{n} M_i M_i' - \mathbb{E}[M_i M_i'] \right\|_2 = O_{\mathrm{P}}(n^{-1/2}) = o_{\mathrm{P}}(\nu_{n,2})$. This gives us that $\left\| \left[\frac{1}{n} \sum_{i=1}^{n} M_i M_i' \right]^{-1} - \mathbb{E}[M_i M_i']^{-1} \right\|_2 = o_{\mathrm{P}}(\nu_{n,2})$.[21] In conjunction with the upper bound

$$\|\hat{Q}_n - Q\|_2 \leq \left\| \left[\frac{1}{n} \sum_{i=1}^{n} M_i M_i' \right]^{-1} - \mathbb{E}[M_i M_i']^{-1} \right\|_2 \left\| \frac{1}{n} \sum_{i=1}^{n} M_i \right\|_2 + \left\| \mathbb{E}[M_i M_i']^{-1} \right\|_2 \left\| \frac{1}{n} \sum_{i=1}^{n} M_i - \mathbb{E} M_i \right\|_2,$$

we conclude that $\|\hat{R}_n - R\|_2 = o_{\mathrm{P}}(\nu_n^{1/2}) = o_{\mathrm{P}}(\nu_{n,p})$, completing the proof. □

Proof of Example 1. Model (18) holds by assumption, as do assumptions 2 and 5. Hence it

[20] Write $\mathbb{P}_n \hat{\varphi}_n \hat{\varphi}_n' - \mathbb{P}_n \varphi \varphi' = \mathbb{P}_n \left[\hat{\varphi}_n (\hat{\varphi}_n - \varphi)' + (\hat{\varphi}_n - \varphi) \varphi' \right]$ and bound the 2-norm by the Frobenius norm.

[21] If A_0 is a non-singular matrix, the inverse map $A \mapsto A^{-1}$ is Lipschitz continuous at A_0.

remains to verify the remaining parts of assumption 4. Part 1 holds since $(M_i)_{i\geq 1}$ is i.i.d. with each component having finite fourth moments, $\mathbb{E}[M_i M_i']$ is positive definite.. Part 3 holds by our independence assumptions. For part 2 note that $\alpha_h = 0$ for $h > 0$. Hence, the summability condition is satisfied taking $r = 4$, whence the moment condition is satisfied by the assumption that $(M_i)_{i\geq 1}$ is i.i.d. with each of its components having finite fourth moments. □

Proof of Example 2. Equation (26) puts the model into the form required by equation (18). Assumptions 2 and 5 hold by hypothesis. Hence it remains to verify the remaining parts of assumption 4. We start by noting that by e.g. theorem 11.3.1. in Brockwell and Davis (1991), there is a sequence of absolutely summable matrices $(\Psi_j)_{j=0}^{\infty}$ such that

$$Z_i = \mu + \sum_{j=0}^{\infty} \Psi_j Y_{i-j},$$

is the unique stationary solution to the difference equation defining the VAR, where $\Psi_0 = I$ and $\mu = \mathbb{E}Z_i$. For part 1, the stationarity of $(Z_i)_{i\in\mathbb{N}}$ holds by the preceding discussion. We note here that since the $(Y_i)_{i\in\mathbb{N}}$ are i.i.d., $(Z_i)_{i\in\mathbb{N}}$ is in fact strictly stationary. It follows immediately that the same holds for $(M_i)_{i\in\mathbb{N}}$. The absolute summability of the autocovariances holds by proposition 10.2 (b) in Hamilton (1994). $\mathbb{E}[M_i M_i'] \succ 0$ holds by our assumptions (see lemma S24). Part 3 holds by the independence of $(\epsilon_i)_{i\in\mathbb{N}}$. For part 2 we note that since $(M_i)_{i\in\mathbb{N}}$ is strictly stationary it has finite $4 + v$ moments due to the same property of the ϵ_i. Therefore taking $r = 4 + v$ will satisfy the required moment conditions. For the mixing conditions note that by theorem 1 of Mokkadem (1988) we have that (M_i) is strong mixing, with α-mixing coefficients which satisfy $\alpha_h = O(q^h)$ for some $q \in (0, 1)$. Hence by the convergence criterion for a geometric series and the fact that $q^{\frac{r-4}{2r}} \in (0, 1)$, we have that

$$\sum_{h=0}^{\infty} \alpha_h^{\frac{r-4}{2r}} \lesssim \sum_{h=0}^{\infty} (q^h)^{\frac{r-4}{2r}} = \sum_{h=0}^{\infty} (q^{\frac{r-4}{2r}})^h < \infty.$$

□

Lemma 2. *Suppose assumption 1-(Non-Singular) holds. Then under the null hypothesis (3), $\hat{S}_n^{SR} = \hat{S}_n + o_P(1)$.*

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

$$\hat{S}_n^{SR} - \hat{S}_n = Z_n' \left[\hat{I}_{\theta_0,n}^{t,\dagger} - \hat{I}_{\theta_0,n}^{-1} \right] Z_n,$$

where $Z_n = \mathbb{G}_n \tilde{\ell}_{\theta_0} + \sqrt{n}\mathbb{P}_n(\hat{\ell}_{\theta_0,n} - \tilde{\ell}_{\theta_0}) = O_P(1)$. We have that each $\hat{\lambda}_{n,i} \xrightarrow{P} \lambda_i > 0$ where $\{\lambda_i\}_{i=1}^{L}$ are the eigenvalues of $\tilde{I}_{\theta_0}$ (ordered non-increasingly) and $\{\hat{\lambda}_{n,i}\}_{i=1}^{L}$ are the non-increasing eigenvalues of $\hat{I}_{\theta_0,n}$. Since $\nu_n \to 0$, it follows that with probability approaching one, $\hat{I}_{\theta_0,n}^{t} = \hat{I}_{\theta_0,n}$ and this matrix is of full rank. Hence, with probability approaching one $\hat{I}_{\theta_0,n}^{t,\dagger} = \hat{I}_{\theta_0,n}^{-1}$, implying that $\hat{I}_{\theta_0,n}^{t,\dagger} - \hat{I}_{\theta_0,n}^{-1} = o_P(1)$, which suffices to complete the proof. □

Appendix B: Density score estimation

In this section we describe a density score estimator based on flexible cubic B-splines. The estimator is also considered in Chen and Bickel (2006) who build on Jin (1992).[22] Letting $\xi_1 < \cdots < \xi_N$ be a knot sequence, the first order B-splines are defined according to $b_i^{(1)}(x) := \mathbf{1}_{[\xi_i, \xi_{i+1})}(x)$. Subsequent order B-splines can be computed according to the recurrence relation

$$b_i^{(\kappa)}(x) = \frac{x - \xi_i}{\xi_{i+\kappa-1} - \xi_i} b_i^{(\kappa-1)}(x) + \frac{\xi_{i+\kappa} - x}{\xi_{i+\kappa} - \xi_{i+1}} b_{i+1}^{(\kappa-1)}(x), \tag{32}$$

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

$$c_i^{(\kappa)}(x) = \frac{\kappa - 1}{\xi_{i+\kappa-1} - \xi_i} b_i^{(\kappa-1)}(x) - \frac{\kappa - 1}{\xi_{i+\kappa} - \xi_{i+1}} b_{i+1}^{(\kappa-1)}(x). \tag{33}$$

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

Let $b_{k,n} = (b_{k,n,1}, \ldots, b_{k,n,B_{k,n}})'$ be a collection of $B_{k,n}$ cubic B-splines and let $c_{k,n} = (c_{k,n,1}, \ldots, c_{k,n,B_{k,n}})'$ be their derivatives: $c_{k,n,i}(x) := \frac{\mathrm{d}b_{k,n,i}(x)}{\mathrm{d}x}$ for each $i \in [B_{k,n}]$. Let $\gamma_k \in \mathbb{R}^{B_{k,n}}$. The knots of the splines, $\xi_{k,n} = (\xi_{k,n,i})_{i=1}^{K_{k,n}}$ are equally spaced in $[\Xi_{k,n}^L, \Xi_{k,n}^U]$ with $\delta_{k,n} := \xi_{k,n,i+1} - \xi_{k,n,i} > 0$.[23] For each (k, n) pair the relationships between the number of knots $(K_{k,n})$, the number of spline functions $(B_{k,n})$ and $\delta_{k,n}$ are given by $B_{k,n} = K_{k,n} - 4$ and $K_{k,n} = 1 + (\Xi_{k,n}^U - \Xi_{k,n}^L)/\delta_{k,n}$.[24]

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

$$\begin{aligned}\int (\phi_k(z) - \gamma_k' b_{k,n}(z))^2 \eta_k(z)\,\mathrm{d}z &= \int \phi_k^2\,\mathrm{d}G_k + \int (\gamma_k' b_{k,n})^2\,\mathrm{d}G_k + 2\int \gamma_k' c_{k,n}(z)\eta_k(z)\,\mathrm{d}z \\ &= G_k\phi_k^2 + \gamma_k' G_k[b_{k,n}b_{k,n}']\gamma_k + 2\gamma_k' G_k c_{k,n}.\end{aligned} \tag{34}$$

The solution to minimising this mean-squared error is given by:[25]

$$\gamma_{k,n} = -G_k[b_{k,n}b_{k,n}']^{-1} G_k c_{k,n}. \tag{35}$$

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

$$\hat{\gamma}_{k,n} := -\left[\frac{1}{n}\sum_{i=1}^{n} b_{k,n}(\epsilon_{k,i}) b_{k,n}(\epsilon_{k,i})'\right]^{-1} \frac{1}{n}\sum_{i=1}^{n} c_{k,n}(\epsilon_{k,i}), \tag{36}$$

[22] The results in this section are based on those in Chen and Bickel (2006) but adapted to our requirements. In particular, we will impose $A = A(\alpha)$ under H_0 and therefore we do not need to account for estimation uncertainty in A; however we do need results which allow us to determine the rate of convergence of our estimate of the efficient information matrix. For our extensions to the ICA model we need a version which applies when we observe only estimates of Y.

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

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

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

and our estimate of ϕ_k:

$$\hat{\phi}_{k,n}(z) := \hat{\gamma}'_{k,n} b_{k,n}(z). \tag{37}$$

We will now show that the estimates $\hat{\phi}_{k,n}$ satisfy assumptions 3 and 5 under regularity conditions on η_k and the choice of knot points. We first state and prove the main results of this section in two propositions; the proofs depend on a number of lemmas which are recorded subsequently.

Proposition 3. *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 ν_n as in assumption 3, $[\Xi^L_{k,n}, \Xi^U_{k,n}] \uparrow \tilde{\Xi} \supset \operatorname{supp}(\eta_k)$ and $\delta_{k,n} \downarrow 0$ such that*

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

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

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

(iv) For each n, $\phi_{k,n}$ is three-times continuously differentiable on $[\Xi^L_{k,n}, \Xi^U_{k,n}]$ and $\|\phi^{(3)}_{k,n}\|^2_\infty \delta^6_{k,n} = o(\nu_n)$;[26]

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

Then, under assumption 2, the estimates $\hat{\phi}_{k,n}$ satisfy assumption 3.

Proof. We have that $\epsilon_{k,i} = A_k Y_i$ and so can write

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n \hat{\phi}_{k,n}(A_k Y_i)A_j Y_i - \frac{1}{n}\sum_{i=1}^n \phi_k(A_k Y_i)A_j Y_i\right| \leq{}& \left|\frac{1}{n}\sum_{i=1}^n \left[\hat{\phi}_{k,n}(\epsilon_{k,i}) - \tilde{\phi}_{k,n}(\epsilon_{k,i})\right]\epsilon_{j,i}\right| \\ &+ \left|\frac{1}{n}\sum_{i=1}^n \left[\tilde{\phi}_{k,n}(\epsilon_{k,i}) - \phi_{k,n}(\epsilon_{k,i})\right]\epsilon_{j,i}\right| \\ &+ \left|\frac{1}{n}\sum_{i=1}^n \left[\phi_{k,n}(\epsilon_{k,i}) - \phi_k(\epsilon_{k,i})\right]\epsilon_{j,i}\right|,\end{aligned} \tag{38}$$

where $\tilde{\phi}_{k,n}(z) := \gamma'_{k,n} b_{k,n}(z)$. We will show that each of these three terms on the right hand side are $o_{P_{\theta_0}}(n^{-1/2})$.

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

$$\begin{aligned}\mathbb{E}_{\theta_0}\left([\phi_{k,n}(\epsilon_k) - \phi_k(\epsilon_k)]^2\epsilon_j^2\right) &= G_k\left[\phi_k(\epsilon_k)^2 \mathbf{1}\{\epsilon_k \notin [\Xi^L_{k,n}, \Xi^U_{k,n}]\}\right] \\ &\leq \left[G_k \phi_k(\epsilon_k)^4\right]^{1/2}\left[G_k \mathbf{1}\{\epsilon_k \notin [\Xi^L_{k,n}, \Xi^U_{k,n}]\}\right]^{1/2} \\ &\to 0.\end{aligned} \tag{39}$$

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

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

$$P_{\theta_0}\left(\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\phi_{k,n}(\epsilon_{k,i}) - \phi_k(\epsilon_{k,i})]\epsilon_{j,i}\right| > \upsilon\right) \leq \frac{n\mathbb{E}_{\theta_0}\left([\phi_{k,n}(\epsilon_k) - \phi_k(\epsilon_k)]^2\epsilon_j^2\right)}{n\upsilon} \to 0.$$

For the second term, we note that by our hypotheses and lemma 5 we have

$$\mathbb{E}_{\theta_0}\left([\tilde{\phi}_{k,n}(\epsilon_k) - \phi_{k,n}(\epsilon_k)]^2\epsilon_j^2\right) = G_k\left([\tilde{\phi}_{k,n}(\epsilon_k) - \phi_{k,n}(\epsilon_k)]^2\right) \leq C^2\delta_{k,n}^6\|\phi_k^{(3)}\|_\infty^2 \to 0, \tag{40}$$

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

$$P_{\theta_0}\left(\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\tilde{\phi}_{k,n}(\epsilon_{k,i}) - \phi_{k,n}(\epsilon_{k,i})]\epsilon_{j,i}\right| > \upsilon\right) \leq \frac{n\mathbb{E}_{\theta_0}\left([\tilde{\phi}_{k,n}(\epsilon_k) - \phi_{k,n}(\epsilon_k)]^2\epsilon_j^2\right)}{n\upsilon} \to 0.$$

For the first term, by Cauchy-Schwarz

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(\epsilon_{k,i}) - \tilde{\phi}_{k,n}(\epsilon_{k,i})\right]\epsilon_{j,i}\right| \leq \|\hat{\gamma}_{k,n} - \gamma_{k,n}\|_2\left\|\frac{1}{n}\sum_{i=1}^{n}b_{k,n}(\epsilon_{k,i})\epsilon_{j,i}\right\|_2 = o_{P_{\theta_0}}(n^{-1/2}),$$

by lemmas 6 and 7.

It remains to prove the second part. Break the sum into components as:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(\left[\hat{\phi}_{k,n}(A_kY_i) - \phi_k(A_kY_i)\right]A_jY_i\right)^2 \leq{}& \frac{4}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(\epsilon_{k,i}) - \tilde{\phi}_{k,n}(\epsilon_{k,i})\right]^2\epsilon_{j,i}^2\\ &+ \frac{4}{n}\sum_{i=1}^{n}\left[\tilde{\phi}_{k,n}(\epsilon_{k,i}) - \phi_{k,n}(\epsilon_{k,i})\right]^2\epsilon_{j,i}^2 \qquad (41)\\ &+ \frac{4}{n}\sum_{i=1}^{n}[\phi_{k,n}(\epsilon_{k,i}) - \phi_k(\epsilon_{k,i})]^2\epsilon_{j,i}^2.\end{aligned}$$

We will show that (1/4 of) each of the right hand side terms is $o_{P_{\theta_0}}(\nu_n)$ under our rate assumptions. For the last term, for any $\upsilon > 0$, by Markov's inequality and (39) we have

$$P_{\theta_0}\left(\left|\frac{1}{n}\sum_{i=1}^{n}[\phi_{k,n}(\epsilon_{k,i}) - \phi_k(\epsilon_{k,i})]^2\epsilon_{j,i}^2\right| > \upsilon\nu_n\right) \lesssim \frac{\left[G_k\mathbf{1}\{\epsilon_k \notin [\Xi_{k,n}^L, \Xi_{k,n}^U]\}\right]^{1/2}}{\upsilon\nu_n} \lesssim \nu_n^{-1}o(\nu_n) = o(1).$$

For the second term, for any $\upsilon > 0$, by Markov's inequality and (40) we have

$$P_{\theta_0}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left[\tilde{\phi}_{k,n}(\epsilon_{k,i}) - \phi_{k,n}(\epsilon_{k,i})\right]^2\epsilon_{j,i}^2\right| > \upsilon\nu_n\right) \leq \nu_n^{-1}O(\delta_{k,n}^6\|\phi_k^{(3)}\|_\infty^2) = o(1).$$

Finally, for the first term in the decomposition, by lemma 7 we have

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(\epsilon_{k,i})-\tilde{\phi}_{k,n}(\epsilon_{k,i})\right]^{2}\epsilon_{j,i}^{2}\leq\|\hat{\gamma}_{k,n}-\gamma_{k,n}\|_{2}^{2}\left[\frac{1}{n}\sum_{i=1}^{n}\|b_{k,n}(\epsilon_{k,i})\|_{2}^{2}\epsilon_{j,i}^{2}\right]=o_{P_{\theta_0}}(\nu_n).$$

□

Proposition 4. *Let $\phi_{k,n}\coloneqq\phi_k\mathbf{1}_{[\Xi^L_{k,n},\Xi^U_{k,n}]}$ and $\Delta_{k,n}\coloneqq\Xi^U_{k,n}-\Xi^L_{k,n}$ and suppose that for ν_n as in assumption 3, $[\Xi^L_{k,n},\Xi^U_{k,n}]\uparrow\tilde{\Xi}\supset\operatorname{supp}(\eta_k)$ and $\delta_{k,n}\downarrow 0$ such that*

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

(ii) For some $\iota>0$, $n^{-1}\Delta_{k,n}^{2+2\iota}\delta_{k,n}^{-(10+2\iota)}=o(\nu_n)$;

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

(iv) For each n, $\phi_{k,n}$ is three-times continuously differentiable on $[\Xi^L_{k,n},\Xi^U_{k,n}]$ and $\|\phi^{(3)}_{k,n}\|^2_\infty\delta^6_{k,n}=o(\nu_n)$;[27]

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

Then, under assumption S1, the estimates $\hat{\phi}_{k,n}(A_k\hat{Y}_{i,n})$ based on $\hat{\epsilon}_{k,i,n}\coloneqq A_k\hat{Y}_{i,n}$ satisfy assumption 5.

Proof. To simplify the notation, first note that $\epsilon_{k,i}=A_kY_i$ and $\hat{\epsilon}_{k,i,n}=A_k\hat{Y}_{i,n}$. We will first demonstrate equation (14) holds. Write

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n})\hat{\epsilon}_{j,i,n}-\phi_k(\epsilon_{k,i})\epsilon_{j,i}\right]&=\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n})[\epsilon_{j,i}+(\hat{\epsilon}_{j,i,n}-\epsilon_{j,i})]-\phi_k(\epsilon_{k,i})\epsilon_{j,i}\right]\\&=\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n})-\phi_k(\epsilon_{k,i})\right]\epsilon_{j,i}+\frac{1}{n}\sum_{i=1}^{n}\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n})\tilde{\epsilon}_{j,i,n},\end{aligned}\tag{42}$$

where $\tilde{\epsilon}_{j,i,n}\coloneqq\hat{\epsilon}_{j,i,n}-\epsilon_{j,i}$. We have that $\tilde{\epsilon}_{j,i,n}=A_jU_nM_i$ and so

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_k(\epsilon_{k,i})\tilde{\epsilon}_{j,i,n}=A_j\sqrt{n}U_n\frac{1}{n}\sum_{i=1}^{n}\phi_k(\epsilon_{k,i})M_i=O_{\mathrm{P}}(1)\times o_{\mathrm{P}}(1)=o_{\mathrm{P}}(1),$$

since $(M_i\phi_k(\epsilon_{k,i}),\mathcal{P}_{k,i})_{i\in\mathbb{N}}$ is a MDS with bounded variances and therefore $\frac{1}{n}\sum_{i=1}^{n}\phi_k(\epsilon_{k,i})M_i=$

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

$o_{\mathrm{P}}(1)$ by e.g. theorem 20.10 of Davidson (1994).[28] Additionally we have that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n}) - \phi_k(\epsilon_{k,i})\right]\tilde{\epsilon}_{j,i,n}\right| \leq \left(\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n}) - \phi_k(\epsilon_{k,i})\right]^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{\epsilon}_{j,i,n}^2\right)^{1/2}$$

$$= o_{\mathrm{P}}(1) \times \left(A_j U_n\left[\frac{1}{n}\sum_{i=1}^{n} M_i M_i'\right]U_n' A_j'\right)^{1/2}$$

$$= o_{\mathrm{P}}(n^{-1/2}),$$

by lemma 10 and assumption S1. Together the preceding two displays demonstrate that the second right hand side term in (42) is $o_{\mathrm{P}}(n^{-1/2})$.

We next show that the first right hand side term in (42) is $o_{\mathrm{P}}(n^{-1/2})$. For this, decompose the term as follows:

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n}) - \phi_k(\epsilon_{k,i})\right]\epsilon_{j,i} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n}) - \tilde{\phi}_{k,n}(\epsilon_{k,i})\right]\epsilon_{j,i} + \frac{1}{n}\sum_{i=1}^{n}\left[\tilde{\phi}_{k,n}(\epsilon_{k,i}) - \phi_{k,n}(\epsilon_{k,i})\right]\epsilon_{j,i},$$

where $\tilde{\phi}_{k,n}(z) \coloneqq \gamma_{k,n}' b_{k,n}(z)$. The second right-hand side term here is $o_{\mathrm{P}}(n^{-1/2})$ by exactly the same argument as in the proof of lemma 3, noting that (the absolute value of) this term can be bounded by the sum of the second two right hand side terms in (38). To handle the first right hand side term, note that each term in the sum can be written as

$$\left[\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n}) - \tilde{\phi}_{k,n}(\epsilon_{k,i})\right]\epsilon_{j,i} = \left[\check{\gamma}_{k,n}'\left[b_{k,n}(\hat{\epsilon}_{k,i,n}) - b_{k,n}(\epsilon_{k,i})\right]\epsilon_{j,i}\right] + \left[(\check{\gamma}_{k,n} - \gamma_{k,n})' b_{k,n}(\epsilon_{k,i})\epsilon_{j,i}\right].$$

By lemmas 9 and 6 we have for the second term:

$$\frac{1}{n}\sum_{i=1}^{n}\left[(\check{\gamma}_{k,n} - \gamma_{k,n})' b_{k,n}(\epsilon_{k,i})\epsilon_{j,i}\right] \leq \|\check{\gamma}_{k,n} - \gamma_{k,n}\|_2 \left\|\frac{1}{n}\sum_{i=1}^{n} b_{k,n}(\epsilon_{k,i})\epsilon_{j,i}\right\|_2 = o_{\mathrm{P}}(n^{-1/2}).$$

It remains to control $\frac{1}{n}\sum_{i=1}^{n}\left[\check{\gamma}_{k,n}'\left[b_{k,n}(\hat{\epsilon}_{k,i,n}) - b_{k,n}(\epsilon_{k,i})\right]\epsilon_{j,i}\right]$. By the mean value theorem for random variables (lemma S21) for each $m \in [B_{k,n}]$ there is a random variable $\bar{\epsilon}_{k,i,n,m}$ which lies on the line connecting $\hat{\epsilon}_{k,i,n}$ and $\epsilon_{k,i}$ such that

$$b_{k,n,m}(\hat{\epsilon}_{k,i,n}) - b_{k,n,m}(\epsilon_{k,i}) = c_{k,n,m}(\bar{\epsilon}_{k,i,n,m})[\hat{\epsilon}_{k,i,n} - \epsilon_{k,i}] = c_{k,n,m}(\bar{\epsilon}_{k,i,n,m}) A_k U_n M_i, \tag{43}$$

P-a.s.. Next we have that

$$\frac{1}{n}\sum_{i=1}^{n} c_{k,n,m}(\epsilon_{k,i}) M_{s,i}\epsilon_{j,i} \lesssim \delta_{k,n}^{-2}\frac{1}{n}\sum_{i=1}^{n} M_{s,i}\epsilon_{j,i} = O_{\mathrm{P}}\left(\delta_{k,n}^{-2} n^{-1/2}\right),$$

since $(M_i\epsilon_{j,i}, \mathcal{E}_{j,i})_{i\in\mathbb{N}}$ is a MDS[29] and therefore serially uncorrelated and so we can argue that

[28] Here $(\mathcal{P}_{k,i})_{i\in\mathbb{N}}$ is the sequence of filtrations with respect to which $(M_i\phi_k(\epsilon_{k,i}))_{i\in\mathbb{N}}$ is a MDS.

[29] Here $(\mathcal{E}_{k,i})_{i\in\mathbb{N}}$ is the sequence of filtrations with respect to which $(M_i\epsilon_{k,i})_{i\in\mathbb{N}}$ is a MDS.

for any $\varepsilon > 0$ we can choose $R > 0$ large enough such that

$$\mathrm{P}\left(\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} M_{s,i}\epsilon_{j,i}\right| > R\right) \leq \frac{\sup_{i\in\mathbb{N}} \mathbb{E}\left[M_{s,i}\epsilon_{j,i}\right]^2}{R} < \varepsilon.$$

It follows that

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n} \check{\gamma}_{k,n}' c_{k,n}(\epsilon_{k,i})[\hat{\epsilon}_{k,i,n} - \epsilon_{k,i}]\epsilon_{j,i}\right| &\leq \|\check{\gamma}_{k,n}\|_2 \left\|A_k U_n \frac{1}{n}\sum_{i=1}^{n} c_{k,n,m}(\epsilon_{k,i}) M_{s,i}\epsilon_{j,i}\right\|_2 \\ &= O_{\mathrm{P}}(\delta_{k,n} B_{k,n}^{1/2}) \times O_{\mathrm{P}}(\delta_{k,n}^{-2} n^{-1}) \\ &= o_{\mathrm{P}}(n^{-1/2}),\end{aligned} \tag{44}$$

under condition (ii), since we have that $\delta_{k,n}^{-3} B_{k,n}^{1/2} n^{-1/2} \leq \delta_{k,n}^{4} \Delta_{k,n} n^{-1/2} = o(1)$. Lastly,[30]

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n} \check{\gamma}_{k,n}' \left[c_{k,n}(\bar{\epsilon}_{k,i,n}) - c_{k,n}(\epsilon_{k,i})\right] \tilde{\epsilon}_{k,i,n}\epsilon_{j,i}\right| &\leq \|\check{\gamma}_{k,n}\|_2 \times \left\|\frac{1}{n}\sum_{i=1}^{n} \left[c_{k,n}(\bar{\epsilon}_{k,i,n}) - c_{k,n}(\epsilon_{k,i})\right] \tilde{\epsilon}_{k,i,n}\epsilon_{j,i}\right\|_2 \\ &= O_{\mathrm{P}}\left(\delta_{k,n}^{-1} B_{k,n}^{1/2}\right) \times O_{\mathrm{P}}\left(B_{k,n}^{1/2} \delta_{k,n}^{-2} n^{-1}\right) \\ &= o_{\mathrm{P}}(n^{-1/2}),\end{aligned}$$

since $\delta_{k,n}^{-3} B_{k,n} n^{-1/2} \leq \delta_{k,n}^{-4} \Delta_{k,n} n^{-1/2} = o(1)$. This demonstrates that

$$\frac{1}{n}\sum_{i=1}^{n} \check{\gamma}_{k,n}' \left[c_{k,n}(\bar{\epsilon}_{k,i,n}) - c_{k,n}(\epsilon_{k,i})\right] [\hat{\epsilon}_{k,i,n} - \epsilon_{k,i}]\epsilon_{j,i} = o_{\mathrm{P}}(n^{-1/2}). \tag{45}$$

Together, equations (43), (44) and (45) yield that $\frac{1}{n}\sum_{i=1}^{n} \left[\check{\gamma}_{k,n}' \left[b_{k,n}(\hat{\epsilon}_{k,i,n}) - b_{k,n}(\epsilon_{k,i})\right] \epsilon_{j,i}\right] = o_{\mathrm{P}}(n^{-1/2})$ and hence (14) holds.

[30] Here $c_{k,n}(\bar{\epsilon}_{k,i,n})$ is to be understood as the vector with components $c_{k,n,m}(\bar{\epsilon}_{k,i,n,m})$. Moreover, for each $m \in [B_{k,n}]$ using lemma 8, the definition of $\bar{\epsilon}_{k,i,n,m}$ and assumption S1 we have

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n} \left[c_{k,n,m}(\bar{\epsilon}_{k,i,n}) - c_{k,n,m}(\epsilon_{k,i})\right] [\hat{\epsilon}_{k,i,n} - \epsilon_{k,i}]\epsilon_{j,i}\right| &\leq \left(\frac{1}{n}\sum_{i=1}^{n} \epsilon_{j,i}^2\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^{n} \delta_{k,n}^{-4}[\hat{\epsilon}_{k,i,n} - \epsilon_{k,i}]^2[\hat{\epsilon}_{k,i,n} - \epsilon_{k,i}]^2\right)^{1/2} \\ &= O_{\mathrm{P}}(1) \times \left(\delta_{k,n}^{-4}\|A_k\|_2^4 \|U_n\|_2^4 \frac{1}{n}\sum_{i=1}^{n} \|M_i\|_2^4\right)^{1/2} \\ &= O_{\mathrm{P}}(1) \times O_{\mathrm{P}}(\delta_{k,n}^{-2} n^{-1}).\end{aligned}$$

It remains to prove that equation (15) holds. We first re-write the sum as

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n})\hat{\epsilon}_{j,i,n} - \phi_k(\epsilon_{k,i})\epsilon_{j,i}\right)^2 \\
&\qquad = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n})\left[\hat{\epsilon}_{j,i,n} - \epsilon_{j,i}\right] + \left[\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n}) - \phi_k(\epsilon_{k,i})\right]\epsilon_{j,i}\right)^2 \qquad (46)\\
&\qquad \lesssim \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n})\left[\hat{\epsilon}_{j,i,n} - \epsilon_{j,i}\right]\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(\left[\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n}) - \phi_k(\epsilon_{k,i})\right]\epsilon_{j,i}\right)^2.
\end{aligned}$$

For the first right hand side term note that since $\sum_{m=1}^{B_{k,n}} b_{k,n,m}(x)^2 \leq 1$ (see e.g. (36) on p. 96 of de Boor, 2001), using lemma 9 we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n})\left[\hat{\epsilon}_{j,i,n} - \epsilon_{j,i}\right]\right)^2 &\leq \frac{1}{n}\sum_{i=1}^{n}\|\check{\gamma}_{k,n}\|_2^2\|b_{k,n}(\hat{\epsilon}_{k,i,n})A_jU_nM_i\|_2^2 \\
&\lesssim \|\check{\gamma}_{k,n}\|_2^2\|U_n\|_2^2\frac{1}{n}\sum_{i=1}^{n}\|M_i\|_2^2 \qquad (47)\\
&= O_{\mathrm{P}}\left(\delta_{k,n}^{-2}B_{k,n}n^{-1}\right) \\
&= o_{\mathrm{P}}(\nu_n),
\end{aligned}$$

since (for sufficiently large n) $\delta_{k,n}^{-2}B_{k,n}n^{-1} \leq \delta_{k,n}^{-3}\Delta_{k,n}n^{-1/2} = o(1)$ by condition (ii). For the second RHS term in (46) start by bounding it by:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left(\left[\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n}) - \phi_k(\epsilon_{k,i})\right]\epsilon_{j,i}\right)^2 \leq &\frac{4}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n}) - \tilde{\phi}_{k,n}(\epsilon_{k,i})\right]^2\epsilon_{j,i}^2 \\
&+ \frac{4}{n}\sum_{i=1}^{n}\left[\tilde{\phi}_{k,n}(\epsilon_{k,i}) - \phi_{k,n}(\epsilon_{k,i})\right]^2\epsilon_{j,i}^2 \qquad (48)\\
&+ \frac{4}{n}\sum_{i=1}^{n}\left[\phi_{k,n}(\epsilon_{k,i}) - \phi_k(\epsilon_{k,i})\right]^2\epsilon_{j,i}^2.
\end{aligned}$$

That the second two terms on the right hand side in equation (48) are $o_{\mathrm{P}}(\nu_n)$ follows exactly as the corresponding argument in the proof of lemma 3.[31] For the first term, again using lemma 9 and the fact that $\sum_{m=1}^{B_{k,n}} b_{k,n,m}(x)^2 \leq 1$ along with lemma 8, we can upper bound

[31] See equation (41) and the two subsequent displays.

the term by

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n})-\tilde{\phi}_{k,n}(\epsilon_{k,i})\right]^2\epsilon_{j,i}^2\\
&\quad\lesssim \frac{1}{n}\sum_{i=1}^{n}\left[\tilde{\gamma}_{k,n}'\left(b_{k,n}(\hat{\epsilon}_{k,i,n})-b_{k,n}(\epsilon_{k,i})\right)\epsilon_{j,i}\right]^2+\frac{1}{n}\sum_{i=1}^{n}\left[(\tilde{\gamma}_{k,n}-\gamma_{k,n})'b_{k,n}(\epsilon_{k,i})\epsilon_{j,i}\right]^2\\
&\quad\leq \frac{1}{n}\sum_{i=1}^{n}\|\tilde{\gamma}_{k,n}\|_2^2\|b_{k,n}(\hat{\epsilon}_{k,i,n})-b_{k,n}(\epsilon_{k,i})\|_2^2\epsilon_{j,i}^2+\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\gamma}_{k,n}-\gamma_{k,n}\|_2^2\|b_{k,n}(\epsilon_{k,i})\|_2^2\epsilon_{j,i}^2 \qquad (49)\\
&\quad\lesssim \|\tilde{\gamma}_{k,n}\|_2^2\delta_{k,n}^{-2}B_{k,n}\|U_n\|_2^2\frac{1}{n}\sum_{i=1}^{n}\|M_i\|_2^2\epsilon_{j,i}^2+\|\tilde{\gamma}_{k,n}-\gamma_{k,n}\|_2^2\frac{1}{n}\sum_{i=1}^{n}\epsilon_{j,i}^2\\
&\quad= O_{\mathrm{P}}\left(\delta_{k,n}^{-4}B_{k,n}^2n^{-1}\right)+O_{\mathrm{P}}\left(\frac{B_{k,n}^2\log B_{k,n}}{\delta_{k,n}^8 n}\right)\\
&\quad= o_{\mathrm{P}}(\nu_n),
\end{aligned}$$

since $\delta_{k,n}^{-4}B_{k,n}^2n^{-1}\leq\delta_{k,n}^{-6}\Delta_{k,n}^2n^{-1}=o(\nu_n)$ and $B_{k,n}^2\log B_{k,n}\delta_{k,n}^{-8}n^{-1}\leq\delta_{k,n}^{-10}\Delta_{k,n}^2\log(\delta_{k,n}^{-1}\Delta_{k,n})n^{-1}=o(\nu_n)$ by condition (ii). This shows that the remaining term in equation (48) has the required rate, which implies that the remaining term in (46) has the required rate and therefore (15) holds. □

Lemma 3. *The smallest eigenvalue of the $B_{k,n}\times B_{k,n}$ Gram matrix $\tilde{\Gamma}_{k,n}:=\int b_{k,n}b_{k,n}'\,\mathrm{d}\lambda$ satisfies*

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

for a $\upsilon>0$.

Proof. Since $b_{k,n,m}(x)b_{k,n,s}(x)$ is non-zero only for $|m-s|\leq 3$ and each $b_{k,n,m}$ is non-zero only on $[\xi_{k,n,m},\xi_{k,n,m+4})]$ (e.g. (20) p. 91 of de Boor, 2001), $\tilde{\Gamma}_{k,n}$ is a symmetric banded Toeplitz matrix.[32] Its entries can be computed by direct integration:

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

For simplicity of notation let $f_0:=\frac{151}{315}$, $f_1:=f_{-1}:=\frac{397}{1680}$, $f_2:=f_{-2}:=\frac{1}{42}$ and $f_3:=f_{-3}:=\frac{1}{5040}$ and let $f_s:=0$ for $|s|>3$ Now, let $f(\theta):=\sum_{s=-3}^{3}f_se^{i(s\theta)}$. Then, $\tilde{\Gamma}_{k,n}/\delta_{k,n}$ is then the matrix generated by f in the sense that $\tilde{\Gamma}_{k,n}/\delta_{k,n}=\mathscr{T}_n(f):=\sum_{s=-\min(B_{k,n}-1,3)}^{\min(B_{k,n}-1,3)}f_kJ_n^s$ where each J_n^s is the $B_{k,n}\times B_{k,n}$ matrix which is zero everywhere except for the (i,j)-th entries where

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

$i - j = s$, where it has a value of 1.[33] Since $f \in L_1([-\pi, \pi])$ and is real on $[-\pi, \pi]$ by Theorem 6.1 in Garoni and Serra-Capizzano (2017) we have that $\lambda_{\min}(\tilde{\Gamma}_{k,n}) = \delta_{k,n}\lambda_{\min}(\tilde{\Gamma}_{k,n}/\delta_{k,n}) \geq \delta_{k,n} \inf_{\theta \in [-\pi,\pi]} f(\theta) = \delta_{k,n}\upsilon$, where $\upsilon := \inf_{\theta \in [-\pi,\pi]} f(\theta) \geq 1/20$. □

Lemma 4. *Suppose $\xi \in \mathbb{R}^{N+1}$ such that $a = \xi_0 < \xi_1 < \cdots < \xi_N = b$, $h := \max_{i \in [N]} \xi_i - \xi_{i-1}$, and let $\mathscr{G}_k(\xi)$ be the linear space formed by degree k splines with knots ξ. Then, if $f \in C^{k-1}[a, b]$ we have that*

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

where c_k depends only on k.

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

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

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

Proof. By the definition of $\tilde{\phi}_{k,n}$ and lemma 4 we have

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

The first inequality comes from the fact that we can equivalently see $\gamma_{k,n} = -G_k[b_{k,n}b'_{k,n}]^{-1}G_k c_{k,n}$ as the solution to a version of the mean-squared error problem based on equation (34) where we only integrate over the support of $\phi_{k,n}$ since this is also the support of $b_{k,n}$ and $c_{k,n}$. □

Lemma 6 (Cf. Lemma A.3, Chen and Bickel, 2006)**.** *Under assumption 2 we have for $k \neq j$,*

$$\left\| \frac{1}{n} \sum_{i=1}^n b_{k,n}(\epsilon_{k,i})\epsilon_{j,i} \right\|_2 = O_{P_{\theta_0}}(n^{-1/2}).$$

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

$$\mathbb{E}_{\theta_0} \left(\left\| \frac{1}{n} \sum_{i=1}^n b_{k,n}(\epsilon_{k,i})\epsilon_{j,i} \right\|_2^2 \right) = \frac{1}{n} \mathbb{E}_{\theta_0} \left(\sum_{m=1}^{B_{k,n}} b_{k,n,m}(\epsilon_k)^2 \right) \leq \frac{1}{n}$$

Fix $\epsilon > 0$ and take $M > 0$ large enough such that $1/M^2 < \epsilon$. Markov's inequality yields

$$P_{\theta_0} \left(\sqrt{n} \left\| \frac{1}{n} \sum_{i=1}^n b_{k,n}(\epsilon_{k,i})\epsilon_{j,i} \right\|_2 > M \right) \leq \frac{\mathbb{E}_{\theta_0} \left(n \left\| \frac{1}{n} \sum_{i=1}^n b_{k,n}(\epsilon_{k,i})\epsilon_{j,i} \right\|_2^2 \right)}{M^2} \leq \frac{1}{M^2} < \epsilon.$$

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

□

Lemma 7 (Cf. Lemma A.2, Chen and Bickel, 2006)**.** *Let $\hat{\gamma}_{k,n}$ and $\gamma_{k,n}$ be defined as in equations* (36) *and* (35) *respectively. Suppose that conditions (ii), (iii) and (v) of proposition 3 and assumption 2 hold. Then, if we define*

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

and

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

we have that

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

(ii) $\|\hat{C}_{k,n} - C_{k,n}\|_2 = O_{P_{\theta_0}}\left(\sqrt{\frac{B_{k,n}\log B_{k,n}}{n\delta_{k,n}^2}}\right)$,

(iii) $\|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_2 = O_{P_{\theta_0}}\left(\sqrt{\frac{B_{k,n}\log B_{k,n}}{n}}\right)$,

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

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

In particular, $\|\hat{\gamma}_{k,n} - \gamma_{k,n}\|_2 = O_{P_{\theta_0}}(n^{-1/2}\Delta_{k,n}\delta_{k,n}^{-4}(\Delta_{k,n}\delta_{k,n}^{-1})^{\iota}) = o_{P_{\theta_0}}(1)$ *and* $\|\hat{\Gamma}_{k,n}\|_2 = o_{P_{\theta_0}}(1)$.

Proof. The proof follows the relevant parts of the proof of lemma A.2 in Chen and Bickel (2006). Firstly, from the representation of the derivative of the cubic spline in (32) we can write $c_{k,n,i} = \left(b_{k,n,i}^{(3)} - b_{k,n,i+1}^{(3)}\right)/\delta_{k,n}$. We have, for large enough $n \in \mathbb{N}$,

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

where the last inequality is due to (20) on p. 91 in de Boor (2001) and the fact that splines (of any order) take values in $[0,1]$.[34] It follows immediately that for large enough $n \in \mathbb{N}$,

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

from which (i) follows immediately.

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

$$P_{\theta_0}\left(\left|\frac{1}{n}\sum_{i=1}^n c_{k,n,m}(\epsilon_{k,i}) - G_k c_{k,n,m}\right| \geq t\right) \leq 2\exp\left(\frac{-n^2t^2}{2n\delta_{k,n}^{-2}}\right) = 2\exp(-nt^2\delta_{k,n}^2/2).$$

Therefore,

$$\begin{aligned} P_{\theta_0}\left(\|\hat{C}_{k,n} - C_{k,n}\|_2 \geq t\right) &\leq \sum_{m=1}^{B_{k,n}} P_{\theta_0}\left(\left|\frac{1}{n}\sum_{i=1}^n c_{k,n,m}(\epsilon_{k,i}) - G_k c_{k,n,m}\right| \geq \frac{t}{\sqrt{B_{k,n}}}\right) \\ &\leq 2B_{k,n}\exp(-nt^2 B_{k,n}^{-1}\delta_{k,n}^2/2), \end{aligned}$$

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

$$P_{\theta_0}\left(\|\hat{C}_{k,n} - C_{k,n}\|_2 \geq t\right) \leq 2B_{k,n}^{-1} \to 0,$$

yielding (ii).

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

$$P_{\theta_0}\left(\left|\frac{1}{n}\sum_{i=1}^n b_{k,n,m}(\epsilon_{k,i})b_{k,n,s}(\epsilon_{k,i}) - G_k b_{k,n,m}b_{k,n,s}\right| \geq t\right) \leq 2\exp\left(\frac{-n^2t^2}{n}\right) = 2\exp(-nt^2).$$

Therefore, since $\|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_2 \leq \|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_F$ and both $\hat{\Gamma}_{k,n}$ and $\Gamma_{k,n}$ are zero for all (m,s) entries where $|m-s| > 3$ (de Boor, 2001, (20), p. 91) we have that

$$\begin{aligned} P_{\theta_0}\left(\|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_2 \geq t\right) &\leq P_{\theta_0}\left(\|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_F \geq t\right) \\ &\leq \sum_{m=1}^{B_{k,n}} \sum_{s=\max(m-3,1)}^{\min(B_{k,n},m+3)} P_{\theta_0}\left(\left|\frac{1}{n}\sum_{i=1}^n b_{k,n,m}(\epsilon_{k,i})b_{k,n,s}(\epsilon_{k,i}) - G_k b_{k,n,m}b_{k,n,s}\right| \geq \frac{t}{\sqrt{7B_{k,n}}}\right) \\ &\leq 14B_{k,n}\exp\left(-nt^2B_{k,n}^{-1}\right). \end{aligned}$$

[34] This is evident from their definition in (32). See also property (36) (p. 96) of de Boor (2001).

Putting $t = \sqrt{\frac{2B_{k,n}\log B_{k,n}}{7n}}$ we obtain

$$P_{\theta_0}\left(\|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_2 \geq t\right) \leq 14B_{k,n}^{-1} \to 0,$$

yielding (iii).

Since $\Gamma_{k,n}$ is symmetric and positive (semi-)definite we have that $\|\Gamma_{k,n}\|_2 \leq \|\Gamma_{k,n}\|_\infty = \max_{m=1,\ldots,B_{k,n}} \sum_{s=1}^{B_{k,n}} G_k b_{k,n,m} b_{k,n,s}$.[35] Then, since for any $z \in \mathbb{R}$, each row of $b_{k,n}(z)b_{k,n}(z)'$ has at most 7 non-zero entries,[36] all of which are bounded above by 1 we have

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

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

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

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

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

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

which yields (v).

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

$$\|\hat{C}_{k,n}\|_2 \leq \|\hat{C}_{k,n} - C_{k,n}\|_2 + \|C_{k,n}\|_2 = O_{P_{\theta_0}}\left(\sqrt{\frac{B_{k,n}\log B_{k,n}}{n\delta_{k,n}^2}}\right) + O\left(\delta_{k,n}\sqrt{B_{k,n}}\right) = O_{P_{\theta_0}}\left(\delta_{k,n}\sqrt{B_{k,n}}\right),$$

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

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

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

$$\begin{aligned}
\|\hat{\Gamma}_{k,n}^{-1}\|_2 &\leq \|\Gamma_{k,n}^{-1}(I + [\hat{\Gamma}_{k,n} - \Gamma_{k,n}]\Gamma_{k,n}^{-1})^{-1}\|_2 \\
&\leq \|\Gamma_{k,n}^{-1}\|_2 \|(I + [\hat{\Gamma}_{k,n} - \Gamma_{k,n}]\Gamma_{k,n}^{-1})^{-1}\|_2 \\
&\leq \|\Gamma_{k,n}^{-1}\|_2 \left(1 - \|[\hat{\Gamma}_{k,n} - \Gamma_{k,n}]\Gamma_{k,n}^{-1}\|_2\right)^{-1} \\
&\leq \|\Gamma_{k,n}^{-1}\|_2 \left(1 - \|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_2 \|\Gamma_{k,n}^{-1}\|_2\right)^{-1} \\
&= O_{P_{\theta_0}}(\delta_{k,n}^{-2}).
\end{aligned} \tag{50}$$

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

$$\begin{aligned}
\|\hat{\gamma}_{k,n} - \gamma_{k,n}\|_2 &= \|\hat{\Gamma}_{k,n}^{-1}\hat{C}_{k,n} - \Gamma_{k,n}^{-1}C_{k,n}\|_2 \\
&\leq \|(\hat{\Gamma}_{k,n}^{-1} - \Gamma_{k,n}^{-1})\hat{C}_{k,n}\|_2 + \|\Gamma_{k,n}^{-1}(\hat{C}_{k,n} - C_{k,n})\|_2 \\
&\leq \|\Gamma_{k,n}^{-1}\|_2 \|\Gamma_{k,n} - \hat{\Gamma}_{k,n}\|_2 \|\hat{\Gamma}_{k,n}^{-1}\|_2 \|\hat{C}_{k,n}\|_2 + \|\Gamma_{k,n}^{-1}\|_2 \|\hat{C}_{k,n} - C_{k,n}\|_2 \\
&= O_{P_{\theta_0}}\left(\sqrt{\frac{B_{k,n}^2 \log B_{k,n}}{\delta_{k,n}^6 n}}\right) + O_{P_{\theta_0}}\left(\sqrt{\frac{B_{k,n} \log B_{k,n}}{\delta_{k,n}^6 n}}\right) \\
&= o_{P_{\theta_0}}(1),
\end{aligned}$$

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

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

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

$$\|\hat{\Gamma}_{k,n}\|_2 \leq \|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_2 + \|\Gamma_{k,n}\|_2 = O_{P_{\theta_0}}\left(\sqrt{\frac{B_{k,n}\log B_{k,n}}{n}}\right) + O(\delta_{k,n}) = o_{P_{\theta_0}}(1),$$

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

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

□

Lemma 8. *Suppose that* $\check{\epsilon}_{k,i,u} := A_k[Y_i + UM_i]$ *where* $u := \mathrm{vec}(U)$ *for any* $k \in [K]$.[37] *For any such* k *and any* $u, v \in \mathbb{R}^{Kd_M}$ *we have for* $p \geq 1$,

(i) $|\check{\epsilon}_{k,i,u} - \check{\epsilon}_{k,i,v}| \lesssim \|u - v\|_p \|M_i\|_p$,

(ii) $|b_{k,n,m}(\check{\epsilon}_{k,i,u}) - b_{k,n,m}(\check{\epsilon}_{k,i,v})| \lesssim \delta_{k,n}^{-1}\|u - v\|_p \|M_i\|_p$,

[37] See assumption S1 for the definition of M_i.

(iii) $|c_{k,n,m}(\check{\epsilon}_{k,i,u}) - c_{k,n,m}(\check{\epsilon}_{k,i,v})| \lesssim \delta_{k,n}^{-2}\|u-v\|_p\|M_i\|_p$.

Proof. For (i) note that we have

$$\begin{aligned}|\check{\epsilon}_{k,i,u} - \check{\epsilon}_{k,i,v}| &= |A_k[Y_i + UM_i - Y_i - VM_i]| \\ &\leq \|A_k\|_p\|U-V\|_p\|M_i\|_p \\ &\lesssim \|u-v\|_p\|M_i\|_p.\end{aligned}$$

For (ii), use (i) in conjunction with the derivative expression in definition (33) which reveals that $b_{k,n,m}$ is Lipschitz with constant $2\delta_{k,n}^{-1}$. For (iii), the same argument suffices upon noting that the derivative expression implies that $c_{k,n,m}$ is Lipschitz with constant $4\delta_{k,n}^{-2}$. □

Lemma 9. *Suppose that assumption S1 holds along with conditions (ii), (iii) and (v) of proposition 4. Define*

$$\check{\Gamma}_{k,n} := \frac{1}{n}\sum_{i=1}^{n} b_{k,n}(\hat{\epsilon}_{k,i,n})b_{k,n}(\hat{\epsilon}_{k,i,n})', \qquad \Gamma_{k,n} := G_k b_{k,n} b_{k,n}',$$

and

$$\check{C}_{k,n} := \frac{1}{n}\sum_{i=1}^{n} c_{k,n}(\hat{\epsilon}_{k,i}), \quad C_{k,n} := G_k c_{k,n}.$$

then we have:

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

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

(iii) $\|\check{\Gamma}_{k,n} - \Gamma_{k,n}\|_2 = O_{\mathrm{P}}\left(\sqrt{\frac{B_{k,n}(\log B_{k,n}\vee\delta_{k,n}^{-2})}{n}}\right)$.

(iv) $\|\Gamma_{k,n}\|_2 = O_{\mathrm{P}}(\delta_{k,n})$,

(v) $\|\Gamma_{k,n}^{-1}\|_2 = O_{\mathrm{P}}(\delta_{k,n}^{-2})$,

In particular, if $\check{\gamma}_{k,n} := \check{\Gamma}_{k,n}^{-1}\check{C}_{k,n}$ *we have that*

$$\|\check{\gamma}_{k,n} - \gamma_{k,n}\|_2 = O_{\mathrm{P}}\left(\sqrt{\frac{B_{k,n}^2\log B_{k,n}}{\delta_{k,n}^8 n}}\right) = o_{\mathrm{P}^*}(1),$$

and $\|\check{\Gamma}_{k,n}\|_2 = o_{\mathrm{P}}(1)$.

Proof. Results (i), (iv) and (v) follow directly from the corresponding results in lemma 7 on noting that the required conditions (ii), (iii), (v) are the same as the corresponding conditions in lemma 7. Given this, we first prove (ii) and (iii).

For (ii), by lemma 8 we have for any $m \in [B_{k,n}]$ that

$$\left|\frac{1}{n}\sum_{i=1}^{n} c_{k,n,m}(\hat{\epsilon}_{k,i,n}) - c_{k,n,m}(\epsilon_{k,i})\right| \lesssim \frac{1}{n}\sum_{i=1}^{n} \delta_{k,n}^{-2}\|U_n\|_1\|M_i\|_1 = O_{\mathrm{P}}(\delta_{k,n}^{-2}n^{-1/2})\times O_{\mathrm{P}}(1) = O_{\mathrm{P}}(\delta_{k,n}^{-2}n^{-1/2}),$$

and therefore $\|\check{C}_{k,n} - \hat{C}_{k,n}\|_2 = O_{\mathrm{P}}(B_{k,n}^{1/2}\delta_{k,n}^{-2}n^{-1/2})$. Hence using (ii) of lemma 7 we can bound our term by

$$\begin{aligned}\|\check{C}_{k,n} - C_{k,n}\|_2 &\le \|\check{C}_{k,n} - \hat{C}_{k,n}\|_2 + \|\hat{C}_{k,n} - C_{kn}\|_2 \\ &= O_{\mathrm{P}}\left(B_{k,n}^{1/2}\delta_{k,n}^{-2}n^{-1/2}\right) + O_{\mathrm{P}}\left(\sqrt{\frac{B_{k,n}\log B_{k,n}}{n\delta_{k,n}^2}}\right) \\ &= O_{\mathrm{P}}\left(\sqrt{\frac{B_{k,n}\log B_{k,n}}{n\delta_{k,n}^2}}\right).\end{aligned}$$

For (iii), similarly using lemma 8 we have for any $m \in [B_{k,n}]$ that

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^{n} b_{k,n,m}(\hat{\epsilon}_{k,i,n})b_{k,n,s}(\hat{\epsilon}_{k,i,n}) - b_{k,n,m}(\epsilon_{k,i})b_{k,n,s}(\epsilon_{k,i})\right| \\ &\quad\le \frac{1}{n}\sum_{i=1}^{n} |b_{k,n,m}(\hat{\epsilon}_{k,i,n}) - b_{k,n,m}(\epsilon_{k,i})||b_{k,n,s}(\hat{\epsilon}_{k,i,n})| + \frac{1}{n}\sum_{i=1}^{n} |b_{k,n,m}(\epsilon_{k,i})||b_{k,n,s}(\hat{\epsilon}_{k,i,n}) - b_{k,n,s}(\epsilon_{k,i})| \\ &\quad\lesssim \frac{1}{n}\sum_{i=1}^{n} \delta_{k,n}^{-1}\|U_n\|_1\|M_i\|_1 \\ &\quad= O_{\mathrm{P}}(\delta_{k,n}^{-1}n^{-1/2}).\end{aligned}$$

Since $b_{k,n,m}(x)b_{k,n,s}(x) = 0$ for (m, s) with $|m - s| > 3$ (de Boor, 2001, (20), p. 91) we have that $\|\check{\Gamma}_{k,n} - \hat{\Gamma}_{k,n}\|_2 = O_{\mathrm{P}}(B_{k,n}^{1/2}\delta_{k,n}^{-1}n^{-1/2})$ and so by (iii) of lemma 7 we have

$$\begin{aligned}\|\check{\Gamma}_{k,n} - \Gamma_{k,n}\|_2 &\le \|\check{\Gamma}_{k,n} - \hat{\Gamma}_{k,n}\|_2 + \|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_2 \\ &= O_{\mathrm{P}}(B_{k,n}^{1/2}\delta_{k,n}^{-1}n^{-1/2}) + O_{\mathrm{P}}\left(\sqrt{\frac{B_{k,n}\log B_{k,n}}{n}}\right) \\ &= O_{\mathrm{P}}\left(\sqrt{\frac{B_{k,n}(\log B_{k,n} \vee \delta_{k,n}^{-2})}{n}}\right)\end{aligned}$$

Using the just derived results along we have

$$\|\check{C}_{k,n}\|_2 \le \|\check{C}_{k,n} - C_{k,n}\|_2 + \|C_{k,n}\|_2 = O_{\mathrm{P}}\left(\sqrt{\frac{B_{k,n}\log B_{k,n}}{n\delta_{k,n}^2}}\right) + O\left(\delta_{k,n}\sqrt{B_{k,n}}\right) = O_{\mathrm{P}}\left(\delta_{k,n}\sqrt{B_{k,n}}\right),$$

and by an analogous argument to in that of equation (50)

$$\|\check{\Gamma}_{k,n}^{-1}\|_2 = O_{\mathrm{P}}(\delta_{k,n}^{-2}).$$

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

$$\begin{aligned}
\|\check{\gamma}_{k,n} - \gamma_{k,n}\|_2 &= \|\check{\Gamma}_{k,n}^{-1}\check{C}_{k,n} - \Gamma_{k,n}^{-1}C_{k,n}\|_2 \\
&\leq \|(\check{\Gamma}_{k,n}^{-1} - \Gamma_{k,n}^{-1})\check{C}_{k,n}\|_2 + \|\Gamma_{k,n}^{-1}(\check{C}_{k,n} - C_{k,n})\|_2 \\
&\leq \|\Gamma_{k,n}^{-1}\|_2\|\Gamma_{k,n} - \check{\Gamma}_{k,n}\|_2\|\check{\Gamma}_{k,n}^{-1}\|_2\|\check{C}_{k,n}\|_2 + \|\Gamma_{k,n}^{-1}\|_2\|\check{C}_{k,n} - C_{k,n}\|_2 \\
&= O_{\mathrm{P}}\left(\sqrt{\frac{B_{k,n}^2 \log B_{k,n}}{\delta_{k,n}^8 n}}\right) + O_{\mathrm{P}}\left(\sqrt{\frac{B_{k,n} \log B_{k,n}}{\delta_{k,n}^6 n}}\right) \\
&= o_{\mathrm{P}}(1),
\end{aligned}$$

by condition (ii) of lemma 4, since we have $B_{k,n} \leq \Delta_{k,n}\delta_{k,n}^{-1}$ and hence the dominant term above vanishes since for all large enough n,

$$\sqrt{\frac{B_{k,n}^2 \log B_{k,n}}{\delta_{k,n}^8 n}} \leq n^{-1/2}\Delta_{k,n}\delta_{k,n}^{-5}\log(\Delta_{k,n}\delta_{k,n}^{-1}) \leq n^{-1/2}\Delta_{k,n}\delta_{k,n}^{-5}(\Delta_{k,n}\delta_{k,n}^{-1})^{\iota} = o(1).$$

Finally, by (iii), and (iv) and condition (ii) of lemma 4 we have

$$\|\check{\Gamma}_{k,n}\|_2 \leq \|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_2 + \|\Gamma_{k,n}\|_2 = O_{\mathrm{P}}\left(\sqrt{\frac{B_{k,n}(\log B_{k,n} \vee \delta_{k,n}^{-2})}{n}}\right) + O(\delta_{k,n}) = o_{\mathrm{P}}(1),$$

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

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

□

Lemma 10. *In the setting of lemma 4, for $k \in [K]$ we have*

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n}) - \phi_k(\epsilon_{k,i})\right]^2 = o_{\mathrm{P}}(1).$$

Proof. We can upper bound our term:[38]

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n})-\phi_k(\epsilon_{k,i})\right]^2 \leq \frac{8}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n})-\tilde{\phi}_{k,n}(\epsilon_{k,i})\right]^2+\frac{8}{n}\sum_{i=1}^{n}\left[\tilde{\phi}_{k,n}(\epsilon_{k,i})-\phi_{k,n}(\epsilon_{k,i})\right]^2 + \frac{8}{n}\sum_{i=1}^{n}\left[\phi_{k,n}(\epsilon_{k,i})-\tilde{\phi}_k(\epsilon_{k,i})\right]^2,$$

where $\tilde{\phi}_{k,n}$ and $\phi_{k,n}$ are defined as in lemma 3.[39] The proof of the convergence of the latter two terms is very similar to that in lemma 3. By our hypotheses and Cauchy-Schwarz:

$$\begin{aligned} G_k\left([\phi_{k,n}(\epsilon_k)-\phi_k(\epsilon_k)]^2\right) &= G_k\left[\phi_k(\epsilon_k)^2\mathbf{1}\{\epsilon_k\notin[\Xi^L_{k,n},\Xi^U_{k,n}]\}\right] \\ &\leq \left[G_k\phi_k(\epsilon_k)^4\right]^{1/2}\left[G_k\mathbf{1}\{\epsilon_k\notin[\Xi^L_{k,n},\Xi^U_{k,n}]\}\right]^{1/2} \\ &\to 0. \end{aligned} \tag{51}$$

Similarly, by our hypotheses and lemma 5, as $n\to\infty$,[40]

$$G_k\left([\tilde{\phi}_{k,n}(\epsilon_k)-\phi_{k,n}(\epsilon_k)]^2\right)=G_k\left([\tilde{\phi}_{k,n}(\epsilon_k)-\phi_{k,n}(\epsilon_k)]^2\right)\leq C^2\delta^6_{k,n}\|\phi^{(3)}_k\|^2_\infty\to 0. \tag{52}$$

Using the preceding two displays in conjunction with Markov's inequality, for any $\upsilon>0$, we have:

$$\mathrm{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}[\phi_{k,n}(\epsilon_{k,i})-\phi_k(\epsilon_{k,i})]^2\right|>\upsilon\right)\leq\frac{G_k\left([\phi_{k,n}(\epsilon_k)-\phi_k(\epsilon_k)]^2\right)}{\upsilon}\to 0,$$

and

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

which deals with the last two terms in our upper bound. Finally, for the first term in the

[38] Throughout the proof we will use notation introduced in the proof of lemma 4 without comment.

[39] The former is defined during the proof.

[40] Note that (iv) of the hypotheses of lemma 3 is also (iv) of the hypotheses of lemma 4.

decomposition, using lemmas 8 and 9 we have:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(\hat{\epsilon}_{k,i,n})-\tilde{\phi}_{k,n}(\epsilon_{k,i})\right]^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[(\check{\gamma}_{k,n}-\gamma_{k,n})'\,b_{k,n}(\hat{\epsilon}_{k,i,n})+\gamma_{k,n}'\left(b_{k,n}(\hat{\epsilon}_{k,i,n})-b_{k,n}(\epsilon_{k,i})\right)\right]^2 \\
&\lesssim (\check{\gamma}_{k,n}-\gamma_{k,n})'\left[\frac{1}{n}\sum_{i=1}^{n} b_{k,n}(\hat{\epsilon}_{k,i,n})b_{k,n}(\hat{\epsilon}_{k,i,n})'\right](\check{\gamma}_{k,n}-\gamma_{k,n}) \\
&\quad + \|\gamma_{k,n}\|_2^2\frac{1}{n}\sum_{i=1}^{n}\|b_{k,n}(\hat{\epsilon}_{k,i,n})-b_{k,n}(\epsilon_{k,i})\|_2^2 \\
&\lesssim \|\check{\gamma}_{k,n}-\gamma_{k,n}\|_2\|\check{\Gamma}_{k,n,u}\|_2\|\check{\gamma}_{k,n}-\gamma_{k,n}\|_2 \\
&\quad + \|\gamma_{k,n}\|_2^2 B_{k,n}\delta_{k,n}^{-2}\|U_n\|_2^2\frac{1}{n}\sum_{i=1}^{n}\|M_i\|_2^2 \\
&= o_{\mathrm{P}}(1)+O_{\mathrm{P}}\left(\delta_{k,n}^{-2}B_{k,n}\right)\times O_{\mathrm{P}}(B_{k,n}\delta_{k,n}^{-2}n^{-1}) \\
&= o_{\mathrm{P}}(1),
\end{aligned}$$

where we note that $\delta_{k,n}^{-4}B_{k,n}^2 n^{-1} \leq \Delta_{k,n}^2\delta_{k,n}^{-6}n^{-1} = o(1)$, by condition (ii) and the fact that $B_{k,n} \leq \Delta_{k,n}\delta_{k,n}^{-1}$. □

Appendix C: Tables and figures

Table 1: True Error Distributions

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

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

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

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

Notes: The table shows the empirical rejection frequencies for the S_n^{SR} test based on $S = 5.000$ Monte Carlo replications for the baseline ICA model. The test has nominal size $a = 0.05$. The columns denote the sample size n, the dimension of the ICA model K, the number of B-splines B and the choice for densities ϵ_k, for $k > 1$, where the numbers correspond to the different densities listed in Table 1.

Table 3: Empirical Rejection Frequencies Alternative Tests for Baseline ICA

Test	1	2	3	4	5	6	7	8	9	10
W	0.257	0.231	0.187	0.092	0.282	0.076	0.022	0.176	0.188	0.252
LM	0.072	0.090	0.075	0.065	0.109	0.063	0.069	0.065	0.066	0.087
LR	0.011	0.035	0.045	0.050	0.045	0.035	0.021	0.000	0.001	0.026
LR^G	0.428	0.243	0.174	0.057	0.406	0.019	0.005	0.989	0.960	0.296
LR^L	0.164	0.141	0.106	0.092	0.149	0.168	0.345	0.117	0.112	0.161
W_p	-	0.376	0.271	0.095	0.362	0.114	0.021	0.207	0.150	0.448
LM_p	-	0.129	0.102	0.132	0.086	0.051	0.068	0.226	0.250	0.147
LR_p	-	0.070	0.116	0.055	0.099	0.055	0.021	0.018	0.050	0.102
LR_p^G	-	0.411	0.289	0.062	0.448	0.047	0.005	0.962	0.950	0.511
LR_p^L	-	0.223	0.280	0.231	0.254	0.163	0.345	0.320	0.100	0.426

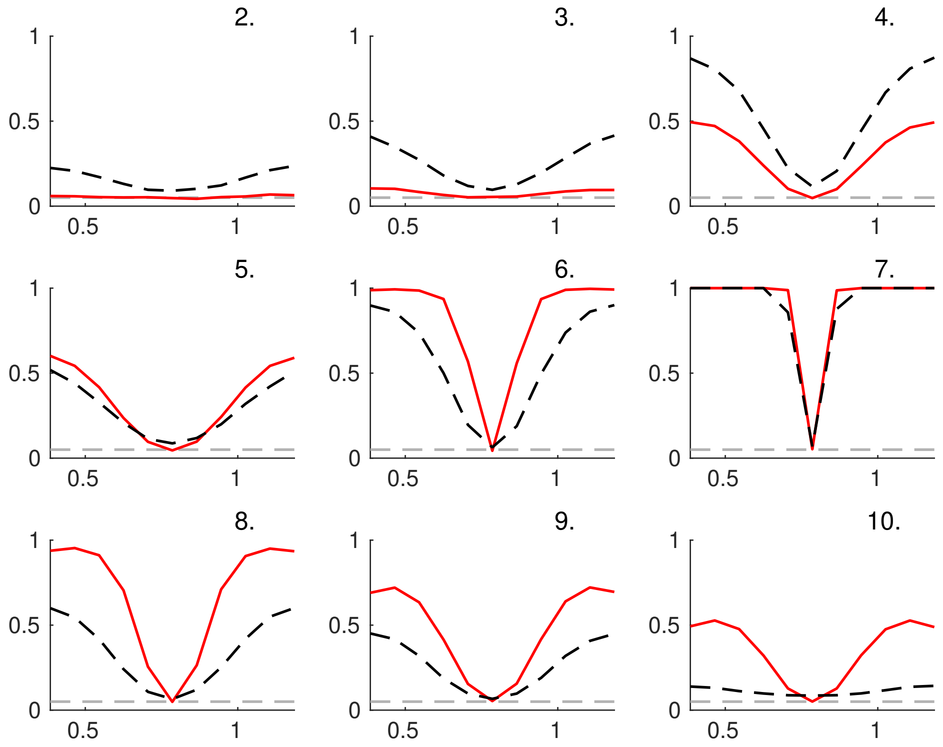
Notes: The table shows the empirical rejection frequencies based on $S = 5.000$ Monte Carlo replications for the baseline ICA model with $n = 500$ and $K = 2$. All tests have nominal size $a = 0.05$. The first column indicates the test. In particular, W denotes the MLE-based Wald test, LM denotes the MLE-based Lagrange multiplier test, LR denotes the MLE-based likelihood ratio test, LR^G denotes the likelihood ratio test based on the psuedo-maximum likelihood estimator of Gouriéroux, Monfort and Renne (2017), LR^L denotes the likelihood ratio test based on the GMM estimator of Lanne and Luoto (2019*a*). Finally, the subscript p denotes the same test computed conditional on passing the Jarque-Berra pre-test. The remaining columns denote the choice for densities ϵ_k, for $k \geq 2$, where the numbers correspond to the different densities listed in Table 1.

Table 4: Empirical Rejection Frequencies $\hat{S}_n^{SR}$ test for SVAR

n	K	q	1	2	3	4	5	6	7	8	9	10
200	2	1	0.069	0.089	0.088	0.113	0.081	0.086	0.167	0.072	0.072	0.070
200	2	2	0.085	0.091	0.094	0.133	0.087	0.093	0.180	0.079	0.082	0.075
200	2	4	0.112	0.126	0.120	0.164	0.117	0.114	0.206	0.106	0.104	0.110
200	3	1	0.090	0.095	0.111	0.163	0.093	0.095	0.303	0.069	0.070	0.073
200	3	2	0.093	0.105	0.110	0.162	0.107	0.104	0.311	0.079	0.088	0.080
200	3	4	0.126	0.135	0.147	0.208	0.126	0.116	0.303	0.122	0.117	0.109
500	2	1	0.062	0.061	0.066	0.089	0.051	0.069	0.109	0.056	0.055	0.052
500	2	2	0.057	0.062	0.067	0.093	0.058	0.062	0.099	0.057	0.058	0.051
500	2	4	0.070	0.072	0.083	0.106	0.068	0.072	0.110	0.063	0.064	0.059
500	3	1	0.059	0.069	0.063	0.109	0.061	0.072	0.162	0.053	0.047	0.040
500	3	2	0.056	0.064	0.077	0.111	0.066	0.070	0.156	0.056	0.058	0.051
500	3	4	0.084	0.086	0.088	0.136	0.070	0.073	0.167	0.081	0.074	0.063
1000	2	1	0.056	0.050	0.057	0.067	0.045	0.052	0.076	0.045	0.048	0.041
1000	2	2	0.050	0.052	0.049	0.067	0.048	0.050	0.080	0.050	0.047	0.043
1000	2	4	0.058	0.057	0.062	0.083	0.049	0.053	0.074	0.052	0.055	0.044
1000	3	1	0.043	0.046	0.055	0.091	0.045	0.052	0.102	0.044	0.040	0.045
1000	3	2	0.050	0.049	0.054	0.084	0.046	0.059	0.100	0.043	0.045	0.048
1000	3	4	0.054	0.061	0.059	0.091	0.051	0.058	0.117	0.058	0.052	0.042

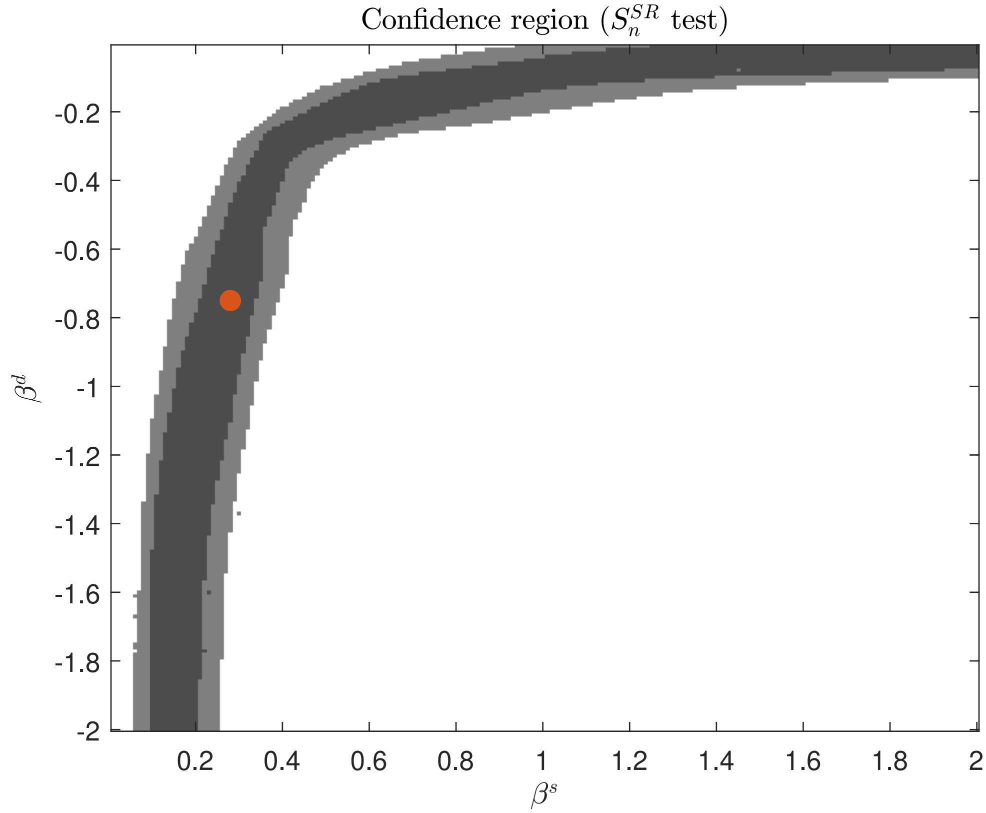
Notes: The table shows the empirical rejection frequencies for the S_n^{SR} test based on $S = 5.000$ Monte Carlo replications for the SVAR model. The test has nominal size $a = 0.05$. The columns denote the sample size n, the dimension of the ICA model K, the number of lags included q and the choice for densities ϵ_k, for $k \geq 2$, where the numbers correspond to the different densities listed in Table 1. The S_n^{SR} test was implemented using $B = 6$ B-splines.

Figure 1: Power Baseline ICA model: $n = 500$



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

Figure 2: Confidence region Labor Elasticities



Notes: Confidence regions (light gray is 95% and dark grey is 67%) for β^d and β^s in the model for US labor.