Robust and efficient inference for non-regular semiparametric models

Adam Lee*

Job Market Paper

This version: January 6, 2022

[Click here for the latest version]

Abstract

This paper considers hypothesis testing problems in semiparametric models which may be non-regular for certain values of a potentially *infinite dimensional* nuisance parameter. I establish that, under mild regularity conditions, tests based on the efficient score function provide locally uniform size control and enjoy minimax optimality properties. This approach is applicable to situations with (i) identification failures, (ii) boundary problems and (iii) distortions induced by the use of regularised estimators. Full details are worked out for two examples: a single index model where the link function may be relatively flat and a linear simultaneous equations model that is (weakly) identified by non-Gaussian errors. In practice the tests are easy to implement and rely on χ^2 critical values. I illustrate the approach by using the linear simultaneous equations model to examine the labour supply decisions of men in the US. I find a small but positive effect of wage increases on hours worked for hourly paid workers, but no effect for salaried workers.

JEL classification: C10, C12, C14, C21, C39

Keywords: Hypothesis testing, local asymptotics, uniformity, semiparametric models, weak identification, boundary, regularisation, single-index, simultaneous equations.

*Universitat Pompeu Fabra (UPF) and Barcelona GSE. Email: adam.lee@upf.edu. Website: adamjclee.github.io. I thank my advisor, Geert Mesters, for his guidance. I have benefitted from discussions with and comments from Christian Brownlees, Bjarni G. Einarsson, Kirill Evdokimov, Lukas Hoesch, Vladislav Morozov, Katerina Petrova, Barbara Rossi, André B. M. Souza, Philipp Tiozzo as well as participants at the 2021 Barcelona GSE PhD Jamboree, the 2021 Symposium of the Spanish Economics Association, the 2021 SNDE Workshop for Young Researchers and the statistics seminar at UPF. All errors are my own.

1 Introduction

In many econometric models, the behaviour of commonly used inference procedures can depend crucially on the value of nuisance parameters. There are many cases where the asymptotic distributions of test statistics derived using standard (fixed parameter) arguments provide poor approximations to the finite sample distribution for certain values of nuisance parameters. When this occurs, the corresponding tests justified by such asymptotic arguments may have (finite sample) size far in excess of the nominal level.

In this paper I develop a general framework for conducting inference on a finite dimensional parameter in a semiparametric model, robust to (sequences of) values of a possibly infinite dimensional nuisance parameter which may invalidate standard inference methods. In particular, the main contribution of this paper is to show that semiparametric score tests based on the efficient score function (e.g. Bickel, Klaassen, Ritov, and Wellner, 1998; van der Vaart, 2002) are robust under mild assumptions which allow for, among others, (i) identification failure, (ii) nuisance parameters on the boundary and (iii) the use of regularised estimates of nuisance parameters.

Importantly – and unlike other general approaches put forward in the robust inference literature (e.g. Andrews and Guggenberger, 2009, McCloskey, 2017 and Elliott, Müller, and Watson, 2015) – this approach permits the nuisance parameter which causes standard inferential approaches to break down to be *infinite dimensional*.

A key benefit of this approach is that this efficient score test does not sacrifice power in order to obtain this robustness: when classical regularity conditions hold, the test enjoys classical optimality properties. Additionally, I demonstrate that the test is minimax optimal in some cases which fall in-between classical regularity conditions and the weaker conditions under which the robustness results of this paper are obtained. Such results apply, for example, when the parameter of interest is underidentified. Moreover these tests are often easy to compute and require only χ^2 critical values.

The semiparametric models I consider are parametrised by a pair $\gamma = (\theta, \eta)$ where θ is the parameter of interest and η collects all nuisance parameters (and is therefore typically infinite dimensional). γ fully parametrises the distribution of the observed data and I write the corresponding probability law as P_γ. This setup permits a large range of models regularly used in practice and includes both traditional parametric models and models defined by moment conditions as special cases.

The theoretical results of this paper are derived under a few high level conditions, for which some more primitive conditions are given subsequently. The main condition is local asymptotic normality (LAN) of the model, which implicitly defines score functions for θ and η. LAN specifies that the logarithms of certain likelihood ratios posses a local quadratic approximation and – in the i.i.d. case considered in this paper – can be demonstrated to hold under an L_2-differentiability condition known as "differentiability in quadratic mean" (DQM).[1] Such conditions are common in the semiparametric statistical theory as expounded by e.g. Bickel et al. (1998) or van der Vaart (2002).[2] This literature usually complements LAN (or DQM) with additional regularity conditions, such as (a) the non-singularity of information matrices and (b) all parameters lying in the interior of the parameter space.[3] These conditions rule out a number of cases of interest in econometrics. For example, (a) the non-singularity of the information matrix is often violated when the parameter of interest is under- or un-identified; (b) many model specifications permit nuisance parameters to lie on the boundary. Fortunately, as I show in this paper, valid inference can be conducted without these additional conditions.[4]

With the LAN condition in hand, the *efficient score function* (for the parameter of interest) can be defined as the orthogonal projection (in L_2) of the score function for θ on the orthocomplement of the set of score functions for η. This efficient score function is the basis of the robust inferential theory put forward in this paper. The main test statistic I consider, the *efficient score statistic*, is the quadratic form of an estimate of the efficient score function, weighted by a (pseudo-)inverse of its (estimated) variance matrix.[5] The key insight I exploit is that – under the null – the limiting distribution of the efficient score function is the same regardless of the (local) nuisance parameter sequence along which the limit is taken. This directly leads to robustness of the efficient score test against such sequences and consequently that such tests control size in a (locally) uniform manner over certain compact subsets. In contrast, there are many models in which this property fails to hold for commonly used test statistics: different sequences of nuisance parameters consistent with the null hypothesis result in different limiting distributions.

Moving from size to power, the efficient score test has attractive optimality properties if the possible local nuisance parameter values are indexed by a linear space.[6] Firstly, if the covariance matrix of the efficient score function is non-singular then the efficient score test is asymptotically uniformly most powerful within the class of asymptotically invariant tests as defined and demonstrated by Choi et al. (1996).[7] Moreover, if the covariance matrix of the

[1]See e.g. Le Cam and Yang (2000, Chapters 6 and 7).

[2]Similar quadratic expansions of an objective function have also been previously used to analyse nonstandard models in econometrics. See, for instance, Andrews (2001); Andrews and Cheng (2012).

[3]Cf. e.g. Definitions 2.1.1, 2.1.2 and 3.1.1 of Bickel et al. (1998).

[4]Cf. section 6.9 of Le Cam and Yang (2000) where the authors explicitly discuss a number of simplifying assumptions which are often made but are not essential. Their point (v), that "the points ... are interior points of $\Theta \in \mathbb{R}^k$" is clearly directly relevant to the case (b) with parameters potentially on the boundary. For (a), where un- or under-identification of the parameter of interest may cause singularity of the information matrix, cf. Le Cam and Yang, 2000, example (a), pp. 56 - 57.

[5]When the variance matrix is non-singular, the corresponding *efficient score test* is the same as the "effective score test" of Choi, Hall, and Schick (1996). Additionally, the efficient score statistic can be viewed as the semiparametric analogue of Neyman's C(α) statistic (Neyman, 1959, 1979).

[6]This is often – but not always – the case. It fails, for example, at boundary points of the parameter space. See Rieder (2014) for a discussion and some optimality results in such cases.

[7]For scalar parameters the asymptotic invariance can be replaced by asymptotic unbiasedness for two-sided

efficient score function has positive rank, I establish that the test enjoys a local asymptotic minimax optimality property. In addition to the standard full rank case, this situation may arise when the parameter of interest is underidentified.

I work out the details of the application of the general theory to two econometric models: a single index model where the link function may be relatively flat compared to sampling variation and a linear simultaneous equations model where identification may be weak when an identifying assumption of non-Gaussianity is close to failing. In each case, the models have nonstandard features which can invalidate some standard approaches to inference. For each model I give primitive conditions that allow (i) derivation of the efficient score function and (ii) a demonstration that the high level conditions required for the application of the previously developed theory are satisfied. Crucially, the assumptions imposed do not carve out parts of the parameter space which cause problems for other testing approaches.

Firstly, I consider a single index model (SIM). The SIM is a popular model in econometrics as it retains a large amount of flexibility whilst successfully combating the curse of dimensionality. Identification of parameters in the index function requires a number of assumptions, including the non-constancy of the link function. As is usual with points of identification failure, if the link function is sufficiently close to constancy relative to the sample size, a weak identification problem obtains. Importantly, the identification status of the parameter of interest in this model depends on the link function, an infinite dimensional nuisance parameter. Additionally regularised estimation is required to perform inference in this model. I demonstrate that the efficient score test provides (locally uniformly) valid size control in spite of these issues.

Secondly, I examine a semiparametric linear simultaneous equations model (LSEM). The LSEM is a foundational model in econometrics, used to analyse equilibrium relationships. As is well known, the simultaneity problem precludes the identification of all structural parameters from observed data without further restrictions, leading researchers to adopt alternative methods (e.g. analysing only one equation with the help of instrumental variable techniques); see Dhrymes (1994) for an in-depth review.

In fact, the identification status of the structural parameters of interest depends on the true error distribution (an infinite dimensional nuisance parameter). In particular, if no more than one of the (mutually independent) error components is Gaussian the structural parameters are identified as a consequence of the Darmois-Skitovich Theorem (Comon, 1994).[8] If multiple components are Gaussian the structural parameters may be under- or un-identified and standard inferential approaches may fail to control size. As is typical in models with points of identification failure, such behaviour is also observed if the true error distributions are sufficiently to close to Gaussianity, relative to sampling variation.

tests; for one-sided tests the asymptotic optimality holds over all tests of correct asymptotic level.

[8]Strictly speaking the identification result is up to column permutations and sign changes of the matrix which transforms the structural shocks into reduced form shocks.

In addition to these potential identification problems, regularised estimation is required to handle the non-parametric part of the model, leading to regularisation bias. I demonstrate that despite the presence of these non-standard features, the efficient score test provides (locally uniformly) valid and efficient inference in the LSEM model, providing researchers with a direct approach to conduct inference on structural parameters in linear simultaneous systems without needing to employ, for example, instrumental variables approaches.

I conduct a large scale simulation study based on each example. The results verify that the asymptotic size results obtained provide a good guide to finite sample size, with the efficient score test always being correctly sized, including in cases where alternative procedures fail to correctly control size. The simulation studies also highlight the power of this testing approach and suggest that the asymptotic approximations provide a good guide to finite sample power, with finite sample power curves and surfaces matching the predictions of the asymptotic theory.

To illustrate the practical application of the approach, I use the LSEM to examine the labour supply behaviour of US men. Wages and hours are typically considered to form a simultaneous system. If the distribution of the error terms in this system is not (local to) Gaussian, this approach permits identification of the structural parameters of interest in the presence of this simultaneity without, for instance, instrumental variables.[9] I find a small but positive effect of wage increases on hours worked for hourly paid workers, but no effect for salaried workers.

1.1 Relation to the literature

This paper is primarily a contribution to the literature on general approaches to robust inference methods for statistical and econometric models with non-standard asymptotic behaviour in part of the parameter space.

A number of papers analyse size-correction methods to provide inference valid uniformly over nuisance parameter values. For instance, Andrews and Guggenberger (2009, 2010a,b) analyse the use of resampling methods and data-dependent critical values to provide uniformly correct size control over the parameter space; McCloskey (2017) provides alternative size correction approaches based on Bonferroni bounds, which can improve the power of such size corrected tests. The approaches proposed in the cited papers are designed for models in which a statistic has a limiting distribution which is discontinuous in a finite-dimensional nuisance parameter.[10] This setup is very general but differs from the one considered in the present paper on a number of key points: (i) in this paper, the parameter which may cause

[9] If it *were* local to Gaussian, then as the results of this paper show, the testing procedures used would continue to be correctly sized.

[10] In related work, Andrews, Cheng, and Guggenberger (2020) provide some general results to establish the (uniform) size of tests and (uniform) coverage probabilities of confidence sets based on (pointwise) asymptotic distributions which are discontinuous in some function of a parameter.

standard inferential approaches to suffer from size distortions can be infinite dimensional; (ii) rather than size-correcting tests based on a specific test statistics which have parameter discontinuous asymptotic distributions, I suggest the use of the the efficient score statistic which always has a χ^2 distribution and hence the tests always use χ^2 critical values. There is not complete overlap between the class of models considered in this paper and those to which the methods in these papers are applicable: the efficient score test remains valid in cases where the asymptotic distribution of (other) test statistics may depend on the particular local sequence of infinite dimensional nuisance parameters. Conversely, the example of an autoregressive model with a root which may be local to unity studied in Andrews and Guggenberger (2009) does not satisfy the high-level conditions I impose as such models are locally asymptotically quadratic (LAQ) but not LAN (Jeganathan, 1995; Jansson, 2008).

Romano and Shaikh (2012) provide high level conditions under which bootstrap and subsampling procedures yield tests and confidence sets with (uniformly) correct size and coverage probabilities in a very general class of models. Their approach differs substantially from the approach in this paper, using resampling schemes to provide appropriate quantiles to conduct tests and construct confidence sets for the values of general parameters of interest defined on the model. As a result, their approach can deal with more general parameters of interest than are considered in this paper. On the other hand, there are cases in which the procedure outlined in this paper correctly controls size, but subsampling and bootstrapping approaches fail to do so, for example, subsampling TSLS t-type statistics in IV regression models with weak instruments (Andrews and Guggenberger, 2010a) and subsampling Wald-type statistics in models with nuisance parameters near the boundary (Andrews and Guggenberger, 2010b).

Elliott et al. (2015) provide nearly optimal tests for models which have a Gaussian shift limit experiment (locally to the true parameter) with part of the shift vector being a nuisance parameter. Their tests correctly control size and (approximately) maximise weighted average power given a weighting function (over the nonstandard region of the parameter space). Their approach requires the nuisance parameter to be finite dimensional and is quite different from the one proposed in this paper, though it shares some common threads, being based on a least favourable approach in a Gaussian shift limit experiment.[11]

For numerous classes of nonstandard inference problems a large literature exists analysing the problem at hand and providing particular solutions. There are too many such examples to provide a full account here; instead I provide a selective summary of the literature pertaining to those non-standard features relevant to the examples I consider in detail in this paper, comprising (a) identification robust inference, (b) inference in models with boundary constraints and (c) inference post a model selection or regularisation step.

[11] I do not consider least favourable distributions explicitly, however the efficient score function can be considered to correspond to an approximately least favourable submodel; see §25.11 in van der Vaart (1998).

Inference robust to identification problems has been considered in various settings by, inter alia, Stock and Wright (2000); Kleibergen (2005); Andrews and Cheng (2012, 2013); Andrews and Mikusheva (2015, 2016a,b, forthcoming); Han and McCloskey (2019); Andrews and Guggenberger (2019).[12] Dufour (1997) provides some impossibility results. Chen, Christensen, and Tamer (2018) consider semiparametric models in which parameters may be only partially identified and suggest inferential procedures based on a Monte Carlo simulation approach. Kaji (2021) puts forward a general theory of weak identification in semiparametric models and focusses on efficient estimation rather than robust inference.

A long considered problem is inference in models with boundary constraints, which has been studied by, amongst others, Chernoff (1954); Geyer (1994); Andrews (2000, 2001); Andrews and Guggenberger (2010a,b); Chen, Ning, Ning, Liang, and Bandeen-Roche (2017); Ketz (2018); Cavaliere, Nielsen, Pedersen, and Rahbek (2020). An antecedent to the approach of this paper in the case of nuisance parameters potentially on (or close to) the boundary can be found in Andrews (2001, p. 698) where the nuisance parameters are split into those which satisfy a block diagonality condition with respect to the other parameters and those which do not. The author of that paper then notes that those which satisfy the block diagonality condition "may or may not lie on the boundary of the parameter space". I exploit a similar idea, as the efficient score function is orthogonal to *all* nuisance scores by construction.

Inference post model selection or regularisation is also problem with a long history, which has become increasingly important in recent years due to the increasing availability of "big data". Leeb and Pötscher (2005) analyse in detail some of the difficulties associated with inference post model selection; additional demonstrations along with applications of some of the size correction approaches previously mentioned can be found in Andrews and Guggenberger (2010a); McCloskey (2020). Chernozhukov, Hansen, and Spindler (2015) outline an approach to post model selection / post regularisation inference which uses an approach similar to the one proposed in this paper with their class of "Neyman orthogonalised" statistics also being a generalisation of the C(α) approach of Neyman (1959, 1979).[13] The development in their paper is framed somewhat differently and focusses on post-regularisation inference in problems defined by a finite vector of known moment conditions with a larger class of test statistics, whereas I consider a more general class of inference problems with potentially non-standard features but only one test statistic.[14]

[12] There is also a large literature on robust inference in models defined by moment inequalities (and partially identified models more generally). Additionally a further sub-literature exists on subvector inference for weakly identified parameters. I do not consider subvector inference in this sense in this paper, though I note here that Chaudhuri and Zivot (2011) used the efficient score corresponding to a GMM model as a way to improve power in projection-based subvector inference with weak identification.

[13] See also Belloni, Chernozhukov, Fernández-Val, and Hansen, 2017 and Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins, 2018.

[14] In many models, the test statistic considered in this paper would belong to the general class they consider.

The general approach to inference outlined in this paper is based on the efficient score function which, along with its variance matrix (the "efficient information matrix"), is a key quantity in the literature on semiparametric efficiency. Textbook treatments of this framework can be found in Bickel et al. (1998); van der Vaart (2002) and van der Vaart (1998, Chapter 25). The efficient score test was shown to be optimal (in certain classes of tests) by Choi et al. (1996). These ideas have been widely used in statistics and econometrics since their introduction, particularly to determine efficiency bounds in semiparametric models and construct estimators which attain them.

I now briefly turn to the specific examples I consider. The first – inference in the single index model with potential identification failure – is related to the (previously summarised) literatures on inference with potential identification problems and inference post-regularisation as well as the literature on single index models and extensions thereof. Such models have been widely studied by, amongst others, Ichimura (1993); Newey and Stoker (1993); Ma and Zhu (2013).

The second example I consider, the LSEM, is related to the (previously summarised) literatures on inference with potential identification problems and inference post-regularisation as well as the statistical literature on independent components analysis (ICA) modelling. The ICA model has long been used in a number of fields as an approach to the analysis of data forming systems of simultaneous equations; see Hyvärinen, Karhunen, and Oja (2001) for many examples.[15] By adding covariates to the ICA model a class of linear simultaneous equations models is obtained. Such systems of equations have a long history in econometrics; see the introduction of Lee and Mesters (2021a) for a summary.[16] A semiparametric approach to the ICA model was considered in Amari and Cardoso (1997); Chen and Bickel (2006). Lee and Mesters (2021a) consider a semiparametric approach to the LSEM which uses the approach discussed in this paper to conduct tests robust to potential identification failure. Concretely, they consider testing when the (fixed) distribution of the error terms may be arbitrarily close to Gaussianity but this distribution is not permitted to change with the sample size. They provide simulation evidence of a weak identification problem when the error distribution is sufficiently close to Gaussianity (relative to the sample size), but their theoretical work assumes a fixed error distribution and consequently does not cover weak identification. In contrast, in this paper, I explicitly model weak identification and obtain size results which are valid locally uniformly over subsets of the parameter space.

[15]The ICA model relates observables Y and errors ϵ according to $Y = A^{-1}\epsilon, \quad \mathbb{E}\epsilon = 0,\ \mathbb{V}\epsilon = I$ where ϵ has independent components.

[16]More recently such models have also been adopted in econometrics as an approach to SVAR modelling, with an assumption of non-Gaussianity imposed to identify the matrix required to obtain the structural shocks from the reduced form shocks. A recent summary of this approach is given by Montiel Olea, Plagborg-Møller, and Qian (2021). Also see, inter alia, Gouriéroux, Monfort, and Renne (2017, 2019); Lanne and Lütkepohl (2010); Lanne, Meitz, and Saikkonen (2017); Lanne and Luoto (2021); Bekaert, Engstrom, and Ermolov (2019, 2020); Fiorentini and Sentana (2021, 2020); Davis and Ng (2021). Velasco (2020) considers the more general SVARMA case. In this paper I do not consider dynamics for simplicity.

1.2 Outline

The remainder of this paper is organised as follows. Section 2 describes the setting of the paper, explains the intuition underlying the testing approach and introduces a number of examples. Section 3 formalises the heuristic definitions given previously, develops the theoretical contributions of this paper under high level conditions and provides some lower-level conditions and constructions sufficient for their validity. Two examples are worked out in detail in sections 4 and 5; these sections also discuss the results from several simulation studies. Section 6 highlights the results from an empirical study into the labour supply decisions of US men. Section 7 concludes and discusses possible extensions.

2 Heuristic explanation and examples

I now provide a heuristic discussion of the efficient score test, focussing on the underlying intuition, and provide a number of examples to demonstrate the breadth of applicability of my framework. I purposely omit all formal definitions and assumptions, which are provided in section 3 below.

The parameter of interest is $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ and the goal is to construct (asymptotically) correctly sized tests for the hypothesis $H_0 : \theta = \theta_0$ or confidence sets for θ which have correct (asymptotic) coverage probability over a range of data generating processes (DGPs) consistent with the null hypothesis.

I suppose that the researcher observes a random sample $(W_i)_{i=1}^n$. The considered probability model for the distribution of each such observation W_i is given by

$$\mathcal{P} = \{P_\gamma : \gamma \in \Gamma\}, \quad \Gamma = \Theta \times \mathcal{H}, \tag{1}$$

where $\gamma = (\theta, \eta)$ with η collecting all the remaining parameters required to fully describe the distribution of the data (given θ). In the classical parametric setting η is finite dimensional; in the semiparametric models which are the focus of this paper it may be infinite dimensional.

Analogously to the parametric case, it is possible to define *score functions* for all of the parameters in semiparametric models (see section 3 for the details). Let $\dot{\ell}_\gamma$ be the (vector of) score functions for θ and $\mathscr{H}_\gamma = \{B_\gamma h : h \in H\}$ a collection of score functions for η.[17] All score functions are mean zero and have finite variance. The *efficient score function* is defined as the orthogonal projection (in L_2) of the scores for θ onto the orthogonal complement of the scores for η:

$$\tilde{\ell}_\gamma = \dot{\ell}_\gamma - \Pi\left(\dot{\ell}_\gamma \middle| \overline{\operatorname{lin}}\, \mathscr{H}_\gamma\right), \tag{2}$$

[17] The score functions are indexed by elements h in a set H. In the parametric case this set could be taken as the integers from 1 to the (finite) number of elements in η. In the case where η is infinite dimensional, the indexing set H will typically also be infinite dimensional.

where $\overline{\text{lin}}\,\mathscr{H}_\gamma$ denotes the closed linear span of the set $\mathscr{H}_\gamma$.[18] This operation removes from $\dot{\ell}_\gamma$ that part which can explained by score functions in $\mathscr{H}_\gamma$. The corresponding variance matrix, the *efficient information matrix* is

$$\tilde{\mathcal{I}}_\gamma = \int \tilde{\ell}_\gamma \tilde{\ell}_\gamma' \, \mathrm{d}P_\gamma.$$

Analytical derivation of the efficient score function for specific models can be complex, however due to the central role of the efficient score function in the literature on semi parametrically efficient estimation the efficient score function has already been derived for a large number of popular models.[19]

As a direct consequence of the definition in (2), $\int \tilde{\ell}_\gamma \, \mathrm{d}P_\gamma = 0$ and hence the efficient score function provides a d_θ-dimensional vector of moment condition on which one can base inference about θ. In general, constructing estimators and tests based on the efficient score function is attractive as these have well established optimality properties (e.g. Bickel et al., 1998; van der Vaart, 2002; Choi et al., 1996). In some of the examples considered in this paper, the conditions which are required to obtain such results may fail. For instance, if θ is unidentified, no consistent estimator of θ can exist, let alone asymptotically efficient estimators. Nevertheless, I will show that in such situations tests based on the efficient score function can be used to conduct valid inference provided some mild conditions are satisfied.

To introduce the test statistic, let $\hat{\ell}_{n,\theta}$ and $\hat{\mathcal{I}}_{n,\theta}$ denote estimates of $\tilde{\ell}_\gamma$ and $\tilde{\mathcal{I}}_\gamma$ respectively. The efficient score statistic (for a given θ) is given by

$$\hat{S}_{n,\theta} = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\ell}_{n,\theta}(W_i) \right)' \hat{\mathcal{I}}_{n,\theta}^{\dagger} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\ell}_{n,\theta}(W_i) \right),$$

where "$\dagger$" denotes the Moore-Penrose pseudo-inverse. Supposing that mild assumptions hold, I show that, under $H_0 : \theta = \theta_0$, $\hat{S}_{n,\theta_0}$ converges in distribution to a χ^2_r random variable where $r = \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$. Importantly (i) this convergence holds under any local sequence of nuisance parameters and (ii) the assumptions imposed do not require θ to be identified, allow η to be on the boundary of the parameter space and allow for the estimates to depend on regularised estimators of η. Based on this convergence, the efficient score test is performed by comparison of $\hat{S}_{n,\theta_0}$ to the appropriate quantile of the $\chi^2_{r_n}$ distribution where $r_n = \operatorname{rank}(\hat{\mathcal{I}}_{n,\theta_0})$ and confidence sets for θ can be constructed by inverting the test.

Intuitively there are two features of the efficient score statistic which are responsible for this result. The first is that the null value θ_0 is imposed in the construction of the statistic which precludes the need for θ to be identifiable or consistently estimable. This is key in

[18] The projection in the preceding display should be understood componentwise.

[19] Additionally guidance and a large number of examples can be found in Newey (1990), Bickel et al. (1998) and van der Vaart (1998, Chapter 25).

models with potential identification failures, where such requirements can fail. Second, the orthogonal projection in the definition of the efficient score function ensures that

$$\int \tilde{\ell}_\gamma \, B_\gamma h \, \mathrm{d}P_\gamma = 0 \quad \text{for all} \quad B_\gamma h \in \mathscr{H}_\gamma, \tag{3}$$

i.e. the efficient score function is uncorrelated with the scores $B_\gamma h$ for the nuisance parameters (in each direction h). Similar properties have been shown to alleviate size distortions in a number of settings, including those caused by identification issues (Kleibergen, 2005), boundary effects (Andrews, 2001) and regularised estimation of nuisance parameters (Chernozhukov et al., 2015, 2018). Property (3) has a fundamentally important role more generally in models with nuisance parameters in order to obtain the same limiting distribution regardless of the local sequence of nuisance parameters under which the limit is taken (cf. Hall and Mathiason, 1990; Choi et al., 1996).[20]

In addition to the robustness properties that (3) gives the efficient score test, (3) is also important for its power optimality properties – reflecting the original development of the C(α) test by Neyman (1959). If the efficient information matrix has full rank – as is usually the case in well identified models – and local perturbations to the nuisance parameters are indexed by a linear space, the efficient score test belongs to the class of asymptotically uniformly most powerful invariant tests (AUMPI) as described and demonstrated in Choi et al. (1996). Moreover, if the efficient information matrix has positive rank, there are directions against which non-trivial local power can be attained. I demonstrate that the efficient score test is minimax optimal in this scenario, in that there is no alternative test which provides higher power in a minimax sense.

To illustrate the broad applicability of these results, I now present two different examples to show (i) how commonly used econometric models can be placed into the framework required by (1) and (ii) how certain (local) sequences of nuisance parameters η can cause problems for commonly used inferential procedures. Following this I briefly discuss a number of other important examples in econometrics for which the inferential approach in this paper could be useful.

Example 1 (Single-index model)**.** Consider the single-index regression model (e.g Ichimura, 1993; Horowitz, 2009)

$$Y = f(X_1 + X_2\theta) + \epsilon, \quad \mathbb{E}(\epsilon|X) = 0,$$

where $f : \mathbb{R} \to \mathbb{R}$ belongs to some function class $\mathscr{F}$, X_1 and X_2 are continuously distributed random variables and ϵ is an unobserved error term. $(\epsilon, X) \sim \zeta$ for some Lebesgue density function ζ which ensures that the conditional mean restriction indicated above is satisfied. Such single-index models are popular as they relax the commonly imposed linear structure

[20]See also the discussions comparing Rao's score test and Neyman's C(α) test on page 133 of Andrews and Mikusheva (2015) and page 492 of Kocherlakota and Kocherlakota (1991).

of linear regression models but avoid the curse of dimensionality by ensuring the argument of f is a scalar. The density of an observation $W = (Y, X) \in \mathbb{R}^3$ is

$$p_\gamma(W) = \zeta(Y - f(X_1 + X_2\theta), X),$$

and the corresponding model is given by $\mathcal{P} = \{P_\gamma : \gamma \in \Theta \times \mathcal{H}\}$ for some open $\Theta \subset \mathbb{R}$ and $\mathcal{H} = (f, \zeta) \in \mathscr{F} \times \mathscr{Z}$, where the latter set restricts the possible distribution of (ϵ, X).

As discussed in Horowitz (2009), θ is unidentified when f is a constant function. Weak identification can therefore occur when f is sufficiently close to constancy (relative to the sample size). The potential identification failure here is due to an *infinite dimensional* nuisance parameter and therefore robust approaches to inference designed for cases where identification failure is caused by a finite dimensional nuisance parameter do not apply. Derivations of the efficient score function for the model above (and various extensions) can been found in the literature, see e.g. Newey and Stoker (1993); Ma and Zhu (2013); Kuchibhotla and Patra (2020). The efficient score test permits inference on θ to be performed which is robust to potential identification failure; full details are given in section 4. $\triangle$

Example 2 (Simple linear simultaneous equations model)**.** Suppose that the $K \times 1$ vector W satisfies

$$W = A(\theta)^{-1}\epsilon,$$

where $A(\theta)$ is a rotation matrix parametrised by $\theta \in \Theta$ and ϵ a $K \times 1$ vector of independent structural shocks each with mean zero and unit variance. Let $\eta = (\eta_1, \ldots, \eta_K) \in \mathcal{H}$ denote the densities of the components of ϵ. This yields the model

$$\mathcal{P} = \{P_\gamma : \gamma = (\theta, \eta) \in \Gamma = \Theta \times \mathcal{H}\},$$

where P_γ has Lebesgue density $p_\gamma(W) = \prod_{k=1}^K \eta_k (A_k(\theta)W)$.[21]

If all ϵ_k are Gaussian, $A(\theta)$ is not identified and hence the same is true of θ. In contrast, if (at least) $K-1$ of the components of ϵ have non-Gaussian distributions, $A(\theta)$ is identified up to sign changes and column permutations (Comon, 1994). Appropriate restrictions on the signs and labelling of the elements then result in identification of θ. However, if the non-Gaussian distributions of the ϵ_k are sufficiently close to Gaussian, θ is only weakly identified and inference methods which assume non-Gaussianity can suffer from size distortions.

The efficient score test avoids these size distortions by fixing $\theta = \theta_0$ under the null and orthogonalising with respect to (the scores for) η. In section 5, I show that the conclusions of these heuristic arguments hold formally in a considerably richer class of LSEMs. I also show that inference based on the efficient score test is minimax optimal in these models, including in cases where θ is underidentified.

[21] $A_k(\theta)$ is the k-th row of $A(\theta)$.

The identification problem in this example is caused by an *infinite dimensional* nuisance parameter and therefore robust approaches to inference designed for cases where identification failure is caused by a finite dimensional nuisance parameter do not apply. $\triangle$

Other examples

In addition to the preceding examples, robust inference on a large variety of other models of interest in econometrics can be conducted using the approach in this paper, pending verification of the high-level conditions in the next section. I briefly discuss four such cases here.

Firstly, consider inference on the slope parameters θ associated with the endogenous variables in an instrumental variables regression model. As is well known, many standard tests are unreliable in instrumental variable regression models if the instruments are weak (Andrews, Stock, and Sun, 2019). In contrast, the efficient score test could be used to provide valid inference in this model. In this model – unlike examples 1 or 2 – the lack of identification is caused by a finite dimensional parameter. Nevertheless, due to potential heteroskedasticity, the efficient score in this model depends on an infinite dimensional object, the heteroskedastic function. The resulting test does not coincide with any of the "standard" weak-IV robust tests, such as the AR, LM and CLR statistics (e.g. Anderson and Rubin, 1949; Staiger and Stock, 1997; Moreira, 2003; Kleibergen, 2002, 2007).

Secondly, consider the classical linear errors-in-variables model (as in, for example, equation (1.1) of Bickel and Ritov, 1987 or equation (1) of Ben-Moshe, 2020). As discussed by numerous authors (e.g. Reiersøl, 1950; Willassen, 1979; Bickel and Ritov, 1987; Ben-Moshe, 2020), identification of the regression coefficients may depend on (joint) distributional properties of the covariates, structural errors and measurement errors. These can include, for example, independence restrictions and non-Gaussianity assumptions on the latent covariates (Reiersøl, 1950; Willassen, 1979). Similarly to example 2, on verification of the high-level conditions in the next section, the inferential framework in this paper could be used to perform inference which will remain valid if, for instance, the distribution of the latent covariates is sufficiently close to Gaussianity that the regression coefficients become weakly identified. As in examples 1 and 2, this is a case of non-regularity caused by an infinite dimensional parameter.

As a third example, consider the mixed proportional hazard model, a common model used in duration analysis which allows for unobserved heterogeneity (see van den Berg, 2001, for a review). As was demonstrated by Hahn (1994), in the case where the baseline hazard function is Weibull, the efficient information matrix (for the Euclidean parameters) is singular, and no regular estimator sequence for these parameters can exist.[22] Pending verification of the high-level conditions in the next section, the inferential framework outlined

[22] Hahn (1994) also derives the efficient score function for this model.

in this paper could be used to perform inference which will remain valid if the baseline hazard function is (close to) Weibull. As in examples 1 and 2, this is a case of non-regularity caused by an infinite dimensional parameter.

Finally, as is well known, models with nuisance parameters on or close to the boundary can cause standard testing approaches to be unreliable (Andrews, 2001; Elliott et al., 2015; Ketz, 2018). Similar problems may arise in models where nuisance functions are estimated with shape restrictions imposed (cf. Chetverikov, Santos, and Shaikh, 2018, section 3). Due to the orthogonality between the scores for the parameter of interest and the nuisance scores, these restrictions do not affect the limiting distribution of the efficient score statistic and hence inferential approach in this paper will remain valid in these models – pending the verification of the high-level conditions in the next section. Depending on the model and the restriction under consideration, this case of non-regularity may be caused by either a finite-dimensional parameter or an infinite-dimensional parameter.

The next section describes the high level theory and provides a set of mild assumptions under which the efficient score test provides robust inference and has power optimality properties. Thereafter I revisit and generalise examples 1, 2 and work out the details for implementation.

3 Theory

In this section I formalise inference based on the efficient score statistic. First I set out the high-level assumptions which will be required throughout and formally define the efficient score test and associated confidence sets. Second, I perform an asymptotic analysis of the size properties of this test and the coverage of the associated confidence sets. Third, I demonstrate that this test has power optimality properties in a number of scenarios. Finally I provide a number of conditions and constructions which are sufficient for the high-level assumptions and often simpler to verify. In what follows I will often use operator notation for integrals, e.g. for a function f and a probability measure P, $Pf := \int f \,\mathrm{d}P$. $\mathbb{P}_n$ denotes the empirical measure of the sample $(W_i)_{i=1}^n$, so $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(W_i)$.[23]

3.1 Model setup and maintained assumptions

The first assumption that I impose merely formalises the model of interest as discussed in section 2 and stipulates that the observed data form a random sample.

Assumption M (Model and sampling)**.** Let $(W_i)_{i=1}^n$ be independent copies of a $\mathcal{W}$-valued random element W, with $\mathcal{W}$ a Polish space, all defined on an underlying probability space

[23] See appendix section A for additional details and notational conventions.

$(\Omega, \mathcal{F}, \mathrm{P})$.[24] The considered model for the law of W on $(\mathcal{W}, \mathcal{B}(\mathcal{W}))$ is

$$\mathcal{P} := \{P_\gamma : \gamma \in \Gamma\},$$

where Γ has the product form $\Gamma = \Theta \times \mathcal{H}$ for Θ an open subset of $\mathbb{R}^{d_\theta}$ and $\mathcal{H}$ a metric space. A typical value $\gamma \in \Gamma$ will be written as $\gamma = (\theta, \eta)$ where $\theta \in \Theta$ and $\eta \in \mathcal{H}$. Each $P_\gamma \in \mathcal{P}$ is dominated by a common σ-finite measure ν. $\diamond$

The next assumption is the key requirement. It imposes that the model satisfies a LAN condition (e.g. van der Vaart, 1998, Chapter 7; Le Cam and Yang, 2000, Chapter 6), where the parameter $\gamma = \gamma_n$ can change with the sample size n. In order to state this assumption, some notation is required. For any $P_\gamma \in \mathcal{P}$ I write p_γ for its density with respect to ν and for any two points $\gamma_1, \gamma_2 \in \Gamma$, $\Lambda_n(\gamma_1, \gamma_2)$ denotes the log-likelihood ratio:

$$\Lambda_n(\gamma_1, \gamma_2) := \log \prod_{i=1}^{n} \frac{p_{\gamma_1}}{p_{\gamma_2}}. \tag{4}$$

The LAN requirement is imposed as follows.

Assumption LAN (Local asymptotic normality)**.** Let $(\gamma_n)_{n\in\mathbb{N}}$ be a sequence in Γ which converges to a point $\gamma \in \Gamma$ and H_η a subset of a Banach space, H, which includes 0.

For any sequence $\tau_n \to \tau$ with each $\tau_n, \tau \in \mathbb{R}^{d_\theta}$, any sequence $h_n \to h$ with $h_n, h \in H_\eta$, a convergent sequence of $d_\theta \times d_\theta$ matrices δ_n and sequences $\eta_n(h_n) \to \eta$ with each $\eta_n(h_n) \in \mathcal{H}$, define

$$\gamma_n(\tau_n, h_n) := (\theta_n + \delta_n \tau_n, \eta_n(h_n)),$$

and suppose that

(i) the sequence $(P_{\gamma_n(\tau_n, h_n)})_{n\geq 1}$ is (eventually) in $\mathcal{P}$,

(ii) the associated log-likelihood ratio satisfies

$$\Lambda_n(\gamma_n(\tau_n, h_n), \gamma_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\tau' \dot{\ell}_{\gamma_n} + B_{\gamma_n} h\right] - \frac{1}{2} P_{\gamma_n} \left[\tau' \dot{\ell}_{\gamma_n} + B_{\gamma_n} h\right]^2 + o_{P_{\gamma_n}}(1), \tag{5}$$

for a sequence of functions $(\dot{\ell}_{\gamma_n})_{n\in\mathbb{N}}$ with each $\dot{\ell}_{\gamma_n} \in L_2^0(P_{\gamma_n})$ and a sequence of linear maps $(B_{\gamma_n})_{n\in\mathbb{N}}$ with each $B_{\gamma_n} : H_\eta \to L_2^0(P_{\gamma_n})$ such that $\tau' \dot{\ell}_{\gamma_n} + B_{\gamma_n} h$ is uniformly square P_{γ_n}-integrable.

$\diamond$

In what follows I use the notation $P_{\gamma_n, \tau_n, h_n}$ for $P_{\gamma_n(\tau_n, h_n)}$. The functions $\tau' \dot{\ell}_{\gamma_n} + B_{\gamma_n} h$ will (collectively) be called "score functions", as will the vector $\dot{\ell}_{\gamma_n}$ (the "score functions for

[24] A Polish space is a separable completely metrisable topological space. Let d be a metric such that $(\mathcal{W}, d)$ is a complete (separable) metric space. $\mathcal{B}(\mathcal{W})$ is the Borel σ-algebra on $(\mathcal{W}, d)$.

θ") and the functions $B_{\gamma_n}h$ (the "score functions for η"). Such functions play the same role as score functions in classical parametric models in which – under regularity conditions – a similar LAN condition holds (e.g. van der Vaart, 1998, Theorem 7.2).

Assumption LAN stipulates that the likelihood ratios $\Lambda_n(\gamma_n(\tau_n, h_n), \gamma_n)$ admit a local quadratic approximation with a particular form. It is important to clarify the roles of the different sequences of parameters present in these likelihood ratios. I refer to $(\gamma_n)_{n\in\mathbb{N}}$ as the "base sequence" and the components $\delta_n\tau_n$ and $\eta_n(h_n) - \eta_n$ as "local perturbations" to the elements of this base sequence respectively:

$$\begin{aligned}\gamma_n(\tau_n, h_n) &= \gamma_n + (\delta_n\tau_n,\ \eta_n(h_n) - \eta_n) \\ &= \Big(\ \theta_n + \underbrace{\delta_n\tau_n}_{\text{local perturbation of }\theta_n},\quad \eta_n + \underbrace{\eta_n(h_n) - \eta_n}_{\text{local perturbation of }\eta_n}\ \Big).\end{aligned}$$

That γ_n is permitted to vary with n has two important implications. Firstly, replacing a fixed θ with a convergent sequence $\theta_n \to \theta$ permits the demonstration that confidence sets constructed by inverting the efficient score test are *uniformly* valid over compact subsets of Θ. Secondly, this permits local power analysis in situations where the rate of information accumulation is non-standard.[25]

The separation of the local perturbation of θ_n into a "rate" term δ_n and a "direction" term τ_n is not strictly necessary but clarifies the role each plays in the subsequent power results. Due to the (possible) infinite dimensionality of the nuisance parameters η_n, the form of the local perturbation may be complex and generally will be model dependent, but the role of h_n is analogous to that of τ_n, i.e. it is the "direction" term in the perturbation.

Assumption LAN requires that for any permitted sequence of local perturbations, the measures P_{γ_n,τ_n,h_n} eventually belong to the model and (5) holds. That these hold over all such local sequences is key for the size results below which demonstrate that the efficient score test controls size *locally uniformly*, i.e. over any compact set of local perturbation directions consistent with the null. I emphasise that in the size and power results below LAN is only assumed to hold along certain specified base sequences $(\gamma_n)_{n\in\mathbb{N}}$ which are defined in the relevant results.

It is also important to note that assumption LAN concerns only the model $\mathcal{P}$ and perturbation spaces H_η, both of which are chosen by the researcher. This includes the choice of the metric on $\mathcal{H}$, which – particularly in the infinite dimensional case – has implications for

[25]For instance, one key feature of weak or semi-strong identification (in the terminology of Andrews and Cheng, 2012) is that the information that can be learned about the parameter of interest accrues at a rate slower than the "usual" $\sqrt{n}$; robust tests can then often be built on top of "rescaling" arguments: some part of γ_n changes with the sample size, causing a slower rate of information acquisition, which can be compensated for by a "slower" rate sequence δ_n — i.e. the local alternatives are "closer" than in the "usual" $\sqrt{n}$ case (Cf. Antoine and Renault, 2009, 2011; Andrews and Mikusheva, 2015). The prototypical "weak identification" case is usually the limiting case of this argument, where $\delta_n \not\to 0$ and the "local" alternatives are, in a sense, "fixed" alternatives.

the uniformity results obtained below, which hold over compact sets. Specifically, choosing a stronger metric on $\mathcal{H}$ will often simplify the demonstration that assumption LAN holds, but leads to "fewer" compact sets and therefore weaker uniformity results.[26]

Finally, rather than establishing LAN directly, one may establish that the relevant submodels are differentiable in quadratic mean (see assumption DQM below), which then implies assumption LAN (under assumption M; see proposition 3.10). A detailed analysis of the relationship between conditions of these types is given by Le Cam (1986, Chapter 17, section 3); see also Strasser (1985, Theorem 75.9).

I now introduce the next assumption, which concerns the limits of the scores.

Assumption CM(i) (Convergence of moments (i))**.** In the setting of assumption LAN suppose that there exists a vector of functions $\dot{\ell}_\gamma \in L_2^0(P_\gamma)$ and a bounded linear map $B_\gamma : H_\eta \to L_2^0(P_\gamma)$ such that for each $(\tau, h) \in \mathbb{R}^{d_\theta} \times H_\eta$

$$\lim_{n\to\infty} P_{\gamma_n} \left[\tau' \dot{\ell}_{\gamma_n} + B_{\gamma_n} h\right]^2 = P_\gamma \left[\tau' \dot{\ell}_\gamma + B_\gamma h\right]^2 .$$

$\diamond$

The uniform integrability required by assumption LAN may directly imply that assumption CM(i) holds; see subsection 3.4 for some sufficient conditions.

With the quantities introduced in the preceding assumptions, the efficient score function can be formally defined. First define the *tangent sets* for η as

$$\mathscr{H}_\gamma := \{B_\gamma h : h \in H_\eta\}, \qquad \text{for} \qquad \gamma \in \{\gamma\} \cup \{\gamma_n : n \in \mathbb{N}\}.$$

The efficient score functions are defined as the orthogonal projections of the score functions for θ, i.e. the $\dot{\ell}_{\gamma_n}$ and $\dot{\ell}_\gamma$ onto the orthocomplement of $\mathscr{H}_{\gamma_n}$ and $\mathscr{H}_\gamma$ respectively. The corresponding efficient information matrices are the expectations of the outer products of these (vectors of) functions:

$$\tilde{\ell}_\gamma := \dot{\ell}_\gamma - \Pi_\gamma \left(\dot{\ell}_\gamma \mid \overline{\text{lin}}\, \mathscr{H}_\gamma\right), \quad \tilde{\mathcal{I}}_\gamma := P_\gamma \left[\tilde{\ell}_\gamma \tilde{\ell}_\gamma'\right], \qquad \text{for} \qquad \gamma \in \{\gamma\} \cup \{\gamma_n : n \in \mathbb{N}\},$$

where $\Pi_\gamma(\cdot|\mathcal{S})$ is the orthogonal projection on $\mathcal{S} \subset L_2(P_\gamma)$.

I assume the same uniform integrability moment convergence conditions on the efficient scores that have been imposed on the scores for θ and η.

Assumption CM(ii) (Convergence of moments (ii))**.** Suppose that assumption CM(i) holds and moreover that $\|\tilde{\ell}_{\gamma_n}\|_2^2$ is uniformly P_{γ_n}-integrable and $\lim_{n\to\infty} \tilde{\mathcal{I}}_{\gamma_n} = \tilde{\mathcal{I}}_\gamma$. $\diamond$

[26] More formally, if d_1 and d_2 are metrics on $\mathcal{H}$ with d_1 stronger than d_2 (i.e. every open subset of $\mathcal{H}$ with respect to d_2 is also open with respect to d_1), then if a set $H' \subset \mathcal{H}$ is compact with respect to d_1, then it is compact with respect to d_2.

The definition of the efficient score function ensures that $P_\gamma \tilde{\ell}_\gamma = 0$, since both $\dot{\ell}_\gamma$ and the elements of $\overline{\text{lin}}\ \mathscr{H}_\gamma$ are mean zero by assumption LAN. In other words, the efficient score function provides d_θ moment conditions on which inference about θ can be based.

In many cases, the efficient score function will not be formed only of observed or known quantities, but will need to be estimated. The following two conditions impose what is required of these estimates and complete the collection of high-level assumptions.

Assumption E (Estimation)**.** Let $(\gamma_n)_{n\in\mathbb{N}}$ be as in assumption LAN and suppose that for an estimator $\hat{\ell}_{n,\theta_n}$

$$\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\theta_n} - \tilde{\ell}_{\gamma_n}\right] = o_{P_{\gamma_n}}(1), \tag{6}$$

and for an estimator $\hat{\mathcal{I}}_{n,\theta_n}$

$$\left\|\hat{\mathcal{I}}_{n,\theta_n} - \tilde{\mathcal{I}}_\gamma\right\|_2 = o_{P_{\gamma_n}}(1). \tag{7}$$

◇

Assumption R (Rank convergence)**.** Let $(\gamma_n)_{n\in\mathbb{N}}$ be as in assumption LAN and suppose that the estimator $\hat{\mathcal{I}}_{n,\theta_n}$ of assumption E satisfies

$$P_{\gamma_n}\left(\text{rank}(\hat{\mathcal{I}}_{n,\theta_n}) = \text{rank}(\tilde{\mathcal{I}}_\gamma)\right) \to 1. \tag{8}$$

◇

That the first condition of assumption E, equation (6), can hold is often related to the specific structure of the efficient score function, particularly the fact that it is orthogonalised with respect to the nuisance scores. The second condition (7) requires consistency of an estimator of the efficient information matrix $\tilde{\mathcal{I}}_\gamma$. If the latter is non-singular and (7) holds, then (8) holds automatically.[27] If $\tilde{\mathcal{I}}_\gamma$ is rank deficient, (8) must be established separately. A construction which can ensure this holds, given an initial estimator with known convergence rate is given in subsection 3.4.

The fact that assumption R is required is due to the fact that the Moore-Penrose pseudo-inverse (which I denote by $M^\dagger$ for an arbitrary matrix M) is not continuous. However, if $E_n \to 0$ such that $M + E_n$ has the same rank as M, then $(M + E_n)^\dagger \to M^\dagger$.[28]

Verification of equations (6) and (7) is model specific and typically requires the application of various stochastic limit theorems. Incorporating estimates of Euclidean parts of the nuisance parameter can typically be achieved relatively simply via discretisation arguments if a $\sqrt{n}$-consistent estimator is available; see the example in section 5 below. For nonparametric parts, sample splitting can often be used to provide estimators for which the verification of the required conditions is relatively straightforward.

[27] See Lemma C.7.

[28] See e.g. Ben-Israel and Greville (2003, Section 6.6) and Cf. Andrews (1987).

3.2 The efficient score test

In this section, I define the efficient score test, which forms the basis of the inferential approach suggested in this paper. Two different definitions are required: one for a (scalar) one-sided hypothesis and one for a two-sided hypothesis.

For the purposes of testing a two-sided hypothesis at level $\alpha \in (0, 1)$, the efficient score statistic at θ is defined as

$$\hat{S}_{n,\theta} := \left(\sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta}\right)' \hat{\mathcal{I}}^{\dagger}_{n,\theta} \left(\sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta}\right). \tag{9}$$

The efficient score test can then be defined as

$$\phi_{n,\theta} := \mathbf{1}\left\{\hat{S}_{n,\theta} > c_n\right\}, \tag{10}$$

where c_n is the $1 - \alpha$ quantile of the $\chi^2_{r_n}$ distribution, with $r_n := \operatorname{rank}(\hat{\mathcal{I}}_{n,\theta})$. The confidence set corresponding to the efficient score test is denoted by $\hat{C}_n$ and defined as

$$\hat{C}_n := \{\theta \in \Theta : \phi_{n,\theta} = 0\} = \left\{\theta \in \Theta : \hat{S}_{n,\theta} \leq c_n\right\}. \tag{11}$$

For the purposes of testing a one-sided hypothesis for a scalar parameter, i.e. when $d_\theta = 1$ and $\alpha \in (0, 1/2]$, I instead define the efficient score statistic at θ as

$$\hat{S}_{n,\theta} := \left(\sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta}\right) \sqrt{\hat{\mathcal{I}}^{\dagger}_{n,\theta}}, \tag{12}$$

and define the corresponding test as

$$\phi_{n,\theta} := \mathbf{1}\left\{\hat{S}_{n,\theta} > z_\alpha\right\}, \tag{13}$$

where z_α is the $1 - \alpha$ quantile of the $\mathcal{N}(0, 1)$ distribution. A confidence set can again be constructed by test inversion as

$$\hat{C}_n := \{\theta \in \Theta : \phi_{n,\theta} = 0\} = \left\{\theta \in \Theta : \hat{S}_{n,\theta} \leq z_\alpha\right\}. \tag{14}$$

The use of the same notation for these different objects should not cause any confusion as only one of the two is applicable to any given testing problem and hence which is meant will be clear from context.

3.3 Asymptotic properties

I now derive the asymptotic properties of the efficient score test and test inversion confidence sets. I first state a weak convergence result along local alternatives, which follows directly from standard stochastic limit theorems and Le Cam's third lemma. Following this size results are given in section 3.3.1 and power results in section 3.3.2.[29]

Proposition 3.1. *Suppose that assumptions M, LAN and CM(i) hold. Then, the sequences of product measures* $\left(P^n_{\gamma_n}\right)_{n\in\mathbb{N}}$ *and* $\left(P^n_{\gamma_n,\tau_n,h_n}\right)_{n\in\mathbb{N}}$ *are mutually contiguous. If also assumption CM(ii) holds, then under* P_{γ_n,τ_n,h_n}

$$\sqrt{n}\mathbb{P}_n\tilde{\ell}_{\gamma_n} \rightsquigarrow \mathcal{N}(\tilde{\mathcal{I}}_\gamma\tau, \tilde{\mathcal{I}}_\gamma).$$

If, additionally, (6) *of assumption E holds, then also under* P_{γ_n,τ_n,h_n}

$$\sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta_n} \rightsquigarrow \mathcal{N}(\tilde{\mathcal{I}}_\gamma\tau, \tilde{\mathcal{I}}_\gamma).$$

The key takeaway from the preceding proposition is that the limiting distributions depend on τ but not on h (or $(h_n)_{n\in\mathbb{N}}$): by its construction the efficient score function has an invariance property with regard to the local nuisance perturbations.

3.3.1 Size results

The invariance property discussed in the preceding paragraph is precisely what ensures that the size of the efficient score test does not depend on the particular local nuisance perturbation along which the limit is taken.[30]

Proposition 3.2. *Suppose that assumptions M, LAN, CM(ii), E and R hold for a sequence* $(\gamma_n)_{n\in\mathbb{N}} \subset \Gamma$ *with limit* $\gamma \in \Gamma$ *and where* $\theta_n = \theta_0$ *for all* $n \in \mathbb{N}$*. Then, for any compact subset* H'_η *of* H_η*,*

$$\lim_{n\to\infty} \sup_{h\in H'_\eta} P^n_{\gamma_n,0,h}\phi_{n,\theta_0} \leq \alpha.$$

The preceding proposition demonstrates that the efficient score test is correctly sized uniformly over local perturbations consistent with the null. Note that this result specifies that the high-level conditions need hold only along the specified base sequence with $\gamma_n = (\theta_0, \eta_n) \to (\theta_0, \eta) = \gamma$. This result immediately implies that the efficient score test is correctly sized along any sequence of local perturbations of $\gamma_n = (\theta_0, \eta_n)$ with $\tau_n = 0$ and $h_n \to h$ in H_η.[31]

[29] Readers primarily interested in the robustness results may safely skip section 3.3.2.

[30] In fact this property can be shown to hold rather more generally, for $\breve{\ell}_{\gamma_n}$ in place of $\tilde{\ell}_{\gamma_n}$ as long as $P_{\gamma_n}[\breve{\ell}_{\gamma_n} B_{\gamma_n} h] = 0$ for all $h \in H_\eta$. If $\breve{\ell}_{\gamma_n} \neq \tilde{\ell}_{\gamma_n}$ this would typically result in a less powerful test and hence I do not explicitly consider this case in the theoretical results. Nevertheless this observation can be particularly useful in cases when the efficient score function is hard to estimate. See e.g. the treatment of heteroskedasticity in section 4 below.

[31] In a metric space the union of a convergent sequence and its limit is compact.

An analogous result holds for confidence sets constructed by test inversion, provided the high level conditions hold along sequences of the form $\gamma_n = (\theta_n, \eta_n) \to (\theta, \eta) = \gamma$, for any convergent sequence $\theta_n \to \theta$ (in a compact subset of Θ) and a specified $\eta_n \to \eta$.

Proposition 3.3. *Let Θ' be a compact subset of Θ. Fix a convergent sequence $(\eta_n)_{n\in\mathbb{N}}$ and denote its limit by η. Suppose that assumptions M, LAN, CM(ii), E and R hold for any sequence $(\gamma_n)_{n\in\mathbb{N}}$ where each $\gamma_n := (\theta_n, \eta_n)_{n\in\mathbb{N}} \subset \Theta' \times \mathcal{H}$ with $\theta_n \to \theta \in \Theta'$. Then, for any compact subset H'_η of H_η,*

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta'} \inf_{h\in H'_\eta} P^n_{(\theta,\eta_n),0,h}(\theta \in \hat{C}_n) \geq 1-\alpha.$$

3.3.2 Power results

In the scalar case I consider both one-sided tests of the form $H_0 : \theta > \theta_0$ against $H_1 : \theta \leq \theta_0$ and two-sided tests, i.e. $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. These results are essentially standard (Cf. Choi et al., 1996), with the key difference being that here they are stated with γ_n potentially changing with n. Whilst this is a potentially useful strengthening, it simply reflects the corresponding change in the assumptions – i.e. assumption LAN is assumed to hold along such sequences – with the arguments following in the usual way.[32] The first result concerns the power of one-sided tests.

Proposition 3.4. *Suppose that assumptions M, LAN, and CM(i) hold. Additionally suppose that H_η is a linear subspace of H and $\tilde{\mathcal{I}}_\gamma > 0$. Then, for any $\alpha \in (0,1)$, any sequence of asymptotically level-α tests $(\psi_n)_{n\in\mathbb{N}}$ for $H_0 : \tau \leq 0$ against $H_1 : \tau > 0$, i.e. any sequence of tests $\psi_n : \mathcal{W}^n \to [0,1]$ such that*

$$\limsup_{n\to\infty} P^n_{\gamma_n,\tau,h}\psi_n \leq \alpha \qquad \text{for all } \tau \leq 0,\ h \in H_\eta$$

is subject to the power bound

$$\limsup_{n\to\infty} P^n_{\gamma_n,\tau_n,h_n}\psi_n \leq 1 - \Phi\left(z_\alpha - \tilde{\mathcal{I}}_\gamma^{1/2}\tau\right), \tag{15}$$

for all $\tau_n \to \tau > 0$ and $h_n \to h \in H_\eta$ where z_α is the $1-\alpha$ quantile of the standard normal distribution and Φ is the standard normal CDF.

Any sequence of tests $\psi_n : \mathcal{W}^n \to [0,1]$ of asymptotic level α which attains the power bound (15) is called "asymptotically locally uniformly most powerful of level-α". The efficient score test attains this bound under the assumptions of section 3.1, provided that H_η is a linear subspace and $\tilde{\mathcal{I}}_\gamma > 0$.

[32] In particular the proofs are based on convergence of a particular sequence of experiments to a Gaussian shift limit experiment. The construction of the relevant sequence of experiments is given in section B.

Corollary 3.5. *Suppose that assumptions M, LAN, CM(ii), E hold, with $\gamma_n = (\theta_0, \eta_n) \to (\theta_0, \eta) = \gamma$. Additionally suppose that H_η is a linear subspace of H, $\tilde{\mathcal{I}}_\gamma > 0$ and $\alpha \in (0,1)$. Then the sequence of tests $(\phi_{n,\theta_0})_{n\in\mathbb{N}}$ is asymptotically locally uniformly most powerful of level-α for the hypothesis $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$, i.e. it is asymptotically level–α and achieves the power bound in (15) for any $\tau_n \to \tau > 0$ and any $h_n \to h \in H_\eta$.*

A similar result holds for two-sided tests, with the claim of optimality holding in the class of tests which are (asymptotically) unbiased and of level-α.

Proposition 3.6. *Suppose that assumptions M, LAN, CM(i) hold. Additionally suppose that H_η is a linear subspace of H and $\tilde{\mathcal{I}}_\gamma > 0$. Then, for any $\alpha \in (0,1)$, any sequence of asymptotically unbiased, level-α tests $(\psi_n)_{n\in\mathbb{N}}$ for $H_0 : \tau = 0$ against $H_1 : \tau \neq 0$, i.e. any sequence of tests $\psi_n : \mathcal{W}^n \to [0,1]$ such that*

$$\limsup_{n\to\infty} P^n_{\gamma_n,0,h}\psi_n \leq \alpha \qquad \text{for all } h \in \mathfrak{H}_\gamma,$$

and

$$\liminf_{n\to\infty} P^n_{\gamma_n,\tau,h}\psi_n \geq \alpha \qquad \text{for all } \tau \neq 0,\ h \in H_\eta$$

is subject to the power bound

$$\limsup_{n\to\infty} P^n_{\gamma_n,\tau_n,h_n}\psi_n \leq 1 - \Phi\left(z_{\alpha/2} - \tilde{\mathcal{I}}_\gamma^{1/2}\tau\right) + 1 - \Phi\left(z_{\alpha/2} + \tilde{\mathcal{I}}_\gamma^{1/2}\tau\right) \tag{16}$$

for all $\tau_n \to \tau \neq 0$ and $h_n \to h \in H_\eta$, where z_α is the $1-\alpha$ quantile of the standard normal distribution and Φ is the standard normal CDF.

Any asymptotically unbiased sequence of tests $\psi_n : \mathcal{W}^n \to [0,1]$ of asymptotic level α which attains the power bound (15) is called "asymptotically locally uniformly most powerful unbiased of level-α". The efficient score test attains this bound under the same assumptions as for the one-sided case.

Corollary 3.7. *Suppose that assumptions M, LAN, CM(ii) and E hold, with $\gamma_n = (\theta_0, \eta_n) \to (\theta_0, \eta) = \gamma$. Additionally suppose that H_η is a linear subspace of H, $\tilde{\mathcal{I}}_\gamma > 0$ and $\alpha \in (0,1)$. Then the sequence of tests $(\phi_{n,\theta_0})_{n\in\mathbb{N}}$ is asymptotically locally uniformly most powerful unbiased of level-α for the hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, i.e. it is asymptotically unbiased and of level-α and achieves the power bound in (16) for any $\tau_n \to \tau \neq 0$ and any $h_n \to h \in H_\eta$.*

For multivariate hypotheses I consider maximin optimality.[33] The difference between the power bound given here and what might be called the "usual" case (Cf. Theorem 13.5.4 of

[33] For an alternative approach which restricts the class of tests to those satisfying a rotation invariance condition see Choi et al. (1996).

Lehmann and Romano (2005) for the parametric case) is that I do not require the efficient information matrix to be positive definite. Rather I consider a restricted class of directions along which θ may be approached. Specifically, letting $N(\tilde{\mathcal{I}}_\gamma)$ denote the nullspace of $\tilde{\mathcal{I}}_\gamma$, the permitted directions are $\tau \in N(\tilde{\mathcal{I}}_\gamma)^\perp$ rather than $\tau \in \mathbb{R}^{d_\theta}$. Note that these coincide if (and only if) $\tilde{\mathcal{I}}_\gamma \succ 0$ and hence the "usual" case is a special case of this result. The generalisation given here is useful for models in which the parameter of interest may be underidentified.[34]

Proposition 3.8. *Suppose that assumptions M, LAN and CM(i) hold. Additionally suppose that H_η is a linear subspace of H and $r := \operatorname{rank}(\tilde{\mathcal{I}}_\gamma) > 0$. Then, for any $\alpha \in (0,1)$, any sequence of asymptotically level-α tests $(\psi_n)_{n\in\mathbb{N}}$ for $H_0 : \tau = 0$ against $H_1 : \tau \neq 0$, i.e. any sequence of tests $\psi_n : \mathcal{W}^n \to [0,1]$ such that*

$$\limsup_{n\to\infty} P^n_{\gamma_n,0,h}\psi_n \le \alpha \qquad \textit{for all } h \in H_\eta$$

is subject to the power bound

$$\limsup_{n\to\infty} \inf_{(\tau,h)\in M_a} P^n_{\gamma_n,\tau,h}\psi_n \le 1 - \mathrm{P}\left(\chi^2_r(a) \le c_{r,\alpha}\right), \tag{17}$$

for all $a > 0$, where $M_a := \{(\tau,h) \in N(\tilde{\mathcal{I}}_\gamma)^\perp \times H_\eta : \tau'\tilde{\mathcal{I}}_\gamma\tau \ge a\}$, $c_{r,\alpha}$ is the $1-\alpha$ quantile of the χ^2_r distribution and $\chi^2_r(a)$ denotes a non-central χ^2 random variable with r degrees of freedom and non-centrality a.

Any sequence of tests $\psi_n : \mathcal{W}^n \to [0,1]$ of asymptotic level α which attains the power bound (15) over all compact subsets of M_a is called "asymptotically maximin of level-α".[35] The efficient score test is asymptotically maximin of level-α under the assumptions in section 3.1, provided that H_η is a linear subspace and $\operatorname{rank}(\tilde{\mathcal{I}}_\gamma) > 0$.

Corollary 3.9. *Suppose that assumptions M, LAN, CM(ii), E and R hold, with $\gamma_n = (\theta_0, \eta_n) \to (\theta_0, \eta) = \gamma$. Additionally suppose that H_η is a linear subspace of H, $r := \operatorname{rank}(\tilde{\mathcal{I}}_\gamma) > 0$ and $\alpha \in (0,1)$. Then the sequence of tests $(\phi_{n,\theta_0})_{n\in\mathbb{N}}$ is asymptotically maximin of level-α for the hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ over all compacts, in the sense that for any compact $K_a \subset M_a$*

$$\lim_{n\to\infty} \inf_{(\tau,h)\in K_a} P^n_{\gamma_n,\tau,h}\phi_{n,\theta_0} = 1 - \mathrm{P}\left(\chi^2_r(a) \le c_{r,\alpha}\right). \tag{18}$$

There are two key takeaways from this result. Firstly, when the efficient information matrix is rank deficient, the efficient score test continues to enjoy non-trivial power in

[34]For details of the construction of the sequence of experiments used to establish this result see appendix section B.

[35]Cf. Section 13.5.3 of Lehmann and Romano (2005) for the terminology

certain directions.[36] Secondly the power it achieves is – in a certain sense – optimal.[37]

3.4 Sufficient conditions for the assumptions

In the i.i.d. setting it is well known that differentiability in quadratic mean (e.g. van der Vaart, 2002, Definition 1.6) is a sufficient condition for a LAN expansion like that in equation (5) with a fixed $\gamma \in \Gamma$ (e.g. Bickel et al., 1998; Le Cam and Yang, 2000; van der Vaart, 2002). In the setting of interest here, a suitably adapted version of this condition also suffices for assumption LAN.[38]

Assumption DQM (Differentiability in quadratic mean)**.** Let $(\gamma_n)_{n\in\mathbb{N}}$ be a sequence in Γ which converges to a point $\gamma \in \Gamma$ and H_η a subset of a Banach space, H, which includes 0.

For any sequence $\tau_n \to \tau$ with each $\tau_n, \tau \in \mathbb{R}^{d_\theta}$, any sequence $h_n \to h$ with $h_n, h \in H_\eta$, a convergent sequence of $d_\theta \times d_\theta$ matrices δ_n and sequences $\eta_n(h_n) \to \eta$ with each $\eta_n(h_n) \in \mathcal{H}$, define $\gamma_n(\tau_n, h_n)$ as in assumption LAN and suppose that

(i) the sequence $(P_{\gamma_n(\tau_n,h_n)})_{n\geq 1}$ is (eventually) in $\mathcal{P}$,

(ii) for some sequence of measurable functions $(g_n)_{n\in\mathbb{N}}$ such that $(g_n^2)_{n\in\mathbb{N}}$ are uniformly P_{γ_n}-integrable and $P_{\gamma_n} g_n = o(n^{-1/2})$,

$$\int \left[\sqrt{n}(\sqrt{p_{\gamma_n(\tau_n,h_n)}} - \sqrt{p_{\gamma_n}}) - \frac{1}{2} g_n \sqrt{p_{\gamma_n}}\right]^2 \mathrm{d}\nu \to 0. \tag{19}$$

◇

Proposition 3.10. *Suppose assumptions M and DQM hold. Moreover suppose that for a sequence of functions $(\dot{\ell}_{\gamma_n})_{n\in\mathbb{N}}$ with each $\dot{\ell}_{\gamma_n} \in L_2^0(P_{\gamma_n})$ and a sequence of linear maps $(B_{\gamma_n})_{n\in\mathbb{N}}$ with each $B_{\gamma_n} : H_\eta \to L_2^0(P_{\gamma_n})$,*

$$P_{\gamma_n}\left[\tau' \dot{\ell}_{\gamma_n} + B_{\gamma_n} h - g_n\right]^2 \to 0.$$

Then assumption LAN holds.

The addtional condition in the display in proposition 3.10 allows DQM to be shown with any sequence g_n such that the L_2 distance between g_n and the scores $\tau' \dot{\ell}_{\gamma_n} + B_{\gamma_n} h$ vanishes as $n \to \infty$.

[36] This is demonstrated in a specific example in section 5.5.

[37] Nevertheless, if one has a particular direction against which one wishes to direct power, or – more generally – a weighting function over alternatives, a criterion based on weighted average power would seem more appropriate. Cf. e.g. Elliott et al. (2015); Montiel Olea (2020).

[38] Results of this nature are known to hold see e.g. Strasser (1985, Chapter 74) or van der Vaart (1988, A.2). I provide this formulation to facilitate the demonstration of the version of LAN assumed in this paper.

I next record two conditions useful for checking the integral convergence required in CM(ii), once the uniform square P_{γ_n}-integrability has been established. The first can be obtained as an immediate corollary of a (stronger) result of Feinberg, Kasyanov, and Zgurovsky (2016), who establish a uniform (over Borel sets) version of the integral convergence. The second is effectively the standard result that weak convergence and uniform integrability imply convergence of moments, where the condition of continuous convergence is imposed to ensure the weak convergence of the appropriate laws.

Lemma 3.11. *Suppose that $(P_n)_{n\in\mathbb{N}}$ is a sequence of probability measures which converges in total variation to P.*[39] *If $(f_n)_{n\in\mathbb{N}}$ is a sequence of functions in $L_1(P_n)$ such that (a) $f_n \xrightarrow{P} f \in L_1(P)$ and (b) $(f_n)_{n\in\mathbb{N}}$ is uniformly P_n-integrable, then $P_nf_n \to Pf$.*

Lemma 3.12. *Let S be a metric space and suppose that $(P_n)_{n\in\mathbb{N}}$ is a sequence of measures on $(S, \mathcal{B}(S))$ which converge weakly to P. Suppose that $(f_n)_{n\in\mathbb{N}}$ is a sequence of real-valued functions with each $f_n \in L_1(P_n)$ which (a) converge continuously to $f \in L_1(P)$ and (b) are uniformly P_n-integrable.*[40] *Then $P_nf_n \to Pf$.*

Assumption R requires the estimate of the efficient information matrix, $\hat{\mathcal{I}}_{n,\theta_n}$, to have the same rank as $\tilde{\mathcal{I}}_\gamma$ with P_{γ_n}-probability approaching one. The following construction is sufficient to guarantee this; it requires knowledge of the rate of convergence to zero of the difference (in the spectral norm) of an estimator $\check{\mathcal{I}}_{n,\theta_n}$ and a matrix $\mathcal{I}_n$ where $\mathcal{I}_n \to \tilde{\mathcal{I}}_\gamma$ and $\operatorname{rank}(\mathcal{I}_n) = \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$ for all sufficiently large n. As there is nothing special about the limit being the efficient information matrix here, the construction is given more generally.[41]

In particular, suppose that the sequence of (random) positive semi-definite (symmetric) matrices $(\check{M}_n)_{n\in\mathbb{N}}$ (of fixed dimension $L \times L$) satisfy

$$P_n\left(\|\check{M}_n - M_n\|_2 < \nu_n\right) \to 1, \tag{20}$$

for a sequence $(P_n)_{n\in\mathbb{N}}$ of probability measures, a known non-negative sequence $\nu_n \to 0$ and a sequence of deterministic matrices $M_n \to M$ with $\operatorname{rank}(M_n) = \operatorname{rank}(M)$ for all sufficiently large n.[42] Let $\check{M}_n = \check{U}_n\check{\Lambda}_n\check{U}_n'$ be the corresponding eigendecompositions and define

$$\hat{M}_n \coloneqq \check{U}_n\Lambda_n(\nu_n)\check{U}_n' \ , \tag{21}$$

[39] Each P_n and P are defined on a common measurable space $(S, \mathcal{B}(S))$.

[40] Continuous convergence requires $f_n(s_n) \to f(s)$ for all $(s_n)_{n\in\mathbb{N}} \subset S$ with $s_n \to s \in S$. Here this is equivalent to compact convergence of the f_n to a continuous limit f (cf. Remmert, 1991, Chapter 3, §1, Section 5).

[41] A similar construction appears as part of Theorem 2 in Lee and Mesters (2021a). If the (non-zero) eigenvalues of $\tilde{\mathcal{I}}_\gamma$ can be computed, a simpler truncation approach can be utilised, cf. Proposition 2 in Lütkepohl and Burda (1997).

[42] (20) is implied by $\|\check{M}_n - M_n\| = o_{P_{\gamma_n}}(\nu_n)$ for any matrix norm. Moreover, the existence of such a sequence $(\nu_n)_{n\in\mathbb{N}}$ is guaranteed if $\|\check{M}_n - M_n\|_2 \to 0$ in P_n-probability, however its explicit knowledge is necessary to perform the subsequent construction.

where $\Lambda_n(\nu_n)$ is a diagonal matrix with the ν_n-truncated eigenvalues of $\check{M}_n$ on the main diagonal and $\check{U}_n$ is the matrix of corresponding orthonormal eigenvectors. That is, if $(\check{\lambda}_{n,i})_{i=1}^{L}$ denote the non-increasing eigenvalues of $\check{M}_n$, then the (i,i)-th element of $\Lambda_n(\nu_n)$ is $\check{\lambda}_{n,i}\mathbf{1}(\check{\lambda}_{n,i} \geq \nu_n)$.

Proposition 3.13. *If* (20) *holds,* $M_n \to M$ *and for all* n *greater than some* $N \in \mathbb{N}$ $\operatorname{rank}(M_n) = \operatorname{rank}(M)$*, then* $\hat{M}_n \xrightarrow{P_n} M$ *and*

$$P_n\left(\operatorname{rank}(\hat{M}_n) = \operatorname{rank}(M)\right) \to 1, \tag{22}$$

where $\hat{M}_n$ *is defined as in* (21)*.*

Assumption T. Let $(\gamma_n)_{n\in\mathbb{N}}$ be a sequence in Γ with a limit $\gamma \in \Gamma$, $(\tilde{\mathcal{I}}_n)_{n\in\mathbb{N}}$ a deterministic sequence of matrices with $\tilde{\mathcal{I}}_n \to \tilde{\mathcal{I}}_\gamma$ and $\operatorname{rank}(\tilde{\mathcal{I}}_n) = \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$ for all n exceeding some $N \in \mathbb{N}$ and suppose that the sequence $(\check{\mathcal{I}}_{n,\theta_n})_{n\in\mathbb{N}}$ satisfies

$$P_{\gamma_n}\left(\|\check{\mathcal{I}}_{n,\theta_n} - \tilde{\mathcal{I}}_n\|_2 < \nu_n\right) \to 1. \tag{23}$$

$\diamond$

Corollary 3.14. *If assumption T holds, the estimate* $\hat{\mathcal{I}}_{n,\theta_n}$ *formed by truncating the eigendecompositions of* $\check{\mathcal{I}}_{n,\theta_n}$ *at* ν_n*, as in* (21)*, satisfies equation* (7) *and assumption R.*

In practice equation (23) is likely to be established by demonstrating that $\|\check{\mathcal{I}}_{n,\theta_n} - \tilde{\mathcal{I}}_n\| = o_{P_{\gamma_n}}(\nu_n)$.[43] As this condition concerns only asymptotic behaviour, there is wide scope for different possible sequences which have the same asymptotic behaviour but rather different behaviour in finite samples. Simulation experiments designed to replicate various possible DGPs for the case under consideration may provide some guidance.

4 Single index model

In this section I provide details of the application of the theory of section 3 to a more general version of the single index model in example 1.

Consider the single index (regression) model (SIM), where $W = (Y, X)$ with

$$Y = f(X_1 + X_2'\theta) + \epsilon, \quad \mathbb{E}[\epsilon|X] = 0, \tag{24}$$

for $X = (X_1, X_2) \in \mathbb{R}^K$ a vector of covariates such that $(\epsilon, X) \sim \zeta$ for some Lebesgue density

[43] For any matrix norm $\|\cdot\|$.

ζ and some unknown link function f.[44] As recorded in Theorem 2.1 of Horowitz (2009), f and θ are identified in this model if f is differentiable, not constant on the support of $X_1+X_2'\theta$ and the support of X is not contained in a proper linear subspace of $\mathbb{R}^K$. By utilising the inferential approach developed in section 3, this section provides an inferential approach for θ in model (24) which is robust to failure of these assumptions, and – perhaps more importantly – robust in a setting where f is relatively flat when compared with sampling variation, leading to weak identification of θ.

The first step of the analysis is to formally specify the model under consideration and establish some primitive assumptions under which the results will be obtained. The basic model setup is given by the following assumption.

Assumption SIM. Suppose that $W = (Y, X) \in \mathbb{R}^{1+K}$ satisfies (24) and

(i) $\Theta \subset \mathbb{R}^{d_\theta}$ is open,
(ii) $(\epsilon, X) \sim \zeta$ where $\zeta \in \mathscr{Z}$,
(iii) $f \in \mathscr{F}$,

where $\mathscr{Z}$ and $\mathscr{F}$ are defined as follows. Let $\mathscr{X} \subset \mathbb{R}^K$ be closed, $\phi(\epsilon, X) := \frac{\partial \log \zeta(e,X)}{\partial e}(\epsilon, X)$ the log-density score in the first argument of ζ and $\rho > 0$. Then $\mathscr{Z}$ is the collection:

$$\mathscr{Z} := \left\{ \zeta \in L_1(\mathbb{R}^{1+K}) : \zeta \geq 0, \int_{\mathbb{R}\times\mathscr{X}} \zeta \, \mathrm{d}\lambda = 1, \text{ if } (e, Z) \sim \zeta \text{ then (26)}, \ \zeta \text{ satisfies (25)} \right\},$$

where $L_1(\mathbb{R}^{1+K})$ is the space of Lebesgue integrable functions on $\mathbb{R}^{1+K}$ and

$$e \mapsto \sqrt{\zeta(e, X)} \text{ is continuously differentiable } \lambda-\text{a.e.}, \tag{25}$$

$$\mathbb{E}[\epsilon|X] = 0, \quad \mathbb{E}[(\phi(\epsilon, X)^{2+\rho} + 1)\|X\|_2^{2+\rho}] < \infty, \quad \mathbb{E}[XX'] \succ 0. \tag{26}$$

$\mathscr{F} := C_b^1(\mathscr{D})$ is the class of functions which are bounded and continuously differentiable with bounded derivative λ-a.e. on $\mathscr{D} := \{X_1 + X_2'\theta : \theta \in \Theta, x \in \mathscr{X}\}$.

The model is given by $\mathcal{P} = \{P_\gamma : \gamma \in \Gamma\}$ for $\Gamma = \Theta \times \mathcal{H}$ with $\mathcal{H} = \mathscr{F} \times \mathscr{Z}$ where each P_γ is the probability measure on $\mathbb{R}^{1+K}$ corresponding to the Lebesgue density $p_\gamma(W) = \zeta(Y - f(X_1 + X_2'\theta), X)$. ◇

Part (ii) of the preceding assumption restricts the class of density functions which govern the distribution of the error term and covariates in (24). The key restrictions it imposes are (a) the required conditional mean restriction $\mathbb{E}[\epsilon|X] = 0$, (b) the existence of some

[44] This particular specification of the single index model is relatively simple. More complex versions of this model (e.g. with a more general index specification or a linear component $Z'\xi$) could be analysed using similar techniques. The form used here is deliberately chosen to retain only the key aspect of the model relevant to this paper: that θ may be unidentified or weakly identified for certain values of f, an infinite dimensional nuisance parameter.

moments of specific functions of the data, and (c) a smoothness condition on the density function. Part (iii) restricts the link function f to belong to a specified class of functions; the restrictions imposed on f by this assumption are relatively weak and common in the literature on single index models.[45] Note that these restrictions do not rule out f being constant on $\mathscr{D}$: if $f(v) = c$ for all $v \in \mathscr{D}$ and some $c \in \mathbb{R}$, $f \in \mathscr{F}$.

4.1 Verification of the modelling assumptions

Given a random sample $(W_i)_{i=1}^n$ satisfying assumption SIM, assumption M holds. To establish assumptions LAN and CM(ii) I first need to specify the local perturbations to the nuisance parameter η for which the quadratic approximation will hold.

The considered local perturbations to the nuisance parameters take the form

$$\eta_n(h) := (f + t_n h_1,\ \zeta(1 + t_n h_2)), \quad t_n = n^{-1/2}, \tag{27}$$

with $h_1 \in \dot{\mathscr{F}} := C_b^1(\mathscr{D})$, the set of real valued functions on $\mathbb{R}$ which are continuously differentiable and bounded λ-a.e. on $\mathscr{D}$, and $h_2 \in \dot{\mathscr{Z}}_\eta$ where

$$\dot{\mathscr{Z}}_\eta := \left\{ h_2 \in C_b^{1|1}(\mathbb{R}^{1+K}) : \mathbb{E}[h_2(\epsilon, Z)] = 0,\ \mathbb{E}[\epsilon h_2(\epsilon, X)|X] = 0 \text{ if } (\epsilon, X) \sim \zeta \right\},$$

for $C_b^{1|1}(\mathbb{R}^{1+K})$ is the space of functions $h_2 : \mathbb{R}^{1+K} \to \mathbb{R}$ which are bounded λ-a.e. and such that $e \mapsto h_2(e, X)$ is continuously differentiable with bounded derivative λ-a.e.. The perturbation directions for η are $H_\eta := \dot{\mathscr{F}} \times \dot{\mathscr{Z}}_\eta$ which is a linear subspace of $L_\infty(\lambda) \times L_\infty(G) =: H$, for λ the Lebesgue measure on $\mathbb{R}$. Equip H with the norm $\|h\| = \|h_1\|_{\lambda,\infty} + \|h_2\|_{G,\infty}$.

I now establish that the model is differentiable in quadratic mean and hence (by Proposition 3.10) locally asymptotically normal.

Proposition 4.1. *Suppose that assumption SIM holds, $\theta_n \to \theta \in \Theta$ and $\eta \in \mathcal{H}$ and consider the sequence defined by $\gamma_n = (\theta_n, \eta) \in \Gamma$. Let $\delta_n = I/\sqrt{n}$, $\tau_n \to \tau$, $h_n \in H_\eta$ with $h_n \to h \in H_\eta$ and define $\eta_n : H_\eta \to \mathcal{H}$ as in (27). Then assumption DQM holds with score functions $g_n = \tau \dot{\ell}_{\gamma_n} + B_{\gamma_n} h$ where for $V_{\theta_n} := X_1 + X_2'\theta_n$, $e_n := Y - f(V_{\theta_n})$,*

$$\begin{aligned} \dot{\ell}_{\gamma_n}(W) &:= -\phi(e_n, X) f'(V_{\theta_n}) X_2 \\ [B_{\gamma_n} h](W) &:= -\phi(e_n, X) h_1(V_{\theta_n}) + h_2(e_n, X). \end{aligned}$$

The efficient score function for this model was derived by Newey and Stoker (1993) and is given in the following Proposition.

[45] Cf. Assumption 4.1 in Newey and Stoker (1993); Assumptions A0 – A2 in Kuchibhotla and Patra (2020).

Proposition 4.2. *Consider the sequence $(\gamma_n)_{n\in\mathbb{N}}$ of Proposition 4.1, suppose that assumption SIM holds and*

$$\mathbb{E}[\epsilon\phi(\epsilon, X)|X] = -1, \quad \mathbb{E}[\phi(\epsilon, X)^2|X] < C < \infty, \quad 0 < c < \mathbb{E}[\epsilon^2|X] < C < \infty. \tag{28}$$

Additionally suppose there exists a function $\tilde{m} : \mathbb{R} \to \mathbb{R}$ which is bounded and continuously differentiable with bounded derivative such that $\mathbb{E}[\epsilon\tilde{m}(\epsilon)|X]$ is bounded away from zero uniformly in X. Then assumption CM(ii) holds and for $\omega(X) := \mathbb{E}[\epsilon^2|X]^{-1}$ the efficient score function is

$$\tilde{\ell}_{\gamma_n} := \omega(X)(Y - f(V_{\theta_n}))f'(V_{\theta_n})\left[X_2 - \frac{\mathbb{E}\left[\omega(X)X_2|V_{\theta_n}\right]}{\mathbb{E}\left[\omega(X)|V_{\theta_n}\right]}\right].$$

The (conditional) moment conditions in (28) are standard. The first is a particular case of the (conditional) generalised information equality; it will hold provided differentiation and integration can be interchanged appropriately. The second and third provide uniform bounds on some conditional expectation functions. Existence of the function $\tilde{m}$ is a weak condition; see Assumption 4.2 and the subsequent discussion in Newey and Stoker (1993, p. 1210).

4.2 Implementation of the efficient score test

I now consider estimation of the efficient score function just described in order to satisfy assumptions E and R. Estimation in the (conditionally) heteroskedastic case introduces technical difficulties which are essentially unrelated to the problem studied in this paper and therefore I initially focus on the (conditionally) homoskedastic case and subsequently note that this belongs to a more general class of statistics which remain robust under heteroskedasticity though are typically not power optimal.[46]

Suppose that $\sigma^2 := \mathbb{E}[\epsilon^2|X] = \mathbb{E}[\epsilon^2] > 0$. Under this simplification, the efficient score function is:

$$\tilde{\ell}_{\gamma_n} := \sigma^{-2}(Y - f(V_{\theta_n})f'(V_{\theta_n})\left[X_2 - Z(V_{\theta_n})\right],$$

where $Z(V_{\theta_n}) := \mathbb{E}\left[X_2|V_{\theta_n}\right]$.

To estimate the nonparametric parts of the efficient score function I will use split-sample estimators. Let $N^{(1)} = \{1, \ldots, \lfloor n/2 \rfloor\}$ and $N^{(2)} = [n] \setminus N^{(1)}$. For $i \in [n]$ let N_{-i} denote whichever of $N^{(1)}$ or $N^{(2)}$ that does *not* contain i. The class of estimators considered have the following form:

[46]The class contains a member which achieves the power bound under appropriate conditions but is not feasible as it requires knowledge of the optimal weighting function $\omega(X)$. Cf. the approach taken for estimation in Ichimura (1993).

$$
\begin{aligned}
\hat{f}_{n,i} &\coloneqq \hat{f}_n(V_{\theta_n,i}) \coloneqq \check{f}_n(V_{\theta_n,i}, \hat{\xi}_{1,n,i}) & \hat{\xi}_{1,n,i} &\coloneqq \xi_{1,n}((W_j)_{j\in N_{-i}}), \\
\widehat{f'}_{n,i} &\coloneqq \widehat{f'}_n(V_{\theta_n,i}) \coloneqq \check{f'}_n(V_{\theta_n,i}, \hat{\xi}_{2,n,i}) & \hat{\xi}_{2,n,i} &\coloneqq \xi_{2,n}((W_j)_{j\in N_{-i}}), \\
\hat{Z}_{n,i} &\coloneqq \hat{Z}_n(V_{\theta_n,i}) \coloneqq \check{Z}_n(V_{\theta_n,i}, \hat{\xi}_{3,n,i}) & \hat{\xi}_{3,n,i} &\coloneqq \xi_{3,n}((W_j)_{j\in N_{-i}}),
\end{aligned} \tag{29}
$$

where each $\hat{\xi}_{j,n,i}$ is a (random) vector whose dimension may increase with the sample size. This class of estimators includes, for example, series estimators (of conditional moment functions and their derivatives) as considered by e.g. Newey (1997); Belloni, Chernozhukov, Chetverikov, and Kato (2015); Chen and Christensen (2015); Cattaneo, Farrell, and Feng (2020).[47] In this case, e.g. $f(V_{\theta_n})$ is the conditional expectation of Y given V_{θ_n} and estimates of $f(V_{\theta_n,i})$ and $\widehat{f'}(V_{\theta_n,i})$ can be given as

$$
\hat{f}_n(V_{\theta_n,i}) = \check{f}_n(v, \hat{\xi}_{1,n,i}) = q_n(V_{\theta_n,i})'\hat{\xi}_{1,n,i}, \quad \widehat{f'}_n(V_{\theta_n,i}) = \check{f'}_n(V_{\theta_n,i}, \hat{\xi}_{2,n,i}) = \left[q_n'(V_{\theta_n,i})\right]' \hat{\xi}_{2,n,i},
$$

where q_n is a K_n-vector of basis functions from $\mathbb{R} \to \mathbb{R}$, q_n' their derivatives and

$$
\hat{\xi}_{1,n,i} = \hat{\xi}_{2,n,i} = \left(\sum_{j\in N_{-i}} q_n(V_{\theta_n,j}) q_n(V_{\theta_n,j})' \right)^{-1} \left(\sum_{j\in N_{-i}} q_n(V_{\theta_n,j}) Y_j \right).
$$

Similar estimators can be constructed for $Z(V_{\theta_n})$ which is the conditional expectation of X_2 given V_{θ_n}.

Given such estimators I form an estimate of σ^2 as

$$
\hat{\sigma}_n^2 \coloneqq \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{f}_{n,i})^2,
$$

and the estimates

$$
\hat{\ell}_{n,\theta_n}(W_i) \coloneqq \hat{\sigma}_n^{-2} \left(Y_i - \hat{f}_{n,i} \right) \widehat{f'}_{n,i} \left[X_{2,i} - \hat{Z}_{n,i} \right], \quad \check{\mathcal{I}}_{n,\theta_n} \coloneqq \frac{1}{n} \sum_{i=1}^n \hat{\ell}_{n,\theta_n}(W_i) \hat{\ell}_{n,\theta_n}(W_i)'. \tag{30}
$$

Let $\hat{\mathcal{I}}_{n,\theta_n}$ be the eigendecomposition-truncated version of $\check{\mathcal{I}}_{n,\theta_n}$ at ν_n (analogously to (21)), where $(\nu_n)_{n\in\mathbb{N}}$ is a non-negative sequence converging to zero. With these estimators assumptions E and R can be shown to hold under conditions on the sequence $(\nu_n)_{n\in\mathbb{N}}$ and the following high-level condition which assumes certain (probabilistic) rates of convergence

[47] This class of estimators also includes, for example, kernel estimators.

hold for

$$\mathcal{R}_{1,n,i} \coloneqq \left(\int \left[\check{f}_n(v, \hat{\xi}_{1,n,i}) - f(v)\right]^2 \mathrm{d}\mathcal{V}_n(v)\right)^{1/2},$$
$$\mathcal{R}_{2,n,i} \coloneqq \left(\int \left[\check{f'}_n(v, \hat{\xi}_{2,n,i}) - f'(v)\right]^2 \mathrm{d}\mathcal{V}_n(v)\right)^{1/2},$$
$$\mathcal{R}_{3,n,i} \coloneqq \left(\int \left\|\check{Z}_n(v, \hat{\xi}_{3,n,i}) - Z(v))\right\|_2^2 \mathrm{d}\mathcal{V}_n(v)\right)^{1/2},$$

where $\mathcal{V}_n$ is the distribution of V_{θ_n}.

Assumption SIM-NP(i). Suppose that $\mathscr{X}$ is a compact set, equation (28) holds, $\sigma^2 \coloneqq \mathbb{E}[\epsilon^2|X] = \mathbb{E}[\epsilon^2]$, $\mathbb{E}[\epsilon^4] < \infty$ and with P_{γ_n}-probability approaching one for $l \in [3]$ and each $i \in [n]$, $\mathcal{R}_{l,n,i} \leq r_n = o(n^{-1/4})$. $\diamond$

The rates in assumption SIM-NP(i) are attainable under reasonable regularity conditions. For example, series (linear sieve) estimators of f, f' and Z can attain these rates given sufficient smoothness of the target function and other regularity conditions. See, inter alia, Belloni et al. (2015); Chen and Christensen (2015); Cattaneo et al. (2020); Huang and Su (2021). This assumption is sufficient for the estimator of σ^{-2} to be $\sqrt{n}$-consistent.

Lemma 4.3. *Suppose that assumption SIM holds and $\sigma^2 \coloneqq \mathbb{E}[\epsilon^2|X] = \mathbb{E}[\epsilon^2] \in (0, \infty)$ and let $(\gamma_n)_{n\in\mathbb{N}}$ be as in Proposition 4.1. If $\mathbb{E}[\epsilon^4] < \infty$ and with P_{γ_n}-probability approaching one, $\mathcal{R}_{1,n,i} \leq r_n = o(n^{-1/4})$, then $\sqrt{n}(\hat{\sigma}_n^{-2} - \sigma^{-2}) = O_{P_{\gamma_n}}(1)$.*

In the general, heteroskedastic, case I consider a related estimator, where – as in Ichimura (1993) – a *known* weighting function $\breve{\omega}(X)$ is utilised in place of the unknown $\omega(X)$. In particular, I estimate the function

$$\breve{\ell}_{\gamma_n}(W) \coloneqq \breve{\omega}(X)(Y - f(V_{\theta_n}))f'(V_{\theta_n}) \left[X_2 - \frac{\mathbb{E}\left[\breve{\omega}(X)X_2|V_{\theta_n}\right]}{\mathbb{E}\left[\breve{\omega}(X)|V_{\theta_n}\right]}\right].$$

Clearly if $\breve{\omega} = \omega$, $\breve{\ell}_{\gamma_n}$ coincides with the efficient score function and hence power optimality results are available if the conditions outlined in section 3 hold. In the case where $\breve{\omega} \neq \omega$ the resulting statistic will not be power optimal, but will retain the locally uniform size control properties of the efficient score statistic.

In the heteroskedastic case, I replace the function $Z(V_{\theta_n}) \coloneqq \mathbb{E}[X_2|V_{\theta_n}]$ with $Z_1(V_{\theta_n})/Z_2(V_{\theta_n})$ where $Z_1(V_{\theta_n}) \coloneqq \mathbb{E}[\breve{\omega}(X)X_2|V_{\theta_n}]$ and $Z_2(V_{\theta_n}) \coloneqq \mathbb{E}[\breve{\omega}(X)|V_{\theta_n}]$. Let $\hat{f}_{n,i}$ and $\widehat{f'}_{n,i}$ be as in (29) and similarly define

$$\begin{aligned} \hat{Z}_{1,n,i} &\coloneqq \hat{Z}_{1,n}(V_{\theta_n,i}) \coloneqq \check{Z}_{1,n}(V_{\theta_n,i}, \hat{\xi}_{3,n,i}) \qquad & \check{\xi}_{3,n,i} &\coloneqq \xi_{3,n}((W_j)_{j\in N_{-i}}) \\ \hat{Z}_{2,n,i} &\coloneqq \hat{Z}_{2,n}(V_{\theta_n,i}) \coloneqq \check{Z}_{2,n}(V_{\theta_n,i}, \hat{\xi}_{4,n,i}) \qquad & \check{\xi}_{4,n,i} &\coloneqq \xi_{4,n}((W_j)_{j\in N_{-i}}). \end{aligned} \tag{31}$$

With these estimates I can form an estimate of $\breve{\ell}_{\gamma_n}$ and $\Upsilon_{\gamma_n} := P_{\gamma_n}\breve{\ell}_{\gamma_n}\breve{\ell}'_{\gamma_n}$ according to

$$\breve{\ell}_{n,\theta_n}(W_i) := \breve{\omega}(X_i)\left(Y_i - \hat{f}_{n,i}\right)\widehat{f'}_{n,i}\left[X_{2,i} - \frac{\hat{Z}_{1,n,i}}{\hat{Z}_{2,n,i}}\right], \quad \hat{\Upsilon}_{n,\theta_n} := \frac{1}{n}\sum_{i=1}^{n}\breve{\ell}_{n,\theta_n}(W_i)\breve{\ell}_{n,\theta_n}(W_i)'. \tag{32}$$

Let $\check{\Upsilon}_{n,\theta_n}$ be the eigendecomposition-truncated version of $\hat{\Upsilon}_{n,\theta_n}$ at ν_n (analogously to (21)). The test statistic that will be used in this case (for testing a two-sided hypothesis) is

$$\check{S}_{n,\theta} := \left(\sqrt{n}\mathbb{P}_n\breve{\ell}_{n,\theta}\right)'\check{\Upsilon}^{\dagger}_{n,\theta}\left(\sqrt{n}\mathbb{P}_n\breve{\ell}_{n,\theta}\right), \tag{33}$$

with the test and confidence then being defined analogously to (10) and (11) with $\check{S}_{n,\theta}$ in place of $\hat{S}_{n,\theta}$. Denote these respectively by $\check{\phi}_{n,\theta_0}$ and $\check{C}_n$. This test will be called the "pseudo efficient score test" in what follows. Let $\breve{\mathcal{R}}_{l,n,i} := \mathcal{R}_{l,n,i}$ for $l = 1, 2$ and define

$$\breve{\mathcal{R}}_{3,n,i} := \left(\int \left\|\check{Z}_{1,n}(v, \hat{\xi}_{3,n,i}) - Z_1(v))\right\|_2^2 \,\mathrm{d}\mathcal{V}_n(v)\right)^{1/2}$$
$$\breve{\mathcal{R}}_{4,n,i} := \left(\int \left(\check{Z}_{2,n}(v, \hat{\xi}_{4,n,i}) - Z_2(v))\right)^2 \,\mathrm{d}\mathcal{V}_n(v)\right)^{1/2}.$$

In the heteroskedastic case, assumption SIM-NP(i) is replaced by the following assumption:

Assumption SIM-NP(ii). Suppose that $\mathscr{X}$ is a compact set, equation (28) holds, $\mathbb{E}[\epsilon^4] < \infty$, $\breve{\omega} : \mathbb{R}^K \to (\underline{\omega}, \overline{\omega})$ is a known function and with P_{γ_n}-probability approaching one for $l \in [4]$ and each $i \in [n]$, $\breve{\mathcal{R}}_{l,n,i} \leq r_n = o(n^{-1/4})$. ◇

The rates required by this assumption are attainable under reasonable regularity conditions; cf. the discussion following assumption SIM-NP(i).

4.3 Asymptotic properties

I start by detailing the asymptotic properties of the efficient score statistic in the homoskedastic case.

Proposition 4.4. *Suppose that assumptions SIM, SIM-NP(i) hold and there exists a function $\tilde{m}$ as in Proposition 4.2. Consider the sequence $(\gamma_n)_{n\in\mathbb{N}}$ of proposition 4.1, suppose the observations form an i.i.d. sample and $\hat{\ell}_{n,\theta_n}$ and $\hat{\mathcal{I}}_{n,\theta_n}$ are as in (30), with $0 \leq \nu_n \to 0$ such that $r_n + n^{-1/2}\log(n)^{1/2+\kappa} = o(\nu_n)$ for some $\kappa > 0$. Then assumptions M, LAN, CM(ii), E and R hold.*

With the estimators $\hat{\ell}_{n,\theta_n}$ and $\hat{\mathcal{I}}_{n,\theta_n}$ the efficient score statistic, test and confidence set can be defined as in section 3.2. The following results demonstrate that the efficient score test is optimal under strong-identification asymptotics and provides robust size control and the corresponding confidence sets robust coverage, including under asymptotics in which the function f is local to a constant (function) at rate $\sqrt{n}$, corresponding to a setting where θ is weakly identified.

Corollary 4.5. *In the setting of Proposition 4.4, let H_η' be a compact subset of H_η. Then, the efficient score test satisfies*

$$\limsup_{n\to\infty} \sup_{h\in H_\eta'} P^n_{\gamma_n,0,h}\phi_{n,\theta_0} \leq \alpha,$$

and, for any compact $\Theta' \subset \Theta$, the corresponding test inversion confidence sets satisfy

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta} \inf_{h\in H_\eta'} P^n_{\gamma_n,0,h}(\theta \in \hat{C}_n) \geq 1-\alpha.$$

Corollary 4.6. *In the setting of Proposition 4.4, suppose additionally that* $\operatorname{rank}(\tilde{\mathcal{I}}_\gamma) > 0$. *If $d_\theta = 1$, then the efficient score test is locally asymptotically uniformly most powerful unbiased. If $d_\theta > 1$, then the efficient score test is locally asymptotically maximin.*

I now establish a similar uniform size control result for the heteroskedastic case, with the psuedo efficient score test defined immediately following (33).

Proposition 4.7. *Suppose that that assumptions SIM, SIM-NP(ii) hold and there exists a function $\tilde{m}$ as in Proposition 4.2. Consider the sequence $(\gamma_n)_{n\in\mathbb{N}}$ of proposition 4.1, suppose the observations form an i.i.d. sample and $\check{\ell}_{n,\theta_n}$ and $\check{\Upsilon}_{n,\theta_n}$ are as in (32), with $0 \leq \nu_n \to 0$ such that $r_n + n^{-1/2}\log(n)^{1/2+\kappa} = o(\nu_n)$ for some $\kappa > 0$. Let H_η' be a compact subset of H_η. Then, the psuedo efficient score test satisfies*

$$\limsup_{n\to\infty} \sup_{h\in H_\eta'} P^n_{\gamma_n,0,h}\check{\phi}_{n,\theta_0} \leq \alpha,$$

and, for any compact $\Theta' \subset \Theta$, the corresponding test inversion confidence sets satisfy

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta} \inf_{h\in H_\eta'} P^n_{\gamma_n,0,h}(\theta \in \check{C}_n) \geq 1-\alpha.$$

I remark here that if $\breve{\omega} = \omega$ then each $\check{\ell}_{\gamma_n} = \tilde{\ell}_{\gamma_n}$. In this situation, if the rank of $\Upsilon_\gamma = \tilde{\mathcal{I}}_\gamma$ is positive, then in the setting of Proposition 4.7 the (pseudo) efficient score test is is locally asymptotically uniformly most powerful unbiased if $d_\theta = 1$ and locally asymptotically maximin if $d_\theta > 1$. However, as this is infeasible in the heteroskedastic case, I do not state a formal power result.

4.4 Simulation study

I conduct a simulation study to examine the finite sample properties of the efficient score test. I draw $n \in \{200, 400, 600, 800\}$ samples from model (24) for a number of different functions f and distributions ζ. I set $K = 1$ throughout and examine finite sample size using 5000 Monte Carlo replications, at a nominal level of 5%. In each case I test the null $H_0 : \theta = 1$.

Overall the simulation experiments suggest the asymptotic results of section 4.3 provide a good guide to the performance of the efficient score test (and psuedo efficient score test) in finite samples.

4.4.1 Homoskedastic case

Initially I consider the homoskedastic case. The error term is taken as either (1) $\epsilon \sim \mathcal{N}(0, 1)$ or (2) $\epsilon|\xi \sim \sqrt{5}(-1)^{\xi}\,\mathrm{Beta}(2, 3)$, $\xi \sim \mathrm{Bernoulli}(1/2)$. In both cases $\mathbb{E}\epsilon = 0$ and $\mathbb{V}\epsilon = 1$. The covariates are drawn as either (a) $X_k = Z_k$ or (b) $X = (Z_1, 0.2Z_1 + 0.4Z_2 + 0.8)$ where $Z_k \sim U(-1, 1)$ for $k = 1, 2$. The link functions considered take the form $f(v) = \delta f^{\star}(v)$ for $f^{\star} \in \{v \mapsto c_1(1 + \exp(-v))^{-1}, v \mapsto c_2 \exp(-v^2), v \mapsto c_3v^2\}$, $\delta \in (0, 1)$.[48] Each of these functions has a different shape; the scalars c_i $(i = 1, 2, 3)$ vary across the functions $f^{\star}$ and distributions for X and are chosen so that the variance of $f^{\star}(V_\theta)$ equals 4 under H_0: $\theta = 1$, whilst δ is taken the same for all functions and used to scale this variance.[49]

To examine the finite sample size of the proposed test, the efficient score function and efficient information matrix are estimated as in (30), with split-sample (penalised) smoothing cubic B-splines used to estimate each of $\hat{f}$, $\widehat{f'}$ and $\hat{Z}$.[50] I truncate the efficient information matrix at machine precision. Additionally I consider a Wald statistic estimated using an Ichimura (1993) style estimator, which uses the same estimates of $\hat{f}$, $\widehat{f'}$ and $\hat{Z}$ as the efficient score statistic.[51] The finite sample empirical rejection frequencies are reported in tables 1 - 4. In all specifications considered the efficient score provides good size control, whereas the Wald statistic based on the Ichimura (1993) type estimator described above displays substantial over-rejection, particularly for small δ.

To analyse the finite sample power of the efficient score test I consider the finite sample rejection frequency of the efficient score test of $\theta = 1$ for a grid of values around θ. Specifically, I take 21 equally spaced values between 0.875 and 1.125 and all other parameters are the same as for the simulations used to investigate finite sample size. Figures 5 - 8 plot the

[48] The first of these is the standard Logistic CDF.

[49] The scaling constants c are calculated in closed form for the case (a) with $X = (Z_1, Z_2)$. In the correlated case (b), evaluation of the integrals becomes substantially more complex and so simulated values are used, based on 10,000,000 draws.

[50] In particular I use the `smooth.spline` function in R with its default knot choice and penalty settings.

[51] This approach estimates θ by minimising the criterion $\theta \mapsto \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_{n,i}(V_\theta))^2$; the estimates of $\widehat{f'}$ and $\hat{Z}$ are necessary to construct the asymptotic variance.

finite sample power function of the efficient score test, which demonstrate that – as expected – higher δ leads to higher power for the same distance from the null.

4.4.2 Heteroskedastic case

I now consider the heteroskedastic case. I consider two specifications for the error term: (1) $\epsilon \sim \mathcal{N}(0, s_1 \log(2+(X_1+X_2\theta)^2))$ and (2) $\epsilon \sim \mathcal{N}(0, s_2(1+5\sin(X_2)^2))$ where the constants s_i $(i = 1, 2)$ are chosen such that in each case $\mathbb{V}(\epsilon) = 1$ (unconditionally) under $H_0 : \theta = 1$.[52] The distributions for the covariates and the link functions used are the same as in the homoskedastic case.

To examine the finite sample size of the proposed test, the pseudo-efficient score function and its variance matrix are estimated as in (32) with split-sample (penalised) smoothing cubic B-splines used to estimate each of $\hat{f}$, $\widehat{f'}$, $\hat{Z}_1$ and $\hat{Z}_2$.[53] As in the homoskedastic case I truncate the variance matrix at machine precision. Additionally I consider a Wald statistic estimated using an Ichimura (1993) style estimator, which uses the same nonparametric estimates as the psuedo-efficient score statistic.[54]

The finite sample rejection frequencies with $\breve{\omega}(X)$ is taken as the infeasible truth $\omega(X)$ are reported in tables 5 - 8, whilst tables 9 - 12 report the finite sample size where $\breve{\omega}(X) = 1$. The results demonstrate qualitatively the same conclusions as the homoskedastic case, with the pseudo efficient score statistic always providing good size control, unlike the Wald statistic, which displays large over-rejection, particularly for small δ.

As in the homoskedastic case, to analyse the finite sample power of the pseudo efficient score test I consider the finite sample rejection frequency of the efficient score test of $\theta = 1$ for a grid of values around θ. As in the homoskedastic case, I consider 21 equally spaced values between 0.875 and 1.125 with all other parameters the same as for the simulations used to investigate finite sample size. Figures 9 - 12 plot the finite sample power curves. Similar observations apply as in the homoskedastic case, with higher δ leading to higher power for a given distance from the null. Moreover, as expected, the optimal (but infeasible) weighting scheme delivers higher power, though the difference seems to be relatively small for the designs considered.

5 Linear simultaneous equations models

In this section, I work out the details of the application of the theory developed in section 3 to a class of linear simultaneous equations models (LSEMs) where identification is based on

[52]These are determined by simulation with 10,000,000 draws.
[53]See footnote 50.
[54]This approach estimates θ by minimising the criterion $\theta \mapsto \frac{1}{n}\sum_{i=1}^{n} \breve{\omega}(X_i)(Y_i - \hat{f}_{n,i}(V_\theta))^2$; the estimates of $\widehat{f'}$, $\hat{Z}_1$, $\hat{Z}_2$ are necessary to construct the asymptotic variance.

an assumption of mutually independent and non-Gaussian errors. Under this assumption, no external information (e.g. instrumental variables) is required in order to identify the parameter of interest.

Consider the following linear simultaneous equations model (LSEM)

$$Y = RX + V, \quad V = A(\theta,\sigma)^{-1}\epsilon, \quad \mathbb{E}\epsilon = 0, \mathbb{V}\epsilon = I, \tag{34}$$

where the K components of ϵ are mutually independent, $X = (1, \tilde{X}')'$ is a vector of covariates independent of ϵ. R is a $K \times L$ matrix of regression coefficients and $A(\theta,\sigma)$ is a $K \times K$ invertible matrix. For later convenience I collect the Euclidean nuisance parameters R and σ into one vector: $\beta := (\beta_1', \beta_2')' := (\sigma', \mathrm{vec}(R)')'$.

As is well known, in simultaneous equations models of this form the elements of the mixing matrix, $A(\theta,\sigma)$, are not identified without further restrictions. However, if no more than one component of ϵ is Gaussian, the elements of the matrix $A(\theta,\sigma)$ are identified up to column permutation and sign changes (Comon, 1994). Imposition of sign restrictions and labelling of the shocks can then yield identification of the elements of $A(\theta,\sigma)$ which – assuming an identifiable parametrisation – yields that of θ.

Nevetheless, the identifying assumption that no more than one component of ϵ is Gaussian is not innocuous. In particular, depending on the parametrisation of the model, if this assumption fails, θ may be underidentified or completely unidentified. Moreover, as is typical in models with points of identification failure, the impact of the potential identification problem here is not binary. “Weak non-Gaussianity”, where the error distribution is sufficiently close to Gaussianity relative to sampling uncertainty, can cause problems for inference methods which assume non-Gaussianity to obtain identification.[55] In this section I extend the analysis of Lee and Mesters (2021a) to demonstrate that inference based on the efficient score test is (i) robust to weak identification (in addition to underidentification and complete unidentification) and (ii) minimax optimal if θ is identified or underidentified.[56]

The first step of the analysis is to formally set up the model under consideration. Let η_0 denote the (Lebesgue) density of $\tilde{X}$ and for each $k = 1, \ldots, K$ let η_k be the (Lebesgue) density of ϵ_k and define ϕ_k as the log-density scores, i.e. $\phi_k(e) := \frac{\mathrm{d}\log\eta_k(s)}{\mathrm{d}s}(e)$. I will require a number of moments of (functions of) ϵ and $\tilde{X}$ to satisfy certain conditions.[57] In particular, for each $k \in [K]$ and some $\delta > 0$

$$\mathbb{E}\epsilon_k = 0,\ \mathbb{E}\epsilon_k^2 = 1,\ \mathbb{E}|\epsilon_k|^{4+\delta} < \infty,\ \mathbb{E}|\phi_k(\epsilon_k)|^{4+\delta} < \infty,\ \mathbb{E}\epsilon_k^4 - 1 > (\mathbb{E}\epsilon_k^3)^2, \tag{35}$$

[55]See Lee and Mesters (2021a) for simulation evidence of this phenomenon.

[56]Lee and Mesters (2021a) provide simulation evidence of a weak identification problem in this class of models, but their theoretical work only considers robustness against fixed distributions under which θ may be identified, underidentified or unidentified and does not cover weak identification.

[57]These conditions are the same as imposed in Lee and Mesters (2021a). Additionally I note that such fourth-moment conditions are common for conducting inference on variance parameters (e.g. White, 1980).

and

$$\mathbb{E}\tilde{X}\tilde{X}' \succ 0, \quad \mathbb{E}\|\tilde{X}\|_2^{4+\delta} < \infty. \tag{36}$$

These moment restrictions are used to characterise the DGPs permitted by the model. Specifically, the density functions η_k and η_0 are assumed to belong (respectively) to the sets $\mathscr{G}$ and $\mathscr{X}$ which are defined as follows:

$$\mathscr{G} := \left\{ g \in L_1(\mathbb{R}) : g \geq 0, \int g \, \mathrm{d}\lambda = 1, \sqrt{g} \in C^1(\mathbb{R}), \text{ if } \epsilon_k \sim g \text{ then } (35) \right\}, \tag{37}$$

$$\mathscr{X} := \left\{ g \in L_1(\mathbb{R}^{L-1}) : g \geq 0, \int g \, \mathrm{d}\lambda^{L-1} = 1, \text{ if } \tilde{X} \sim g \text{ then } (36) \right\}, \tag{38}$$

where $L_1(\mathbb{R}^d)$ denotes the space of integrable functions on $\mathbb{R}^d$ with respect to the Lebesgue measure (which is denoted by λ^d or λ if the dimension is clear from context) and $C^1(\mathbb{R})$ denotes the space of functions $\mathbb{R} \to \mathbb{R}$ which are continuously differentiable λ-a.e.. Finally the parameter $\beta = (\sigma', \operatorname{vec}(R)')'$ is assumed to belong to $\mathscr{B} \subset \mathbb{R}^{d_\beta}$. I will consider two restrictions on $\mathscr{B}$. Firstly it will be permitted to be an (otherwise unrestricted) open set. Alternatively – to explicitly handle the case of sign restrictions (or non-negativity restrictions on variances) – it will be permitted to have the form

$$\mathscr{B} = \mathscr{B}_1 \times \mathscr{B}_2, \qquad \mathscr{B}_1 = \prod_{l=1}^{d_\sigma} \mathscr{B}_{1,l}, \tag{39}$$

where $\mathscr{B}_2 \subset \mathbb{R}^{KL}$ is open and each $\mathscr{B}_{1,l} \subset \mathbb{R}$ is either open or one of $(-\infty, 0]$ or $[0, \infty)$.

The assumptions imposed on the LSEM model (34) are summarised as follows:

Assumption LSEM. $W = (Y, \tilde{X})$ satisfies (34) where the K components of ϵ have marginal densities η_k ($k \in [K]$). Let the density of $\tilde{X}$ be η_0.[58]

- (i) $\Theta \subset \mathbb{R}^{d_\theta}$ is an open set and $\mathscr{B} \subset \mathbb{R}^{d_\beta}$ is either open or has the form $\mathscr{B}_1 \times \mathscr{B}_2$ where these factors are as described following (39).
- (ii) The components of ϵ are mutually independent and ϵ is independent of X.
- (iii) $\eta_k \in \mathscr{G}$ for each $k \in [K]$ and $\eta_0 \in \mathscr{X}$, for $\mathscr{G}$ and $\mathscr{X}$ defined in (37) and (38) respectively.
- (iv) The function $(\theta, \sigma) \mapsto A(\theta, \sigma)$ is continuously differentiable with l-th partial derivative $D_{1,l}(\theta, \sigma)$ and the functions $(\theta, \sigma) \mapsto D_{1,l}(\theta, \sigma) A(\theta, \sigma)^{-1}$ are Lipschitz continuous.

The model is given by $\mathcal{P} = \{P_\gamma : \gamma \in \Gamma = \Theta \times \mathcal{H}\}$ with $\mathcal{H} := \mathscr{B} \times \mathscr{X} \times \prod_{k=1}^K \mathscr{G}$ and where each P_γ has (Lebesgue) density

$$p_\gamma(W) = |\det(A(\theta, \sigma))| \prod_{k=1}^K \eta_k(A_k[Y - RX]) \times \eta_0(\tilde{X}). \tag{40}$$

[58] Each η_k is a density with respect to Lebesgue measure on the appropriate Euclidean space.

◇

The moment and smoothness conditions imposed by part (iii) of assumption LSEM are reasonably weak, as are the smoothness conditions in (iv). The independence in (ii) is, however, restrictive. Mutual independence of the components of ϵ is a testable assumption in applications (Matteson and Tsay, 2017; Amengual, Fiorentini, and Sentana, 2021). The independence of $\tilde{X}$ and ϵ could be replaced by a conditional moment restriction, for which the general approach outlined in this paper would continue to hold, but the analysis below would need to be redone under this alternative assumption, with the efficient score function taking a different form.

5.1 Verification of the modelling assumptions

Assumption LSEM coupled with the assumption that the observed data comprises an i.i.d. sample $(W_i)_{i=1}^n$ ensures that assumption M holds. I next show that assumption DQM holds, which is sufficient to imply assumption LAN by proposition 3.10.

For any $l \in [d_\theta + d_\sigma]$ and any $(k, j) \in [K]^2$, let $\zeta_{l,k,j} := [D_{1,l}(\theta, \sigma)]_k [A^{-1}]'_j$. Additionally write $D_{2,l}$ for the derivative of R with respect to the l-th component of $\beta_2 = \operatorname{vec}(R)$. $C_b^1(\mathbb{R})$ denotes the space of functions $\mathbb{R} \to \mathbb{R}$ which are bounded, continuously differentiable and have bounded derivatives λ-a.e. and $C_b(\mathbb{R}^L)$ denotes the space of functions $\mathbb{R}^L \to \mathbb{R}$ which are bounded and continuous λ^L-a.e.. Define the sets $\dot{\mathscr{G}}_{\eta,k}$ and $\dot{\mathscr{Z}}_\eta$ as:

$$\dot{\mathscr{G}}_{\eta,k} := \left\{ h_k \in C_b^1(\mathbb{R}) : \int h_k \, \mathrm{d}G_k = \int h_k \iota \, \mathrm{d}G_k = \int h_k \kappa \, \mathrm{d}G_k = 0 \right\}, \tag{41}$$

$$\dot{\mathscr{Z}}_\eta := \left\{ h_0 \in C_b(\mathbb{R}^{L-1}) : \int h_0 \, \mathrm{d}G_0 = 0 \right\} \tag{42}$$

where G_k is the measure on $\mathbb{R}$ corresponding to η_k ($k \in [K]$), G_0 the measure on $\mathbb{R}^{L-1}$ corresponding to η_0, ι denotes the identity function and $\kappa(e) := e^2 - 1$. Let

$$H_\eta := \prod_{l=1}^{d_\sigma} \mathscr{V}_l \times \mathbb{R}^{KL} \times \dot{\mathscr{Z}}_\eta \times \prod_{k=1}^{K} \dot{\mathscr{G}}_{\eta,k} \subset H := \mathbb{R}^{d_\beta} \times L_\infty(\lambda^{L-1}) \times \prod_{k=1}^{K} L_\infty(\lambda), \tag{43}$$

where each $\mathscr{V}_l = \mathbb{R}$ if β is an interior point of $\mathscr{B}$ and otherwise (i) $\mathscr{V}_l = [0, \infty)$ if $\mathscr{B}_{1,l} = [0, \infty)$ and $\sigma_l = 0$ or (ii) $\mathscr{V}_l = (-\infty, 0]$ if $\mathscr{B}_{1,l} = (-\infty, 0]$ and $\sigma_l = 0$. H is equipped with the norm $\|h\| := \|b\|_2 + \|h_0\|_{\lambda^{L-1},\infty} + \sum_{k=1}^K \|h_k\|_{\lambda,\infty}$, for $h = (b, h_0, \dots, h_K) \in H$.[59] H_η is a linear subspace of H whenever β is an interior point of $\mathscr{B}$.

The sequences of base parameters considered are $\gamma_n = (\theta_n, \eta)$, with local perturbations

[59] Each of the factors defining H is a Banach space (with the corresponding norm as just indicated) and hence the same is true of H when equipped with the indicated norm.

of the form $\theta_n + \tau_n/\sqrt{n} \to \theta$ with $\tau_n \to \tau$ and

$$\eta_n(h_n) := (\beta_1 + t_n b_{1,n}, \beta_2 + t_n b_{2,n}, \eta_0(1 + t_n h_{n,0}), \eta_1(1 + t_n h_{n,1}), \ldots, \eta_K(1 + t_n h_{n,K})) \quad (44)$$

with $h_n \to h$ (all in H_η); note that $\eta_n(h_n) \to \eta$.

The following proposition establishes the quadratic mean differentiability of the model and hence LAN in view of Proposition 3.10.

Proposition 5.1. *Suppose that assumption LSEM holds, $\theta_n \to \theta \in \Theta$ and $\eta \in \mathcal{H}$ and consider the sequence defined by $\gamma_n = (\theta_n, \eta) \in \Gamma$. Let $\delta_n = I/\sqrt{n}$, $t_n := n^{-1/2}$, $\tau_n \to \tau$, $h_n := (b_n, h_{n,0}, h_{n,1}, \ldots, h_{n,K})$ (with $b_n = (b_{1,n}', b_{2,n}')'$), with $h_n \to h$, and define $\eta_n : H_\eta \to \mathcal{H}$ as in (44). Then assumption DQM holds, with $g_n := \tau' \dot{\ell}_{\gamma_n} + B_{\gamma_n} h$ where for $l = 1, \ldots, d_\theta$,*

$$\dot{\ell}_{\gamma_n,l}(W) := \sum_{k=1}^{K} \left[\zeta_{l,k,k,n}(\phi_k(A_{n,k}V)A_{n,k}V + 1) + \sum_{j=1, j\neq k}^{K} \zeta_{l,k,j,n}\phi_k(A_{n,k}V)A_{n,j}V \right],$$

$$\begin{aligned}[B_{\gamma_n}h](W) := &\sum_{m=d_\theta+1}^{d_\theta+d_{b_1}} b_{1,m} \sum_{k=1}^{K} \left[\zeta_{m,k,k,n}(\phi_k(A_{n,k}V)A_{n,k}V + 1) + \sum_{j=1, j\neq k}^{K} \zeta_{m,k,j,n}\phi_k(A_{n,k}V)A_{n,j}V \right] \\ &+ \sum_{k=1}^{K} \phi_k(A_{n,k}V) \left[-A_{n,k} \sum_{l=1}^{d_\beta} b_{2,l} D_{2,l} X \right] + h_0(\tilde{X}) + \sum_{k=1}^{K} h_k(A_{n,k}V),\end{aligned}$$

with $A_n := A(\theta_n, \sigma)$, $V := Y - RX$.

In order to simplify the expression of the the efficient score function, I suppose the following moment conditions on ϕ_k hold.

$$\mathbb{E}\phi_k(\epsilon_k) = 0, \ \mathbb{E}\phi_k(\epsilon_k)\epsilon_k = -1, \ \mathbb{E}\phi_k(\epsilon_k)\epsilon_k^2 = 0, \ \mathbb{E}\phi_k(\epsilon_k)\epsilon_k^3 = -3. \quad (45)$$

These moment conditions are weak; if (35) holds then a sufficient condition for (45) to hold is that the tails of the densities satisfy $\eta_k(x) = o(x^{-3})$.[60]

Proposition 5.2. *Suppose that assumption LSEM and equation (45) hold and consider the sequence $(\gamma_n)_{n\in\mathbb{N}}$ of Proposition 5.1. Then assumption CM(ii) holds and (provided the inverse in the subsequent display exists) the efficient score function, $\tilde{\ell}_{\gamma_n}$, is given by*

$$\tilde{\ell}_{\gamma_n} = \tilde{\ell}_{\gamma_n,1} - \left[P_{\gamma_n} \tilde{\ell}_{\gamma_n,1} \tilde{\ell}_{\gamma_n,2}' \right] \left[P_{\gamma_n} \tilde{\ell}_{\gamma_n,2} \tilde{\ell}_{\gamma_n,2}' \right]^{-1} \tilde{\ell}_{\gamma_n,2}, \quad (46)$$

[60] See Lemma S8 in Lee and Mesters (2021b). Alternatively, these conditions will hold provided differentiation and integration can be appropriately interchanged.

where for $l = 1, \ldots, d_\theta$, $m = 1, \ldots, d_{b_1}$, $s = 1, \ldots, d_{b_2}$, $v \coloneqq V - RX$ *and* $\mu \coloneqq \mathbb{E}X$,

$$
\begin{aligned}
\tilde{\ell}_{\gamma_n,1,l}(W) &= \sum_{k=1}^{K} \left[\zeta_{l,k,k,n} \left(\tau_{k,1} A_{n,k} V + \tau_{k,2} \kappa(A_{n,k}V) \right) + \sum_{j=1, j \neq k}^{K} \zeta_{l,k,j,n} \phi_k(A_{n,k}V) A_{n,j} V \right] \\
\tilde{\ell}_{\gamma_n,2,m}(W) &= \sum_{k=1}^{K} \left[\zeta_{m,k,k,n} \left(\tau_{k,1} A_{n,k} V + \tau_{k,2} \kappa(A_{n,k}V) \right) + \sum_{j=1, j \neq k}^{K} \zeta_{m,k,j,n} \phi_k(A_{n,k}V) A_{n,j} V \right] \\
\tilde{\ell}_{\gamma_n,2,d_{b_1}+s}(w) &= \sum_{k=1}^{K} [-A_{n,k} D_{2,s}] \left[(x - \mu) \phi_k(A_{n,k}V) - \mu \left(\varsigma_{k,1} A_{n,k} V + \varsigma_{k,2} \kappa(A_{n,k}V) \right) \right],
\end{aligned}
$$

and

$$
\tau_k \coloneqq M_k^{-1} \begin{pmatrix} 0 \\ -2 \end{pmatrix}, \quad \varsigma_k \coloneqq M_k^{-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \textit{with } M_k \coloneqq \begin{pmatrix} 1 & P_{\gamma_n}(A_{n,k}V)^3 \\ P_{\gamma_n}(A_{n,k}V)^3 & P_{\gamma_n}(A_{n,k}V)^4 - 1 \end{pmatrix}.
$$

The preceding proposition requires the inverse of the variance matrix of $\tilde{\ell}_{\gamma_n,2}$ to exist. This is only necessary for the projection to be expressed in this precise form; if the matrix in question is singular, one can drop linearly dependent (in $L_2(P_{\gamma_n})$) elements from $\tilde{\ell}_{\gamma_n,2}$ until it is nonsingular. Additionally note that M_k is not indexed by n; under P_{γ_n}, $A_{n,k}V \sim \eta_k$ and so the moments making up M_k are constant in n.

5.2 Implementation of the efficient score test

Next I impose conditions which are sufficient for the construction of estimates of the efficient score function and efficient information matrix which satisfy assumptions E and R. First, I suppose that there is an appropriate estimator of each log density score ϕ_k available.

Assumption DSE. Suppose that $(\beta_n)_{n \in \mathbb{N}} \subset \mathscr{B}$ is a deterministic sequence with $\sqrt{n}(\beta_n - \beta) = O(1)$. Let $\gamma_n' \coloneqq (\theta_n, \beta_n, \eta)$, $A_n \coloneqq A(\theta_n, \beta_{1,n})$ and $V_{n,i} \coloneqq Y_i - R_n X_i$. The array of estimates $(\hat{\phi}_{n,k}(A_{n,k}V_{n,i}))_{n \in \mathbb{N}, i \leq n}$ satisfies

$$
\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} \left[\hat{\phi}_{k,n}(A_{n,k}V_{n,i}) - \phi_k(A_{n,k}V_{n,i}) \right] U_{n,i} &= o_{P_{\gamma_n'}}(n^{-1/2}) \\
\frac{1}{n} \sum_{i=1}^{n} \left(\left[\hat{\phi}_{n,k}(A_{n,k}V_{n,i}) - \phi_k(A_{n,k}V_{n,i}) \right] U_{n,i} \right)^2 &= o_{P_{\gamma_n'}}(\nu_n^2),
\end{aligned}
\tag{47}
$$

for any $(U_{n,i})_{n \in \mathbb{N}, i \leq n}$ such that for each $n \in \mathbb{N}$, under $P_{\gamma_n'}$, the $U_{n,i} \in L_2^0(P_{\gamma_n'})$, are i.i.d. with marginal distribution G_u and are independent of each $A_{n,k}V_{n,j}$, and where $0 \leq \nu_n \to 0$

satisfies $\mathsf{v}_n = o(\nu_n)$ with

$$\mathsf{v}_n := \begin{cases} n^{-1/2}\log(n)^{1/2+\rho} & \text{if } \delta \geq 4 \\ n^{(1-p)/(p)} & \text{otherwise} \end{cases}, \tag{48}$$

for $p := \min\{1 + \delta/4, 2\}$ and some $\rho > 0$. ◇

Lee and Mesters (2021a, Appendix B) propose an appropriate estimator of ϕ_k using cubic B-splines – based on the density score estimator of Chen and Bickel (2006) – and demonstrate that it satisfies assumption DSE under assumption LSEM and some mild additional restrictions on η.

Given such an estimator, $\hat{\phi}_{n,k}$, of each ϕ_k and a $\xi_n := (\theta_n, \beta_n)$, the efficient score functions in Proposition 5.2 can be estimated by replacing each $\phi_k(A_k v)$ with $\hat{\phi}_{n,k}(A_{n,k}V_{n,k})$ and replacing each τ_k, ς_k and μ by their sample counterparts:

$$\begin{aligned} \hat{\ell}_{\xi_n,1,l}(W_i) &:= \sum_{k=1}^{K}\left[\zeta_{l,k,k,n}\left(\hat{\tau}_{n,k,1}e_{n,k,i} + \hat{\tau}_{n,k,2}\kappa(e_{n,k,i})\right) + \sum_{j=1, j\neq k}^{K}\zeta_{l,k,j,n}\hat{\phi}_{n,k}(e_{n,k,i})e_{n,j,i}\right] \\ \hat{\ell}_{\xi_n,2,m}(W_i) &:= \sum_{k=1}^{K}\left[\zeta_{m,k,k,n}\left(\hat{\tau}_{n,k,1}e_{n,k,i} + \hat{\tau}_{n,k,2}\kappa(e_{n,k,i})\right) + \sum_{j=1, j\neq k}^{K}\zeta_{m,k,j,n}\hat{\phi}_{n,k}(e_{n,k,i})e_{n,j,i}\right] \\ \hat{\ell}_{\xi_n,2,d_{b_1}+s}(W_i) &:= \sum_{k=1}^{K}[-A_{n,k}D_{2,s}]\left[(X_i - \bar{X}_n)\hat{\phi}_{n,k}(e_{n,k,i}) - \bar{X}_n\left(\hat{\varsigma}_{n,k,1}e_{n,k,i} + \hat{\varsigma}_{n,k,2}\kappa(e_{n,k,i})\right)\right], \end{aligned} \tag{49}$$

where $e_{n,k,i} := A_{n,k}V_{n,i}$, $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ and

$$\hat{\tau}_{n,k} := \hat{M}_{n,k}^{-1}\begin{pmatrix} 0 \\ -2 \end{pmatrix}, \quad \hat{\varsigma}_{n,k} := \hat{M}_{n,k}^{-1}\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \text{with } \hat{M}_{n,k} := \begin{pmatrix} 1 & \frac{1}{n}\sum_{i=1}^{n} e_{n,k,i}^3 \\ \frac{1}{n}\sum_{i=1}^{n} e_{n,k,i}^3 & \frac{1}{n}\sum_{i=1}^{n} e_{n,k,i}^4 - 1 \end{pmatrix}.$$

In practice, β is unknown but estimates can be formed using a discretised version of an estimator for β which is $\sqrt{n}$-consistent under P_{γ_n}. In model (34), $\beta_2 = \operatorname{vec}(R)$ can be estimated by OLS. Appropriate estimators of $\sigma = \beta_1$ depend on the parametrisation of the matrix $A(\theta, \sigma)$ but can usually be constructed from the sample analogue of the equality $\mathbb{E}(VV') = A(\theta, \sigma)^{-1}(A(\theta, \sigma)^{-1})'$ for a given θ and estimate of R.[61]

Suppose $\hat{\beta}_n$ is a $\sqrt{n}$-consistent estimate of β and let $\bar{\beta}_n$ be the estimate which replaces $\hat{\beta}_n$ by the closest value in $n^{-1/2}C\mathbb{Z}^{d_\beta} \cap \mathscr{B}$.[62] Let $\bar{\xi}_n := (\theta_n, \bar{\beta}_n)$ and define the estimates

$$\begin{aligned} \hat{\ell}_{n,\theta_n} &:= \hat{\ell}_{\bar{\xi}_n,1} - \left[\mathbb{P}_n\hat{\ell}_{\bar{\xi}_n,1}\hat{\ell}'_{\bar{\xi}_n,2}\right]\left[\mathbb{P}_n\hat{\ell}_{\bar{\xi}_n,2}\hat{\ell}'_{\bar{\xi}_n,2}\right]^{-1}\hat{\ell}_{\bar{\xi}_n,2} \\ \check{\mathcal{I}}_{n,\theta_n} &:= \mathbb{P}_n\hat{\ell}_{\bar{\xi}_n,1}\hat{\ell}'_{\bar{\xi}_n,1} - \left[\mathbb{P}_n\hat{\ell}_{\bar{\xi}_n,1}\hat{\ell}'_{\bar{\xi}_n,2}\right]\left[\mathbb{P}_n\hat{\ell}_{\bar{\xi}_n,2}\hat{\ell}'_{\bar{\xi}_n,2}\right]^{-1}\left[\mathbb{P}_n\hat{\ell}_{\bar{\xi}_n,2}\hat{\ell}'_{\bar{\xi}_n,1}\right], \end{aligned} \tag{50}$$

[61] Such initial estimators can often be refined by one step updates, see e.g. §25.8 in van der Vaart (1998).
[62] For an abritrary constant $C > 0$.

and let $\hat{\mathcal{I}}_{n,\theta_n}$ be the eigendecomposition-truncated version of $\check{\mathcal{I}}_{n,\theta_n}$ at ν_n analogously to (21) (with ν_n as in assumption DSE).

5.3 Asymptotic properties

The following proposition demonstrates that the estimation procedure outlined in the previous subsection satisfies the conditions required for the theory in section 3 to apply.

Proposition 5.3. *Suppose that assumptions LSEM, DSE and equation (45) hold and that the observations form an i.i.d. sample. Consider the sequence $(\gamma_n)_{n\in\mathbb{N}}$ of Proposition 5.1. Suppose the inverse in (46) exists, $\theta \mapsto \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$ is locally constant at γ, $\hat{\beta}_n$ is a $\sqrt{n}$-consistent estimate for β under P_{γ_n} and $\hat{\ell}_{n,\theta_n}$, $\hat{\mathcal{I}}_{n,\theta_n}$ are as in equation (50). Then assumptions M, LAN, CM(ii), E and R hold.*[63]

The preceding proposition requires the rank of $\tilde{\mathcal{I}}_\gamma$ to be locally constant in θ at γ. This reflects the situation under study in which the identification status of θ is determined by η. Note that since the rank function is lower semi-continuous and non-negative integer valued, there is always a small enough neighbourhood on which the rank is bounded below by $\operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$. Therefore the force of the restriction is only that on some neighbourhood the rank cannot strictly exceed $\operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$, which is evidently the case for full rank $\tilde{\mathcal{I}}_\gamma$. For rank deficient $\tilde{\mathcal{I}}_\gamma$, the assumption has force.[64]

Given the definition of the efficient score and efficient information matrix estimators in (50) and supposing the hypothesis of interest is two-sided, the efficient score statistic and test can be defined as in equations (9) and (10). Since the required conditions have been established above, the results on size and power of the efficient score test – as established in section 3 – apply directly.

Corollary 5.4. *In the setting of proposition 5.3, let H_η' be a compact subset of H_η. Then the efficient score test satisfies*

$$\limsup_{n\to\infty} \sup_{h\in H_\eta'} P^n_{\gamma_n,0,h}\phi_{n,\theta_0} \leq \alpha,$$

and, for any compact $\Theta' \subset \Theta$, the corresponding test inversion confidence sets satisfy

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta'} \inf_{h\in H_\eta'} P^n_{(\theta,\eta),0,h}(\theta \in \hat{C}_n) \geq 1-\alpha.$$

Corollary 5.4 is the key results as regards robust inference in the presence of possible weak under- or un-identification of θ, as may occur when the components of η are sufficiently close

[63]Where the scores and paths in assumption LAN are as in proposition 5.1.

[64]From this discussion it is evident that an alternative way of stating this restriction would be that $\theta \mapsto \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$ is upper semi-continuous (or continuous) at γ.

to Gaussianity relative to the sample size. The results demonstrate that the efficient score has correct asymptotic size uniformly over local perturbations of the nuisance parameters and the corresponding (test inversion) confidence sets are uniformly valid over compact subsets of Θ and local perturbations of the nuisance parameters.

As the perturbation sets H_η are linear spaces whenever $\beta \in \operatorname{int} \mathscr{B}$, if this condition holds the efficient score test has optimality properties in the fully- and under- identified cases

Corollary 5.5. *In the setting of proposition 5.3 suppose additionally that β is an interior point of $\mathscr{B}$ and* $\operatorname{rank}(\tilde{\mathcal{I}}_\gamma) > 0$. *If $d_\theta = 1$, then the efficient score test is locally asymptotically uniformly most powerful unbiased. If $d_\theta > 1$, then the efficient score test is locally asymptotically maximin.*

I next examine the finite sample performance of the efficient score test in two explicit versions of the LSEM via two simulation studies. In the first study I consider a scalar parameter and focus on potential weak identification as may occur under error distributions close to Gaussianity. In the second I consider a two dimensional parameter which is underidentified under Gaussianity.

5.4 Simulation study (i)

Consider model (34), with $K = 2$, $L = 2$ and let the mixing matrix $A(\theta, \sigma)$ be

$$A(\theta,\sigma) = \begin{bmatrix} \sigma_2^{-1} & 0 \\ 0 & \sigma_3^{-1} \end{bmatrix} \begin{bmatrix} 1 & -\theta \\ -\sigma_1 & 1 \end{bmatrix}.$$

The null hypothesis under consideration is that $H_0 : \theta = 0$. When both ϵ_1 and ϵ_2 are close to Gaussianity, θ in this model will be only weakly identified.

To shed light on the finite sample performance of the efficient score test, I draw 5000 samples from this model for a range of different sample sizes and distributions for the error components ϵ_1 and ϵ_2. The $\tilde{X}$ variables are drawn as independent standard normals and $\beta_1 = \sigma = (0.7, 1.0, 3.0)$, $\beta_2 = \operatorname{vec}(R) = (1, 2, -1, -3/2)'$. Table 13 tabulates the considered error distributions for ϵ_1 and ϵ_2. 3 different distributions are considered for ϵ_1 and 10 for ϵ_2.[65] In particular, I consider a fixed distribution for ϵ_1 and examine the finite sample behaviour of the efficient score test as the distribution of ϵ_2 approaches Gaussianity, starting from 3 non-Gaussian distributions, each with a different shape.

To implement the efficient score test, I estimate each ϕ_k using the B-spline based estimator described in Appendix B of Lee and Mesters (2021a), which is adapted from a similar estimator proposed by Chen and Bickel (2006).[66] The remaining (Euclidean) nuisance parameters are estimated in two ways: (i) $\beta_2 = \operatorname{vec}(R)$ is estimated by OLS, with an estimate

[65]The density functions of these distributions are plotted in figures 1 - 3.

[66]In each simulation design, I use 6 cubic B-splines and set the upper and lower knots to be the 95th and

of β_1 recovered from the empirical variance matrix of the residuals $Y_i - \hat{R}X_i$. (ii) These OLS-based estimates are used to estimate the efficient score function for β, and then a one-step update is made based on this preliminary efficient score.[67]

With all the required nuisance parameters estimated, the efficient score function is constructed as in equation (50), the efficient score statistic is conducted as in equation (9) and the test performed as in equation (10) at a nominal level of 5%.[68]

The empirical rejection frequencies for the efficient score test conducted with (i) OLS-based estimates of the Euclidean nuisance parameters and (ii) one-step updates of these estimates are recorded in tables 14 - 16; each table corresponds to a different distribution for ϵ_1. The table of primary interest is table 14, with $\epsilon_1 \sim \mathcal{N}(0, 1)$ as this corresponds to a potentially weakly identified setting. As this table demonstrates, the efficient score test appears to demonstrate valid size control for all sample sizes and choices of η_2 considered. The version of the efficient score test with one-step updates provides reasonable size control, though demonstrates slight over-rejection in a number of cases. This finding holds also in each tables 15 - 16.

Tables 14 – 16 also contain size results for a number of alternative testing approaches. Two are Wald and LM tests based on a pseudo-maximum likelihood approach, inspired by the approach in Gouriéroux et al. (2017).[69] Here, a density is chosen for each of the error components and standard psuedo-maximum likelihood tests are performed. Following Gouriéroux et al. (2017) I choose a (normalised) $t(5)$ distribution for both ϵ_1 and ϵ_2 in this simulation experiment. As might be expected, the Wald statistic does not control size at the nominal level and displays both under- and over-rejection (depending on η_2) in table 14. Its performance in the settings recorded in tables 15 and 16 is mixed, demonstrating an ability to control size when at least one psuedo-density is sufficiently close to the truth, and substantial over-rejection otherwise. In contrast, the LM statistic (which imposes the null value of θ) does correctly control size for each choice of η_2 in tables 14 – 16.

The final two tests are Wald and LM tests based on a GMM framework in which higher moments of the error terms are used to provide identifying information. The moments used were drawn from Lanne and Luoto (2021).[70] Specifically, the (nine) moment conditions utilised are:

$$\mathbb{E}[\epsilon_1 \tilde{X}] = \mathbb{E}[\epsilon_2 \tilde{X}] = \mathbb{E}[\kappa(\epsilon_1)] = \mathbb{E}[\kappa(\epsilon_2)] = \mathbb{E}[\epsilon_1 \epsilon_2] = \mathbb{E}[\epsilon_1^3 \epsilon_2] = \mathbb{E}[\epsilon_1^2 \epsilon_2^2 - 1] = 0.$$

Neither of these GMM based tests (based on these moments) achieve finite sample size close

5th percentile of the samples, respectively adjusted up and down by $\log(\log n)$, truncated at the maximum (respectively minimum) sample value.

[67]I note that in the construction of the test θ is fixed throughout and so considered known.

[68]The information matrix eigenvalues are truncated at machine precision.

[69]Gouriéroux et al. (2017) consider a similar problem but in a SVAR setting.

[70]Like Gouriéroux et al. (2017), Lanne and Luoto (2021) consider a SVAR setting.

to nominal in the simulation experiments, as can be seen in tables 14 – 16. In the latter two tables, where weak identification is not present, the finite sample sizes of these tests appear to be reducing towards the nominal level as n increases, but remain substantially above the nominal level in each simulation design considered.

I perform a further simulation experiment based on this model to document the failure of size control of the score test based on the score functions for the Euclidean parameters $(\theta', \beta_1', \beta_2')'$. The relevant scores take the form

$$\dot{\ell}_{\gamma,l}(W) := \sum_{k=1}^{K}\left[\zeta_{l,k,k}(\phi_k(A_kV)A_k + 1) + \sum_{j=1, j\neq k}^{K} \zeta_{l,k,j}\phi_k(A_kV)A_jV\right]$$
$$\dot{\ell}_{\gamma,m}(W) := \sum_{k=1}^{K}[-A_kD_{b,l}X]\phi_k(A_kV),$$

for $l = 1, \dots, d_\theta, d_\theta+1, \dots, d_\theta + d_{\beta_1}$ and $m = d_\theta + d_{\beta_1} + 1, \dots, d_\theta + d_{\beta_1} + d_{\beta_2}$.[71] Let $\dot{\ell}^1_\gamma$ denote the first d_θ elements, and $\dot{\ell}^2_\gamma$ the remainder. Let $\dot{S}_{n,\theta}$ be the statistic formed analogously to (9) but based on an estimated version of $\dot{\ell}^1_\gamma - \dot{I}_{12}\dot{I}_{22}^{-1}\dot{\ell}^2_\gamma$, with $\dot{I}_\gamma = P_\gamma \dot{\ell}_\gamma \dot{\ell}_\gamma'$, rather than $\tilde{\ell}_\gamma$.

Since score functions have finite second moments,

$$\sqrt{n}\mathbb{P}_n\left[\dot{\ell}^1_\gamma - \dot{I}_{12}\dot{I}_{22}^{-1}\dot{\ell}^2_\gamma\right] \rightsquigarrow \mathcal{N}(0, \dot{I}_{\gamma,11} - \dot{I}_{\gamma,12}\dot{I}_{\gamma,22}^{-1}\dot{I}_{\gamma,21}),$$

and hence *if* $\dot{\ell}_\gamma$ and $\dot{I}_\gamma$ could be replaced by estimates with conditions analogous to those in assumption E and R holding, the test based on $\dot{S}_{n,\theta}$ would correctly control size.

Table 17 demonstrates that this is not the case, with the efficient score based tests controlling size, whilst the analogous tests based on $\dot{\ell}_\gamma$ (with the same estimator of ϕ_k) do not.[72] The key problem here is the bias caused by the regularised estimation of ϕ_k which is present in the estimate of $\dot{\ell}_\gamma$. This bias is removed by the orthogonal projection onto the nuisance score space in the definition of $\tilde{\ell}_\gamma$.

Following the size results, I compared the power of the two efficient score tests to that of the psuedo-ML based LM test which also was able to correctly control size in all designs considered. Figures 13 - 15 plot the results, corresponding to $\epsilon_1 \sim \{\mathcal{N}(0,1),\ t'(5),\ \mathcal{SN}'(0,1,4)\}$ respectively where t' and $\mathcal{SN}'$ denote the standardised version of the indicated distribution.

These finite sample power curves show that the power provided by any of the tests considered declines as the density η_2 approaches Gaussianity, particularly in the potentially weakly identified case where $\epsilon_1 \sim \mathcal{N}(0,1)$ (figure 13) in which available power appears low. In constrast, in figures 14 and 15 where there is no (weak) identification issue, the efficient

[71] Cf. proposition 5.1.

[72] In this simulation design, ϵ_1 and ϵ_2 have the same distribution, and are at a fixed distance from Gaussianity to focus on the problem of plugging in an estimate of a non-parametric parameter, rather than potential identification problems.

score tests apear to provide good finite sample power, with the version based on one-step updated estimates providing slightly higher power. The pseudo-maximum likelihood LM test also provides good power in cases where the chosen pseudo-densities are close to the truth. In particular, it slightly exceeds the power of the efficient score tests when ϵ_2 has a (standardised) t distribution in figures 13 and 14. Nevertheless, the efficient score test is competitive and provides close to identical power in the first row of figure 14, despite the pseudo density matching the truth in the first panel. Moreover, in cases where the psuedo-density is far from the truth, the power of the efficient score test is substantially higher than that provided by the pseudo-ML LM test (see, in particular, the third row of figure 14 and each row of figure 15).

5.5 Simulation study (ii)

In this second simulation study I consider the power available in a LSEM where the structural parameter of interest is underidentified. Specifically suppose that the data satisfies (34) where for $\theta = (a, b)$ with $a \neq b$ and $\beta_1 = (\sigma_1, \sigma_2) \in (0, \infty)^2$,

$$A(\theta, \beta_1) = \begin{bmatrix} \sigma_1^{-1} & 0 \\ 0 & \sigma_2^{-1} \end{bmatrix} \begin{bmatrix} 1 & -a \\ 1 & -b \end{bmatrix},$$

and there is one, zero-mean, unit variance X variable with coefficients $R = 0$. By explicit calculation, the efficient information matrix in this model takes the form

$$\tilde{\mathcal{I}}_\gamma = \frac{1}{(a-b)^2} \begin{bmatrix} \mathbb{E}[\phi_1(\epsilon_1)^2]c & -1 \\ -1 & \mathbb{E}[\phi_2(\epsilon_2)^2]c^{-1} \end{bmatrix}, \quad c := \left({}^{\sigma_2}/_{\sigma_1}\right)^2. \tag{51}$$

I consider three distributions from which to draw each ϵ_k: (i) $\mathcal{N}(0,1)$, (ii) $t'(5)$ - a (standardised) t distribution with 5 degrees of freedom and (iii) $st'(5,2)$ a (standardised) skew t distribution constructed as in Fernandez and Steel (1998) with 5 degrees of freedom and skewness parameter 2.[73] These correspond to (i) $\mathbb{E}[\phi_k(\epsilon_k)^2] = 1$, (ii) $\mathbb{E}[\phi_k(\epsilon_k)^2] = 1.25$ and (iii) $\mathbb{E}[\phi_k(\epsilon_k)^2] \approx 2.54$ respectively.

In the standard normal case (i), $\tilde{\mathcal{I}}_\gamma$ has eigenvalues $\lambda_1 = (c + c^{-1})/(a-b)^2$, $\lambda_2 = 0$ and a one-dimensional hyperplane as its nullspace: $N(\tilde{\mathcal{I}}_\gamma) = \{x \in \mathbb{R}^2 : cx_1 = x_2\}$. In cases (ii) and (iii), the matrix is positive definite and so $N(\tilde{\mathcal{I}}_\gamma) = \{0\}$.

Consider testing $\theta = \theta_0 = (a, b) = (1/2, 1/4)$, where $\sigma_1 = \sigma_2 = 1$ and hence the nullspace is the line $x_1 = x_2$. I take $n \in \{600, 1000, 1400\}$ and draw simulation samples according to (34) with $\theta = \theta_0 + \tau/\sqrt{n}$ and $X \sim \mathcal{N}(0, 1)$. β_2 is estimated by OLS and β_1 by GMM using the three moment conditions implied by the relationship $\mathbb{E}[VV'] =$

[73] The density functions of these distributions are plotted in figure 4.

$A(\theta, \beta_1)^{-1}(A(\theta, \beta_1)^{-1})'$. These estimates are used to construct estimates of the efficient score function and information matrix as in (50). In each case I truncate at machine precision.

The finite sample and asymptotic power surfaces are plotted in figures 16 - 18. Figure 16 demonstrates the expected trivial power along the hyperplane $N(\tilde{\mathcal{I}}_\gamma)$ in the Gaussian case, with power otherwise increasing in $\|\tau\|$. In contrast, figures 17 and 18 depict the full rank case, in which trivial power is found only at the point $\tau = 0$.[74] In all three figures, comparison of the finite sample power surface to the asymptotic power surface in the bottom right suggests that the asymptotic power results provide a good approximation to finite sample power.

6 Empirical study

In this section I use the LSEM of section 5 to analyse the relationship between hourly wages and hours worked, using non-Gaussianity in the data to identify the structural parameter of interest. There is a large literature on the estimation of labour supply equations, which takes note of many econometric challenges; see Blundell, MaCurdy, and Meghir (2007); Keane (2011) for detailed reviews. Two of the most notable difficulties include potential endogeneity and heterogeneity.[75]

A common labour supply specification is the semi-log formulation (e.g. equation (2.8) in Blundell et al., 2007):

$$H = \theta \log W + \gamma_1' X + \epsilon_1, \tag{52}$$

where H is hours of work, W the wage rate and X contains an intercept along with additional explanatory variables. As noted in e.g. Blundell et al. (2007), due to correlation with unobserved characteristics, wages are unlikely to be exogenous.[76] In order to take account of this potential endogeneity, I stack equation (52) with an equation for log wages of the form

$$\log W = \sigma_1 H + \gamma_2' X + \epsilon_2. \tag{53}$$

Once re-arranged, these two equations form a simultaneous system as in section 5:

$$Y = RX + V, \quad V = A(\theta, \sigma)^{-1}\epsilon, \tag{54}$$

where $Y = (H, \log W)$ and X collects (exogenous) covariates. I take the matrix A to be

[74] Which, of course, is exactly the nullspace of $\tilde{\mathcal{I}}_\gamma$ in this case.

[75] A further difficulty is the fact that individuals select into the labour force. I focus only on the intensive margin; a more realistic model would take account of potential selection biases which are ignored in the subsequent analysis.

[76] Blundell et al. (2007, p. 4676) write that “Wages may well be endogenous because unobservables affecting preferences for work may well be correlated with unobservables affecting productivity and hence wages”. Additionally, the “division bias” highlighted by Borjas (1980) provides another source of potential endogeneity.

parametrised as

$$A(\theta, \sigma) = D(\sigma_2, \sigma_3)^{-1} S(\theta, \sigma_1) = \begin{bmatrix} \sigma_2^{-1} & 0 \\ 0 & \sigma_3^{-1} \end{bmatrix} \begin{bmatrix} 1 & -\theta \\ -\sigma_1 & 1 \end{bmatrix}, \tag{55}$$

which permits the errors terms in V to be correlated with each other. Pre-multiplying Y by $S(\theta, \sigma)$ yields that θ is measure of the effect of a change in the log wage on hours.[77] The components $\beta = (\sigma', \operatorname{vec}(R)')'$, are estimable by OLS and from the variance of V, $\Sigma = A(\theta, \sigma)^{-1}[A(\theta, \sigma)^{-1}]'$.

Since the model is invariant to sign changes and column permutations of the A matrix, I choose the sign of the variance parameters σ_2, σ_3 to fix the column permutation and sign.[78] With these restrictions – and given a value of θ – the β parameters are estimable by standard techniques.[79] A confidence set for θ can then be constructed by inverting the efficient score test over a grid of possible values.

The data are taken from the CEPR uniform extracts of the (US) CPS outgoing rotation group (ORG) (Center for Economic and Policy Research, 2020). I select the analysis sample similarly to Bick, Blandin, and Rogerson (2021). In particular, I use the subset of the data between 2000 and 2007, restricted to males 25 – 64, who are employed (excluding the self-employed) with one job and at least 10 usual hours of work per week. Any observations with imputed values for hours or wages are dropped. This procedure leaves just under 200,000 pooled observations. As explanatory variables I include age and its square along with dummy variables for the year, education level (no high school, high school, some college, college degree or advanced degree), race (white, black, hispanic, other), whether the individual is married, whether they have children under the age of 18 in their household ("kids") and the interaction of kids with married.

I split the data between individuals who are paid hourly and those who are salaried. The former group consists of approximately 78,000 observations whilst the latter has approximately 120,000 observations. I construct confidence intervals for θ by inverting the efficient score test. Specifically, for each θ in a grid of 200 equally spaced points between -1.5 and 1.5, I calculate the efficient score statistic (as described in section 5) and form the confidence set consisting of those points θ for which the efficient score test does not reject at the 5% level.

This procedure yields confidence intervals of $[-0.069, 0.083]$ for salaried individuals and $[0.26, 0.43]$ for hourly paid individuals. These are plotted in figure 19; figure 20 plots the values of the efficient score statistic over the considered grid. The results suggest that the

[77] $\beta_{1,1}$ measures the effect of a change in hours on the log wage, but is not considered a parameter of interest in this exercise.

[78] If $\theta \geq 0$ I take $\sigma_2 \geq 0$; if $\theta < 0$ I take $\sigma_2 < 0$ and check that $\sigma_1 \geq 0$. σ_3 is always taken as positive.

[79] R can be estimated by OLS and σ from the system $S(\theta, \sigma_1)\Sigma S(\theta, \sigma_1)' = D(\sigma_2, \sigma_3)D(\sigma_2, \sigma_3)'$, with population moments replaced by sample equivalents based on the OLS residuals.

choice of hours by salaried individuals is less sensitive to changes in their hourly wage rate than for those who are paid by the hour. In particular, for the former group, the null hypothesis of no effect cannot be rejected at the 5% significance level.

Histograms of the residuals from (54) – with θ taken as the value which minimises the efficient score statistic – are plotted in figure 21 and overlaid with a $\mathcal{N}(0,1)$ density function. This figure is clearly suggestive of substantial non-Gaussianity in ϵ_1 (but not ϵ_2), providing identifying power for θ.[80]

The confidence intervals here suggest a similar qualitative response of labour supply to changes in wage rates as has been found previously in the literature, i.e. the effect of hourly wages on hours is small for men.[81]

7 Discussion

In this paper I demonstrated that score-type statistics based on the efficient score function can be used to perform uniformly valid inference in a wide class of models. A high level framework was provided in order to develop the theoretical results, based on the local asymptotic normality (LAN) framework of Le Cam.

The version of this framework considered permits many models and scenarios in which standard testing procedures fail to correctly control size, as demonstrated via specific examples. This class includes models which may suffer from identification problems, models where nuisance parameters may lie on the boundary of the parameter space and models which need a regularisation step for their estimation. I demonstrated that the efficient score test enjoys locally uniformly valid size control. Moreover, I showed that a number of standard testing optimality results continue to hold in this setup and demonstrated a minimax optimality result which applies in cases where, for example, the parameter of interest is underidentified.

A number of examples were studied in detail to demonstrate the applicability of the suggested framework and how the conditions it requires may be shown to hold. Simulation studies based on these examples suggest that the asymptotic results obtained provide a useful guide to finite sample performance. The simulations also show that – in the cases considered – the procedures based on the efficient score statistic perform better than alternative procedures. I applied the linear simultaneous equations model example to the study of the labour supply decision of US men. This approach permits the study of simultaneous systems without interventions or instruments.

[80] A Jarque and Bera (1980) test rejects the null that ϵ_1 is Gaussian at all standard significance levels.

[81] A coefficient of e.g. $\theta = 0.3$ in the semi-log specification of (52) implies that, for instance, a 50% increase in the wage rate would increase hours worked by $0.3 \times \log(1.5) \approx 0.12$. For summaries of elasticities found in the literature, see e.g. Table 3 & Figure 2 of Bargain and Peichl (2016) and Table 3.1 & Figure 3.2 of Evers, de Mooij, and van Vuuren (2005).

The treatment in the current paper is restricted to cases where the observed data forms a random sample. This restriction was made to remove inessential complications in the derivation of the results. With these now established in the baseline i.i.d. case, an interesting potential extension would be to extend these results to other sampling schemes. An additional drawback of the current treatment is that the parameter of interest θ is required to be a bona fide parameter of the model as opposed to a function of the model parameters. An extension to permit this scenario could be provided along the lines of Susyanto and Klaassen (2017). Such extensions are left for future work.

References

Amari, S.-I. and Cardoso, J.-F. (1997), "Blind source separation-semiparametric statistical approach," *IEEE Transactions on Signal Processing*, 45, 2692–2700.

Amengual, D., Fiorentini, G., and Sentana, E. (2021), "Moment tests of independent components," .

Anderson, T. W. and Rubin, H. (1949), "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *The Annals of Mathematical Statistics*, 20, 46 – 63.

Andrews, D. W., Cheng, X., and Guggenberger, P. (2020), "Generic results for establishing the asymptotic size of confidence sets and tests," *Journal of Econometrics*, 218, 496–531.

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

— (2000), "Inconsistency of the Bootstrap When a Parameter is on the Boundary of the Parameter Space," *Econometrica*, 68, 399–405.

— (2001), "Testing When a Parameter is on the Boundary of the Maintained Hypothesis," *Econometrica*, 69, 683–734.

Andrews, D. W. K. and Cheng, X. (2012), "Estimation and Inference With Weak, Semi-Strong, and Strong Identification," *Econometrica*, 80, 2153–2211.

— (2013), "Maximum likelihood estimation and uniform inference with sporadic identification failure," *Journal of Econometrics*, 173, 36–56.

Andrews, D. W. K. and Guggenberger, P. (2009), "Hybrid and Size-Corrected Subsampling Methods," *Econometrica*, 77, 721–762.

— (2010a), "Applications of subsampling, hybrid, and size-correction methods," *Journal of Econometrics*, 158, 285–305.

— (2010b), "Asymptotic Size and a Problem with Subsampling and with the m out of n Bootstrap," *Econometric Theory*, 26, 426–468.

— (2019), "Identification- and singularity-robust inference for moment condition models," *Quantitative Economics*, 10, 1703–1746.

Andrews, I. and Mikusheva, A. (2015), "Maximum likelihood inference in weakly identified dynamic stochastic general equilibrium models," *Quantitative Economics*, 6, 123–152.

— (2016a), "Conditional Inference With a Functional Nuisance Parameter," *Econometrica*, 84, 1571–1612.

— (2016b), "A Geometric Approach to Nonlinear Econometric Models," *Econometrica*, 84, 1249–1264.

— (forthcoming), "Optimal Decision Rules for Weak GMM," *Econometrica*.

Andrews, I., Stock, J. H., and Sun, L. (2019), "Weak Instruments in Instrumental Variables Regression: Theory and Practice," *Annual Review of Economics*, 11, 727–753.

Antoine, B. and Renault, E. (2009), "Efficient GMM with nearly-weak instruments," *The Econometrics Journal*, 12, S135–S171.

— (2011), "Efficient Inference with Poor Instruments: A General Framework," in *Handbook of Empirical Economics and Finance*, eds. Ullah, A. and Giles, D. E. A., CRC Press.

Bargain, O. and Peichl, A. (2016), "Own-wage labor supply elasticities: variation across time and estimation methods," *IZA Journal of Labor Economics*, 5, 1–31.

Bekaert, G., Engstrom, E., and Ermolov, A. (2019), "Macro Risks and the Term Structure of Interest Rates," *Working paper*.

— (2020), "Aggregate Demand and Aggregate Supply Effects of COVID-19: A Real-time Analysis," *Working paper*.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015), "Some new asymptotic theory for least squares series: Pointwise and uniform results," *Journal of Econometrics*, 186, 345–366.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017), "Program Evaluation and Causal Inference With High-Dimensional Data," *Econometrica*, 85, 233–298.

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

Ben-Moshe, D. (2020), "Identification of linear regressions with errors in all variables," *Econometric Theory*, 1–31.

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

Bick, A., Blandin, A., and Rogerson, R. (2021), "Hours and Wages," Working Paper.

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

Bickel, P. J. and Ritov, Y. (1987), "Efficient Estimation in the Errors in Variables Model," *The Annals of Statistics*, 15, 513 – 540.

Billingsley, P. (1995), *Probability and Measure*, Wiley.

— (1999), *Convergence of Probability Measures*, Wiley.

Blundell, R., MaCurdy, T., and Meghir, C. (2007), "Chapter 69 Labor Supply Models: Unobserved Heterogeneity, Nonparticipation and Dynamics," Elsevier, vol. 6 of *Handbook of Econometrics*, pp. 4667–4775.

Bogachev, V. I. (2007), *Measure Theory*, Springer Berlin Heidelberg.

Borjas, G. (1980), "The Relationship between Wages and Weekly Hours of Work: The Role of Division Bias," *Journal of Human Resources*, 15, 409–423.

Cattaneo, M. D., Farrell, M. H., and Feng, Y. (2020), "Large sample properties of partitioning-based series estimators," *The Annals of Statistics*, 48, 1718 – 1741.

Cavaliere, G., Nielsen, H. B., Pedersen, R. S., and Rahbek, A. (2020), "Bootstrap inference on the boundary of the parameter space, with application to conditional volatility models," *Journal of Econometrics*.

Center for Economic and Policy Research (2020), "CPS ORG Uniform Extracts, Version 2.5," .

Chaudhuri, S. and Zivot, E. (2011), "A new method of projection-based inference in GMM with weakly identified nuisance parameters," *Journal of Econometrics*, 164, 239–251.

Chen, A. and Bickel, P. J. (2006), "Efficient independent component analysis," *The Annals of Statistics*, 34, 2825 – 2855.

Chen, X. and Christensen, T. M. (2015), "Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions," *Journal of Econometrics*, 188, 447–465.

Chen, X., Christensen, T. M., and Tamer, E. (2018), "Monte Carlo Confidence Sets for Identified Sets," *Econometrica*, 86, 1965–2018.

Chen, Y., Ning, J., Ning, Y., Liang, K.-Y., and Bandeen-Roche, K. (2017), "On pseudolikelihood inference for semiparametric models with boundary problems," *Biometrika*, 104, 165–179.

Chernoff, H. (1954), "On the distribution of the likelihood ratio," *Annals of Mathematical Statistics*, 25, 573–578.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. K., and Robins, J. (2018), "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21, C1–C68.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015), "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach," *Annual Review of Economics*, 7, 649–688.

Chetverikov, D., Santos, A., and Shaikh, A. M. (2018), "The Econometrics of Shape Restrictions," *Annual Review of Economics*, 10, 31–63.

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

Comon, P. (1994), "Independent component analysis, A new concept?" *Signal Processing*, 36, 287–314.

Conway, J. B. (1985), *A course in functional analysis*, New York, NY, USA: Springer.

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

Davis, R. and Ng, S. (2021), "Time Series Estimation of the Dynamic Effects of Disaster-Type Shock," ArXiv:2107.06663.

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

Drabek, P. and Milota, J. (2007), *Methods of Nonlinear Analysis: Applications to Differential Equations*, Birkhäuser Advanced Texts Basler Lehrbücher, Birkhäuser Basel.

Dufour, J.-M. (1997), "Some Impossibility Theorems in Econometrics With Applications to Structural and Dynamic Models," *Econometrica*, 65, 1365–1387.

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

Elliott, G., Müller, U. K., and Watson, M. W. (2015), "Nearly Optimal Tests When a Nuisance Parameter Is Present Under the Null Hypothesis," *Econometrica*, 83, 771–811.

Evers, M., de Mooij, R. A., and van Vuuren, D. J. (2005), "What Explains the Variation in Estimates of Labour Supply Elasticities?" CESifo Working Paper Series 1633, CESifo.

Feinberg, E. A., Kasyanov, P. O., and Zgurovsky, M. Z. (2016), "Uniform Fatou's lemma," *Journal of Mathematical Analysis and Applications*, 444, 550–567.

Fernandez, C. and Steel, M. F. J. (1998), "On Bayesian Modeling of Fat Tails and Skewness," *Journal of the American Statistical Association*, 93, 359–371.

Fiorentini, G. and Sentana, E. (2020), "Specification tests for non-Gaussian maximum likelihood estimators," Working paper.

— (2021), "Discrete Mixtures of Normals Pseudo Maximum Likelihood Estimators of Structural Vector Autoregressions," Working paper.

Geyer, C. J. (1994), "On the asymptotics of constrained M-estimation," *Annals of Statistics*, 22, 1993–2010.

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

— (2019), "Identification and Estimation in Non-Fundamental Structural VARMA Models," *The Review of Economic Studies*, 87, 1915–1953.

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

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

Han, S. and McCloskey, A. (2019), "Estimation and inference with a (nearly) singular Jacobian," *Quantitative Economics*, 10, 1019–1068.

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

Huang, J. Z. and Su, Y. (2021), "Asymptotic Properties of Penalized Spline Estimators in Concave Extended Linear Models: Rates of Convergence," ArXiv:2105.06367.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001), *Independent Component Analysis*, Wiley, New York.

Ichimura, H. (1993), "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models," *Journal of Econometrics*, 58, 71–120.

Jansson, M. (2008), "Semiparametric Power Envelopes for Tests of the Unit Root Hypothesis," *Econometrica*, 76, 1103–1142.

Jarque, C. M. and Bera, A. K. (1980), "Efficient tests for normality, homoscedasticity and serial independence of regression residuals," *Economics Letters*, 6, 255–259.

Jeganathan, P. (1995), "Some Aspects of Asymptotic Theory with Applications to Time Series Models," *Econometric Theory*, 11, 818–887.

Kaji, T. (2021), "Theory of Weak Identification in Semiparametric Models," *Econometrica*, 89, 733–763.

Keane, M. P. (2011), "Labor Supply and Taxes: A Survey," *Journal of Economic Literature*, 49, 961–1075.

Ketz, P. (2018), "Subvector inference when the true parameter vector may be near or at the boundary," *Journal of Econometrics*, 207, 285–306.

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

— (2005), "Testing Parameters in GMM Without Assuming that They Are Identified," *Econometrica*, 73, 1103–1123.

— (2007), "Generalizing weak instrument robust IV statistics towards multiple parameters, unrestricted covariance matrices and identification statistics," *Journal of Econometrics*, 139, 181–216.

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

Kuchibhotla, A. K. and Patra, R. K. (2020), "Efficient estimation in single index models through smoothing splines," *Bernoulli*, 26, 1587 – 1618.

Lanne, M. and Luoto, J. (2021), "GMM Estimation of Non-Gaussian Structural Vector Autoregression," *Journal of Business & Economic Statistics*, 39, 69–81.

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

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

Le Cam, L. M. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer Series in Statistics, New York, NY, USA: Springer.

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

Lee, A. and Mesters, G. (2021a), "Robust Non-Gaussian Inference for Linear Simultaneous Equations Models," Working Paper.

— (2021b), "Supplement to "Robust Non-Gaussian Inference for Linear Simultaneous Equations Models"," Working Paper.

Leeb, H. and Pötscher, B. M. (2005), "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21–59.

Lehmann, E. L. and Romano, J. P. (2005), *Testing Statistical Hypotheses*, New York, NY, USA: Springer, 3rd ed.

Lütkepohl, H. and Burda, M. M. (1997), "Modified wald tests under nonregular conditions," *Journal of Econometrics*, 78, 315–332.

Ma, Y. and Zhu, L. (2013), "Doubly robust and efficient estimators for heteroscedastic partially linear single-index models allowing high dimensional covariates," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75, 305–322.

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

Marron, J. S. and Wand, M. P. (1992), "Exact Mean Integrated Squared Error," *The Annals of Statistics*, 20, 712 – 736.

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

McCloskey, A. (2017), "Bonferroni-based size-correction for nonstandard testing problems," *Journal of Econometrics*, 200, 17–35.

— (2020), "Asymptotically Uniform Tests After Consistent Model Selection in the Linear Regression Model," *Journal of Business & Economic Statistics*, 38, 810–825.

Montiel Olea, J. L. (2020), "Admissible, Similar Tests: A Characterization," *Econometric Theory*, 36, 347–366.

Montiel Olea, J. L., Plagborg-Møller, M., and Qian, E. (2021), "SVAR Identification From Higher Moments: Has the Simultaneous Causality Problem Been Solved?" *Working paper*, prepared for AEA Papers and Proceedings.

Moreira, M. J. (2003), "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71, 1027–1048.

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

— (1991), "Estimation of Tobit models under conditional symmetry," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics*, eds. Barnett, W. A., Powell, J., and Tauchen, G. E., Cambridge University Press, International Symposia in Economic Theory and Econometrics.

— (1997), "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79, 147–168.

Newey, W. K. and Stoker, T. M. (1993), "Efficiency of Weighted Average Derivative Estimators and Index Models," *Econometrica*, 61, 1199–1223.

Neyman, J. (1959), "Optimal Asymptotic Tests of Composite Statistical Hypotheses," in *Probability and Statistics, the Harald Cramér Volume*, ed. Grenander, U., New York, USA: Wiley.

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

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

Reiersøl, O. (1950), "Identifiability of a Linear Relation between Variables Which Are Subject to Error," *Econometrica*, 18, 375–389.

Remmert, R. (1991), *Theory of Complex Functions*, Springer.

Rieder, H. (2014), "One-Sided Confidence About Functionals Over Tangent Cones," ArXiv:1412.1701.

Romano, J. P. and Shaikh, A. M. (2012), "On the uniform asymptotic validity of subsampling and the bootstrap," *The Annals of Statistics*, 40, 2798 – 2822.

Royden, H. and Fitzpatrick, P. (2010), *Real Analysis*, Pearson Prentice Hall.

Rudin, W. (1991), *Functional analysis*, McGraw Hill, Inc., 2nd ed.

Staiger, D. and Stock, J. H. (1997), “Instrumental Variables Regression with Weak Instruments,” *Econometrica*, 65, 557–586.

Stock, J. H. and Wright, J. H. (2000), “GMM with Weak Identification,” *Econometrica*, 68, 1055–1096.

Strasser, H. (1985), *Mathematical Theory of Statistics: Statistical Experiments and Asymptotic Decision Theory*, De Gruyter studies in mathematics, W. de Gruyter.

Susyanto, N. and Klaassen, C. A. J. (2017), “Semiparametrically efficient estimation of constrained Euclidean parameters,” *Electronic Journal of Statistics*, 11, 3120 – 3140.

van den Berg, G. J. (2001), “Chapter 55 - Duration Models: Specification, Identification and Multiple Durations,” Elsevier, vol. 5 of *Handbook of Econometrics*, pp. 3381–3460.

van der Vaart, A. W. (1988), *Statistical Estimation in Large Parameter Spaces*, no. 44 in CWI Tracts, Amsterdam: Centrum voor Wiskunde en Informatica.

— (1991), “An Asymptotic Representation Theorem,” *International Statistical Review / Revue Internationale de Statistique*, 59, 97–121.

— (1998), *Asymptotic Statistics*, New York, NY, USA: Cambridge University Press.

— (2002), “Semiparametric Statistics,” in *Lectures on Probability Theory and Statistics: Ecole d'Eté de Probabilités de Saint-Flour XXIX - 1999*, ed. Bernard, P., Springer.

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

Velasco, C. (2020), “Identification and estimation of Structural VARMA models using higher order dynamics,” ArXiv:2009.04428.

White, H. (1980), “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity,” *Econometrica*, 48, 817–838.

Willassen, Y. (1979), “Extension of Some Results by Reiersøl to Multivariate Models,” *Scandinavian Journal of Statistics*, 6, 89–91.

A Notation & conventions

$A \coloneqq B$ means that A is defined to be B. $A \subset B$ indicates that A is a subset of B. All vector spaces are over the real field $\mathbb{R}$. Given a positive integer K, $[K] \coloneqq \{1, \ldots, K\}$. For any Euclidean parameter, say α, d_α denotes the dimension of the space in which it lives. Similarly for a vector of functions κ, d_κ is the number of component functions. For a sequence $(x_n)_{n\in\mathbb{N}}$, $(x_n)_{n\in\mathbb{N}} \subset \mathcal{X}$ denotes that each $x_n \in \mathcal{X}$. For any matrix M, $\|M\|_2$ is its spectral norm and $M^\dagger$ is its Moore-Penrose pseudo-inverse. "$\succeq$" is used to denote the Loewner partial order; that is, given two Hermitian matrices A, B, $A \succeq B$ iff $A - B$ is positive semi-definite and $A \succ B$ iff $A - B$ is positive definite. If A is a linear operator, $N(A)$ is its nullspace. Given a topological space S, $\mathcal{B}(S)$ is its Borel σ-algebra. Weak convergence is denoted by "$\rightsquigarrow$". Operator notation is often used for integrals: $Pf \coloneqq \int f \,\mathrm{d}P$. $\mathbb{P}_n$ denotes the empirical measure of a given sample and $\mathbb{G}_n$ the empirical process. Throughout this paper & unless otherwise noted the sample considered is denoted by $(W_i)_{i=1}^n \in \mathcal{W}^n$, hence $\mathbb{P}_n f = \int f \,\mathrm{d}\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n f(W_i)$. For a sequence of functions $(f_n)_{n\in\mathbb{N}}$ with each f_n having domain $\mathcal{W}^n$ and a sequence of probability measures $(P_n)_{n\in\mathbb{N}}$ on $\mathcal{W}$, convergence statements will often be written as $f_n \rightsquigarrow f$ under P_n. This is shorthand for weak convergence under the product measures P_n^n. If X has distribution G, I write $X \sim G$. If g is the density of G (with respect to some σ-finite measure), I also write $X \sim g$. $X \simeq Y$ indicates that X and Y have the same distribution. $L_p(P)$ denotes the space of functions f such that $P|f|^p < \infty$. In the case where $f = (f_1, \ldots, f_K)$ is a vector of functions $f \in L_p(P)$ denotes that each $f_i \in L_p(P)$ for $i = 1, \ldots, K$. $L_p^0(P)$ is the subspace of $L_p(P)$ whose members f satisfy $Pf = 0$. Given a (closed) subspace S of a Hilbert space H, the orthogonal projection of a function $f \in H$ onto S is denoted by $\Pi(f|S)$.

B Additional details and proofs of results in the main text

B.1 Details and proofs for section 3

B.1.1 Construction of the sequence of experiments

In order to discuss power I use the limits of experiments framework of Le Cam (see e.g. chapter 9 of van der Vaart (1998) for an introduction). Under the additional assumption that $\mathscr{H}_\gamma$ is a linear space, I will obtain a Gaussian shift limit experiment on a particular inner-product space.[82]

To state the proposition, I need to define the inner-product space that will be used to parametrise the experiments. Let $N(A)$ denote the null space of a linear transformation A; in particular $N(\tilde{\mathcal{I}}_\gamma)$ denotes the null space of the matrix $\tilde{\mathcal{I}}_\gamma$. For the nuisance perturbations, h, it is more convenient to parametrise directly by the scores $g = B_\gamma h$. For each $g = B_\gamma h \in \mathscr{H}_\gamma$ let $\mathfrak{h}_{g,\eta} \coloneqq \{h \in H_\eta : B_\gamma h = g\}$. Suppose that $\mathscr{H}_\gamma$ is a linear subspace of $L_2(P_\gamma)$ and note that it is therefore a dense subspace of a its completion (which is a Hilbert space). It therefore has an orthonormal basis, $(g_k)_{k\in\mathbb{N}}$.[83] For each element g_k in this basis select (arbitrarily) an

[82]That is, the limit experiment is the restriction of a Gaussian shift experiment on a specific Hilbert space to the inner-product space of interest. See e.g. Le Cam (1986, Chapter 9, section 3) or Strasser (1985, Chapter 11) for an introduction to Gaussian shift experiments on Hilbert spaces.

[83]See footnote 93.

element $h_k = h_{g_k}$ from each $\mathfrak{h}_{g_k,\eta}$. For any other element $g \in \mathscr{H}_\gamma$ choose $h_g = \sum_{k\in\mathbb{N}} a_k h_k$ where $g = \sum_{k\in\mathbb{N}} a_k g_k$. Denote the collection of such h_g as $\mathfrak{H}_\gamma := \{h_g : g \in \mathscr{H}_\gamma\} \subset H_\eta$.[84] I will consider sequences of experiments, where each consists of measures of the form $P_{\gamma_n,\tau,g} = P_{\gamma_n,\tau,h}$ for $\tau \in N(\tilde{\mathcal{I}}_\gamma)^\perp$ and $g \in \mathscr{H}_\gamma$, $h = h_g \in \mathfrak{H}_\gamma$ (with $\gamma = \lim_{n\to\infty}\gamma_n$); that is to say, these experiments are parametrised by the (inner-product) space $\mathbb{H}_\gamma := N(\tilde{\mathcal{I}}_\gamma)^\perp \times \mathscr{H}_\gamma$ equipped with the inner-product given below in (56).

The choice of a particular "representative" $h = h_g$ for each score $g = B_\gamma h \in \mathscr{H}_\gamma$ as in the preceding construction is a technical point which will not impede statements being made about the behaviour of tests along sequences with $h_n \to h \in H_\eta \setminus \mathfrak{H}_\gamma$ due to the following lemma.

Lemma B.1. *Suppose that assumptions M, LAN, CM(i) hold and that $(\psi_n)_{n\in\mathbb{N}}$ is a sequence of tests on $\mathcal{W}^n$ (i.e. each $\psi_n : \mathcal{W}^n \to [0,1]$).*

(i) If $(\tau_n)_{n\in\mathbb{N}} \subset \mathbb{R}^{d_\theta}$ and $(h_n)_{n\in\mathbb{N}} \subset H_\eta$ are convergent sequences with limits $\tau \in \mathbb{R}^{d_\theta}$ and $h \in H_\eta$ respectively, then

$$\limsup_{n\to\infty} \left[P^n_{\gamma_n,\tau_n,h_n}\psi_n - P^n_{\gamma_n,\tau,h}\psi_n\right] = 0.$$

(ii) If $h_1, h_2 \in H_\eta$ are such that $B_\gamma h_1 = B_\gamma h_2$ and $h_1 - h_2 \in H_\eta$, then for any convergent sequences $(\tau_n)_{n\in\mathbb{N}} \subset \mathbb{R}^{d_\theta}$, $(h_{1,n})_{n\in\mathbb{N}} \subset H_\eta$, $(h_{2,n})_{n\in\mathbb{N}} \subset H_\eta$ with limits $\tau \in \mathbb{R}^{d_\theta}$ and $h_1, h_2 \in H_\eta$ respectively,

$$\Lambda_n(\gamma_n(\tau_n, h_{1,n}), \gamma_n) - \Lambda_n(\gamma_n(\tau_n, h_{2,n}), \gamma_n) = o_{P_{\gamma_n}}(1),$$

and

$$\limsup_{n\to\infty} \left[P^n_{\gamma_n,\tau_n,h_{1,n}}\psi_n - P^n_{\gamma_n,\tau_n,h_{2,n}}\psi_n\right] = 0.$$

With the setup previously described the following result concerning convergence of experiments can be stated. This result is straightforward given the assumptions made, and is quite standard, aside from potentially one key aspect: the definition of the indexing set of the sequence of experiments — that $\tau \in N(\tilde{\mathcal{I}}_\gamma)^\perp$. This ensures that the inner-product in equation (56) *is* an inner-product. If $N(\tilde{\mathcal{I}}_\gamma)^\perp$ was replaced by $\mathbb{R}^{d_\theta}$ and $\operatorname{rank}(\tilde{\mathcal{I}}_\gamma) < d_\theta$, the map in (56) would only be a positive-semidefinite Hermitian form.[85]

Proposition B.2. *Suppose that assumptions M, LAN and CM(i) hold and that $\mathscr{H}_\gamma$ is a linear subspace of $L_2(P_\gamma)$. Suppose that $\operatorname{rank}(\tilde{\mathcal{I}}_\gamma) > 0$ and let $\mathbb{H}_\gamma := N(\tilde{\mathcal{I}}_\gamma)^\perp \times \mathscr{H}_\gamma$. If the map $\langle\cdot,\cdot\rangle_{\mathbb{H}_\gamma} : \mathbb{H}_\gamma \times \mathbb{H}_\gamma \to \mathbb{R}$ is defined by*

$$\langle(\tau_1, g_1), (\tau_2, g_2)\rangle := \langle \tau_1'\dot{\ell}_\gamma + g_1, \tau_2'\dot{\ell}_\gamma + g_2\rangle_{P_\gamma}, \tag{56}$$

then $(\mathbb{H}_\gamma, \langle\cdot,\cdot\rangle)$ is an inner-product space. In addition, the sequence of experiments $(\mathscr{E}_n)_{n\in\mathbb{N}}$, where each

$$\mathscr{E}_n := \left(\mathcal{W}^n, \mathcal{B}(\mathcal{W}^n), \left\{P^n_{\gamma_n,\tau,g} : (\tau, g) \in \mathbb{H}_\gamma\right\}\right), \tag{57}$$

[84] I will suppose that the $h_g = h_0$ chosen to correspond to $g = 0$ is $h_g = h_0 = 0$. Note that if B_γ is injective there is only one such h_g for each $g \in \mathscr{H}_\gamma$.

[85] That is, $\langle(\tau, g), (\tau, g)\rangle = 0$ whilst $(\tau, g) \neq 0$ would be possible. In particular, $\langle(\tau, 0), (\tau, 0)\rangle = 0$ would hold for all $\tau \in N(\tilde{\mathcal{I}}_\gamma)$, which has positive dimension whenever $\operatorname{rank}(\tilde{\mathcal{I}}_\gamma) < d_\theta$.

converges weakly to a Gaussian shift on $(\mathbb{H}_\gamma, \langle\cdot,\cdot\rangle)$.

B.1.2 Proofs

Proof of proposition 3.1. To simplify the notation, let $g_n := \tau'\dot{\ell}_{\gamma_n} + B_{\gamma_n}h$ and $g := \tau'\dot{\ell}_\gamma + B_\gamma h$. Let $\{W_{n,k} : k \leq n, n \in \mathbb{N}\}$ be a triangular array, where each row $W_{n,1}, \ldots, W_{n,n}$ $(n \in \mathbb{N})$ is independently and identically distributed, with each random vector $W_{n,k}$ having law P_{γ_n}. Let $\{Z_{n,k} : k \leq n, n \in \mathbb{N}\}$ be the array defined by $Z_{n,k} := \left(\tilde{\ell}_{\gamma_n}(W_{n,k})', g_n(W_{n,k})\right)'$. The rows of this array are i.i.d. with $\mathbb{E}Z_{n,k} = 0$ and $\mathbb{V}Z_{n,k} = \begin{bmatrix} \tilde{\mathcal{I}}_{\gamma_n} & \tilde{\mathcal{I}}_{\gamma_n}\tau \\ \tau'\tilde{\mathcal{I}}_{\gamma_n} & P_{\gamma_n}g_n^2 \end{bmatrix}$ (for each k, n). [86] By assumption CM(ii)

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{V}Z_{n,k} = \mathbb{V}Z_{n,1} \to \begin{bmatrix} \tilde{\mathcal{I}}_\gamma & \tilde{\mathcal{I}}_\gamma\tau \\ \tau'\tilde{\mathcal{I}}_\gamma & \sigma^2_{\tau,h} \end{bmatrix}, \tag{58}$$

where

$$\sigma^2_{\tau,h} := P_\gamma g^2 = P_\gamma[\tau'\dot{\ell}_\gamma + B_\gamma h]^2 = \lim_{n\to\infty} P_{\gamma_n}\left[\tau'\dot{\ell}_{\gamma_n} + B_{\gamma_n}h\right]^2 = \lim_{n\to\infty} P_{\gamma_n}g_n^2, \tag{59}$$

and hence (101) is satisfied. Moreover assumptions LAN and CM(ii) together yield that $(\|Z_{n,1}\|_2^2)_{n\in\mathbb{N}}$ is uniformly integrable and hence as the rows are identically distributed, (102) holds. It then follows by lemma C.1 that under P_{γ_n} we have

$$\sqrt{n}\mathbb{P}_n\left(\tilde{\ell}'_{\gamma_n}, \tau'\dot{\ell}_{\gamma_n} + B_{\gamma_n}h\right)' \rightsquigarrow \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tilde{\mathcal{I}}_\gamma & \tilde{\mathcal{I}}_\gamma\tau \\ \tau'\tilde{\mathcal{I}}_\gamma & \sigma^2_{\tau,h} \end{pmatrix}\right). \tag{60}$$

Combining equations (5), (58), (59) and (60) we have

$$\left(\sqrt{n}\mathbb{P}_n\tilde{\ell}'_{\gamma_n},\ \Lambda_n(\gamma_n(\tau_n, h_n), \gamma_n)\right)' \rightsquigarrow \mathcal{N}\left(\begin{pmatrix} 0 \\ -\frac{1}{2}\sigma^2_{\tau,h} \end{pmatrix}, \begin{pmatrix} \tilde{\mathcal{I}}_\gamma & \tilde{\mathcal{I}}_\gamma\tau \\ \tau'\tilde{\mathcal{I}}_\gamma & \sigma^2_{\tau,h} \end{pmatrix}\right). \tag{61}$$

The marginal convergence of the likelihood ratio yields that $(P^n_{\gamma_n})_{n\in\mathbb{N}}$ and $(P^n_{\gamma_n,\tau_n,h_n})_{n\in\mathbb{N}}$ are mutually contiguous (e.g. van der Vaart and Wellner, 1996, Example 3.10.6). We remark here that a completely analogous argument to the foregoing applied to the array $\{g_n(W_{n,k}) : k \leq n, n \in \mathbb{N}\}$ yields this same marginal convergence under assumption CM(i) rather than assumption CM(ii) and hence the mutual contiguity of these sequences of measures continues to hold under this weaker condition, as claimed in the statement of the proposition.

By Le Cam's third lemma (e.g. van der Vaart and Wellner, 1996, Example 3.10.8) it follows from (61) that under P_{γ_n,τ_n,h_n}

$$\sqrt{n}\mathbb{P}_n\tilde{\ell}_{\gamma_n} \rightsquigarrow \mathcal{N}(\tilde{\mathcal{I}}_\gamma\tau, \tilde{\mathcal{I}}_\gamma).$$

Equation (6), the mutual contiguity and Le Cam's first lemma (e.g. van der Vaart, 1998, Lemma 6.4) allow us to conclude that

$$\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\theta_n} - \tilde{\ell}_{\gamma_n}\right] = o_{P_{\gamma_n,\tau_n,h_n}}(1)$$

[86] We have that $P_{\gamma_n}\tilde{\ell}_{\gamma_n}\dot{\ell}'_{\gamma_n} = \tilde{\mathcal{I}}_{\gamma_n}$ (e.g. Rudin, 1991, Theorem 12.14). $P_{\gamma_n}\tilde{\ell}_{\gamma_n}[B_{\gamma_n}h] = 0$ by the construction of the efficient score function.

It follows that under P_{γ_n,τ_n,h_n}

$$\sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta_n} = \sqrt{n}\mathbb{P}_n\tilde{\ell}_{\gamma_n} + \sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\theta_n} - \tilde{\ell}_{\gamma_n}\right] \rightsquigarrow \mathcal{N}(\tilde{\mathcal{I}}_\gamma\tau, \tilde{\mathcal{I}}_\gamma).$$

□

Proof of lemma B.1. For (i), use (5) to obtain that under P_{γ_n}

$$\Lambda_n(\gamma_n(\tau_n, h_n), \gamma_n(\tau, h)) = \Lambda_n(\gamma_n(\tau_n, h_n), \gamma_n) - \Lambda_n(\gamma_n(\tau, h), \gamma_n) = o_{P_{\gamma_n}}(1),$$

and so by the continuous mapping theorem, the mutual contiguity of $\left(P^n_{\gamma_n}\right)_{n\in\mathbb{N}}$ and $\left(P^n_{\gamma_n,\tau,h}\right)_{n\in\mathbb{N}}$ (Proposition 3.1) and Le Cam's first lemma (e.g. van der Vaart, 1998, Lemma 6.4)

$$\exp(\Lambda_n(\gamma_n(\tau_n, h_n), \gamma_n(\tau, h))) \rightsquigarrow 1, \qquad \text{under } P_{\gamma_n,\tau,h}.$$

Since ψ_n since it is bounded between 0 and 1, it is tight under $P_{\gamma_n,\tau,h}$ and hence by Prohorov's theorem (e.g. Billingsley, 1999, Theorem 5.1) for any subsequence $(n_j)_{j\in\mathbb{N}}$ of $(n)_{n\in\mathbb{N}}$ there is a further subsequence $(n_k)_{k\in\mathbb{N}}$ such that $\psi_{n_k} \rightsquigarrow \psi$ for some $\psi \in [0, 1]$ under $P_{\gamma_n,\tau,h}$. In conjunction with the preceding display, Slutsky's lemma yields

$$(\psi_n, \exp(\Lambda_n(\gamma_n(\tau_n, h_n), \gamma_n(\tau, h)))) \rightsquigarrow (\psi, 1) \qquad \text{under } P_{\gamma_n,\tau,h}.$$

By Le Cam's third lemma (e.g. van der Vaart, 1998, Theorem 6.6) we have that under P_{γ_n,τ_n,h_n}, the law of ψ_{n_k} converges weakly to the law of ψ in the preceding display. Since each $\psi_n \in [0, 1]$ it is both uniformly $P_{\gamma_n,\tau,h}$-integrable and uniformly P_{γ_n,τ_n,h_n}-integrable. These observations imply that

$$\lim_{k\to\infty}\left[P^{n_k}_{\gamma_{n_k},\tau_{n_k},g_{n_k}}\psi_n - P^{n_k}_{\gamma_{n_k},\tau,h}\psi_{n_k}\right] = 0.$$

Since the original subsequence $(n_j)_{j\in\mathbb{N}}$ was arbitrary, this holds also for the original sequence.

For (ii), from (5), assumption CM(i) and the hypothesis that $B_\gamma h_1 = B_\gamma h_2$

$$\begin{aligned}\Lambda_n(\gamma_n(\tau_n, h_{1,n}), \gamma_n(\tau_n, h_{2,n})) &= \Lambda_n(\gamma_n(\tau_n, h_{1,n}), \gamma_n) - \Lambda_n(\gamma_n(\tau_n, h_{2,n}), \gamma_n)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n B_{\gamma_n}(h_1 - h_2) + o_{P_{\gamma_n}}(1).\end{aligned}$$

$h := h_1 - h_2 \in H_\eta$ by assumption. Let $h_n := h$ for each $n \in \mathbb{N}$ and form g_n as in the proof of proposition 3.1 with $\tau = 0$. Argue analogously to the the proof of proposition 3.1 (noting that for this purpose assumption CM(i) rather than CM(ii) is sufficient) to obtain

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n B_{\gamma_n}(h_1 - h_2) = \frac{1}{\sqrt{n}}\sum_{i=1}^n B_{\gamma_n}h \rightsquigarrow \mathcal{N}(0, \sigma^2_{0,h}), \qquad \text{under } P_{\gamma_n},$$

with $\sigma^2_{0,h} = P_\gamma\left[B_\gamma h\right]^2 = P_\gamma 0^2 = 0$. It follows from the two preceding displays that

$$\Lambda_n(\gamma_n(\tau_n, h_{1,n}), \gamma_n(\tau_n, h_{2,n})) = \Lambda_n(\gamma_n(\tau_n, h_{1,n}), \gamma_n) - \Lambda_n(\gamma_n(\tau_n, h_{2,n}), \gamma_n) = o_{P_{\gamma_n}}(1).$$

With this in hand, the second part of (ii) can be established by an argument analogous to that used to establish (i). $\square$

Proof of proposition B.2. That $\mathbb{H}_\gamma$ is a linear space is clear. Moreover, linearity, coordinate symmetry, and positive semi-definiteness of the map in (56) are clear from its definition. It remains to prove that it is positive definite. Let Π denote the projection onto $\operatorname{cl}\mathscr{H}_\gamma \subset L_2(P_\gamma)$. Then, we can re-write

$$\langle(\tau, g), (\tau, g)\rangle = \tau'\tilde{\mathcal{I}}_\gamma\tau + \langle \tau'\Pi\dot{\ell}_\gamma + g, \tau'\Pi\dot{\ell}_\gamma + g\rangle_{P_\gamma}. \tag{62}$$

This is strictly positive whenever $\tau \in N(\tilde{\mathcal{I}}_\gamma)^\perp \setminus \{0\}$.[87] If instead $\tau = 0$ but $g \neq 0$ it is positive since $\langle\cdot,\cdot\rangle_{P_\gamma}$ is an inner product. Thus $\langle\cdot,\cdot\rangle_{\mathbb{H}_\gamma}$ is an inner product and $(\mathbb{H}_\gamma, \langle\cdot,\cdot\rangle)$ is an inner-product space. Denote the completion of this space with respect to the norm induced by $\langle\cdot,\cdot\rangle$ as $(\overline{\mathbb{H}_\gamma}, \langle\cdot,\cdot\rangle)$.

A Gaussian shift on $(\mathbb{H}_\gamma, \langle\cdot,\cdot\rangle)$ is the restriction to $\mathbb{H}_\gamma$ of the standard Gaussian shift experiment of the Hilbert space $(\overline{\mathbb{H}_\gamma}, \langle\cdot,\cdot\rangle)$. Define

$$L_n(\tau, g) := \Lambda(\gamma_n(\tau, h_g), \gamma_n) + \frac{1}{2}\|(\tau, g)\|^2, \tag{63}$$

and note that equation (59) and the marginal convergence of the log-likelihood (cf. equation (60)):

$$\sqrt{n}\mathbb{P}_n\tau'\dot{\ell}_{\gamma_n} + B_{\gamma_n}h_g \rightsquigarrow \mathcal{N}\left(0, \sigma^2_{\tau,g}\right) \quad \text{under } P_{\gamma_n}, \tag{64}$$

remain valid in this setting, where we write $\sigma^2_{\tau,g}$ for σ^2_{τ,h_g}.[88] By equation (59)

$$\|(\tau, g)\|^2 = \sigma^2_{\tau,g} = P_\gamma\left[\tau'\dot{\ell}_\gamma + g\right]^2 = \lim_{n\to\infty} P_{\gamma_n}\left[\tau'\dot{\ell}_{\gamma_n} + B_{\gamma_n}h_g\right]^2. \tag{65}$$

Equations (5), (63) and (65) allow us to write

$$L_n(\tau, g) = \sqrt{n}\mathbb{P}_n\left[\tau'\dot{\ell}_{\gamma_n} + B_{\gamma_n}h_g\right] + o_{P_{\gamma_n}}(1),$$

and hence by (64),

$$L_n(\tau, g) \rightsquigarrow \mathcal{N}\left(0, \|(\tau, g)\|^2\right) \quad \text{under } P_{\gamma_n}, \quad \text{for any } (\tau, g) \in \mathbb{H}_\gamma. \tag{66}$$

Moreover, for any $(\tau_1, g_1), (\tau_2, g_2) \in \mathbb{H}_\gamma$ and any $a_1, a_2 \in \mathbb{R}$ we have, where $R_{n,i} = o_{P_{\gamma_n}}(1)$

[87] Suppose $\tau \in N(\tilde{\mathcal{I}}_\gamma)^\perp$ and $\tau'\tilde{\mathcal{I}}_\gamma\tau = 0$. The latter implies that $\tilde{\mathcal{I}}_\gamma^{1/2}\tau = 0$, and hence $\tilde{\mathcal{I}}_\gamma\tau = \tilde{\mathcal{I}}_\gamma^{1/2}\tilde{\mathcal{I}}_\gamma^{1/2}\tau = 0$; i.e. $\tau \in N(\tilde{\mathcal{I}}_\gamma)$. Since τ is also in $N(\tilde{\mathcal{I}}_\gamma)^\perp$ we must have $\tau'\tau = 0$, i.e. $\tau = 0$.

[88] Proposition 3.1 requires assumption CM(ii) rather than the weaker CM(i). It is easy to see that an analogous argument as to that given in the proof of proposition 3.1 concerned only with marginal weak convergence of the log-likelihood in equation (60) holds under the weaker condition.

for $i = 1, 2, 3$,

$$\begin{aligned}
&a_1 L_n(\tau_1, g_1) + a_2 L_n(\tau_2, g_2) - L_n(a_1\tau_1 + a_2\tau_2, a_1 g_1 + a_2 g_2) \\
&\quad = a_1\sqrt{n}\mathbb{P}_n\left[\tau_1'\dot{\ell}_{\gamma_n} + B_{\gamma_n} h_{g_1}\right] + a_1 R_{n,1} - a_2\sqrt{n}\mathbb{P}_n\left[\tau_2'\dot{\ell}_{\gamma_n} + B_{\gamma_n} h_{g_2}\right] + a_2 R_{n,2} \\
&\qquad - \sqrt{n}\mathbb{P}_n\left[(a_1\tau_1 + a_2\tau_2)'\dot{\ell}_{\gamma_n} + B_{\gamma_n}[a_1 h_{g_1} + a_2 h_{g_2}]\right] + R_{n,3} \\
&\quad = a_1 R_{n,1} + a_2 R_{n,2} + R_{n,3} \\
&\quad = o_{P_{\gamma_n}}(1).
\end{aligned}$$

That is,

$$\begin{aligned}
a_1 L_n(\tau_1, g_1) + a_2 L_n(\tau_2, g_2) - L_n(a_1\tau_1 + a_2\tau_2, a_1 g_1 + a_2 g_2) = o_{P_{\gamma_n}}(1), \\
\text{whenever } a_1, a_2 \in \mathbb{R},\ (\tau_1, g_1), (\tau_2, g_2) \in \mathbb{H}_\gamma.
\end{aligned} \tag{67}$$

By imitating the proof of Theorem 69.4 in Strasser (1985), one obtains that the experiment

$$\mathscr{E} = (\Omega, \mathcal{F}, \{G_{\tau,g} : (\tau, g) \in \mathbb{H}_\gamma\}) \tag{68}$$

is the restriction to $\mathbb{H}_\gamma$ of a Gaussian shift experiment on $(\overline{\mathbb{H}_\gamma}, \langle\cdot,\cdot\rangle)$ if and only if the stochastic process $(L(\tau, g))_{(\tau,g)\in\mathbb{H}_\gamma}$, defined by

$$L(\tau, g) = \Lambda((\tau, h_g), (0, 0)) + \frac{1}{2}\|(\tau, g)\|^2, \tag{69}$$

with $\Lambda((\tau, h_g), (0, 0))$ the log-likelihood ratio of $G_{\tau,g}$ and $G_{(0,0)}$, is the restriction to $\mathbb{H}_\gamma$ of a standard Gaussian process defined on $\overline{\mathbb{H}_\gamma}$ under $G_{(0,0)}$.[89] Combining equations (66) and (67) we have that for any $K \in \mathbb{N}$, $a \in \mathbb{R}^K$ and $(\tau_k, g_k) \in \mathbb{H}_\gamma$ (for $k = 1, \dots, K$) we have that under P_{γ_n}

$$\sum_{k=1}^K a_k L_n(\tau_k, g_k) \rightsquigarrow \sum_{k=1}^K a_k L^*(\tau_k, g_k) = L^*\left(\sum_{k=1}^K a_k(\tau_k, g_k)\right), \tag{70}$$

for a square integrable stochastic process L^* defined on $\mathbb{H}_\gamma$. Thus we have convergence of the finite dimensional marginal distributions of L_n to those of L^* by the Cramér-Wold theorem. Imitating the proof of Theorem 68.4 in Strasser (1985) yields that a square integrable stochastic process L defined on $\mathbb{H}_\gamma$ is the restriction to $\mathbb{H}_\gamma$ of a standard Gaussian process defined on $\overline{\mathbb{H}_\gamma}$ if and only if L is linear and has a $\mathcal{N}(0, \|(\tau, g)\|^2)$ marginal distribution for each $(\tau, g) \in \mathbb{H}_\gamma$. Since our process L^* satisfies these conditions, it follows that it is such a restriction of a standard Gaussian process. Therefore we have convergence of the finite dimensional distributions of $(L_n(\tau, g))_{(\tau,g)\in\mathbb{H}_\gamma}$ to those of (the restriction to $\mathbb{H}_\gamma$ of) a standard Gaussian process (on $(\overline{\mathbb{H}_\gamma}, \langle\cdot,\cdot\rangle)$). By (63) and (69) this implies the convergence of the finite dimensional distributions of $(\Lambda_n(\gamma_n(\tau, h_g), \gamma_n))_{(\tau,g)\in\mathbb{H}_\gamma}$ to those of $(\Lambda((\tau, h_g), (0, 0)))_{(\tau,g)\in\mathbb{H}_\gamma}$. With this in hand, the proof is completed by an appeal to Theorem 61.6 of Strasser (1985), upon noting that that the sequence of experiments $(\mathscr{E}_n)_{n\in\mathbb{N}}$ is contiguous (see e.g. Strasser, 1985, Definition 61.1) by an analogous argument as used to prove the contiguity claimed in proposition 3.1 and the transitivity of (mutual) contiguity. □

[89] Such a standard Gaussian process is a square integrable stochastic process such that all its finite dimensional distributions are Gaussian with $\mathbb{E}L(\tau_1, g_1) = 0$ and $\mathbb{E}[L(\tau_1, g_1)L(\tau_2, g_2)] = \langle(\tau_1, g_1), (\tau_2, g_2)\rangle$ for all $(\tau_1, g_1), (\tau_2, g_2) \in \overline{\mathbb{H}_\gamma}$.

Lemma B.3. *Suppose that assumptions M, LAN, CM(ii), E and R hold for a sequence $(\gamma_n)_{n\in\mathbb{N}} \subset \Gamma$ with limit $\gamma \in \Gamma$. Then, for any $h_n \to h$ with each $h_n, h \in H_\eta$*

$$\lim_{n\to\infty} P^n_{\gamma_n,0,h_n}\phi_{n,\theta_n} = \begin{cases} \alpha & \text{if } \operatorname{rank}(\tilde{\mathcal{I}}_\gamma) > 0 \\ 0 & \text{if } \operatorname{rank}(\tilde{\mathcal{I}}_\gamma) = 0 \end{cases}.$$

Proof of Lemma B.3. By proposition 3.1 we have that under $P_{\gamma_n,0,h_n}$

$$\sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta_n} \rightsquigarrow \mathcal{N}(0,\tilde{\mathcal{I}}_\gamma).$$

Equations (7), (8) and Lemma C.6 imply that $\|\hat{\mathcal{I}}^\dagger_{n,\theta_n} - \tilde{\mathcal{I}}^\dagger_\gamma\|_2 = o_{P_{\gamma_n}}(1)$. The mutual contiguity established in proposition 3.1 along with Le Cam's first lemma (e.g. van der Vaart, 1998, Lemma 6.4) ensures that this result and equation (6) also hold under $P_{\gamma_n,0,h_n}$:

$$\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\theta_n} - \tilde{\ell}_{\gamma_n}\right] = o_{P_{\gamma_n,0,h_n}}(1) \quad \text{and} \quad \|\hat{\mathcal{I}}^\dagger_{n,\theta_n} - \tilde{\mathcal{I}}^\dagger_\gamma\|_2 = o_{P_{\gamma_n,0,h_n}}(1).$$

Write $\hat{Z}_n := \sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta_n}$. We have

$$\hat{Z}_n = \sqrt{n}\mathbb{P}_n\tilde{\ell}_{\gamma_n} + \sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\theta_n} - \tilde{\ell}_{\gamma_n}\right] \rightsquigarrow Z \sim \mathcal{N}(0,\tilde{\mathcal{I}}_\gamma)$$

under $P_{\gamma_n,0,h_n}$. We now cover the case of one-sided and two-sided tests separately. In the case of a two-sided test, the continuous mapping theorem implies that

$$\hat{S}_{n,\theta_n} = \hat{Z}_n'\hat{\mathcal{I}}^\dagger_{n,\theta_n}\hat{Z}_n \rightsquigarrow Z'\tilde{\mathcal{I}}^\dagger_\gamma Z =: S \sim \chi^2_r,$$

under $P_{\gamma_n,0,h_n}$ where $r = \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$.[90]

Let c_n be the $1-\alpha$ quantile of the $\chi^2_{r_n}$ distribution and c the $1-\alpha$ quantile of the χ^2_r distribution. We have $P_{\gamma_n}\{c_n = c\} = P_{\gamma_n}\{r_n = r\} \to 1$ by assumption. This implies that $c_n - c \to 0$ in P_{γ_n}-probability and hence by the mutual contiguity and Le Cam's first lemma, also under $P_{\gamma_n,0,h_n}$. By continuous mapping once more we have $\hat{S}_{n,\theta_n} - c_n \rightsquigarrow S - c$ under $P_{\gamma_n,0,h_n}$.

Now, consider first the case where $r > 0$. In this case, since the χ^2_r distribution is continuous the portmanteau theorem gives

$$P_{\gamma_n,0,h_n}\phi_{n,\theta_n} = P_{\gamma_n,0,h_n}\left(\hat{S}_{n,\theta_n} - c_n > 0\right) \to L\left(S - c > 0\right) = \alpha,$$

where L is the law of S. In the case where instead $r = 0$ we note that on the sets $\{r_n = r\} = \{r_n = 0\}$ we have that $\hat{\mathcal{I}}^\dagger_{n,\theta_n} = 0$ and $c_n = 0$ and hence do not reject since $\hat{S}_{n,\theta_n} = 0 \le c_n = 0$. It follows that $P_{\gamma_n,0,h_n}\phi_{n,\theta_n} \le 1 - P_{\gamma_n,0,h_n}\{r_n = r\} \to 0$.

Finally consider a one-sided test with $d_\theta = 1$ and $1 - \alpha \in [1/2, 1)$. By the continuous mapping theorem,

$$\hat{S}_{n,\theta_n} = \hat{Z}_n\sqrt{\hat{\mathcal{I}}^\dagger_{n,\theta_n}} \rightsquigarrow Z\sqrt{\tilde{\mathcal{I}}^\dagger_\gamma}.$$

If $r = \operatorname{rank}(\tilde{\mathcal{I}}_\gamma) = 1$, then $Z\sqrt{\tilde{\mathcal{I}}^\dagger_\gamma} = Z/\sqrt{\tilde{\mathcal{I}}_\gamma} \sim \mathcal{N}(0,1)$ and since this distribution is

[90] The distributional result is given by, for example, Theorem 9.2.2 in Rao and Mitra (1971).

continuous, the portmanteau theorem yields

$$P_{\gamma_n,0,h_n}\phi_{n,\theta_n} \to 1-\Phi(z_\alpha)=\alpha,$$

where Φ is the CDF of the standard normal distribution. If, instead $r=0$, then again on the sets where $r_n = \operatorname{rank}(\hat{\mathcal{I}}_{n,\theta_n}) = 0$ we have that $\hat{\mathcal{I}}_{n,\theta_n} = \hat{\mathcal{I}}^\dagger_{n,\theta_n} = 0$ and so $\hat{S}_{n,\theta_n} = 0 \leq z_\alpha$ and hence we do not reject. It follows that $P_{\gamma_n,0,h_n}\phi_{n,\theta_n} \leq 1 - P_{\gamma_n,0,h_n}\{r_n = r\} \to 0$. $\square$

Lemma B.4. *Suppose that assumptions M, LAN, CM(ii), E and R hold for a convergent sequence $(\gamma_n)_{n\in\mathbb{N}} \subset \Gamma$ with limit $\gamma \in \Gamma$. Suppose we are given a convergent sequences $h_{n_k} \to h \in H_\eta$ with $(h_{n_k})_{k\in\mathbb{N}} \subset H_\eta$. If the limit*

$$\mathcal{S} \coloneqq \lim_{k\to\infty} P^{n_k}_{\gamma_{n_k},0,h_{n_k}}\phi_{n_k,\theta_{n_k}} \tag{71}$$

exists, then $\mathcal{S} = \alpha \times \mathbf{1}\{\operatorname{rank}(\tilde{\mathcal{I}}_\gamma) > 0\}$.

Proof. The idea is to construct a new sequence to which proposition B.3 can be applied.[91] For all m with $m \in [n_k, n_{k+1}) \cap \mathbb{N}$ for some $k \in \mathbb{N}$ put $h^*_m = h_{n_k}$. For $m = 1,\ldots,n_1$, put $h^*_m = h_{n_1}$. For each m let $\gamma^*_m = \gamma_m$. By construction $h^*_m \to h$, and by our hypotheses and proposition B.3 we may conclude that

$$\lim_{m\to\infty} P^m_{\gamma^*_m,0,h^*_m}\phi_{m,\theta^*_m} = s_\gamma \coloneqq \begin{cases} \alpha & \text{if } \operatorname{rank}(\tilde{\mathcal{I}}_\gamma) > 0 \\ 0 & \text{if } \operatorname{rank}(\tilde{\mathcal{I}}_\gamma) = 0 \end{cases}.$$

Fix an arbitrary $\varepsilon > 0$. There is a $M \in \mathbb{N}$ such that for all $m \geq M$, $\left|P^m_{\gamma^*_m,0,h^*_m}\phi_{m,\theta^*_m} - s_\gamma\right| < \varepsilon/2$. By (71) there is a $K \in \mathbb{N}$ such that if $k \geq K$, $\left|\mathcal{S} - P^{n_k}_{\gamma_{n_k},0,h_{n_k}}\phi_{n_k,\theta_{n_k}}\right| < \varepsilon/2$. Hence for any k sufficiently large that $m = n_k \geq M$ and $k \geq K$ we have

$$|\mathcal{S} - s_\gamma| \leq \left|\mathcal{S} - P^m_{\gamma^*_m,0,h^*_m}\phi_{m,\theta^*_m}\right| + \left|P^m_{\gamma^*_m,0,h^*_m}\phi_{m,\theta^*_m} - s_\gamma\right| < \left|\mathcal{S} - P^{n_k}_{\gamma_{n_k},0,h_{n_k}}\phi_{n_k,\theta_{n_k}}\right| + \frac{\varepsilon}{2} < \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, the inequality $|\mathcal{S} - s_\gamma| < \varepsilon$ can be obtained for any $\varepsilon > 0$ and hence taking the limit as $\varepsilon \downarrow 0$ completes the proof. $\square$

Proof of proposition 3.2. There is a sequence $(h_n)_{n\in\mathbb{N}} \subset H'_\eta$ and a subsequence $(n_j)_{j\in\mathbb{N}}$ of $(n)_{n\in\mathbb{N}}$ such that

$$\mathcal{S} \coloneqq \limsup_{n\to\infty} \sup_{h\in H'_\eta} P^n_{\gamma_n,0,h}\phi_{n,\theta_n} = \limsup_{n\to\infty} P^n_{\gamma_n,0,h_n}\phi_{n,\theta_n} = \lim_{j\to\infty} P^{n_j}_{\gamma_{n_j},0,h_{n_j}}\phi_{n_j,\theta_{n_j}}$$

There is a further subsequence $(n_k)_{k\in\mathbb{N}}$ such that $h_{n_k} \to h$ and $\mathcal{S} = \lim_{k\to\infty} P^{n_k}_{\gamma_{n_k},0,h_{n_k}}\phi_{n_k,\theta_{n_k}}$. Applying lemma B.4 yields that $\mathcal{S} = \alpha \times \mathbf{1}\{\operatorname{rank}(\tilde{\mathcal{I}}_\gamma) > 0\}$. Since an analogous argument can be made to obtain the same conclusion but with “”lim inf” replacing “”lim sup” in the definition of $\mathcal{S}$, we obtain the desired result. $\square$

Lemma B.5. *Fix a convergent sequence $(\eta_n)_{n\in\mathbb{N}}$ and denote its limit by η. Suppose that assumptions M, LAN, CM(ii), E and R hold for any sequence $(\gamma_n)_{n\in\mathbb{N}}$ where each $\gamma_n \coloneqq$*

[91] This construction is based on that used in the proofs of e.g. Lemma 6 in Andrews and Guggenberger (2010b), Lemma 2.1 in Andrews and Cheng (2012).

$(\theta_n, \eta_n)_{n\in\mathbb{N}} \subset \Theta' \times \mathcal{H} =: \Gamma'$ *with* $\theta_n \to \theta \in \Theta' \subset \Theta$. *Suppose we are given convergent sequences* $\gamma_{n_k} \to \gamma$ *with* $(\gamma_{n_k})_{k\in\mathbb{N}} \subset \Gamma'$ *and* $h_{n_k} \to h$ *with* $(h_{n_k})_{k\in\mathbb{N}} \subset H_\eta$. *If the limit*

$$\mathcal{S} := \lim_{k\to\infty} P^{n_k}_{\gamma_{n_k},0,h_{n_k}} \phi_{n_k,\theta_{n_k}} \tag{72}$$

exists, then $\mathcal{S} \le \alpha$.

Proof. The idea is to construct a new sequence to which proposition B.3 can be applied.[92] For all m with $m \in [n_k, n_{k+1}) \cap \mathbb{N}$ for some $k \in \mathbb{N}$ put $\theta^*_m = \theta_{n_k}$ and $h^*_m = h_{n_k}$. For $m = 1, \ldots, n_1$, put $\theta^*_m = \theta_{n_1}$ and $h^*_m = h_{n_1}$. For each m let $\gamma^*_m = (\theta^*_m, \eta_m)$. By construction $\gamma^*_m \to \gamma$ through Γ' and $h^*_m \to h$, and by our hypotheses and proposition B.3 we may conclude that

$$\lim_{m\to\infty} P^m_{\gamma^*_m,0,h^*_m} \phi_{m,\theta^*_m} \le \alpha.$$

Fix an arbitrary $\varepsilon > 0$. There is a $M \in \mathbb{N}$ such that for all $m \ge M$, $P^m_{\gamma^*_m,0,h^*_m} \phi_{m,\theta^*_m} \le \alpha + \varepsilon/2$. By (72) there is a $K \in \mathbb{N}$ such that if $k \ge K$, $\left| \mathcal{S} - P^{n_k}_{\gamma_{n_k},0,h_{n_k}} \phi_{n_k,\theta_{n_k}} \right| < \varepsilon/2$. Hence for any k sufficiently large that $m = n_k \ge M$ and $k \ge K$ we have

$$\mathcal{S} \le \left| \mathcal{S} - P^m_{\gamma^*_m,0,h^*_m} \phi_{m,\theta^*_m} \right| + P^m_{\gamma^*_m,0,h^*_m} \phi_{m,\theta^*_m} < \left| \mathcal{S} - P^{n_k}_{\gamma_{n_k},0,h_{n_k}} \phi_{n_k,\theta_{n_k}} \right| \alpha + \frac{\varepsilon}{2} \le \alpha + \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, we can obtain the inequality $\mathcal{S} \le \alpha + \varepsilon$ for any $\varepsilon > 0$ and hence taking the limit as $\varepsilon \downarrow 0$ completes the proof. □

Proof of proposition 3.3. There are sequences $(\theta_n)_{n\in\mathbb{N}} \subset \Theta'$ and $(h_n)_{n\in\mathbb{N}} \subset H'_\eta$ and a subsequence $(n_j)_{j\in\mathbb{N}}$ of $(n)_{n\in\mathbb{N}}$ such that

$$\mathcal{S} := \liminf_{n\to\infty} \inf_{\theta\in\Theta'} \inf_{h\in H'_\eta} P^n_{(\theta,\eta_n),0,h}(\theta \in \hat{C}_n) = \lim_{j\to\infty} P^{n_j}_{(\theta_{n_j},\eta_{n_j}),0,h_{n_j}}(\theta_{n_j} \in \hat{C}_{n_j}).$$

There is a further subsequence $(n_k)_{k\in\mathbb{N}}$ of $(n_j)_{j\in\mathbb{N}}$ such that $\theta_{n_j} \to \theta \in \Theta'$ and $h_{n_j} \to h \in H'_\eta$. We also clearly have

$$\mathcal{S} = \lim_{k\to\infty} P^{n_k}_{(\theta_{n_k},\eta_{n_k}),0,h_{n_k}}(\theta_{n_k} \in \hat{C}_{n_k}) = 1 - \lim_{k\to\infty} P^{n_k}_{(\theta_{n_k},\eta_{n_k}),0,h_{n_k}} \phi_{n_k,\theta_{n_k}}. \tag{73}$$

Apply lemma B.5 to conclude that $1 - \mathcal{S} \le \alpha$, and rearrange to obtain the desired result. □

Proof of proposition 3.4. By (both parts of) lemma B.1, it suffices to show that

$$\limsup_{n\to\infty} P^n_{\gamma_n,\tau,h}\psi_n \le 1 - \Phi\left(z_\alpha - \tilde{\mathcal{I}}_\gamma^{1/2}\tau\right) \qquad \text{for all } \tau > 0,\ h \in \mathfrak{H}_\gamma. \tag{74}$$

Since $d_\theta = 1$ and $\tilde{\mathcal{I}}_\gamma > 0$, $N(\tilde{\mathcal{I}}_\gamma)^\perp = \mathbb{R}$. Let $\tilde{g} = (g_k)_{k\in\mathbb{N}} \subset \mathscr{H}_\gamma$ be an orthonormal basis of $\operatorname{cl} \mathscr{H}_\gamma$.[93] Consider the subspace $\mathcal{G}^m := \operatorname{Span}\{g_1, \ldots, g_m\}$, and let Π^m denote the orthogonal

[92] See footnote 91.

[93] Such a basis always exists: by assumption M, $\mathcal{W}$ is Polish. Take a metric d such that $(\mathcal{W}, d)$ is a complete (separable) metric space. By Theorem 1.3 in Billingsley (1999), P_γ is tight. By Proposition 7.14.12 in Bogachev (2007) this is a sufficient condition for separability of P_γ which is equivalent to separability of the $L_p(P_\gamma)$ spaces for $p \in (0, \infty)$ (e.g. Bogachev, 2007, Exercise 4.7.63). $\operatorname{cl} \mathscr{H}_\gamma$ is therefore separable as a subset of $L_2(P_\gamma)$. Choose a countable dense subset in $\mathscr{H}_\gamma$ and apply Gram-Schmidt to obtain an orthonormal basis which satisfies the the desired property.

projection onto $\mathcal{G}^m$. Fix $b = (\tau, g_b) \in (0, \infty) \times \mathscr{H}_\gamma =: K_1$ and any $\varepsilon > 0$.[94] By lemma C.2 we can take $m \in \mathbb{N}$ large enough that $\left\|(\Pi^m - \Pi)\dot{\ell}_\gamma\right\|_{P_\gamma,2} < \varepsilon$. Now consider the restriction of $\mathscr{E}$ to $\mathbb{R} \times \mathcal{G}^m$ for any $m \in \mathbb{N}$.[95] Choose $a = (0, g_a)$ with $g_a = \Pi^m \left(\tau \Pi \dot{\ell}_\gamma + g_b\right) = \tau \left(\Pi^m \Pi \dot{\ell}_\gamma\right) + g_b$ and note that by Lemma 28.1 of Strasser (1985) any test ψ of level$-\alpha$ of H_0 against H_1 satisfies

$$G_b \psi \leq 1 - \Phi\left(z_\alpha - \|b - a\|\right)$$

Expand the square of the norm using the Pythagorean theorem to obtain

$$\|b - a\|^2 = \tau^2 \tilde{\mathcal{I}}_\gamma + \tau^2 \left\|(\Pi^m - \Pi)\dot{\ell}_\gamma\right\|_{P_\gamma,2}^2 = \tau^2 \tilde{\mathcal{I}}_\gamma + \tau^2 \varepsilon^2.$$

Hence we have

$$G_b \psi \leq 1 - \Phi\left(z_\alpha - \sqrt{\tau^2 \tilde{\mathcal{I}}_\gamma + \tau^2 \varepsilon^2}\right).$$

Since $\varepsilon > 0$ was arbitrary, we can take the limit as $\varepsilon \downarrow 0$ to obtain

$$G_b \psi \leq 1 - \Phi\left(z_\alpha - \tilde{\mathcal{I}}_\star^{1/2} \tau\right), \tag{75}$$

which holds for all $b \in K_1$, since the choice of $b \in K_1$ was arbitrary. Moreover, since the test ψ was an arbitrary test of level-α, this power bound holds for all level-α tests in $\mathscr{E}$.

By proposition B.2 the the sequence of experiments $(\mathscr{E}_n)_{n \in \mathbb{N}}$ defined in (57) converge to the dominated experiment $\mathscr{E}$. (74) then follows on combining the power bound given by (75) with Theorem 7.2 in van der Vaart (1991). □

Proof of corollary 3.5. Since $\tilde{\mathcal{I}}_\gamma > 0$ and $d_\theta = 1$, assumption R is automatically satisfied given assumption E. By proposition 3.1 we have that

$$\sqrt{n} \mathbb{P}_n \hat{\ell}_{n,\theta_0} / \hat{\mathcal{I}}_{n,\theta_0}^{1/2} \rightsquigarrow \mathcal{N}(\tilde{\mathcal{I}}_\gamma^{1/2} \tau, 1), \text{ under } P_{\gamma_n, \tau_n, h_n}.$$

Hence by the portmanteau theorem

$$\lim_{n\to\infty} P^n_{\gamma_n,\tau_n,h_n} \phi_n = \lim_{n\to\infty} P^n_{\gamma_n,\tau_n,h_n} (\sqrt{n} \mathbb{P}_n \hat{\ell}_{n,\theta_0} / \hat{\mathcal{I}}_{n,\theta_0}^{1/2} > z_\alpha) = 1 - \Phi(z_\alpha - \tilde{\mathcal{I}}_\gamma^{1/2} \tau).$$

For $\tau \leq 0$, $1 - \Phi(z_\alpha - \tilde{\mathcal{I}}_\star^{1/2} \tau) \leq \alpha$; hence this test is level-α as claimed. For any $\tau > 0$, it attains the power bound in equation (15). □

Proof of proposition 3.6. The proof is is very similar to that of proposition 3.4. By lemma B.1 it suffices to show that for all $\tau \neq 0$ and $h \in \mathfrak{H}_\gamma$

$$\limsup_{n\to\infty} P^n_{\gamma_n,\tau,h} \psi_n \leq 1 - \Phi\left(z_{\alpha/2} - \tilde{\mathcal{I}}_\gamma^{1/2} \tau\right) + 1 - \Phi\left(z_{\alpha/2} + \tilde{\mathcal{I}}_\gamma^{1/2} \tau\right). \tag{76}$$

Since $d_\theta = 1$ and $\tilde{\mathcal{I}}_\gamma > 0$, $N(\tilde{\mathcal{I}}_\gamma)^\perp = \mathbb{R}$. Let $\tilde{g}$, $\mathcal{G}^m$ and Π^m be defined as in the proof of proposition 3.4 and consider the restriction of $\mathscr{E}$ to $L^m := \mathbb{R} \times \mathcal{G}^m$ for some $m \in \mathbb{N}$ which contains $(\tau, g) \in K_1 = \{(\tau, g) : \tau \neq 0,\ h \in \mathscr{H}_\gamma\}$.[96] This is a finite dimensional (hence closed)

[94] We can always change the choice of the orthonormal basis such that g_b lies in (each) $\mathcal{G}^m$.
[95] See equations (68), (69) and the surrounding text for the definitions of $\mathscr{E}$ and $G_{\tau,g}$.
[96] See footnote 94.

subspace of $\overline{\mathbb{H}_\gamma}$ (the completion of $\mathbb{H}_\gamma$) and so is a Hilbert space. Hence this restriction is a finite dimensional (standard) Gaussian shift. Take $f : \mathbb{R} \times \mathcal{G}^m \to \mathbb{R}$ as $f(\tau, g) = \tau$ and let $\Sigma^m := P_\gamma \left([I - \Pi^m]\dot{\ell}_\gamma\right)^2$, which can be ensured positive by taking $m \in \mathbb{N}$ sufficiently large.[97] Then, letting $g \in \mathcal{G}^m$ be such that $g = -\Pi^m \dot{\ell}_\gamma \in \mathcal{G}^m$, $e = (1, g)/\sqrt{\Sigma^m}$ is a unit vector in $\mathbb{R} \times \mathcal{G}^m \subset \mathbb{H}_\gamma$, orthogonal to $N(f) = \{(0, g) : g \in \mathcal{G}^m\}$ and has $f(e) = 1/\sqrt{\Sigma^m} > 0$. Thus, by Theorem 28.8 of Strasser (1985), any unbiased test ψ of level-α has power bounded by

$$G_{\tau,g}\psi \leq 1 - \Phi(z_{\alpha/2} - (\Sigma^m)^{1/2}\tau) + 1 - \Phi(z_{\alpha/2} + (\Sigma^m)^{1/2}\tau).$$

Since $\Sigma^m \to \tilde{\mathcal{I}}_\gamma$ as $m \to \infty$, by continuity we obtain that

$$G_{\tau,g}\psi \leq 1 - \Phi(z_{\alpha/2} - \tilde{\mathcal{I}}_\gamma^{1/2}\tau) + 1 - \Phi(z_{\alpha/2} + \tilde{\mathcal{I}}_\gamma^{1/2}\tau). \tag{77}$$

Since the point $(\tau, g) \in K_1$ was arbitrary, this bound holds for all K_1.

By proposition B.2 the sequence of experiments $(\mathscr{E}_n)_{n\in\mathbb{N}}$ converges to the dominated experiment $\mathscr{E}$. Let $\pi_n(\tau, g) := P^n_{\gamma_n,\tau,g}\psi_n \in [0, 1]$. Fix a $(\tau, g) \in K_1$ and let $(n_j)_{j\in\mathbb{N}}$ be a subsequence of $(n)_{n\in\mathbb{N}}$ along which $\limsup_{n\to\infty} P^n_{\gamma_n,\tau,g}\psi_n = \lim_{j\to\infty} P^{n_j}_{\gamma_{n_j},\tau,g}\psi_{n_j}$. Since $[0, 1]^{\mathbb{H}_\gamma}$ is compact in the product topology there is a subnet $(n_{j(\alpha)})_{\alpha\in A}$ of the subsequence $(n_j)_{j\in\mathbb{N}}$ and a function $\pi : \mathbb{H}_\gamma \to [0, 1]$ such that $\lim_{\alpha\in A} \pi_{n_{j(\alpha)}}(\tau, g) = \pi(\tau, g)$ for every $(\tau, g) \in \mathbb{H}_\gamma$. By Theorem 7.1 in van der Vaart (1991) there is a test ψ in $\mathscr{E}$ with power function π. By our hypotheses and the pointwise convergence we have that for any $\tau \neq 0$ and any $g_1, g_2 \in \mathscr{H}_\gamma$

$$\pi(0, g_1) = \lim_{\alpha\in A} \pi_{n_{j(\alpha)}}(0, g_1) \leq \alpha \leq \lim_{\alpha\in A} \pi_{n_{j(\alpha)}}(\tau, g_2) = \pi(\tau, g_2).$$

It follows that ψ is unbiased and hence combining

$$\limsup_{n\to\infty} P^n_{\gamma_n,\tau,g}\psi_n = \limsup_{n\to\infty} \pi_n(\tau, g) = \lim_{j\to\infty} \pi_{n_j}(\tau, g) = \lim_{\alpha\in A} \pi_{n_{j(\alpha)}}(\tau, g) = \pi(\tau, g)$$

with the power bound given by (77) we obtain (76).[98] □

Proof of corollary 3.7. Since $\tilde{\mathcal{I}}_\gamma > 0$ and $d_\theta = 1$, assumption R is automatically satisfied given assumption E. By proposition 3.1 we have that

$$\sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta_0}/\hat{\mathcal{I}}^{1/2}_{n,\theta_0} \rightsquigarrow \mathcal{N}(\tilde{\mathcal{I}}_\gamma^{1/2}\tau, 1), \text{ under } P_{\gamma_n,\tau_n,h_n}.$$

Let the $1 - \alpha$ quantile of the χ^2_1 distribution be denoted by c_α. By assumption R holds and the contiguity noted in proposition 3.1 we have that $P_{\gamma_n,\tau_n,h_n}(\hat{r}_n = 1) \to 1$ and hence $c_n \to c_\alpha$ in P_{γ_n,τ_n,h_n}-probability. Hence by the portmanteau theorem

$$\lim_{n\to\infty} P^n_{\gamma_n,\tau_n,h_n}\phi_{n,\theta_0} = 1 - \Phi(z_{\alpha/2} - \tilde{\mathcal{I}}_\star^{1/2}\tau) + 1 - \Phi(z_{\alpha/2} + \tilde{\mathcal{I}}_\star^{1/2}\tau),$$

which is exactly the power bound given by equation (16). For $\tau = 0$, $1 - \Phi(z_{\alpha/2}) + 1 - \Phi(z_{\alpha/2}) = \alpha$; hence this test is level-α as claimed. It is unbiased since the last right hand side expression in the preceding display exceeds α for any $\tau \neq 0$. □

Lemma B.6. *If $(\overline{\mathbb{H}_\gamma}, \langle\cdot,\cdot\rangle)$ is the completion of $(\mathbb{H}_\gamma, \langle\cdot,\cdot\rangle)$, then*

[97] By lemma C.2 we have that $\Sigma^m \to \tilde{\mathcal{I}}_\gamma > 0$ as $m \to \infty$.

[98] Where $g = B_\gamma h$ for the $h \in \mathfrak{H}_\gamma$ in the latter.

(i) we can take $\overline{\mathbb{H}_\gamma}$ to be $N(\tilde{\mathcal{I}}_\gamma)^\perp \times \operatorname{cl} \mathscr{H}_\gamma$;

(ii) $(\tau_n, g_n)_{n\in\mathbb{N}} \subset \mathbb{H}_\gamma$ converges to $(\tau, g) \in \overline{\mathbb{H}_\gamma}$ if and only if $\tau_n \to \tau \in N(\tilde{\mathcal{I}}_\gamma)^\perp$ and $g_n \to g \in \operatorname{cl} \mathscr{H}_\gamma$.

Proof. We first note that $(x, y) \mapsto x'\tilde{\mathcal{I}}_\gamma y$ defines an inner-product on $N(\tilde{\mathcal{I}}_\gamma)^\perp$. Linearity and symmetry are obvious. Positive definiteness was established in footnote 87. On $\mathbb{R}^{d_\theta}$ it defines a positive-semidefinite Hermitian form and thus induces a semi-norm by $\|x\| := \sqrt{x'\tilde{\mathcal{I}}_\gamma x}$.

By the Pythagorean theorem we can decompose the square of the $\overline{\mathbb{H}_\gamma}$ norm as follows

$$\|(\tau_n, g_n) - (\tau, g)\|^2 = (\tau_n - \tau)'\tilde{\mathcal{I}}_\gamma(\tau_n - \tau) + \|(\tau_n - \tau)'\Pi\dot{\ell}_\gamma + g_n - g\|^2_{P_\gamma,2}. \tag{78}$$

We start with the first claim. Suppose that $(\tau_n, g_n)_{n\in\mathbb{N}} \subset \mathbb{H}_\gamma$ is a Cauchy sequence. By (78) we must have that $(\tau_n - \tau_m)'\tilde{\mathcal{I}}_\gamma(\tau_n - \tau)m \to 0$ as $n, m \to \infty$. Let UDU' be an eigendecomposition of $\tilde{\mathcal{I}}_\gamma^{1/2}$ with eigenvalues $\lambda_1, \ldots, \lambda_{d_\theta}$ in decreasing order. Then the eigenvectors u_j for $j > r$ are in the null space of $\tilde{\mathcal{I}}_\gamma^{1/2}$ and so that of $\tilde{\mathcal{I}}_\gamma$. Letting U_1 be the $d_\theta \times r$ matrix of the first r columns of U and U_2 the remaining columns, we then have that $\|\tau_n - \tau_m\|_2 = \|U'(\tau_n - \tau_m)\|_2 = \|U_1'(\tau_n - \tau_m)\|_2$. Let $\tilde{\tau}_{n,m} := U_1'(\tau_n - \tau_m)$ and note that by hypothesis

$$(\tau_n - \tau_m)'\tilde{\mathcal{I}}_\gamma(\tau_n - \tau_m) = \sum_{i=1}^{r} \lambda_i \tilde{\tau}^2_{n,m,i} \to 0.$$

Since the λ_i are all positive this implies that $\|\tilde{\tau}_{n,m}\|_2 \to 0$, i.e. $\tau_n - \tau_m \to 0$. Since this is a Cauchy sequence in $N(\tilde{\mathcal{I}}_\gamma)^\perp$, which is a closed subspace of $\mathbb{R}^{d_\theta}$, it follows that τ_n has a limit, say $\tau^* \in N(\tilde{\mathcal{I}}_\gamma)^\perp$. From this and that $\left\|(\tau_n - \tau_m)'\Pi\dot{\ell}_\gamma + g_n - g_m\right\|_{P_\gamma,2} \to 0$ (as $m, n \to \infty$) we can also conclude that $(g_n)_{n\in\mathbb{N}}$ is Cauchy in $L_2(P_\gamma)$ and hence has a limit, say $g^* \in \operatorname{cl} \mathscr{H}_\gamma$.[99] Hence all such Cauchy sequences have limits in $N(\tilde{\mathcal{I}}_\gamma)^\perp \times \operatorname{cl} \mathscr{H}_\gamma$ and so this is complete under the relevant norm.

To complete the proof we will now show that $(\tau_n, g_n)_{n\in\mathbb{N}} \subset \mathbb{H}_\gamma$ converges to $(\tau, g) \in N(\tilde{\mathcal{I}}_\gamma)^\perp \times \operatorname{cl} \mathscr{H}_\gamma$ if and only if $\tau_n \to \tau \in N(\tilde{\mathcal{I}}_\gamma)^\perp$ and $g_n \to g \in \operatorname{cl} \mathscr{H}_\gamma$. Since this ensures that $N(\tilde{\mathcal{I}}_\gamma)^\perp \times \operatorname{cl} \mathscr{H}_\gamma = \operatorname{cl} \mathbb{H}_\gamma$, this is the smallest closed set containing $\mathbb{H}_\gamma$, which completes the proof of the first part, and hence the second.

Suppose first that $(\tau_n, g_n)_{n\in\mathbb{N}} \subset \mathbb{H}_\gamma$ converges to $(\tau, g) \in N(\tilde{\mathcal{I}}_\gamma)^\perp \times \operatorname{cl} \mathscr{H}_\gamma$. Then since each $\tau_n - \tau \in \mathbb{N}(\tilde{\mathcal{I}}_\gamma)^\perp$ we can argue as above via the same eigendecomposition (replacing τ_m with τ) to obtain that $\tau_n - \tau \to 0$. An argument analogous to that in footnote 99 (replace g_m with g) can be used to show the convergence $g_n \to g$ in the $L_2(P_\gamma)$ norm.

For the converse, suppose that $\tau_n \to \tau$ and $g_n \to g$. It follows immediately that $(\tau_n - \tau)'\tilde{\mathcal{I}}_\gamma(\tau_n - \tau) \to 0$ and $\|(\tau_n - \tau)'\Pi\dot{\ell}_\gamma\|_{P_\gamma,2} \to 0$. Using (78) we have

$$\|(\tau_n, g_n) - (\tau, g)\|^2 \lesssim (\tau_n - \tau)'\tilde{\mathcal{I}}_\gamma(\tau_n - \tau) + \|(\tau_n - \tau)'\Pi\dot{\ell}_\gamma\|^2_{P_\gamma,2} + \|g_n - g\|^2_{P_\gamma,2} = o(1).$$

□

[99] By the reverse triangle inequality we have

$$\lim_{n,m\to\infty} \|g_n - g_m\|_{P_\gamma,2} \leq \lim_{n,m\to\infty} \left\|(\tau_n - \tau_m)'\Pi\dot{\ell}_\gamma + g_n - g_m\right\|_{P_\gamma,2} = 0.$$

Proof of proposition 3.8. Let $\tilde{M}_a := \{(\tau, h) \in M_a : h \in \mathfrak{H}_\gamma\}$. We clearly have that

$$\limsup_{n\to\infty} \inf_{(\tau,h)\in M_a} P^n_{\gamma_n,\tau,h}\psi_n \le \limsup_{n\to\infty} \inf_{(\tau,h)\in \tilde{M}_a} P^n_{\gamma_n,\tau,h}\psi_n,$$

so it will suffice to demonstrate the upper bound in

$$\limsup_{n\to\infty} \inf_{(\tau,h)\in \check{M}_a} P^n_{\gamma_n,\tau,g}\psi_n = \limsup_{n\to\infty} \inf_{(\tau,h)\in \tilde{M}_a} P^n_{\gamma_n,\tau,h}\psi_n \le 1 - \mathrm{P}\left(\chi^2_r(a) \le c_{r,\alpha}\right), \tag{79}$$

where $\check{M}_a := \{(\tau, g) \in \mathbb{H}_\gamma : \tau'\tilde{\mathcal{I}}_\gamma\tau \ge a\}$. We first observe that if $(\tau, g) \in \overline{\mathbb{H}_\gamma}$ then $\tau \in N(\tilde{\mathcal{I}}_\gamma)^\perp$. Define $f : \overline{\mathbb{H}_\gamma} \to \mathbb{R}^{d_\theta}$ by $f(\tau, g) := \tau$ and let $L_0 := N(f)$. Let Π_0 denote the orthogonal projection onto L_0 in $\overline{\mathbb{H}_\gamma}$ and Π the orthogonal projection onto $\operatorname{cl}\mathscr{H}_\gamma$ in $L_2(P_\gamma)$. The (finite dimensional) subspace $L_0^\perp \subset \overline{\mathbb{H}_\gamma}$ consists of vectors

$$L_0^\perp = \left\{(\tau, -\tau'\Pi\dot{\ell}_\gamma) \in \overline{\mathbb{H}_\gamma}\right\}.$$

It follows from lemma B.6 that this has dimension r, since we can take $\overline{\mathbb{H}_\gamma} = N(\tilde{\mathcal{I}}_\gamma)^\perp \times \operatorname{cl}\mathscr{H}_\gamma$.

Consider the orthogonal projection onto L_0: we must have $\langle(\tau, g) - \Pi_0(\tau, g), (0, g')\rangle = 0$ for all $(0, g') \in L_0$. This implies that $\Pi_0(\tau, g) = (0, \tilde{g})$ must satisfy $\tilde{g} = \tau'\Pi\dot{\ell}_\gamma + g$. It follows that $\|(\tau, g) - \Pi_0(\tau, g)\|^2 = \tau'\tilde{\mathcal{I}}_\gamma\tau$. Define

$$\overline{M}_a = \left\{(\tau, h) \in \overline{\mathbb{H}_\gamma} : \tau'\tilde{\mathcal{I}}_\gamma\tau \ge a\right\},$$

and let $\overline{M}'_a$ be the set defined analogously to $\overline{M}_a$ where "=" replaces "≥". We note here that $\overline{M}_a = \operatorname{cl}\check{M}_a$. For this, note firstly that any convergent $(t_n, g_n)_{n\in\mathbb{N}} \subset \overline{M}_a$ converges in $\overline{M}_a$ and hence this is a closed set.[100] It follows that $\operatorname{cl}\check{M}_a \subset \overline{M}_a$. Suppose that this inclusion were strict. Then there must be a point $(\tau, g) \in \overline{M}_a$ which is not the limit of a sequence $(\tau_n, g_n)_{n\in\mathbb{N}} \subset \check{M}_a$. There must exist a sequence $(\tau_n, g_n)_{n\in\mathbb{N}} \subset \mathbb{H}_\gamma$ with $(\tau_n, g_n) \to (\tau, g)$. By the argument in footnote 100 we have that $\tau_n'\tilde{\mathcal{I}}_\gamma\tau_n \to \tau\tilde{\mathcal{I}}_\gamma\tau$. If the difference $e_n := \tau\tilde{\mathcal{I}}_\gamma\tau - \tau_n'\tilde{\mathcal{I}}_\gamma\tau_n \to 0$ is always negative there is nothing to do. Else take a sequence $(\tau_n', 0)_{n\in\mathbb{N}} \subset \mathbb{H}_\gamma$ which converges to $(0, 0)$ and satisfies $\tau_n'\tilde{\mathcal{I}}_\gamma\tau_n \ge \max\{e_n, 0\}$.[101] Then $(\tau_n + \tau_n', g_n)_{n\in\mathbb{N}} \subset \check{M}_a$ and converges to (τ, g). Hence no such point can exist and the two sets are equal.

Consider the testing problem of $K_0' = \{0\}$ against $K_1' = L_0^\perp \setminus \{0\}$ in the standard Gaussian shift experiment on $L_0^\perp$. For any $a' \ge a$ and any level$-\alpha$ test ψ we have by Theorem 30.2 of Strasser (1985) that (Cf. Strasser, 1985, Theorem 71.10)

$$\inf_{t\in\overline{M}'_{a'}} G_t\psi \le \inf_{t\in\overline{M}'_{a'}\cap L_0^\perp} G_t\psi \le \mathrm{P}\left(\chi^2_r(a') > c_{r,\alpha}\right).$$

[100] That $(\tau, g) \in \overline{\mathbb{H}_\gamma}$ is clear since the latter is complete and hence closed. It remains to show that if $\tau_n\tilde{\mathcal{I}}_\gamma\tau_n \ge a$ for each $n \in \mathbb{N}$ then also $\tau\tilde{\mathcal{I}}_\gamma\tau \ge a$. For this, we note that if $(\tau_n, g_n) \to (\tau, g)$ then by lemma B.6 we have that $\tau_n \to \tau$. $(x, y) \mapsto x'\tilde{\mathcal{I}}_\gamma y$ defines a positive-semidefinite Hermitian form over $\mathbb{R}^{d_\theta}$ and thus induces a semi-norm $\|x\| := \sqrt{x'\tilde{\mathcal{I}}_\gamma x}$. Hence by the reverse triangle inequality

$$|\|\tau_n\| - \|\tau\|| \le \|\tau_n - \tau\| \to 0.$$

That is $\|\tau_n\| \to \|\tau\|$ and hence by the continuity of $x \mapsto x^2$ we have $\tau_n\tilde{\mathcal{I}}_\gamma\tau_n = \|\tau_n\|^2 \to \|\tau\|^2 = \tau'\tilde{\mathcal{I}}_\gamma\tau$.

[101] An explicit construction of such a sequence can be given based on the eigendecomposition of $\tilde{\mathcal{I}}_\gamma$.

Since $\overline{M}_a = \operatorname{cl} \check{M}_a$ and $t \mapsto G_t\psi$ is continuous, taking the infimum over $a' \geq a$ yields[102]

$$\inf_{t \in \check{M}_a} G_t\psi = \inf_{t \in \overline{M}_a} G_t\psi \leq \mathrm{P}\left(\chi_r^2(a) > c_{r,\alpha}\right) =: \mathcal{R}. \tag{80}$$

By proposition B.2 $(\mathscr{E}_n)_{n\in\mathbb{N}}$ converges to $\mathscr{E}$. Suppose that (17) does not hold for all sequences of asymptotically level-α tests for $H_0 : \tau = 0$ against $H_1 : \tau \in N(\tilde{\mathcal{I}}_\gamma)^\perp \setminus \{0\}$ in $\mathscr{E}_n$. Then there is such a sequence of tests $(\psi_n)_{n\in\mathbb{N}}$ and a subsequence $(n_j)_{j\in\mathbb{N}}$ such that for some $\varepsilon > 0$

$$\lim_{j\to\infty} \inf\{\pi_{n_j}(\tau, h) : (\tau, h) \in N(\tilde{\mathcal{I}}_\gamma)^\perp \times H_\eta,\ \tau'\tilde{\mathcal{I}}_\gamma\tau \geq a\} \geq \mathcal{R} + \varepsilon,$$

where $\pi_n(\tau, h) := P^n_{\gamma_n,\tau,h}\psi_n$. Since $[0,1]^{N(\tilde{\mathcal{I}}_\gamma)^\perp \times H_\eta}$ is compact in the product topology there is a subnet $(n_{j(\alpha)})_{\alpha\in A}$ of the subsequence $(n_j)_{j\in\mathbb{N}}$ and a function $\pi : \mathbb{N}(\tilde{\mathcal{I}}_\gamma)^\perp \times H_\eta \to [0,1]$ such that $\lim_{\alpha\in A} \pi_{n_{j(\alpha)}}(\tau, h) = \pi(\tau, h)$ for every $(\tau, h) \in N(\tilde{\mathcal{I}}_\gamma)^\perp \times H_\eta$. Combine this with the preceding display to conclude that for any $(\tau, h) \in N(\tilde{\mathcal{I}}_\gamma)^\perp \times H_\eta$ with $\tau'\tilde{\mathcal{I}}_\gamma\tau \geq a$ we have

$$\pi(\tau, h) = \lim_{\alpha\in A} \pi_{n_{j(\alpha)}}(\tau, h) \geq \lim_{\alpha\in A} \inf\{\pi_{n_{j(\alpha)}}(\tau, h) : (\tau, h) \in N(\tilde{\mathcal{I}}_\gamma)^\perp \times H_\eta,\ \tau'\tilde{\mathcal{I}}_\gamma\tau \geq a\} \geq \mathcal{R} + \varepsilon.$$

However, by Theorem 7.1 in van der Vaart (1991) there is a test ψ in $\mathscr{E}$ with power function π and it follows from our hypothesis that this test is of level-α, since for any $g \in \mathscr{H}_\gamma$ there is a $h \in \mathfrak{H}_\gamma$ with $B_\gamma h = g$ and so

$$G_{0,g}\psi = \pi(0, h) = \lim_{\alpha\in A} \pi_{n_{j(\alpha)}}(\tau, h) \leq \limsup_n \pi_n(\tau, h) \leq \alpha.$$

Then by the preceding two displays we have $G_{0,g}\psi \leq \alpha$ for any $(0, g) \in \mathbb{H}_\gamma$ and for any $(\tau, g) \in \check{M}_a$

$$G_{\tau,g}\psi = \pi(\tau, h_g) \geq \mathcal{R} + \varepsilon,$$

which contradicts (80).

□

Proof of corollary 3.9. By proposition 3.1 we have that for $\tau_n \to \tau$ and $h_n \to h$,

$$\sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta_0} \rightsquigarrow \mathcal{N}(\tilde{\mathcal{I}}_\gamma\tau, \tilde{\mathcal{I}}_\gamma), \text{ under } P_{\gamma_n,\tau_n,h_n}.$$

As in the proof of proposition B.3, equations (7), (8) and Lemma C.6 imply that $\|\hat{\mathcal{I}}^\dagger_{n,\theta_n} - \tilde{\mathcal{I}}^\dagger_\gamma\|_2 = o_{P_{\gamma_n}}(1)$. The mutual contiguity established in proposition 3.1 along with Le Cam's first lemma (e.g. van der Vaart, 1998, Lemma 6.4) ensures that this result and equation (6) also hold under P_{γ_n,τ_n,h_n}:

$$\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\theta_0} - \tilde{\ell}_{\gamma_n}\right] = o_{P_{\gamma_n,\tau_n,h_n}}(1) \quad \text{and} \quad \|\hat{\mathcal{I}}^\dagger_{n,\theta_0} - \tilde{\mathcal{I}}^\dagger_\gamma\|_2 = o_{P_{\gamma_n,\tau_n,h_n}}(1).$$

Write $\hat{Z}_n := \sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta_0}$. We have

$$\hat{Z}_n = \sqrt{n}\mathbb{P}_n\tilde{\ell}_{\gamma_n} + \sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\theta_0} - \tilde{\ell}_{\gamma_n}\right] \rightsquigarrow Z \sim \mathcal{N}(\tilde{\mathcal{I}}_\gamma\tau, \tilde{\mathcal{I}}_\gamma)$$

[102] The continuity of the indicated map follows directly from the fact that a Gaussian shift experiment is continuous in the total variation norm.

under P_{γ_n,τ_n,h_n}. The continuous mapping theorem and Theorem 9.2.3 of Rao and Mitra (1971) imply that

$$\hat{S}_{n,\theta_0} = \hat{Z}_n' \hat{\mathcal{I}}_{n,\theta_0}^{\dagger} \hat{Z}_n \rightsquigarrow Z' \tilde{\mathcal{I}}_{\gamma}^{\dagger} Z =: S \sim \chi_r^2(\tau' \tilde{\mathcal{I}}_{\gamma} \tau),$$

under P_{γ_n,τ_n,h_n} where $r = \operatorname{rank}(\tilde{\mathcal{I}}_{\gamma})$.

Let c_n be the $1-\alpha$ quantile of the $\chi_{r_n}^2$ distribution and c the $1-\alpha$ quantile of the χ_r^2 distribution. We have $P_{\gamma_n}\{c_n = c\} = P_{\gamma_n}\{r_n = r\} \to 1$ by assumption. This implies that $c_n - c \to 0$ in P_{γ_n}-probability and hence by the mutual contiguity and Le Cam's first lemma, also under P_{γ_n,τ_n,h_n}. By continuous mapping once more we have $\hat{S}_{n,\theta_0} - c_n \rightsquigarrow S - c$ under P_{γ_n,τ_n,h_n}. Hence by the portmanteau theorem

$$\lim_{n\to\infty} P_{\gamma_n,\tau_n,h_n}^n \phi_{n,\theta_0} = 1 - \mathrm{P}\left(\chi_r^2\left(\tau\tilde{\mathcal{I}}_{\gamma}\tau\right) \leq c\right). \tag{81}$$

For $\tau = 0$, $1 - \mathrm{P}\left(\chi_r^2(0) \leq c\right) = \alpha$; hence this test is level-α as claimed.

Let $K_a \subset M_a$ be compact and suppose $(\tau_n, h_n)_{n\in\mathbb{N}} \subset K_a$ is such that $\tau_n \to \tau$ and $h_n \to h$. Then, by equation (81) we have that

$$\lim_{n\to\infty} P_{\gamma_n,\tau_n,h_n}^n \phi_{n,\theta_0} = \mathrm{P}(\chi_r^2\left(\tau'\tilde{\mathcal{I}}_{\gamma}\tau\right) > c) \geq \mathrm{P}(\chi_r^2(a) > c) =: \mathcal{R}. \tag{82}$$

Taking a constant sequence in K_a with $\tau'\tilde{\mathcal{I}}_{\gamma}\tau = a$ we obtain from the preceding display that $\limsup_{n\to\infty} \inf_{(\tau,h)\in K_a} P_{\gamma_n,\tau,h}^n \phi_{n,\theta_0} \leq \lim_{n\to\infty} P_{\gamma_n,\tau,h}^n \phi_{n,\theta_0} = \mathcal{R}$. It follows that if equation (18) does not hold then there is a sequence $(\tau_n, h_n)_{n\in\mathbb{N}} \subset K_a$ and a subsequence $(n_j)_{j\in\mathbb{N}}$ of $(n)_{n\in\mathbb{N}}$ such that

$$\mathcal{S} = \lim_{j\to\infty} P_{\gamma_{n_j},\tau_{n_j},h_{n_j}}^{n_j} \phi_{n_j,\theta_0} < \mathcal{R}. \tag{83}$$

Take a further subsequence $(n_k)_{k\in\mathbb{N}}$ along which $\tau_n \to \tau$ and $h_n \to h$ with $(\tau, h) \in K_a$. Construct new sequences $(h_m^*)_{m\in\mathbb{N}}$ and $(\tau_m^*)_{m\in\mathbb{N}}$ as follows. For all $m \in [n_k, n_{k+1}) \cap \mathbb{N}$ for some $k \in \mathbb{N}$ put $\tau_m^* = \tau_{n_k}$ and $h_m^* = h_{n_k}$. For $m = 1, \ldots, n_1$ put $\tau_m^* = \tau_{n_1}$ and $h_m^* = h_{n_1}$. By construction we have that $\tau_m^* \to \tau$ and $h_m^* \to h$. By (82) we have that

$$\lim_{m\to\infty} P_{\gamma_m,\tau_m^*,h_m^*}^m \phi_{m,\theta_0} \geq \mathcal{R}.$$

Fix an arbitrary $\varepsilon > 0$. There is an $M \in \mathbb{N}$ such that for all $m \geq M$ we have $P_{\gamma_m,\tau_m^*,h_m^*}^m \phi_{m,\theta_0} \geq \mathcal{R} - \varepsilon/2$. Hence for any k sufficiently large that $m = n_k \geq M$ we have

$$\mathcal{S} = \mathcal{S} - P_{\gamma_{n_k},\tau_{n_k},h_{n_k}}^{n_k} \phi_{n_k,\theta_0} + P_{\gamma_m,\tau_m^*,h_m^*}^m \phi_{m,\theta_0} \geq \mathcal{S} - P_{\gamma_{n_k},\tau_{n_k},h_{n_k}}^{n_k} \phi_{n_k,\theta_0} + \mathcal{R} - \varepsilon/2.$$

This holds for all large enough k and so taking the limit first as $k \to \infty$ and then as $\varepsilon \downarrow 0$ yields that $\mathcal{S} \geq \mathcal{R}$. But this contradicts equation (83). $\square$

Proof of proposition 3.10. By proposition A.8 in van der Vaart (1988) and our assumptions

$$\begin{aligned}
\Lambda_n(\gamma_n(\tau_n, h_n), \gamma_n) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n g_n - \frac{1}{2} P_{\gamma_n} g_n^2 + o_{P_{\gamma_n}}(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\tau' \dot{\ell}_{\gamma_n} + B_{\gamma_n} h\right] - \frac{1}{2} P_{\gamma_n} \left[\tau' \dot{\ell}_{\gamma_n} + B_{\gamma_n} h\right]^2 + o_{P_{\gamma_n}}(1).
\end{aligned}$$

since $\frac{1}{2}P_{\gamma_n}\left[\tau'\dot{\ell}_{\gamma_n} + B_{\gamma_n}h\right]^2 - \frac{1}{2}P_{\gamma_n}g_n^2 = \frac{1}{2}P_{\gamma_n}\left(f_n^2 - g_n^2\right) \to 0$, where $f_n := \tau'\dot{\ell}_{\gamma_n} + B_{\gamma_n}h$, as

$$\left|P_{\gamma_n}\left(f_n^2 - g_n^2\right)\right| = \left|\|f_n\|_{P_{\gamma_n},2}^2 - \|g_n\|_{P_{\gamma_n},2}^2\right| \leq \|f_n - g_n\|_{P_{\gamma_n},2}^2 + 2\|f_n - g_n\|_{P_{\gamma_n},2}\|g\|_{P_{\gamma_n},2} \to 0,$$

as $(g_n)_{n\in\mathbb{N}}$ is uniformly square P_{γ_n}-integrable and hence $P_{\gamma_n}g_n^2 \leq M$ for some $M \in (0,\infty)$.

It remains to show that $(f_n)_{n\in\mathbb{N}}$ is uniformly square P_{γ_n}-integrable. The preceding display yields that $P_{\gamma_n}f_n^2 = P_{\gamma_n}g_n^2 - P_{\gamma_n}(g_n^2 - f_n^2) = P_{\gamma_n}g_n^2 + o(1)$. Hence there is an $N \in \mathbb{N}$ such that $n > N$ has $P_{\gamma_n}f_n^2 \leq M + 1$. It follows that $P_{\gamma_n}f_n^2 \leq K < \infty$ with $K := \max\{M+1, P_{\gamma_1}f_1^2, \ldots, P_{\gamma_N}f_N^2\}$. Let $\varepsilon > 0$ be given and note that there is a $\delta > 0$ such that if $P_{\gamma_n}(A) < \delta$ we have $P_{\gamma_n}(g_n^2\mathbf{1}_A) < \varepsilon/4$.[103] Hence

$$P_{\gamma_n}\left(f_n^2\mathbf{1}_A\right) \leq 2P_{\gamma_n}\left((f_n - g_n)^2\mathbf{1}_A\right) + 2P_{\gamma_n}\left(g_n^2\mathbf{1}_A\right) = o(1) + \frac{\varepsilon}{2}.$$

Hence there is an $N' \in \mathbb{N}$ such that for all $n \geq N'$ we have $P_{\gamma_n}(f_n^2\mathbf{1}_A) < \varepsilon$ if $P_{\gamma_n}(A) < \delta$. By Markov's inequality we have that for $K' > K/\delta$, $P_{\gamma_n}(f_n^2 > K') \leq P_{\gamma_n}f_n^2/K' \leq \delta$ and hence for all $n \geq N'$, $P_{\gamma_n}(f_n^2\mathbf{1}\{f_n^2 > K'\}) < \varepsilon$. That is, $(f_n)_{n\in\mathbb{N}}$ is asymptotically uniformly square P_{γ_n}-integrable, which implies that $(f_n)_{n\in\mathbb{N}}$ is uniformly square P_{γ_n}-integrable.[104] □

Proof of lemma 3.11. This is implied by Corollary 2.9 of Feinberg et al. (2016). □

Proof of lemma 3.12. Define Q_n, Q respectively as the pushforward measures of P_n under f_n and P under f. By the extended continuous mapping theorem of van der Vaart and Wellner (1996, Theorem 1.11.1), $Q_n \rightsquigarrow Q$ and by hypothesis,

$$\lim_{M\to\infty}\sup_{n\in\mathbb{N}}\int_{|x|>M}|x|\,\mathrm{d}Q_n(x) = \lim_{M\to\infty}\sup_{n\in\mathbb{N}}\int_{|f_n(s)|>M}|f(s)|\,\mathrm{d}P_n(s) = 0.$$

The result now follows from the equivalence of (ii) and (iii) in Proposition A.6.1 of Bickel et al. (1998). □

Proof of proposition 3.13. Throughout let $\hat{r}_n := \operatorname{rank}(\hat{M}_n)$, $r := \operatorname{rank}(M)$, $R_n := \{\hat{r}_n = r\}$ and $\lambda_l, \lambda_{n,l}, \check{\lambda}_{n,l}$ and $\hat{\lambda}_{n,l}$ respectively the l-th largest eigenvalue of M, M_n, $\check{M}_n$ and $\hat{M}_n$.

Start with the case $r = 0$. By Weyl's perturbation theorem and the fact that $M_n = 0$ for all n larger than some $N \in \mathbb{N}$,

$$P_n(R_n) = P_n\left(\max_{l=1,\ldots,L}|\check{\lambda}_{n,l}| < \nu_n\right) \geq P_n(\|\check{M}_n - M_n\|_2 < \nu_n) \to 1.$$

On the sets R_n we have that $\hat{M}_n = 0 = M$ and so $\hat{M}_n \xrightarrow{P_n} M$ as $P(R_n) \to 1$.

Now suppose that $r > 0$. let $\underline{\nu} := \lambda_r/2 > 0$ and note that (20) implies that $\|\check{M}_n - M_n\|_2 = o_{P_n}(1)$ and so, by Weyl's perturbation theorem (e.g. Bhatia, 1997, Corollary III.2.6),

[103] Given $\varepsilon > 0$, take $M < \infty$ large enough that $P_n(g_n^2\mathbf{1}\{g_n^2 > M\}) < \varepsilon/8$ for all $n \in \mathbb{N}$ and let $\delta < \varepsilon/(8M)$. Then if $P_{\gamma_n}(A) < \delta$ we have

$$P_\gamma(g_n^2\mathbf{1}_A) \leq P_{\gamma_n}(g_n^2\mathbf{1}_A\mathbf{1}\{g_n^2 \leq M\}) + P_{\gamma_n}(g_n^2\mathbf{1}_A\mathbf{1}\{g_n^2 > M\}) \leq MP_{\gamma_n}(A) + P_n(g_n^2\mathbf{1}\{g_n^2 > M\}) < \varepsilon/4.$$

[104] Increase K' to K'' as necessary to ensure that also $P_{\gamma_n}(f_n^2\mathbf{1}\{f_n^2 > K''\}) < \varepsilon$ for all $1 \leq n < N'$.

$\max_{l=1,\dots,L} |\check{\lambda}_{n,l} - \lambda_{n,l}| \le \|\check{M}_n - M_n\|_2 = o_{P_n}(1)$. Hence, defining $E_n := \{\check{\lambda}_{n,r} \ge \nu_n\}$, for n large enough such that $\nu_n < \underline{\nu}$ and $\|M_n - M\|_2 < \underline{\nu}/2$ we have

$$P_n(E_n) = P_n\left(\check{\lambda}_{n,r} \ge \nu_n\right) \ge P_n\left(\check{\lambda}_{n,r} \ge \underline{\nu}\right) \ge P_n\left(|\check{\lambda}_{n,r} - \lambda_{n,r}| < \underline{\nu}/2\right) \to 1.$$

If $r = L$ we have that $R_n \supset E_n$ and therefore $P_n(R_n) \to 1$. Additionally, if $\check{\lambda}_{n,L} \ge \nu_n$ then $\hat{\lambda}_{n,l} = \check{\lambda}_{n,l}$ for each $l \in [L]$ and hence $\hat{M}_n = \check{M}_n$, implying $\|\hat{M}_n - M\|_2 \le \|\check{M}_n - M_n\|_2 + \|M_n - M\|_2 = o_{P_n}(1)$.

Now suppose instead that $r < L$ and define $F_n := \{\check{\lambda}_{n,r+1} < \nu_n\}$. It follows by Weyl's perturbation theorem and the fact that $\lambda_{n,l} = 0$ for $l > r$ and $n \ge N$ that as $n \to \infty$

$$P_n(F_n) = P_n(\check{\lambda}_{n,r+1} < \nu_n) \ge P_n(\|\check{M}_n - M_n\|_2 < \nu_n) \to 1.$$

Since $R_n \supset E_n \cap F_n$, this implies that $P_n(R_n) \to 1$ as $n \to \infty$. Additionally, if $\check{\lambda}_{n,r} \ge \nu_n$, $\check{\lambda}_{n,r+1} < \nu_n$ and $\|\check{M}_n - M\|_2 \le \upsilon$, we have that $\hat{\lambda}_{n,k} = \check{\lambda}_{n,k}$ for $k \le r$ and $\hat{\lambda}_{n,l} = 0 = \lambda_l$ for $l > r$ and so

$$\|\Lambda_n(\nu_n) - \Lambda\|_2 = \max_{l=1,\dots,r} |\hat{\lambda}_{n,l} - \lambda_l| = \max_{l=1,\dots,r} |\check{\lambda}_{n,l} - \lambda_l| \le \|\check{\Lambda}_n - \Lambda\|_2 \le \|\check{M}_n - M\|_2 \le \upsilon,$$

and hence $\{\|\check{M}_n - M\|_2 \le \upsilon\} \cap E_n \cap F_n \subset \{\|\Lambda_n(\nu_n) - \Lambda\|_2 \le \upsilon\}$, from which it follows that $\Lambda_n(\nu_n) \xrightarrow{P_n} \Lambda$ as $\|\check{M}_n - M\|_2 \le \|\check{M}_n - M_n\|_2 + \|M_n - M\|_2 \xrightarrow{P_n} 0$. Suppose that $(\lambda_1, \dots, \lambda_r)$ consists of s distinct eigenvalues with values $\lambda^1 > \lambda^2 > \dots > \lambda^s$ and multiplicities $\mathfrak{m}_1, \dots, \mathfrak{m}_s$ (each at least one).[105] $\lambda^{s+1} = 0$ 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 $\Pi_k := \sum_{i=1}^{\mathfrak{m}_k} u_{l_i^k} u'_{l_i^k}$.[106] Total eigenprojections are continuous.[107] Therefore, if we construct $\Pi_{n,k}$ in in an analogous fashion to Π_k but replace columns of U with columns of $\check{U}_n$, we have $\Pi_{n,k} \xrightarrow{P_n} \Pi_k$ for each $k \in [s+1]$ since $\check{M}_n \xrightarrow{P_n} M$. Spectrally decompose M as $M = \sum_{k=1}^{s} \lambda^k \Pi_k$, where the sum runs to s rather than $s+1$ since $\lambda^{s+1} = 0$. Then,

$$\hat{M}_n = \sum_{k=1}^{s+1} \sum_{i=1}^{\mathfrak{m}_k} \hat{\lambda}_{n,l_i^k} u_{n,l_i^k} u'_{n,l_i^k} = \sum_{k=1}^{s+1} \sum_{i=1}^{\mathfrak{m}_k} (\hat{\lambda}_{n,l_i^k} - \lambda^k) u_{n,l_i^k} u'_{n,l_i^k} + \sum_{k=1}^{s} \lambda^k \Pi_{n,k},$$

whence

$$\|\hat{M}_n - M\|_2 \le \sum_{k=1}^{s+1} \sum_{i=1}^{\mathfrak{m}_k} |\hat{\lambda}_{n,l_i^k} - \lambda^k| \|u_{n,l_i^k} u'_{n,l_i^k}\|_2 + \sum_{k=1}^{s} |\lambda^k| \|\Pi_{n,k} - \Pi_k\|_2 \xrightarrow{P_n} 0,$$

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

Proof of corollary 3.14. Apply proposition 3.13 with $\check{\mathcal{I}}_{n,\theta_n} = \check{M}_n$, $\hat{\mathcal{I}}_{n,\theta_n} = \hat{M}_n$, $\tilde{\mathcal{I}}_n = M_n$, $\tilde{\mathcal{I}}_\gamma = M$ and $P_{\gamma_n} = P_n$. □

[105] The superscripts on the λs are indices, not exponents.
[106] See e.g Chapter 8.8 of Magnus and Neudecker (2019).
[107] E.g. Theorem 8.7 of Magnus and Neudecker (2019).

B.2 Additional miscellaneous results

Lemma B.7. *Suppose that assumption M holds and assumptions LAN and CM(i) hold along a convergent sequence $(\gamma_n)_{n\in\mathbb{N}}$ with $\gamma_n := (\theta_n, \eta) \to \gamma \in \Gamma$,that $\eta = (\eta_1, \eta_2)$ with $\eta_1 \in \mathcal{H}_1 \subset \mathbb{R}^{d_{\eta_1}}$ and that the efficient score function takes the form*

$$\tilde{\ell}_{\gamma_n} = \breve{\ell}_{\gamma_n,1} - \breve{I}_{\gamma_n,12}\breve{I}^{-1}_{\gamma_n,22}\breve{\ell}_{\gamma_n,2}, \quad \breve{I}_{\gamma_n} := P_{\gamma_n}\breve{\ell}_{\gamma_n}\breve{\ell}'_{\gamma_n},$$

for a L-dimensional vector of functions $\breve{\ell}_{\gamma_n} := \left(\breve{\ell}'_{\gamma_n,1}, \breve{\ell}'_{\gamma_n,2}\right)'$. Suppose that $\tilde{\mathcal{I}}_{\gamma_n} \to \tilde{\mathcal{I}}_\gamma$ and $\operatorname{rank}(\tilde{\mathcal{I}}_{\gamma_n}) = \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$ for all sufficiently large $n \in \mathbb{N}$. Moreover, suppose that along any sequence $(\gamma'_n)_{n\in\mathbb{N}}$ with $\gamma'_n := (\theta_n, (\eta_{n,1}, \eta_2)) \to \gamma$ where $\sqrt{n}\|\eta_{n,1} - \eta_1\| = O(1)$,

(i) $P_{\gamma'_n}\breve{\ell}_{\gamma'_n} = o(n^{-1/2})$,

(ii) $(\|\breve{\ell}_{\gamma_n}\|_2^2)_{n\in\mathbb{N}}$ *is uniformly* $P_{\gamma'_n}$*-integrable,*

(iii) $\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\xi_n} - \breve{\ell}_{\gamma'_n}\right] = o_{P_{\gamma'_n}}(1)$,

(iv) $\nu_n^{-1}\|\hat{I}_{n,\xi_n} - \breve{I}_{\gamma_n}\|_2 = o_{P_{\gamma'_n}}(1)$,

(v) $\int\left[\breve{\ell}_{\gamma'_n,l}\sqrt{p_{\gamma'_n}} - \breve{\ell}_{\gamma_n,l}\sqrt{p_{\gamma_n}}\right]^2 \mathrm{d}\nu \to 0$ *for each* $l \in [L]$,

with $\xi_n := (\theta_n, \eta_{n,1})$. Finally suppose that $\hat{\eta}_{n,1}$ satisfies $\sqrt{n}\|\hat{\eta}_{n,1} - \eta_1\| = O_{P_{\gamma_n}}(1)$. Then if $\bar{\xi}_n := (\theta_n, \bar{\eta}_{n,1})$ where $\bar{\eta}_{n,1}$ is the version of $\hat{\eta}_{n,1}$ discretised on $n^{-1/2}C\mathbb{Z}^{d_{\eta_1}} \cap \mathcal{H}_1$,

$$\hat{\ell}_{n,\theta_n} := \hat{\ell}_{n,\bar{\xi}_n,1} - \hat{I}_{n,\bar{\xi}_n,12}\hat{I}^{-1}_{n,\bar{\xi}_n,22}\hat{\ell}_{n,\bar{\xi}_n,2}, \quad \check{\mathcal{I}}_{n,\theta_n} := \hat{I}_{n,\bar{\xi}_n,11} - \hat{I}_{n,\bar{\xi}_n,12}\hat{I}^{-1}_{n,\bar{\xi}_n,22}\hat{I}_{n,\bar{\xi}_n,21}, \tag{84}$$

and $\hat{\mathcal{I}}_{n,\theta_n}$ is the eigendecomposition-truncated version of $\check{\mathcal{I}}_{n,\theta_n}$ at ν_n analogously to (21), then assumptions E and R hold.

Proof. Define $b_n := \sqrt{n}(\eta_{n,1} - \eta_1)$. Take an arbitrary subsequence $(n_m)_{m\in\mathbb{N}}$ of $(n)_{n\in\mathbb{N}}$ and a further subsequence $(n_k)_{k\in\mathbb{N}}$ along which $b_{n_k} \to b \in \mathbb{R}^{d_{\eta_1}}$. Construct a "full" sequence $(b^\star_n)_{n\in\mathbb{N}}$ according to $b^\star_{n_k} := b_{n_k}$ for all $k \in \mathbb{N}$ and for all $m \in \mathbb{N}$ such that $m \notin \{n_k : k \in \mathbb{N}\}$ set $b^\star_m := b^\star_{m-1}$ (arbitrarily put $b_0 = 0$). Constructed in this manner $b^\star_n \to b$ as $n \to \infty$ and hence $\beta^\star_{n,1} := \eta + \sqrt{n}b^\star_n$ is a deterministic sequence satisfying $\sqrt{n}(\eta^\star_{n,1} - \eta) = O(1)$. Note that we can write $\gamma^\star_n := (\theta_n, (\eta^\star_{n,1}, \eta_2))$ as $\gamma^\star_n = \gamma_n(0, h^\star_n)$ for $h^\star_n := (b^\star_n, 0)$. Since conditions (i) - (v) are valid along $(\gamma'_n)_{n\in\mathbb{N}}$ formed with an arbitrary deterministic $\sqrt{n}$-consistent sequence $(\eta_{n,1})_{n\in\mathbb{N}}$, they apply along $(\gamma^\star_n)_{n\in\mathbb{N}}$ in particular. Since LAN holds, these observations, in conjunction with Proposition A.10 in van der Vaart (1988) yield that

$$\sqrt{n}\mathbb{P}_n\left[\breve{\ell}_{\gamma^\star_n} - \breve{\ell}_{\gamma_n}\right] + \breve{I}_{\gamma_n}(0', (b^\star_n)')' = o_{P_{\gamma_n}}(1).$$

This clearly implies also that

$$\sqrt{n_k}\mathbb{P}_{n_k}\left[\breve{\ell}_{\gamma'_{n_k}} - \breve{\ell}_{\gamma_{n_k}}\right] + \breve{I}_{\gamma_{n_k}}(0', b'_{n_k})' = \sqrt{n_k}\mathbb{P}_{n_k}\left[\breve{\ell}_{\gamma^\star_{n_k}} - \breve{\ell}_{\gamma_{n_k}}\right] + \breve{I}_{\gamma_{n_k}}(0', (b^\star_{n_k})')' = o_{P_{\gamma_{n_k}}}(1),$$

and therefore, as the original subsequence $(n_m)_{m\in\mathbb{N}}$ was arbitrary,

$$\sqrt{n}\mathbb{P}_n\left[\breve{\ell}_{\gamma'_n} - \breve{\ell}_{\gamma_n}\right] + \sqrt{n}\breve{I}_{\gamma_n}(0', (\eta_{n,1} - \eta)')' = o_{P_{\gamma_n}}(1). \tag{85}$$

Moreover we have by Proposition 3.1 that $(P^n_{\gamma_n})_{n\in\mathbb{N}}$ and $(P^n_{\gamma^\star_n})_{n\in\mathbb{N}}$ are mutually contiguous.

Hence the same is true of $(P^{n_k}_{\gamma_{n_k}})_{k\in\mathbb{N}}$ and $(P^{n_k}_{\gamma^\star_{n_k}})_{k\in\mathbb{N}} = (P^{n_k}_{\gamma'_{n_k}})_{k\in\mathbb{N}}$. This observation in conjunction with (iii), (iv) and the fact that our initial subsequence $(n_m)_{m\in\mathbb{N}}$ was arbitrary yields the conclusion that

$$\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\xi_n} - \breve{\ell}_{\gamma'_n}\right] = o_{P_{\gamma_n}}(1), \quad \text{and} \quad \left\|\hat{I}_{n,\xi_n} - \breve{I}_{\gamma_n}\right\|_2 = o_{P_{\gamma_n}}(\nu_n). \tag{86}$$

Now, for $\eta_1^\sharp \in \mathcal{H}_1$ let

$$R_{1,n}(\eta_1^\sharp) := \sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\xi_n^\sharp} - \breve{\ell}_{\gamma_n}\right] + \sqrt{n}\breve{I}_{\gamma_n}(0', (\eta_1^\sharp - \eta)'), \quad R_{2,n}(\eta_1^\sharp) := \nu_n^{-1}\left[\hat{I}_{n,\xi_n^\sharp} - \breve{I}_{\gamma_n}\right]$$

where $\xi_n^\sharp := (\theta_n, \eta_1^\sharp)$, $\eta^\sharp := (\eta_1^\sharp, \eta_2)$ and $\gamma_n^\sharp := (\theta_n, \eta^\sharp)$. Let As $\bar{\beta}_n$ is discretised on $n^{-1/2}C\mathbb{Z}^{d_{\eta_1}} \cap \mathcal{H}_1$ from $\hat{\eta}_{1,n}$ it remains $\sqrt{n}$-consistent under P_{γ_n} and hence for any $\varepsilon > 0$ there is an $M \in (0, \infty)$ and N such that for all $n \geq N$, $P_{\gamma_n}(\sqrt{n}\|\bar{\eta}_{n,1} - \eta_1\|_2 > M) < \varepsilon$. If $\sqrt{n}\|\bar{\eta}_{n,1} - \eta_1\|_2 \leq M$ then $\bar{\eta}_{n,1} \in \mathfrak{S}_n := \{\eta_1^\flat \in n^{-1/2}C\mathbb{Z}^{d_{\eta_1}} \cap \mathcal{H}_1 : \|\eta_1^\flat - \eta_1\|_2 \leq M/\sqrt{n}\}$. For any fixed M, $\mathfrak{S}_n$ has a finite number of elements bounded independently of n, call this number $\overline{\mathfrak{S}}$. For $R_n \in \{R_{1,n}, R_{2,n}\}$, any $\upsilon > 0$ and $n \geq N$

$$\begin{aligned}
P_{\gamma_n}(\|R_n(\bar{\eta}_{n,1})\| > \upsilon) &\leq \varepsilon + \sum_{\eta_{n,1}\in\mathfrak{S}_n} P_{\gamma_n}(\{\|R_n(\eta_{n,1})\| > \upsilon\} \cap \{\bar{\eta}_{n,1} = \eta_{n,1}\}) \\
&\leq \varepsilon + \overline{\mathfrak{S}} P_{\gamma_n}\left(\|R_n(\eta_{n,1}^*)\| > \upsilon\right),
\end{aligned}$$

where $\eta_{n,1}^* \in \mathfrak{S}_n$ maximises $\eta_1 \mapsto P_{\gamma_n}(\|R_n(\eta_1)\| > \upsilon)$. Since $(\eta_{n,1}^*)_{n\in\mathbb{N}}$ is deterministic and $\sqrt{n}$-consistent for η_1, $P_{\gamma_n}\left(\|R_n(\eta_{n,1}^*)\| > \upsilon\right) \to 0$ by equations (85) & (86). It follows that $\|R_{i,n}(\bar{\eta}_{n,1})\| = o_{P_{\gamma_n}}(1)$ for $i \in \{1,2\}$. It follows that $\|\hat{\mathcal{K}}_{\bar{\xi}_n} - \tilde{\mathcal{K}}_{\gamma_n}\|_2 \xrightarrow{P_{\gamma_n}} 0$ where

$$\tilde{\mathcal{K}}_{\gamma_n} := \left[I \;\; -\breve{I}_{\gamma_n,12}\breve{I}^{-1}_{\gamma_n,22}\right], \quad \hat{\mathcal{K}}_{\bar{\xi}_n} := \left[I \;\; -\hat{I}_{n,\bar{\xi}_n,12}\hat{I}^{-1}_{n,\bar{\xi}_n,22}\right],$$

with the partitions of the matrices $\hat{I}_{\bar{\xi}_n}$, $\breve{I}_{\gamma_n}$ corresponds to the partition of the vectors $\hat{\ell}_{n,\bar{\xi}_n} = (\hat{\ell}'_{n,\bar{\xi}_n,1}, \hat{\ell}'_{n,\bar{\xi}_n,2})'$, $\breve{\ell}_{\gamma_n} = (\tilde{\ell}'_{\gamma_n,1}, \tilde{\ell}'_{\gamma_n,2})'$, $\bar{\xi}_n := (\theta_n, \bar{\eta}_{n,1})$ and $\breve{I}^{-1}_{\gamma_n,22}$ exists by assumption. Using these results, (84) and the uniform P_{γ_n}-integrability of $\|\breve{\ell}_{\gamma_n}\|_2^2$,

$$\begin{aligned}
&\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\theta_n} - \tilde{\ell}_{\gamma_n}\right] \\
&\quad = \left(\hat{\mathcal{K}}_{\bar{\xi}_n} - \tilde{\mathcal{K}}_{\gamma_n}\right)\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\bar{\xi}_n} - \breve{\ell}_{\gamma_n}\right] + \tilde{\mathcal{K}}_{\gamma_n}\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\bar{\xi}_n} - \breve{\ell}_{\gamma_n}\right] + \left(\hat{\mathcal{K}}_{\bar{\xi}_n} - \tilde{\mathcal{K}}_{\gamma_n}\right)\sqrt{n}\mathbb{P}_n\breve{\ell}_{\gamma_n} \\
&\quad = -\left[I \;\; -\breve{I}_{\gamma_n,12}\breve{I}^{-1}_{\gamma_n,22}\right]\begin{bmatrix}\breve{I}_{\gamma_n,11} & \breve{I}_{\gamma_n,12} \\ \breve{I}_{\gamma_n,21} & \breve{I}_{\gamma_n,22}\end{bmatrix}\begin{bmatrix}0 \\ \sqrt{n}(\bar{\eta}_{n,1} - \eta_1)\end{bmatrix} + o_{P_{\gamma_n}}(1) \\
&\quad = o_{P_{\gamma_n}}(1),
\end{aligned}$$

which gives (6). To show that equation (7) and assumption R hold, Corollary 3.14 indicates that it suffices to show that the requirements of assumption T are satisifed. For this note that by assumption $\tilde{\mathcal{I}}_{\gamma_n} \to \tilde{\mathcal{I}}_\gamma$ with $\operatorname{rank}(\tilde{\mathcal{I}}_{\gamma_n}) = \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$ for all sufficiently large $n \in \mathbb{N}$ and (23) follows from $\|R_{2,n}(\bar{\eta}_{n,1})\| = o_{P_{\gamma_n}}(1)$. □

B.3 Proofs for section 4

Throughout this section I use the notation $\iota(\theta, X) := X_1 + X_2'\theta$.

Proof of Proposition 4.1. Fix arbitrary $\tau_n \to \tau \in \mathbb{R}^{d_\theta}$ and $h_n \to h \in H_\eta$. The perturbed law is P_{γ_n,τ_n,h_n} with density

$$p_{\gamma_n,\tau_n,h_n}(W) := \zeta(e_n, X)(1 + h_{n,2}(e_n, X)/\sqrt{n}),$$

where $e_n := Y - f(\iota(\theta_n + n^{-1/2}\tau_n, X)) - n^{-1/2}h_{n,1}(\iota(\theta_n + n^{-1/2}\tau_n, X))$. Since Θ is are open and $\theta_n \to \theta$, $\theta_n + n^{-1/2}\tau_n \in \Theta$ for all large enough $n \in \mathbb{N}$. The restrictions on $\dot{\mathscr{F}}$ ensure that $f + n^{-1/2}h_{n,1} \in \mathscr{F}$. The restrictions on $\dot{\mathscr{Z}}_\eta$ along with the norm on H suffice to ensure that $\zeta(1 + h_{n,2}/\sqrt{n}) \in \mathscr{Z}$. Specifically, for all large enough n, $\zeta(1 + h_{n,2}/\sqrt{n}) \geq 0$ (λ-a.e.) since $h_{n,2}$ is bounded (λ-a.e.) and the conditions on $\dot{\mathscr{Z}}$ ensure that $\int \zeta(1+h_{n,2}/\sqrt{n})\,\mathrm{d}\lambda = \int \zeta\,\mathrm{d}\lambda + \frac{1}{\sqrt{n}}\int h_{n,2}\zeta\,\mathrm{d}\lambda = 1$. Continuous differentiability (λ-a.e.) of $e \mapsto \sqrt{\zeta(1 + h_{n,2}/\sqrt{n})(e, X)}$ follows from the same requirement on $\sqrt{\zeta}$ and $h_{n,2}$, the boundedness of $h_{n,2}$ (which ensures that eventually $1 + h_{n,2}/\sqrt{n}$ is bounded away from zero λ-a.e.) and the chain rule. Finally it remains to check the conditions in (26). For any $A \in \sigma(Z)$, letting G denote the measure corresponding to ζ

$$\begin{aligned}\int_A \epsilon\zeta(\epsilon, X)(1 + h_{n,2}(\epsilon, X)/\sqrt{n})\,\mathrm{d}\lambda &= \int_A \epsilon\,\mathrm{d}G + \frac{1}{\sqrt{n}}\int_A \epsilon h_{n,2}(\epsilon, X)\,\mathrm{d}G \\ &= \int_A \mathbb{E}[\epsilon|X]\,\mathrm{d}G + \frac{1}{\sqrt{n}}\int_A \mathbb{E}[\epsilon h_{n,2}(\epsilon, X)|X]\,\mathrm{d}G \\ &= 0,\end{aligned}$$

and hence $\mathbb{E}[\epsilon|X] = 0$ (a.s. under $\zeta(1 + h_{n,2}/\sqrt{n})$). For the rest, firstly let $m(\epsilon, X)$ be non-negative and integrable under G. By the (λ-a.e.) boundedness of $h_{n,2}$ (by $\bar{h}_2$, say)

$$\int m(\epsilon, X)\zeta(\epsilon, X)(1 + h_{n,2}(\epsilon, X)/\sqrt{n})\,\mathrm{d}\lambda \leq \left(1 + \frac{\bar{h}_2}{\sqrt{n}}\right)\int m(\epsilon, X)\,\mathrm{d}G < \infty.$$

Secondly, note that by Jensen's inequality

$$\left\|\int XX'\zeta(1 + h_{n,2}/\sqrt{n})\,\mathrm{d}\lambda - \int XX'\,\mathrm{d}G\right\|_2 \leq \frac{\bar{h}_2}{\sqrt{n}}\left\|\int XX'\,\mathrm{d}G\right\| \leq \frac{\bar{h}_2}{\sqrt{n}}\int \|X\|_2^2\,\mathrm{d}G \to 0,$$

which implies that for all large enough n, $\int XX'\zeta(1 + h_{n,2}/\sqrt{n})\,\mathrm{d}\lambda \succ 0$.

To establish (19), first let $\gamma \in \Gamma$, $u = (\tau, h) \in \mathbb{R}^{d_\theta} \times H_\eta$, $t \in (0, \infty)$ and $\varphi := \varphi(u) := (\tau, h_1, \zeta h_2)$ and let $\Delta_\gamma(\varphi) := \frac{1}{2}[\tau\dot{\ell}_\gamma + B_\gamma h]\sqrt{p_\gamma}$. By arguing analogously to the preceding paragraph it is seen that for all t in a sufficiently small neighbourhood $\mathscr{U}$ of 0 in $[0, \infty)$, $p_{\gamma+t\varphi}$ is a probability density. $t \mapsto \sqrt{p_{\gamma+t\varphi}}$ is continuously differentiable λ-a.e. by the corresponding conditions imposed on $e \mapsto \sqrt{\zeta(e, X)}$ and $e \mapsto h_3(e, X)$. For $t \in \mathscr{U}$, define $e(t) = Y - f(\iota(\theta(t), X)) - th_1(\iota(\theta(t), X))$ with $\theta(t) := \theta + t\tau$. Define $g(t) := \frac{\partial}{\partial s}_{|s=t} \log p_{\gamma+s\varphi}$

and note

$$g(t) = -\phi(e(t), X)\,[f'(\iota(\theta(t), X))X_2'\tau + h_1(\iota(\theta(t), X)) + th_1'(\iota(\theta(t), X))X_2'\tau] \\ + \frac{h_2(e(t), X) + th_2'(e(t), X)\,[f'(\iota(\theta(t), X))X_2'\tau + h_1(\iota(\theta(t), X)) + th_1'(\iota(\theta(t), X))X_2'\tau]}{1 + th_2(e(t), X)}.$$

By taking $\mathscr{U}$ smaller if necessary suppose that $1 + th_2 > c > 0$, and $|f'|$, $|h_1|$, $|h_1'|$, $|h_2|$ and $|h_2'|$ are bounded by $C \in (0, \infty)$ λ-a.e.. Let $t_n \to t$ through $\mathscr{U}$ and note that $g(t_n) \to g(t)$ λ-a.e. by the continuity and continuous differentiability assumptions. For any $t \in \mathscr{U}$

$$\int |g(t)|^{2+\rho}\,\mathrm{d}P_{\gamma+t\varphi} \lesssim \int (\phi(\epsilon, X)^{2+\rho} + 1)\|X\|_2^{2+\rho}\zeta(\epsilon, X)\,\mathrm{d}\lambda < \infty,$$

which can be used in conjunction with Markov's inequality to obtain the uniform $P_{\gamma+t_n\varphi}$-integrability of $(g(t_n)^2)_{n\in\mathbb{N}}$. Since also $p_{\gamma+t_n\varphi} \to p_{\gamma+t\varphi}$ λ-a.e. as is easily verified by inspection, Lemma 3.11 implies that $\int g(t_n)^2\,\mathrm{d}P_{\gamma+t_n\varphi} \to \int g(t)^2\,\mathrm{d}P_{\gamma+t\varphi}$. By Lemma 1.8 in van der Vaart (2002)

$$\lim_{t\downarrow 0}\left\|\frac{\sqrt{p_{\gamma+t\varphi}} - \sqrt{p_\gamma}}{t} - \Delta_\gamma(\varphi)\right\|_{\lambda,2} = 0. \tag{87}$$

Next let $(\delta_n)_{n\in\mathbb{N}} \subset [0, 1]$ be an arbitrary sequence, $t_n \downarrow 0$ and define $\boldsymbol{\gamma}_n := \gamma_n + \delta_n t_n \varphi_n$ for $\varphi_n := \varphi(u_n)$ with $u_n \to u \in \mathbb{R}^{d_\theta} \times H_\eta$. Define $\tilde{e}_n := Y - f(\iota(\tilde{\theta}_n, X)) - \delta_n t_n h_{n,1}(\iota(\tilde{\theta}_n, X))$ with $\tilde{\theta}_n := \theta_n + \delta_n t_n \tau_n$,

$$\phi_n := \phi(\tilde{e}_n, X) + \frac{\delta_n t_n h_{n,2}'(\tilde{e}_n, X)}{1 + \delta_n t_n h_{n,2}(\tilde{e}_n, X)}.$$

Then, $\Delta_{\boldsymbol{\gamma}_n}(\varphi_n) := \frac{1}{2}[\tau_n'\dot{\ell}_{\boldsymbol{\gamma}_n} + B_{\boldsymbol{\gamma}_n}h_n]\sqrt{p_{\boldsymbol{\gamma}_n}}$, with

$$\begin{aligned} p_{\boldsymbol{\gamma}_n}(W) &= \zeta(\tilde{e}_n, X)(1 + \delta_n t_n h_{n,2}(\tilde{e}_n, X)) \\ \dot{\ell}_{\boldsymbol{\gamma}_n}(W) &= -\phi_n f'(\iota(\tilde{\theta}_n, X))X_2 \\ [B_{\boldsymbol{\gamma}_n}h](W) &= -\phi_n h_{n,1}(\iota(\tilde{\theta}_n, X)) + h_{n,2}(\tilde{e}_n, X). \end{aligned}$$

It may be verified by inspection that $\Delta_{\boldsymbol{\gamma}_n}(\varphi_n) \to \Delta_\gamma(\varphi)$ λ-a.e. under our assumptions. Argue analogously to the demonstration that $\int g(t_n)^2\,\mathrm{d}P_{\gamma+t_n\varphi} \to \int g(t)^2\,\mathrm{d}P_{\gamma+t\varphi}$ above to conclude $\|\Delta_{\boldsymbol{\gamma}_n}(\varphi_n)\|_{\lambda,2}^2 \to \|\Delta_\gamma(\varphi)\|_{\lambda,2}^2$ and hence by Proposition 2.29 in van der Vaart (1998),

$$\|\Delta_{\boldsymbol{\gamma}_n}(\varphi_n) - \Delta_\gamma(\varphi)\|_{\lambda,2} \to 0. \tag{88}$$

Now we establish (19). First suppose that $\theta_n = \theta$ for all $n \in \mathbb{N}$, let $u_n \to u$ be arbitrary, put $\varphi_n := \varphi(u_n)$, $\varphi := \varphi(u)$ and $t_n \downarrow 0$. For all large enough n, $\gamma + t_n\varphi_n \in \Gamma$ and so using (87) and the mean-value theorem (e.g. Drabek and Milota, 2007, Theorem 3.2.7), for such n

$$\begin{aligned}\left\|\frac{\sqrt{p_{\gamma+t_n\varphi_n}}-\sqrt{p_\gamma}}{t_n}-\Delta_\gamma(\varphi)\right\|_{\lambda,2} &\le \left\|\frac{\sqrt{p_{\gamma+t_n\varphi_n}}-\sqrt{p_{\gamma+t_n\varphi}}}{t_n}\right\|_{\lambda,2}+\left\|\frac{\sqrt{p_{\gamma+t_n\varphi}}-\sqrt{p_\gamma}}{t_n}-\Delta_\gamma(\varphi)\right\|_{\lambda,2} \\ &\le \sup_{\delta\in[0,1]}\left\|\Delta_{\gamma+\delta t_n(\varphi_n-\varphi)(\varphi_n-\varphi)}\right\|_{\lambda,2}+o(1) \\ &= o(1),\end{aligned} \tag{89}$$

where the last step uses that for any sequence $(\delta_n)_{n\in\mathbb{N}}\subset[0,1]$, $\|\Delta_{\gamma+\delta_n t_n(\varphi_n-\varphi)}(\varphi_n-\varphi)-\Delta_\gamma(0)\|_{\lambda,2}\to 0$ by (88) and $\Delta_\gamma(0)=0$. Now consider an arbitrary sequence $\theta_n\to\theta$ and $\gamma_n=(\theta_n,\eta)$. Using (89) and applying the mean-value theorem at each $n\in\mathbb{N}$ gives

$$\begin{aligned}\left\|\frac{\sqrt{p_{\gamma_n+t_n\varphi_n}}-\sqrt{p_{\gamma_n}}}{t_n}-\Delta_{\gamma_n}(\varphi)\right\|_{\lambda,2} &\le |t_n^{-1}|\sup_{\delta\in[0,1]}\|\Delta_{\gamma_n+\delta_n t_n\varphi_n}(t_n\varphi_n)-t_n\Delta_{\gamma_n}(\varphi)\|_{\lambda,2} \\ &= \sup_{\delta\in[0,1]}\|\Delta_{\gamma_n+\delta_n t_n\varphi_n}(\varphi_n)-\Delta_{\gamma_n}(\varphi)\|_{\lambda,2}.\end{aligned}$$

By (88), for some sequence $(\delta_n)_{n\in\mathbb{N}}\subset[0,1]$[108]

$$\begin{aligned}&\limsup_{n\to\infty}\sup_{\delta\in[0,1]}\|\Delta_{\gamma_n+\delta t_n\varphi_n}(\varphi_n)-\Delta_{\gamma_n}(\varphi)\|_{\lambda,2} \\ &\quad\le \limsup_{n\to\infty}\|\Delta_{\gamma_n+\delta_n t_n\varphi_n}(\varphi_n)-\Delta_\gamma(\varphi)\|_{\lambda,2}+\limsup_{n\to\infty}\|\Delta_{\gamma_n}(\varphi)-\Delta_\gamma(\varphi)\|_{\lambda,2} \\ &\quad= o(1).\end{aligned}$$

Combine the two preceding displays and take $t_n=n^{-1/2}$ to yield (19):

$$\left\|\sqrt{n}\left(\sqrt{p_{\gamma_n,\tau_n,h_n}}-\sqrt{p_{\gamma_n}}\right)-\frac{1}{2}g_n\sqrt{p_{\gamma_n}}\right\|_{\lambda,2}=\left\|\frac{\sqrt{p_{\gamma_n+t_n\varphi_n}}-\sqrt{p_\gamma}}{t_n}-\Delta_{\gamma_n}(\varphi)\right\|_{\lambda,2}=o(1).$$

To conclude we note that Lemma 1.8 in van der Vaart (2002) along with (87) applied for each γ_n separately yields that $P_{\gamma_n}g_n=0$. The uniform square P_{γ_n}-integrability of g_n follows by Lemma C.8 on noting that by (88) (applied with $\delta_n=t_n=0$ and $u_n=0$) $P_{\gamma_n}g_n^2\to P_\gamma g^2$ (where $g:=\tau\dot{\ell}_\gamma+B_\gamma h$), and $p_{\gamma_n}\to p_\gamma$ λ-a.e.. Linearity of each B_{γ_n} is clear. □

Lemma B.8. *In the setting of Proposition 4.2, let G be the measure on $\mathbb{R}^{1+K}$ corresponding to ζ and $U=(\epsilon,X)\sim\zeta$. Let $\mathscr{N}:=\left\{-\phi(\epsilon,X)h_1(\iota(\theta,X))+h_2(\epsilon,X):h_1\in\dot{\mathscr{F}},h_2\in\dot{\mathscr{Z}}_\eta\right\}$. The closed linear span of $\mathscr{N}$ in $L_2(G)$ is*

$$\overline{\operatorname{lin}}\,\mathscr{N}=\{q\in L_2(G):\mathbb{E}[q(U)]=0,\ \mathbb{E}[\epsilon q(U)|X]=\mathbb{E}[\epsilon q(U)|\iota(\theta,X)]\}.$$

Proof.[109] Let $h_1\in\dot{\mathscr{F}}$ and $h_2\in\dot{\mathscr{Z}}_\eta$. The definition of the sets $\dot{\mathscr{F}}$, $\dot{\mathscr{Z}}_\eta$ and (26) ensure that $\mathscr{N}\subset L_2(G)$. Taking $h_1=0$ and $h_2=0$, we have that $\mathbb{E}[-\phi(\epsilon,X)h_1(\iota(\theta,X))]=0$ by Proposition 4.1. $\mathbb{E}[h_2(\epsilon,X)]=0$ by definition. Additionally, we have by (28)

$$\mathbb{E}[-\epsilon\phi(U)h_1(\iota(\theta,X))+\epsilon h_2(U)|X]=h_1(\iota(\theta,X)),$$

[108] On the right hand side take $\varphi_n=\varphi$ and $\delta_n=0$.

[109] Cf. the proof of Lemma A.1 in Newey and Stoker (1993, pp. 1219 – 1220).

and since $\sigma(\iota(\theta, X)) \subset \sigma(X)$, by (28) and the law of iterated expectations

$$\mathbb{E}[-\epsilon\phi(U)h_1(\iota(\theta, X)) + \epsilon h_2(U)|\iota(\theta, X)] = h_1(\iota(\theta, X)).$$

Hence $\mathscr{N} \subset \{q \in L_2(G) : \mathbb{E}[q(U)] = 0,\ \mathbb{E}[\epsilon q(U)|X] = \mathbb{E}[\epsilon q(U)|\iota(\theta, X)]\}$. Both sets are clearly linear spaces, hence it suffices to show that the latter is the closure of the former. Suppose that $q \in \{q \in L_2(G) : \mathbb{E}[q(U)] = 0,\ \mathbb{E}[\epsilon q(U)|X] = \mathbb{E}[\epsilon q(U)|\iota(\theta, X)]\}$.

It follows from the defintion of $\tilde{m}$ that $\bar{m}(U) := \tilde{m}(\epsilon) - \mathbb{E}[\tilde{m}(\epsilon)|X]$ is bounded and $e \mapsto \bar{m}((e, X))$ is continuously differentiable with bounded derivative. For any bounded function $U \mapsto \tilde{\mathfrak{q}}(U)$ such that $e \mapsto \tilde{\mathfrak{q}}((e, X))$ is continuously differentiable with bounded deriviatives, define $\bar{\mathfrak{q}}(U) := \tilde{\mathfrak{q}}(U) - \mathbb{E}[\tilde{\mathfrak{q}}(U)|X]$ and put for a bounded function $a : \mathbb{R} \to \mathbb{R}$ where a is continuously differentiable with bounded derivative,

$$\mathfrak{q}(U) := \bar{\mathfrak{q}}(U) - \bar{m}(\epsilon) \left[\mathbb{E}[\bar{m}(\epsilon)\epsilon|X]\right]^{-1} \left[\mathbb{E}[\bar{\mathfrak{q}}(U)\epsilon|X] - a(\iota(\theta, X))\right].$$

By construction, $\mathfrak{q}$ is bounded, $e \mapsto \mathfrak{q}((e, X))$ is continuously differentiable with bounded derivative, $\mathbb{E}\left[\mathfrak{q}(U)|X\right] = 0$ and $\mathbb{E}\left[\epsilon\mathfrak{q}(U)|X\right] = 0$. Hence $\mathfrak{q} \in \dot{\mathscr{Z}}_\eta$. For any $\varepsilon > 0$, by Lemma C.7 of Newey (1991), there are $\tilde{\mathfrak{q}}$, a and ψ such that $\tilde{\mathfrak{q}}$ and a satisfy the conditions required for the construction of $\mathfrak{q}$ above and $\|q - \tilde{\mathfrak{q}}\|_{G,2}^2 < \varepsilon$, $\|\mathbb{E}[\epsilon q|\iota(\theta, X)] - a(\iota(\theta, X))\|_{G,2}^2 < \varepsilon$ and $\|\mathbb{E}[q|X] - \psi(X)\|_{G,2}^2 < \varepsilon$.[110] The proof is completed by arguing as in display (A.11) of Newey and Stoker (1993, p. 1220). $\square$

Proof of Proposition 4.2. Lemma B.8 establishes the closed linear span of the nuisance tangent set. The orthogonal projection (in $L_2(G)$) of a function onto the orthocomplement of this set is given by Lemma A.2 in Newey and Stoker (1993). In particular, for $U = (\epsilon, X) \sim G$ and $V_n := \iota(\theta_n, X)$, the projection $\Pi\left(-\phi(U)f'(V_n)X_2\middle|\mathscr{N}^\perp\right)$ has the form

$$\begin{aligned}
&\omega(X)\epsilon\left[\mathbb{E}[-\epsilon\phi(U)f'(V_n)X_2)|X] - \frac{\mathbb{E}\left[-\omega(X)\epsilon\phi(U)f'(V_n)X_2)|V_n\right]}{\mathbb{E}\left[\omega(X)|V_n\right]}\right] \\
&\quad = \omega(X)\epsilon f'(V_n)\left[\mathbb{E}\left[X_2\mathbb{E}\left[-\epsilon\phi(U)|X\right]|V_n\right] - \frac{\mathbb{E}\left[\omega(X)X_2\mathbb{E}\left[-\epsilon\phi(U)|X\right]|V_n\right]}{\mathbb{E}\left[\omega(X)|V_n\right]}\right] \\
&\quad = \omega(X)\epsilon f'(V_n)\left[\mathbb{E}\left[X_2|V_n\right] - \frac{\mathbb{E}\left[\omega(X)X_2|V_n\right]}{\mathbb{E}\left[\omega(X)|V_n\right]}\right],
\end{aligned}$$

where the last equality is by (28). As $(Y - f(V_n), X) \sim G$ under P_{γ_n}, the claimed form of the efficient score function follows. $\square$

Proof of Lemma 4.3. We first show that $\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) = O_{P_{\gamma_n}}(1)$. For $\tilde{\epsilon}_i^2 := \epsilon_i^2 - \sigma^2$ we have

$$\begin{aligned}
\sqrt{n}|\hat{\sigma}_n^2 - \sigma^2| &\lesssim \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\epsilon}_i^2 + \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(f(V_{n,i}) - \hat{f}_{n,i}\right)^2 \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\epsilon}_i^2 + \frac{1}{\sqrt{n}}\left[\sum_{i\in N^{(1)}} \left(f(V_{n,i}) - \hat{f}_{n,i}\right)^2 + \sum_{i\in N^{(2)}} \left(f(V_{n,i}) - \hat{f}_{n,i}\right)^2\right],
\end{aligned}$$

[110] I.e. a is bounded, continuously differentiable with bounded derivative and $\tilde{\mathfrak{q}}$ is bounded and $e \mapsto \tilde{\mathfrak{q}}((e, X))$ is continuously differentiable with bounded deriviatives.

The first right hand side term is $O_{P_{\gamma_n}}(1)$ by the CLT. Next define $\tilde{f}_{n,i} := (f(V_{n,i}) - \hat{f}_{n,i})$ and $\mathcal{C}_n := (W_j)_{j \in N_{-i}}$. On a set E_n with $P_{\gamma_n}(E_n) \to 1$ we have $\mathbb{E}[\tilde{f}_{n,i}^2 | \mathcal{C}_n] \le \mathcal{R}_{1,n,i} \le r_n^2 = o(n^{-1/2})$ and hence by Markov's inequality, the second and third terms are $o_{P_{\gamma_n}}(1)$. Finally note that

$$\sqrt{n}|\hat{\sigma}_n^{-2} - \sigma^{-2}| = \frac{\sqrt{n}|\hat{\sigma}^2 - \sigma^2|}{|\hat{\sigma}_n^2 \sigma^2|} = o_{P_{\gamma_n}}(1),$$

by $\sqrt{n}|\hat{\sigma}_n^2 - \sigma^2| = O_{P_{\gamma_n}}(1)$ and since for some $c > 0$, $\sigma^2 > c$ and with P_{γ_n}-probability approaching 1, $\hat{\sigma}_n^2 > c$ and so $1/|\hat{\sigma}_n^2 \sigma^2| = O_{P_{\gamma_n}}(1)$. $\square$

Proof of Proposition 4.4. That assumptions M, LAN and CM(ii) hold follows from Propositions 3.10, 4.1 and 4.2. We next show (6) holds. Let $\mathcal{C}_n$ be some collection of random vectors. Let $\delta_n \to 0$, $\delta_n' \to 0$. For a triangular array of random vectors $(R_{n,i})_{n \in \mathbb{N}, i \le n}$ if with P_{γ_n}-probability approaching one either (a) $\mathbb{E}[\|R_{n,i}\|_2 | \mathcal{C}_n] \le \delta_n n^{-1/2}$ or (b) for each element $R_{n,i,s}$ of $R_{n,i}$ and each $j \le n'$, $\mathbb{E}[R_{n,i,s} R_{n,j,s} | \mathcal{C}_n] = 0$ (P_{γ_n}-a.s.) and $\mathbb{E}[R_{n,i,s}^2 | \mathcal{C}_n] \le \delta_n'$ then by Markov's inequality, $\frac{1}{\sqrt{n}} \sum_{i=1}^{n'} R_{n,i} = o_{P_{\gamma_n}}(1)$ for $n' \le n$. We establish that (a) or (b) holds for terms which sum to $\hat{\ell}_{n,\theta_n}(W_i) - \tilde{\ell}_{\gamma_n}(W_i)$. Abbreviate $Z_{n,i} := Z(V_{n,i})$ and let

$$\begin{aligned}
R_{1,n,i} &:= (\hat{f}_{n,i} - f(V_{n,i})) f'(V_{n,i})(X_{2,i} - Z_{n,i}) \\
R_{2,n,i} &:= (Y_i - f(V_{n,i})) \left(f'(V_{n,i}) - \widehat{f'}_{n,i} \right) (X_{2,i} - Z_{n,i}) \\
R_{3,n,i} &:= (Y_i - f(V_{n,i})) \widehat{f'}_{n,i} \left(\hat{Z}_{n,i} - Z_{n,i} \right) \\
R_{4,n,i} &:= (\hat{f}_{n,i} - f(V_{n,i})) \left(f'(V_{n,i}) - \widehat{f'}_{n,i} \right) (X_{2,i} - Z_{n,i}) \\
R_{5,n,i} &:= (\hat{f}_{n,i} - f(V_{n,i})) \widehat{f'}_{n,i} \left(\hat{Z}_{n,i} - Z_{n,i} \right)
\end{aligned}$$

For some $a_j \in \{-1, 1\}$, we have that

$$\begin{aligned}
\frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\ell}_{n,\theta_n}(W_i) - \tilde{\ell}_{\gamma_n}(W_i) &= \sqrt{n}(\hat{\sigma}_n^{-2} - \sigma^{-2}) \sigma^2 \frac{1}{n} \sum_{i=1}^n \tilde{\ell}_{\gamma_n}(W_i) \\
&\quad + \hat{\sigma}_n^{-2} \sum_{j=1}^5 a_j \frac{1}{\sqrt{n}} \left[\sum_{i \in N^{(1)}} R_{j,n,i} + \sum_{i \in N^{(2)}} R_{j,n,i} \right].
\end{aligned}$$

The first term on the right hand side is $o_{P_{\gamma_n}}(1)$ by Lemma 4.3 and Proposition 3.1. For the second right hand side term first note that Lemma 4.3 also implies that $\hat{\sigma}_n^{-2} = O_{P_{\gamma_n}}(1)$. Let E_n be sets on which conditions (i) and (ii) in assumption SIM-NP(i) hold with $P_{\gamma_n}(E_n) \to 1$. For $j \in [3]$ we will show that (b) holds on E_n (for $i \in N^{(1)}$ or $i \in N^{(2)}$). That these terms are conditionally mean zero follows from the construction of the estimates. Specifically, using the fact that each $\hat{f}_{n,i}$, $\widehat{f'}_{n,i}$, $\hat{Z}_{n,i}$ is $\sigma(V_{n,i}, \{W_j\}_{j \in N_{-i}})$ measurable, independence, the LIE, Lemma C.5, $\mathbb{E}[\epsilon_i | X_i] = 0$ and $\mathbb{E}[(X_{2,i} - Z_{n,i}) | V_{n,i}] = 0$, it follows that each $\mathbb{E}[R_{j,n,i,s} R_{j,n,k,s} | \mathcal{C}_n] = 0$ for $j \in [3]$ and $k \notin N_{-i}$ with $\mathcal{C}_n = (W_j)_{j \in N^{(1)}}$ for $i \in N^{(2)}$ and $\mathcal{C}_n = (W_j)_{j \in N^{(2)}}$ for $i \in N^{(1)}$. Similar arguments along with the (P_{γ_n}-a.s.) boundedness of X_2 and assumption SIM-NP(i) show that on E_n each component $\mathbb{E}[R_{j,n,i,s}^2 | \mathcal{C}_n] \le r_n^2$. For $j \in \{4, 5\}$ (a) holds on E_n as by SIM-NP(i), on E_n, each $\mathbb{E}[\|R_{j,n,i}\|_2 | \mathcal{C}_n] \lesssim \mathcal{R}_{l,n,i} \mathcal{R}_{k,n,i} \le r_n^2 = o(n^{-1/2})$ for $l, k \in [3]$.

For the second part we will verify assumption T, which suffices to establish (7) and assumption R by Corollary 3.14. Note first that by (28) and assumption SIM-NP(i) the elements of $\tilde{\ell}_{\gamma_n}$ satisfy $\mathbb{E}[\tilde{\ell}^4_{\gamma_n,l}] = \mathbb{E}[(\epsilon_i f'(V_{n,i})\omega(X_i)(X_{2,i} - Z_{n,i}))^4] \lesssim \mathbb{E}[\epsilon_i^4] < \infty$ and so by Cauchy-Schwarz and e.g. Theorem 2.5.11 in Durrett (2019), $\frac{1}{n}\sum_{i=1}^n \tilde{\ell}_{\gamma_n,l}\tilde{\ell}_{\gamma_n,k} - \mathbb{E}\tilde{\ell}_{\gamma_n,l}\tilde{\ell}_{\gamma_n,k} = O_{P_{\gamma_n}}(n^{-1/2}\log(n)^{1/2+\kappa})$ for any $\kappa > 0$. The distributional observation that under P_{γ_n}, $(Y - f(V_n), X) \sim G$ and the form of $\tilde{\ell}_{\gamma_n}$ then implies that $\tilde{\mathcal{I}}_{\gamma_n} = \tilde{\mathcal{I}}_{\gamma}$ and hence

$$\left\| \frac{1}{n}\sum_{i=1}^n \tilde{\ell}_{\gamma_n}\tilde{\ell}'_{\gamma_n} - \tilde{\mathcal{I}}_{\gamma}\right\|_2 \leq \left\| \frac{1}{n}\sum_{i=1}^n \tilde{\ell}_{\gamma_n}\tilde{\ell}'_{\gamma_n} - \tilde{\mathcal{I}}_{\gamma}\right\|_F = O_{P_{\gamma_n}}(n^{-1/2}\log(n)^{1/2+\kappa}). \tag{90}$$

Secondly, write

$$\frac{1}{n}\sum_{i=1}^n \left(\hat{\ell}_{n,\theta_n,l} - \tilde{\ell}_{\gamma_n,l}\right)^2 \lesssim \hat{\sigma}_n^{-4}\sum_{j=1}^5 \frac{1}{n}\left(\sum_{i\in N^{(1)}} R^2_{j,n,i,l} + \sum_{i\in N^{(2)}} R^2_{j,n,i,l}\right) + (\hat{\sigma}_n^{-2} - \sigma^2)^2\sigma^4\mathbb{P}_n\tilde{\ell}^2_{\gamma_n,l}.$$

By Lemma 4.3, $\hat{\sigma}_n^{-4} = O_{P_{\gamma_n}}(1)$. Under assumptions SIM and SIM-NP(i), on E_n, each $\mathbb{E}[R^2_{j,n,i,l}|\mathcal{C}_n] \lesssim r_n^2$ as noted above. Since $r_n = o(\nu_n)$, Markov's inequality then implies that $\frac{1}{n}\sum_{i\in N^{(s)}} R^2_{j,n,i,l} = o_{P_{\gamma_n}}(\nu_n^2)$ for $s = 1, 2$. By Lemma 4.3 and equation (58), the second RHS term is $O_{P_{\gamma_n}}(n^{-1})$. Adding and subtracting and using Cauchy-Schwarz yields

$$\left\| \frac{1}{n}\sum_{i=1}^n \hat{\ell}_{n,\theta_n}\hat{\ell}'_{n,\theta_n} - \tilde{\ell}_{\gamma_n}\tilde{\ell}'_{\gamma_n}\right\|_2 \leq \left\| \frac{1}{n}\sum_{i=1}^n \hat{\ell}_{n,\theta_n}\hat{\ell}'_{n,\theta_n} - \tilde{\ell}_{\gamma_n}\tilde{\ell}'_{\gamma_n}\right\|_F = o_{P_{\gamma_n}}(\nu_n). \tag{91}$$

Combine (90) and (91) to see that assumption T is satisfied with any sequence $(\nu_n)_{n\in\mathbb{N}}$ as in the statement of the proposition. □

Proof of Proposition 4.7. Let $V_n := \iota(\theta_n, X)$. We first note that (i) $\breve{\ell}_{\gamma_n} \in L_2^0(P_{\gamma_n})$ and (ii) $P_{\gamma_n}\left[\breve{\ell}_{\gamma_n}B_{\gamma_n}h\right] = 0$ for all $h \in H_\eta$. For (i) use the LIE to obtain that if $W \sim P_{\gamma_n}$

$$\mathbb{E}\breve{\ell}_{\gamma_n}(W) = \mathbb{E}\left[\mathbb{E}[\epsilon|X]f'(V_n)\breve{\omega}(X)\left(X_2 - \frac{\mathbb{E}\left[\breve{\omega}(X)X_2|V_n\right]}{\mathbb{E}\left[\breve{\omega}(X)|V_n\right]}\right)\right] = 0,$$

and note that by boundedness of $\breve{\omega}$ (above and below), f', compactness of $\mathscr{X}$ we have $\mathbb{E}\breve{\ell}_{\gamma_n,k}(W)^4 < \infty$ for each $k = 1, \ldots, K-1$ which implies (i) and moreover that $\|\breve{\ell}_{\gamma_n}\|_2^2$ is uniformly P_{γ_n}-integrable. For (ii), if $W \sim P_{\gamma_n}$ then by the LIE, definition of $\dot{\mathscr{Z}}_\eta$ and (28)

$$\begin{aligned}
\mathbb{E}\left[\breve{\ell}_{\gamma_n}(W)[B_{\gamma_n}h](W)\right] &= \mathbb{E}\left[\mathbb{E}[\epsilon h_2(\epsilon, X)|X]f'(V_n)\breve{\omega}(X)\left(X_2 - \frac{\mathbb{E}\left[\breve{\omega}(X)X_2|V_n\right]}{\mathbb{E}\left[\breve{\omega}(X)|V_n\right]}\right)\right] \\
&\quad + \mathbb{E}\left[-\mathbb{E}[\epsilon\phi(\epsilon, X)|X]f'(V_n)\breve{\omega}(X)\left(X_2 - \frac{\mathbb{E}\left[\breve{\omega}(X)X_2|V_n\right]}{\mathbb{E}\left[\breve{\omega}(X)|V_n\right]}\right)\right] \\
&= \mathbb{E}\left[f'(V_n)\mathbb{E}\left[\breve{\omega}(X)X_2 - \frac{\breve{\omega}(X)\mathbb{E}\left[\breve{\omega}(X)X_2|V_n\right]}{\mathbb{E}\left[\breve{\omega}(X)|V_n\right]}\middle| V_n\right]\right] \\
&= 0.
\end{aligned}$$

The distributional observation that under P_{γ_n}, $(Y - f(V_n), X) \sim G$ and the form of $\breve{\ell}_{\gamma_n}$ then implies that $\Upsilon_{\gamma_n} = \Upsilon_\gamma$. Using this, along with (a) and (b) above, we can argue analogously to as in the proof of Proposition 3.1 (with $\tilde{\ell}_{\gamma_n}$ replaced by $\breve{\ell}_{\gamma_n}$ and $\tilde{\mathcal{I}}_{\gamma_n}$ replaced by Υ_{γ_n}) to conclude that under P_{γ_n,τ_n,h_n}, $\sqrt{n}\mathbb{P}_n\breve{\ell}_{\gamma_n} \rightsquigarrow \mathcal{N}(\Upsilon_\gamma\tau, \Upsilon_\gamma)$. Arguing as in the proofs of Propositions 3.2, 3.3 and Lemmas B.3, B.4, B.5 reveals that this suffices for the result provided we show that equations (6), (7) and (8) hold with $\breve{\ell}_{n,\theta_n}$ replacing $\hat{\ell}_{n,\theta_n}$, $\breve{\ell}_{\gamma_n}$ replacing $\tilde{\ell}_{\gamma_n}$, $\breve{\Upsilon}_{n,\theta_n}$ replacing $\hat{\mathcal{I}}_{n,\theta_n}$ and Υ_γ replacing $\tilde{\mathcal{I}}_\gamma$.

To this end we argue as in the proof of Proposition 4.4. Let $\mathcal{C}_n$ be some collection of random vectors, $\delta_n \to 0$ and $\delta_n' \to 0$. For any triangular array of random vectors $(R_{n,i})_{n\in\mathbb{N},i\leq n}$ if with P_{γ_n}-probability approaching one either (a) $\mathbb{E}[\|R_{n,i}\|_2|\mathcal{C}_n] \leq \delta_n n^{-1/2}$ or (b) for each element $R_{n,i,s}$ of $R_{n,i}$ and any $j \leq n'$, $\mathbb{E}[R_{n,i,s}R_{n,j,s}|\mathcal{C}_n] = 0$ (P_{γ_n}-a.s.) and $\mathbb{E}[R_{n,i,s}^2|\mathcal{C}_n] \leq \delta_n'$ then by Markov's inequality, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n'} R_{n,i} = o_{P_{\gamma_n}}(1)$ for $n' \leq n$. We establish that (a) or (b) holds for terms which sum to $\breve{\ell}_{n,\theta_n}(W_i) - \breve{\ell}_{\gamma_n}(W_i)$. Abbreviate $Z_{l,n,i} \coloneqq Z_l(V_{n,i})$ for $l \in [2]$ and let

$$
\begin{aligned}
R_{1,n,i} &\coloneqq (\hat{f}_{n,i} - f(V_{n,i}))f'(V_{n,i})\breve{\omega}(X_i)(X_{2,i} - Z_{n,i}) \\
R_{2,n,i} &\coloneqq (Y_i - f(V_{n,i}))\left(f'(V_{n,i}) - \widehat{f'}_{n,i}\right)\breve{\omega}(X_i)(X_{2,i} - Z_{n,i}) \\
R_{3,n,i} &\coloneqq (Y_i - f(V_{n,i}))\widehat{f'}_{n,i}\breve{\omega}(X_i)\left(\hat{Z}_{n,i} - Z_{n,i}\right) \\
R_{4,n,i} &\coloneqq (\hat{f}_{n,i} - f(V_{n,i}))\left(f'(V_{n,i}) - \widehat{f'}_{n,i}\right)\breve{\omega}(X_i)(X_{2,i} - Z_{n,i}) \\
R_{5,n,i} &\coloneqq (\hat{f}_{n,i} - f(V_{n,i}))\widehat{f'}_{n,i}\breve{\omega}(X_i)\left(\hat{Z}_{n,i} - Z_{n,i}\right),
\end{aligned}
$$

with $Z_{n,i} \coloneqq Z_{1,n,i}/Z_{2,n,i}$ and $\hat{Z}_{n,i} \coloneqq \hat{Z}_{1,n,i}/\hat{Z}_{2,n,i}$. For some $a_j \in \{-1, 1\}$, we have that

$$
\frac{1}{\sqrt{n}}\sum_{i=1}^n \breve{\ell}_{n,\theta_n}(W_i) - \breve{\ell}_{\gamma_n}(W_i) = \sum_{j=1}^5 a_j \frac{1}{\sqrt{n}}\left[\sum_{i\in N^{(1)}} R_{j,n,i} + \sum_{i\in N^{(2)}} R_{j,n,i}\right].
$$

Note also that

$$
\hat{Z}_{n,i} - Z_{n,i} = \frac{(\hat{Z}_{1,n,i} - Z_{1,n,i})Z_{2,n,i} + (Z_{2,n,i} - \hat{Z}_{2,n,i})Z_{1,n,i}}{\hat{Z}_{2,n,i}Z_{2,n,i}},
$$

and by assumption SIM-NP(ii) there is a sequence of sets E_n with $P_{\gamma_n}(E_n) \to 1$ such that each $\breve{\mathcal{R}}_{l,n,i} \leq r_n$ and each $\hat{f}_{n,i}$, $\widehat{f'}_{n,i}$, $\hat{Z}_{1,n,i,k}$ are bounded uniformly in i and for all large enough $n \in \mathbb{N}$ and $\hat{Z}_{2,n,i}$ is bounded below and above, uniformly in i and for all large enough $n \in \mathbb{N}$. From this it follows that $\mathbb{E}\left[\|\hat{Z}_{n,i} - Z_{n,i}\|_2^2|\mathcal{C}_n\right] \lesssim r_n^2 = o(n^{-1/2})$ on E_n where $\mathcal{C}_n = (W_j)_{j\in N^{(1)}}$ for $i \in N^{(2)}$ and $\mathcal{C}_n = (W_j)_{j\in N^{(2)}}$ for $i \in N^{(1)}$. Combining these observations we obtain that for $j \in \{4, 5\}$, on E_n, $\mathbb{E}[\|R_{j,n,i}\|_2|\mathcal{C}_n] \lesssim r_n^2 = o(n^{-1/2})$, which establishes (a). For $j \in [3]$ we establish (b). Specifically, using the fact that each $\hat{f}_{n,i}$, $\widehat{f'}_{n,i}$, $\hat{Z}_{n,i}$ is $\sigma(V_{n,i}, \{W_j\}_{j\in N_{-i}})$ measurable, independence, the LIE, Lemma C.5, $\mathbb{E}[\epsilon_i|X_i] = 0$ and $\mathbb{E}[\breve{\omega}(X_i)(X_{2,i} - Z_{n,i})|V_{n,i}] = 0$, it follows that $\mathbb{E}[R_{j,n,i,s}R_{j,n,k,s}|\mathcal{C}_n] = 0$ for $j \in [3]$ and $k \notin N_{-i}$ with $\mathcal{C}_n$ as above. Similar arguments along with the (P_{γ_n}-a.s.) boundedness of X_2 and the probabilistic rate and boundedness observations above show that on E_n each component $\mathbb{E}[R_{j,n,i,s}^2|\mathcal{C}_n] \lesssim r_n^2$. For the second part we will verify assumption T, which suffices to establish the required modifications of (7) and (8) by Corollary 3.14. Note first that as noted above the components of $\breve{\ell}_{\gamma_n}$ satisfy $\mathbb{E}[\breve{\ell}^4_{\gamma_n,l}] < \infty$ and so by Cauchy-Schwarz and e.g. Theorem 2.5.11 in Durrett (2019), $\frac{1}{n}\sum_{i=1}^n \breve{\ell}_{\gamma_n,l}\breve{\ell}_{\gamma_n,k} - \mathbb{E}\breve{\ell}_{\gamma_n,l}\breve{\ell}_{\gamma_n,k} = O_{P_{\gamma_n}}(n^{-1/2}\log(n)^{1/2+\kappa})$ for any $\kappa > 0$. As noted above $\Upsilon_{\gamma_n} = \Upsilon_\gamma$ and hence

$$\left\| \frac{1}{n}\sum_{i=1}^n \breve{\ell}_{\gamma_n}\breve{\ell}_{\gamma_n}' - \Upsilon_\gamma \right\|_2 \le \left\| \frac{1}{n}\sum_{i=1}^n \breve{\ell}_{\gamma_n}\breve{\ell}_{\gamma_n}' - \Upsilon_\gamma \right\|_F = O_{P_{\gamma_n}}(n^{-1/2}\log(n)^{1/2+\kappa}). \tag{92}$$

Secondly, write

$$\frac{1}{n}\sum_{i=1}^n \left(\breve{\ell}_{n,\theta_n,l} - \breve{\ell}_{\gamma_n,l}\right)^2 \lesssim \sum_{j=1}^5 \frac{1}{n}\left(\sum_{i\in N^{(1)}} R^2_{j,n,i,l} + \sum_{i\in N^{(2)}} R^2_{j,n,i,l}\right).$$

As noted above, on E_n, each $\mathbb{E}[R^2_{j,n,i,l}|\mathcal{C}_n] \lesssim r_n^2$. Since $r_n = o(\nu_n)$, Markov's inequality then implies that $\frac{1}{n}\sum_{i\in N^{(s)}} R^2_{j,n,i,l} = o_{P_{\gamma_n}}(\nu_n^2)$ for $s = 1,2$. Adding and subtracting and using Cauchy-Schwarz yields

$$\left\| \frac{1}{n}\sum_{i=1}^n \breve{\ell}_{n,\theta_n}\breve{\ell}_{n,\theta_n}' - \breve{\ell}_{\gamma_n}\breve{\ell}_{\gamma_n}' \right\|_2 \le \left\| \frac{1}{n}\sum_{i=1}^n \breve{\ell}_{n,\theta_n}\breve{\ell}_{n,\theta_n}' - \breve{\ell}_{\gamma_n}\breve{\ell}_{\gamma_n}' \right\|_F = o_{P_{\gamma_n}}(\nu_n). \tag{93}$$

Combine (92) and (93) to see that assumption T is satisfied with any sequence $(\nu_n)_{n\in\mathbb{N}}$ as in the statement of the proposition. $\square$

B.4 Proofs for section 5

Proof of proposition 5.1. Fix arbitrary $\tau_n \to \tau \in \mathbb{R}^{d_\theta}$ and $h_n \to h \in H_\eta$. Since Θ is open and $\theta_n + \tau_n/\sqrt{n} \to \theta \in \Theta$ for sufficiently large n, $\theta_n + \tau_n/\sqrt{n} \in \Theta$. The construction of H_η ensures that also $\beta + b_n/\sqrt{n} \in \mathscr{B}$ for large enough n. The restrictions on $\dot{\mathscr{Z}}_\eta$ and $\dot{\mathscr{G}}_{\eta,k}$ along with the norm on H suffice to ensure that $\eta_0(1 + t_n h_{n,0}) \in \mathscr{Z}$ and each $\eta_k(1 + t_n h_{n,k}) \in \mathscr{G}$. Specifically, for $k = 0, 1, \ldots, K$, the convergence in ensures the exists of an $M \in (0,\infty)$ such that, for all large enough n, $|h_{n,k}| \le M$ and $M/\sqrt{n} < 1$, λ-a.e.. This ensures that each $(1 + t_n h_{n,k}) > 0$ and hence $\eta_k(1 + t_n h_{n,k}) \ge 0$ (λ-a.e.). Moreover, the positivity of $1 + t_n h_{n,k}$ in combination with the continuous differentiability of $e \mapsto \sqrt{\eta_k(e)}$ and the fact that the square-root function is continuously differentiable away from 0, yields (via the chain rule) that $\sqrt{\eta_k(1 + t_n h_{n,k})}$ is continuously differentiable λ-a.e. (for $k \in [K]$). Moreover, for $k \in [K] \cup \{0\}$,

$$\int \eta_k(1 + t_n h_{n,k})\,\mathrm{d}\lambda = \int \eta_k\,\mathrm{d}\lambda + t_n \int h_{n,k}\eta_k\,\mathrm{d}\lambda = 1 + t_n \int h_k\,\mathrm{d}G_k = 1.$$

Additionally by Jensen's inequality

$$\left\| \int \tilde{X}\tilde{X}'\zeta(1 + h_{n,0}/\sqrt{n})\,\mathrm{d}\lambda - \int \tilde{X}\tilde{X}'\,\mathrm{d}G \right\|_2 \le \frac{M}{\sqrt{n}}\left\| \int \tilde{X}\tilde{X}'\,\mathrm{d}G \right\| \le \frac{M}{\sqrt{n}}\int \|\tilde{X}\|_2^2\,\mathrm{d}G \to 0,$$

which implies that for all large enough n, $\int \tilde{X}\tilde{X}'\zeta(1+h_{n,1}/\sqrt{n})\,\mathrm{d}\lambda \succ 0$. By the boundedness of each $h_{n,k}$ for large enough n, for such n and any non-negative function f with $G_k f < \infty$,

$$\int \eta_k(1+t_n h_{n,k}) f\,\mathrm{d}\lambda \leq (1+t_n M) G_k f < \infty.$$

Applying this with $k = 0$ and $f(\tilde{x}) = \|\tilde{x}\|_2^{4+\delta}$ completes the demonstration that $\eta_0(1 + t_n h_{n,0}) \in \mathscr{X}$ for all large enough n. Similarly applying it with $k \in [K]$ and $f(e) = |e|^{4+\delta}$ & $f(e) = |\phi_k(e)|^{4+\delta}$ ensures that the finite moment requirements in (35) are satisfied under $\eta_k(1 + t_n h_{n,k})$ for large enough n. By the definitions of $\mathscr{G}$ and $\dot{\mathscr{G}}_{\eta,k}$,

$$\int \iota\, \eta_k(1+t_n h_{n,k})\,\mathrm{d}\lambda = \int \kappa\, \eta_k(1+t_n h_{n,k})\,\mathrm{d}\lambda = 0,$$

verifying that the first two conditions of (35) hold under $\eta_k(1 + t_n h_{n,k})$. Lastly, since $G_k|\epsilon_k|^{4+\delta} < \infty$, the boundedness of $|h_{n,k}|$ ensures that

$$\int e^4 t_n h_{n,k}(e)\,\mathrm{d}G_k(e) \to 0, \quad \left[\int e^3 t_n h_{n,k}(e)\,\mathrm{d}G_k(e)\right]^2 \to 0$$

which, combined with $\mathbb{E}\epsilon_k^4 - 1 > (\mathbb{E}\epsilon_k^3)^2$ implies that for large enough n,

$$\int e^4(1+t_n h_{n,k}(e))\,\mathrm{d}G_k(e) - 1 > \left[\int e^3(1+t_n h_{n,k}(e))\,\mathrm{d}G_k(e)\right]^2,$$

completing the verification that $\eta_k(1 + t_n h_{n,k}) \in \mathscr{G}$ for all large enough n.

The next step is to establish (19). Firstly, for any given $u := (\tau, h) \in \mathbb{R}^{d_\theta} \times H_\eta$ let $\varphi := \varphi(u) := (\tau, b_1, b_2, \eta_0 h_0, \ldots, \eta_K h_K)$. Then, for any $\gamma \in \Gamma$, $t \in [0, \infty)$ and $u \in \mathbb{R}^{d_\theta} \times H_\eta$, define $q_{\gamma,t,u} := p_{\gamma+t\varphi}$ and $q_\gamma := q_{\gamma,0,0} = p_\gamma$. Finally, let $\Delta_\gamma(\varphi) := \frac{1}{2}[\tau'\dot{\ell}_\gamma + B_\gamma h]\sqrt{p_\gamma}$. For any $\gamma \in \Gamma$ and any $u \in \mathbb{R}^{d_\theta} \times H_\eta$, by Lemma S4 in Lee and Mesters (2021b),

$$\lim_{t\downarrow 0}\left\|\frac{\sqrt{p_{\gamma+t\varphi}} - \sqrt{p_\gamma}}{t} - \Delta_\gamma(\varphi)\right\|_{\lambda,2} = \lim_{t\downarrow 0}\left\|\frac{\sqrt{q_{\gamma,t,u}} - \sqrt{q_\gamma}}{t} - \frac{1}{2}[\tau'\dot{\ell}_\gamma + B_\gamma h]\sqrt{q_\gamma}\right\|_{\lambda,2} = 0. \quad (94)$$

In order to strengthen this directional differentiability into the result required by (19), we first establish an intermediate result. Let $(\updelta_n)_{n\in\mathbb{N}} \subset [0,1]$ be an arbitrary sequence, $t_n \downarrow 0$ and define $\boldsymbol{\gamma}_n := \gamma_n + \updelta_n t_n \varphi_n$ for $\varphi_n := \varphi(u_n)$ with $u_n \to u \in \mathbb{R}^{d_\theta} \times H_\eta$. Define also $A_n := A(\theta_n + \updelta_n t_n \tau_n, \beta_1 + \updelta_n t_n b_{n,1})$, $D_{1,l,n} := D_{1,l}(\theta_n + \updelta_n t_n \tau_n, \beta_1 + \updelta t_n b_{n,1})$, $\zeta_{l,k,j,n} := [D_{1,l,n}]_k [A_n^{-1}]_j'$, R_n is such that $\operatorname{vec}(R_n) = \beta_2 + \updelta_n t_n b_{n,2}$, $V_n := Y - R_n X$ and finally

$$\phi_{k,n} := \phi_k + \frac{\updelta_n t_n h_{n,k}'}{1 + \updelta_n t_n h_{n,k}}.$$

We will show that $\|\Delta_{\boldsymbol{\gamma}_n}(\varphi_n) - \Delta_\gamma(\varphi)\|_{\lambda,2} \to 0$ $(*)$. By Proposition 2.29 in van der Vaart (1998) it suffices to show that (i) $\Delta_{\boldsymbol{\gamma}_n}(\varphi_n) \to \Delta_\gamma(\varphi)$ λ-a.e. and (ii) $\limsup_{n\to\infty} \|\Delta_{\boldsymbol{\gamma}_n}(\varphi_n)\|_{\lambda,2}^2 \leq$

$\|\Delta_\gamma(\varphi)\|^2_{\lambda,2} < \infty$. We have that $\Delta_{\gamma_n}(\varphi_n) := \frac{1}{2}[\tau_n'\dot{\ell}_{\gamma_n} + B_{\gamma_n}h_n]\sqrt{p_{\gamma_n}}$, with

$$p_{\gamma_n}(W) = |\det(A_n)| \prod_{k=1}^{K} [\eta_k(1+\delta_n t_n h_{n,k})]\,(A_{n,k}V_n) \times [\eta_0(1+\delta_n t_n h_{n,0})](\tilde{X})$$

$$\dot{\ell}_{\gamma_n,l}(W) = \sum_{k=1}^{K} \zeta_{l,k,k,n}\left[\phi_{k,n}\left(A_{n,k}V_n\right)A_{n,k}V_n + 1\right] + \sum_{k=1}^{K}\sum_{j=1,\ j\neq k}^{K} \zeta_{l,k,j,n}\phi_{k,n}\left(A_{n,k}V_n\right)A_{n,j}V_n$$

$$\begin{aligned}[B_{\gamma_n}h_n](W) = h_{n,0}(\tilde{X}) &+ \sum_{k=1}^{K} h_{n,k}(A_{n,k}V_n) - \sum_{l=1}^{d_{\beta_2}} b_{n,2,l}\sum_{k=1}^{K}\phi_{k,n}\left(A_{n,k}V_n\right)A_{n,k}D_{2,l}X \\ &+ \sum_{m=d_\theta+1}^{d_\theta+d_{\beta_1}} b_{n,1,m}\left[\sum_{k=1}^{K}\zeta_{m,k,k,n}\left[\phi_{k,n}\left(A_{n,k}V_n\right)A_{n,k}V_n+1\right]\right] \\ &+ \sum_{m=d_\theta+1}^{d_\theta+d_{\beta_1}} b_{n,1,m}\left[\sum_{k=1}^{K}\sum_{j=1,\ j\neq k}^{K}\zeta_{m,k,j,n}\phi_{k,n}\left(A_{n,k}V_n\right)A_{n,j}V_n\right].\end{aligned}$$

Note first that there is a $N \in \mathbb{N}$ such that for $n \geq N$ each $|h_{n,k}|$ and $|h'_{n,k}|$ is bounded above λ-a.e. by some $\bar{h} \in (0,\infty)$. This implies that $\phi_{k,n} \to \phi_k$ λ-a.e. The assumed continuity of $D_{1,l}$ and A imply that $A_n \to A$ and each $\zeta_{l,j,k,n} \to \zeta_{l,j,k}$ and it is clear from its definition that $V_n \to V := Y - RX$. Inspection of the preceding display in light of these observations reveals that (i) holds. For (ii), the finiteness of $\|\Delta_\gamma(\varphi)\|^2_{\lambda,2} = {}^1\!/\!_4 P_\gamma[\tau'\dot{\ell}_\gamma + B_\gamma h]^2$ follows from Lemma 1.7 of van der Vaart (2002) and (94). For the remaining inequality it suffices to show that $P_{\gamma_n}\left[\tau_n\dot{\ell}_{\gamma_n} + B_{\gamma_n}h_n\right]^2 \to P_\gamma\left[\tau\dot{\ell}_\gamma + B_\gamma h\right]^2$. This will follow by Lemma 3.11 if we show that (a) P_{γ_n} converges to P_γ in total variation, (b) $g_n' := \tau_n\dot{\ell}_{\gamma_n} + B_{\gamma_n}h_n \in L_2(P_{\gamma_n})$ and $g := \tau\dot{\ell}_\gamma + B_\gamma h \in L_2(P_\gamma)$, (c) $g_n' \to g$ in P_γ-probability and (d) $(g_n')_{n\in\mathbb{N}}$ is uniformly square P_{γ_n}-integrable.[111] For (a), note that inspection of the preceding display reveals that $p_{\gamma_n} \to p_\gamma$ λ-a.e.. Hence, $P_{\gamma_n} \to P_\gamma$ in total variation by Scheffé's theorem. (b) follows from the fact that (94) holds for each $\gamma \in \Gamma$, $\tau \in \mathbb{R}^{d_\theta}$, $h \in H_\eta$ and Lemma 1.7 in van der Vaart (2002). For (c) note that inspection of the preceding display once more gives that $g_n' \to g$ λ-a.e. and hence P_γ-a.s. as $P_\gamma \ll \lambda$. Finally, for (d), let $\rho = 2+\delta/2$ where $\delta > 0$ is as in (35) & (36). Let N be large enough that for $n \geq N$, $t_n \in [0,1)$, each $|h_{n,k}|, |h'_{n,k}| \leq \bar{h} \in (0,\infty)$, each $|\tau_{n,l}| \leq 2|\tau_l|$, $|\varsigma_{n,l}| \leq 2|\varsigma_l|$ $\|A_n\|_2 \leq 2\|A\|_2$, each $|\zeta_{l,k,j,n}| \leq 2|\zeta_{l,k,j}|$, $|\phi_{n,k}| \leq |\phi_k| + \bar{h}$ and $P_{\gamma_n} \in \mathcal{P}$.[112] It suffices to show that $\sup_{n\geq N} P_{\gamma_n}|g_n'|^\rho < \infty$. In particular, by Hölder's inequality (and given the bounds just discussed holding for $n \geq N$), it is enough to show that each of $P_{\gamma_n}|\phi_{n,k}(A_{n,k}V_n)A_{n,j}V_n|^\rho$ for all $(k,j) \in [K]^2$ and $P_{\gamma_n}|\phi_{n,k}(A_{n,k}V_n)A_{n,k}D_{2,l}X|^\rho$ for all $k \in [K]$ and $l \in [d_{\beta_2}]$ are bounded independently of n (for $n \geq N$). Note that under

[111] Since we are interested only in the limiting behaviour, we can replace any P_{γ_n} which are not probability measures with $P_{\gamma_{n'}}$ where n' indicates the first index for which all subsequent elements of the sequence are probability measures. That each P_{γ_n} is a probability measure for n sufficiently large can be established analogously to the same for P_{γ_n,τ_n,h_n}, which was established at the start of this proof, upon replacing the t_n used in the argument there with $\delta_n t_n$.

[112] See footnote 111.

P_{γ_n}, $A_{n,k}V_n \sim \eta_k(1+\delta_n t_n h_{n,k})$ and $\tilde{X} \sim \eta_0(1+\delta_n t_n h_{n,0})$. By Cauchy-Schwarz we have

$$P_{\gamma_n}\left[|\phi_{n,k}(A_{n,k}V_n)|^\rho |A_{n,j}V_n|^\rho\right] \leq P_{\gamma_n}|\phi_{n,k}(A_{n,k}V_n)|^{4+\delta} P_{\gamma_n}|A_{n,j}V_n|^{4+\delta},$$
$$P_{\gamma_n}\left[|\phi_{n,k}(A_{n,k}V_n)|^\rho |A_{n,k}D_{2,l}X|^\rho\right] \leq P_{\gamma_n}|\phi_{n,k}(A_{n,k}V_n)|^{4+\delta} P_{\gamma_n}|A_{n,k}D_{2,l}X|^{4+\delta}.$$

For $n \geq N$, $\eta_k(1+\delta_n t_n h_{n,k}) \leq \eta_k(1+\bar{h})$ and so by (35) & (36), for a constant C which does not depend on n,

$$P_{\gamma_n}|A_{n,j}V_n|^{4+\delta} \leq (1+\bar{h})\int e^{4+\delta}\eta_j(e)\,\mathrm{d}\lambda < \infty,$$
$$P_{\gamma_n}|\phi_{n,k}(A_{n,k}V_n)|^{4+\delta} \leq C(1+\bar{h})\int \left[|\phi_k(e)|^{4+\delta} + \bar{h}^{d+\delta}\right]\eta_k(e)\,\mathrm{d}\lambda < \infty,$$
$$P_{\gamma_n}|A_{n,j}D_{2,l}X|^{4+\delta} \leq (1+\bar{h})[2\|A\|_2\|D_{2,l}\|_2]^{4+\delta}\int \|(1,\tilde{x}')\|_2^{4+\delta}\eta_0(\tilde{x})\,\mathrm{d}\lambda < \infty.$$

As each right hand side term in the preceding display does not depend on n, this completes the demonstration of (d) and hence of $(*)$.

We now establish (19). Suppose first that $\theta_n = \theta$ and let $u_n \to u$ be arbitrary and put $\varphi_n := \varphi(u_n)$, $\varphi := \varphi(u)$ and $t_n \downarrow 0$. Also let $g_\gamma := \tau'\dot{\ell}_\gamma + B_\gamma h$. For large enough n, $\gamma + \varphi_n \in \Gamma$ and so applying (94) and the mean value theorem (e.g. Drabek and Milota, 2007, Theorem 3.2.7) for all such n,

$$\begin{aligned}
&\left\|t_n^{-1}\left(\sqrt{q_{\gamma,t_n,u_n}} - \sqrt{q_\gamma}\right) - \frac{1}{2}g_\gamma\sqrt{q_\gamma}\right\|_{\lambda,2} \\
&\quad\leq \left\|t_n^{-1}\left(\sqrt{q_{\gamma,t_n,u_n}} - \sqrt{q_{\gamma,t_n,u}}\right)\right\|_{\lambda,2} + \left\|t_n^{-1}\left(\sqrt{q_{\gamma,t_n,u}} - \sqrt{q_\gamma}\right) - \frac{1}{2}g_\gamma\sqrt{q_\gamma}\right\|_{\lambda,2} \\
&\quad\leq \sup_{\delta\in[0,1]}\left\|\Delta_{\gamma+\delta t_n(\varphi_n-\varphi)}(\varphi_n - \varphi)\right\|_{\lambda,2} + o(1).
\end{aligned} \tag{95}$$

For any sequence $(\delta_n)_{n\in\mathbb{N}} \subset [0,1]$ we have that $\|\Delta_{\gamma+\delta_n t_n(\varphi_n-\varphi)}(\varphi_n-\varphi) - \Delta_\gamma(0)\|_{\lambda,2} \to 0$ by $(*)$ and $\|\Delta_\gamma(0)\|_{\lambda,2} = 0$.[113] It follows that $\limsup_{n\to\infty}\sup_{\delta\in[0,1]}\left\|\Delta_{\gamma+\delta t_n(\varphi_n-\varphi)}(\varphi_n-\varphi)\right\|_{\lambda,2} = 0$ and hence

$$\left\|\frac{\sqrt{p_{\gamma+t_n\varphi_n}} - \sqrt{p_\gamma}}{t_n} - \Delta_\gamma(\varphi)\right\|_{\lambda,2} = \left\|\frac{\sqrt{q_{\gamma,t_n,u_n}} - \sqrt{q_\gamma}}{t_n} - \frac{1}{2}g_\gamma\sqrt{q_\gamma}\right\|_{\lambda,2} = o(1), \tag{96}$$

which we note holds for any $\gamma \in \Gamma$, since such γ was arbitrary. Now, consider an arbitrary sequence $\theta_n \to \theta$ and $\gamma_n = (\theta_n, \eta)$. Using (96) and applying the mean value theorem at each $n \in \mathbb{N}$ gives (e.g. Drabek and Milota, 2007, Theorem 3.2.7)

$$\begin{aligned}
\left\|t_n^{-1}\left(\sqrt{q_{\gamma_n,t_n,u_n}} - \sqrt{q_{\gamma_n}}\right) - \frac{1}{2}g_{\gamma_n}\sqrt{q_{\gamma_n}}\right\|_{\lambda,2} &\leq |t_n^{-1}|\sup_{\delta\in[0,1]}\|\Delta_{\gamma_n+\delta t_n\varphi_n}(t_n\varphi_n) - t_n\Delta_{\gamma_n}(\varphi)\|_{\lambda,2} \\
&= \sup_{\delta\in[0,1]}\|\Delta_{\gamma_n+\delta t_n\varphi_n}(\varphi_n) - \Delta_{\gamma_n}(\varphi)\|_{\lambda,2}\,.
\end{aligned} \tag{97}$$

[113] The latter observation follows directly from the definition of Δ_γ

By $(*)$ we have for some sequence $(\delta_n)_{n\in\mathbb{N}} \subset [0,1]$,[114]

$$
\begin{aligned}
\limsup_{n\to\infty} \sup_{\delta\in[0,1]} &\|\Delta_{\gamma_n+\delta t_n\varphi_n}(\varphi_n) - \Delta_{\gamma_n}(\varphi)\|_{\lambda,2} \\
&\le \limsup_{n\to\infty} \|\Delta_{\gamma_n+\delta_n t_n\varphi_n}(\varphi_n) - \Delta_{\gamma}(\varphi)\|_{\lambda,2} + \limsup_{n\to\infty} \|\Delta_{\gamma_n}(\varphi) - \Delta_{\gamma}(\varphi)\|_{\lambda,2} \\
&= o(1).
\end{aligned}
$$

Combining this with (97) and taking $t_n = n^{-1/2}$ yields

$$
\left\|\sqrt{n}\left(\sqrt{p_{\gamma_n,\tau_n,h_n}} - \sqrt{p_{\gamma_n}}\right) - \frac{1}{2} g_{\gamma_n}\sqrt{p_{\gamma_n}}\right\|_{\lambda,2} = \left\|t_n^{-1}\left(\sqrt{q_{\gamma_n,t_n,u_n}} - \sqrt{q_{\gamma_n}}\right) - \frac{1}{2} g_{\gamma_n}\sqrt{q_{\gamma_n}}\right\|_{\lambda,2} = o(1),
$$

which implies (19).

Finally we demonstrate that $P_{\gamma_n} g_n = 0$ and the uniform square P_{γ_n}-integrability of the score functions g_n. That $P_{\gamma_n} g_n = 0$ and $g_n \in L_2(P_{\gamma_n})$ follows from (94) applied separately for each $n \in \mathbb{N}$ (with $\gamma = \gamma_n$) and Lemma 1.7 in van der Vaart (2002). The uniform square P_{γ_n}-integrability of $(g_n)_{n\in\mathbb{N}}$ follows from the uniform square P_{γ_n}-integrability of $(g_n')_{n\in\mathbb{N}}$ established in (d) above applied with $\delta_n = 0$, any $t_n \downarrow 0$ and $u_n = 0$. □

Proof of proposition 5.2. The claim regarding the form of the efficient score function follows from proposition 5.1, Lemma 3 of Lee and Mesters (2021a) and Lemma C.4.

For assumption CM(ii), fix $\tau \in \mathbb{R}^{d_\theta}$ and $h \in H_\eta$ and let $g_n = \tau'\dot{\ell}_{\gamma_n} + B_{\gamma_n} h$ and $g := \tau'\dot{\ell}_\gamma + B_\gamma h$ where $\dot{\ell}_\gamma$ and B_γ are defined analogously to in Proposition 5.1 but with $A = A(\theta, \beta_1)$ in place of $A_n = A(\theta_n, \beta_1)$. During the demonstration of $(*)$ in the proof of Proposition 5.1 it was shown that $\lim_{n\to\infty} P_{\gamma_n}(g_n')^2 = P_\gamma g^2$. Applying this result with $\delta_n = 0$, any $t_n \downarrow 0$ and $u_n = 0$ yields $\lim_{n\to\infty} P_{\gamma_n} g_n^2 = P_\gamma g^2$.

A similar argument can be used for the efficient score function. Let $\breve{\ell}_\gamma := (\tilde{\ell}_{\gamma,1}', \tilde{\ell}_{\gamma,2}')'$. Applied with $\delta_n = 0$, any $t_n \downarrow 0$ and $u_n = 0$, (a) in the proof of Proposition 5.1 yields that $P_{\gamma_n} \to P_\gamma$ in total variation. Since the components of $\breve{\ell}_\gamma$ and $\breve{\ell}_\gamma$ are defined as orthogonal projections onto subspaces of $L_2(P_{\gamma_n})$ and $\in L_2(P_\gamma)$ respectively, they lie in these spaces. Inspection of the form of each element of $\breve{\ell}_{\gamma_n}$ and $\breve{\ell}_\gamma$ reveals that $\breve{\ell}_{\gamma_n} \to \breve{\ell}_\gamma$ λ-a.e. and hence P_γ-a.s. as $P_\gamma \ll \lambda$. Let $\rho = 2 + \delta/2$ where δ is as in (35) & (36). Let $N \in \mathbb{N}$ be large enough that for $n \ge N$, each $|\tau_{n,l}| \le 2|\tau_l|$, $|\varsigma_{n,l}| \le 2|\varsigma_l|$, $\|A_n\|_2 \le 2\|A\|_2$, each $|\zeta_{l,k,j,n}| \le 2|\zeta_{l,k,j}|$ and $P_{\gamma_n} \in \mathcal{P}$. To show that $\breve{\ell}_{\gamma_n,l}^2$ is uniformly P_{γ_n}-integrable for each $l \in [d_\theta + d_\beta]$ it suffices to show that $\sup_{n\ge N} P_{\gamma_n}|\breve{\ell}_{\gamma_n,l}|^\rho < \infty$ for each such l. In particular, by Hölder's inequality (and given the bounds just discussed holding for $n \ge N$) it is sufficient to show that each of (for all $(k,j) \in [K]^2$ with $k \ne j$ and $s \in [d_{\beta_2}]$)

$$
P_{\gamma_n}|A_{n,k}V_n|^\rho,\ P_{\gamma_n}|\kappa(A_{n,k}V_n)|^\rho,\ P_{\gamma_n}|\phi_k(A_{n,k}V_n)A_{j,n}V_n|^\rho,\ P_{\gamma_n}|A_{n,k}D_{2,s}(X-\mu)\phi_k(A_{k,n}V_n)|^\rho,
$$

are bounded independently of n (for $n \ge N$). Under P_{γ_n} $A_{n,k}V_n \sim \eta_k$ and $X \sim \eta_0$. Using independence, Hölder's inequality and (35) & (36) for constants $C_1, C_2 \in (0,\infty)$ independent

[114] On the right hand side take the trivial sequences $\varphi_n = \varphi$ and $\delta_n = 0$.

of n

$$
\begin{aligned}
P_{\gamma_n}|A_{n,k}V_n|^\rho &= \int e^\rho \,\mathrm{d}G_k(e) < \infty \\
P_{\gamma_n}|\kappa(A_{n,k}V_n)|^\rho &\le C_1 \int (e^{4+\delta}+1)\,\mathrm{d}G_k(e) < \infty \\
P_{\gamma_n}|\phi_k(A_{n,k}V_n)A_{j,n}V_n|^\rho &= \int |\phi_k(e_k)|^\rho \,\mathrm{d}G_k(e_k) \int |e_j|^\rho \,\mathrm{d}G_j(e_j) < \infty \\
P_{\gamma_n}|A_{n,k}D_{2,s}(X-\mu)\phi_k(A_{k,n}V_n)|^\rho &\le C_2 \int (\|(1,\tilde{x})\|_2^\rho + \|\mu\|_2^\rho)\,\mathrm{d}G_0(\tilde{x}) \int |\phi_k(e_k)|^\rho \,\mathrm{d}G_k(e_k) < \infty.
\end{aligned}
$$

Since each right hand side term in the preceding display does not depend on n, this establishes the uniform P_{γ_n}-integrability of each $\breve{\ell}^2_{\gamma_n,l}$. By Cauchy-Schwarz, the continuous mapping theorem and Lemma 3.11 it then follows that $P_{\gamma_n}\left[\breve{\ell}_{\gamma_n}\breve{\ell}'_{\gamma_n}\right] \to P_\gamma\left[\breve{\ell}_\gamma\breve{\ell}'_\gamma\right]$. To complete the argument, note that the convergence just established along with the uniform P_{γ_n}-integrability of each $\breve{\ell}^2_{\gamma_n,l}$ implies that also each component $\tilde{\ell}^2_{\gamma_n,l}$ (for $l \in [d_\theta]$) is uniformly P_{γ_n}-integrable and so the same holds for $\|\tilde{\ell}_{\gamma_n}\|_2^2$. Again by definition each component $\tilde{\ell}_{\gamma_n,l} \in L_2(P_{\gamma_n})$ and $\tilde{\ell}_{\gamma,l} \in L_2(P_\gamma)$ and so using the uniform P_{γ_n}-integrability just established, (46), $P_{\gamma_n}\left[\breve{\ell}_{\gamma_n}\breve{\ell}'_{\gamma_n}\right] \to P_\gamma\left[\breve{\ell}_\gamma\breve{\ell}'_\gamma\right]$, Cauchy-Schwarz, the continuous mapping theorem and Lemma 3.11 once more we may conclude that $\lim_{n\to\infty}\tilde{\mathcal{I}}_{\gamma_n} = \tilde{\mathcal{I}}_\gamma$.

It remains to check the boundedness of B_γ, which follows directly as

$$
\|B_\gamma h\|_{P_\gamma,2} \lesssim \|b_1\|_2 + \|b_2\|_2 + \sum_{k=1}^K \|h_k\|_{G_k,2} \lesssim \|b\|_2 + \sum_{k=1}^K \|h_k\| = \|h\|.
$$

□

Proof of proposition 5.3. That assumption M holds is a consequence of the model setup in assumption LSEM & the sampling assumption. Assumption CM(ii) follows by proposition 5.2. Assumption DQM holds by proposition 5.1, the proof of which also shows that the scores $\dot{\ell}_{\gamma_n} \in L_2^0(P_{\gamma_n})$ & $B_{\gamma_n} : H_\eta \to L_2^0(P_{\gamma_n})$. Then proposition 3.10 applied with $g_n = \tau'\dot{\ell}_{\gamma_n} + B_{\gamma_n}h$ yields that assumption LAN holds.

It remains to show that assumptions E and R hold.[115] Suppose that $(\beta_n)_{n\in\mathbb{N}} \subset \mathscr{B}$ is a deterministic $\sqrt{n}$-consistent sequence for β (as in assumption DSE) and let $\hat{\ell}_{\xi_n,1}$ & $\hat{\ell}_{\xi_n,2}$ be formed as in equation (49). Let $\gamma'_n := (\theta_n, \eta_n)$ with $\eta_n := (\beta_n, \eta_0, \ldots, \eta_K)$. Let $\breve{\ell}_\gamma := (\tilde{\ell}'_{\gamma,1}, \tilde{\ell}'_{\gamma,2})'$ and $\check{\ell}_{\xi_n} := (\hat{\ell}'_{\xi_n,1}, \hat{\ell}'_{\xi_n,2})'$. Components of $\check{\ell}_{\xi_n}$ have one of two forms:

$$
\hat{\ell}_{\xi_n,m,l}(W_i) = \sum_{k=1}^K \left[\zeta_{l,k,k,n}\left(\hat{\tau}_{n,k,1}e_{n,k,i} + \hat{\tau}_{n,k,2}\kappa(e_{n,k,i})\right) + \sum_{j=1,j\neq k}^K \zeta_{l,k,j,n}\hat{\phi}_{n,k}(e_{n,k,i})e_{n,j,i}\right],
$$

$$
\hat{\ell}_{\xi_n,2,d_{b_1}+s}(W_i) = \sum_{k=1}^K [-A_{n,k}D_{2,s}]\left[(X_i - \bar{X}_n)\hat{\phi}_{n,k}(e_{n,k,i}) - \bar{X}_n\left(\hat{\varsigma}_{n,k,1}e_{n,k,i} + \hat{\varsigma}_{n,k,2}\kappa(e_{n,k,i})\right)\right]
$$

(with $m = 1$ and $l \in [d_\theta]$ or $m = 2$ and $l \in [d_{\beta_1}]$ and $s \in [d_{\beta_2}]$). Under $P_{\gamma'_n}$, $e_{n,k,i} \simeq \epsilon_k$ and

[115]The argument in this section proceeds similarly to the relevant parts of the proofs of Theorem 2 & Proposition 2 of Lee and Mesters (2021a).

$e_{n,j,i} \simeq \epsilon_k$. Therefore, by assumptions LSEM and DSE, $\frac{1}{n}\sum_{i=1}^n \left[\hat{\phi}_{n,k}(e_{n,k,i}) - \phi_k(e_{n,k,i})\right] e_{n,j,i} = o_{P_{\gamma_n'}}(n^{-1/2})$ and $\frac{1}{n}\sum_{i=1}^n \left[\hat{\phi}_{n,k}(e_{n,k,i}) - \phi_k(e_{n,k,i})\right](X_i - \mu) = o_{P_{\gamma_n'}}(n^{-1/2})$. Additionally, since $(e_{n,k,i})_{i=1}^n$ and $(\kappa(e_{n,k,i}))_{i=1}^n$ and $(\phi_k(e_{n,k,i}))_{n\in\mathbb{N}}$ are i.i.d. samples from mean zero distributions with finite variance under $P_{\gamma_n'}$ given assumption LSEM and equation (45), it follows that $\frac{1}{\sqrt{n}}\sum_{i=1}^n a_{n,k,i} = O_{P_{\gamma_n'}}(1)$, for $a_{n,k,i} \in \{e_{n,k,i}, \kappa(e_{n,k,i}), \phi_k(e_{n,k,i})\}$. The argument of Lemma 7 in Lee and Mesters (2021a) implies that $\|\hat{\varkappa}_{n,k} - \varkappa_k\|_2 = o_{P_{\gamma_n'}}(\nu_n) = o_{P'_{\gamma_n}}(1)$ for $\varkappa \in \{\tau, \varsigma\}$ where ν_n is defined as in assumption DSE.[116] Since $\tilde{X} \sim \eta_0$ under $P_{\gamma_n'}$, $\frac{1}{n}\sum_{i=1}^n X_i - \mu = o_{P_{\gamma_n'}}(1)$ by the LLN. The continuity of A and $D_{1,l}$ yields that each $\zeta_{l,k,j,n} \to \zeta_{l,k,j}$ and hence are bounded. Combining these observations yields that

$$\sqrt{n}\mathbb{P}_n\left[\check{\ell}_{\xi_n} - \check{\ell}_{\gamma_n'}\right] = o_{P_{\gamma_n'}}(1). \tag{98}$$

Let $\hat{I}_{\xi_n} := \mathbb{P}_n\check{\ell}_{\xi_n}\check{\ell}_{\xi_n}'$, $\check{I}_{\gamma_n'} := \mathbb{P}_n\check{\ell}_{\gamma_n'}\check{\ell}_{\gamma_n'}'$ and $\check{I}_{\gamma_n} := P_{\gamma_n}\check{\ell}_{\gamma_n}\check{\ell}_{\gamma_n}'$. Firstly, let $m, r \in \{1,2\}$ and l, s be indices such that $\hat{\ell}_{\xi_n,m,l}$ and $\hat{\ell}_{\xi_n,r,s}$ are components of $\check{\ell}_{\xi_n}$. Let $\hat{U}_{n,i,m,l} := \hat{\ell}_{\xi_n,m,l}(W_i)$, $\tilde{U}_{n,i,m,l} := \tilde{\ell}_{\xi_n,m,l}(W_i)$ and $D_{n,i,m,l} := \hat{U}_{n,i,m,l} - \tilde{U}_{n,i,m,l}$. By Cauchy-Schwarz, assumptions LSEM, DSE, (45) and arguing analogously to Lemma 8 of Lee and Mesters (2021a)

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^n D_{n,i,l,m}\tilde{U}_{n,i,r,s}\right| &\le \left(\frac{1}{n}\sum_{i=1}^n \tilde{U}_{n,i,r,s}^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^n D_{n,i,l,m}^2\right)^{1/2} = o_{P'_{\gamma_n}}(\nu_n)\\
\left|\frac{1}{n}\sum_{i=1}^n \tilde{U}_{n,i,l,m}D_{n,i,r,s}\right| &\le \left(\frac{1}{n}\sum_{i=1}^n \hat{U}_{n,i,l,m}^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^n D_{n,i,r,r}^2\right)^{1/2} = o_{P'_{\gamma_n}}(\nu_n),
\end{aligned}$$

and hence $\mathscr{R}_{1,n} := \|\hat{I}_{\xi_n} - \check{I}_{\xi_n}\|_2 \le \|\hat{I}_{\xi_n} - \check{I}_{\xi_n}\|_F = o_{P_{\gamma_n'}}(\nu_n)$.[117] Next let

$$Q_{n,i,l,m,r,s} := \tilde{\ell}_{\gamma_n',l,m}(W_i)\tilde{\ell}_{\gamma_n',r,s}(W_i) - \tilde{\ell}_{\gamma_n,l,m}(W_i)\tilde{\ell}_{\gamma_n,r,s}(W_i),$$

and let $\check{Q}_{n,i,l,m,r,s}$ be defined analogously except with each $e_{n,k,i}$ replaced by $\epsilon_{i,k}$. Note that the distribution of $Q_{n,i,l,m,r,s}$ under $P_{\gamma_n'}$ is the same as that of $\check{Q}_{n,i,l,m,r,s}$ under the product measure $G = \prod_{k=0}^K G_k$. Therefore, arguing analogously to the corresponding part of the proof of proposition 2 in Lee and Mesters (2021a), using their Lemma 6 and Theorems 2.5.11 & 2.5.12 in Durrett (2019) gives that $\mathscr{R}_{2,n} := \|\check{I}_{\xi_n} - \check{I}_{\gamma_n}\|_2 = o_{P'_{\gamma_n}}(\nu_n)$. Combining this with the result for $\mathscr{R}_{1,n}$ we have that

$$\|\hat{I}_{\xi_n} - \check{I}_{\gamma_n}\|_2 = o_{P_{\gamma_n'}}(\nu_n). \tag{99}$$

[116] The Lemma as stated does not apply directly since it is for the case where $\theta_n = \theta$. Regardless, since $e_{n,k,i} \sim \eta_k$ and $\tilde{X} \sim \eta_0$ under $P_{\gamma_n'}$ the argument also holds in our case.

[117] Similarly to footnote 116, whilst Lemma 8 in Lee and Mesters (2021a) cannot be directly applied since it assumes $\theta_n = \theta$, the underlying argument continues to apply here as it is based on the fact that under the relevant measure (here $P_{\gamma_n'}$) $e_{n,k,i} \sim \eta_k$ and $\tilde{X} \sim \eta_0$. Moreover their assumptions 5 & 6 hold under assumptions LSEM, DSE and (45).

Next we demonstrate that for each pair m, l indexing and element of $\breve{\ell}_{\gamma_n}$ we have

$$\int [\tilde{\ell}_{\gamma_n',m,l}\sqrt{p_{\gamma_n'}} - \tilde{\ell}_{\gamma_n,m,l}\sqrt{p_{\gamma_n}}]^2 \,\mathrm{d}\lambda \to 0. \tag{100}$$

Note that λ-a.e. each $\tilde{\ell}_{\gamma_n',m,l}\sqrt{p_{\gamma_n'}} \to \tilde{\ell}_{\gamma,m,l}\sqrt{p_\gamma}$ and $\tilde{\ell}_{\gamma_n,m,l}\sqrt{p_{\gamma_n}} \to \tilde{\ell}_{\gamma,m,l}\sqrt{p_\gamma}$ by the assumed continuity of A, each $D_{1,l}$, each η_k and each ϕ_k and the form of these functions. Hence by Proposition 2.29 in van der Vaart (1998) it suffices to show that $\int \tilde{\ell}^2_{\gamma_n',m,l}\,\mathrm{d}P_{\gamma_n'} \to \int \tilde{\ell}^2_{\gamma,m,l}\,\mathrm{d}P_\gamma$ and $\int \tilde{\ell}^2_{\gamma_n,m,l}\,\mathrm{d}P_{\gamma_n} \to \int \tilde{\ell}^2_{\gamma,m,l}\,\mathrm{d}P_\gamma$, since $\tilde{\ell}_{\gamma,m,l} \in L_2(P_\gamma)$ by its definition. Define $Q_{n,i,l,m} := \tilde{\ell}^2_{\gamma_n,m,l}$, $Q'_{n,l,m} := \tilde{\ell}^2_{\gamma_n',m,l}$ and $\breve{Q}_{n,l,m}$, $\breve{Q}'_{n,l,m}$ which are defined analogously except with each $e_{n,k,i}$ replaced by $\epsilon_{i,k}$. Under P_{γ_n}, $Q_{n,l,m}$ has the same distribution as $\breve{Q}_{n,l,m}$ has under G; similarly under $P_{\gamma_n'}$, $Q'_{n,l,m}$ has the same distribution as $\breve{Q}'_{n,l,m}$ has under G. Hence, $\int \tilde{\ell}^2_{\gamma_n',m,l}\,\mathrm{d}P_{\gamma_n'} = \int Q'_{n,m,l}\,\mathrm{d}G$ and $\int \tilde{\ell}^2_{\gamma_n,m,l}\,\mathrm{d}P_{\gamma_n} = \int Q_{n,m,l}\,\mathrm{d}G$. This observation and the the continuity of A and each $D_{1,l}$ is sufficient for the required integral convergence to hold.[118] We note that the same argument which yielded the uniform P_{γ_n}-integrability of $\|\tilde{\ell}_{\gamma_n}\|_2^2$ in the proof of Proposition 5.2 can be used to show that that $\|\tilde{\ell}_{\gamma_n'}\|_2^2$ is uniform $P_{\gamma_n'}$-integrable. Since $\theta \mapsto \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$ is locally constant, for all sufficiently large $n \in \mathbb{N}$ we have $\operatorname{rank}(\tilde{\mathcal{I}}_{\gamma_n}) = \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$. $\tilde{\mathcal{I}}_{\gamma_n} \to \tilde{\mathcal{I}}_\gamma$ (which holds as we have shown that assumption CM(ii) does). The proof is completed by applying Lemma B.7. □

Proof of corollary 5.4. This follows from propositions 5.3, 3.2 and 3.3. □

Proof of corollary 5.5. This follows from proposition 5.3 and corollaries 3.7 & 3.9, on noting that H_η – as defined in equation (43) – is a linear subspace of H whenever $\beta \in \operatorname{int} \mathscr{B}$. □

C Supporting results

Lemma C.1. *Let $\{Z_{n,k} : k \leq n, n \in \mathbb{N}\}$ be a triangular array of $L-$dimensional random vectors, such that each row is independent with $\mathbb{E}[Z_{n,k}] = 0$ and $\Sigma_{n,k} := \mathbb{E}\left[Z_{n,k}Z_{n,k}'\right]$ exists. Suppose that*

$$\frac{1}{n}\sum_{k=1}^n \Sigma_{n,k} \to \Sigma_\star, \tag{101}$$

with $\Sigma_\star$ positive semi-definite (and finite) and that for each $\varepsilon > 0$

$$\frac{1}{n}\sum_{k=1}^n \mathbb{E}\left[\|Z_{n,k}\|^2 \mathbf{1}\{\|Z_{n,k}\| \geq \varepsilon\sqrt{n}\}\right] \to 0. \tag{102}$$

Then

$$\frac{1}{\sqrt{n}}\sum_{k=1}^n Z_{n,k} \rightsquigarrow \mathcal{N}(0, \Sigma_\star).$$

Proof. Put $\xi_{n,k} := Z_{n,k}/\sqrt{n}$ for $k \leq n$ and $\xi_{n,k} := 0$ otherwise. Fix $a \in \mathbb{R}^L$. For each $n \in \mathbb{N}$, let $\mathcal{F}_{n,k} = \sigma(\xi_{n,t} : t \leq k)$ for $k \leq n$ and $\mathcal{F}_{n,k} = \mathcal{F}_{n,n}$ otherwise. The adapted sequence $(a'\xi_{n,k}, \mathcal{F}_{n,k})_{k\in\mathbb{N}}$ is clearly a martingale difference sequence by the independence, mean zero and (square) integrability of each $Z_{n,k}$. Moreover, the sums $\sum_{k=1}^\infty a'\xi_{n,k} = \sum_{k=1}^n a'\xi_{n,k}$ and

[118] See the corresponding part of the proof of proposition 2 in Lee and Mesters (2021a) for additional details.

$\sum_{k=1}^{\infty}\mathbb{E}[(a'\xi_{n,k})^2] = \sum_{k=1}^{n}\mathbb{E}[(a'\xi_{n,k})^2]$ trivially converge with probability 1 for each $n \in \mathbb{N}$. By linearity and continuity we have that

$$\sum_{k=1}^{\infty}\mathbb{E}[(a'\xi_{n,k})^2] = \sum_{k=1}^{n}\mathbb{E}[(a'\xi_{n,k})^2] = a'\left[\frac{1}{n}\sum_{k=1}^{n}\Sigma_{n,k}\right]a \to a'\Sigma_\star a \geq 0.$$

Next, suppose that $a \neq 0$ and let $\varepsilon > 0$. We have that $\{|a'Z_{n,k}| \geq \varepsilon\sqrt{n}\} \subset \{\|Z_{n,k}\| \geq \varepsilon\sqrt{n}/\|a\|\}$ and therefore

$$\sum_{k=1}^{\infty}\mathbb{E}\left[(a'\xi_{n,k})^2\mathbf{1}\{|a'\xi_{n,k}| \geq \varepsilon\}\right] \leq \|a\|^2\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\left[\|Z_{n,k}\|^2\mathbf{1}\{\|Z_{n,k}\| \geq \varepsilon\sqrt{n}/\|a\|\}\right] \to 0,$$

by assumption.[119] Noting the assumed independence, the conditions of Theorem 18.1 of Billingsley (1999) are satisfied and hence

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}a'Z_{n,k} = \sum_{k=1}^{\infty}a'\xi_{n,k} \rightsquigarrow \mathcal{N}(0, a'\Sigma_\star a).$$

The claimed result then follows by an application of the Cramér-Wold theorem. $\square$

Remark C.1. Lemma C.1 is, of course, completely standard. I record it here because I have been unable to find a reference for a multivariate CLT for triangular arrays which permits a positive *semi*-definite limiting variance matrix.

Lemma C.2. *Let $\mathcal{G}$ be a closed subspace of $L_2(P)$ where the latter is separable and let $(g_m)_{m\in\mathbb{N}}$ denote an orthonormal basis in $\mathcal{G}$. Let for $m \in \mathbb{N}$, let Π_m denote the orthogonal projection on $\mathcal{G}_m := \operatorname{lin}\{g_1, \ldots, g_m\}$ and let Π denote the orthogonal projection on $\mathcal{G}$. Then, for any $X \in L_2(P)$ we have that $\Pi_m X \to \Pi X$ in $L_2(P)$ as $m \to \infty$.*

Proof. We first note that the formulation in the lemma is well-defined: every subspace of a separable metric space is itself separable (see e.g. Proposition 26, section 9.6 of Royden and Fitzpatrick, 2010, p. 204-205). Since a closed subspace of a Hilbert space is also a Hilbert space (with the same inner product), it follows that $\mathcal{G}$ is separable and therefore possesses an orthonormal basis (e.g. Theorem 11, Section 16.3 of Royden and Fitzpatrick, 2010, p. 317-318). Since any finite dimensional subset of a Hilbert space is closed, the orthogonal projection operators Π_m are well defined. Throughout $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ will denote the inner product in $L_2(P)$.

By proposition I.4.7 in Conway (1985, p. 15) we have that

$$\Pi_m X = \sum_{k=1}^{m}\langle X, g_k\rangle g_k.$$

ΠX is the unique vector in $\mathcal{G}$ such that $\langle X - \Pi X, g\rangle = 0$ for all $g \in \mathcal{G}$ (see e.g. I.2.6 - I.2.8 in Conway, 1985, p. 9-10). Now, let $Y = \sum_{k=1}^{\infty}\langle X, g_k\rangle g_k$ which converges by e.g. lemma I.4.12 in Conway (1985, p. 16). By continuity and linearity of the inner product we

[119] In the case that $a = 0$ this limit trivially holds.

then have that for any g_j

$$\langle X - Y, g_j \rangle = \langle X, g_j \rangle - \sum_{k=1}^{\infty} \langle X, g_k \rangle \langle g_k, g_j \rangle = \langle X, g_j \rangle - \langle X, g_j \rangle = 0.$$

Using linearity and continuity of the inner product once more permits the conclusion that $\langle X - Y, g \rangle = 0$ for any $g \in \mathcal{G}$. Hence $Y = \Pi X$. Then, we have $\Pi X - \Pi_m X = \sum_{k=m+1}^{\infty} \langle X, g_k \rangle g_k = Y - \sum_{k=1}^{m} \langle X, g_k \rangle g_k$ which converges to 0 in $L_2(P)$ by the convergence of $\sum_{k=1}^{m} \langle X, g_k \rangle g_k$ to Y. □

Lemma C.3. *Let X be an integrable random variable and Z a random element in a metric space $\mathcal{Z}$, both defined on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$. Then $\mathbb{E}[X|Z] = 0$ (P-almost surely) if and only if $\mathbb{E}[Xf(Z)] = 0$ for all square integrable functions $f : \mathcal{Z} \to \mathbb{R}$ such that $Xf(Z)$ is integrable.*

Proof. Suppose that $\mathbb{E}[X|Z] = 0$. We have

$$\mathbb{E}[Xf(Z)] = \mathbb{E}[\mathbb{E}[Xf(Z)|Z]] = \mathbb{E}[\mathbb{E}[X|Z]f(Z)] = 0.$$

Conversely suppose that $\mathbb{E}[Xf(Z)] = 0$ for all square-integrable functions $f : \mathcal{Z} \to \mathbb{R}$ with $Xf(Z)$ integrable. Let Y be any of the conditional expectations $\mathbb{E}[X|Z]$ and let $A \in \sigma(Z)$. There is a set $B \in \mathcal{B}(\mathbb{R})$ such that $A = Z^{-1}(B)$. Put f as the indicator $f(z) := \mathbf{1}\{z \in B\}$. Clearly $\mathbb{E}f(Z)^2 \leq 1$ and $Xf(Z)$ is integrable. Then, by definition,

$$\int_A Y \,\mathrm{dP} = \int_A X \,\mathrm{dP} = \int Xf(Z) \,\mathrm{dP} = \mathbb{E}[Xf(Z)] = 0.$$

Now, suppose $\{Y \neq 0\}$ has positive measure. Then one of $\{Y > 0\}$ or $\{Y < 0\}$ must. Say the first, the argument for the latter is analogous. This is $\{Y > 0\} = E = \cup_{n \geq 1} E_n$ for $E_n := \{Y > 1/n\}$. So one E_k at least has positive measure. So $\int_E Y \,\mathrm{dP} \geq \int_{E_k} Y \,\mathrm{dP} \geq \int_{E_k} 1/k \,\mathrm{dP} = \mathrm{P}(E_k)/k > 0$. But this is a contradiction since $E \in \sigma(Z)$. □

Lemma C.4. *Let $\dot{\ell}$ and $\dot{\kappa}$ be L- and K- dimensional vectors of functions in $L_2(P)$ respectively. Define $\mathscr{B} := \mathrm{lin}\{\dot{\kappa}_1, \ldots, \dot{\kappa}_K\}$ and suppose that $\mathscr{G}$ is a subspace of $L_2(P)$. For any closed subspace $S \subset L_2(P)$, denote the orthogonal projection of $X \in L_2(P)$ on S by $\Pi(X \mid S)$. Then if $\breve{X} := \Pi\left(X \mid \mathscr{G}^{\perp}\right)$ we have*

$$\tilde{\ell} := \Pi\left(\dot{\ell} \mid [\mathscr{B} + \mathscr{G}]^{\perp}\right) = \breve{\ell} - \Pi\left(\breve{\ell} \mid \mathrm{lin}\{\breve{\kappa}_1, \ldots, \breve{\kappa}_K\}\right). \tag{103}$$

Moreover, if $\tilde{I} := P\left[\tilde{\ell}\tilde{\ell}'\right]$ and $\breve{J} := P\left[\left(\breve{\ell}', \breve{\kappa}'\right)'\left(\breve{\ell}', \breve{\kappa}'\right)\right]$ and $\breve{J}_{22}$ is positive-definite then

$$\tilde{\ell} = \breve{\ell} - \breve{J}_{12}\breve{J}_{22}^{-1}\breve{\kappa}, \quad \text{and} \quad \tilde{I} = \breve{J}_{11} - \breve{J}_{12}\breve{J}_{22}^{-1}\breve{J}_{21}. \tag{104}$$

Proof. The proof of the first claim is as discussed on p. 74 of Bickel et al. (1998). As there, noting that $\mathscr{G} \subset \operatorname{lin} \mathscr{B} + \mathscr{G}$ and using their equation (A.2.11) (p. 428) we obtain

$$\begin{aligned}
\tilde{\ell} &= \dot{\ell} - \Pi\left(\dot{\ell} \mid \mathscr{G}\right) - \Pi\left(\dot{\ell} \mid (\mathscr{B} + \mathscr{G}) \cap \mathscr{G}^{\perp}\right) \\
&= \dot{\ell} - \Pi\left(\dot{\ell} \mid \mathscr{G}\right) - \Pi\left(\dot{\ell} - \Pi\left(\dot{\ell} \mid \mathscr{G}\right) \mid (\mathscr{B} + \mathscr{G}) \cap \mathscr{G}^{\perp}\right) \\
&= \breve{\ell} - \Pi\left(\breve{\ell} \mid (\mathscr{B} + \mathscr{G}) \cap \mathscr{G}^{\perp}\right).
\end{aligned}$$

Now, suppose that $f \in \operatorname{lin}\{\breve{\kappa}_1, \ldots, \breve{\kappa}_K\}$. Then we have

$$f = \sum_{k=1}^{K} a_k \breve{\kappa}_k = \sum_{k=1}^{K} a_k \dot{\kappa}_k - \sum_{k=1}^{K} a_k \Pi(\dot{\kappa}_k \mid \mathscr{G}) \in \operatorname{lin} \mathscr{B} + \mathscr{G},$$

and moreover, since each $\breve{\kappa}_k \in \mathscr{G}^{\perp}$, linearity of the inner product implies the same holds for f. Hence $f \in (\mathscr{B} + \mathscr{G}) \cap \mathscr{G}^{\perp}$. For the reverse containment, suppose that $f \in (\mathscr{B} + \mathscr{G}) \cap \mathscr{G}^{\perp}$. Then, we have for some $g \in \mathscr{G}$ that

$$f = \sum_{k=1}^{K} a_k \dot{\kappa}_k + g.$$

Now, suppose that $g \neq -\sum_{k=1}^{K} a_k \Pi\left(\dot{\kappa}_k \mid \mathscr{G}\right)$, and hence $g = -\sum_{k=1}^{K} a_k \Pi\left(\dot{\kappa}_k \mid \mathscr{G}\right) + h \neq 0$ for some $h \in \mathscr{G}$ with $h \neq 0$. Then

$$\langle f, h\rangle = \sum_{k=1}^{K} a_k \langle \dot{\kappa}_k, h\rangle - \sum_{k=1}^{K} a_k \langle \Pi(\dot{\kappa}_k \mid \mathscr{G}), h\rangle + \langle h, h\rangle = \sum_{k=1}^{K} a_k \langle \breve{\kappa}_k, h\rangle + \langle h, h\rangle = \langle h, h\rangle > 0,$$

which is a contradiction to $f \in \mathscr{G}^{\perp}$. Hence we must have $g = -\sum_{k=1}^{K} a_k \Pi\left(\dot{\kappa}_k \mid \mathscr{G}\right)$ and therefore $f = \sum_{k=1}^{K} a_k \breve{\kappa}_k \in \operatorname{lin}\{\breve{\kappa}_1, \ldots, \breve{\kappa}_K\}$. It follows that $(\mathscr{B} + \mathscr{G}) \cap \mathscr{G}^{\perp} = \operatorname{lin}\{\breve{\kappa}_1, \ldots, \breve{\kappa}_K\}$ which, in conjunction with the first display of the proof, yields (103).

Next, if $\breve{J}_{22}$ is positive definite, then the formulae in in (104) are well-defined. For the left hand side note that we have

$$P\left[\left(\breve{\ell} - \breve{J}_{12} \breve{J}_{22}^{-1} \breve{\kappa}\right) \breve{\kappa}'\right] = \breve{J}_{12} - \breve{J}_{12} \breve{J}_{22}^{-1} \breve{J}_{22} = \breve{J}_{12} - \breve{J}_{12} = 0,$$

implying that $\breve{\ell} - \breve{J}_{12} \breve{J}_{22}^{-1} \breve{\kappa}$ is the orthogonal projection of $\breve{\ell}$ onto the orthocomplement of $\operatorname{lin}\{\breve{\kappa}_1, \ldots, \breve{\kappa}_K\}$ (e.g. Conway, 1985, Theorem I.2.6) and hence satisfies the condition given in (103). The formula on the right hand side of (104) then follows by elementary calculations. □

Lemma C.5. *Suppose that X is an integrable random variable on $(\Omega, \mathcal{F}, P)$, $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ and $\sigma(\sigma(X) \cup \mathcal{H})$ is independent of $\mathcal{G}$. Then, almost surely $\mathbb{E}[X | \sigma(\mathcal{G} \cup \mathcal{H})] = E[X | \mathcal{H}]$.*

Proof. (i) $\mathbb{E}(X|\mathcal{H})$ is $\sigma(\mathcal{G} \cup \mathcal{H})$ measurable since $\mathbb{E}(X|\mathcal{H})$ is $\mathcal{H}$-measurable by definition. (ii) $\mathbb{E}(X|\mathcal{H})$ is integrable by definition of conditional expectation. (iii) We demonstrate that for each $A \in \sigma(\mathcal{G} \cup \mathcal{H})$,

$$\int_A \mathbb{E}(X|\mathcal{H}) \, \mathrm{d}P = \int_A X \, \mathrm{d}P.$$

Let $\mathcal{M} = \{B \cap C : B \in \mathcal{G}, C \in \mathcal{H}\}$. This is closed under intersections and contains Ω.

Additionally, we have that $\mathcal{G} \cup \mathcal{H} \subset \mathcal{M} \subset \sigma(\mathcal{G} \cup \mathcal{H})$ and therefore, $\sigma(\mathcal{M}) = \sigma(\mathcal{G} \cup \mathcal{H})$. Hence, by Theorem 34.1 in Billingsley (1995) it is sufficient to demonstrate $\int_{B \cap C} \mathbb{E}(X|\mathcal{H})\, \mathrm{d}P = \int_{B \cap C} X\, \mathrm{d}P$ for $B \in \mathcal{G}$ and $C \in \mathcal{H}$. To this end, suppose that $X \geq 0$ (without loss of generality, since the following argument can be applied to the two positive parts $X = X^+ + X^-$ separately and linearity used to conclude otherwise). Then, we have that

$$\int_{B \cap C} X\, \mathrm{d}P = \mathbb{E}(\mathbf{1}_B \mathbf{1}_C X) = \mathbb{E}(\mathbf{1}_B)\mathbb{E}(\mathbf{1}_C X),$$

since $\mathcal{G}$ is independent of $\sigma(\sigma(X) \cup \mathcal{H})$. Additionally,

$$\begin{aligned}
\int_{B \cap C} \mathbb{E}[X|\mathcal{H}]\, \mathrm{d}P &= \mathbb{E}\left(\mathbf{1}_B \mathbf{1}_C \mathbb{E}[X|\mathcal{H}]\right) \\
&= \mathbb{E}(\mathbf{1}_B)\mathbb{E}\left[\mathbf{1}_C \mathbb{E}[X|\mathcal{H}]\right] \\
&= \mathbb{E}(\mathbf{1}_B)\mathbb{E}\left[\mathbb{E}[\mathbf{1}_C X|\mathcal{H}]\right] \\
&= \mathbb{E}(\mathbf{1}_B)\mathbb{E}(\mathbf{1}_C X),
\end{aligned}$$

using the independence between $\mathcal{G}$ and $\mathcal{H}$, 10.10 in Davidson (1994) and the LIE. □

Lemma C.6 (Cf. Theorem 2 in Andrews, 1987)**.** *Suppose that equations* (7) *and* (8) *hold. Then,*

$$\|\hat{\mathcal{I}}^{\dagger}_{n,\theta_n} - \tilde{\mathcal{I}}^{\dagger}_{\gamma}\|_2 = o_{P_{\gamma_n}}(1).$$

Proof. Let $r := \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$ and let $\mathcal{M}$ denote the set of $d_\theta \times d_\theta$ matrices with rank r. Fix $\varepsilon > 0$ and let $\delta > 0$ be small enough that whenever $M \in \mathcal{M}$ is such that $\|\tilde{\mathcal{I}}_\gamma - M\|_2 < \delta$ we have $\|\tilde{\mathcal{I}}^{\dagger}_\gamma - M^{\dagger}\|_2 < \varepsilon$.[120] It follows that for each $n \in \mathbb{N}$,

$$\left\{\|\hat{\mathcal{I}}^{\dagger}_{n,\theta_n} - \tilde{\mathcal{I}}^{\dagger}_{\gamma}\|_2 \geq \varepsilon\right\} \subset \left\{\|\hat{\mathcal{I}}_{n,\theta_n} - \tilde{\mathcal{I}}_{\gamma}\|_2 \geq \delta\right\} \cup \left\{\operatorname{rank}(\hat{\mathcal{I}}_{n,\theta_n}) \neq r\right\},$$

and so

$$P_{\gamma_n}\left(\|\hat{\mathcal{I}}^{\dagger}_{n,\theta_n} - \tilde{\mathcal{I}}^{\dagger}_{\gamma}\|_2 \geq \varepsilon\right) \leq P_{\gamma_n}\left(\|\hat{\mathcal{I}}_{n,\theta_n} - \tilde{\mathcal{I}}_{\gamma}\|_2 \geq \delta\right) + P_{\gamma_n}\left(\operatorname{rank}(\hat{\mathcal{I}}_{n,\theta_n}) \neq r\right) \to 0.$$

□

Lemma C.7. *Suppose that equation* (7) *holds and* $\tilde{\mathcal{I}}_\gamma \succ 0$*. Then assumption* R *holds.*

Proof. The function $M \mapsto \operatorname{rank}(M)$ is lower-semicontinuous on the set of matrices of any (fixed) dimension. There is a $\delta > 0$ such that on the set $\{\|\hat{\mathcal{I}}_{n,\theta_n} - \tilde{\mathcal{I}}_\gamma\|_2 < \delta\}$, $d_\theta \geq \operatorname{rank}(\hat{\mathcal{I}}_{n,\theta_n}) \geq \operatorname{rank}(\tilde{\mathcal{I}}_\gamma) - 1/2 > d_\theta - 1$, implying $\operatorname{rank}(\tilde{\mathcal{I}}_\gamma) = d_\theta = \operatorname{rank}(\hat{\mathcal{I}}_{n,\theta_n})$. Hence, by (7)

$$P_{\gamma_n}\left(\operatorname{rank}(\hat{\mathcal{I}}_{n,\theta_n}) = \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)\right) \leq P_{\gamma_n}(\{\|\hat{\mathcal{I}}_{n,\theta_n} - \tilde{\mathcal{I}}_\gamma\|_2 < \delta\}) \to 1.$$

□

Lemma C.8. *Suppose that S is a Polish space and $(P_n)_{n\in\mathbb{N}}$ is a sequence of probability measures which converges in total variation to P, with each P_n and P defined on $(S, \mathcal{B}(S))$. If $(f_n)_{n\in\mathbb{N}}$ is a sequence of non-negative functions in $L_1(P_n)$ such that (a) $f_n \xrightarrow{P} f \in L_1(P)$ and (b) $P_n f_n \to Pf$ then $(f_n)_{n\in\mathbb{N}}$ is uniformly P_n-integrable.*

[120] See e.g. section 6.6 in Ben-Israel and Greville (2003).

Proof. Condition (a) and $P_n \xrightarrow{TV} P$ together imply that $Q_n \rightsquigarrow Q$ where Q_n is the pushforward measure of P_n under f_n and Q the same of P under f. Let $h \in C_b(S)$. By change of variables (e.g. Bogachev, 2007, Theorem 3.6.1) $\int h \, \mathrm{d}Q_n = \int h(f_n) \, \mathrm{d}P_n$ and $\int g \, \mathrm{d}Q = \int h(f) \, \mathrm{d}P$. By (a) and the bounded convergence theorem, $\int h(f_n) \, \mathrm{d}P \to \int h(f) \, \mathrm{d}P$. By $P_n \xrightarrow{TV} P$

$$\left| \int h(f_n) \, \mathrm{d}P_n - \int h(f_n) \, \mathrm{d}P \right| \leq 2\bar{h} \sup \left\{ \left| \int g \, \mathrm{d}P_n - \int g \, \mathrm{d}P \right| \right\} \to 0,$$

where $|h| \leq \bar{h} \in (0, \infty)$ and the supremum is taken over all measurable g with $0 \leq g \leq 1$. Hence $Q_n \rightsquigarrow Q$ as claimed. This, in conjunction with (b), Theorem 3.6 of Billingsley (1999) and translating terms yields the result. □

D Tables & figures

D.1 Empirical rejection frequencies (ERF)

D.1.1 SIM

Table 1: ERF (%) $\epsilon \sim \mathcal{N}(0,1)$, $X_k \sim U(-1,1)$

		$\hat{S}$			W		
n	δ^{-1}	$f=\delta f_1$	$f=\delta f_2$	$f=\delta f_3$	$f=\delta f_1$	$f=\delta f_2$	$f=\delta f_3$
200	$\sqrt{1}$	5.24	6.58	6.14	15.94	14.38	18.92
400	$\sqrt{1}$	5.38	5.20	5.40	10.28	10.14	13.82
600	$\sqrt{1}$	5.50	5.70	5.14	8.06	7.88	11.22
800	$\sqrt{1}$	4.74	4.76	5.36	6.94	7.78	10.28
200	$\sqrt{2}$	5.46	5.36	5.38	17.62	15.18	19.90
400	$\sqrt{2}$	5.58	5.68	5.58	12.72	10.26	14.58
600	$\sqrt{2}$	4.60	5.48	5.42	10.66	9.14	13.20
800	$\sqrt{2}$	5.20	5.34	5.74	9.20	8.98	10.60
200	$\sqrt{4}$	5.22	5.50	5.62	20.86	19.10	24.62
400	$\sqrt{4}$	4.98	5.86	5.60	14.68	12.62	17.04
600	$\sqrt{4}$	4.92	5.20	5.52	12.80	9.82	15.10
800	$\sqrt{4}$	5.48	4.96	6.02	10.48	9.32	13.08
200	$\sqrt{8}$	5.12	5.34	5.60	16.28	22.52	26.20
400	$\sqrt{8}$	5.98	5.50	5.12	19.48	16.12	19.98
600	$\sqrt{8}$	5.62	5.00	6.48	15.24	14.18	16.94
800	$\sqrt{8}$	4.98	5.54	5.40	13.02	11.76	14.42
200	$\sqrt{16}$	4.82	5.64	5.22	12.28	20.08	21.76
400	$\sqrt{16}$	5.28	5.30	6.02	15.66	18.66	23.66
600	$\sqrt{16}$	4.58	5.46	5.62	19.30	15.68	19.64
800	$\sqrt{16}$	5.30	5.56	5.32	17.02	14.68	17.62

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 4.4. $f_1(v) = c_1(1+\exp(-v))^{-1}$, $f_2(v) = c_2\exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1,2,3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null.

Table 2: ERF (%), $\epsilon|\xi \sim \sqrt{5}(-1)^{\xi}\,\text{Beta}(2,3)$, $\xi \sim \text{Bernoulli}(1/2)$, $X_k \sim U(-1,1)$

		$\hat{S}$			W		
n	δ^{-1}	$f=\delta f_1$	$f=\delta f_2$	$f=\delta f_3$	$f=\delta f_1$	$f=\delta f_2$	$f=\delta f_3$
200	$\sqrt{1}$	4.82	5.56	5.94	14.72	12.88	16.98
400	$\sqrt{1}$	5.74	4.96	5.50	10.28	10.68	12.42
600	$\sqrt{1}$	4.78	4.98	5.08	7.98	8.52	10.56
800	$\sqrt{1}$	5.14	4.88	5.34	7.06	7.78	9.58
200	$\sqrt{2}$	4.82	5.84	5.94	17.06	15.38	19.58
400	$\sqrt{2}$	5.14	5.86	5.52	11.86	10.02	14.20
600	$\sqrt{2}$	5.18	5.26	5.46	9.72	9.22	12.84
800	$\sqrt{2}$	5.04	5.12	5.40	8.72	8.60	11.90
200	$\sqrt{4}$	5.26	5.48	5.78	19.84	18.44	22.34
400	$\sqrt{4}$	5.64	5.38	5.62	15.18	12.20	16.02
600	$\sqrt{4}$	6.18	5.66	5.64	10.92	10.18	15.18
800	$\sqrt{4}$	4.88	5.26	4.84	10.12	9.52	13.24
200	$\sqrt{8}$	5.10	5.38	5.08	15.36	20.18	25.64
400	$\sqrt{8}$	4.66	5.58	4.96	19.08	16.20	20.44
600	$\sqrt{8}$	5.22	4.92	5.52	15.14	13.08	16.36
800	$\sqrt{8}$	5.10	4.98	5.66	12.64	11.00	14.78
200	$\sqrt{16}$	5.28	4.76	5.60	12.58	18.62	21.90
400	$\sqrt{16}$	5.54	5.56	5.34	15.38	19.14	23.40
600	$\sqrt{16}$	5.24	5.20	5.32	18.08	14.98	20.26
800	$\sqrt{16}$	4.92	5.30	5.02	17.54	13.60	18.08

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 4.4. $f_1(v) = c_1(1+\exp(-v))^{-1}$, $f_2(v) = c_2\exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null.

Table 3: ERF (%), $\epsilon \sim \mathcal{N}(0,1)$, $X = (Z_1, 0.2Z_1 + 0.4Z_2 + 0.8)$, $Z_k \sim U(-1,1)$

		$\hat{S}$			W		
n	δ^{-1}	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$
200	$\sqrt{1}$	5.28	5.56	6.52	14.74	15.76	14.42
400	$\sqrt{1}$	6.20	5.94	5.96	10.62	10.88	10.68
600	$\sqrt{1}$	5.64	5.62	5.70	9.28	9.00	9.06
800	$\sqrt{1}$	5.10	5.80	5.00	7.28	8.78	8.18
200	$\sqrt{2}$	6.14	5.62	5.80	17.74	20.14	16.92
400	$\sqrt{2}$	5.62	5.96	6.52	12.08	14.02	11.02
600	$\sqrt{2}$	5.70	5.26	5.66	9.72	11.16	9.94
800	$\sqrt{2}$	5.38	5.08	5.78	9.68	10.34	9.02
200	$\sqrt{4}$	6.20	5.44	5.32	20.84	25.02	20.26
400	$\sqrt{4}$	5.64	5.62	5.90	15.70	16.82	14.22
600	$\sqrt{4}$	5.24	5.54	5.88	12.20	13.08	11.32
800	$\sqrt{4}$	5.68	5.74	5.38	11.18	13.14	10.62
200	$\sqrt{8}$	5.42	5.88	5.54	15.70	25.26	16.86
400	$\sqrt{8}$	5.82	5.42	5.32	17.24	21.64	17.42
600	$\sqrt{8}$	5.80	5.84	5.94	15.82	16.56	15.24
800	$\sqrt{8}$	5.44	5.68	5.60	13.14	15.14	13.14
200	$\sqrt{16}$	5.52	5.94	5.86	12.32	20.14	12.94
400	$\sqrt{16}$	6.18	5.68	5.58	16.06	24.22	15.98
600	$\sqrt{16}$	5.76	5.72	5.66	17.90	22.20	16.80
800	$\sqrt{16}$	5.24	5.28	5.02	17.40	19.54	15.38

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 4.4. $f_1(v) = c_1(1+\exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1,2,3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null.

Table 4: ERF (%), $\epsilon|\xi \sim \sqrt{5}(-1)^{\xi}\,\text{Beta}(2,3)$, $\xi \sim \text{Bernoulli}(1/2)$, $X = (Z_1, 0.2Z_1 + 0.4Z_2 + 0.8)$, $Z_k \sim U(-1,1)$

		$\hat{S}$			W		
n	δ^{-1}	$f=\delta f_1$	$f=\delta f_2$	$f=\delta f_3$	$f=\delta f_1$	$f=\delta f_2$	$f=\delta f_3$
200	$\sqrt{1}$	5.26	5.92	6.18	14.78	15.28	13.60
400	$\sqrt{1}$	5.50	5.84	5.54	10.44	10.90	9.50
600	$\sqrt{1}$	5.22	5.70	5.36	8.62	9.14	8.28
800	$\sqrt{1}$	5.26	5.32	5.90	8.26	9.72	8.40
200	$\sqrt{2}$	5.96	6.00	6.02	17.62	19.86	15.54
400	$\sqrt{2}$	5.18	5.16	5.96	12.32	14.40	11.10
600	$\sqrt{2}$	5.22	6.02	5.34	10.86	10.58	9.14
800	$\sqrt{2}$	5.38	4.96	6.02	8.94	10.44	8.36
200	$\sqrt{4}$	5.96	6.26	5.58	20.32	24.04	20.48
400	$\sqrt{4}$	5.78	6.40	6.00	15.26	16.46	13.52
600	$\sqrt{4}$	5.30	5.26	5.60	13.16	13.72	11.06
800	$\sqrt{4}$	5.18	5.62	5.04	10.12	12.38	9.56
200	$\sqrt{8}$	5.72	5.78	5.72	15.14	25.52	16.50
400	$\sqrt{8}$	5.24	5.54	6.14	18.22	21.88	17.82
600	$\sqrt{8}$	5.76	4.96	5.10	15.18	17.34	14.70
800	$\sqrt{8}$	5.46	5.48	5.82	14.26	15.30	13.28
200	$\sqrt{16}$	5.66	5.16	5.96	11.42	20.78	12.82
400	$\sqrt{16}$	5.66	5.84	6.00	15.58	24.86	16.28
600	$\sqrt{16}$	5.00	4.78	5.98	17.44	22.06	16.72
800	$\sqrt{16}$	5.60	5.64	5.36	16.78	19.94	15.90

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 4.4. $f_1(v) = c_1(1+\exp(-v))^{-1}$, $f_2(v) = c_2\exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1,2,3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null.

Table 5: ERF (%), $\epsilon \sim \mathcal{N}(0, s_1 \log(2 + (X_1 + X_2\theta)^2))$, $X_k \sim U(-1, 1)$, $\breve{\omega}(X) = \omega(X)$

		$\hat{S}$			W		
n	δ^{-1}	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$
200	$\sqrt{1}$	6.38	6.64	6.20	18.72	16.76	23.46
400	$\sqrt{1}$	6.24	5.84	6.50	12.34	11.70	17.26
600	$\sqrt{1}$	5.78	5.12	5.72	10.38	10.96	14.70
800	$\sqrt{1}$	5.88	5.58	5.92	8.50	9.94	12.76
200	$\sqrt{2}$	5.76	5.76	6.12	22.62	19.30	25.86
400	$\sqrt{2}$	5.96	6.22	6.26	16.30	13.72	20.08
600	$\sqrt{2}$	5.52	5.46	6.26	14.46	11.70	15.70
800	$\sqrt{2}$	5.34	5.94	5.68	11.26	10.14	14.78
200	$\sqrt{4}$	5.32	5.72	5.44	27.12	24.36	30.40
400	$\sqrt{4}$	5.42	5.96	6.12	21.06	16.28	22.48
600	$\sqrt{4}$	5.24	5.52	5.74	15.50	13.38	19.58
800	$\sqrt{4}$	5.74	5.72	5.76	13.74	11.16	17.78
200	$\sqrt{8}$	5.40	5.64	5.46	19.66	25.36	30.08
400	$\sqrt{8}$	6.60	6.22	6.32	25.42	21.10	28.72
600	$\sqrt{8}$	5.50	5.80	6.60	21.34	17.78	23.80
800	$\sqrt{8}$	5.42	5.84	6.06	17.86	15.58	21.08
200	$\sqrt{16}$	5.86	6.26	5.74	14.06	23.96	25.06
400	$\sqrt{16}$	5.52	6.50	6.46	20.32	23.98	29.78
600	$\sqrt{16}$	5.50	5.74	5.08	25.04	22.00	29.20
800	$\sqrt{16}$	5.28	4.82	5.24	22.90	19.90	25.40

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the psuedo efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 4.4. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1, 2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null.

Table 6: ERF (%), $\epsilon \sim \mathcal{N}(0, s_2(1 + 5\sin(X_1)^2))$, $X_k \sim U(-1, 1)$, $\breve{\omega}(X) = \omega(X)$

		$\hat{S}$			W		
n	δ^{-1}	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$
200	$\sqrt{1}$	5.40	5.94	6.46	19.10	16.34	18.46
400	$\sqrt{1}$	6.68	6.34	7.42	13.36	11.24	13.72
600	$\sqrt{1}$	5.94	6.14	6.00	10.28	8.74	10.88
800	$\sqrt{1}$	5.86	5.70	5.78	8.86	7.68	9.76
200	$\sqrt{2}$	5.12	5.32	5.70	23.74	19.96	22.58
400	$\sqrt{2}$	5.42	6.28	6.62	15.70	12.92	15.72
600	$\sqrt{2}$	5.92	6.00	5.92	12.66	10.44	12.86
800	$\sqrt{2}$	5.68	5.76	5.78	10.38	9.58	11.90
200	$\sqrt{4}$	5.64	6.50	5.94	23.30	22.86	25.92
400	$\sqrt{4}$	5.48	5.82	6.84	19.76	16.60	18.44
600	$\sqrt{4}$	5.82	5.74	6.24	15.70	13.08	14.30
800	$\sqrt{4}$	5.80	5.82	6.18	13.86	12.16	12.54
200	$\sqrt{8}$	5.98	5.70	5.50	14.74	23.00	28.56
400	$\sqrt{8}$	5.48	6.50	5.78	22.32	20.00	23.70
600	$\sqrt{8}$	5.46	5.76	6.24	20.56	16.76	19.02
800	$\sqrt{8}$	5.36	6.00	6.18	17.94	13.74	16.50
200	$\sqrt{16}$	4.96	6.20	5.42	12.96	18.18	26.24
400	$\sqrt{16}$	5.42	6.50	6.70	12.78	21.82	25.66
600	$\sqrt{16}$	5.20	5.86	5.58	18.30	21.24	23.82
800	$\sqrt{16}$	5.06	5.66	5.92	21.44	18.76	20.98

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the psuedo efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 4.4. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1, 2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null.

Table 7: ERF (%), $\epsilon \sim \mathcal{N}(0, s_1 \log(2 + (X_1 + X_2\theta)^2))$, $X = (Z_1, 0.2Z_1 + 0.4Z_2 + 0.8)$, $Z_k \sim U(-1,1)$, $\breve{\omega}(X) = \omega(X)$

		$\hat{S}$			W		
n	δ^{-1}	$f=\delta f_1$	$f=\delta f_2$	$f=\delta f_3$	$f=\delta f_1$	$f=\delta f_2$	$f=\delta f_3$
200	$\sqrt{1}$	5.08	5.98	6.10	15.22	16.62	16.40
400	$\sqrt{1}$	5.06	5.74	5.54	9.62	11.76	12.00
600	$\sqrt{1}$	5.56	5.94	5.84	8.18	11.02	10.86
800	$\sqrt{1}$	5.02	5.58	5.44	8.00	9.02	9.50
200	$\sqrt{2}$	5.70	5.62	5.50	17.94	19.58	19.94
400	$\sqrt{2}$	5.92	5.80	6.06	12.90	13.24	14.08
600	$\sqrt{2}$	6.20	6.02	5.38	9.52	11.22	11.54
800	$\sqrt{2}$	5.60	5.70	5.48	8.78	10.76	9.78
200	$\sqrt{4}$	5.66	6.02	5.50	20.92	24.00	22.28
400	$\sqrt{4}$	5.90	5.68	5.86	16.50	16.98	17.84
600	$\sqrt{4}$	5.08	5.40	5.92	12.20	14.42	14.44
800	$\sqrt{4}$	5.32	4.88	5.72	10.74	11.96	12.54
200	$\sqrt{8}$	5.62	5.36	5.56	18.02	26.58	17.74
400	$\sqrt{8}$	5.90	5.76	5.44	19.70	21.66	20.64
600	$\sqrt{8}$	5.70	5.86	5.76	16.72	17.70	18.04
800	$\sqrt{8}$	5.42	5.18	5.26	13.30	14.92	14.82
200	$\sqrt{16}$	5.20	5.18	5.30	12.16	21.54	15.70
400	$\sqrt{16}$	5.58	5.26	5.80	17.04	25.38	18.52
600	$\sqrt{16}$	5.68	5.42	5.88	18.78	22.58	20.06
800	$\sqrt{16}$	5.08	5.26	5.46	17.80	19.20	18.82

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the psuedo efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 4.4. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1, 2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null.

Table 8: ERF (%), $\epsilon \sim \mathcal{N}(0, s_2(1 + 5\sin(X_1)^2))$, $X = (Z_1, 0.2Z_1 + 0.4Z_2 + 0.8)$, $Z_k \sim U(-1, 1)$, $\breve{\omega}(X) = \omega(X)$

		$\hat{S}$			W		
n	δ^{-1}	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$
200	$\sqrt{1}$	4.74	5.34	5.66	14.86	14.12	17.52
400	$\sqrt{1}$	5.60	6.28	6.12	9.34	10.24	10.12
600	$\sqrt{1}$	5.66	6.00	5.48	6.82	7.76	7.62
800	$\sqrt{1}$	5.82	6.42	5.64	6.70	7.54	6.62
200	$\sqrt{2}$	5.16	5.56	5.84	19.24	17.10	20.10
400	$\sqrt{2}$	6.38	6.14	6.38	11.50	11.92	12.28
600	$\sqrt{2}$	5.62	5.08	6.02	8.34	9.38	9.98
800	$\sqrt{2}$	5.50	6.10	5.50	8.14	8.94	7.42
200	$\sqrt{4}$	5.88	5.58	5.48	24.48	22.80	23.50
400	$\sqrt{4}$	6.10	6.04	6.10	15.04	15.08	15.72
600	$\sqrt{4}$	5.98	6.32	5.84	11.54	11.70	11.30
800	$\sqrt{4}$	5.68	5.82	5.62	9.24	10.88	9.78
200	$\sqrt{8}$	5.48	4.96	5.26	24.04	27.46	24.94
400	$\sqrt{8}$	5.42	5.42	5.94	20.26	19.66	20.38
600	$\sqrt{8}$	5.74	5.58	5.76	16.10	15.44	15.22
800	$\sqrt{8}$	5.40	5.08	5.60	12.26	12.80	13.88
200	$\sqrt{16}$	5.50	4.60	5.10	17.84	22.32	22.20
400	$\sqrt{16}$	5.38	5.80	5.66	20.82	23.02	22.36
600	$\sqrt{16}$	5.54	5.86	5.56	19.78	19.58	21.86
800	$\sqrt{16}$	5.70	5.80	5.60	17.88	17.00	17.14

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the psuedo efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 4.4. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1, 2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null.

Table 9: ERF (%), $\epsilon \sim \mathcal{N}(0, s_1 \log(2 + (X_1 + X_2\theta)^2))$, $X_k \sim U(-1, 1)$, $\breve{\omega}(X) = 1$

		$\hat{S}$			W		
n	δ^{-1}	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$
200	$\sqrt{1}$	4.86	5.74	5.62	22.22	19.88	23.22
400	$\sqrt{1}$	5.64	5.20	6.04	15.80	13.76	18.40
600	$\sqrt{1}$	5.10	5.72	5.50	12.08	12.14	14.68
800	$\sqrt{1}$	5.32	4.88	5.32	10.82	11.06	13.50
200	$\sqrt{2}$	4.68	5.90	5.98	26.22	23.80	27.42
400	$\sqrt{2}$	5.18	5.72	6.44	19.14	15.68	20.74
600	$\sqrt{2}$	5.26	5.72	5.24	15.98	13.20	17.00
800	$\sqrt{2}$	5.28	5.28	6.16	14.00	12.24	16.22
200	$\sqrt{4}$	5.78	5.18	5.54	29.72	27.44	32.58
400	$\sqrt{4}$	5.88	5.46	6.14	24.32	19.24	24.88
600	$\sqrt{4}$	5.34	5.14	6.18	20.10	15.92	19.62
800	$\sqrt{4}$	5.14	5.28	5.10	17.86	14.08	18.12
200	$\sqrt{8}$	6.02	5.74	6.18	23.12	29.98	32.70
400	$\sqrt{8}$	5.44	5.34	5.94	29.00	26.08	29.76
600	$\sqrt{8}$	5.52	5.72	5.04	25.26	20.50	24.70
800	$\sqrt{8}$	5.16	5.70	6.18	21.74	17.42	22.78
200	$\sqrt{16}$	5.48	5.16	5.40	15.62	25.38	28.04
400	$\sqrt{16}$	5.78	5.50	5.86	23.62	28.28	33.34
600	$\sqrt{16}$	5.02	4.74	6.10	28.38	25.90	30.54
800	$\sqrt{16}$	5.00	5.14	5.28	27.00	21.72	26.24

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the psuedo efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 4.4. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1, 2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null.

Table 10: ERF (%), $\epsilon \sim \mathcal{N}(0, s_2(1 + 5\sin(X_1)^2))$, $X_k \sim U(-1,1)$, $\breve{\omega}(X) = 1$

		$\hat{S}$			W		
n	δ^{-1}	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$
200	$\sqrt{1}$	5.10	5.34	6.56	18.52	18.06	22.40
400	$\sqrt{1}$	5.90	5.52	5.26	13.12	12.60	15.68
600	$\sqrt{1}$	5.28	5.10	5.36	9.94	10.36	13.10
800	$\sqrt{1}$	5.08	5.18	5.06	9.10	9.58	12.78
200	$\sqrt{2}$	5.48	5.86	5.86	21.18	19.64	23.92
400	$\sqrt{2}$	5.64	5.14	5.64	15.58	13.28	18.48
600	$\sqrt{2}$	4.70	5.86	5.52	11.58	11.48	14.84
800	$\sqrt{2}$	5.36	5.34	5.20	11.18	10.54	13.80
200	$\sqrt{4}$	4.84	5.22	5.78	21.96	23.54	27.20
400	$\sqrt{4}$	5.52	6.26	6.32	19.00	16.60	20.88
600	$\sqrt{4}$	5.18	5.76	5.14	15.90	13.58	18.66
800	$\sqrt{4}$	5.34	4.88	5.56	13.58	11.90	16.62
200	$\sqrt{8}$	4.86	5.92	5.30	15.86	23.46	27.62
400	$\sqrt{8}$	4.96	5.36	5.78	22.28	20.46	25.90
600	$\sqrt{8}$	5.22	5.66	5.44	19.80	16.18	21.58
800	$\sqrt{8}$	5.10	5.24	5.28	17.08	15.36	19.78
200	$\sqrt{16}$	5.10	5.42	5.68	12.16	17.86	20.54
400	$\sqrt{16}$	5.50	5.70	5.60	13.54	23.14	27.24
600	$\sqrt{16}$	5.54	5.36	5.98	18.22	20.12	25.44
800	$\sqrt{16}$	4.40	5.38	5.00	20.90	18.50	23.26

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the psuedo efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 4.4. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1, 2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null.

Table 11: ERF (%), $\epsilon \sim \mathcal{N}(0, s_1 \log(2 + (X_1 + X_2\theta)^2))$, $X = (Z_1, 0.2Z_1 + 0.4Z_2 + 0.8)$, $Z_k \sim U(-1,1)$, $\breve{\omega}(X) = 1$

		$\hat{S}$			W		
n	δ^{-1}	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$
200	$\sqrt{1}$	6.14	5.46	6.16	17.80	18.88	17.66
400	$\sqrt{1}$	6.24	6.10	5.98	12.54	13.54	12.90
600	$\sqrt{1}$	6.02	5.58	6.44	10.78	11.70	9.96
800	$\sqrt{1}$	5.66	5.42	5.26	10.44	10.90	9.48
200	$\sqrt{2}$	6.08	5.62	5.42	22.22	22.46	20.82
400	$\sqrt{2}$	5.58	5.12	6.00	16.24	16.44	13.68
600	$\sqrt{2}$	5.64	5.66	6.02	12.46	13.22	11.64
800	$\sqrt{2}$	6.08	5.88	5.42	11.96	12.94	10.28
200	$\sqrt{4}$	6.04	5.98	6.12	26.00	28.62	21.70
400	$\sqrt{4}$	5.94	5.60	5.48	19.68	20.80	17.68
600	$\sqrt{4}$	6.10	5.44	5.54	16.54	16.96	14.42
800	$\sqrt{4}$	5.34	5.32	5.74	13.46	15.26	12.44
200	$\sqrt{8}$	5.36	5.72	5.44	19.90	28.44	17.34
400	$\sqrt{8}$	6.36	5.74	5.72	22.72	26.44	20.62
600	$\sqrt{8}$	5.82	5.68	4.98	19.84	20.78	17.80
800	$\sqrt{8}$	4.98	5.36	5.80	17.74	18.96	15.46
200	$\sqrt{16}$	4.90	5.42	5.28	15.04	23.14	15.82
400	$\sqrt{16}$	5.66	5.40	6.06	20.76	28.06	17.40
600	$\sqrt{16}$	5.64	5.26	5.72	22.92	26.58	19.36
800	$\sqrt{16}$	4.84	5.20	5.00	20.84	23.30	18.86

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the psuedo efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 4.4. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1, 2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null.

Table 12: ERF (%), $\epsilon \sim \mathcal{N}(0, s_2(1 + 5\sin(X_1)^2))$, $X = (Z_1, 0.2Z_1 + 0.4Z_2 + 0.8)$, $Z_k \sim U(-1, 1)$, $\breve{\omega}(X) = 1$

		$\hat{S}$			W		
n	δ^{-1}	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$
200	$\sqrt{1}$	6.20	6.34	5.80	18.88	21.20	19.30
400	$\sqrt{1}$	6.36	5.90	5.40	15.04	16.58	13.84
600	$\sqrt{1}$	5.40	5.74	5.24	12.34	14.32	12.60
800	$\sqrt{1}$	5.54	5.66	5.22	10.64	12.30	10.66
200	$\sqrt{2}$	5.72	5.98	6.70	23.24	25.90	23.48
400	$\sqrt{2}$	5.64	6.10	5.82	16.66	19.28	16.56
600	$\sqrt{2}$	5.28	5.22	5.64	14.12	16.08	13.76
800	$\sqrt{2}$	5.92	5.66	6.02	12.94	14.52	11.92
200	$\sqrt{4}$	5.94	6.46	6.12	29.54	29.14	27.76
400	$\sqrt{4}$	6.08	6.16	5.78	21.66	24.08	20.16
600	$\sqrt{4}$	5.10	5.80	5.56	17.50	18.74	14.90
800	$\sqrt{4}$	5.24	5.76	5.32	16.62	18.08	14.58
200	$\sqrt{8}$	6.30	5.96	5.82	26.38	34.50	25.68
400	$\sqrt{8}$	5.64	5.30	5.84	25.76	28.60	24.70
600	$\sqrt{8}$	5.52	5.84	5.72	22.16	23.56	20.06
800	$\sqrt{8}$	5.20	5.74	5.12	18.92	21.02	17.36
200	$\sqrt{16}$	5.44	5.06	6.18	15.94	28.06	18.10
400	$\sqrt{16}$	5.36	5.80	6.50	26.70	33.90	26.38
600	$\sqrt{16}$	5.04	6.00	5.46	26.72	29.04	24.70
800	$\sqrt{16}$	5.46	5.84	5.46	23.14	26.78	22.62

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the psuedo efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 4.4. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1, 2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null.

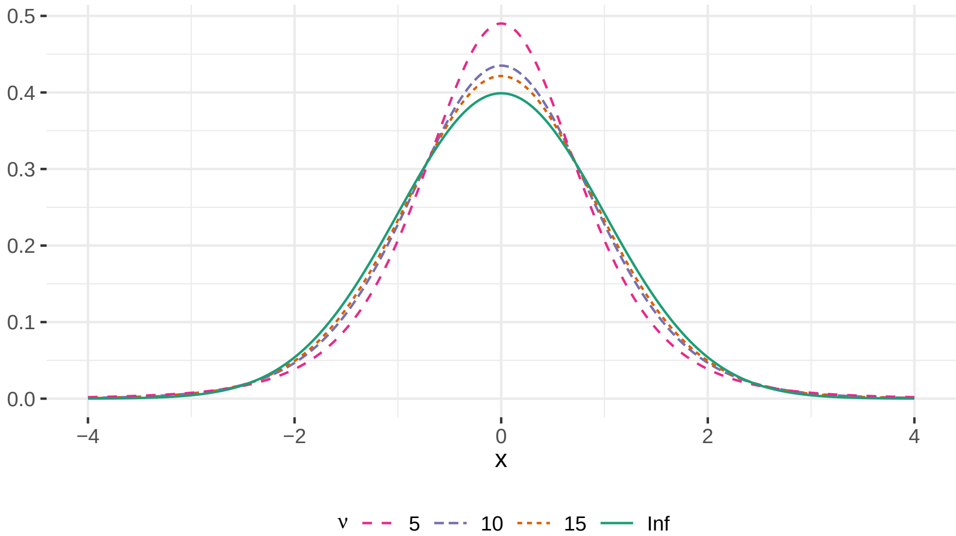
D.1.2 LSEM

Table 13: True error distributions

	η_1		η_2
a	$\mathcal{N}(0,\,1)$	$0-1$	$\mathcal{N}(0,\,1)$
b	$t'(5)$	$1-1$	$t'(5)$
c	$\mathcal{SN}'(0,\,1,\,4)$	$1-2$	$t'(10)$
–	–	$1-3$	$t'(15)$
–	–	$2-1$	$\mathcal{SN}'(0,\,1,\,4)$
–	–	$2-2$	$\mathcal{SN}'(0,\,1,\,3)$
–	–	$2-3$	$\mathcal{SN}'(0,\,1,\,2)$
–	–	$3-1$	${}^{3}\!/_{4}\mathcal{N}(0,1)+{}^{1}\!/_{4}\mathcal{N}({}^{3}\!/_{2},{}^{1}\!/_{9})$
–	–	$3-2$	${}^{17}\!/_{20}\mathcal{N}(0,1)+{}^{3}\!/_{20}\mathcal{N}({}^{3}\!/_{2},{}^{1}\!/_{9})$
–	–	$3-3$	${}^{19}\!/_{20}\mathcal{N}(0,1)+{}^{1}\!/_{20}\mathcal{N}({}^{3}\!/_{2},{}^{1}\!/_{9})$

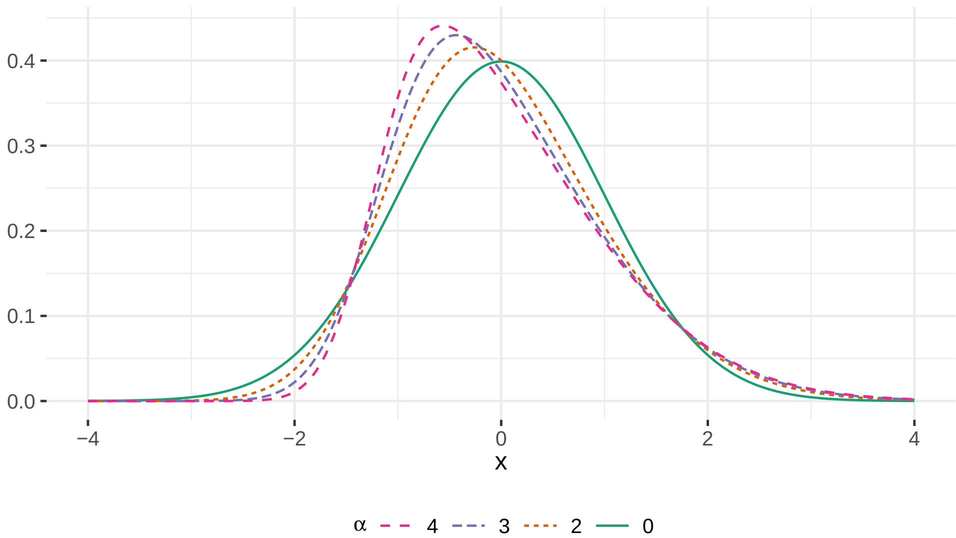
Notes: $\mathcal{SN}(\mu,\sigma,\alpha)$ denotes the skew normal distribution with location μ, scale σ and shape α. t' and $\mathcal{SN}'$ indicate that the corresponding t and skew normal distributions have been normalised to have zero mean and unit variance. The mixutre density in the right hand column is based on the "Skewed bimodal" density in Marron and Wand (1992).

Figure 1: Density function of $t'(\nu)$



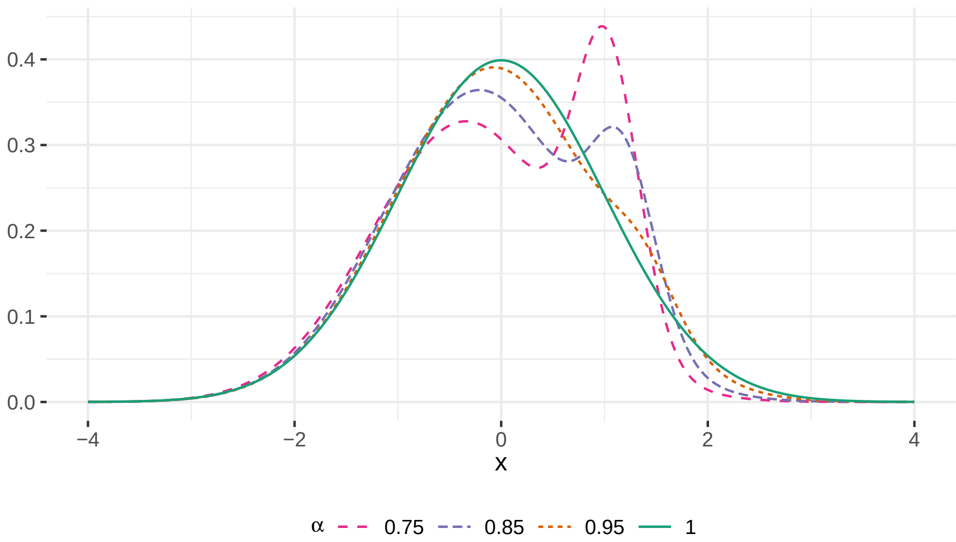
Densities $1 - \mathrm{j}$ for $j = 1, 2, 3$ in table 13.

Figure 2: Density function of $\mathcal{SN}'(0, 1, \alpha)$



Densities $2 - \mathrm{j}$ for $j = 1, 2, 3$ in table 13.

Figure 3: Density function of $\alpha\mathcal{N}(0,1) + (1-\alpha)\mathcal{N}(3/2, 1/9)$



Densities $3-\mathrm{j}$ for $j=1,2,3$ in table 13.

Figure 4: Density functions for distributions used in LSEM simulation study (ii).

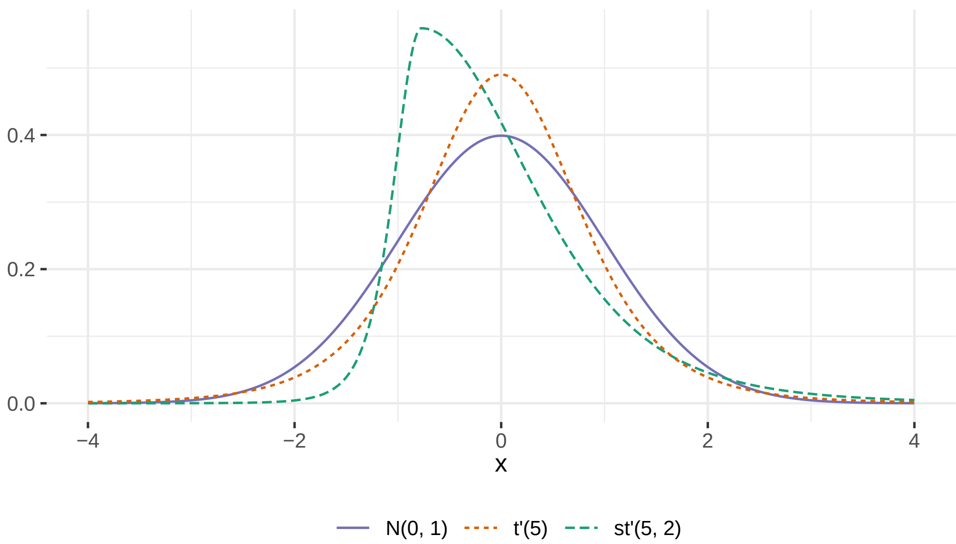


Table 14: Empirical rejection frequencies (%) for LSEM, $\epsilon_1 \sim \mathcal{N}(0,1)$

n	$0-1$	$1-1$	$1-2$	$1-3$	$2-1$	$2-2$	$2-3$	$3-1$	$3-2$	$3-3$
$\hat{S}$										
200	4.74	5.62	5.90	5.50	3.20	3.86	4.48	2.92	3.58	4.62
400	4.78	5.52	4.44	5.16	2.82	3.68	4.66	1.92	3.58	4.24
600	4.60	4.84	4.20	4.74	2.50	3.42	3.76	2.18	3.34	4.56
800	4.56	4.28	4.48	4.12	2.62	2.94	3.56	2.52	3.86	4.16
$\hat{S}^*$										
200	6.94	6.58	6.76	7.26	6.74	6.78	6.46	7.10	7.00	6.88
400	6.82	6.66	6.44	6.76	8.02	7.36	7.74	5.94	7.12	6.46
600	7.04	7.32	5.86	6.58	8.60	7.80	6.68	6.50	6.74	6.82
800	6.68	6.38	6.48	6.04	8.68	7.50	5.84	5.74	7.20	6.82
$\hat{W}$										
200	24.36	4.02	10.10	14.22	18.32	18.80	20.20	58.96	48.82	32.56
400	24.50	1.90	6.20	10.56	15.20	16.32	17.24	74.46	61.24	36.14
600	23.76	2.22	5.24	9.38	14.70	14.58	16.48	84.72	71.18	39.54
800	24.14	2.04	3.52	7.90	13.60	12.62	14.96	90.12	77.34	43.04
$L\hat{M}$										
200	4.96	4.86	4.90	5.32	5.08	5.32	4.78	5.28	5.44	4.74
400	5.42	4.88	5.08	5.30	4.50	5.88	5.14	5.38	4.86	5.10
600	5.14	5.54	5.34	5.28	5.18	5.36	5.32	4.84	5.08	5.22
800	5.14	4.80	4.60	4.82	4.44	4.84	4.78	4.68	5.36	5.42
$\tilde{W}$										
200	27.38	32.18	30.20	29.80	28.28	29.48	28.76	23.10	24.40	25.50
400	25.26	30.24	28.76	27.92	27.60	27.88	26.58	21.94	22.70	22.60
600	23.82	28.14	27.54	28.14	26.02	26.12	26.74	18.76	20.78	21.68
800	23.14	26.86	26.94	25.86	26.62	26.54	25.64	16.88	20.26	20.86
$L\tilde{M}$										
200	30.52	35.66	32.88	31.76	31.16	32.52	31.28	22.90	24.66	28.04
400	21.64	27.26	23.34	23.56	22.84	22.74	22.38	14.66	15.86	17.74
600	16.64	22.86	18.76	19.58	18.30	18.46	20.16	9.10	11.02	14.36
800	14.72	19.68	15.72	15.88	16.60	16.70	16.50	6.84	8.52	11.52

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test computed using OLS estimates of β; $\hat{S}^*$ is the efficient score test computed using 1-step updates from the OLS estimates. $\hat{\mathrm{W}}$, $\hat{\mathrm{LM}}$ denote the Wald and LM tests based on a psuedo-maximum likelihood estimator inspired by Gouriéroux et al. (2017). $\tilde{\mathrm{W}}$ and $\tilde{\mathrm{LM}}$ denote the Wald and LM tests based on a GMM estimator inspired by Lanne and Luoto (2021). Columns 2 – 14 denote the choice of density for ϵ_2, as in Table 13.

Table 15: Empirical rejection frequencies (%) for LSEM, $\epsilon_1 \sim t'(5)$

n	$0-1$	$1-1$	$1-2$	$1-3$	$2-1$	$2-2$	$2-3$	$3-1$	$3-2$	$3-3$
$\hat{S}$										
200	6.16	7.58	6.10	6.00	3.96	4.92	5.40	3.20	4.32	5.74
400	5.40	6.76	5.72	5.86	3.54	4.50	5.06	4.06	3.90	5.34
600	4.96	5.58	5.32	6.06	3.52	4.18	4.82	3.26	4.10	5.50
800	5.04	5.48	5.32	5.58	3.70	4.34	4.78	3.20	4.14	4.80
$\hat{S}^*$										
200	7.24	7.20	6.52	6.88	7.70	7.56	6.92	6.92	7.00	7.20
400	6.38	7.22	6.18	6.52	7.74	6.96	6.70	6.74	6.24	6.52
600	5.64	6.04	5.96	6.72	7.08	6.68	6.28	5.30	5.60	6.42
800	6.12	6.50	6.10	6.32	6.74	7.18	6.40	5.58	5.44	5.68
$\hat{W}$										
200	4.04	1.78	2.42	2.48	3.00	2.78	3.02	9.24	7.46	4.66
400	2.26	2.38	2.14	2.06	2.22	2.44	2.10	6.40	4.10	2.70
600	2.22	2.38	2.32	2.42	2.16	2.20	1.98	5.02	2.98	2.26
800	1.92	2.78	3.14	2.98	2.36	2.32	2.42	3.10	2.78	1.88
$L\hat{M}$										
200	5.20	4.72	4.70	5.00	5.24	5.24	5.46	5.46	5.60	5.42
400	5.40	5.10	5.04	4.80	5.34	4.98	5.30	5.84	5.62	5.14
600	4.78	4.64	4.44	5.18	4.94	5.02	5.22	5.48	5.40	5.12
800	4.82	5.04	5.50	5.40	5.28	4.84	4.38	5.72	5.66	4.48
$\tilde{W}$										
200	24.94	32.26	27.54	26.12	26.34	26.00	25.92	19.88	22.06	22.20
400	20.18	27.78	22.60	21.02	21.04	21.20	20.68	17.38	16.50	19.62
600	17.98	24.62	20.32	19.84	19.52	19.02	17.94	14.96	14.64	16.90
800	16.16	22.20	18.88	18.16	17.66	17.70	16.70	13.42	13.82	15.66
$L\tilde{M}$										
200	37.10	44.10	39.78	39.18	39.44	39.34	37.88	30.94	33.26	34.90
400	29.16	36.58	30.58	29.46	30.74	29.78	29.38	25.06	24.26	27.80
600	23.56	31.82	27.36	26.44	26.52	25.60	24.58	21.06	21.64	23.62
800	21.62	28.30	23.90	23.16	23.22	23.64	21.80	19.20	20.46	21.22

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test computed using OLS estimates of β; $\hat{S}^*$ is the efficient score test computed using 1-step updates from the OLS estimates. $\hat{\mathrm{W}}$, $\hat{\mathrm{LM}}$ denote the Wald and LM tests based on a psuedo-maximum likelihood estimator inspired by Gouriéroux et al. (2017). $\tilde{\mathrm{W}}$ and $\tilde{\mathrm{LM}}$ denote the Wald and LM tests based on a GMM estimator inspired by Lanne and Luoto (2021). Columns 2 – 14 denote the choice of density for ϵ_2, as in Table 13.

Table 16: Empirical rejection frequencies (%) for LSEM, $\epsilon_1 \sim \mathcal{SN}'(0,1,4)$

n	$0-1$	$1-1$	$1-2$	$1-3$	$2-1$	$2-2$	$2-3$	$3-1$	$3-2$	$3-3$
$\hat{S}$										
200	4.90	5.84	5.56	5.48	3.88	4.62	5.16	3.58	4.16	5.08
400	5.22	5.70	5.14	5.00	3.38	4.52	4.92	3.88	4.18	4.54
600	5.18	5.72	4.98	5.52	3.46	4.22	4.78	3.00	4.10	5.08
800	5.10	5.02	5.12	5.22	3.76	3.78	5.08	4.02	3.84	5.02
$\hat{S}^*$										
200	6.02	6.68	6.26	6.40	7.44	6.56	6.74	6.42	6.32	6.18
400	6.34	6.42	6.12	5.84	6.96	6.82	6.46	7.16	6.88	6.32
600	6.34	6.44	5.94	6.40	6.64	6.74	6.26	5.88	6.18	6.16
800	5.58	6.12	6.12	5.86	7.66	5.82	6.42	6.08	5.54	6.40
$\hat{W}$										
200	18.16	2.38	8.40	10.84	14.02	13.36	15.06	47.78	38.68	24.94
400	15.92	1.88	4.82	7.00	10.98	9.78	11.00	60.44	46.74	23.96
600	14.70	1.78	2.76	5.20	8.72	9.00	9.88	66.76	50.96	25.66
800	13.30	2.20	2.68	4.40	6.94	7.44	8.06	73.76	57.02	24.14
$L\hat{M}$										
200	4.84	4.74	5.46	4.36	4.80	5.34	5.46	5.42	5.26	5.16
400	5.44	4.94	5.10	4.26	5.50	5.12	4.26	4.82	5.66	5.42
600	5.02	4.80	5.40	5.30	5.18	4.66	4.88	5.14	5.04	5.02
800	4.98	5.20	4.90	5.58	5.66	4.80	5.70	4.84	5.04	4.90
$\tilde{W}$										
200	27.76	34.48	31.56	29.22	31.88	30.72	30.84	23.16	23.90	26.04
400	24.48	32.28	29.04	28.04	27.94	27.70	27.48	17.80	18.94	23.36
600	20.88	29.54	26.58	24.60	25.72	24.32	23.58	14.38	15.06	18.48
800	20.42	27.94	26.74	23.54	25.42	24.08	23.52	12.50	13.26	16.72
$L\tilde{M}$										
200	35.10	39.54	37.00	36.14	38.18	37.32	38.24	28.98	29.82	33.76
400	27.72	29.62	27.98	28.28	27.94	27.52	27.54	18.90	21.02	25.62
600	21.22	24.22	23.04	21.74	22.74	22.24	22.80	15.42	16.26	19.70
800	20.18	22.34	20.74	18.36	20.48	20.52	21.18	12.18	13.64	17.50

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test computed using OLS estimates of β; $\hat{S}^*$ is the efficient score test computed using 1-step updates from the OLS estimates. $\hat{\mathrm{W}}$, $\mathrm{L\hat{M}}$ denote the Wald and LM tests based on a psuedo-maximum likelihood estimator inspired by Gouriéroux et al. (2017). $\tilde{\mathrm{W}}$ and $\mathrm{L\tilde{M}}$ denote the Wald and LM tests based on a GMM estimator inspired by Lanne and Luoto (2021). Columns 2 – 14 denote the choice of density for ϵ_2, as in Table 13.

Table 17: Empirical rejection frequencies (%) for LSEM

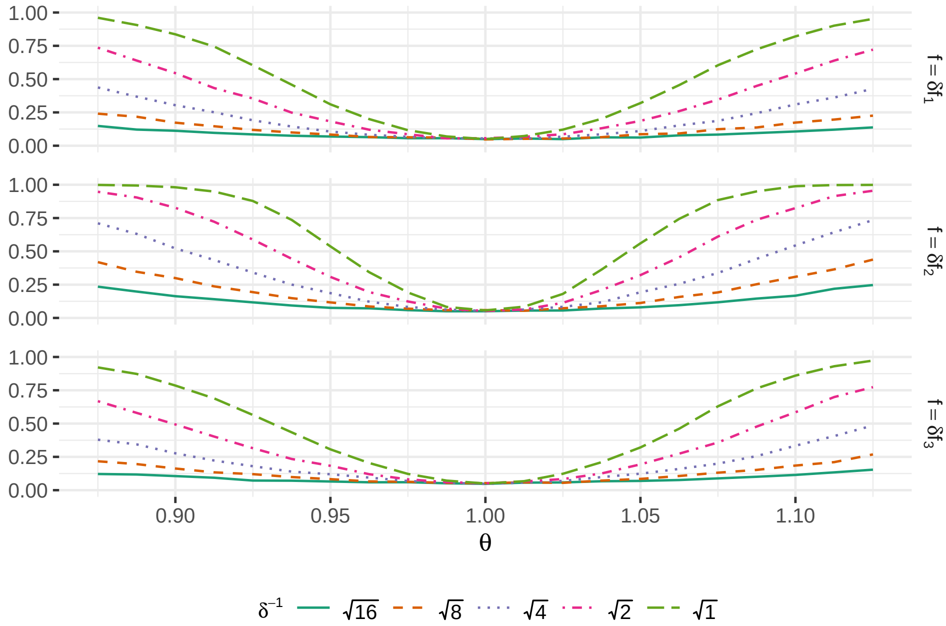
η_1, η_2	n	$\hat{S}$	$\hat{S}^*$	$\dot{S}$	$\dot{S}^*$
1	200	5.20	7.52	7.24	11.84
1	400	4.80	7.24	7.92	12.66
1	600	4.32	6.86	7.58	11.94
1	800	4.32	6.30	7.38	10.76
2	200	7.42	7.68	6.14	9.92
2	400	6.46	6.92	5.48	8.60
2	600	5.56	6.42	5.48	7.98
2	800	5.32	6.24	4.96	7.86
3	200	4.26	7.18	9.10	13.20
3	400	4.06	7.28	8.42	12.68
3	600	3.52	6.90	7.84	12.04
3	800	4.06	7.36	7.56	11.98

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test computed using OLS estimates of β; $\hat{S}^*$ is the efficient score test computed using 1-step updates from the OLS estimates. $\dot{S}$ and $\dot{S}^*$ are score tests based on the score function for the Euclidean parameters using OLS estimates and 1-step updates respectively. The first column denotes the choice of density for both ϵ_1 and ϵ_2 as in the left colum of Table 13.

D.2 Power curves

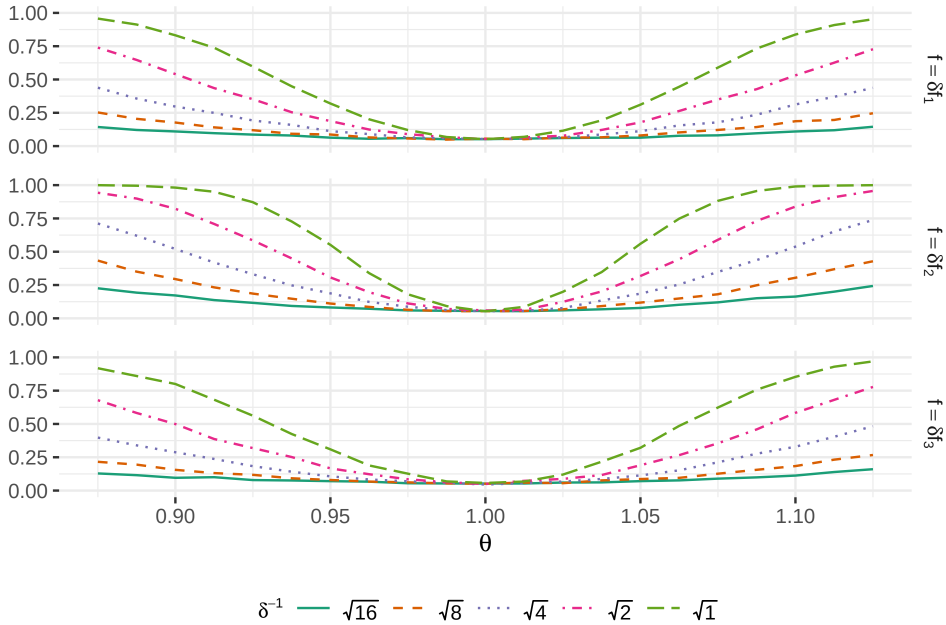
D.2.1 SIM

Figure 5: $\epsilon \sim \mathcal{N}(0,1)$, $X_k \sim U(-1,1)$



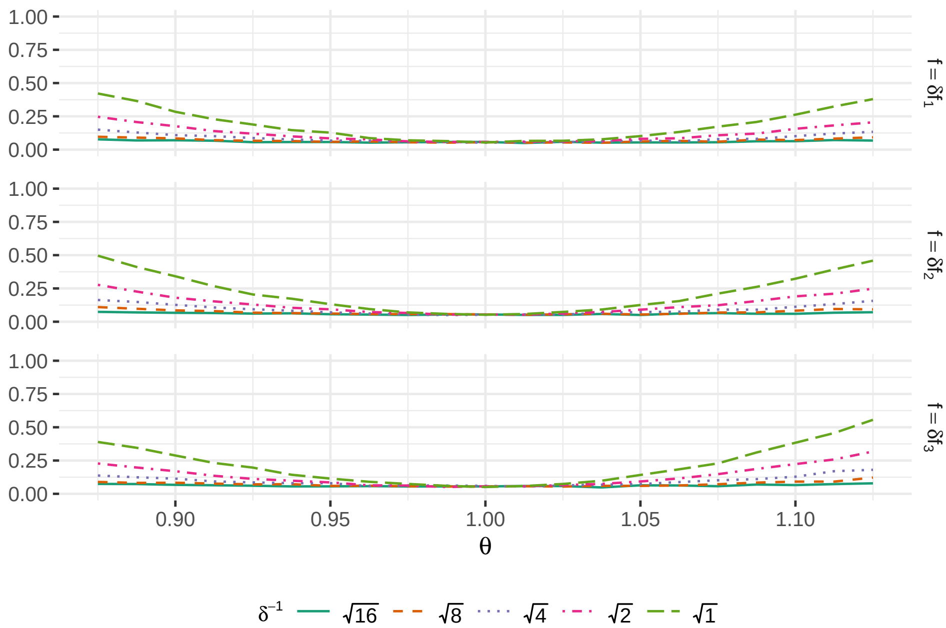
Based on 5000 Monte carlo replications with a sample size of $n = 800$. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null.

Figure 6: $\epsilon|\xi \sim \sqrt{5}(-1)^{\xi}\,\mathrm{Beta}(2,3)$, $\xi \sim \mathrm{Bernoulli}(1/2)$, $X_k \sim U(-1,1)$



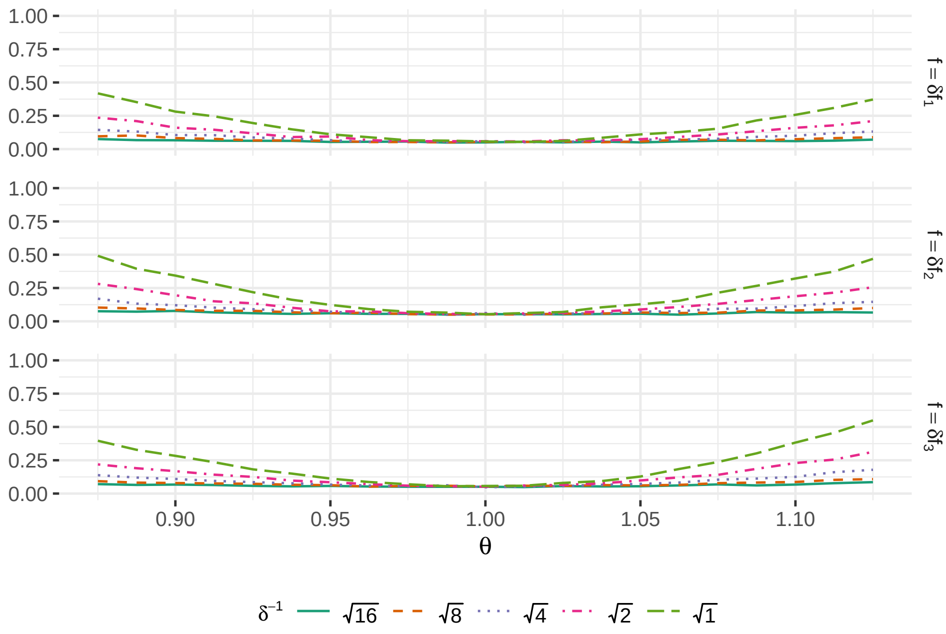
Based on 5000 Monte carlo replications with a sample size of $n = 800$. $f_1(v) = c_1(1+\exp(-v))^{-1}$, $f_2(v) = c_2\exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1,2,3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null.

Figure 7: $\epsilon \sim \mathcal{N}(0,1)$, $X = (Z_1, 0.2Z_1 + 0.4Z_2 + 0.8)$, $Z_k \sim U(-1,1)$



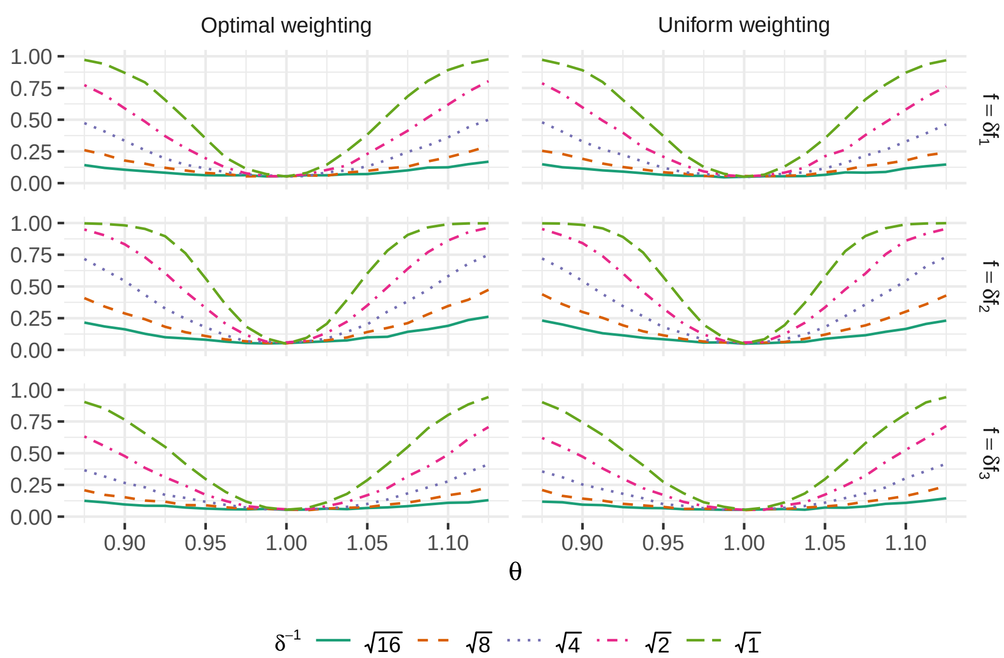
Based on 5000 Monte carlo replications with a sample size of $n = 800$. $f_1(v) = c_1(1+\exp(-v))^{-1}$, $f_2(v) = c_2\exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1,2,3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null.

Figure 8: $\epsilon|\xi \sim \sqrt{5}(-1)^{\xi}\,\mathrm{Beta}(2,3)$, $\xi \sim \mathrm{Bernoulli}(1/2)$ $X = (Z_1, 0.2Z_1 + 0.4Z_2 + 0.8)$, $Z_k \sim U(-1,1)$



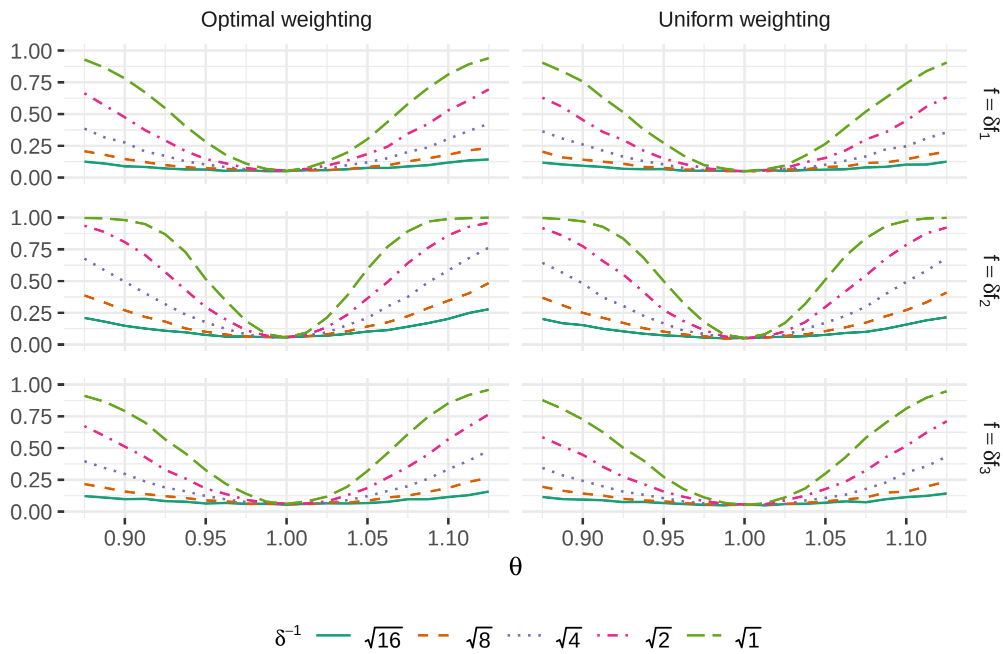
Based on 5000 Monte carlo replications with a sample size of $n = 800$. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null.

Figure 9: $\epsilon \sim \mathcal{N}(0, s_1 \log(2 + (X_1 + X_2\theta)^2))$, $X_k \sim U(-1,1)$



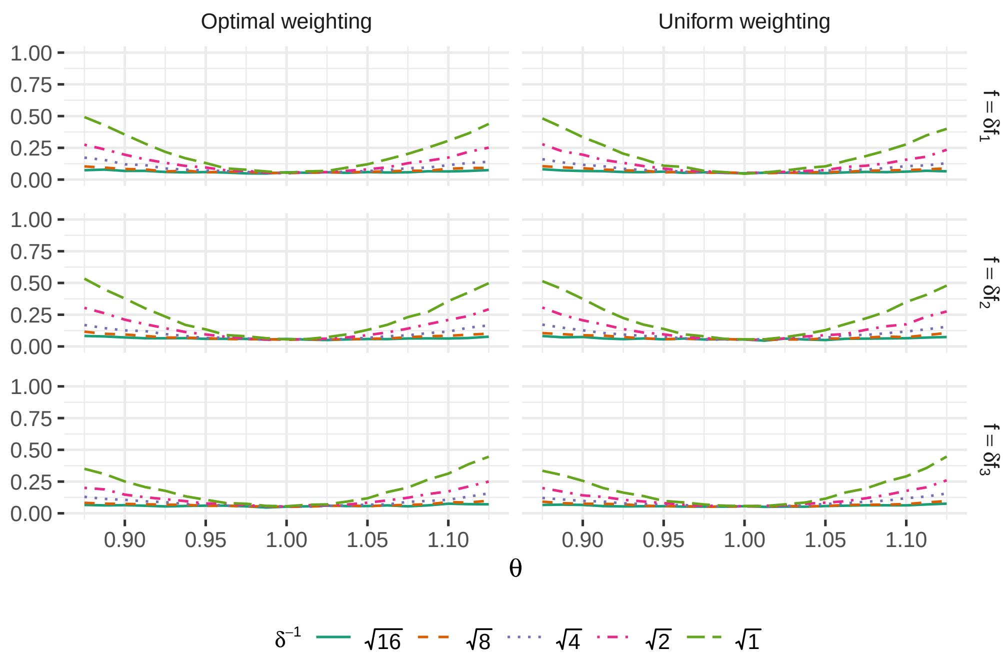
Based on 5000 Monte carlo replications with a sample size of $n = 800$. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1, 2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null. Uniform weighting: $\breve{\omega}(X) = 1$; Optimal weighting: $\breve{\omega}(X) = \omega(X)$.

Figure 10: $\epsilon \sim \mathcal{N}(0, s_2(1+5\sin(X_1)^2))$, $X_k \sim U(-1,1)$



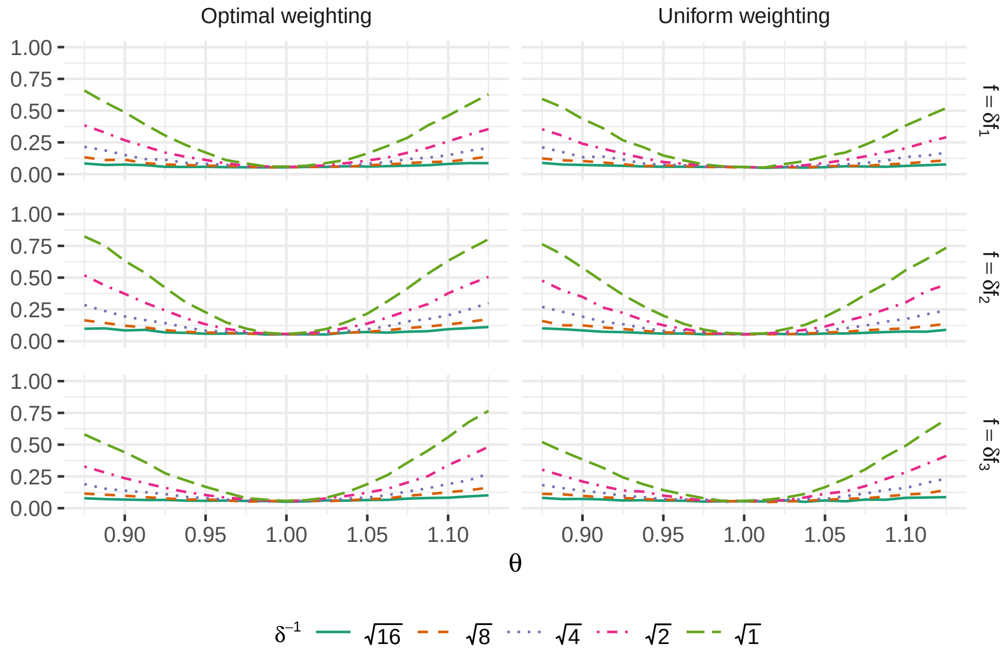
Based on 5000 Monte carlo replications with a sample size of $n = 800$. $f_1(v) = c_1(1+\exp(-v))^{-1}$, $f_2(v) = c_2\exp(-v^2)$, $f_3(v) = c_3v^2$, where the constants c_i $(i = 1,2,3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1,2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null. Uniform weighting: $\breve{\omega}(X) = 1$; Optimal weighting: $\breve{\omega}(X) = \omega(X)$.

Figure 11: $\epsilon \sim \mathcal{N}(0, s_1\log(2+(X_1+X_2\theta)^2))$, $X = (Z_1, 0.2Z_1+0.4Z_2+0.8)$, $Z_k \sim U(-1,1)$



Based on 5000 Monte carlo replications with a sample size of $n = 800$. $f_1(v) = c_1(1+\exp(-v))^{-1}$, $f_2(v) = c_2\exp(-v^2)$, $f_3(v) = c_3v^2$, where the constants c_i $(i = 1,2,3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1,2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null. Uniform weighting: $\breve{\omega}(X) = 1$; Optimal weighting: $\breve{\omega}(X) = \omega(X)$.

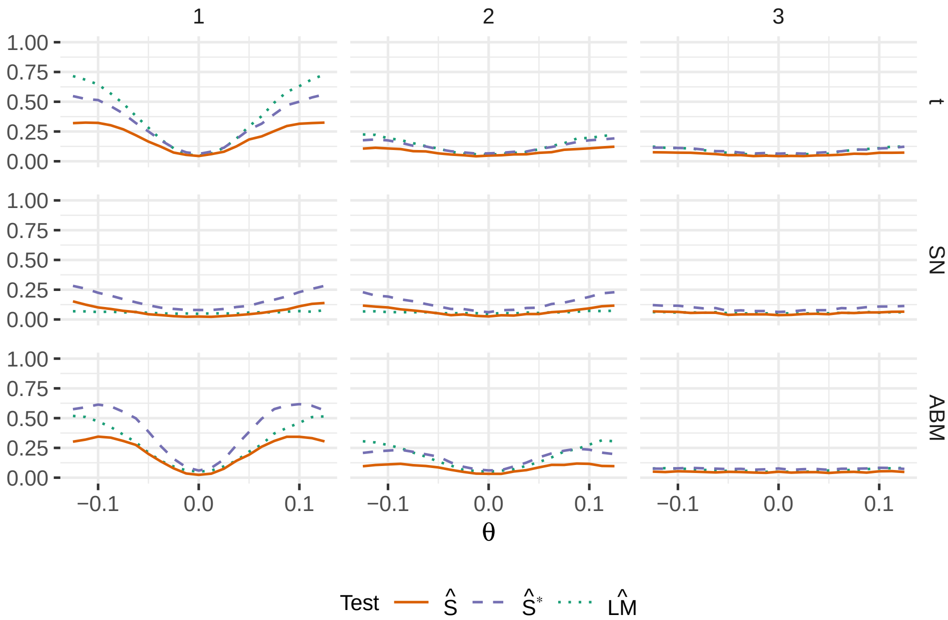
Figure 12: $\epsilon \sim \mathcal{N}(0, s_2(1 + 5\sin(X_1)^2))$, $X = (Z_1, 0.2Z_1 + 0.4Z_2 + 0.8)$, $Z_k \sim U(-1, 1)$



Based on 5000 Monte carlo replications with a sample size of $n = 800$. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1, 2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null. Uniform weighting: $\breve{\omega}(X) = 1$; Optimal weighting: $\breve{\omega}(X) = \omega(X)$.

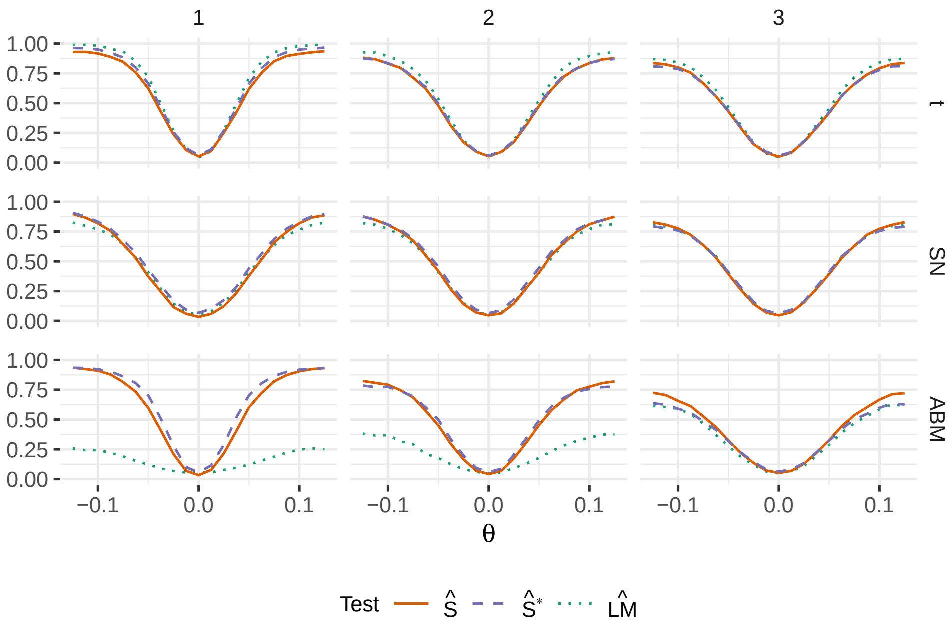
D.2.2 LSEM

Figure 13: Power curves for LSEM (i), $\epsilon_1 \sim \mathcal{N}(0,1)$



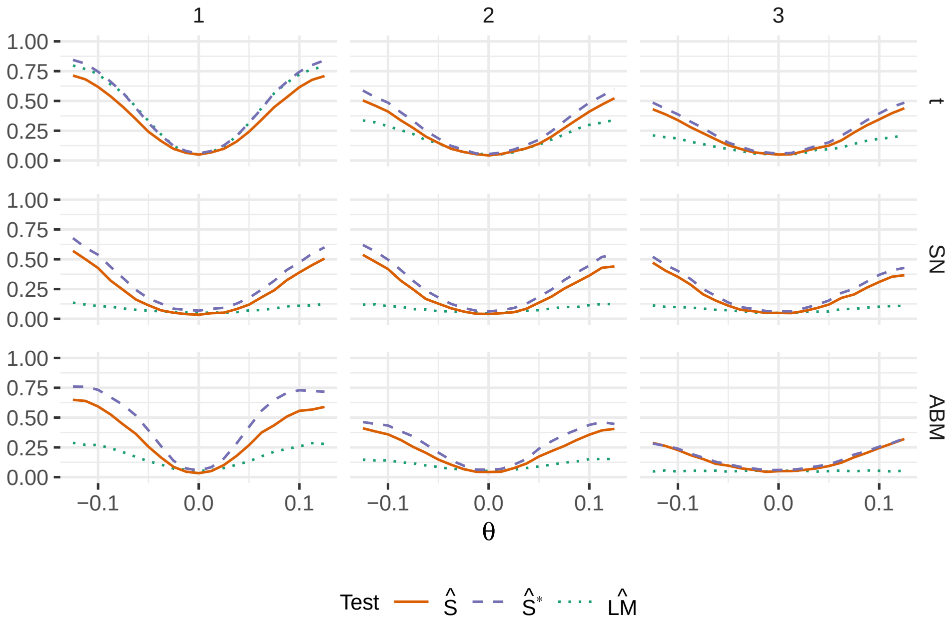
Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test computed using OLS estimates of β; $\hat{S}^*$ is the efficient score test computed using 1-step updates from the OLS estimates. $\hat{\text{LM}}$ denotes the LM test based on a psuedo-maximum likelihood estimator inspired by Gouriéroux et al. (2017). The distribution for ϵ_2 in the $(i,j) - th$ panel has distribution $i - j$ in table 13.

Figure 14: Power curves for LSEM (i), $\epsilon_1 \sim t'(5)$



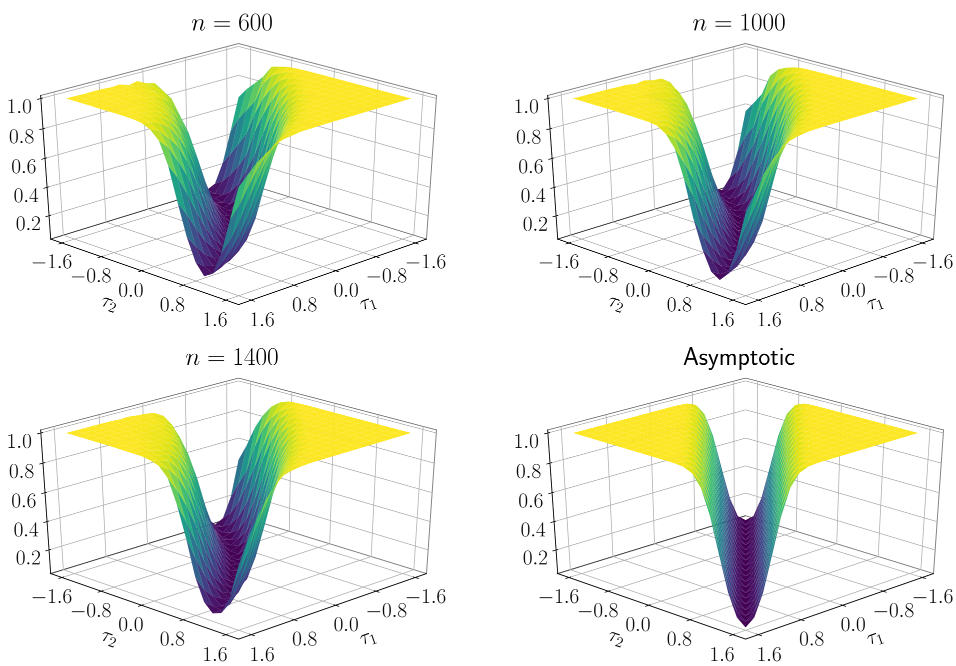
Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test computed using OLS estimates of β; $\hat{S}^*$ is the efficient score test computed using 1-step updates from the OLS estimates. $\hat{\text{LM}}$ denotes the LM test based on a psuedo-maximum likelihood estimator inspired by Gouriéroux et al. (2017). The distribution for ϵ_2 in the $(i,j) - th$ panel has distribution $i - j$ in table 13.

Figure 15: Power curves for LSEM (i), $\epsilon_1 \sim \mathcal{SN}'(0, 1, 4)$



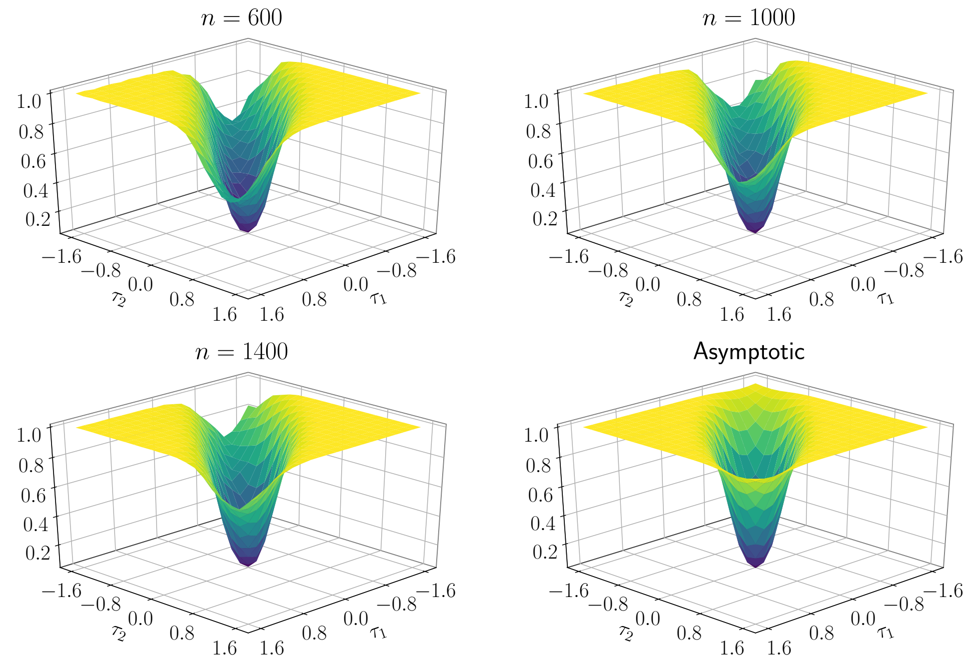
Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test computed using OLS estimates of β; $\hat{S}^*$ is the efficient score test computed using 1-step updates from the OLS estimates. $\hat{\text{LM}}$ denotes the LM test based on a psuedo-maximum likelihood estimator inspired by Gouriéroux et al. (2017). The distribution for ϵ_2 in the $(i, j) - th$ panel has distribution $i - j$ in table 13.

Figure 16: Power surfaces for LSEM (ii), $\eta_1 \sim \mathcal{N}(0, 1)$, $\eta_2 \sim \mathcal{N}(0, 1)$



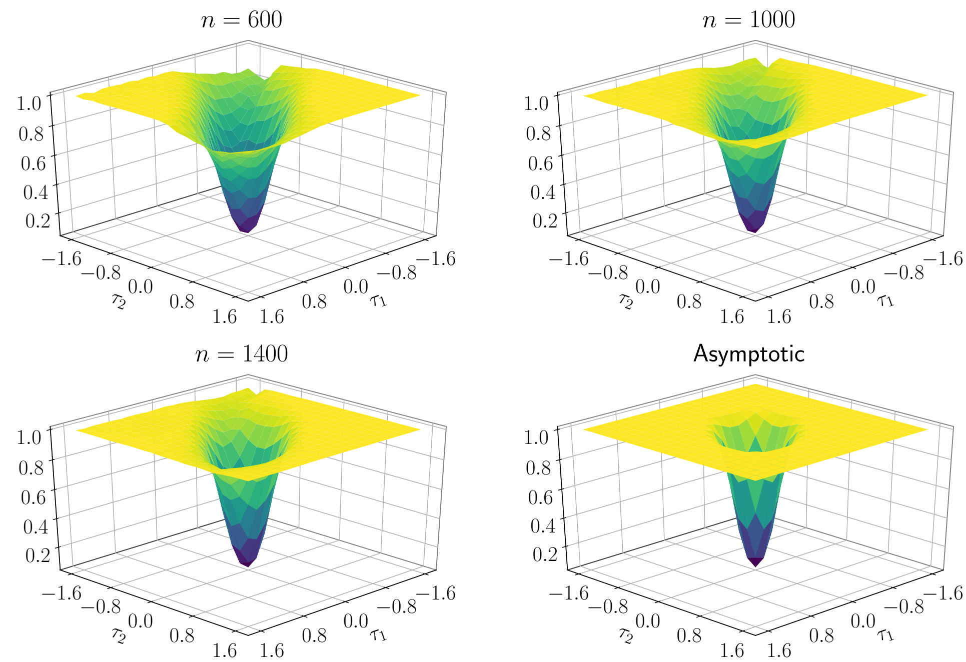
The bottom right panel depicts the asymptotic power surface based on (18) and (51) with $\theta = (a, b) = (1/2, 1/4)$ and $\sigma_1 = \sigma_2 = 1$. The top-left, top-right and bottom-left panels are Monte Carlo version based on 5000 replications of the efficient score test as described in section 5.5, with $n = 600, 1000, 1400$ respectively.

Figure 17: Power surfaces for LSEM (ii), $\eta_1 \sim t'(5)$, $\eta_2 \sim t'(5)$



The bottom right panel depicts the asymptotic power surface based on (18) and (51) with $\theta = (a, b) = (1/2, 1/4)$ and $\sigma_1 = \sigma_2 = 1$. The top-left, top-right and bottom-left panels are Monte Carlo version based on 5000 replications of the efficient score test as described in section 5.5, with $n = 600, 1000, 1400$ respectively. $\eta_k \sim t'(5)$ indicates that each ϵ_k is drawn from a (standardised) t distribution with 5 degrees of freedom.

Figure 18: Power surfaces for LSEM (ii), $\eta_1 \sim st'(5, 2)$, $\eta_2 \sim st'(5, 2)$



The bottom right panel depicts the asymptotic power surface based on (18) and (51) with $\theta = (a, b) = (1/2, 1/4)$ and $\sigma_1 = \sigma_2 = 1$. The top-left, top-right and bottom-left panels are Monte Carlo version based on 5000 replications of the efficient score test as described in section 5.5, with $n = 600, 1000, 1400$ respectively. $\eta_k \sim st'(5, 2)$ indicates that each ϵ_k is drawn from a (standardised) skew t distribution, as in Fernandez and Steel (1998) with 5 degrees of freedom and skewness parameter 2.

D.3 Empirical study

Figure 19: Confidence intervals for θ

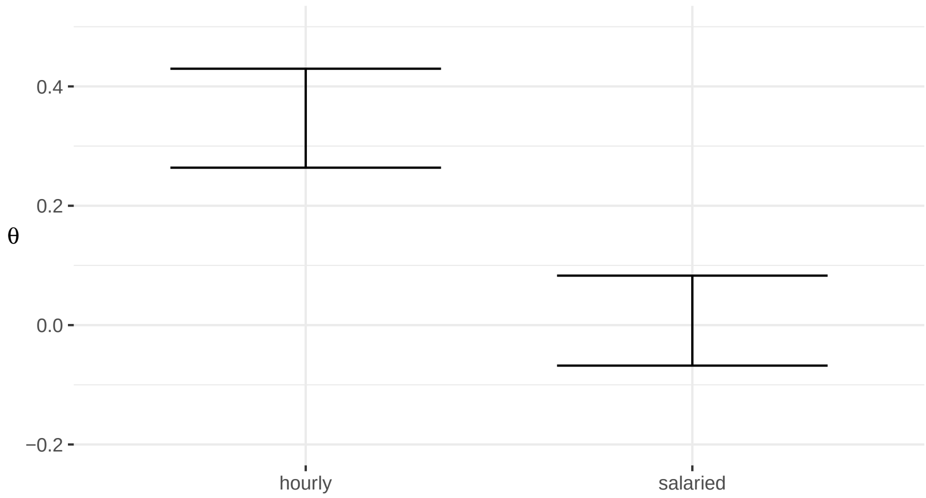
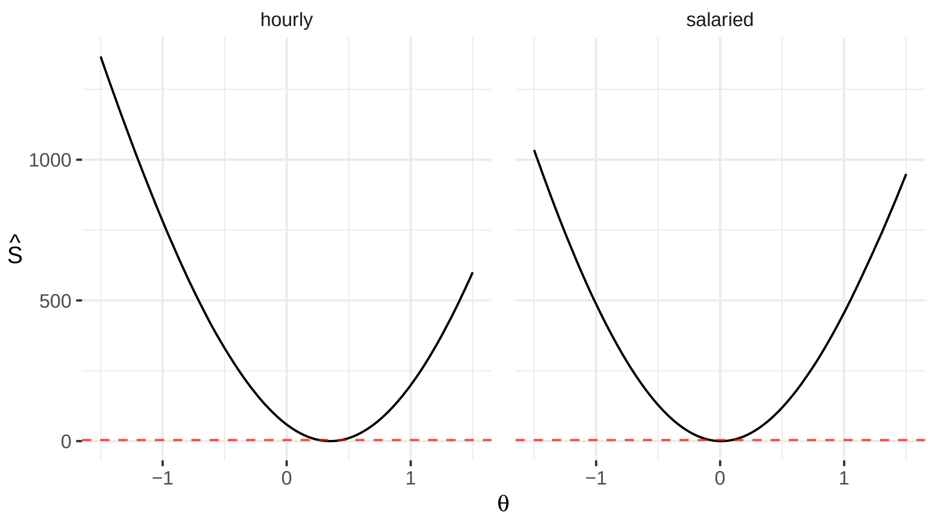
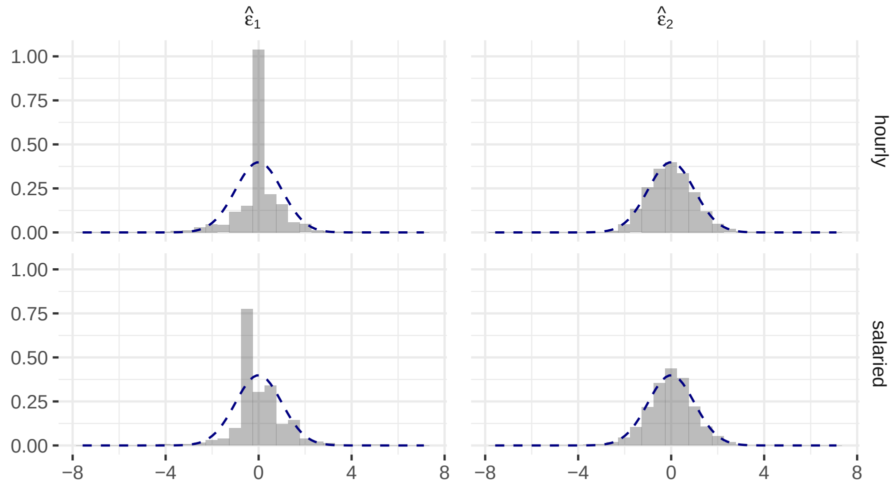


Figure 20: Value of efficient score statistic over θ grid



The red, dashed line is the 95th quantile of the χ^2_1 distribution; values below this are included in the confidence set.

Figure 21: Residuals from LSEM model



Residuals from (54) with θ taken as the value which minimises the efficient score statistic. The dashed blue line is the $\mathcal{N}(0, 1)$ density function.