Conditional Inference with a Functional Nuisance Parameter

By Isaiah Andrews[1] and Anna Mikusheva [2]

Abstract

This paper shows that the problem of testing hypotheses in moment condition models without any assumptions about identification may be considered as a problem of testing with an infinite-dimensional nuisance parameter. We introduce a sufficient statistic for this nuisance parameter in a Gaussian problem and propose conditional tests. These conditional tests have uniformly correct asymptotic size for a large class of models and test statistics. We apply our approach to construct tests based on quasi-likelihood ratio statistics, which we show are efficient in strongly identified models and perform well relative to existing alternatives in two examples.

Key words: weak identification, similar test, conditional inference

First draft: September 2014. This draft: February, 2016.

1 Introduction

Many econometric techniques identify and draw inferences about a structural parameter θ based on a set of moment equalities. In particular, many models imply that some function of the data and model parameters has mean zero when evaluated at the true parameter value θ_0. The current econometric literature devotes a great deal of energy to investigating whether a given set of moment restrictions suffices to uniquely identify the parameter θ, and to studying inference under different identification assumptions. The goal of this paper is to develop a wide variety of tests for the hypothesis that a specific value θ_0 is consistent with the data without making any assumptions about the point identification or strength of identification of the model.

We treat moment equality models as having a functional nuisance parameter. Much work in econometrics focuses on θ as the unknown model parameter, typically belonging to a finite-dimensional parameter space. This is consistent with the tradition from classical statistics, which studied fully-parametric models where the unknown parameter θ

[1]Harvard Society of Fellows. Harvard Department of Economics, Littauer Center 124, Cambridge, MA 02138. Email iandrews@fas.harvard.edu. NSF Graduate Research Fellowship support under grant number 1122374 is gratefully acknowledged.

[2]Department of Economics, M.I.T., 50 Memorial Drive, E52-526, Cambridge, MA, 02142. Email: amikushe@mit.edu. Financial support from the Castle-Krob Career Development Chair and the Sloan Research Fellowship is gratefully acknowledged. We thank Alex Belloni, Victor Chernozhukov, Kirill Evdokimov, Martin Spindler, and numerous seminar participants for helpful discussions. We are grateful to the Editor and three anonymous referees for valuable and constructive comments.

fully described the distribution of the data. By contrast, in moment condition models the joint distribution of the data is typically only partially specified, and in particular the mean of the moment condition at values θ other than θ_0 is typically unknown. In light of this fact we suggest a re-consideration of the parameter space in these semi-parametric models and view the mean function as an unknown (and often infinite-dimensional) parameter. The structural parameter θ_0 corresponds to a zero of this unknown function, and any hypothesis about θ_0 can be viewed as a composite hypothesis with an infinite-dimensional nuisance parameter, specifically the value of the mean function for all other values θ. The mean function determines the identification status of the structural parameter θ. Treating the mean function as a parameter thus allows us to avoid making restrictive assumptions about identification. Corresponding to this infinite-dimensional parameter, we base inference on observation of an infinite-dimensional object, namely the stochastic process given by the sample moment function evaluated at different parameter values θ.

This perspective allows us to study the behavior of a wide variety of tests for the hypothesis that the mean function is equal to zero at θ_0. In a point-identified setting this hypothesis corresponds to testing that θ_0 is the true parameter value, while when point identification fails it corresponds to testing that θ_0 belongs to the identified set. The existing literature proposes a number of tests for this hypothesis but most of these procedures depend on the observed process only through its value, and potentially derivative, at the point θ_0. Examples include the Anderson-Rubin statistic, Kleibergen (2005)'s K statistic, and generalizations and combinations of these. A major reason for restricting attention to tests that depend only on behavior local to θ_0 is that the distributions of these test statistics are independent of the unknown mean function, or depend on it only through a finite-dimensional parameter. Unfortunately, however, restricting attention to the behavior of the process local to θ_0 ignores a great deal of information and so may come at a significant cost in terms of power. Further, this restriction rules out many tests known to have desirable power properties in other settings. In contrast to the previous literature, our approach allows us to consider tests which depend on the full path of the observed process.

In this paper we introduce a large class of tests that may depend on the data in more complicated ways than has been allowed in the previous literature. To ensure size control we introduce a sufficient statistic for the unknown mean function in a Gaussian

problem and condition inference on the realization of this sufficient statistic. The idea of conditioning on a sufficient statistic for a nuisance parameter is a long-standing tradition in statistics and was popularized in econometrics by Moreira (2003), which applied this idea in weakly-identified linear instrumental variables (IV) models. The contribution of our paper is to show how this technique may be applied in contexts with an infinite-dimensional nuisance parameter, allowing its use in a wide range of econometric models. Since the nuisance parameter in our context is a function, our sufficient statistic is a stochastic process. Our proposed approach to testing is computationally feasible and is of similar difficulty to other simulation-based techniques such as the bootstrap.

One test allowed by our approach is the conditional quasi-likelihood ratio (QLR) test. This test makes use of the full path of the observed stochastic process, and its distribution under the null depends on the unknown mean function, which greatly limited its use in the previous literature on inference with nonstandard identification. We make no claim about optimality of the conditional QLR test in general. QLR tests are, however, known to have good properties in some special cases: in well identified (point identified and strongly identified) models QLR tests are asymptotically efficient, while they avoid the power deficiencies of Kleibergen (2005)'s K and related tests under weak identification. Moreover, in linear IV with a single endogenous regressor and homoscedastic errors D. Andrews, Moreira, and Stock (2006) showed that Moreira (2003)'s conditional likelihood ratio (CLR) test, which corresponds to the conditional QLR test in that context, is nearly uniformly most powerful in an important class of tests.

Conditioning on a sufficient statistic for a nuisance parameter, while widely applied, may incur a loss of power by restricting the class of tests permitted. We show, however, that no power loss occurs in well identified models, as in this case our conditional QLR test is asymptotically equivalent to the unconditional QLR test and thus is efficient. We also point out that if one is interested in similar tests (that is, tests with exactly correct size regardless of the mean function) and the set of mean functions is sufficiently rich, all similar tests must be conditional tests of the form we consider.

To justify our approach we show that the conditional tests we propose have uniformly correct asymptotic size over a broad class of models which imposes no restriction on the mean function and so includes a wide range of identification settings. We further extend these results to allow for concentrating out well-identified structural nuisance parameters.

We apply our approach to inference on the coefficients on the endogenous regressors

in the quantile IV model studied by Chernozhukov and Hansen (2005, 2006, 2008) and Jun (2008). We examine the performance of the conditional QLR test in this context and find that it has desirable power properties relative to alternative approaches. In particular, unlike Anderson-Rubin-type tests the conditional QLR test is efficient under strong identification, while unlike tests based on the K statistic it does not suffer from non-monotonic power under weak identification. We find particularly large power gains for the QLR test relative to existing alternatives in cases when the mean function is highly nonlinear. In the Supplementary Appendix we also examine the performance of the conditional QLR test in linear IV with non-homoscedastic errors, find that it outperforms the K and GMM-M tests of Kleibergen (2005), and discuss additional results showing that it is competitive with other tests recently proposed for linear IV motivated by optimality considerations.

As an empirical application of our method we compute confidence sets for nonlinear Euler Equation parameters based on US data. We find that our approach yields much smaller confidence sets than existing alternatives, and in particular allows us to rule out high values of risk aversion allowed by alternative methods. There is, however, evidence of substantial misspecification in this context which may affect the relative performance of different procedures.

In Section 2 we introduce our model and discuss the benefits of formulating the problem using an infinite-dimensional nuisance parameter. Section 3 explains and justifies our conditioning approach and relates our results to previous work. We also discuss power benefits from using information away from the tested structural parameter value. Section 4 establishes the uniform asymptotic validity of our method and proves the asymptotic efficiency of the conditional QLR test in strongly identified settings, while Section 5 discusses the possibility of concentrating out well-identified nuisance parameters. Section 6 reports simulation results for the conditional QLR test in a quantile IV model and gives confidence sets for nonlinear Euler equation parameters based on US data, and Section 7 concludes. Some proofs and additional results may be found in a Supplementary Appendix available on the authors' web-sites.

In the remainder of the paper we denote by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ the minimal and maximal eigenvalues of a square matrix A, respectively, while $\|A\|$ is both the operator norm for a matrix and the Euclidean norm for a vector.

2 Models with functional nuisance parameters

Many testing problems in econometrics can be recast as tests that a vector-valued random function of model parameters has mean zero at a particular point. Following Hansen (1982) suppose we have an economic model which implies that some $k \times 1$-dimensional function $\varphi(X_t;\theta)$ of the data and the $q \times 1$-dimensional parameter θ has mean zero when evaluated at the true parameter value θ_0, $E[\varphi(X_t,\theta_0)] = 0$. Define $g_T(\cdot) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\varphi(X_t,\cdot)$ and let $m_T(\cdot) = E[g_T(X_t,\cdot)]$. Under mild conditions (see e.g. Van der Vaart and Wellner (1996) for i.i.d. data and Dedecker and Louhici (2002) for time series), empirical process theory implies that

$$g_T(\theta) =^d m_T(\theta) + G(\theta) + r_T(\theta), \tag{1}$$

where $G(\cdot)$ is a mean-zero Gaussian process with consistently estimable covariance function $\Sigma(\theta,\tilde{\theta}) = EG(\theta)G(\tilde{\theta})'$, and r_T is a residual term which is negligible for large T in the sense that the process $g_T(\cdot) - m_T(\cdot)$ weakly converges to $G(\cdot)$ as $T \to \infty$. We are interested in testing that θ_0 belongs to the identified set, which is equivalent to testing $H_0 : m_T(\theta_0) = 0$, without any assumption on identification of the parameter θ.

This paper considers equation (1) as a model with an infinite-dimensional nuisance parameter, namely $m_T(\theta)$ for $\theta \neq \theta_0$. Thus our perspective differs from the more classical approach which focuses on θ as the model parameter. This classical approach may be partially derived from the use of parametric models in which θ fully specifies the distribution of the data. By contrast many of the methods used in modern econometrics, including the generalized method of moments (GMM), only partially specify the distribution of the data, and the behavior of $m_T(\theta)$ for θ outside of the identified set is typically neither known nor consistently estimable (due to the $\sqrt{T}$ term embedded in the definition). To formally describe the parameter space for m_T, let $\mathcal{M}$ be the set of functions $m_T(\cdot)$ that may arise in a given model.[3] Let $\mathcal{M}_0$ be the subset of $\mathcal{M}$ containing those functions satisfying $m_T(\theta_0) = 0$. The hypothesis of interest may be formulated as $H_0 : m_T \in \mathcal{M}_0$, which is in general a composite hypothesis with a non-parametric nuisance parameter.

The distribution of most test statistics under the null depends crucially on the nuisance function $m_T(\cdot)$. For example the distribution of the quasi-likelihood ratio (QLR)

[3]Note that the set of functions $\mathcal{M} = \mathcal{M}_T$ may change with the sample size, but we drop the subscript for notational simplicity.

statistic, which for $\widehat{\Sigma}$ an estimator of Σ takes the form

$$QLR = g_T(\theta_0)'\widehat{\Sigma}(\theta_0,\theta_0)^{-1}g_T(\theta_0) - \inf_{\theta} g_T(\theta)'\widehat{\Sigma}(\theta,\theta)^{-1}g_T(\theta), \tag{2}$$

depends in complex ways on the true unknown function $m_T(\cdot)$ except in special cases like the strong identification assumptions introduced in Section 4.2. The same is true of Wald- or t-statistics, or of statistics analogous to QLR constructed using weightings other than $\widehat{\Sigma}(\theta,\theta)^{-1}$, which we call QLR-type statistics. In the literature to date this dependence on m_T has greatly constrained the use of these statistics in non-standard settings, since outside of special cases (for example linear IV, or the models studied by D. Andrews and Cheng (2012)) there has been no way to calculate valid critical values.

Despite these challenges there are a number of tests in the literature that control size for all values of the infinite-dimensional nuisance parameter $m_T(\cdot)$. One well-known example is the S test of Stock and Wright (2000), which is based on the statistic $S = g_T(\theta_0)'\widehat{\Sigma}(\theta_0,\theta_0)^{-1}g_T(\theta_0)$. This is a generalization of the Anderson-Rubin statistic and is asymptotically χ^2_k distributed for all $m_T \in \mathcal{M}_0$. Other examples include Kleibergen (2005)'s K test and its generalizations. Unfortunately, these tests often have deficient power in over-identified settings or when identification is weak, respectively. Several authors have also suggested statistics intended to mimic the behavior of QLR in particular settings, for example the GMM-M statistic of Kleibergen (2005), but the behavior of these statistics differs greatly from that of true QLR statistics in many contexts of interest.

Example 1. Consider the nonlinear Euler equations studied by Hansen and Singleton (1982). The moment function identifying the discount factor δ and the coefficient of relative risk-aversion γ is

$$g_T(\theta) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(\delta\left(\frac{C_t}{C_{t-1}}\right)^{-\gamma}R_t - 1\right)Z_t, \qquad \theta = (\delta,\gamma),$$

where C_t is consumption in period t, R_t is an asset return from period $t-1$ to t, and Z_t is a vector of instruments measurable with respect to information at $t-1$. Under moment and mixing conditions (see for example Theorem 5.2 in Dedecker and Louhici (2002)), the demeaned process $g_T(\cdot) - Eg_T(\cdot)$ will converge to a Gaussian process.

For true parameter value $\theta_0 = (\delta_0,\gamma_0)$ we have $m_T(\theta_0) = Eg_T(\theta_0) = 0$. The value of

$m_T(\theta) = Eg_T(\theta)$ for $\theta \neq \theta_0$ is in general unknown and depends in a complicated way on the joint distribution of the data, which is typically neither known nor explicitly modeled. Further, $m_T(\theta)$ cannot be consistently estimated. Consequently the distribution of QLR and many other statistics which depend on $m_T(\cdot)$ is unavailable unless one is willing to assume the model is well-identified, which is contrary to extensive evidence suggesting identification problems in this context. □

2.1 The mean function m_T in examples

Different econometric settings give rise to different mean functions $m_T(\cdot)$, which in turn determine the identification status of θ. In set-identified models the identified set $\{\theta : m_T(\theta) = 0\}$ might be a collection of isolated points or sets, or even the whole parameter space. In well-identified settings, by contrast, $m_T(\cdot)$ has a unique zero and increases rapidly as one moves away from this point, especially as T becomes large. Common models of weak identification imply that even for T large $m_T(\cdot)$ remains bounded over some non-trivial region of the parameter space.

Consider for example the classical situation (as in Hansen (1982)) where the function $E\varphi(X_t, \cdot)$ is fixed and continuously differentiable with a unique zero at θ_0 and the Jacobian $\frac{\partial E\varphi(X_t,\theta_0)}{\partial\theta}$ has full rank. This is often called a *strongly identified case*, and (under regularity conditions) will imply the strong identification assumptions we introduce in Section 4.2. In this setting the function $m_T(\theta) = \sqrt{T}E\varphi(X_t, \theta)$ diverges to infinity outside of $1/\sqrt{T}$ neighborhoods of θ_0 as the sample size grows. Many statistics, like Wald or QLR-type statistics, use $g_T(\cdot)$ evaluated only at some estimated value $\widehat{\theta}$ and θ_0, and thus in the classical case depend on g_T only through its behavior on a $1/\sqrt{T}$ neighborhood of the true θ_0. Over such neighborhoods $m_T(\cdot)$ is well approximated by $\frac{\partial E\varphi(X_t,\theta_0)}{\partial\theta} \cdot \sqrt{T}(\theta - \theta_0)$, the only unknown component of which, $\frac{\partial E\varphi(X_t,\theta_0)}{\partial\theta}$, is usually consistently estimable. Reasoning along these lines, which we explore in greater detail in Section 4.2, establishes the asymptotic validity of classical tests under strong identification. Thus in strongly identified models the nuisance parameter problem we study here does not arise.

In contrast to the strongly-identified case, *weakly identified* models are often understood as those in which even for T large the mean function m_T fails to dominate the Gaussian process G over a substantial portion of the parameter space. Stock and Wright

(2000) modeled this phenomenon using a drifting sequence of functions. In particular, a simple case of the Stock and Wright (2000) embedding indexes the data-generating process by the sample size and assumes that while the variance of the moment condition is asymptotically constant, the expectation of the moment condition shrinks at rate $1/\sqrt{T}$, so $E\varphi(X_t,\theta) = E_T\varphi(X_t,\theta) = \frac{1}{\sqrt{T}}f(\theta)$ for a fixed function $f(\theta)$. In this case $m_T(\theta) = f(\theta)$ is unknown and cannot be consistently estimated, consistent estimation of θ_0 is likewise impossible, and the whole function $m_T(\cdot)$ is important for the distribution of QLR-type statistics.

By treating m_T as a nuisance parameter, our approach avoids making any assumptions about its behavior. Thus, we can treat both the strongly-identified case described above and the weakly-identified sequences studied by Stock and Wright (2000), as well as set identified models and a wide array of other cases. As we illustrate below this is potentially quite important, as the set $\mathcal{M}$ of mean functions can be extremely rich in examples.

We next discuss the sets $\mathcal{M}$ in several examples. As a starting point we consider the linear IV model where the nuisance function can be reduced to a finite-dimensional vector of nuisance parameters and then consider examples with genuine functional nuisance parameters.

Example 2. (Linear IV) Consider a linear IV model where the data consists of i.i.d. observations on an outcome variable Y_t, an endogenous regressor D_t, and a vector of instruments Z_t. Assume that the identifying moment condition is $E\left[\left(Y_t - D_t'\theta_0\right)Z_t\right] = 0$. This implies that $m_T\left(\theta\right) = \sqrt{T}E\left[Z_tD_t'\right]\left(\theta_0 - \theta\right)$ is a linear function. If $E\left[Z_tD_t'\right]$ is a fixed matrix of full column rank, then θ_0 is point identified and can be consistently estimated using two stage least squares, while if $E\left[Z_tD_t'\right]$ is of reduced rank the identified set is a hyperplane of dimension equal to the rank deficiency of $E\left[Z_tD_t'\right]$. Staiger and Stock (1997) modeled weak instruments by considering a sequence of data-generating processes such that $E\left[Z_tD_t'\right] = \frac{C}{\sqrt{T}}$ for a constant unknown matrix C. Under these sequences the function $m_T(\theta) = C(\theta_0 - \theta)$ is linear and governed by the unknown (and not consistently estimable) parameter C. □

In contrast to the finite-dimensional nuisance parameter obtained in linear IV, in nonlinear models the space of nuisance parameters $m_T(\cdot)$ is typically of infinite dimension.

Example 1 (continued). In the Euler equation example discussed above,

$$m_T(\theta) = \sqrt{T}E\left[\left(\delta(1+R_t)\left(\frac{C_t}{C_{t-1}}\right)^{-\gamma} - 1\right)Z_t\right].$$

Assume for a moment that δ is fixed and known and that R_t and Z_t are constant. In this simplified case the function $m_T(\gamma)$ is a linear transformation of the moment-generating function of $\log(C_t/C_{t-1})$, implying that the set $\mathcal{M}_0$ of mean functions is at least as rich as the set of possible distributions for consumption growth consistent with the null. □

Example 3. (Quantile IV) Suppose we observe i.i.d. data consisting of an outcome variable Y_t, an almost-surely positive endogenous regressor D_t, and instruments Z_t. For U_t a zero-median shock independent of Z_t, suppose Y_t follows $Y_t = \gamma D_t + (D_t + 1)U_t$. These variables obey the quantile IV model of Chernozhukov and Hansen (2005) for all quantiles, and satisfy

$$E\left[\left(\mathbb{I}\{Y_t - \theta_0 D_t \leq 0\} - 1/2\right)Z_t\right] = 0$$

for $\theta_0 = \gamma$, so we can use this moment condition for inference. This moment restriction holds for arbitrary joint distributions of (D_t, Z_t, U_t) provided U_t and Z_t are independent and U_t has median zero. However, different distributions produce different mean functions. In the Supplementary Appendix we consider a weakly identified example with $Z_t = \frac{1}{\sqrt{T}}F(Z_t^*) + (1 - \frac{1}{\sqrt{T}})\eta_t$ and $D_t = \exp\{Z_t^* - U_t\}$, where Z_t^*, U_t, η_t are mutually independent and $E[F(Z_t^*)] = E[\eta_t] = 0$. In this setting we show that

$$m_T(\theta_0 + \delta) = \sqrt{T}E\left[\left(\mathbb{I}\{Y_t - (\theta_0 + \delta)D_t \leq 0\} - 1/2\right)Z_t\right] = E\left[F(Z_t^*)F_U(x(\delta, e^{-Z_t^*}))\right],$$

where $F_U(\cdot)$ is the cdf of U and $x(y, b)$ solves $(1 + be^x)x = y$. Depending on F and the marginal distributions of U_t and Z_t^* one can end up with a wide variety of mean functions in this setting, many of which are highly non-linear. □

Our results also apply outside the GMM context so long as one has a model described by equation (1). In Section 5.1, for example, we apply our results to a quantile IV model where we plug in estimates for nuisance parameters. Our results can likewise be applied to the simulation-based moment conditions considered in McFadden (1989), Pakes and Pollard (1989), and the subsequent literature. More recently Schennach (2014) has shown that models with latent variables can be expressed using simulation-based mo-

ment conditions, allowing for the treatment of an enormous array of additional examples including game-theoretic, moment-inequality, and measurement-error models within the framework studied in this paper.

3 Conditional approach

To construct tests we introduce a sufficient statistic for $m_T(\cdot) \in \mathcal{M}_0$ in a Gaussian problem and suggest conditioning inference on this statistic, thereby eliminating dependence on the nuisance parameter. Moreira (2003) showed that a conditioning approach could be fruitfully applied to inference in linear instrumental variables models, while Kleibergen (2005) extended this approach to GMM statistics which depend only on $g_T(\cdot)$ and its derivative both evaluated at θ_0. In this section we show that conditional tests can be applied far more broadly. We first introduce our approach and describe how to calculate critical values, then justify our procedure in an exact Gaussian problem and discuss power. In Section 4 we show that our tests are uniformly asymptotically correct under more general assumptions.

3.1 Conditional inference

Consider model (1) and let $\widehat{\Sigma}(\cdot,\cdot)$ be a consistent estimator of the covariance function $\Sigma(\cdot,\cdot)$. Let us introduce the process

$$h_T(\theta) = H(g_T, \widehat{\Sigma})(\theta) = g_T(\theta) - \widehat{\Sigma}(\theta,\theta_0)\,\widehat{\Sigma}(\theta_0,\theta_0)^{-1}\,g_T(\theta_0)\,. \tag{3}$$

We show in Section 3.2 that this process is a sufficient statistic for $m_T(\cdot) \in \mathcal{M}_0$ in a Gaussian problem where the residual term in model (1) is exactly zero and the covariance of $G(\cdot)$ is known ($\widehat{\Sigma}(\cdot,\cdot) = \Sigma(\cdot,\cdot)$). Thus the conditional distribution of any statistic $R = R(g_T, \widehat{\Sigma})$ given $h_T(\cdot)$ does not depend on the nuisance parameter $m_T(\cdot)$. Following the classical conditioning approach (see e.g. Lehmann and Romano (2005)) we create a test by comparing R with quantiles of its conditional distribution given the process $h_T(\cdot)$ under the null.

To simulate the conditional distribution of statistic R given $h_T(\cdot)$ we take independent

draws $\xi^* \sim N(0, \widehat{\Sigma}(\theta_0, \theta_0))$ and produce simulated processes

$$g_T^*(\theta) = h_T(\theta) + \widehat{\Sigma}(\theta, \theta_0)\,\widehat{\Sigma}(\theta_0, \theta_0)^{-1}\,\xi^*. \tag{4}$$

We then calculate $R^* = R(g_T^*, \widehat{\Sigma})$, which represents a random draw from the conditional distribution of R given h_T under the null in the exact Gaussian problem. In practice our conditional test can be calculated by simulation as follows. First, take a large number of draws $\xi_b^* \sim N(0, \widehat{\Sigma}(\theta_0, \theta_0))$ for $b = 1, ..., B$. Then, for each draw calculate the statistic $R_b^* = R(g_{T,b}^*, \widehat{\Sigma})$ using the process $g_{T,b}^*$ defined as in equation (4) with ξ_b^* in place of ξ^*. Finally, for $R_{(b)}^*$ the b-th smallest value among $\{R_b^*, b = 1, ..., B\}$, the test rejects if $R(g_T, \widehat{\Sigma})$ exceeds $R_{(\lceil (1-\alpha)B \rceil)}^*$.

3.2 Exact Gaussian problem

In this section we consider an exact Gaussian problem that abstracts from some finite-sample features but leaves the central challenge of inference with an infinite-dimensional nuisance parameter intact. Consider a statistical experiment in which we observe the process $g_T(\theta) = m_T(\theta) + G(\theta)$, where $m_T(\cdot) \in \mathcal{M}$ is an unknown deterministic mean function, and $G(\cdot)$ is a mean-zero Gaussian process with known covariance $\Sigma(\theta, \tilde{\theta}) = EG(\theta)G(\tilde{\theta})'$. We again let $\mathcal{M}$ denote the set of potential mean functions, which is in general infinite-dimensional, and wish to test the hypothesis $H_0 : m_T(\theta_0) = 0$.

Lemma 1 below shows that the process $h_T(\cdot)$ is a sufficient statistic for the unknown function $m_T(\cdot)$ under the null $m_T(\cdot) \in \mathcal{M}_0$. The validity of this statement hinges on the observation that under the null the process $g_T(\cdot)$ can be decomposed into two independent random components – the process $h_T(\cdot)$ and the random vector $g_T(\theta_0)$:

$$g_T(\theta) = h_T(\theta) + \Sigma(\theta, \theta_0)\,\Sigma(\theta_0, \theta_0)^{-1}\,g_T(\theta_0), \tag{5}$$

with the important property that the distribution of $g_T(\theta_0) \sim N(0, \Sigma(\theta_0, \theta_0))$ does not depend on the nuisance parameter $m_T(\cdot)$. In particular, this implies that the conditional distribution of any functional of $g_T(\cdot)$ given $h_T(\cdot)$ does not depend on $m_T(\cdot)$.

Assume we wish to construct a test that rejects the null hypothesis when the statistic $R = R(g_T, \Sigma)$, calculated using the observed $g_T(\cdot)$ and the known covariance $\Sigma(\cdot, \cdot)$, is

large. Define the conditional critical value function $c_\alpha(h_T)$ by

$$c_\alpha\left(\tilde{h}\right) = \min\left\{c : P\left\{R\left(g_T, \Sigma\right) > c \mid h_T = \tilde{h}\right\} \leq \alpha\right\}.$$

Note that the conditional quantile $c_\alpha(\cdot)$ does not depend on the unknown $m_T(\cdot)$, and that for any realization of $h_T(\cdot)$ it can be easily simulated as described above.

Lemma 1 *In the exact Gaussian problem the test that rejects the null hypothesis $H_0 : m_T \in \mathcal{M}_0$ when $R(g_T, \Sigma)$ exceeds the $(1-\alpha)$-quantile $c_\alpha(h_T)$ of its conditional distribution given $h_T(\cdot)$ has correct size. If the conditional distribution of R given h_T is continuous almost surely then the test is conditionally similar given $h_T(\cdot)$. In particular, in this case for any $m_T \in \mathcal{M}_0$ we have that almost surely*

$$P\left\{R(g_T, \Sigma) > c_\alpha(h_T) \mid h_T(\cdot)\right\} = P\left\{R(g_T, \Sigma) > c_\alpha(h_T)\right\} = \alpha.$$

The conditional quantiles $c_\alpha(h_T)$ can be interpreted as data-dependent critical values. Under an almost sure continuity assumption the proposed test is conditionally similar in that it has conditional size α for almost every realization of h_T. It is worth emphasizing that the conditional critical values are simulated assuming the null is true and require no assumption about the behavior of the GMM sample moments under any alternative.

The are many ways to represent a given test using different statistics and data-dependent critical values. In particular, for any statistic $S(g_T)$ such that the test which rejects when $S(g_T) > 0$ has correct size, the test which rejects when the statistic $R(g_T, \Sigma)$ exceeds the random critical value $R(g_T, \Sigma) - S(g_T)$ will have correct size as well, and indeed will be the same test in that it rejects for precisely the same realizations of the data. Nonetheless, there is a sense in which the test we suggest, which rejects when $R(g_T, \Sigma) > c_\alpha(h_T)$, is particularly associated with the test statistic R. Specifically, it is the conditionally similar test most associated with large values of R under any distribution consistent with the null and can be viewed as a best approximation (within the class of size-α conditionally similar tests) to any test based on R which uses a fixed critical value. These properties are formalized in the Lemma below, and are discussed further in the Supplementary Appendix. This result builds on the approach of Moreira and Moreira (2011), though we provide a separate proof.

Let $\Phi_{C,\alpha}$ denote the class of (potentially randomized) size-α tests of $H_0 : m_T(\theta_0) = 0$

which are conditionally similar given h_T. For a given realization of the data a test from this class may use an auxiliary randomization to produce an outcome $\tilde{\phi} \in \{0,1\}$ while also satisfying the conditional size restriction $E[\tilde{\phi}|h_T] = \alpha$ under any $m_T \in \mathcal{M}_0$.[4]

Lemma 2 *Assume that the conditional distribution of R given h_T is almost surely continuous. For any non-decreasing function $F(\cdot)$ and any $m_T \in \mathcal{M}_0$ the test $\phi = \mathbb{I}\{R > c_\alpha(h_T)\}$ solves the problem $\max_{\tilde{\phi} \in \Phi_{C,\alpha}} E\left[\tilde{\phi} F(R)\right]$. If $F(\cdot)$ is strictly increasing, then ϕ is the almost surely unique solution. Moreover, the test ϕ solves the approximation problem $\min_{\tilde{\phi} \in \Phi_{C,\alpha}} E\left[\left(\tilde{\phi} - \phi^*\right)^2\right]$ for $\phi^* = \mathbb{I}\{R > c^*\}$ any test based on R with a non-random critical value c^*.*

Conditional similarity is a very strong restriction and may be hard to justify in some cases as it greatly reduces the class of possible tests. If, however, one is interested in similar tests (tests with exact size α regardless of the value of the nuisance parameter), all such tests will automatically be conditionally similar given a sufficient statistic if the family of distributions for the sufficient statistic under the null is boundedly complete – we refer the interested reader to Moreira (2003) and Section 4.3 of Lehmann and Romano (2005) for further discussion of this point.

If the parameter space for θ is finite ($\Theta = \{\theta_0, \theta_1, ..., \theta_n\}$) the conditions for bounded completeness are well-known and easy to verify. In particular, in this case our problem reduces to that of observing a $k(n+1)$-dimensional Gaussian vector $g_T = (g_T(\theta_0)', ..., g_T(\theta_n)')'$ with unknown mean $(0, \mu_1' = m_T(\theta_1)', ..., \mu_n' = m_T(\theta_n)')'$ and known covariance. If the set $\mathcal{M}$ of possible values for the nuisance parameter $(\mu_1', ..., \mu_n')'$ contains a rectangle with a non-empty interior then the family of distributions for h_T under the null is boundedly complete, and all similar tests are conditionally similar given h_T. A generalization of this statement to cases with infinite-dimensional nuisance parameters is provided in the Supplementary Appendix.

While similarity is still a strong restriction, similar tests have been shown to perform well in other weakly identified contexts, particularly in linear IV: see D. Andrews Moreira and Stock (2008). On a practical level, as we detail below, the presence of the infinite-dimensional nuisance parameter $m_T \in \mathcal{M}_0$ renders many other approaches to constructing valid tests unappealing in the present context, as alternative approaches greatly restrict the set of models considered, the set of test statistics permitted, or both.

[4]Section 3.5 in Lehmann and Romano notes that one can always represent a randomized test in this way.

3.3 Relation to the literature

Moreira (2003) pioneered the conditional testing approach in linear IV models with homoscedastic errors, which are a special case of our Example 2. If we augment Example 2 by assuming that the instruments Z_t are non-random and the reduced form errors are Gaussian with mean zero and known covariance matrix Ω, we obtain a model satisfying the assumptions of the exact Gaussian problem in each sample size. In particular, for each T we observe the process $g_T(\theta) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} (Y_t - D_t'\theta) Z_t$, which is Gaussian with mean function $m_T(\theta) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} E[Z_t D_t'](\theta_0 - \theta)$ and covariance function

$$\Sigma(\theta, \tilde{\theta}) = \left(\frac{1}{T} \sum_{t=1}^{T} Z_t Z_t' \right) (1, -\theta') \Omega (1, -\tilde{\theta}')'.$$

In this case the mean function $m_T(\cdot)$ and the process $g_T(\cdot)$ are both linear in θ. The conditioning process $h_T(\cdot)$ (derived in the Supplementary Appendix) is

$$h_T(\theta) = \frac{1}{\sqrt{T}(1, -\theta_0')\Omega(1, -\theta_0')'} [Z'Y, Z'D] \, \Omega^{-1} \begin{pmatrix} \theta_0' \\ I_p \end{pmatrix} B(\theta - \theta_0),$$

where B is a full rank $p \times p$ matrix. Thus, h_T is linear as well. Moreira (2003) proposed conditioning inference in this model on the statistic $[Z'Y, Z'D] \, \Omega^{-1} (\theta_0, I_p)'$. Thus, in this context the conditioning we propose is equivalent to that suggested by Moreira (2003), and our approach is a direct generalization of Moreira (2003) to nonlinear models. Consequently, when applied to the QLR statistic in homoscedastic linear IV our approach yields the CLR test of Moreira (2003), which D. Andrews, Moreira, and Stock (2006) showed is nearly uniformly most powerful in a class of invariant similar two-sided tests in the homoscedastic Gaussian linear IV model with a single endogenous regressor.

Kleibergen (2005) generalized the conditioning approach of Moreira (2003) to nonlinear GMM models. Kleibergen (2005) restricts attention to tests which depend on the data only through $g_T(\theta_0)$ and $\frac{d}{d\theta} g_T(\theta_0)$, which he assumes to be jointly asymptotically Gaussian. To construct valid tests he makes inferences conditional on a statistic he referred to as D_T, which can be interpreted as the part of $\frac{d}{d\theta} g_T(\theta_0)$ that is independent of $g_T(\theta_0)$. One can easily show, however, that Kleibergen's D_T is the negative of $\frac{d}{d\theta} h_T(\theta_0)$. Moreover, one can decompose $h_T(\cdot)$ into the random matrix $\frac{d}{d\theta} h_T(\theta_0)$ and a process which is independent of both $\frac{d}{d\theta} h_T(\theta_0)$ and $g_T(\theta_0)$, so the conditional distribution

of any function of $g_T(\theta_0)$ and $\frac{d}{d\theta}g_T(\theta_0)$ given $h_T(\cdot)$ is simply its conditional distribution given $\frac{d}{d\theta}h_T(\theta_0)$. Thus, for the class of tests considered in Kleibergen (2005) our conditioning approach coincides with his.[5] Unlike Kleibergen (2005), however, our approach can construct tests that depend on the full process $g_T(\cdot)$, not just on its behavior local to the null. In particular our approach allows us to consider conditional QLR tests, which are outside the scope of Kleibergen's approach in nonlinear models. Kleibergen (2005) introduces what he terms a GMM-M statistic, which coincides with the CLR statistic in homoscedastic linear IV and is intended to extend the properties of the CLR statistic to more general settings, but this statistic unfortunately has behavior quite different from that of a true QLR statistic in some empirically relevant settings, as we demonstrate in an empirical application to the Euler equation example 1 in Section 6.2 and in the linear IV simulations with non-homoscedastic errors given in the Supplementary Appendix.

Unconditional tests with nuisance parameters. In models with finite-dimensional nuisance parameters, working alternatives to the conditioning approach include least favorable and Bonferroni critical values. Least favorable critical values search over the space of nuisance parameters to maximize the $(1-\alpha)$-quantile of the test statistic, and this approach was successfully implemented by D. Andrews and Guggenberger (2009) in models with a finite-dimensional nuisance parameter. Unfortunately, however, in cases with a functional nuisance parameter the least-favorable value is typically unknown and a simulation search is computationally infeasible, rendering this approach unattractive. Bonferroni critical values are similar to least favorable ones, save that instead of searching over the whole space of nuisance parameters we instead search only over some preliminary confidence set. Again, absent additional structure this approach is typically only feasible when the nuisance parameter is of finite dimension. Relatedly, D. Andrews and Cheng (2012) show that in the settings they consider the behavior of estimators and test statistics local to a point of identification failure is controlled by a finite-dimensional nuisance parameter and use this fact to construct critical values for QLR and Wald statistics which control size regardless of the value of this parameter.

Common ways to calculate critical values in other contexts include sub-sampling and the bootstrap. Both of these approaches are known to fail to control size for many

[5]The CQLR tests suggested by D. Andrews and Guggenberger (2014) are also in this class, and depend on the data only through the moment condition and its derivative at the null.

test statistics even in cases with finite-dimensional nuisance parameters, however (see D. Andrews and Guggenberger (2009)), and thus cannot be relied on in the present setting. Indeed, it is straightforward to construct examples demonstrating that neither sub-sampling nor the bootstrap yields valid critical values for the QLR statistic in general.

3.4 Is there useful information outside θ_0?

By allowing tests depending on $g_T(\cdot)$, rather than just on $g_T(\theta_0)$ and $\frac{d}{d\theta}g_T(\theta_0)$, our results enlarge the class of tests and weakly increase the attainable power against any alternative. It is reasonable to ask, however, whether enlarging the class of tests in this way offers any strict power improvements. A full comparison of power envelopes for tests that use the whole process $g_T(\cdot)$ against those that use only the information at θ_0 is beyond the scope of the present paper, but in this section we consider the simpler problem of testing the null that the true structural parameter value is θ_0 against the alternative that it is θ^*. In this context, we show that a feasible test using non-local information can have power exceeding an infeasible power envelope for the tests studied in the previous literature.

Provided the GMM model is correctly specified, testing θ_0 against θ^* corresponds to testing $H_0 : m_T(\theta_0) = 0$ against $H_1 : m_T(\theta_0) \neq 0, m_T(\theta^*) = 0$. Even in this simplified setting both the null and the alternative are composite, with k-dimensional nuisance parameters $\lambda = m_T(\theta^*)$ under the null and $\mu = m_T(\theta_0)$ under the alternative. In the Supplementary Appendix we derive three power envelopes for tests based on $g_T(\theta_0)$ alone. The first power envelope, which we label PE-1, corresponds to the power of the most powerful test against alternative μ. This test rejects when

$$g_T(\theta_0)'\Sigma(\theta_0,\theta_0)^{-1}\mu/\sqrt{\mu'\Sigma(\theta_0,\theta_0)^{-1}\mu} \tag{6}$$

exceeds the $1-\alpha$-quantile of the standard normal distribution and is biased (i.e. has power less than α against some alternatives). The second power envelope (PE-2) corresponds to the power of the most powerful *unbiased* test against alternative μ, which rejects when the square of the expression in (6) exceeds the $1-\alpha$-quantile of an χ^2_1 distribution. Finally, the third power envelope (PE-3) corresponds to the power of the uniformly most powerful test invariant to linear transformations of the moments, which is the S or Anderson-Rubin test rejecting when $S(\theta_0) = g_T(\theta_0)'\Sigma(\theta_0,\theta_0)g_T(\theta_0)$ exceeds the $1-\alpha$ quantile of an χ^2_k distribution. These envelopes are strictly ranked, in that PE-1 always

exceeds PE-2, which in turn always exceeds PE-3. Of these three envelopes only PE-3 is feasible, since PE-1 and PE-2 require knowledge of μ. In the Supplementary Appendix we show that PE-2 is an upper bound on the power of most of the tests studied in the previous literature, including the K, JK and GMM-M tests of Kleibergen (2005).

To demonstrate the value of non-local information we consider one of the new tests allowed by our approach, which rejects when

$$g_T(\theta_0)'\Sigma(\theta_0,\theta_0)^{-1}g_T(\theta_0) - g_T(\theta^*)'\Sigma(\theta^*,\theta^*)^{-1}g_T(\theta^*) > c_\alpha(h_T), \tag{7}$$

for h_T the sufficient statistic for λ under the null and $c_\alpha(h)$ the conditional critical value discussed in this paper. This test is essentially the conditional QLR test when we restrict attention to GMM parameter values $\{\theta_0, \theta^*\}$, so we term this the point-wise QLR (pQLR) test. Note that the pQLR test does not require knowledge of μ and so is feasible in the two-point testing problem we consider here. We show in the Supplementary Appendix that the pQLR test maximizes weighted average power for a family of weight functions in this problem.

For power comparisons it is without loss of generality to normalize $\Sigma(\theta_0,\theta_0) = \Sigma(\theta^*,\theta^*) = I_k$. To obtain a simple parameterization of the covariance structure, and to simplify power comparisons, we focus on the case where $\Sigma(\theta_0,\theta^*) = \rho I_k$.[6] Under this restriction power functions of all tests considered depend only on $\|\mu\|$, k, and ρ. In the Supplementary Appendix we show that under this restriction the pQLR test is the uniformly most powerful similar test based on $(g_T(\theta_0), g_T(\theta^*))$ that is invariant to linear transformations of the moments, and further show that it is unbiased.

Figure 1 depicts the power comparison for $k = 5$, $\rho = 0.3, 0.6, 0.9, 0.99$, and a range of values $\|\mu\|$. When $\rho = 0$ (not shown) the pQLR and S tests are equivalent, while for $\rho > 0$ the power of the pQLR test exceeds that of the S test (and thus PE-3). For small ρ the power curve of the pQLR test lies below PE-2, while for large ρ the power of the pQLR test exceeds this power envelope. Indeed, for ρ close to 1 the power function of the pQLR test approaches PE-1. Note that the assumed structure on $\Sigma(\theta_0,\theta^*)$ plays a role in these results. While the pQLR test continues to have very good performance overall even without this assumption there are some cases, for example particular parameter values in

[6]The matrix $\Sigma(\theta_0,\theta^*)$ will be of this form if, for example, the covariance matrix of $(g_T(\theta_0)', g_T(\theta^*)')'$ can be written as the Kronecker product of a 2×2 matrix with a $k\times k$ matrix.

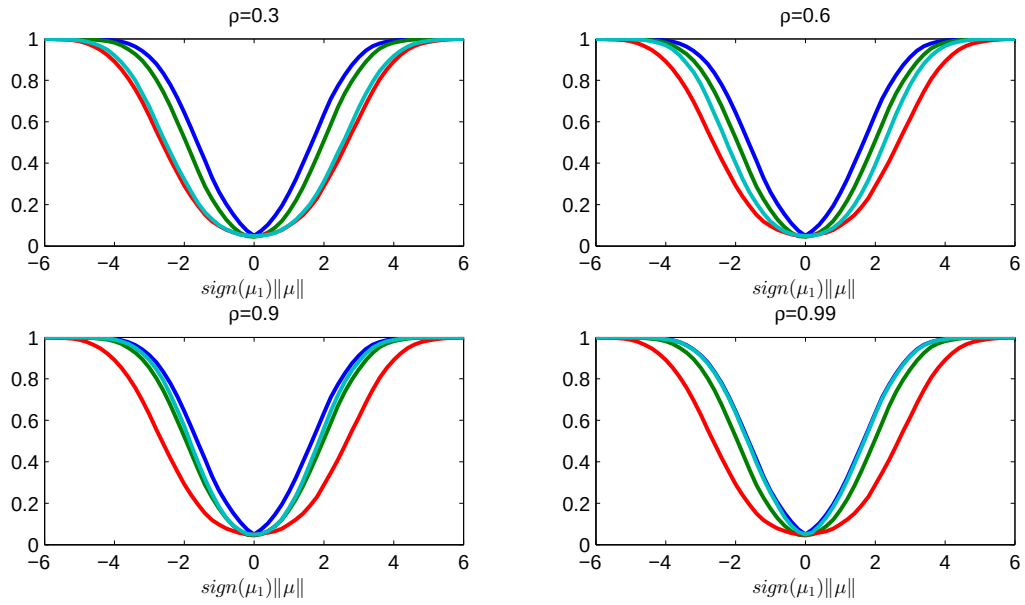

Figure 1: Power envelope (PE-1) for the class of tests based on $g_T(\theta_0)$ (blue), power envelope (PE-2) the class of unbiased tests based on $g_T(\theta_0)$ (green), power functions of the conditional pQLR (cyan) and S (PE-3, red) tests. Upper left panel: $\rho = 0.3$. Upper right panel: $\rho = 0.6$. Lower left panel: $\rho = 0.9$. Lower right panel: $\rho = 0.99$. All tests have size 5%.

the non-homoscedastic IV simulations reported in the Supplementary Appendix, where the power of the pQLR test may be less than that of the S (or AR) test.

To summarize, the newly available pQLR test is unbiased when $\Sigma(\theta_0, \theta^*) = \rho I_k$ and for large ρ has greater power than the *infeasible* best unbiased test (PE-2) based on $g_T(\theta_0)$ which in turn bounds the power attained by most of the tests suggested in the weak identification literature to date. Moreover, the power of the pQLR test approaches the power envelope for the class of all tests based on $g_T(\theta_0)$ (PE-1) even though the tests yielding this envelope are biased and based on knowledge of μ while pQLR is unbiased and feasible. Thus, we see that using information from $g_T(\cdot)$ at points other than θ_0 allows power improvements in testing one structural parameter value against another.

In the Supplementary Appendix we generalize the stylized example above by adding more points to the parameter space and considering the QLR rather than pQLR test. We find that the QLR test is competitive with the infeasible power envelope PE-2 for large ρ, though unlike the pQLR test it does not uniformly dominate PE-2 in any of the cases studied. Further, unlike in the previous section the QLR test can be biased and there are regions of the parameter space where it has less power than the S test.

We do not include power functions for tests which also use $\frac{d}{d\theta} g_T(\theta_0)$, such as K or GMM-M, in Figure 1 since without further restrictions nearly any behavior for the deriva-

tive of the sample moment is consistent with given values for $m_T(\theta_0)$ and $m_T(\theta^*)$. Since the power of the K and GMM-M tests will be very sensitive to this choice, to ensure a realistic comparison we defer power simulations with these tests to the quantile IV simulations below and to the linear IV simulations in the Supplementary Appendix. This section instead considers power envelope PE-2 which, as noted above, gives an upper bound on the power of both the K and GMM-M tests, as well as on most of the other tests studied in the previous literature. The next subsection discusses factors affecting the power of tests in applications.

3.5 Power of feasible tests

The previous subsection considers the problem of testing one structural parameter value against another. In applications, however, we are typically interested in testing $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$, rather than against a specific θ^*. Consequently, rather than using the pQLR test (which, unlike in the two-point testing problem considered above, will typically be badly biased in this setting) we consider the conditional QLR test, which can be viewed as a version of the pQLR test which attempts to "estimate" the alternative θ^*. While common, this "plug-in" approach makes it difficult to obtain finite-sample optimality results for likelihood ratio tests even in parametric models, and we claim no optimality for the QLR test in general.

Since the true value of μ is unknown, feasible tests trying to replicate PE-2 must likewise approximate the direction of μ in some way. If $m_T(\cdot)$ is linear in θ then since $m_T(\theta^*) = 0$ at the true alternative θ^* we know that $\mu = m_T(\theta_0) = \frac{\partial}{\partial\theta} m_T(\theta_0)(\theta_0 - \theta^*)$, so for θ scalar (as we assume here for simplicity) we can replace μ in equation (6) by $\frac{\partial}{\partial\theta} m_T(\theta_0)$ and obtain that the resulting test is the uniformly most powerful unbiased test based on $g_T(\theta_0)$ against any alternative θ^*. Even when $m_T(\cdot)$ is nonlinear in θ tests constructed in this manner will perform well provided the direction of $\frac{\partial}{\partial\theta} m_T(\theta_0)$ is similar to that of $m_T(\theta_0)$. Even this test is infeasible, however, since $\frac{\partial}{\partial\theta} m_T(\theta_0)$ is unknown. $\frac{\partial}{\partial\theta} g_T(\theta_0)$ gives an unbiased estimate of this quantity but, as noted by Kleibergen (2005), $\frac{\partial}{\partial\theta} g_T(\theta_0)$ will typically be correlated with $g_T(\theta_0)$ so simply replacing μ in equation (6) by $\frac{\partial}{\partial\theta} g_T(\theta_0)$ leads to a test which does not in general control size.

To avoid this issue, Kleibergen's K test instead replaces μ by

$$D_T = \frac{\partial}{\partial\theta} g_T(\theta_0) - Cov\left(\frac{\partial}{\partial\theta} g_T(\theta_0), g_T(\theta_0)\right) \Sigma(\theta_0,\theta_0)^{-1} g_T(\theta_0),$$

which as noted above is equal to $-\frac{\partial}{\partial\theta} h_T(\theta_0)$ and is independent of $g_T(\theta_0)$ by construction. Under mild conditions D_T is approximate normally distributed in large samples, $D_T \sim N(\mu_D, \Sigma_D)$, where $\mu_D = \frac{\partial}{\partial\theta} m_T(\theta_0) - Cov(\frac{\partial}{\partial\theta} g_T(\theta_0), g_T(\theta_0))\Sigma(\theta_0,\theta_0)^{-1} m_T(\theta_0)$. As noted in Kleibergen (2005), the K test will be locally asymptotically efficient in well-identified models. By contrast, in weakly identified models three distinct issues arise which may affect the power of this and related tests: (i) D_T may be noisy, with Σ_D large relative to μ_D, (ii) D_T can differ systematically from $\frac{\partial}{\partial\theta} m_T(\theta_0)$, with $\mu_D \not\propto \frac{\partial}{\partial\theta} m_T(\theta_0)$, and (iii) when $m_T(\theta)$ is nonlinear the direction of $\frac{\partial}{\partial\theta} m_T(\theta_0)$ may differ from that of $m_T(\theta_0)$.

Problem (i) arises even in the linear IV model with homoscedastic errors. In that setting it is always the case that $\mu_D \propto m_T(\theta_0) \propto \frac{\partial}{\partial\theta} m_T(\theta_0)$, but μ_D may be small relative to Σ_D, with the result that the K test can have non-monotonic power and behave erratically at particular alternatives. The extent of the problem can be assessed based on the variance-normalized length of D_T, which is precisely what the CLR test of Moreira (2003) does. In particular, I. Andrews (2015) shows that the CLR test can be interpreted as a test based on a convex combination of the S and K statistics, with the weight given to the K statistic increasing in the length of D_T.

Problem (ii) can arise in as simple a setting as linear IV with non-homoscedastic errors, since in that context it is not in general the case that $\mu_D \propto \frac{\partial}{\partial\theta} m_T(\theta_0)$ (again, see I. Andrews (2015) for further discussion). This may potentially result in poor power for the K test, as shown in the non-homoscedastic linear IV simulations given in the Supplementary Appendix. Furthermore, while the CLR test is nearly optimal in homoscedastic linear IV its generalizations, like the GMM-M test of Kleibergen (2005), can too closely resemble the K test in the non-homoscedastic case and thus have poor power. Issue (ii) also arises in a wide range of nonlinear settings and is by no means unique to linear IV.

Problem (iii) is closely tied to nonlinearity of the mean function. In extreme cases one could have that $\frac{\partial}{\partial\theta} m_T(\theta_0)\Sigma(\theta_0,\theta_0)^{-1} m_T(\theta_0) = 0$, in which case the "ideal" score test that replaces μ in equation (6) by $\frac{\partial}{\partial\theta} m_T(\theta_0)$ would have power equal to size, and even without problems (i) and (ii) the K test would perform poorly. By contrast the QLR test is not motivated by approximate linearity and so may perform better in such

cases. Further, Section 4.2 below shows that the conditional QLR test remains locally asymptotically efficient in the well-identified case. In weakly identified models where the $\frac{\partial}{\partial\theta}m_T(\theta_0)$ is approximately proportional to $m_T(\theta_0)$, on the other hand, it seems plausible that the K or GMM-M tests may outperform the QLR test by virtue of exploiting this proportionality. The extent to which either of these possibilities is borne out in practice, and how they interact with problems (i) and (ii) above, will depend on the specifics of the model under consideration and it seems difficult to draw general conclusions. We return to these questions in the quantile IV simulations below.

4 Asymptotic behavior of conditional tests

4.1 Uniform validity

The exact Gaussian problem we study in the previous section assumes away many finite-sample features relevant in empirical work, including non-Gaussianity of g_T and error in estimating the covariance function Σ. In this section we extend our results to allow for these issues and show that our conditioning approach yields uniformly asymptotically valid tests over large classes of models in which the observed process $g_T(\cdot)$ is uniformly asymptotically Gaussian.

Let P be a probability measure describing the distribution of $g_T(\cdot)$, where T denotes the sample size. For each probability law P there is a deterministic mean function $m_{T,P}(\cdot)$, which in many cases will be the expectation $E_P g_T(\cdot)$ of the process $g_T(\cdot)$ under P. We assume that the difference $g_T(\cdot) - m_{T,P}(\cdot)$ converges to a mean zero Gaussian process $G_P(\cdot)$ with covariance function $\Sigma_P(\cdot,\cdot)$ uniformly over the family $\mathcal{P}_0$ of distributions consistent with the null. We formulate this assumption using bounded Lipschitz convergence – see section 1.12 of Van der Vaart and Wellner (1996) for the equivalence between bounded Lipschitz convergence and weak convergence of stochastic processes. For simplicity of notation we suppress the subscript P in all expressions.

Assumption 1 *The difference $g_T(\cdot) - m_T(\cdot)$ converges to a Gaussian process $G(\cdot)$ with mean zero and covariance function $\Sigma(\cdot,\cdot)$ uniformly over $P \in \mathcal{P}_0$, that is:*

$$\lim_{T\to\infty} \sup_{P\in\mathcal{P}_0} \sup_{f\in BL_1} \|E\left[f\left(g_T - m_T\right)\right] - E\left[f\left(G\right)\right]\| = 0,$$

where BL_1 is the set of functionals with Lipschitz constant and supremum norm bounded above by one.

Assumption 2 *The covariance function $\Sigma(\cdot,\cdot)$ is uniformly bounded and positive definite:*

$$1/\bar{\lambda} \leq \inf_{P\in\mathcal{P}_0}\inf_{\theta\in\Theta} \lambda_{\min}\left(\Sigma\left(\theta,\theta\right)\right) \leq \sup_{P\in\mathcal{P}_0}\sup_{\theta\in\Theta} \lambda_{\max}\left(\Sigma\left(\theta,\theta\right)\right) \leq \bar{\lambda},$$

for some finite $\bar{\lambda} > 0$.

Assumption 3 *There is a uniformly consistent estimator $\widehat{\Sigma}(\cdot,\cdot)$ of the covariance function, in that for any $\varepsilon > 0$*

$$\lim_{T\to\infty}\sup_{P\in\mathcal{P}_0} P\left\{\sup_{\theta,\tilde{\theta}}\left\|\widehat{\Sigma}(\theta,\tilde{\theta}) - \Sigma(\theta,\tilde{\theta})\right\| > \varepsilon\right\} = 0.$$

Discussion of Assumptions 1-3 As previously discussed, Assumption 1 imposes a uniform central limit theorem uniformly over $\mathcal{P}_0$. Assumption 2 requires that the covariance function be uniformly bounded and uniformly full rank, which rules out the reduced-rank case considered in D. Andrews and Guggenberger (2014). The possibility of extending the results of the present paper to contexts with possibly degenerate variance is an interesting question for future work. Assumption 3 requires that we have a uniformly consistent estimate for the covariance function. Under smoothness conditions point-wise laws of large numbers yield uniform laws of large numbers as for example in Newey (1991). The same logic applies to dependent data (see for example, Wooldridge (1994)), but one must use a HAC-type estimator.

Suppose we are interested in tests that reject for large values of a statistic R that depends on the moment function $g_T(\cdot)$ and the estimated covariance $\widehat{\Sigma}(\cdot,\cdot)$. Consider the process $h_T(\cdot) = H(g_T, \widehat{\Sigma})$ defined as in equation (3). Since the transformation from $(g_T(\cdot), \widehat{\Sigma}(\cdot,\cdot))$ to $(g_T(\theta_0), h_T(\cdot), \widehat{\Sigma}(\cdot,\cdot))$ is one-to-one, R can be viewed as a functional of $(g_T(\theta_0), h_T(\cdot), \widehat{\Sigma}(\cdot,\cdot))$.

Assumption 4 *The functional $R(\xi, h(\cdot), \Sigma(\cdot,\cdot))$ is defined for all values $\xi \in \mathbb{R}^k$, all k-dimensional functions h with the property that $h(\theta_0) = 0$, and all covariance functions $\Sigma(\cdot,\cdot)$ satisfying Assumption 2. For any fixed $C > 0$, $R(\xi, h, \Sigma)$ is bounded and Lipschitz in ξ, h, and Σ over the set of $(\xi, h(\cdot), \Sigma(\cdot,\cdot))$ with $\xi'\Sigma(\theta_0,\theta_0)^{-1}\xi \leq C$.*

Lemma 3 *The QLR statistic defined in equation (2) satisfies Assumption 4.*

We require that R be sufficiently continuous with respect to $(g_T(\theta_0), h_T(\cdot), \widehat{\Sigma}(\cdot,\cdot))$, which rules out Wald statistics in many models but allows the QLR statistic. Further, this assumption allows QLR-type statistics based on other weighting matrices, and statistics constructed analogously to QLR that use some other norm (for example the sup or L^1 norms) to assess the length of $g_T(\theta)$.

To calculate our conditional critical values, given a realization of h_T we simulate independent draws $\xi \sim N(0, \widehat{\Sigma}(\theta_0, \theta_0))$ and (letting P^* denote the simulation probability) define

$$c_\alpha(h_T, \widehat{\Sigma}) = \inf\left\{c : P^*\left\{\xi : R(\xi, h_T(\cdot), \widehat{\Sigma}(\cdot,\cdot)) \leq c\right\} \geq 1-\alpha\right\}.$$

The test then rejects if $R(g_T(\theta_0), h_T, \widehat{\Sigma}) > c_\alpha(h_T, \widehat{\Sigma})$.

Theorem 1 *If Assumptions 1 - 4 hold, then for any $\varepsilon > 0$ we have*

$$\lim_{T\to\infty} \sup_{P\in\mathcal{P}_0} P\left\{R(g_T(\theta_0), h_T, \widehat{\Sigma}) > c_\alpha(h_T, \widehat{\Sigma}) + \varepsilon\right\} \leq \alpha.$$

Theorem 1 shows that our conditional critical value (increased by an arbitrarily small amount) results in a test that is uniformly asymptotically valid over the large class of distributions $\mathcal{P}_0$. The need for the term ε reflects the possibility that there may be some sequences of distributions in $\mathcal{P}_0$ under which $R - c_\alpha$ converges in distribution to a limit that is not continuously distributed. If we rule out this possibility, for example assuming that the distribution of $R - c_\alpha$ is continuous with uniformly bounded density for all T and all $P \in \mathcal{P}_0$, then the conditional test with $\varepsilon = 0$ is uniformly asymptotically similar in the sense of D. Andrews, Cheng and Guggenberger (2011).

4.2 Strong identification case

Restricting attention to conditionally similar tests rules out many procedures and so could come at a substantial cost in terms of power. In this section, we show that restricting attention to conditionally similar tests does not result in a loss of power if the data are in fact generated from a strongly identified model, by which we mean one satisfying conditions given below. In particular, we establish that under these conditions our conditional QLR test is equivalent to the classical QLR test using χ^2 critical values and so retains the efficiency properties of the usual QLR test.

Assumption 5 *Suppose there exists a subset $\mathcal{P}_{00} \subseteq \mathcal{P}_0$ such that for some sequence of numbers δ_T converging to zero and each $P \in \mathcal{P}_{00}$, there exists a sequence of matrices M_T such that for any $\varepsilon > 0$:*

(i) $\lim_{T\to\infty} \inf_{P\in\mathcal{P}_{00}} \inf_{\|\theta-\theta_0\|>\delta_T} m_T(\theta)'\Sigma(\theta,\theta)^{-1}m_T(\theta) = \infty$,

(ii) $\lim_{T\to\infty} \sup_{P\in\mathcal{P}_{00}} \sup_{|\theta-\theta_0|\leq\delta_T} |m_T(\theta) - M_T(\theta-\theta_0)| = 0$,

(iii) $\lim_{T\to\infty} \inf_{P\in\mathcal{P}_{00}} \delta_T^2 \lambda_{min}(M_T'\Sigma(\theta_0,\theta_0)^{-1}M_T) = \infty$,

(iv) $\lim_{T\to\infty} \sup_{P\in\mathcal{P}_{00}} \sup_{\|\theta-\theta_0\|\leq\delta_T} \|\Sigma(\theta,\theta) - \Sigma(\theta_0,\theta_0)\| = 0$ *and*
$\lim_{T\to\infty} \sup_{P\in\mathcal{P}_{00}} \sup_{\|\theta-\theta_0\|\leq\delta_T} \|\Sigma(\theta,\theta_0) - \Sigma(\theta_0,\theta_0)\| = 0$,

(v) $\lim_{T\to\infty} \sup_{P\in\mathcal{P}_{00}} P\left\{\sup_{\|\theta-\theta_0\|\leq\delta_T} |G(\theta) - G(\theta_0)| > \varepsilon\right\} = 0$,

(vi) there exists a constant C such that $\sup_{P\in\mathcal{P}_{00}} P\{\sup_{\theta\in\Theta} |G(\theta)| > C\} < \varepsilon$.

Discussion of Assumption 5. Assumption 5 defines what we mean by strong identification. Part (i) guarantees that the moment function diverges outside of a shrinking neighborhood of the true parameter value and, together with assumption (vi), implies the existence of consistent estimators. Part (ii) requires that the unknown mean function $m_T(\theta)$ be linearizable on a neighborhood of θ_0, which plays a key role in establishing the asymptotic normality of estimators. Part (iii) follows from parts (i) and (ii) if we require m_T to be uniformly continuously differentiable at θ_0, while parts (iv)-(vi) are regularity conditions closely connected to stochastic equicontinuity. In particular, (iv) requires that the covariance function be continuous at θ_0, while (v) requires that G be equicontinious at θ_0, and (vi) requires that G be bounded almost surely.

Parts (i)-(iii) of Assumption 5 are straightforward to verify in the classical GMM setting. Consider a GMM model as in Section 2.1 that satisfies Assumptions 1 and 2 with mean function $Eg_T(\theta) = m_T(\theta) = T^{\frac{1}{2}-\alpha}m(\theta)$, where $0 \leq \alpha < \frac{1}{2}$ and $m(\theta)$ is a fixed, twice-continuously-differentiable function with $m(\theta) = 0$ iff $\theta = \theta_0$. Assume further that $m(\theta)$ is continuously differentiable at θ_0 with full-rank Jacobian $\frac{\partial}{\partial\theta}m(\theta_0) = M$, and that the parameter space Θ is compact. For $\delta_T = T^{-\gamma}$, $\inf_{\|\theta-\theta_0\|>\delta_T} m_T(\theta)'\Sigma(\theta,\theta)^{-1}m_T(\theta) \approx CT^{1-2\alpha-2\gamma}$ so if $0 < \gamma < \frac{1}{2} - \alpha$, then part (i) of Assumption 5 holds. Taylor expansion shows that

$$\sup_{|\theta-\theta_0|<\delta_T} |m_T(\theta) - M_T(\theta-\theta_0)| \leq T^{1/2-\alpha} q^2 \sup_{\theta\in\Theta} \sup_{i,j} \left|\frac{\partial^2 m(\theta)}{\partial\theta_i\partial\theta_j}\right| \delta_T^2,$$

so for $\gamma > \frac{1}{2}(\frac{1}{2} - \alpha)$ part (ii) holds. Finally, $M_T = T^{\frac{1}{2}-\alpha}M$, thus part (iii) holds if $\gamma < \frac{1}{2} - \alpha$. To summarize, parts (i)-(iii) hold for any γ with $\frac{1}{2}(\frac{1}{2} - \alpha) < \gamma < \frac{1}{2} - \alpha$.

This derivation assumes a common rate of estimation $\frac{1}{2} - \alpha$ for the full parameter vector θ. It is straightforward, however, to show that Assumption 5 may also be satisfied when different elements of the parameter vector are identified at different rates provided that all rates are above $\frac{1}{4}$. This last restriction is a sufficient condition for the existence of a common $\delta_T = T^{-\gamma}$ satisfying all requirements, and is what Antoine and Renault (2009) term the "nearly strong" case.

Theorem 2 *Suppose Assumptions 1-3 and 5 hold, then the QLR statistic defined in equation (2) converges in distribution to a χ^2_q uniformly over $\mathcal{P}_{00}$ as the sample size increases to infinity, while at the same time the conditional critical value $c_\alpha(h_T, \widehat{\Sigma})$ converges in probability to the $1-\alpha$-quantile of an χ^2_q-distribution. Thus while the conditional QLR test controls size uniformly over $\mathcal{P}_0$, under strong identification it is asymptotically equivalent to the classical unconditional QLR test over $\mathcal{P}_{00}$.*

Theorem 2 concerns behavior under the null but can be extended to local alternatives. Define local alternatives to be sequences of alternatives which are contiguous in the sense of Le Cam (see, for example, chapter 10 in Van der Vaart and Wellner (1996)) with sequences in $\mathcal{P}_{00}$ satisfying Assumption 5. By the definition of contiguity, under all such sequences of local alternatives $c_\alpha(h_T, \widehat{\Sigma})$ will again converge to a χ^2_q critical value, implying that our conditional QLR test coincides with the usual QLR test under these sequences.

5 Concentrating out nuisance parameters

As highlighted in Section 2, processes $g_T(\cdot)$ satisfying Assumptions 1-3 arise naturally when one considers normalized moment conditions in GMM estimation. Such processes arise in other contexts as well, however. In particular, one can often obtain such moment functions by "concentrating out" well-identified structural nuisance parameters. This is of particular interest for empirical work, since in many empirical settings one wishes to test a hypothesis concerning a subset of the structural parameters, while the remaining structural (nuisance) parameters are unrestricted. In this section we show that if we have a well-behaved estimate of the structural nuisance parameters (in a sense made precise

below, and somewhat stronger than the conditions of Section 4.2), a normalized moment function based on plugging in this estimator provides a process $g_T(\cdot)$ which satisfies Assumptions 1-3. We then show that these results may be applied to test hypotheses on the coefficients on the endogenous regressors in quantile IV models, treating the parameters on the exogenous controls as strongly-identified nuisance parameters.

In this section we assume that we begin with a $(q+p)$-dimensional structural parameter which can be written as (β, θ), where we are interested in testing a hypothesis $H_0 : \theta = \theta_0$ concerning only the q-dimensional parameter θ. We assume that the parameter space for (β, θ) is the Cartesian product of their individual parameter spaces. The hypothesis of interest is thus that there exists some value β_0 of the nuisance parameter β such that the k-dimensional moment condition $Eg_T^{(L)}(\beta_0, \theta_0) = 0$ holds. Here we use superscript (L) to denote the "long" or non-concentrated moment condition and define a corresponding "long" mean function $m_T^{(L)}(\beta, \theta)$. We assume there exists a function $\beta(\theta)$, which we call the pseudo-true value of parameter β for a given value of θ, satisfying $m_T^{(L)}(\beta(\theta_0), \theta_0) = 0$. For values of θ different from the null value θ_0 the model from which $\beta(\theta)$ comes may be (and often will be) misspecified. This presents no difficulties for us, as our only requirement will be the existence of an estimator $\widehat{\beta}(\theta)$ of $\beta(\theta)$ which is $\sqrt{T}$-consistent and asymptotically normal uniformly over θ. Under additional regularity conditions, we then show that we can use the concentrated moment function $g_T(\theta) = g_T^{(L)}(\widehat{\beta}(\theta), \theta)$ to implement our inference procedure. In what follows we again drop the subscript P for simplicity of notation.

Assumption 6 *There exists a function $\beta(\theta)$ which for all θ belongs to the interior of the parameter space for β and satisfies $m_T^{(L)}(\beta(\theta_0), \theta_0) = 0$, and an estimator $\widehat{\beta}(\theta)$ such that $\left(g_T^{(L)}(\beta, \theta) - m_T^{(L)}(\beta, \theta), \sqrt{T}(\widehat{\beta}(\theta) - \beta(\theta))\right)$ are jointly uniformly asymptotically normal,*

$$\lim_{T\to\infty} \sup_{P\in\mathcal{P}_0} \sup_{f\in BL_1} \left| E_P \left[f \left(\begin{array}{c} g_T^{(L)}(\beta, \theta) - m_T^{(L)}(\beta, \theta) \\ \sqrt{T}(\widehat{\beta}(\theta) - \beta(\theta)) \end{array} \right) \right] - E\left[f(\mathbb{G})\right] \right| = 0.$$

where $\mathbb{G} = (G^{(L)}(\beta, \theta), G_\beta(\theta))$ is a mean-zero Gaussian process with covariance function $\Sigma_L(\beta, \theta, \beta_1, \theta_1)$, such that process $\mathbb{G}$ is uniformly equicontinuous and uniformly bounded over $\mathcal{P}_0$.

Assumption 7 *Assume that the covariance function is uniformly bounded, uniformly*

positive definite, and uniformly continuous in β along $\beta(\theta)$. In particular, for fixed $\bar{\lambda} > 0$ and any sequence $\delta_T \to 0$ we have

$$1/\bar{\lambda} \leq \inf_{P\in\mathcal{P}_0} \inf_{\theta} \lambda_{\min}(\Sigma_L(\beta(\theta),\theta,\beta(\theta),\theta)) \leq \sup_{P\in\mathcal{P}_0} \sup_{\theta} \lambda_{\max}(\Sigma_L(\beta(\theta),\theta,\beta(\theta),\theta)) \leq \bar{\lambda};$$

$$\lim_{T\to\infty} \sup_{P\in\mathcal{P}_0} \sup_{\theta,\theta_1} \sup_{\|\beta-\beta(\theta)\|<\delta_T} \sup_{\|\beta_1-\beta(\theta_1)\|<\delta_T} \|\Sigma_L(\beta,\theta,\beta_1,\theta_1) - \Sigma_L(\beta(\theta),\theta,\beta(\theta_1),\theta_1)\| = 0.$$

Assumption 8 *There is an estimator $\widehat{\Sigma}_L(\beta,\theta,\beta_1,\theta_1)$ of $\Sigma_L(\beta,\theta,\beta_1,\theta_1)$ such that*

$$\lim_{T\to\infty} \sup_{P\in\mathcal{P}_0} P\left\{ \sup_{\beta,\theta,\beta_1,\theta_1} \left\| \widehat{\Sigma}_L(\beta,\theta,\beta_1,\theta_1) - \Sigma_L(\beta,\theta,\beta_1,\theta_1) \right\| > \varepsilon \right\} = 0.$$

Assumption 9 *For some sequence $\delta_T \to \infty$, $\delta_T/\sqrt{T} \to 0$, for each $P \in \mathcal{P}_0$ there exists a deterministic sequence of $k \times p$ functions $M_T(\theta)$ such that:*

$$\lim_{T\to\infty} \sup_{P\in\mathcal{P}_0} \sup_{\theta} \sup_{\sqrt{T}|\beta-\beta(\theta)|\leq\delta_T} \left\| m_T^{(L)}(\beta,\theta) - m_T^{(L)}(\beta(\theta),\theta) - M_T(\theta)\sqrt{T}(\beta-\beta(\theta)) \right\| = 0.$$

We assume that these functions $M_T(\theta)$ are uniformly bounded: $\sup_{P\in\mathcal{P}_0} \sup_\theta \|M_T(\theta)\| < \infty$, and that there exists an estimator $\widehat{M}_T(\theta)$ such that

$$\lim_{T\to\infty} \sup_{P\in\mathcal{P}_0} P\left\{ \sup_{\theta} \left\| \widehat{M}_T(\theta) - M_T(\theta) \right\| > \varepsilon \right\} = 0.$$

Discussion of Assumptions Assumptions 6-8 extend Assumptions 1- 3, adding strong-identification conditions for β. In particular, Assumption 6 states that there exists a consistent and asymptotically normal estimator $\widehat{\beta}(\theta)$ uniformly over θ. The uniform equicontinuity and boundedness of $\mathbb{G}$ are as in Assumption 5 parts (v) and (vi). Assumption 7 additionally guarantees that the rate of convergence for $\widehat{\beta}(\theta)$ is uniformly $\sqrt{T}$, and Assumption 8 guarantees that the covariance function is well-estimable. The estimator $\widehat{\beta}(\theta)$ may come from within the initial set of moment restrictions or from additional moment conditions not used for testing, or even from another data set. For $\theta \neq \theta_0$ the estimator $\widehat{\beta}(\theta)$ may come from a misspecified model and thus fail to consistently estimate the true β_0. This is fine, as noted above, but we require that $\widehat{\beta}(\theta)$ be $\sqrt{T}$-consistent for β_0 when evaluated at θ_0. Note that if the estimator $\widehat{\beta}(\theta)$ is obtained using some subset of the initial moment conditions $\tilde{g}_T^{(L)}$, the covariance matrix Σ_L may be degenerate along some directions, violating Assumption 7. In such cases we should refor-

mulate the moment condition to exclude the redundant directions. For example, suppose $\widehat{\beta}(\theta) = \arg\min_\beta \tilde{g}_T^{(L)}(\beta,\theta)' W(\theta) \tilde{g}_T^{(L)}(\beta,\theta)$, and let $A^\perp(\theta)$ be a $(k-p)\times k$ matrix orthogonal to the $p \times k$ matrix $\frac{\partial \tilde{g}_T^{(L)}(\beta,\theta)}{\partial \beta}' W(\theta)\Big|_{\beta=\widehat{\beta}(\theta)}$. We can use $g_T^{(L)}(\beta,\theta) = A^\perp(\theta)\tilde{g}_T^{(L)}(\beta,\theta)$ as the "long" moment condition.

Assumption 9 supposes that $m_T^{(L)}$ is linearizable in β in the neighborhood of $\beta(\theta)$. In many GMM models $m_T^{(L)}(\beta,\theta) = \sqrt{T} E\varphi^{(L)}(X_t,\beta,\theta)$ and thus we have

$$M_T(\theta) = \frac{\partial}{\partial \beta} E\varphi^{(L)}(X_t,\beta,\theta) \mid_{\beta=\beta(\theta)} .$$

This last expression is typically consistently estimable provided $E\varphi^{(L)}(X_t,\beta,\theta)$ is twice-continuously-differentiable in β, in which case Assumption 9 comes from a Taylor expansion in β around $\beta(\theta)$. Note the close relationship between Assumption 9 and Assumption 5 part (ii).

Theorem 3 *Let Assumptions 6-9 hold, then the moment function $g_T(\theta) = g_T^{(L)}(\widehat{\beta}(\theta),\theta)$, mean function $m_T(\theta) = m_T^{(L)}(\beta(\theta),\theta)$, covariance function*

$$\Sigma(\theta,\theta_1) = (I_k, M_T(\theta))\, \Sigma_L(\beta(\theta),\theta,\beta(\theta_1),\theta_1)\, (I_k, M_T(\theta_1))',$$

and its estimate

$$\widehat{\Sigma}(\theta,\theta_1) = \left(I_k, \widehat{M}_T(\theta)\right) \widehat{\Sigma}_P(\widehat{\beta}(\theta),\theta,\widehat{\beta}(\theta_1),\theta_1) \left(I_k, \widehat{M}_T(\theta_1)\right)',$$

satisfy Assumptions 1-3.

The proof of Theorem 3 may be found in the Supplementary Appendix.

The assumption that the nuisance parameter β is strongly-identified, specifically the existence of a uniformly consistent and asymptotically normal estimator $\widehat{\beta}(\theta)$ and the linearizability of $m_T^{(L)}(\beta,\theta)$ in β, plays a key role here. D. Andrews and Cheng (2012) and I. Andrews and Mikusheva (2016) show that in models with weakly identified nuisance parameters the asymptotic distributions of many statistics will depend on the unknown values of the nuisance parameter, greatly complicating inference. In such cases, rather than concentrating out the nuisance parameter we may instead use the projection method. The projection method tests the continuum of hypotheses $H_0 : \theta = \theta_0, \beta = \beta_0$ for different values of β_0, and rejects the null $H_0 : \theta = \theta_0$ only if all hypotheses of the form

$H_0 : \theta = \theta_0, \beta = \beta_0$ are rejected. Thus, even in cases where the nuisance parameter may be poorly identified one can test $H_0 : \theta = \theta_0$ by applying our conditioning method to test a continuum of hypotheses $H_0 : \theta = \theta_0, \beta = \beta_0$ provided the corresponding $g_T^{(L)}(\beta, \theta)$ processes satisfy Assumptions 1-3.

5.1 Example: quantile IV regression

To illustrate our results on concentrating out nuisance parameters we build on Example 3 above and consider inference on the coefficients on the endogenous regressors in a quantile IV model, where we now allow for the possibility of additional exogenous regressors. This setting has been studied in Chernozhukov and Hansen (2008), where the authors used an Anderson-Rubin-type statistic, and in Jun (2008) where K and J statistics were suggested. Here we propose inference using the conditional QLR test.

We consider an instrumental-variables model of quantile treatment effects as in Chernozhukov and Hansen (2005). Let the data consist of i.i.d. observations on an outcome variable Y_t, a vector of endogenous regressors D_t, a vector of exogenous controls C_t, and a $k \times 1$ vector of instruments Z_t. Following Chernozhukov and Hansen (2006) we assume a linear-in-parameters model for the τ-quantile treatment effect, known up to parameter $\psi = (\beta, \theta)$, and will base inference on the moment condition

$$E\left[\left(\tau - \mathbb{I}\left\{Y_t \leq C_t'\beta_0 + D_t'\theta_0\right\}\right)\begin{pmatrix} C_t \\ Z_t \end{pmatrix}\right] = 0. \tag{8}$$

If we were interested in joint inference on the parameters (β, θ) we could simply view this model as a special case of GMM. In practice, however, we are often concerned with the coefficient θ on the endogenous regressor, so β is a nuisance parameter, and we would prefer to conduct inference on θ alone. To do this we can follow Jun (2008) and obtain for each value θ an estimate $\widehat{\beta}(\theta)$ for β by running a standard, linear-quantile regression of $Y_t - D_t'\theta$ on C_t. In particular, define

$$\widehat{\beta}(\theta) = \arg\min_{\beta} \frac{1}{T}\sum_{t=1}^{T} \rho_\tau(Y_t - D_t'\theta - C_t'\beta),$$

where $\rho_\tau(\cdot)$ is the τ-quantile check function. The idea of estimating $\widehat{\beta}(\theta)$ from simple quantile regression, introduced in Chernozhukov and Hansen (2008), is easy to implement

and computationally feasible. Under mild regularity conditions, $\widehat{\beta}(\theta)$ will be a consistent and asymptotically-normal estimator for the pseudo-true value $\beta(\theta)$ defined by

$$E\left[\left(\tau - \mathbb{I}\left\{Y_t \leq C_t'\beta\left(\theta\right) + D_t'\theta\right\}\right)C_t\right] = 0 \tag{9}$$

for each θ. If we then define the concentrated moment function

$$g_T\left(\theta\right) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(\tau - \mathbb{I}\left\{Y_t \leq C_t'\widehat{\beta}\left(\theta\right) + D_t'\theta\right\}\right)Z_t,$$

mean function

$$m_T\left(\theta\right) = \sqrt{T}E\left[\left(\tau - \mathbb{I}\left\{Y_t \leq C_t'\beta\left(\theta\right) + D_t'\theta\right\}\right)Z_t\right],$$

and the covariance estimator

$$\widehat{\Sigma}(\theta_1,\theta_2) = \frac{1}{T}\sum_{t=1}^{T}\Big[\Big(\tau - \mathbb{I}\{\varepsilon_t(\widehat{\beta}(\theta_1),\theta_1) < 0\}\Big)\Big(\tau - \mathbb{I}\{\varepsilon_t(\widehat{\beta}(\theta_2),\theta_2) < 0\}\Big)\cdot \\ \cdot\Big(Z_t - \widehat{A}(\theta_1)C_t\Big)\Big(Z_t - \widehat{A}(\theta_2)C_t\Big)'\Big],$$

where $\varepsilon(\beta,\theta) = Y_t - D_t'\theta - C_t'\beta$, $\widehat{A}(\theta) = \widehat{M}_T(\theta)\widehat{J}^{-1}(\theta)$, $k(\cdot)$ is a kernel, and

$$\widehat{M}_T(\theta) = \frac{1}{Th_T}\sum_{t=1}^{T}Z_tC_t'k\left(\frac{\varepsilon_t(\widehat{\beta}(\theta),\theta)}{h_T}\right), \widehat{J}(\theta) = \frac{1}{Th_T}\sum_{t=1}^{T}C_tC_t'k\left(\frac{\varepsilon_t(\widehat{\beta}(\theta),\theta)}{h_T}\right),$$

we show in the Supplementary Appendix that these choices satisfy Assumptions 6-9 under the following regularity conditions:

Assumption 10 *(i)* (Y_t, C_t, D_t, Z_t) *are i.i.d.,* $E\left[\|C_t\|^4\right] + E\left[\|D_t\|^{2+\varepsilon}\right] + E\left[\|Z_t\|^4\right]$ *is uniformly bounded above, and the matrix* $E[(C_t', Z_t')(C_t', Z_t')']$ *is full rank.*

(ii) The conditional density $f_{\varepsilon(\theta)}(s|C_t, D_t, Z_t)$ *of* $\varepsilon(\theta) = Y_t - D_t'\theta - C_t'\beta(\theta)$ *is uniformly bounded over the support of* (C_t, D_t, Z_t) *and is twice continuously differentiable at* $s = 0$ *with a second derivative that is uniformly continuous in* θ*;*

(iii) For each θ *the value of* $\beta(\theta)$ *defined in equation (9) is in the interior of the parameter space;*

(iv) $\inf_\theta \lambda_{\min}(J(\theta)) > 0$ *for* $J(\theta) = E\left[f_{\varepsilon(\theta)}(0)C_tC_t'\right]$*;*

(v) The kernel $k(v)$ *is such that* $\sup|k(v)| < \infty$, $\int |k(v)|dv < \infty$, $\int k(v)dv = 1$, *and* $\int k^2(v)dv < \infty$.

Under Assumption 10, one may use the QLR statistic paired with conditional critical values to construct confidence sets for θ in this model. In Section 6 we provide simulation results comparing the performance of the conditional QLR test with known alternatives. Both Chernozhukov and Hansen (2008) and Jun (2008) suggested Anderson-Rubin type statistics for this model which have stable power but are inefficient in over-identified models under strong identification. To overcome this inefficiency, Jun (2008) introduced a K test analogous to that of Kleibergen (2005). This test is locally efficient under strong identification and has good power for small violations of the null hypothesis regardless of identification strength. However, K tests often suffer from substantial declines in power at distant alternatives. To overcome this deficiency a number of approaches to combining the K and AR statistics have been suggested by different authors, including the JK test discussed by Jun (2008), which is expected to improve power against distant alternatives but is inefficient under strong identification. For general GMM models Kleibergen (2005) proposes the GMM-M test, which is motivated by analogy to the CLR test of Moreira (2003) and is based on a data-dependent combination of the AR and K statistics. This test is locally efficient under strong identification, though Jun (2008) does not generalize this approach to his setting. Our approach allows one to use the conditional QLR test, which retains efficiency under strong identification without sacrificing power at distant alternatives.

6 Numerical performance of the conditional QLR test

In this section we examine the performance of the conditional QLR test in numerical examples, first simulating performance in quantile IV and then constructing confidence sets for Euler equation parameters in US data via test inversion.

6.1 Simulations: quantile IV model

We simulate the performance of the QLR test in a quantile IV model with a single endogenous regressor and k instruments. As in Example 3 above we generate outcome

variable Y_t from the location-scale model

$$Y_t = \gamma_1 + \gamma_2 D_t + (\gamma_3 + \gamma_4 D_t)\, U_t,$$

where D_t is almost-surely positive, U_t has median zero, and we have a vector of instruments Z_t which are independent of U_t. For each quantile these variables obey the linear quantile IV model of Chernozhukov and Hansen (2005) with control variable $C_t \equiv 1$. For our simulations we focus on the median, $\tau = \frac{1}{2}$, and use the moment condition (8) for $\beta_0 = \gamma_1$ and $\theta_0 = \gamma_2$. We are interested in inference on the coefficient θ on the endogenous regressor, treating the intercept β as a nuisance parameter as described in Section 5.1.

A wide array of distributions for (D_t, U_t, Z_t) is consistent with the quantile IV moment condition, and as shown in Example 3 above (for a simplified model without an intercept), this can lead to a wide variety of potential mean functions $m_T(\cdot)$. As discussed in Section 3.5 the performance of tests based on local information will depend, among other factors, on whether the direction of $m_T(\theta_0)$ is well approximated by that of $\frac{\partial m_T(\theta_0)}{\partial \theta}$. When this approximation is valid derivative-based procedures like the K, JK, and GMM-M tests may be expected to perform relatively well, while breakdown of this approximation is labeled problem (iii) in Section 3.5 and may yield poor power for these tests. To compare the power of the tests developed in this paper relative to those in the previous literature, we consider two different simulation designs for (D_t, U_t, Z_t). In the "symmetric" simulation design the direction of $m_T(\theta_0)$ coincides with that of $\frac{\partial m_T(\theta_0)}{\partial \theta}$, while in the "asymmetric" simulation design all the issues discussed in Section 3.5 above arise.

Symmetric Simulation Design. We first consider a simulation design in which the instruments play symmetric roles, in the sense that the joint distribution of $(D_t, U_t, Z_{t,1}, ..., Z_{t,k})$ is invariant with respect to the relabeling of the instruments. This implies the direction of the mean function $m_T(\theta)$ is proportional to the vector $(1, ..., 1)'$ for all θ. In this case one can show that in the notation of Section 3.5 above $\mu_D \propto \frac{\partial}{\partial \theta} m_T(\theta_0) \propto m_T(\theta_0)$, so problems (ii) and (iii) discussed in Section 3.5 do not arise. This gives an advantage to the K and related tests that use this proportionality, relative to the conditional QLR test which does not.

For our symmetric design, we draw $(U_t, D_t, Z_t') = (\Phi(\xi_{U,t}) - \frac{1}{2}, \Phi(\xi_{D,t}), \Phi(\xi_{Z_1,t}), ..., \Phi(\xi_{Z_k,t}))$, where $(\xi_{U,t}, \xi_{D,t}, \xi_{Z_1,t}, ..., \xi_{Z_k,t})'$ is a Gaussian vector with mean zero, all variances equal

to one, $\text{cov}(\xi_U, \xi_D) = \rho_S$, $\text{cov}(\xi_D, \xi_{Z_j}) = \pi_S$, and all other covariances are zero, and Φ is the standard normal distribution function. In this model ρ_S measures the endogeneity of the regressor D_t: if $\rho_S = 0$ then there is no endogeneity, and a linear quantile regression of Y_t on D_t and a constant will yield consistent estimates of (β, θ). If, on the other hand, $\rho_S \neq 0$, then we need to adopt a quantile IV strategy to obtain consistent estimates. The parameter π_S controls the strength of identification under the quantile IV approach, and the model will be partially identified when $\pi_S = 0$ and weakly identified when π_S is close to zero.

Asymmetric Simulation Design. The symmetry of the instruments in the simulation design above is quite special, as each instrument brings an independent and equal amount of information about D_t. For our asymmetric design we draw $(U_t, D_t, Z_t') = (\xi_{U,t}, \exp(2\xi_{D,t}), \xi_{Z,t}, \xi_{Z,t}^2, ..., \xi_{Z,t}^k)$, where $(\xi_{U,t}, \xi_{D,t}, \xi_{Z,t})'$ is a Gaussian vector with mean zero, all variances equal to one, $\text{cov}(\xi_{U,t}, \xi_{D,t}) = \rho_A$, $\text{cov}(\xi_{D,t}, \xi_{Z,t}) = \pi_A$, and zero covariance between $\xi_{U,t}$ and $\xi_{Z,t}$. Analogous to the parameters ρ_S and π_S in the symmetric simulation design, the parameters ρ_A and π_A control the degree of endogeneity and the strength of the instruments, respectively. This design corresponds to a case where one begins with a single instrument $\xi_{Z,t}$ and then constructs a number of technical instruments (polynomials in $\xi_{Z,t}$) to gain efficiency. While satisfying the quantile IV moment conditions this data generating process yields a quite nonlinear mean function $m_T(\cdot)$, particularly when the degree of endogeneity ρ_A is large. Thus unlike in the symmetric case, issues (ii) and (iii) discussed in Section 3.5 both arise in this context.

6.1.1 Simulation results

We compare power of tests for $H_0 : \theta = \theta_0$. The conditional QLR test is calculated as described in Section 5.1. For comparison we also calculate the weak-instrument-robust AR, K, and JK tests of Jun (2008), which are based on the same concentrated moment conditions. In Jun (2008)'s simulations the test suggested by Chernozhukov and Hansen (2008) performed quite similarly to Jun's AR test, so here we report results only for Jun's tests. We also consider a variant of the GMM-M test of Kleibergen (2005). As noted by D. Andrews and Guggenberger (2014) the construction of GMM-M tests involves a choice of rank statistic and two options, based on what they term the moment-variance and Jacobian-variance weightings, have been proposed in the literature. Since the derivation

of the Jacobian-variance weighting requires a non-trivial extension of the results of Jun (2008), here we focus on GMM-M tests with the moment-variance weighting. Finally, we report the power of the (infeasible) pQLR test which tests θ_0 against the true alternative at each point.

Our simulations set $\gamma_1 = \gamma_3 = \gamma_4 = 1$ and take the null value θ_0 to be one. We then vary the true value θ by varying γ_2. We draw samples of 1,000 observations from the models above, and vary the endogeneity parameters ρ and the identification parameters π. Here we discuss results for models with ten instruments, while results for models with five instruments are given in the Supplementary Appendix.

Figures 2 and 3 plot power curves for nominal 5% tests in the symmetric and asymmetric simulation designs, respectively. Figure 2 examines $(\rho_S, \pi_S) \in \{(0.25, 0.05), (.25, 0.1), (0.5, 0.05), (0.9, 0.05)\}$, while Figure 3 considers $(\rho_A, \pi_A) \in \{(0.25, 0.2), (.25, 0.4), (0.5, 0.2), (0.9, 0.2)\}$. As expected given the local asymptotic efficiency of the QLR, K, and GMM-M tests in well-identified models, power curves for these tests are quite close together in well-identified cases. Thus we focus here on designs where the instruments are relatively weak, while power curves for designs with stronger instruments are given in the Supplementary Appendix. Simulated size for most of the tests studied is quite close to the nominal size in all designs considered, with the exception of the K and JK tests which have simulated size close to 8% in the symmetric simulation design with $\rho_S = 0.9$. The simulated size for all tests is reported in the Supplementary Appendix. We also calculated size-corrected power curves and found them qualitatively similar to the uncorrected results (these results are available upon request).

In the symmetric simulation design, which favors tests using the orthogonalized derivative D_T, we see that the conditional QLR test performs quite competitively with, and in many cases preferably to, the other tests considered. While the K and JK tests have good power close to the null, often exceeding even the infeasible pQLR test, they suffer from substantial power declines at more distant alternatives, which stems from the high variance of D_T relative to its mean (issue (i) discussed in Section 3.5). By contrast the power of the AR, GMM-M, and QLR tests typically increases as we consider alternatives more distant from the null. The relative performance of the QLR test relative to the AR and GMM-M tests depends on the parameter value considered, with the QLR test generally performing better for higher degrees of endogeneity ρ_S.

In the asymmetric simulation designs, where the derivative of the sample moment

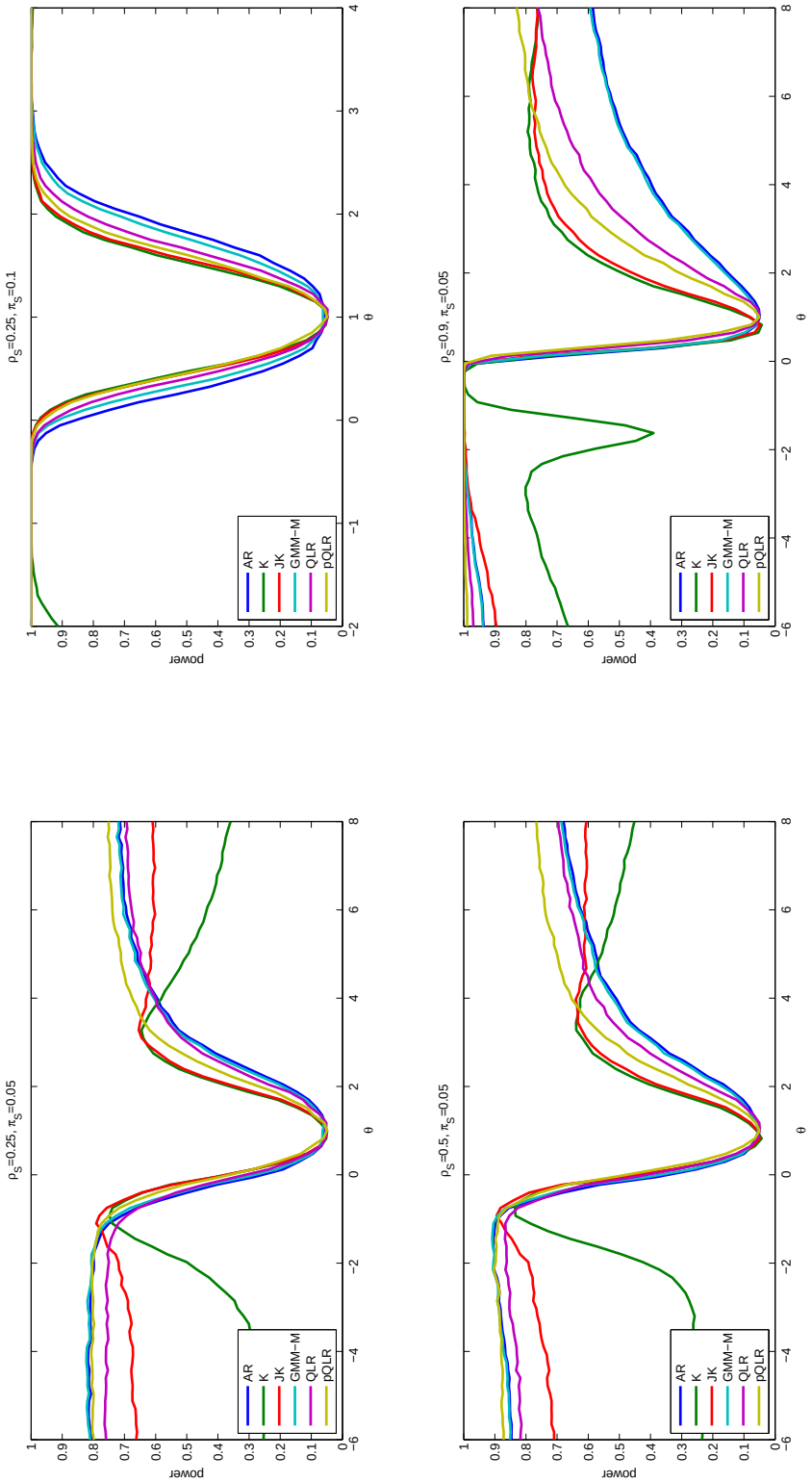

Figure 2: Power of nominal 5% tests in symmetric quantile IV simulation design with ten instruments and 1,000 observations. Based on 2,500 simulation replications and 1,000 draws of conditional critical values.

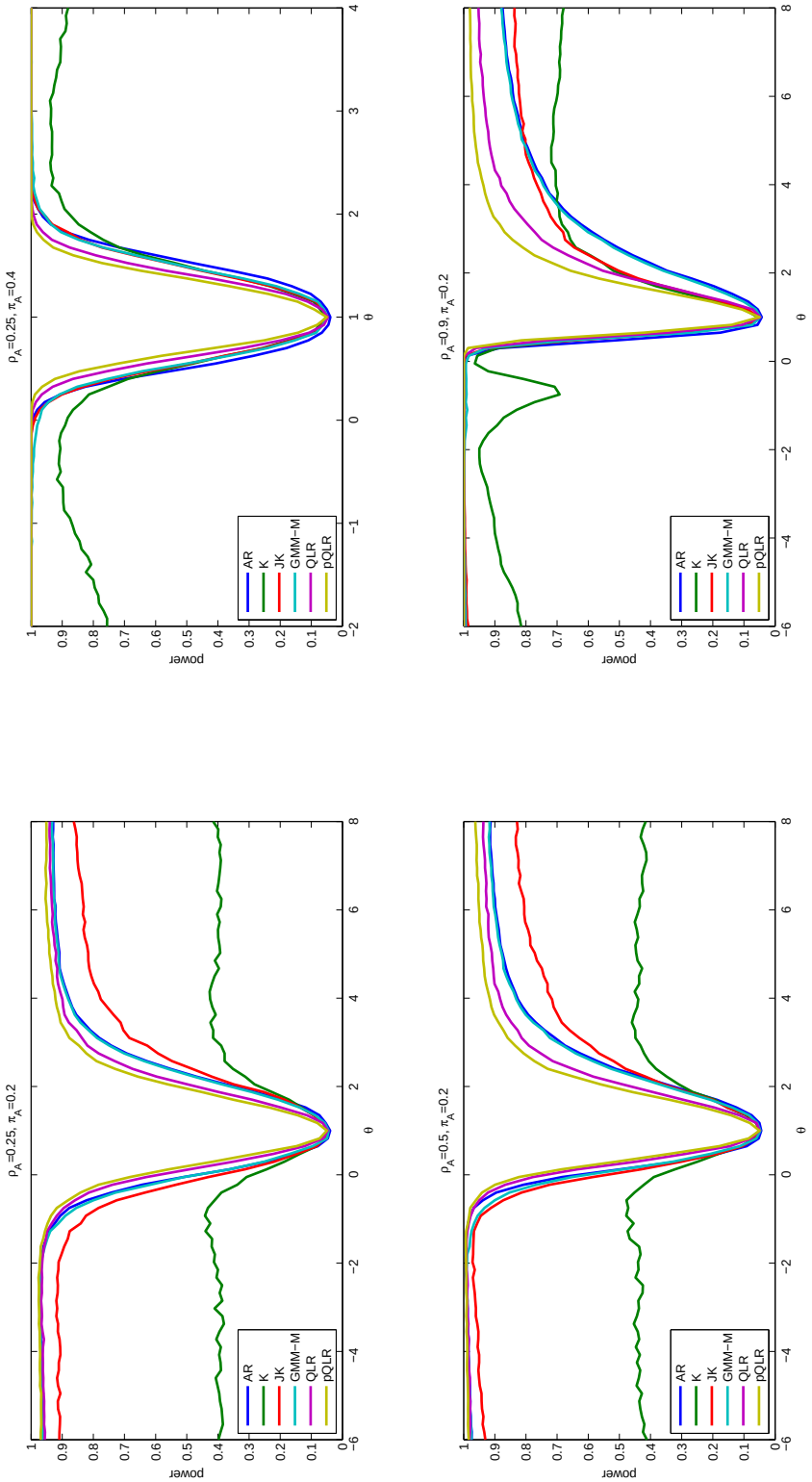

Figure 3: Power of nominal 5% tests in asymmetric quantile IV simulation design with ten instruments and 1,000 observations. Based on 2,500 simulation replications and 1,000 draws of conditional critical values.

condition is less informative about behavior of $m_T(\theta_0)$ under the alternative, we see that the power of the K and JK tests is systematically less than that of the other tests considered, even for alternatives close to the null. This is consistent with fact that all three issues discussed in Section 3.5 arise in this design. Further we see that the QLR test generally has greater power than the AR or GMM-M tests (with the difference between the QLR and GMM-M tests as large as 19% for some parameter values) and that, as in the symmetric simulation design, the difference in power is increasing in the degree of endogeneity.

Finally, we compare the conditional QLR test with the infeasible pQLR test, which as discussed in Section 3.4 is weighted average power optimal in a two-point testing problem and so may be considered a benchmark. The pQLR test has greater power than does the QLR test, though their power curves behave similarly across the designs considered. The largest amount by which the power of the QLR test falls short of that of the pQLR test in the designs reported here is 11%, while the average power shortfall (with respect to uniform weights on θ) never exceeds 5.5%.

6.2 Empirical example: Euler equation

As an empirical example, we invert the QLR and several other robust tests to calculate identification-robust confidence sets based on the nonlinear Euler equation specification discussed in Example 1. Both Stock and Wright (2000), who introduced the concept of weakly identified GMM, and Kleibergen (2005), who proposed the first identification-robust tests for GMM that are locally efficient under strong identification, used the nonlinear Euler equation as their main empirical example. Though substantial evidence suggests that this model may be misspecified, and Stock and Wright (2000) find that results in this context are very sensitive to the choice of specification, the role of this example in the literature nonetheless makes this a useful application to consider.

Following Stock and Wright (2000) we use an extension of the long annual data-set of Campbell and Shiller (1987). Our specification corresponds to the CRRA-1 specification of Stock and Wright (2000), which takes C_t to be aggregate consumption, R_t to be an aggregate stock market return and Z_t to contain a constant, C_{t-1}/C_{t-2}, and R_{t-1}, resulting in a three-dimensional moment condition ($k = 3$) – see Stock and Wright (2000) for details.

While the model implies that $\sqrt{T}g_T(\cdot)$ is a martingale when evaluated at the true parameter value, the QLR statistic also depends on the behavior of g_T away from the null. To estimate all covariance matrices we use the Newey-West estimator with one lag. We could use a martingale-difference covariance estimator in constructing the S statistic, but doing so substantially increases the volume of the joint S confidence set for (δ, γ) so we focus on the HAC formulation for comparability with the other confidence sets studied. We first construct a confidence set for the full parameter vector $\theta = (\delta, \gamma)$ and then consider inference on the risk-aversion coefficient γ alone.

6.2.1 Confidence sets for the full parameter vector

We report joint 90% confidence sets for $\theta = (\delta, \gamma)$ based on inverting QLR, S, K, JK, and GMM-M tests of Stock and Wright (2000) and Kleibergen (2005) in Figure 4.[7] As we can see, the QLR confidence set is substantially smaller than the others considered, largely due to the elimination of disconnected components of the confidence set. To quantify this difference, note that the S, K, JK, and GMM-M confidence sets cover 4.3%, 4.43%, 5.46%, and 4.5% of the parameter space $(\delta, \gamma) \in [0.6, 1.1] \times [-6, 60]$, respectively, while the QLR confidence set covers only 0.64% of the parameter space.

We emphasize that the model is likely misspecified, which may effect the relative sizes of confidence sets. A common critique of the S test is that it has power against both violations of the parametric hypothesis and model misspecification, making it difficult to interpret small confidence sets. The QLR statistic is intended to test only the parametric restriction, under the maintained hypothesis of correct specification, but is not itself robust to model misspecification. Indeed, while we have based our choice of moments on Stock and Wright (2000), as in that paper the results are sensitive to the precise moments chosen, as well as to other other factors like the number of lags used in HAC covariance estimation. This could be further evidence of misspecification in this context, which may affect the relative performance of confidence sets here.

[7] Note that our S confidence set differs from that of Stock and Wright (2000) which, in addition to assuming that the summands in $g_T(\theta_0)$ are serially uncorrelated, also assumes conditional homoscedasticity.

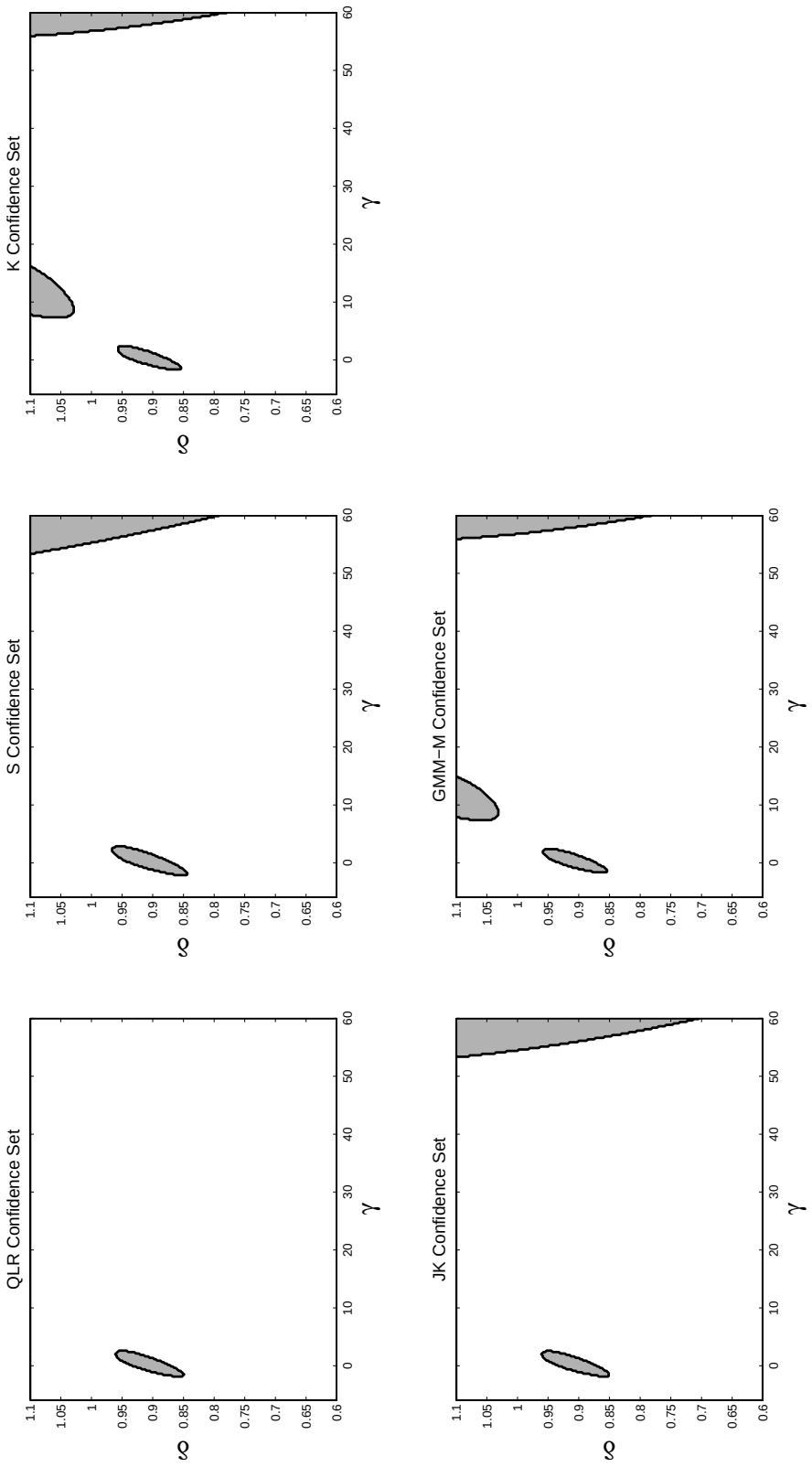

Figure 4: Joint 90% QLR, S, K, JK, and GMM-M confidence sets for risk aversion (γ) and the discount factor (δ) based on annual data, three moment conditions, and 1,000 draws of critical values.

	90% Confidence Set	Length
QLR- Constant Instrument	$[-2, 1.7]$	3.7
QLR- CUE	$[-1.3, 1.9]$	3.2
S	$[-1.6, 2.3]$	3.9
K	$[-1.1, 1.8] \cup [8, 12.3]$	7.2
JK	$[-1.2, 1.9]$	3.1
GMM-M	$[-1.1, 1.8] \cup [8, 12.3]$	7.2

Table 1: 90% confidence sets for risk aversion parameter γ, treating nuisance parameter δ as well identified, based on annual data.

6.3 Confidence sets for risk aversion

Stock and Wright (2000) argued that once one fixes the risk-aversion parameter γ the discount factor δ is well identified. Under this assumption we calculate conditional QLR confidence sets for γ based on two approaches, first by plugging in an estimator for δ based on the moment condition instrumented with a constant and then concentrating out δ using the continuous-updating estimator (CUE), where in each case we modify the moment conditions as discussed in Section 5 to account for this estimation. For comparison we consider the S, K, JK, and GMM-M tests evaluated at the restricted CUE for δ which, as Stock and Wright (2000) and Kleibergen (2005) argue, allow valid inference under the assumption that δ is well identified. The resulting confidence sets are reported in Table 1. Unlike in the joint confidence set case we see that the QLR confidence set is larger than the JK confidence set but is nonetheless the second smallest confidence set out of the five considered. Further, we see that in this application concentrating out the nuisance parameter using the CUE results in a smaller confidence set than does plugging in the estimate based on the moment condition instrumented with a constant.

7 Conclusions

This paper argues that moment equality models without any identification assumptions have a functional nuisance parameter. We introduce a sufficient statistic for this nuisance parameter and construct conditional tests. Our results substantially expand the set of statistics available in weakly- or partially-identified models, and in particular allow the use of quasi-likelihood ratio statistics, which often have superior power properties compared to widely-used Anderson-Rubin type statistics. We show that our tests have uniformly correct asymptotic size over a large class of models and find that the proposed

tests perform well in simulations in a quantile IV model and give smaller confidence sets than existing alternatives in a nonlinear Euler Equation model.

8 References

Andrews, D.W.K. and X. Cheng (2012): "Estimation and Inference with Weak, Semi-strong and Strong Identification," *Econometrica*, 80(5), 2153-2211.

Andrews, D.W.K., X. Cheng, and P. Guggenberger (2011): "Generic Results for Establishing the Asymptotic Size of Confidence Sets and Tests," Cowles Foundation Discussion Paper # 1813.

Andrews, D.W.K. and P. Guggenberger (2009): "Hybrid and Size-Corrected Subsampling Methods," *Econometrica*, 77(3), 721-762.

Andrews D.W.K. and P. Guggenberger (2014): "Identification- and Singlarity-Robust Inference for Moment Condition Models," Cowles Foundation Discussion Paper # 1977.

Andrews, D.W.K., M. Moreira, and J. Stock (2006): "Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression," *Econometrica*, 74(3), 715-752.

Andrews, D.W.K., M. Moreira, and J. Stock (2008): "Efficient Two-sided Nonsimilar Invariant Tests in IV Regression with Weak Instruments," *Journal of Econometrics*, 146, 241-54.

Andrews, I. (2015): "Conditional Linear Combination Tests for Weakly Identified Models," *mimeo*.

Andrews, I. and A. Mikusheva (2016): "A Geometric Approach to Weakly Identified Econometric Models," *Econometrica*, forthcoming.

Antoine, B. and E. Renault (2009): "Efficient GMM with Nearly-Weak Instruments," *Econometrics Journal*, 12, S135-S171.

Campbell, J.Y. and R.J. Shiller (1987): "Cointegration Tests of Present Value Models," *Journal of Political Economy*, 95(5), 1062-1088.

Chernozhukov, V. and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73(1), 245-261.

Chernozhukov, V. and C. Hansen (2006): "Instrumental Quantile Regression Inference for Structural and Treatment Effect Models," *Journal of Econometrics*, 132(2), 491-525.

Chernozhukov, V. and C. Hansen (2008): "Instrumental Variable Quantile Regression: A Robust Inference Approach," *Journal of Econometrics*, 142(1), 379-398.

Dedecker, J. and S. Louhichi (2002): "Maximal Inequalities and Empirical Central Limit Theorems," in H. Dehling, T. Mikosch and M. Sorensen (eds.) *Empirical Process Techniques for Dependent Data*, Boston: Birkhauser, pp. 137-161.

Hansen, L.P. (1982) : "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50(4), 1029-1054.

Hansen, L.P. and K. Singleton (1982): "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica*, 50(5), 1269-1286.

Jun, S.J. (2008): "Weak Identification Robust Tests in an Instrumental Quantile Model," *Journal of Econometrics*, 144(1), 118-138.

Kleibergen, F. (2005): "Testing Parameters in GMM without Assuming that They are Identified," *Econometrica*, 73(4), 1103-1124.

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

McFadden, D. (1989): "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration," *Econometrica*, 57(5), 995-1026.

Moreira, H. and M. Moreira (2011): "Inference with Persistent Regressors," *mimeo*.

Moreira, M. (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71(4), 1027-1048.

Newey, W.K. (1991): "Uniform Convergence in Probability and Stochastic Equicontinuity," *Econometrica*, 59(4), 1161-1167.

Pakes, A. and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57(5), 1027-1057.

Schennach, S. (2014): "Entropic Latent Variable Integration via Simulation," *Econometrica*, 82(1), 345-385.

Staiger, D. and J. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65(3), 557-586.

Stock, J. and J. Wright (2000): "GMM with Weak Identification," *Econometrica*, 68, 1055-96.

Van der Vaart, A.W. and J.A. Wellner (1996): *Weak Convergence and Empirical Processes*. New York: Springer.

Wooldridge, J.M. (1994) "Estimation and Inference for Dependent Processes," *Handbook of Econometrics, vol. 4*, R. Engle and D. McFadden(eds.), Amsterdam: North Holland, pp. 2639-2738.

9 Appendix

Proof of Lemma 1. The proof trivially follows from equation (5) and the observations that the distribution of $g_T(\theta_0) \sim N(0, \Sigma(\theta_0, \theta_0))$ does not depend on $m_T(\cdot)$, the function $\Sigma(\theta, \theta_0)\Sigma(\theta_0, \theta_0)^{-1}$ is deterministic and known, and the vector $g_T(\theta_0)$ is independent of $h_T(\cdot)$. $\square$

Proof of Theorem 1. Let us introduce the process

$$G_h(\theta) = H(G, \Sigma)(\theta) = G(\theta) - \Sigma(\theta, \theta_0)\,\Sigma(\theta_0, \theta_0)^{-1}\,G(\theta_0),$$

and a random variable $\xi = G(\theta_0)$ which is independent of $G_h(\cdot)$. First, we notice that Assumptions 1-3 imply that $\eta_T = (g_T(\theta_0), h_T(\cdot) - m_T(\cdot), \widehat{\Sigma}(\cdot, \cdot))$ converges uniformly to $\eta = (\xi, G_h(\cdot), \Sigma(\cdot, \cdot))$, that is,

$$\lim_{T\to\infty} \sup_{P\in\mathcal{P}_0} \sup_{f\in BL_1} |E_P[f(\eta_T)] - E[f(\eta)]| = 0. \tag{10}$$

We assume here that the distance on the space of realizations is measured as follows: for $\eta_i = (\xi_i, G_{h,i}(\cdot), \Sigma_i(\cdot, \cdot))$ (for $i = 1, 2$),

$$d(\eta_1, \eta_2) = \|\xi_1 - \xi_2\| + \sup_{\theta} \|G_{h,1}(\theta) - G_{h,2}(\theta)\| + \sup_{\theta,\tilde{\theta}} \|\Sigma_1(\theta, \tilde{\theta}) - \Sigma_2(\theta, \tilde{\theta})\|.$$

Statement (10) then follows from the observation that the function which takes $(G(\cdot), \Sigma(\cdot, \cdot))$ to $(\xi, G_h(\cdot), \Sigma(\cdot, \cdot))$ is Lipschitz in (G, Σ) if $|\xi| < C$ for some constant C, provided Σ satisfies Assumption 2.

Note that for any non-random functions $m_T(\cdot)$ random processes $\varsigma_T = (g_T(\theta_0), h_T(\cdot), \widehat{\Sigma}(\cdot, \cdot))$ and $\tilde{\varsigma}_T = (\xi, G_h(\cdot) + m_T, \Sigma(\cdot, \cdot))$ are one-to-one linear transformations of η_T and η. Thus, since bounded Lipschitz functionals of ς_T can also be expressed as bounded Lipschitz functionals of η_T, statement (10) implies:

$$\lim_{T\to\infty} \sup_{P\in\mathcal{P}_0} \sup_{f\in BL_1} |E_P[f(\varsigma_T)] - E[f(\tilde{\varsigma}_T)]| = 0. \tag{11}$$

Let us introduce the function $F(x) = \mathbb{I}\{x < C_1\} + \frac{C_2 - x}{C_2 - C_1}\mathbb{I}\{C_1 \le x < C_2\}$ for some

$0 < C_1 < C_2$ and consider the functional

$$R_F(\xi, h, \Sigma) = R(\xi, h, \Sigma) F(\xi' \Sigma (\theta_0, \theta_0)^{-1} \xi),$$

which is a continuous truncation of the functional $R(\xi, h, \Sigma) = R(g, \Sigma)$. Consider the conditional quantile function corresponding to the new statistic

$$c_{F,\alpha}(h, \Sigma) = \inf \{c : P^* \{\xi : R_F(\xi, h, \Sigma) \leq c\} \geq 1 - \alpha\}.$$

As our next step we show that $c_{F,\alpha}(h, \Sigma)$ is Lipschitz in $h(\cdot)$ and $\Sigma(\cdot, \cdot)$ for all h with $h(\theta_0) = 0$ and Σ satisfying Assumption 2.

Assumption 4 implies that there exists a constant K such that

$$\|R_F(\xi, h_1, \Sigma) - R_F(\xi, h_2, \Sigma)\| \leq K d(h_1, h_2)$$

for all ξ, h_1, h_2 and Σ. Let $c_i = c_{F,\alpha}(h_i, \Sigma)$, then

$$1 - \alpha \leq P^* \{\xi : R_F(\xi, h_1, \Sigma) \leq c_1\} \leq P^* \{\xi : R_F(\xi, h_2, \Sigma) \leq c_1 + K d(h_1, h_2)\}.$$

Thus $c_2 \leq c_1 + K d(h_1, h_2)$. Analogously we get $c_1 \leq c_2 + K d(h_1, h_2)$, implying that $c_{F,\alpha}$ is Lipschitz in h. The same argument shows that $c_{F,\alpha}$ is Lipschitz in Σ.

Assume the conclusion of Theorem 1 does not hold. Then there exists some $\delta > 0$, an infinitely increasing sequence of sample sizes T_i, and a sequence of probability measures $P_{T_i} \in \mathcal{P}_0$ such that for all i

$$P_{T_i} \left\{ R(g_{T_i}(\theta_0), h_{T_i}, \widehat{\Sigma}) > c_\alpha(h_{T_i}, \widehat{\Sigma}) + \varepsilon \right\} > \alpha + \delta.$$

Choose C_1 such that

$$\limsup P_{T_i} \left\{ g_{T_i}(\theta_0)' \widehat{\Sigma} (\theta_0, \theta_0)^{-1} g_{T_i}(\theta_0) \geq C_1 \right\} < \frac{\delta}{2},$$

which can always be chosen since according to Assumption 1 $g_T(\theta_0)$ converges uniformly

to $N(0, \Sigma(\theta_0, \theta_0))$. Since

$$P_T\{R > x\} \leq P_T\{R_F > x\} + P_T\left\{g_T(\theta_0)'\widehat{\Sigma}(\theta_0, \theta_0)^{-1} g_T(\theta_0) \geq C_1\right\},$$

and $c_{F,\alpha}(h_T, \widehat{\Sigma}) < c_\alpha(h_T, \widehat{\Sigma})$ we have that for all i

$$P_{T_i}\left\{R_F(g_{T_i}(\theta_0), h_{T_i}, \widehat{\Sigma}) \geq c_{F,\alpha}(h_{T_i}, \widehat{\Sigma}) + \varepsilon\right\} > \alpha + \frac{\delta}{2}. \tag{12}$$

Denote by $\mathcal{T}_T$ a random variable distributed as $R_F(\xi_T, h_T, \widehat{\Sigma}) - c_{F,\alpha}(h_T, \widehat{\Sigma})$ under the law P_T, and by $\mathcal{T}_{\infty,T}$ a random variable distributed as $R_F(\xi, G_h + m_T, \Sigma) - c_{F,\alpha}(G_h + m_T, \Sigma)$ under the law P_T. The difference between these variables is that the first uses the finite-sample distribution of $(\xi_T, h_T, \widehat{\Sigma})$, while the latter uses its asymptotic counterparts $(\xi, G_h + m_T, \Sigma)$. Equation (11) and the bounded Lipschitz property of the statistic R_F and the conditional critical value imply that

$$\lim_{T\to\infty} \sup_{f \in BL_1} |Ef(\mathcal{T}_T) - Ef(\mathcal{T}_{\infty,T})| = 0. \tag{13}$$

Since $\mathcal{T}_{T_i}$ is a sequence of bounded random variables, by Prokhorov's theorem there exists a subsequence T_j and a random variable $\mathcal{T}$ such that $\mathcal{T}_{T_j} \Rightarrow \mathcal{T}$. By equation (13), $\mathcal{T}_{\infty,T_j} \Rightarrow \mathcal{T}$. Since equation (12) can be written as $P\{\mathcal{T}_{T_i} \geq \varepsilon\} > \alpha + \delta/2$,

$$\liminf P\{\mathcal{T}_{\infty,T_j} > 0\} \geq P\{\mathcal{T} > 0\} \geq P\{\mathcal{T} \geq \varepsilon\} \geq \limsup P\{\mathcal{T}_{T_j} \geq \varepsilon\} \geq \alpha + \frac{\delta}{2}.$$

However, from the definition of quantiles we have

$$P\{\mathcal{T}_{\infty,T_j} > 0\} = P_T\{R_F(\xi, G_h + m_T, \Sigma) > c_{F,\alpha}(G_h + m_T, \Sigma)\} \leq \alpha,$$

since the statistic $\mathcal{T}_{\infty,T_j}$ is the statistic in the exact Gaussian problem and so controls size by Lemma 1. Thus we have reached a contradiction. □

Proof of Theorem 2. As shown in Theorem 1, Assumptions 1-3 imply that the distribution of the QLR statistic is uniformly asymptotically approximated by the distribution of the same statistic in the exact Gaussian problem. Thus, it suffices to prove the statement of Theorem 2 for the exact Gaussian problem only, which is to say when $g_T(\cdot)$ is a Gaussian process with mean $m_T(\cdot)$ and known covariance Σ. In our case

$QLR = R(g_T(\theta_0), h_T, \Sigma)$, where

$$R(\xi, h, \Sigma) = \xi' \Sigma(\theta_0, \theta_0)^{-1} \xi - \inf_{\theta} \left(V(\theta)\xi + h(\theta)\right)' \Sigma(\theta, \theta)^{-1} \left(V(\theta)\xi + h(\theta)\right), \tag{14}$$

and $V(\theta) = \Sigma(\theta, \theta_0) \Sigma(\theta_0, \theta_0)^{-1}$. Denote by $\mathcal{A}$ the event $\mathcal{A} = \{g_T(\theta_0)' \Sigma(\theta_0, \theta_0)^{-1} g_T(\theta_0) < C\}$ and note that by choosing the constant $C > 0$ large enough we can guarantee that the probability of $\mathcal{A}$ is arbitrarily close to one.

Let $\hat{\theta}_T$ be the value at which the optimum in equation (14) is achieved (the case when the optimum may not be achieved may be handled similarly, albeit with additional notation). We first show that $\hat{\theta}_T \to^p \theta_0$. For any a, b we have $(a+b)^2 \geq \frac{a^2}{2} - b^2$, so

$$\begin{aligned}(V(\theta)g_T(\theta_0) + h_T(\theta))' \Sigma(\theta, \theta)^{-1} (V(\theta)g_T(\theta_0) + h_T(\theta)) \geq \frac{1}{2} m_T(\theta)' \Sigma(\theta, \theta)^{-1} m_T(\theta) \\ -(V(\theta)g_T(\theta_0) + h_T(\theta) - m_T(\theta))' \Sigma(\theta, \theta)^{-1} (V(\theta)g_T(\theta_0) + h_T(\theta) - m_T(\theta)).\end{aligned} \tag{15}$$

Assumptions 2 and 5 (vi) guarantee that the second term on the right-hand side of equation (15) is stochastically bounded, so denote this term $A(\theta)$. For any probability $\varepsilon > 0$ there exists a constant C such that

$$\inf_{P \in \mathcal{P}_0} P\left\{\sup_{\theta \in \Theta} A(\theta) \leq C \text{ and } \mathcal{A}\right\} \geq 1 - \varepsilon.$$

Assumption 5(i) implies that there exists T_1 such that for all $T > T_1$ and $P \in \mathcal{P}_0$ we have

$$\inf_{\|\theta - \theta_0\| > \delta_T} m_T(\theta)' \Sigma(\theta, \theta)^{-1} m_T(\theta) > 4C.$$

Putting the last three inequalities together we get that for $T > T_1$ and all $P \in \mathcal{P}_0$

$$P\left\{\inf_{\|\theta - \theta_0\| > \delta_T} (V(\theta)g_T(\theta_0) + h_T(\theta))' \Sigma(\theta, \theta)^{-1} (V(\theta)g_T(\theta_0) + h_T(\theta)) > C \text{ and } \mathcal{A}\right\} \geq 1 - \varepsilon.$$

This implies that $\sup_{P \in \mathcal{P}_0} P\left\{\|\hat{\theta}_T - \theta_0\| > \delta_T\right\} \leq \varepsilon$ for all $T > T_1$.

As our second step we show that for any $\varepsilon > 0$

$$\begin{aligned}\lim_{T \to \infty} \sup_{P \in \mathcal{P}_0} P\left\{\left| \inf_{\|\theta - \theta_0\| < \delta_T} g_T(\theta)' \Sigma(\theta, \theta)^{-1} g_T(\theta) \right.\right. \\ \left.\left. - \inf_{\|\theta - \theta_0\| < \delta_T} \tilde{g}_T(\theta)' \Sigma(\theta_0, \theta_0)^{-1} \tilde{g}_T(\theta) \right| > \varepsilon \right\} = 0,\end{aligned} \tag{16}$$

where we replace the process $g_T(\theta) = V(\theta)g_T(\theta_0) + h_T(\theta)$ by the process $\tilde{g}_T(\theta) = g_T(\theta_0) + m_T(\theta)$ with the same mean function $m_T(\theta)$ and covariance $\tilde{\Sigma}(\theta, \theta_1) = \Sigma(\theta_0, \theta_0)$ for all θ, θ_1. For this new process we have $\tilde{V}(\theta) = I$ and $\tilde{h}_T(\theta) = m_T(\theta)$. To verify equation (16), restrict attention to the event $\mathcal{A}$ for some large $C > 0$. The functional that transforms $(g_T(\theta_0), h, \Sigma(\theta, \theta), V(\cdot))$ to $\inf_{\|\theta-\theta_0\|<\delta_T}(V(\theta)g_T(\theta_0) + h(\theta))'\Sigma(\theta, \theta)^{-1}(V(\theta)g_T(\theta_0) + h(\theta))$ is Lipschitz in h, V and $\Sigma(\theta, \theta)$ on $\mathcal{A}$. Thus,

$$\begin{aligned}
&\left| \inf_{\|\theta-\theta_0\|<\delta_T} g_T(\theta)'\Sigma(\theta,\theta)^{-1}g_T(\theta) - \inf_{\|\theta-\theta_0\|<\delta_T} \tilde{g}_T(\theta)'\Sigma(\theta_0,\theta_0)^{-1}\tilde{g}_T(\theta) \right| \\
&\leq K_1 \sup_{\|\theta-\theta_0\|\leq\delta_T} |h_T(\theta) - m_T(\theta)| + K_2 \sup_{\|\theta-\theta_0\|\leq\delta_T} \|\Sigma(\theta,\theta) - \Sigma(\theta_0,\theta_0)\| \\
&\qquad + K_3 \sup_{\|\theta-\theta_0\|\leq\delta_T} \|\Sigma(\theta,\theta_0) - \Sigma(\theta_0,\theta_0)\| .
\end{aligned}$$

Note, however, that $h_T(\theta) - m_T(\theta) = G(\theta) - \Sigma(\theta, \theta_0)\Sigma(\theta_0, \theta_0)^{-1}G(\theta_0)$. Assumptions 5 (iv) and (v) therefore imply equation (16).

As our third step, we linearly approximate m_T using Assumption 5 (ii), which implies that for any $\varepsilon > 0$

$$\begin{aligned}
&\lim_{T\to\infty} \sup_{P\in\mathcal{P}_0} P\left\{ \left| \inf_{\|\theta-\theta_0\|<\delta_T} (g_T(\theta_0) + m_T(\theta))'\Sigma(\theta_0,\theta_0)^{-1}(g_T(\theta_0) + m_T(\theta)) \right.\right. \\
&\left.\left. - \inf_{\|\theta-\theta_0\|<\delta_T} (g_T(\theta_0) + M_T(\theta-\theta_0))'\Sigma(\theta_0,\theta_0)^{-1}(g_T(\theta_0) + M_T(\theta-\theta_0)) \right| > \varepsilon \right\} = 0.
\end{aligned}$$

Indeed, on the set $\mathcal{A}$ we have that $\inf_{\|\theta-\theta_0\|<\delta_T}(g_T(\theta_0) + m(\theta))'\Sigma(\theta_0, \theta_0)^{-1}(g_T(\theta_0) + m(\theta))$ is Lipschitz in m.

So far we have shown that QLR is asymptotically equivalent to

$$QLR_1 = g_T(\theta_0)'\Sigma(\theta_0,\theta_0)^{-1}g_T(\theta_0) - \inf_{\|\theta-\theta_0\|<\delta_T} (g_T(\theta_0)+M_T(\theta-\theta_0))'\Sigma(\theta_0,\theta_0)^{-1}(g_T(\theta_0)+M_T(\theta-\theta_0)),$$

and in particular that $QLR - QLR_1 \to^p 0$ as $T \to \infty$. Note, however, that statistic

$$QLR_2 = g_T(\theta_0)'\Sigma(\theta_0,\theta_0)^{-1}g_T(\theta_0) - \inf_{\theta}(g_T(\theta_0)+M_T(\theta-\theta_0))'\Sigma(\theta_0,\theta_0)^{-1}(g_T(\theta_0)+M_T(\theta-\theta_0))$$

is χ^2_q distributed provided M_T is full rank. The difference between QLR_1 and QLR_2 is

in the area of optimization, and the optimizer in QLR_2 is

$$\theta^* = (M_T'\Sigma(\theta_0,\theta_0)^{-1}M_T)^{-1}M_T'\Sigma(\theta_0,\theta_0)^{-1}g_T(\theta_0) \sim N(0,(M_T'\Sigma(\theta_0,\theta_0)^{-1}M_T)^{-1}).$$

Assumption 5 (iii) guarantees that $\|\theta^*\|/\delta_T$ converges uniformly to zero in probability, and thus that

$$\lim_{T\to\infty}\sup_{P\in\mathcal{P}_0} P\{\|\theta^*-\theta_0\|>\delta_T\}=0.$$

As a result, $QLR_1 - QLR_2 \to^p 0$, which proves that $QLR \Rightarrow \chi^2_q$ uniformly over $\mathcal{P}_0$. The convergence of the conditional critical values is proved in a similar way.□