OVERIDENTIFICATION IN REGULAR MODELS

By

Xiaohong Chen and Andres Santos

April 2015

COWLES FOUNDATION DISCUSSION PAPER NO. 1999

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Overidentification in Regular Models*

Xiaohong Chen†
Department of Economics
Yale University
xiaohong.chen@yale.edu

Andres Santos‡
Department of Economics
UC San Diego
a2santos@ucsd.edu

October, 2014

Abstract

In models defined by unconditional moment restrictions, specification tests are possible and estimators can be ranked in terms of efficiency whenever the number of moment restrictions exceeds the number of parameters. We show that a similar relationship between potential refutability of a model and semiparametric efficiency is present in a much broader class of settings. Formally, we show a condition we name *local overidentification* is required for both specification tests to have power against local alternatives and for the existence of both efficient and inefficient estimators of regular parameters. Our results immediately imply semiparametric conditional moment restriction models are typically locally overidentified, and hence their proper specification is locally testable. We further study nonparametric conditional moment restriction models and obtain a simple characterization of local overidentification in that context. As a result, we are able to determine when nonparametric conditional moment restriction models are locally testable, and when plug-in and two stage estimators of regular parameters are semiparametrically efficient.

Keywords: Overidentification, semiparametric efficiency, specification testing, nonparametric conditional moment restrictions.

*We thank Stephane Bonhomme, Phil Haile, Hide Ichimura, Yuichi Kitamura, Patrick Kline, Peter Phillips, Jim Powell, Richard Smith, and other econometrics seminar participants at Yale for helpful discussions. Any errors are the responsibility of the authors.

†Research supported by Cowles Foundation for Research in Economics.

‡Research supported by NSF Grant SES-1426882.

1 Introduction

Early research on identification recognized the possibility that the distribution P of the observed data might not belong to the set of distributions $\mathbf{P}$ implied by the posited model (Fisher, 1922). Specifications for which this prospect existed were named *observationally restrictive* or *overidentified* by Koopmans and Riersol (1950), who emphasized such models could in principle be refuted by the data. As underscored by Koopmans and Riersol (1950), however, being observationally restrictive is a necessary but not sufficient condition for testability, with examples existing of models that are simultaneously overidentified and untestable (Romano, 2004).

Fortunately, the gap between overidentification and testability proved to be small in the generalized method of moments (GMM) framework. In work originating with Anderson and Rubin (1949) and Sargan (1958), and culminating in Hansen (1982), overidentification in unconditional moment restriction models was equated with the number of moment restrictions exceeding the dimension of the parameter of interest. Under mild regularity conditions, such a surplus of restrictions was shown to enable the construction of both specification tests and more efficient estimators. While perhaps intuitive, the discovery that a simple common condition is instrumental for both specification testing and the availability of more efficient estimators is upon introspection not obvious. Is this close link between efficiency and potential refutability of the GMM model coincidental? Or is it reflective of a deeper principle applicable to a broader class of models?

The need to elucidate the relationship between specification testing, semiparametric efficiency, and overidentification is saliently illustrated by the literature studying nonparametric and semiparametric models. In the latter context, diverse definitions of overidentification exist whose mutual consistency is unclear. Florens et al. (2007), for instance, identifies overidentification with specification testing and states "*... the term overidentification is ill-chosen. If one defines it precisely, one actually obtains the notion of a hypothesis ... This identity between overidentification and hypothesis explains why the term overidentification is associated with the idea of a test*". In contrast, Powell (1994) defines just identification in terms more closely linked to estimation "*... in a nonparametric model, the parameters of interest can be said to be just-identified, in that they are defined by a unique functional of the joint distribution of the data*". Consistent with Powell (1994), Newey and Powell (1999) in turn relate overidentification to efficiency considerations in two stage estimation problems by asserting "*... the efficient estimator for a given first step nonparametric estimator will often be fully efficient, attaining the semiparametric efficiency bound for the model ... Full efficiency occurs because the first step is just identified ...*". While these different definitions of overidentification are concordant in the context of GMM, their compatibility in nonparametric and semiparametric models is to the best of our knowledge unknown. Are these views of overidentification in fact implications of a common condition as in GMM? If such a common condition is indeed available, is it then simple enough to assess when nonparametric conditional moment restriction models are overidentified?[1]

In this paper, we introduce a concept we name *local overidentification* and show that it is in fact responsible for an inherent link between specification testing and semiparametric efficiency analysis that is present in models far beyond the scope of Hansen (1982). The notion of local overidentification arises naturally from the study of the limiting experiment generated by parametric perturbations to a (data) distribution P that belongs to the model $\mathbf{P}$ (LeCam, 1986). As is well understood in the literature on limiting experiments, a fundamental role is played by the *tangent set* $T(P)$, which consists of the scores corresponding to the parametric submodels of $\mathbf{P}$ that contain P (Bickel et al., 1993). Heuristically, $T(P)$ represents the directions from which P may be approached from within the model $\mathbf{P}$. In particular, whenever the closure of $T(P)$ in the mean squared norm equals the set of all possible scores, the model $\mathbf{P}$ is locally consistent with any parametric specification and hence we say P is *locally just identified* by $\mathbf{P}$. In contrast, whenever there exist scores that do not belong to the closure of $T(P)$, the model $\mathbf{P}$ is locally inconsistent with some parametric specification and hence we say P is *locally overidentified* by $\mathbf{P}$. While these definitions can in principle be more generally applied, we focus on models that are regular – in the sense that $T(P)$ is linear[2] – due to the necessity of this condition in semiparametric efficiency analysis (van der Vaart, 1989). When specialized to unconditional moment restriction models, local overidentification is equivalent to the standard GMM overidentification requirement that the number of moment restrictions exceed the number of parameters.

Due to its fundamental role in the limiting experiment, local overidentification is intrinsically related to the local asymptotic behavior of both regular estimators and specification tests. We show, for example, that the local power function of any specification test is also the power function of a test of whether the score of the underlying deviation belongs to the closure of the tangent set $T(P)$. Since we define local just identification as the closure of $T(P)$ equaling the set of all possible scores, it follows that the local power of a specification test cannot exceed its local size whenever P is locally just identified by $\mathbf{P}$ – i.e. proper specification is locally untestable under local just identification. Conversely, whenever P is locally overidentified by $\mathbf{P}$ there exist scores that are uncorrelated with all the scores in $T(P)$, and we show how a specification test whose local power exceeds its local size may be constructed by employing them. Hence, in analogy to Sargan (1958) and Hansen (1982) we conclude that, subject to the availability of a set of scores orthogonal to $T(P)$, proper specification is locally testable if

[1]We thank Stephane Bonhomme for asking the question of whether the nonparametric instrumental variables regression of Newey and Powell (2003) is overidentified.

[2]We stress that a model $\mathbf{P}$ being regular does not imply that all the parameters underlying the model are regular (or root-n estimable, where n is the sample size). In fact, the underlying parameters of a regular model $\mathbf{P}$ may themselves not be identified or not be root-n estimable.

and only if P is locally overidentified by $\mathbf{P}$. While the required set of orthogonal scores is in principle unknown, we show how its estimation is implicitly done by the J-test of Hansen (1982), and more generally devise a method for its estimation by following Hausman (1978).

The connection between local overidentification and the local asymptotic behavior of regular estimators can be readily established by exploiting results from the semiparametric efficiency literature (Bickel et al., 1993). For instance, Newey (1990) shows asymptotically linear and regular estimators of a common Euclidean parameter must share the same influence function whenever the closure of $T(P)$ equals the set of all possible scores. Building on this result, we establish that asymptotically linear and regular estimators of any (possibly infinite dimensional) parameter must be asymptotically equivalent whenever P is locally just identified by $\mathbf{P}$. Conversely, we show that if P is locally overidentified by $\mathbf{P}$, then parameters that admit at least one asymptotically linear and regular estimator may in fact be estimated by multiple asymptotically distinct estimators. Thus, alternative asymptotically linear and regular estimators of a common parameter may only be asymptotically different if and only if P is locally overidentified. It is worth emphasizing this conclusion pertains to any regular parameter and hence applies to simple cases such as means or cumulative distribution functions, as well as to more complex ones like average derivatives (whenever the parameter is regular).

We deduce from the described results that local overidentification is instrumental for both the existence of locally nontrivial specification tests and the semiparametric efficiency analysis of regular estimators. It follows that the equivalence between efficiency and potential refutability of a model found in Hansen (1982) is not coincidental, but rather the reflection of a deeper principle applicable to all regular models. In particular, these results enable us to immediately conclude that semiparametric conditional moment restriction models are typically locally overidentified because they allow for both inefficient and efficient estimators (Ai and Chen, 2003; Chen and Pouzo, 2009). Thus, according to our specification testing results, a locally nontrivial specification test of these models exists and may be constructed by comparing efficient and inefficient estimators of a common parameter as in Hausman (1978).

In order to further illustrate the utility of our results, we characterize local overidentification in nonparametric conditional moment restriction models (Chen and Pouzo, 2012). Heuristically, we establish that local overidentification in such models is equivalent to the existence of a nonconstant transformation of the conditioning variable that is uncorrelated with the span of the derivative of the conditional expectation with respect to the nonparametric parameter. In the leading example of nonparametric instrumental variables regression, local overidentification demands the existence of a non-constant function of the instrument that is uncorrelated with all possible transformations of the endogenous regressor – formally, the joint distribution of the instrument and endogenous regressor must not be complete with respect to the endogenous regressor. Hence, while the nonparametric instrumental variables regression model may be locally overidentified, it follows that local overidentification cannot be determined by simply counting the number of instruments as in Hansen (1982). Moreover, we further conclude from our general results that regular plug-in functionals of a nonparametric instrumental variables regression are not automatically efficient, while regular plug-in functionals of nonparametric conditional means generally are.[3] Analogously, two stage estimation approaches in which the first stage parameter is identified by a conditional moment restriction may be inefficient when P is locally overidentified by the first stage conditional moment restriction model – a situation that may arise when the first stage parameter is semiparametric or a function of endogenous variables. Whenever P is locally just identified by the first stage conditional moment restriction model, however, two stage estimation can indeed be semiparametrically efficient; see Newey and Powell (1999) and Ackerberg et al. (2014).

The remainder of the paper is organized as follows. Section 2 formally defines local overidentification, while Sections 3 and 4 discuss the connections to testing and efficiency respectively. Section 5 applies our results to characterize local overidentification in nonparametric conditional moment restriction models and studies its implications. We briefly conclude in Section 6. All proofs are contained in the Appendix.

2 Local Overidentification

In this section, we introduce basic notation and formally define local overidentification.

2.1 Basic Notation

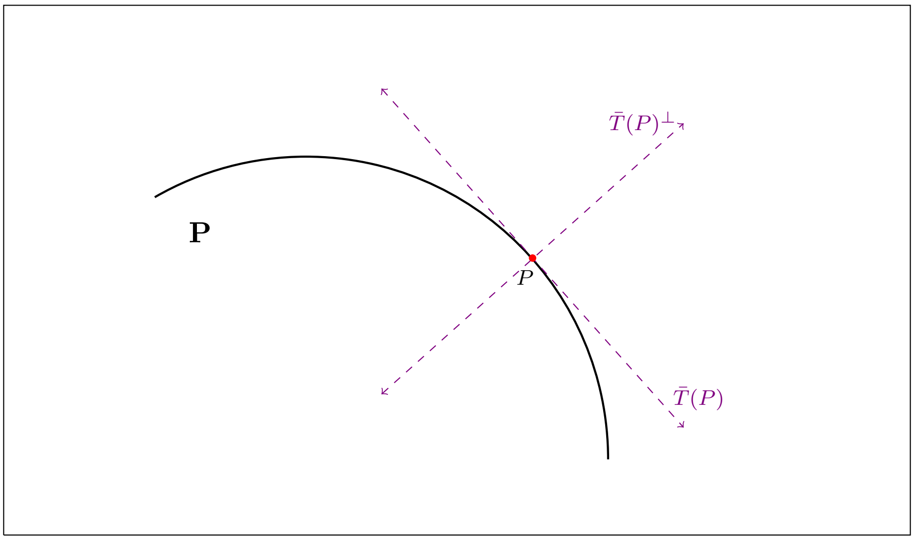
Throughout, we assume the data consists of an i.i.d. sample $\{X_i\}_{i=1}^n$ with each observation $X_i \in \mathbf{R}^{d_x}$ distributed according to $P \in \mathbf{P}$. The set of probability measures $\mathbf{P}$ represents the model under consideration and may be parametric, semiparametric, or fully nonparametric depending on the maintained model assumptions.

Our analysis is local in nature and hence we need to introduce suitable perturbations to the distribution P. Following the literature on limiting experiments (LeCam, 1986), we consider arbitrary smooth parametric likelihoods, which we formally define by:

Definition 2.1. *A "path" $t \mapsto P_{t,g}$ is a function defined on a neighborhood $N \subseteq \mathbf{R}$ of*

[3] Our results in particular imply nonparametric conditional mean and conditional quantiles models are locally just identified.

Figure 1: Model **P** and Tangent Space at P



zero such that $P_{t,g}$ is a probability measure on $\mathbf{R}^{d_x}$ for every $t \in N$, $P_{0,g} = P$, and

$$\lim_{t \to 0} \int [\frac{1}{t}(dP_{t,g}^{1/2} - dP^{1/2}) - \frac{1}{2} g dP^{1/2}]^2 = 0 \ . \tag{1}$$

The function $g : \mathbf{R}^{d_x} \to \mathbf{R}$ is referred to as the "score" of the path $t \mapsto P_{t,g}$. ■

Thus, a path $t \mapsto P_{t,g}$ is simply a parametric model that passes through P and is smooth in the sense of satisfying (1) or, equivalently, of being differentiable in quadratic mean.[4] In our asymptotic analysis, the only relevant characteristic of a path $t \mapsto P_{t,g}$ is its score g, which is why we emphasize its importance through the notation. It is evident from Definition 2.1 that any score g must have mean zero and be square integrable with respect to P. In other words, all scores must belong to the space L_0^2 given by

$$L_0^2 \equiv \{g : \int g dP = 0 \text{ and } \|g\|_{L^2} < \infty\} \qquad \|g\|_{L^2}^2 \equiv \int g^2 dP \ . \tag{2}$$

The implication that $g \in L_0^2$, however, is solely the result of $P_{t,g}$ being restricted to be a probability measure for all t in a neighborhood of zero. If we in addition demand that $P_{t,g}$ belong to the model **P**, then the set of feasible scores reduces to

$$T(P) \equiv \{g \in L_0^2 : \text{ (1) holds for some } t \mapsto P_{t,g} \in \mathbf{P}\} \ , \tag{3}$$

which is often referred to as the *tangent set* at P. Its closure in L_0^2 under the norm $\|\cdot\|_{L^2}$ is in turn termed the *tangent space* and denoted by $\bar{T}(P)$.

[4]The integral should be understood as $\int [\frac{1}{t}((\frac{dP_{t,g}}{d\mu_t})^{1/2} - (\frac{dP}{d\mu_t})^{1/2}) - \frac{1}{2} g (\frac{dP}{d\mu_t})^{1/2}]^2 d\mu_t$ where μ_t is any σ-finite positive measure dominating $(P_t + P)$. The choice of μ_t does not affect the value of the integral.

Whenever $\bar{T}(P)$ is a vector subspace, it is useful to define its orthogonal complement

$$\bar{T}(P)^{\perp} \equiv \{g \in L_0^2 : \int g f dP = 0 \text{ for all } f \in \bar{T}(P)\} \ . \tag{4}$$

Together, the tangent space $\bar{T}(P)$ and its complement $\bar{T}(P)^{\perp}$ form a decomposition of the space of all possible scores L_0^2. Formally, every $g \in L_0^2$ satisfies the equality

$$g = \Pi_T(g) + \Pi_{T^{\perp}}(g) \ , \tag{5}$$

where $\Pi_T(g)$ and $\Pi_{T^{\perp}}(g)$ denote the metric projections under $\|\cdot\|_{L^2}$ onto $\bar{T}(P)$ and $\bar{T}(P)^{\perp}$ respectively. Intuitively, $\Pi_T(g)$ corresponds to the component of a score g that is in accord with the model $\mathbf{P}$, while $\Pi_{T^{\perp}}(g)$ represents the component orthogonal to $\mathbf{P}$; see Bickel et al. (1993). Figure 1 illustrates this standard construction.

As a final piece of notation, for an arbitrary set $\mathbf{A}$ we define the space $\ell^{\infty}(\mathbf{A})$ by

$$\ell^{\infty}(\mathbf{A}) \equiv \{f : \mathbf{A} \to \mathbf{R} \text{ s.t. } \sup_{a \in \mathbf{A}} |f(a)| < \infty\} \ , \tag{6}$$

which is endowed with the norm $\|f\|_{\infty} \equiv \sup_{a \in \mathbf{A}} |f(a)|$ – i.e. $\ell^{\infty}(\mathbf{A})$ is simply the set of bounded functions defined on $\mathbf{A}$. In particular, setting $[d] \equiv \{1, \ldots, d\}$ for any integer d, it follows that $\ell^{\infty}([d])$ denotes the set of bounded sequences with d elements.

2.2 Main Definition

We formalize our discussion so far by imposing the following Assumption:

Assumption 2.1. *(i) $\{X_i\}_{i=1}^n$ is an i.i.d. sequence with $X_i \in \mathbf{R}^{d_x}$ distributed according to $P \in \mathbf{P}$; (ii) $T(P)$ is linear – i.e. if $g, f \in T(P)$, $a, b \in \mathbf{R}$, then $ag + bf \in T(P)$.*

While the i.i.d. requirement in Assumption 2.1(i) is not strictly necessary, we impose it in order to streamline exposition. An extension of our results to certain non i.i.d. models can be accomplished by generalizing our setting to Gaussian shift experiments; see van der Vaart and Wellner (1989). In turn, Assumption 2.1(ii) requires the model $\mathbf{P}$ to be regular at P in the sense that its tangent set $T(P)$ be linear. This requirement is satisfied by numerous models in econometrics, and it is either implicitly or explicitly imposed whenever semiparametric efficient estimators are justified through the convolution theorem (van der Vaart, 1989). Nonetheless, Assumption 2.1(ii) does rule out certain partially identified settings such as missing data problems (Manski, 2003), mixture models (van der Vaart, 1989), and instances in which a parameter is on a boundary (Andrews, 1999). In the latter three cases, the tangent set $T(P)$ is often not linear but a convex cone instead – a setting that enables a partial extension of our results concerning the local testability of the model; see Remark 3.2. However, since linearity of $T(P)$

plays an essential role in the theory of semiparametric efficiency, our results concerning estimation are not directly generalizable to nonlinear tangent sets.

Given the introduced notation, we next formally define local overidentification.

Definition 2.2. *If $L_0^2 = \bar{T}(P)$, then we say P is locally just identified by the model $\mathbf{P}$. Conversely, if $\bar{T}(P) \subsetneq L_0^2$, then we say P is locally overidentified by the model $\mathbf{P}$.* ■

Intuitively, P is locally overidentified by a model $\mathbf{P}$ if $\mathbf{P}$ yields meaningful restrictions on the scores that can be generated by parametric submodels. Conversely, P is locally just identified by the model $\mathbf{P}$ when the sole imposed restriction is that the scores have mean zero and a finite second moment – a quality common to the scores of all paths regardless of whether they belong to the model $\mathbf{P}$ or not. Definition 2.2 is inherently local in that it concerns only the "shape" of $\mathbf{P}$ at the point P rather than $\mathbf{P}$ in its entirety as would be appropriate for a global notion of overidentification (Koopmans and Riersol, 1950); see Remark 2.1. It is also worth emphasizing that local overidentification concerns solely a relationship between the distribution P and the model $\mathbf{P}$. As a result, it is possible for P to be locally overidentified despite underlying parameters of the model being partially identified – an observation that simply reflects the fact that partially identified models may still be refuted by the data (Manski, 2003; Arellano et al., 2012).

Remark 2.1. Koopmans and Riersol (1950) refer to a model $\mathbf{P}$ as being overidentified whenever there exists the possibility that P does not belong to $\mathbf{P}$. This criterion leads to a global definition of overidentification, whereby $\mathbf{P}$ is deemed overidentified if it is a strict subset of the set of all probability measures (on $\mathbf{R}^{d_x}$). Regrettably, as emphasized by Koopmans and Riersol (1950), this global definition is too general to have strong implications on statistical analysis; see for instance Romano (2004) for examples of models $\mathbf{P}$ that are both overidentified and untestable. In contrast, despite also being a population concept, local overidentification is able to provide a stronger connection to the testability of $\mathbf{P}$ and the performance of regular estimators albeit at the cost of conducting a local rather than global analysis. ■

In what follows, we demonstrate the utility of the proposed definition by studying the fundamental role local overidentification plays in both specification testing and semiparametric efficiency analysis. Before proceeding, however, we first illustrate the introduced concepts in the generalized methods of moments framework. We will repeatedly return to this application in order to obtain further intuition for our results.

GMM Illustration. Let $\Theta \subseteq \mathbf{R}^{d_\beta}$ denote the parameter space and $\rho : \mathbf{R}^{d_x} \times \mathbf{R}^{d_\beta} \to \mathbf{R}^{d_\rho}$ be a known moment function with $d_\beta \leq d_\rho$. The model $\mathbf{P}$ then consists of the set

$$\mathbf{P} \equiv \{P : \int \rho(\cdot, \beta) dP = 0 \text{ for some } \beta \in \Theta\} ; \tag{7}$$

i.e. the maintained assumption is that there exists a β that zeroes the moment conditions. Let $\beta(P)$ solve $\int \rho(\cdot, \beta(P))dP = 0$ when $P \in \mathbf{P}$. Assuming ρ is differentiable in β for simplicity, define $\Gamma(P) \equiv \int \nabla_\beta \rho(\cdot, \beta(P))dP$, $\Omega(P) \equiv \int \rho(\cdot, \beta(P))\rho(\cdot, \beta(P))'dP$, and

$$S(P) \equiv [I_{d_\rho} - \Gamma(P)(\Gamma(P)'\Omega(P)^{-1}\Gamma(P))^{-1}\Gamma(P)'\Omega(P)^{-1}] , \tag{8}$$

where I_{d_ρ} denotes a $d_\rho \times d_\rho$ identity matrix, and both $\Gamma(P)$ and $\Omega(P)$ are assumed to have full rank. By direct calculation it is then possible to show that

$$\begin{aligned} \bar{T}(P)^\perp &= \{g \in L_0^2 : g = \lambda' S(P)\rho(\cdot, \beta(P)) \text{ for some } \lambda \in \mathbf{R}^{d_\rho}\} \\ \bar{T}(P) &= \{g \in L_0^2 : \int gf dP = 0 \text{ for all } f \in \bar{T}(P)^\perp\} . \end{aligned} \tag{9}$$

Thus, P is locally overidentified by the model $\mathbf{P}$ if and only if $\bar{T}(P)^\perp \neq \{0\}$, or equivalently if and only if $S(P) \neq 0$ which yields the usual condition $d_\rho > d_\beta$. ■

3 Testing

Under appropriate regularity conditions, the requirement that the number of moments exceed the number of parameters can be shown to be equivalent to the existence of locally nontrivial specification tests in models defined by unconditional moment restrictions. In this section, we show that in regular models an analogous relationship exists between the local overidentification of P and the local testability of the model $\mathbf{P}$.

3.1 Testing Setup

A specification test for a model $\mathbf{P}$ is a test of the null hypothesis that P belongs to $\mathbf{P}$ against the alternative that it does not – i.e. it is a test of the hypotheses

$$H_0 : P \in \mathbf{P} \qquad\qquad H_1 : P \notin \mathbf{P} . \tag{10}$$

We denote an arbitrary (possibly randomized) test of (10) by $\phi_n : \{X_i\}_{i=1}^n \to [0, 1]$, which recall is a function specifying for each realization of the data $\{X_i\}_{i=1}^n$ a corresponding probability of rejecting the null hypothesis.[5] Our interest is in examining the local behavior of such tests along local perturbations to a distribution $P \in \mathbf{P}$. More precisely, we aim to characterize the limiting local power functions of tests ϕ_n when at sample size n each X_i is distributed according to $P_{1/\sqrt{n},g}$ for some path $t \mapsto P_{t,g}$. To this end, however, it is necessary to restrict attention to tests ϕ_n whose limiting local power is well defined. Therefore, for any path $t \mapsto P_{t,g}$ we set $P^n_{1/\sqrt{n},g} \equiv \bigotimes_{i=1}^n P_{1/\sqrt{n},g}$

[5]A non-randomized test is therefore one where ϕ_n only takes values 1 (reject) or 0 (fail to reject).

and define the limiting local power function π of a test ϕ_n to be given by

$$\lim_{n\to\infty} \int \phi_n dP^n_{1/\sqrt{n},g} \equiv \pi(g) \ , \tag{11}$$

where in (11) we have implicitly assumed that ϕ_n is such that the limit indeed exists for any path $t \mapsto P_{t,g}$. The existence of a limiting local power function π is a mild requirement which can be easily verified, for example, when tests are based on comparing statistics to critical values; see Remark 3.1. It is also worth noting that, as emphasized in the notation, any limiting power function π must only depend on the score g and be independent of any other characteristics of the path $t \mapsto P_{t,g}$.[6]

Remark 3.1. Tests ϕ_n are often constructed by comparing a statistic T_n to an estimate $\hat{c}_{1-\alpha}$ of the $1-\alpha$ quantile of its asymptotic distribution. Such tests are of the form

$$\phi_n(\{X_i\}_{i=1}^n) = 1\{T_n > \hat{c}_{1-\alpha}\} \ , \tag{12}$$

and can be shown to satisfy (11) provided (i) $(T_n, \frac{1}{\sqrt{n}}\sum_i g(X_i))$ converges in distribution under P^n for any $g \in L^2_0$, and (ii) the limiting distribution of T_n is continuous. ■

3.2 Limiting Experiment

Intuitively, the limiting local power function π of a level α test ϕ_n of (10) that can control size locally to $P \in \mathbf{P}$ must not exceed α along submodels $t \mapsto P_{t,g} \in \mathbf{P}$ (see (11)). Since by definition the score g of any submodel $t \mapsto P_{t,g} \in \mathbf{P}$ belongs to the tangent set $T(P)$, local size control is therefore tantamount to the requirement that

$$\pi(g) \leq \alpha \text{ for all } g \in T(P) \tag{13}$$

or, equivalently, that $\pi(g) \leq \alpha$ whenever $\Pi_{T^\perp}(g) = 0$. In contrast, a path $t \mapsto P_{t,g}$ that approaches $P \in \mathbf{P}$ from outside the model $\mathbf{P}$ should be such that its score g does not belong to $T(P)$ or, equivalently, we should expect $\Pi_{T^\perp}(g) \neq 0$ – see Figure 2. In short, these heuristics suggest π may be viewed as the power function of a level α test for

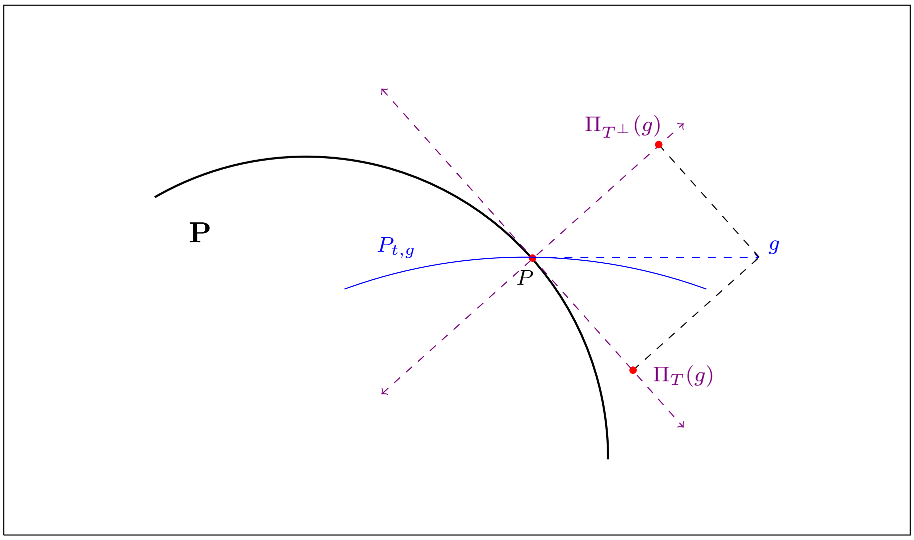
$$H_0 : \Pi_{T^\perp}(g) = 0 \qquad H_1 : \Pi_{T^\perp}(g) \neq 0 \ . \tag{14}$$

In this section, we formalize this discussion by establishing that the limiting local power function π of any test ϕ_n of (10) is also the power function of a test of (14).

The key step necessary to relate π to the hypothesis testing problem in (14) is to embed the latter in a concrete statistical experiment. To this end, let $d_T \equiv \dim\{\bar{T}(P)\}$

[6] This follows from the fact that the product measures of any two local paths that share the same score must converge in the Total Variation metric; see Lemma B.1 in the Appendix.

Figure 2: The Score of a Path outside **P**



and $d_{T^\perp} \equiv \dim\{\bar{T}(P)^\perp\}$ denote the dimensions of the tangent space $\bar{T}(P)$ and its orthogonal complement $\bar{T}(P)^\perp$, and note that d_T and $d_{T^\perp}$ may be infinite. Under Assumption 2.1(ii) both $\bar{T}(P)$ and $\bar{T}(P)^\perp$ are Hilbert spaces and there therefore exist orthonormal bases $\{\psi_k^T\}_{k=1}^{d_T}$ and $\{\psi_k^\perp\}_{k=1}^{d_{T^\perp}}$ for $\bar{T}(P)$ and $\bar{T}(P)^\perp$ respectively. For each $g \in L_0^2$ we then define a probability measure Q_g on $\mathbf{R}^{d_T} \times \mathbf{R}^{d_{T^\perp}}$ to be equal to the law of $(Y, Z) \in \mathbf{R}^{d_T} \times \mathbf{R}^{d_{T^\perp}}$ where $Y \equiv (Y_1, \ldots, Y_{d_T})'$ and $Z \equiv (Z_1, \ldots, Z_{d_{T^\perp}})'$ are such that all $\{Y_k\}_{k=1}^{d_T}$ and $\{Z_k\}_{k=1}^{d_{T^\perp}}$ are mutually independent and distributed according to

$$\begin{aligned} Y_k &\sim N(\int g\psi_k^T dP, 1) \text{ for } 1 \leq k \leq d_T \\ Z_k &\sim N(\int g\psi_k^\perp dP, 1) \text{ for } 1 \leq k \leq d_{T^\perp} \ ; \end{aligned} \tag{15}$$

i.e. Y and Z are possibly infinite dimensional vectors with independent coordinates that are normally distributed with unknown means depending on g and known unit variance.

The local behavior of specification tests can then be understood through the related problem of testing (14) based on a single observation (Y, Z) whose distribution is known to belong to the family $\{Q_g : g \in L_0^2\}$. Formally, Theorem 3.1 establishes that if π is the limiting local power function of a level α test ϕ_n of (10), then there exists a level α test ϕ of (14) based on a single observation (Y, Z) whose power function is also π.

Theorem 3.1. *Let Assumption 2.1 hold and suppose that ϕ_n satisfies* (11) *and*

$$\lim_{n\to\infty} \int \phi_n dP_{1/\sqrt{n},g}^n \leq \alpha \tag{16}$$

for any submodel $t \mapsto P_{t,g} \in \mathbf{P}$. Then there is a level α test $\phi : (Y, Z) \to [0, 1]$ of the *hypothesis in* (14) *based on a single observation* (Y, Z) *such that for any path* $t \mapsto P_{t,g}$

$$\pi(g) \equiv \lim_{n\to\infty} \int \phi_n dP^n_{1/\sqrt{n},g} = \int \phi dQ_g \ . \tag{17}$$

The principal utility of Theorem 3.1 is that it enables us to understand the local asymptotic behavior of specification tests by studying the simpler testing problem in (14). For instance, if optimality results are available for the testing problem in (14), then these can be combined with Theorem 3.1 to obtain local power bounds for specification tests. To this end, it is convenient to note that since $\{\psi_k^\perp\}_{k=1}^{d_{T^\perp}}$ is an orthonormal basis for $\bar{T}(P)^\perp$, the null hypothesis that $\Pi_{T^\perp}(g) = 0$ is equivalent to Z having mean zero under Q_g. Corollary 3.1 illustrates these points, by employing Theorem 3.1 and maximin power bounds for tests on the mean of Z to obtain maximin local power bounds for specification tests when $\bar{T}(P)^\perp$ is finite dimensional.

Corollary 3.1. *Let* $\mathcal{X}^2(B)$ *follow a noncentral chi-squared distribution with* $d_{T^\perp}$ *degrees of freedom and noncentrality parameter* B*, and let* $\chi^2_{1-\alpha}$ *denote the* $1-\alpha$ *quantile of* $\mathcal{X}^2(0)$*. If Assumption 2.1 holds,* ϕ_n *satisfies* (11) *and* (16)*, and* $d_{T^\perp} < \infty$*, then*

$$\limsup_{n\to\infty} \inf_{\{g\in L^2_0: \|\Pi_{T^\perp}(g)\|_{L^2} \geq B\}} \int \phi_n dP^n_{1/\sqrt{n},g} \leq P(\mathcal{X}^2(B) \geq \chi^2_{1-\alpha}) \ . \tag{18}$$

Intuitively, Corollary 3.1 establishes an upper bound on the minimum local power a test may have along paths $t \mapsto P_{t,g}$ which are in a local sense a distance B away from the model $\mathbf{P}$.[7] Thus, if a specification test is shown to attain the bound in (18), then its local power function is maximin. We next return to the generalized methods of moment model to illustrate our results and conclude the J-test is in fact locally maximin.

GMM Illustration (cont). For $\hat{\Omega}(P)^{-1}$ a consistent estimator for $\Omega(P)^{-1}$, it is customary in GMM to conduct a specification test by comparing the J-statistic

$$J_n \equiv \inf_{\beta\in\Theta} (\sqrt{n} \int \rho(\cdot,\beta) dP^n)' \hat{\Omega}(P)^{-1} (\sqrt{n} \int \rho(\cdot,\beta) dP^n) \tag{19}$$

to the $1-\alpha$ quantile of a chi-squared distribution with $(d_\rho - d_\beta)$ degrees of freedom (denoted $\chi^2_{1-\alpha}$). Under standard regularity conditions, the limiting local power function of this specification test exists and for any score $g \in L^2_0$ it is equal to

$$\pi(g) = P(\|\Omega(P)^{-\frac{1}{2}} S(P)(\mathbb{G}_0 + \int \rho(\cdot, \beta(P)) g dP)\|^2 \geq \chi^2_{1-\alpha}) \ , \tag{20}$$

where $\mathbb{G}_0 \sim N(0, \Omega(P))$ and $S(P)$ is as defined in (8). Moreover, it can be shown by direct calculation that $S(P)'\Omega(P)^{-1}S(P) = \Omega(P)^{-\frac{1}{2}} M(P) \Omega(P)^{-\frac{1}{2}}$ for a symmetric

[7]This follows by noting $\|\Pi_{T^\perp}(g)\|_{L^2} = \inf_{f\in T(P)} \|g - f\|_{L^2}$ and interpreting local distance between different paths $t \mapsto P_{t,g}$ as the $\|\cdot\|_{L^2}$ distance of their respective scores.

idempotent matrix $M(P)$ of rank $d_\rho - d_\beta$. Thus, letting $\{\gamma_k\}_{k=1}^{d_\rho - d_\beta}$ denote the eigenvectors of $M(P)$ corresponding to nonzero eigenvalues, it then follows from (20) that

$$\|\Omega(P)^{-\frac{1}{2}}S(P)(\mathbb{G}_0 + \int \rho(\cdot,\beta(P))gdP)\|^2 = \sum_{k=1}^{d_\rho - d_\beta}(\gamma_k'\Omega(P)^{-\frac{1}{2}}\mathbb{G}_0 + \int(\gamma_k'\Omega(P)^{-\frac{1}{2}}\rho(\cdot,\beta(P)))gdP)^2 \ . \quad (21)$$

However, $\{\gamma_k'\Omega(P)^{-\frac{1}{2}}\rho(\cdot,\beta(P))\}_{k=1}^{d_\rho - d_\beta}$ forms an orthonormal basis for $\bar{T}(P)^\perp$ (see (9)) while $\{\gamma_k'\Omega(P)^{-\frac{1}{2}}\mathbb{G}_0\}_{k=1}^{d_\rho - d_\beta}$ are independent standard normal random variables. From (20) and (21) we therefore obtain for $Z \in \mathbf{R}^{d_\rho - d_\beta}$ distributed according to (15) that

$$\pi(g) = P(\|Z\|^2 \geq \chi^2_{1-\alpha}) \ . \quad (22)$$

Hence, in accord to Theorem 3.1, the local power function of a J-test has a dual interpretation as the power function of a Wald test for whether Z has mean zero. In particular, it immediately follows that the J-test attains the optimality bound of Corollary 3.1. Similarly, it is possible to employ Theorem 3.1 to conclude the J-test is optimal among specification tests whose local power functions is invariant in $\|\Pi_{T^\perp}(g)\|_{L^2}$. ■

3.3 Specification Testing

Having characterized the limiting experiment in Theorem 3.1, we can now develop the connection between local overidentification and the local behavior of specification tests. A first immediate conclusion of our analysis is that proper model specification is locally untestable whenever P is locally just identified by the model $\mathbf{P}$.

Corollary 3.2. *Let Assumption 2.1 hold and suppose ϕ_n satisfies* (11) *and* (16). *If P is locally just identified by the model $\mathbf{P}$, then for any path $t \mapsto P_{t,g}$ it follows that*

$$\lim_{n\to\infty} \int \phi_n dP^n_{1/\sqrt{n},g} \leq \alpha \ . \quad (23)$$

Corollary 3.2 establishes that if P is locally just identified, then the local power of a specification test that can locally control size (as in (16)) cannot exceed its level. Intuitively, whenever P is locally just identified by $\mathbf{P}$, the set of scores $T(P)$ which correspond to paths $t \mapsto P_{t,g} \in \mathbf{P}$ is dense in the set of all possible scores and as a result every path is locally on the "boundary" of the null hypothesis; see also Romano (2004) for a nonlocal analogue. The result is straightforward to derive from Theorem 3.1 by noting that under local just identification $\bar{T}(P) = L^2_0$ and $\bar{T}(P)^\perp = \{0\}$, which implies $\Pi_{T^\perp}(g) = 0$ for all possible scores $g \in L^2_0$. Therefore, the null hypothesis in (14) is satisfied for all paths $t \mapsto P_{t,g}$ regardless of whether they belong to $\mathbf{P}$ or not, and thus by Theorem 3.1 the limiting local power of a test ϕ_n cannot exceed α.

In order to complete the analogy to the generalized methods of moments setting, it remains to be shown that the converse to Corollary 3.2 also holds – namely, that if P is locally overidentified, then there exists a specification test with nontrivial local power. We will establish the desired converse under the following high level assumption:

Assumption 3.1. *There is a known* $\mathcal{F} \equiv \{f_k\}_{k=1}^{d_F} \subseteq \bar{T}(P)^{\perp}$ *with* $\sum_{k=1}^{d_F} \int f_k^2 dP < \infty$.

Assumption 3.1 is untenable because the availability of a subset $\mathcal{F}$ of the orthogonal complement $\bar{T}(P)^{\perp}$ in principle requires knowledge of the tangent set $T(P)$ and thus of P itself. Nonetheless, under model specific regularity conditions, suitable estimators $\hat{\mathcal{F}}$ for appropriate subsets $\mathcal{F}$ of $\bar{T}(P)^{\perp}$ are often available; see the discussion of Hansen (1982) below and of Hausman (1978) in Section 4.3. At present, we therefore abstract from the construction of such estimators $\hat{\mathcal{F}}$ and directly impose Assumption 3.1 so as to ease exposition of the connection between specification testing and local overidentification. We also note that in Assumption 3.1 the dimension of $\mathcal{F}$ may be infinite ($d_F = \infty$). In this case the condition $\sum_{k=1}^{d_F} \int f_k^2 dP < \infty$ implicitly requires the variance of $f_k(X_i)$ to decrease to zero sufficiently fast with k.

Given the presumed availability of the set $\mathcal{F}$, we may define the vector of means

$$\mathbb{G}_n \equiv (\frac{1}{\sqrt{n}} \sum_{i=1}^{n} f_1(X_i), \ldots, \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f_{d_F}(X_i))' \tag{24}$$

which belong to $\mathbf{R}^{d_F}$ and is potentially infinite dimensional ($d_F = \infty$). Assumption 3.1, however, ensures that even when d_F is infinite the vector $\mathbb{G}_n$ belongs to the space $\ell^\infty([d_F])$ P^n-almost-surely, where recall $[d_F] \equiv \{1, \ldots, d_F\}$ and $\ell^\infty([d_F])$ denotes the set of bounded functions on $[d_F]$.[8] Moreover, since each $f \in \mathcal{F}$ is such that $E[f(X_i)] = 0$ when X_i is distributed according to P, the vector $\mathbb{G}_n$ is properly centered and it can therefore be expected to converge to a centered Gaussian measure under P^n. In contrast, if $\{X_i\}_{i=1}^n$ is distributed according to $P^n_{1/\sqrt{n},g}$ for a path $t \mapsto P_{t,g}$ approaching P from outside $\mathbf{P}$, then $\mathbb{G}_n$ should not be properly centered and instead converge to a non-centered Gaussian measure. The following Lemma formalizes these heuristics and provides the foundation for a specification test with nontrivial local power.

Lemma 3.1. *Let Assumptions 2.1, 3.1 hold, for any path* $t \mapsto P_{t,g}$ *let* $L_{n,g}$ *denote the law of* $\{X_i\}_{i=1}^n$ *under* $\bigotimes_{i=1}^n P_{1/\sqrt{n},g}$, *and define* $\Delta_g \equiv \{\int f_k g dP\}_{k=1}^{d_F}$. *Then, there is a tight centered Gaussian process* $\mathbb{G}_0 \in \ell^\infty([d_F])$ *such that for any path* $t \mapsto P_{t,g}$:

$$\mathbb{G}_n \stackrel{L_{n,g}}{\to} \mathbb{G}_0 + \Delta_g \ . \tag{25}$$

[8] The space $\ell^\infty([d_F])$ may be identified with $\mathbf{R}^{d_F}$ when $d_F < \infty$ and with $\ell^\infty(\mathbb{N})$ (the space of bounded sequences) when $d_F = \infty$.

Moreover, (i) $\Delta_g = 0$ (in $\ell^\infty([d_F])$) whenever $t \mapsto P_{t,g} \in \mathbf{P}$, and (ii) If in addition cl{lin{$\mathcal{F}$}} $= \bar{T}(P)^\perp$, then it also follows that $\Delta_g \neq 0$ (in $\ell^\infty([d_F])$) whenever

$$\liminf_{n\to\infty} \inf_{Q\in\mathbf{P}} n \int [dQ^{1/2} - dP^{1/2}_{1/\sqrt{n},g}]^2 > 0 \ . \tag{26}$$

Lemma 3.1 corroborates the presence of a shift $\Delta_g \equiv \{\int f_k g dP\}_{k=1}^{d_F}$ in the asymptotic distribution of $\mathbb{G}_n$ whenever $\{X_i\}_{i=1}^n$ is distributed according to $P^n_{1/\sqrt{n},g}$ in place of P^n. The explicit formulation of Δ_g as a vector of covariances between the functions $f \in \mathcal{F}$ and the score g of the path $t \mapsto P_{t,g}$ is particularly important for testing purposes. On the one hand, if the path $t \mapsto P_{t,g}$ approaches P from within the model $\mathbf{P}$, then $g \in T(P)$ by definition and thus $\Delta_g = 0$ as a result of $\mathcal{F} \subset \bar{T}(P)^\perp$. On the other hand, if a path $t \mapsto P_{t,g}$ does not approach P "too fast" from outside $\mathbf{P}$ (see (26)) and $\mathcal{F}$ is sufficiently rich in the sense that the closure of its linear span (cl{lin{$\mathcal{F}$}}) equals $\bar{T}(P)^\perp$, then g must be correlated with some $f \in \mathcal{F}$ implying $\Delta_g \neq 0$. Intuitively, the vector of sample moments $\mathbb{G}_n$ can therefore be employed to detect whether a local perturbation approaches P from within the model $\mathbf{P}$ or from outside of it.

By equating the null hypothesis in (14) to $\mathbb{G}_n$ having mean zero, Lemma 3.1 suggests a specification test with nontrivial asymptotic local power may be built by testing whether $\mathbb{G}_n$ has mean zero. Following this intuition, we construct a locally nontrivial specification test by utilizing a function $\Psi : \ell^\infty([d_F]) \to \mathbf{R}_+$ to reduce the vector of sample means $\mathbb{G}_n$ to a scalar test statistic $\Psi(\mathbb{G}_n)$ – for instance, we may set $\Psi(\mathbb{G}_n) = \|\mathbb{G}_n\|_\infty$ when $d_F = \infty$. In general, we require the function Ψ to satisfy the following condition:

Assumption 3.2. *(i) $\Psi : \ell^\infty([d_F]) \to \mathbf{R}_+$ is continuous, convex; (ii) $\Psi(0) = 0$, $\Psi(b) = \Psi(-b)$ for all $b \in \ell^\infty([d_F])$; (iii) $\{b \in \ell^\infty([d_F]) : \Psi(b) \leq t\}$ is bounded for all $t > 0$.*

Given Assumption 3.2, the following Theorem shows that a specification test based on rejecting whenever the statistic $\Psi(\mathbb{G}_n)$ is large indeed has nontrivial local power.

Theorem 3.2. *Let Assumptions 2.1, 3.1, 3.2 hold, and for $\alpha \in (0, 1)$ let $c_{1-\alpha}$ denote the $1-\alpha$ quantile of $\Psi(\mathbb{G}_0)$. If $c_{1-\alpha} > 0$, then for any path $t \mapsto P_{t,g} \in \mathbf{P}$ it follows that*

$$\lim_{n\to\infty} P^n_{1/\sqrt{n},g}(\Psi(\mathbb{G}_n) > c_{1-\alpha}) = \alpha \ . \tag{27}$$

Moreover, if in addition cl{lin{$\mathcal{F}$}} $= \bar{T}(P)^\perp$ and a path $t \mapsto P_{t,g}$ satisfies (26), then

$$\liminf_{n\to\infty} P^n_{1/\sqrt{n},g}(\Psi(\mathbb{G}_n) > c_{1-\alpha}) \geq P(\Psi(\mathbb{G}_0 + \Delta_g) > c_{1-\alpha}) > \alpha \ . \tag{28}$$

It follows from Lemma 3.1 that the asymptotic distribution of $\Psi(\mathbb{G}_n)$ equals $\Psi(\mathbb{G}_0)$ whenever $\{X_i\}_{i=1}^n$ is distributed according to $P^n_{1/\sqrt{n},g}$ for any path $t \mapsto P_{t,g} \in \mathbf{P}$. Therefore, letting $c_{1-\alpha}$ denote the $1-\alpha$ quantile of $\Psi(\mathbb{G}_0)$, it is straightforward to establish the first claim of Theorem 3.2 that a test that rejects whenever $\Psi(\mathbb{G}_n)$ is larger than $c_{1-\alpha}$ locally controls size (see (27)). Establishing that such a test also has nontrivial local power (see (28)) is substantially more challenging and follows from a strengthening of Anderson's inequality due to Lewandowski et al. (1995). Finally, we note that while linearity of $T(P)$ plays a crucial in the study of regular estimators (van der Vaart, 1989), the present section's results concerning local testability of $\mathbf{P}$ can in fact be partially extended to settings in which $T(P)$ is not linear; see Remark 3.2.

Remark 3.2. In an important class of irregular models, $\bar{T}(P)$ is not a vector space but a convex cone instead. In such a setting it is convenient to define the polar cone

$$\bar{T}(P)^{-} \equiv \{f \in L_0^2 : \int fgdP \leq 0 \text{ for all } g \in \bar{T}(P)\} \ , \tag{29}$$

which equals $\bar{T}(P)^{\perp}$ when $\bar{T}(P)$ is a vector space. Employing Moreau's decomposition (Moreau, 1962) it is then possible to generalize Lemma 3.1 provided the condition $\mathcal{F} \subset \bar{T}(P)^{\perp}$ is replaced by $\mathcal{F} \subset \bar{T}(P)^{-}$. In particular, result (25) continues to hold, but with $\Delta_g \leq 0$ whenever $t \mapsto P_{t,g} \in \mathbf{P}$, and $\Delta_g \nleq 0$ whenever the path $t \mapsto P_{t,g}$ satisfies (26) and the convex cone generated by $\mathcal{F}$ in L_0^2 equals $\bar{T}(P)^{-}$. This extension of Lemma 3.1 suggests an analogue to Theorem 3.2 can be established by employing test statistics $\Psi(\mathbb{G}_n)$ of the null hypothesis that the mean $\mathbb{G}_n$ is negative (instead of zero). However, while given the availability of a suitable class $\mathcal{F}$ it is straightforward to construct a test that locally controls size (as in (27)), the resulting test will have power greater than size against some but not all local alternatives (as in (28)). The latter weaker result is due to the nonexistence of nontrivial unbiased tests in these problems (Lehmann, 1952). ■

Together, Corollary 3.2 and Theorem 3.2 complete the analogy to the generalized method of moments setting. Namely, they imply that there exists a locally nontrivial specification test if and only if P is locally overidentified by the model $\mathbf{P}$. The specification test in Theorem 3.2 is infeasible insofar as it requires knowledge of a set of functions $\mathcal{F}$ satisfying Assumption 3.1 and whose linear span is dense in the orthogonal complement $\bar{T}(P)^{\perp}$. These requirement can be dispensed with under additional regularity conditions, as we next illustrate by showing that the J-test of Hansen (1982) is in fact asymptotically equivalent to the test developed in Theorem 3.2.

GMM Illustration (cont). For the J-statistic J_n and the matrix $S(P)$ as defined in (19) and (8) respectively, it is straightforward to show that

$$J_n = \|\sqrt{n} \int \Omega(P)^{-\frac{1}{2}} S(P) \rho(\cdot, \beta(P)) dP^n \|^2 + o_p(1) \ . \tag{30}$$

Hence, for $e_k \in \mathbf{R}^{d_\rho}$ a vector whose k^{th} coordinate is one and all other coordinates are zero, and the set $\mathcal{F} \equiv \{e_k'\Omega(P)^{-\frac{1}{2}}S(P)\rho(\cdot,\beta(P))\}_{k=1}^{d_\rho}$, we obtain from (30) that

$$J_n = \Psi(\mathbb{G}_n) + o_p(1) \ , \tag{31}$$

where $\Psi(b) = \|b\|^2$ for any $b \in \ell^\infty([d_\rho])$. Moreover, by (9) we can also conclude that $\mathcal{F} \subset \bar{T}(P)^\perp$ and $\mathrm{cl}\{\mathrm{lin}\{\mathcal{F}\}\} = \bar{T}(P)^\perp$. Thus, the J-test is asymptotically equivalent to a version of the infeasible test developed in Theorem 3.2. Intuitively, J_n may be interpreted as employing an estimate $\hat{\mathcal{F}}$ of the unknown class $\mathcal{F}$ instead of $\mathcal{F}$ itself. ■

4 Estimation

In Section 3 we argued that local overidentification plays a fundamental role in determining whether a model is locally testable. In what follows, we show that local overidentification is also essential in determining whether regular parameters admit asymptotically distinct linear and regular estimators. We therefore conclude that, subject to regularity conditions, local overidentification is equivalent to both the local testability of the model and the existence of efficient (and inefficient) estimators. Hence, in accord with Hansen (1982), our results imply semiparametric efficiency considerations are only of importance when the model is locally testable.

4.1 Estimation Setup

We represent a parameter as the value a known mapping $\theta : \mathbf{P} \to \mathbf{B}$ takes at the unknown distribution $P \in \mathbf{P}$. In the construction of Hausman tests, it will be particularly useful to allow for $\theta(P)$ to be infinite dimensional and we therefore only require $\mathbf{B}$ to be a Banach space with norm $\|\cdot\|_\mathbf{B}$. The dual space of $\mathbf{B}$ is denoted by $\mathbf{B}^*$ and defined as

$$\mathbf{B}^* \equiv \{b^* : \mathbf{B} \to \mathbf{R} : b^* \text{ is linear and } \|b^*\|_{\mathbf{B}^*} < \infty\} \qquad \|b^*\|_{\mathbf{B}^*} \equiv \sup_{\|b\|_\mathbf{B} \leq 1} |b^*(b)|; \tag{32}$$

i.e. the dual space $\mathbf{B}^*$ is the set of continuous linear functionals operating on $\mathbf{B}$.

An estimator $\hat{\theta}_n : \{X_i\}_{i=1}^n \to \mathbf{B}$ for the parameter $\theta(P)$ is then simply a function mapping the data into the space $\mathbf{B}$ where $\theta(P)$ belongs. It is evident that given a consistent estimator $\hat{\theta}_n$ of $\theta(P)$ it is always possible to construct an alternative consistent estimator – for instance, by considering $\hat{\theta}_n + b/\sqrt{n}$ for any $b \in \mathbf{B}$. Addressing the question of whether $\theta(P)$ admits a "unique" estimator therefore requires us to in some manner constrain the class of estimators under consideration. We accomplish this goal by focusing attention on estimators that are both *regular* and *asymptotically linear*. In the present setting, regularity and asymptotic linearity are defined as follows:

Definition 4.1. *$\hat{\theta}_n : \{X_i\}_{i=1}^n \to \mathbf{B}$ is a regular estimator of $\theta(P)$ if there is a tight $\mathbb{Z}$ such that $\sqrt{n}\{\hat{\theta}_n - \theta(P_{1/\sqrt{n},g})\} \stackrel{L_{n,g}}{\to} \mathbb{Z}$ for any $t \mapsto P_{t,g} \in \mathbf{P}$ and $L_{n,g} \equiv \bigotimes_{i=1}^n P_{1/\sqrt{n},g}$.* ■

Definition 4.2. *$\hat{\theta}_n : \{X_i\}_{i=1}^n \to \mathbf{B}$ is an asymptotically linear estimator of $\theta(P)$ if*

$$\sqrt{n}\{\hat{\theta}_n - \theta(P)\} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \nu(X_i) + o_p(1) \tag{33}$$

under $P^n \equiv \bigotimes_{i=1}^n P$ for some $\nu : \mathbf{R}^{d_x} \to \mathbf{B}$ satisfying $b^*(\nu) \in L_0^2$ for any $b^* \in \mathbf{B}^*$. ■

By restricting attention to regular estimators, as in Definition 4.1, we focus on $\sqrt{n}$-consistent estimators whose asymptotic distribution is invariant to local perturbations of the data generating process. We note that the perturbations we consider are only under paths within the model $\mathbf{P}$ since the parameter $\theta(P_{1/\sqrt{n},g})$ may not be defined when $P_{1/\sqrt{n},g} \notin \mathbf{P}$; see the GMM discussion below. Asymptotic linearity, as in Definition 4.2, in turn imposes the existence of an influence function ν for which (33) holds in $\mathbf{B}$. The condition that $b^*(\nu) \in L_0^2$ for any element $b^* \in \mathbf{B}^*$ implies $\sqrt{n}\{b^*(\hat{\theta}_n) - b^*(\theta(P))\}$ converges to a centered normal distribution on $\mathbf{R}$ under P^n. Thus, any estimator $\hat{\theta}_n$ satisfying (33) that converges in distribution in $\mathbf{B}$ must do so to a centered Gaussian measure. In particular, we note that this requirement implies that if $\hat{\theta}_n$ is asymptotically linear and regular, then $\hat{\theta}_n + b/\sqrt{n}$ for $0 \neq b \in \mathbf{B}$ cannot be.

We illustrate these concepts in the generalized methods of moments context.

GMM Illustration (cont). The canonical example for a parameter $\theta(P)$ in this setting is $\beta(P)$ – i.e. the element $\beta \in \Theta$ solving $\int \rho(\cdot, \beta(P))dP = 0$. In this case $\mathbf{B} = \mathbf{R}^{d_\beta}$ and $\theta(P) = \beta(P)$ is clearly defined for all $P \in \mathbf{P}$ but not for $P \notin \mathbf{P}$. The generalized methods of moments estimator of Hansen (1982) is then both regular and asymptotically linear under standard conditions. For an example of an infinite dimensional parameter, we may let $\mathbf{B} = \ell^\infty(\mathbf{R}^{d_x})$ and consider estimating the c.d.f. of P so that

$$\theta(P) \equiv t \mapsto P(X_i \leq t) \ . \tag{34}$$

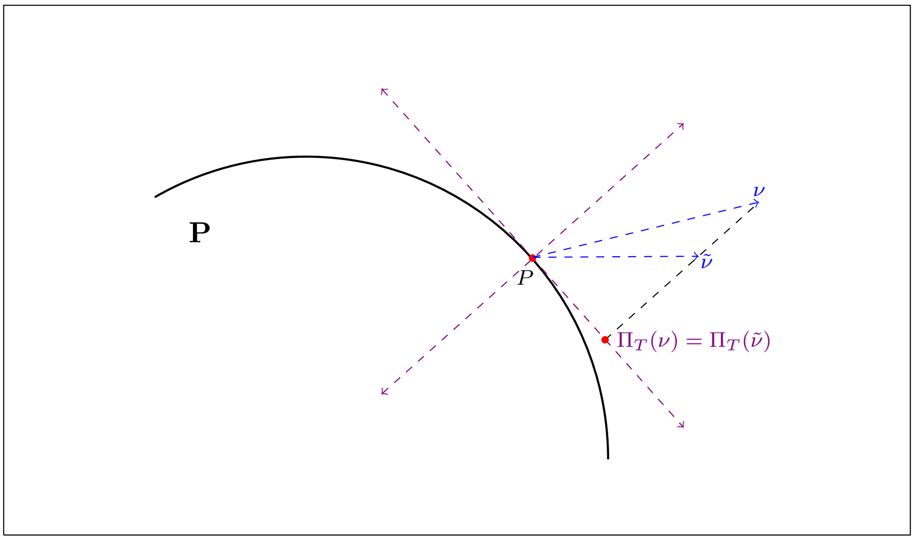
Regular and asymptotically linear estimators for this parameter include the empirical c.d.f. and the empirical likelihood estimator studied in Yuan et al. (2014). ■

4.2 Multiplicity of Estimators

Given the introduced notation, we next examine the relationship between local overidentification of P and estimation of a parameter $\theta(P)$ that admits at least one asymptotically linear regular estimator. To formalize our discussion, we first impose:

Assumption 4.1. *(i) $\theta : \mathbf{P} \to \mathbf{B}$ is a known map, and $\mathbf{B}$ a Banach space $\mathbf{B}$ with norm $\|\cdot\|_{\mathbf{B}}$; (ii) There is an asymptotically linear regular estimator $\hat{\theta}_n : \{X_i\}_{i=1}^n \to \mathbf{B}$ of $\theta(P)$.*

Figure 3: The Projection of Influence Functions



While most commonly employed estimators are regular and asymptotically linear, it is worth noting that their existence imposes restrictions on the map $\theta : \mathbf{P} \to \mathbf{B}$. In particular, as shown by van der Vaart (1991b), the existence of an estimator $\hat{\theta}_n$ satisfying Assumption 4.1(ii) and the linearity of $T(P)$ imposed in Assumption 2.1(ii) together imply that the map $\theta : \mathbf{P} \to \mathbf{B}$ must be pathwise differentiable relative to $T(P)$. We emphasize, however, that the existence of a parameter $\theta(P)$ and an estimator $\hat{\theta}_n$ satisfying Assumption 4.1 imposes restrictions only on the map $\theta : \mathbf{P} \to \mathbf{B}$ but not directly on the model $\mathbf{P}$. Moreover, we also note that our results apply to any parameter $\theta(P)$ for which Assumption 4.1 holds. Thus, $\theta(P)$ should not be solely thought of as an intrinsic characteristic of the model $\mathbf{P}$, but rather as any "smooth" function of $P \in \mathbf{P}$. In particular, we emphasize that under Assumption 2.1 one can always find a map $\theta : \mathbf{P} \to \mathbf{B}$ and an estimator $\hat{\theta}_n$ for which Assumption 4.1 holds[9] – a point that is particularly useful in the construction of Hausman tests and the analysis of nonparametric conditional moment restriction models; see Sections 4.3 and 5.2 respectively.

Under Assumption 4.1, the following theorem establishes the connection between estimation and local overidentification by showing asymptotically linear regular estimators are up to first order unique if and only if P is locally just identified by $\mathbf{P}$.

Theorem 4.1. *Let Assumptions 2.1 and 4.1 hold. It then follows that:*

(i) If P is locally just identified by the model $\mathbf{P}$ and $\tilde{\theta}_n : \{X_i\}_{i=1}^n \to \mathbf{B}$ is an asymptotically linear regular estimator of $\theta(P)$, then $\sqrt{n}\{\hat{\theta}_n - \tilde{\theta}_n\} = o_p(1)$ in $\mathbf{B}$.

[9] For example, the parameter $\theta(P) = \int f dP$ for a bounded function f always admits the sample mean $\frac{1}{n}\sum_i f(X_i)$ as an asymptotically linear and regular estimator

(ii) If P is locally overidentified by the model $\mathbf{P}$, then there exists an asymptotically linear regular estimator $\tilde{\theta}_n : \{X_i\}_{i=1}^n \to \mathbf{B}$ such that $\sqrt{n}\{\hat{\theta}_n - \tilde{\theta}_n\} \stackrel{L}{\to} \Delta \neq 0$ in $\mathbf{B}$.

The heuristics behind Theorem 4.1 can most easily be understood in the special case where $\theta(P)$ is scalar valued. When $\mathbf{B} = \mathbf{R}$ the influence functions ν and $\tilde{\nu}$ of alternative asymptotically linear regular estimators $\hat{\theta}_n$ and $\tilde{\theta}_n$ are both elements of L_0^2. As such, both ν and $\tilde{\nu}$ may be projected onto the tangent space $\bar{T}(P)$ and, crucially, by regularity their projections must agree; see Figure 3.[10] Theorem 4.1(i) immediately follows from this observation for if P is locally just identified, then $\bar{T}(P) = L_0^2$ and thus the influence functions of $\hat{\theta}_n$ and $\tilde{\theta}_n$ must in fact coincide implying the estimators are first order equivalent.[11] Analogously, if P is locally overidentified so that $\bar{T}(P) \subsetneq L_0^2$, then Theorem 4.1(ii) may be established by constructing a regular asymptotically linear estimator $\tilde{\theta}_n$ whose influence function differs from that of $\hat{\theta}_n$ on the orthogonal complement $\bar{T}(P)^{\perp}$. The generalization of these heuristics to an infinite dimensional Banach space $\mathbf{B}$ can in turn be accomplished by employing the dual space $\mathbf{B}^*$ and exploiting the asymptotic tightness of $\hat{\theta}_n$ to reduce the analysis to the scalar case.

The proof of Theorem 4.1 is closely related to standard arguments found in the semiparametric efficiency literature (Bickel et al., 1993). In particular, when $\theta(P)$ is scalar valued efficiency is also naturally studied through $\bar{T}(P)$ and the decomposition

$$\nu = \Pi_T(\nu) + \Pi_{T^{\perp}}(\nu) \ , \tag{35}$$

where $\Pi_T(\nu)$ may be understood as the efficient component and $\Pi_{T^{\perp}}(\nu)$ as a "noise" factor extraneous to the model $\mathbf{P}$. Whenever P is locally just identified all asymptotically linear regular estimators lack a noise factor ($\Pi_{T^{\perp}}(\nu) = 0$) and are thus not only equivalent but also efficient. Intuitively, efficiency gains are only possible when P is locally overidentified and there thus exists model information to be exploited in estimation. In contrast, whenever P is locally overidentified asymptotically linear regular estimators may differ and be inefficient by possessing a "noise" component ($\Pi_{T^{\perp}}(\nu) \neq 0$) – the estimator $\tilde{\theta}_n$ in Theorem 4.1 is in fact constructed precisely in this manner. This intrinsic relationship between efficiency and local overidentification can be exploited to characterize local overidentification in models $\mathbf{P}$ for which semiparametric efficiency results are already available in the literature; see Remak 4.1.

Remark 4.1. Whenever P is locally just identified, Theorem 4.1(i) implies that for any function $f \in L^2 \equiv \{h : \mathbf{R}^{d_x} \to \mathbf{R} : \|h\|_{L^2} < \infty\}$, the mean parameter $\theta_f(P) \equiv \int f dP$

[10]Formally, both $\Pi_T(\nu)$ and $\Pi_T(\tilde{\nu})$ must equal the Riesz representor of the pathwise derivative $\dot{\theta} : \bar{T}(P) \to \mathbf{R}$. See, e.g., Proposition 3.3.1 in Bickel et al. (1993) for a statement and proof.

[11]The fact that if $\bar{T}(P) = L_0^2$ then the influence functions must be unique had been previously noted by Newey (1990) (p.106) for the case of $\mathbf{B} = \mathbf{R}$ and Newey (1994) (Theorem 2.1) for the case of $\mathbf{B} = \mathbf{R}^d$ for a fixed finite $d < \infty$.

can be efficiently estimated by its sample analogue:

$$\hat{\theta}_{f,n} \equiv \frac{1}{n}\sum_{i=1}^{n} f(X_i) \ . \tag{36}$$

It is useful to note that the converse to this statement is also true. Namely, if $\hat{\theta}_{f,n}$ is an efficient estimator of $\theta_f(P)$ for all $f \in L^2$, then P must be locally just identified. In fact, for P to be locally just identified it suffices that $\hat{\theta}_{f,n}$ be efficient for all $f \in \mathcal{D}$ for any dense subset $\mathcal{D}$ of L^2.[12] We will exploit this relationship in Section 5 to derive necessary and sufficient conditions for P to be locally just identified in nonparametric conditional moment restriction models. ■

In the context of GMM, Theorem 4.1 is easily illustrated through well known results.

GMM Illustration (cont). Let $\hat{W}_n$ be a sequence of $d_\rho \times d_\rho$ matrices converging in probability under P^n to a positive definite matrix W, and define the estimator

$$\hat{\beta}_n^W \in \arg\min_{\beta\in\Theta} \ (\sqrt{n}\int \rho(\cdot,\beta)dP^n)'\hat{W}_n(\sqrt{n}\int \rho(\cdot,\beta)dP^n) \ , \tag{37}$$

which is both regular and asymptotically linear under standard conditions. Setting $\theta(P) \equiv \beta(P)$ we then recover from Theorem 4.1(i) the well known result that the asymptotic distribution of $\hat{\beta}_n^W$ does not depend on W when P is just identified ($d_\rho = d_\beta$). Theorem 4.1, however, further implies the conclusion is true for all parameters $\theta : \mathbf{P} \to \mathbf{B}$ satisfying Assumption 4.1 – for example, for $\theta(P)$ the c.d.f. of X_i as in (34). ■

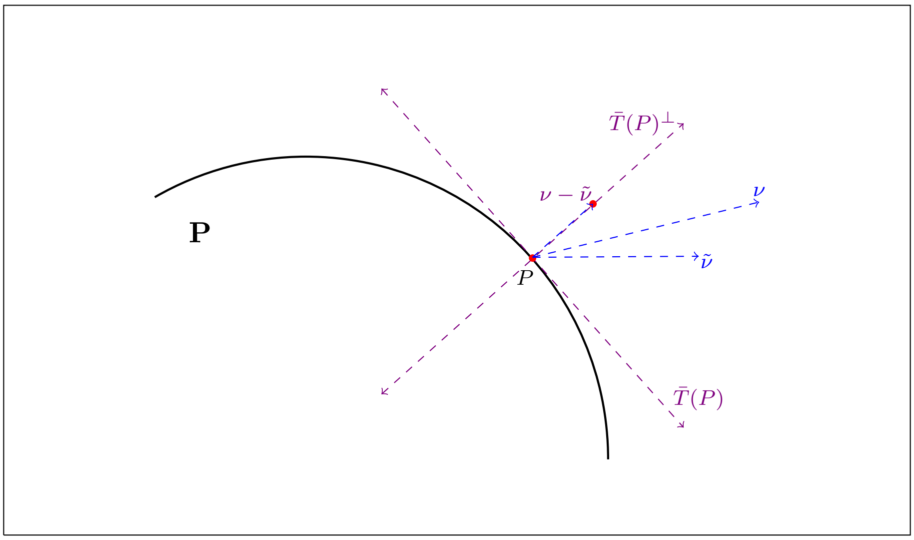
4.3 Hausman Tests

Our results so far have established the equivalence of local overidentification and the existence of both locally nontrivial specification tests and of asymptotically different estimators for a common regular parameter. The latter two concepts were also intrinsically linked by the seminal work of Hausman (1978), who proposed conducting specification tests through the comparison of alternative estimators of a common parameter. In what follows, we revisit Hausman (1978) and show how the principles devised therein can be employed to implement the infeasible specification test of Theorem 3.2.

The construction of a Hausman test requires the existence of two asymptotically linear regular estimators $\hat{\theta}_n$ and $\tilde{\theta}_n$ for a parameter $\theta(P) \in \mathbf{B}$, and we therefore impose:

Assumption 4.2. *(i) There are two asymptotically linear regular estimators $\hat{\theta}_n$ and $\tilde{\theta}_n$ of the parameter $\theta(P) \in \mathbf{B}$ with influence functions ν and $\tilde{\nu}$ respectively; (ii) Under $P^n \equiv \bigotimes_{i=1}^n P$, $(\sqrt{n}\{\hat{\theta}_n - \theta(P)\}, \sqrt{n}\{\tilde{\theta}_n - \theta(P)\})$ converges in distribution on $\mathbf{B} \times \mathbf{B}$.*

Figure 4: The Difference of Influence Functions



The connection between a Hausman test and the results on specification testing from Section 3.3 is most easily illustrated when the parameter $\theta(P)$ is a scalar ($\mathbf{B} = \mathbf{R}$). In such a setting, the influence functions ν and $\tilde{\nu}$ of $\hat{\theta}_n$ and $\tilde{\theta}_n$ both belong to L_0^2 and in addition their projections onto the tangent space $\bar{T}(P)$ coincide ($\Pi_T(\nu) = \Pi_T(\tilde{\nu})$); recall Figure 3. The latter crucial property, however, is equivalent to the difference of the influence functions $(\nu - \tilde{\nu})$ belonging to the orthogonal complement $\bar{T}(P)^{\perp}$; see Figure 4. Moreover, by asymptotic linearity of $\hat{\theta}_n$ and $\tilde{\theta}_n$ it also follows that

$$\sqrt{n}\{\hat{\theta}_n - \tilde{\theta}_n\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\nu(X_i) - \tilde{\nu}(X_i)\} + o_p(1) \tag{38}$$

under P^n, and hence $\sqrt{n}\{\hat{\theta}_n - \tilde{\theta}_n\}$ is asymptotically equivalent to the mean of a function $f \equiv (\nu - \tilde{\nu})$ in the orthogonal complement $\bar{T}(P)^{\perp}$ (compare to (24)). Intuitively, the difference $\sqrt{n}\{\hat{\theta}_n - \tilde{\theta}_n\}$ therefore provides us with an estimator of the sample mean of an element $f \in \bar{T}(P)^{\perp}$, and a Hausman test can in turn be interpreted as a feasible version of the test developed in Section 3.3.

The above discussion can be generalized to allow for nonscalar $\theta(P)$ ($\mathbf{B} \neq \mathbf{R}$) by employing the dual space $\mathbf{B}^*$. In particular, note that for any $b^* \in \mathbf{B}^*$, $b^*(\theta(P))$ is a scalar valued parameter and $b^*(\hat{\theta}_n)$ and $b^*(\tilde{\theta}_n)$ are both regular and asymptotically linear estimators with influence functions $b^* \circ \nu$ and $b^* \circ \tilde{\nu}$ respectively. Thus, by our preceding discussion, $(b^* \circ \nu - b^* \circ \tilde{\nu}) \in \bar{T}(P)^{\perp}$ and in addition

$$\sqrt{n}\{b^*(\hat{\theta}_n) - b^*(\tilde{\theta}_n)\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{b^*(\nu(X_i)) - b^*(\tilde{\nu}(X_i))\} + o_p(1) \tag{39}$$

[12]See Lemma B.7 in the Appendix for a formal statement.

under P^n. Heuristically, every $b^* \in \mathbf{B}^*$ may therefore be employed to estimate the sample mean of the function $b^* \circ (\nu - \tilde{\nu}) \in \bar{T}(P)^{\perp}$ through the difference $\sqrt{n}\{b^*(\hat{\theta}_n) - b^*(\tilde{\theta}_n)\}$. More generally, for any norm bounded subset $\mathbb{U}$ of $\mathbf{B}^*$ we may collect all such estimates implied by $b^* \in \mathbb{U}$ into a stochastic process $\hat{\mathbb{G}}_n \in \ell^\infty(\mathbb{U})$ defined by

$$\hat{\mathbb{G}}_n(b^*) \equiv \sqrt{n}\{b^*(\hat{\theta}_n) - b^*(\tilde{\theta}_n)\} \ . \tag{40}$$

As the following Theorem shows, the stochastic process $\hat{\mathbb{G}}_n \in \ell^\infty(\mathbb{U})$ is asymptotically equivalent to the empirical process generated by $\mathcal{G} \equiv \{b^* \circ (\nu - \tilde{\nu}) : b^* \in \mathbb{U}\}$ and in this manner mimics the role of the process $\mathbb{G}_n$ in Lemma 3.1.

Theorem 4.2. *Let Assumptions 2.1, 4.1(i), and 4.2 hold, and $\mathbb{U} \subset \mathbf{B}^*$ be norm bounded. For any path $t \mapsto P_{t,g}$ let $L_{n,g}$ denote the law of $\{X_i\}_{i=1}^n$ under $\bigotimes_{i=1}^n P_{1/\sqrt{n},g}$, and define $\Delta_g \in \ell^\infty(\mathbb{U})$ pointwise by $\Delta_g(b^*) \equiv \int b^* \circ (\nu - \tilde{\nu}) g dP$. Then, there exists a tight centered Gaussian process $\mathbb{G}_0 \in \ell^\infty(\mathbb{U})$ such that for any path $t \mapsto P_{t,g}$ we have*

$$\hat{\mathbb{G}}_n \stackrel{L_{n,g}}{\to} \mathbb{G}_0 + \Delta_g \ . \tag{41}$$

Moreover, $b^* \circ (\nu - \tilde{\nu}) \in \bar{T}(P)^{\perp}$ for all $b^* \in \mathbb{U}$ and hence (i) $\Delta_g = 0$ whenever $t \mapsto P_{t,g} \in$ $\mathbf{P}$, and (ii) If in addition $cl\{lin\{\{b^* \circ (\nu - \tilde{\nu})\}_{b^* \in \mathbb{U}}\}\} = \bar{T}(P)^{\perp}$, then $\Delta_g \neq 0$ whenever

$$\liminf_{n \to \infty} \inf_{Q \in \mathbf{P}} n \int [dQ^{1/2} - dP^{1/2}_{1/\sqrt{n},g}]^2 > 0 \ . \tag{42}$$

In accord with Lemma 3.1, the asymptotic distribution of $\hat{\mathbb{G}}_n$ exhibits a drift Δ_g when the data $\{X_i\}_{i=1}^n$ is distributed according to $P^n_{1/\sqrt{n},g}$ in place of P^n. The functional form of Δ_g is one of covariances between the score g and the functions $\{b^* \circ (\nu - \tilde{\nu})\}_{b^* \in \mathbb{U}}$, which as argued belong to $\bar{T}(P)^{\perp}$. Hence, following the logic of Lemma 3.1, we again conclude that $\Delta_g = 0$ if the path $t \mapsto P_{t,g}$ approaches P from within the model $\mathbf{P}$. In contrast, if the path $t \mapsto P_{t,g}$ does not approach $\mathbf{P}$ "too fast" (see (42)), then the drift Δ_g will be nonzero provided the closure of the linear span of $\{b^* \circ (\nu - \tilde{\nu})\}_{b^* \in \mathbb{U}}$ equals $\bar{T}(P)^{\perp}$. Whenever $\bar{T}(P)^{\perp}$ is infinite dimensional, this condition necessitates $\theta(P)$ to be itself infinite dimensional, thus motivating our focus on Banach space valued parameters.

Remark 4.2. Theorem 4.2(ii) requires the closure of the linear span of $\{b^* \circ (\nu - \tilde{\nu})\}_{b^* \in \mathbb{U}}$ to equal $\bar{T}(P)^{\perp}$. More generally, however, such closure may be equal to a strict subspace of $\bar{T}(P)^{\perp}$ – for instance whenever the dimension of $\mathbf{B}$ is smaller than that of $\bar{T}(P)^{\perp}$; see the GMM discussion below. In such a case, the drift Δ_g is nonzero if and only if the projection of g onto $\mathrm{cl}\{\mathrm{lin}\{\{b^* \circ (\nu - \tilde{\nu})\}_{b^* \in \mathbb{U}}\}\}$ is nonzero. Thus, if $\{b^* \circ (\nu - \tilde{\nu})\}_{b^* \in \mathbb{U}}$ is not sufficiently rich, then there exists a path $t \mapsto P_{t,g}$ satisfying (42) and for which $\Delta_g = 0$ in $\ell^\infty(\mathbb{U})$. As a result, a Hausman test based on $\hat{\mathbb{G}}_n$ will still provide local size control but have power no larger than size against certain local alternatives. ■

Theorem 4.2 provides the foundation for constructing a specification test, much in the same manner Lemma 3.1 was exploited to build an infeasible test based on the untenable knowledge of a subset $\mathcal{F}$ of $\bar{T}(P)^{\perp}$. In particular, comparing a test statistic $\Psi(\hat{\mathbb{G}}_n)$ to its asymptotic $1-\alpha$ quantile yields a locally unbiased test for any continuous subconvex function $\Psi : \ell^{\infty}(\mathbb{U}) \to \mathbf{R}_+$. Corollary 4.1 illustrates this point by providing an analogue to Theorem 3.2. While we abstract from the problem of estimating asymptotic critical values, we note that the use of the bootstrap is automatically justified whenever it is valid for the asymptotic joint law of the original estimators $\hat{\theta}_n$ and $\tilde{\theta}_n$ on $\mathbf{B} \times \mathbf{B}$.

Corollary 4.1. *Let Assumptions 2.1, 3.2, 4.1(i), 4.2 hold, and for any $\alpha \in (0,1)$ let $c_{1-\alpha}$ denote the $1-\alpha$ quantile of $\Psi(\mathbb{G}_0)$. Further let $\mathbb{U} \equiv \{b_k^*\}_{k=1}^{d_F}$ be norm bounded, and suppose $\sum_{k=1}^{d_F} \int (b_k^* \circ (\nu - \tilde{\nu}))^2 dP < \infty$. If $c_{1-\alpha} > 0$, then for any path $t \mapsto P_{t,g} \in \mathbf{P}$:*

$$\lim_{n\to\infty} P^n_{1/\sqrt{n},g}(\Psi(\hat{\mathbb{G}}_n) > c_{1-\alpha}) = \alpha \ . \tag{43}$$

If in addition $cl\{lin\{\{b_k^* \circ (\nu - \tilde{\nu})\}_{k=1}^{d_F}\}\} = \bar{T}(P)^{\perp}$ and $t \mapsto P_{t,g}$ satisfies (42), then

$$\liminf_{n\to\infty} P^n_{1/\sqrt{n},g}(\Psi(\hat{\mathbb{G}}_n) > c_{1-\alpha}) \geq P(\Psi(\mathbb{G}_0 + \Delta_g) > c_{1-\alpha}) > \alpha \ . \tag{44}$$

We conclude by illustrating the construction of Hausman tests in the GMM setting.

GMM Illustration (cont). Recall that for every $P \in \mathbf{P}$, $\beta(P) \in \Theta$ is the parameter solving $\int \rho(\cdot, \beta(P))dP = 0$ and consider the construction of a Hausman test based on alternative estimators of $\theta(P) \equiv \beta(P)$. In particular, for different $d_\rho \times d_\rho$ positive definite matrices W_1 and W_2, and $\hat{\beta}_n^W$ the estimator defined in (37), it follows that

$$\sqrt{n}\{\hat{\beta}_n^{W_1} - \hat{\beta}_n^{W_2}\} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \{V_{W_1}(P) - V_{W_2}(P)\}\rho(X_i, \beta(P)) + o_p(1) \tag{45}$$

under P^n, where $V_W(P) \equiv (\Gamma(P)'W\Gamma(P))^{-1}\Gamma(P)'W$. Since in this case $\mathbf{B} = \mathbf{R}^{d_\beta}$, the dual space equals $\mathbf{B}^*$ equals $\mathbf{R}^{d_\beta}$ and as indicated by Theorem 4.2 $\lambda_\beta'\{V_{W_1}(P) - V_{W_2}(P)\}\rho(\cdot, \beta(P))$ indeed belongs to $\bar{T}(P)^{\perp}$ for any $\lambda_\beta \in \mathbf{R}^{d_\beta}$.[13] Letting $e_k \in \mathbf{R}^{d_\beta}$ denote the vector whose k^{th} coordinate is one and all other coordinated are zero, we may then set $\mathbb{U} \equiv \{e_k\}_{k=1}^{d_\beta}$ and $\Psi : \ell^{\infty}([d_\beta]) \to \mathbf{R}_+$ to be given by $\Psi(b) = \|b\|^2$ to obtain

$$\Psi(\hat{\mathbb{G}}_n) = \|\sqrt{n}\{\hat{\beta}_n^{W_1} - \hat{\beta}_n^{W_2}\}\|^2 \ . \tag{46}$$

We note, however, that the linear span of $\{e_k'\{V_{W_1}(P) - V_{W_2}(P)\}\rho(\cdot, \beta(P))\}_{k=1}^{d_\beta}$ is at most a d_β-dimensional subspace of $\bar{T}(P)^{\perp}$. Thus, since in contrast the dimension of $\bar{T}(P)^{\perp}$ is $d_\rho - d_\beta$ (see (9)), it follows that a Hausman test based on (46) will fail to have power against certain local alternatives when $d_\rho > 2d_\beta$; see Remark 4.2. Such a

[13]This follows by noting that for any $\lambda_\beta \in \mathbf{R}^{d_\beta}$ we may set $\lambda_\rho \equiv \{V_{W_1}(P)' - V_{W_2}(P)'\}\lambda_\beta$ to obtain $\lambda_\beta'\{V_{W_1}(P) - V_{W_2}(P)\} = \lambda_\rho' S(P)$, and hence $\lambda_\beta'\{V_{W_1}(P) - V_{W_2}(P)\}\rho(\cdot, \beta(P)) \in \bar{T}(P)^{\perp}$ by (9).

problem can be easily addressed by either considering higher dimensional parameters, such as the c.d.f. of P, or by comparing more than two estimators – e.g. by employing $(\sqrt{n}\{\hat{\beta}_n^{W_1} - \hat{\beta}_n^{W_2}\}, \sqrt{n}\{\hat{\beta}_n^{W_1} - \hat{\beta}_n^{W_3}\})$ for different matrices W_1, W_2, and W_3. ■

5 Nonparametric Conditional Moment Models

In order to demonstrate the utility of our general results, we next study a broad class of nonparametric conditional moment restriction models.

5.1 General Setup

In what follows, we set $X = (Y, Z, W)$ for an outcome variable $Y \in \mathbf{R}^{d_y}$, a potentially endogenous variable $Z \in \mathbf{R}^{d_z}$, and an exogenous instrument $W \in \mathbf{R}^{d_w}$. For a known function $\rho : \mathbf{R}^{d_y} \times \mathbf{R} \to \mathbf{R}$, the distribution P of X is then assumed to satisfy

$$E[\rho(Y, h_P(Z))|W] = 0 \tag{47}$$

for some unknown function $h_P : \mathbf{R}^{d_z} \to \mathbf{R}$. The model is nonparametric in that we only require $h_P(Z)$ to have a second moment – i.e. we assume $h_P \in L^2$.

As previously noted by Chen and Pouzo (2012), model (47) may be studied without requiring differentiability of $\rho : \mathbf{R}^{d_y} \times \mathbf{R} \to \mathbf{R}$ provided the conditional expectation is appropriately "smooth". To formalize this differentiability requirement, we let L^2_Z and L^2_W denote the functions in L^2 depending only on Z and W respectively, and define

$$m(W, h) \equiv E[\rho(Y, h(Z))|W] \ . \tag{48}$$

The following Assumption then imposes the required restrictions on the function ρ.

Assumption 5.1. *(i) $E[\{\rho(Y, h(Z))\}^2] < \infty$ for all $h \in L^2_Z$; (ii) $m(W, \cdot) : L^2_Z \to L^2_W$ is pathwise differentiable at h_P with derivative $\nabla m(W, h_P)[s] \equiv \frac{\partial}{\partial \tau} m(W, h_P + \tau s)|_{\tau=0}$; (iii) The map $\nabla m(W, h_P) : L^2_Z \to L^2_W$ is linear and continuous; (iv) There exist constants $\eta, M \in \mathbf{R}$ such that $0 < \eta < M < \infty$ and $P(\eta < Var\{\rho(Y, h_P(Z))|W\} < M) = 1$.*

Assumption 5.1(i) and Jensen's inequality imply that $m(W, h) \in L^2_W$ for all $h \in L^2_Z$, and hence we may view $m(W, \cdot)$ as a map from L^2_Z into L^2_W. Given the codomain space of the map $m(W, \cdot)$, in Assumption 5.1(ii) we further require that $m(W, \cdot) : L^2_Z \to L^2_W$ be pathwise (Gateaux) differentiable at h_P. In turn, Assumption 5.1(iii) imposes that the pathwise derivative $\nabla m(W, h_P) : L^2_Z \to L^2_W$ be linear and continuous in its direction – a property that is not guaranteed by pathwise differentiability. Finally, Assumption 5.1(iv) demands that the conditional variance of $\rho(Y, h_P(Z))$ given W be bounded from above and away from zero almost surely. The conditions of Assumption 5.1 are commonly imposed in the analysis of efficient estimation in conditional moment restriction models. We similarly require them because we will heavily rely on results from the efficiency literature in our analysis; see Chamberlain (1992); Ai and Chen (2003, 2012).

The range of the derivative $\nabla m(W, h_P) : L^2_Z \to L^2_W$ plays a fundamental role in deriving a characterization of local overidentification. We therefore introduce the notation

$$\mathcal{R} \equiv \{f \in L^2_W : f = \nabla m(W, h_P)[s] \text{ for some } s \in L^2_Z\} \ , \tag{49}$$

and let $\bar{\mathcal{R}}$ denote the closure of $\mathcal{R}$ in L^2_W – i.e. $\bar{\mathcal{R}}$ denotes the subset of L^2_W that can be arbitrarily well approximated by functions of the form $\nabla m(W, h_P)[s]$ for some $s \in L^2_Z$. It is worth noting that when Z does not equal W, the presence of an ill-posed inverse problem and the closed graph theorem imply $\mathcal{R}$ cannot equal L^2_W. Nonetheless, the closure $\bar{\mathcal{R}}$ may still equal L^2_W even when an ill-posed inverse problem is present.

We illustrate the introduced concepts through the nonparametric instrumental variables model (NPIV) of Newey and Powell (2003), Hall and Horowitz (2005), and Darolles et al. (2011), and the nonparametric quantile instrumental variables model (NPQIV) examined by Chernozhukov et al. (2007), Horowitz and Lee (2007), and Chen et al. (2014).

Example 5.1. (NPIV) The NPIV model corresponds to setting $\rho(y, u) = y - u$ for all $y, u \in \mathbf{R}$, in which case the restriction in (47) reduces to

$$E[Y - h_P(Z)|W] = 0 \ . \tag{50}$$

The map $m(W, \cdot) : L^2_Z \to L^2_W$ is then given by $m(W, h) \equiv E[Y - h(Z)|W]$ which, since it is linear, is trivially pathwise differentiable with derivative $\nabla m(W, h_P)[s] = -E[s(Z)|W]$ for any $s \in L^2_Z$. In this context, the range $\mathcal{R}$ of $\nabla m(W, h_P)$ is therefore given by

$$\mathcal{R} = \{f \in L^2_W : f(W) = E[s(Z)|W] \text{ for some } s \in L^2_Z\} \ . \tag{51}$$

We also observe that if $Z = W$, then (50) reduces to the mean regression model. ■

Example 5.2. (NPQIV) Setting $\rho(y, u) = 1\{y \leq u\} - \tau$ for $\tau \in (0, 1)$ in (47) yields

$$P(Y \leq h_P(Z)|W) = \tau \ , \tag{52}$$

which corresponds to the NPQIV model. Thus, in this context $m(W, \cdot) : L^2_Z \to L^2_W$ is given by $m(W, h) \equiv P(Y \leq h(Z)|W) - \tau$ for all $h \in L^2_Z$. If Y is continuously distributed conditional on (Z, W) with a density $f_{Y|Z,W}$ that is both bounded and continuous, then it is possible to show $m(W, \cdot) : L^2_Z \to L^2_W$ is pathwise differentiable with

$$\nabla m(W, h_P)[s] = E[f_{Y|Z,W}(h_P(Z)|Z, W)s(Z)|W] \tag{53}$$

for any $s \in L^2_Z$. Therefore, in this model the range $\mathcal{R}$ of $\nabla m(W, h_P)$ is equal to

$$\mathcal{R} = \{f \in L^2_W : f(W) = E[f_{Y|Z,W}(h_P(Z)|Z, W)s(Z)|W] \text{ for some } s \in L^2_Z\} . \tag{54}$$

It is worth noting that when $Z = W$ (52) reduces to a nonparametric quantile regression model and (53) simplifies to $\nabla m(Z, h_P)[s] = s(Z) f_{Y|Z}(h_P(Z)|Z)$. ■

5.2 Characterization

By relating local overidentification to estimation, Theorem 4.1 enables us to leverage existing results to analyze the nonparametric conditional moment restriction model in (47). In particular, efficient estimation in these models is well understood with Ai and Chen (2012) notably deriving the efficiency bound for a general class of functions of P. While their results cover intrinsically interesting parameters such as average derivatives, for our purposes it is convenient to, for any $f \in L^2$, focus on the simpler parameter

$$\theta_f(P) \equiv \int f dP . \tag{55}$$

As argued in Remark 4.1, local just identification is equivalent to the sample mean being an efficient estimator of $\theta_f(P)$ for all f in any dense subset $\mathcal{D}$ of L^2. Thus, we may characterize local just identification by obtaining necessary and sufficient conditions for the efficiency bound for $\theta_f(P)$, denoted Ω^*_f, to equal the asymptotic variance of the sample mean, i.e. $\text{Var}\{f(X)\}$. The following proposition exploits the explicit formulation for the efficiency bound Ω^*_f derived in Ai and Chen (2012) to accomplish this goal.

Proposition 5.1. *Let Assumption 5.1 hold. Then, there exists a dense subset $\mathcal{D}$ of L^2 such that $Var\{f(X)\} = \Omega^*_f$ for all $f \in \mathcal{D}$ if and only if $\bar{\mathcal{R}} = L^2_W$.*

Proposition 5.1 establishes that sample means are efficient if and only if the closure of the range of $\nabla m(W, h_P) : L^2_Z \to L^2_W$ is equal to L^2_W. Hence, we conclude that local just identification in a nonparametric conditional moment restriction model is equivalent to the requirement $\bar{\mathcal{R}} = L^2_W$ while local overidentification corresponds to the case $\bar{\mathcal{R}} \subsetneq L^2_W$. Heuristically, we may thus understand local overidentification as the existence of transformations of the instrument ($f \in L^2_W$) that cannot be mimicked by a Z induced local change in the conditional expectation ($\nabla m(W, h_P)[s]$ for $s \in L^2_Z$). This characterization of local overidentification is analogous to the one obtained in the GMM context, where L^2_W corresponds to the codomain of the restrictions ($\mathbf{R}^{d_\rho}$) and $\bar{\mathcal{R}}$ corresponds to the range of the derivative of the moment restrictions ($\mathbf{R}^{d_\beta}$). Thus, local overidentification ($\bar{\mathcal{R}} \subsetneq L^2_W$) maps into the case $\mathbf{R}^{d_\beta} \subsetneq \mathbf{R}^{d_\rho}$ – i.e. $d_\beta < d_\rho$. In Remarks 5.1 and 5.2 below, we discuss an alternative characterization of local overidentification, as well as the implications of imposing additional restrictions on h_P.

Remark 5.1. Since L^2_Z and L^2_W are both Hilbert spaces, the requirement that $\bar{\mathcal{R}}$ be equal to L^2_W can also be expressed in terms of the adjoint to $\nabla m(W, h_P) : L^2_Z \to L^2_W$, which we denote by $\nabla m(W, h_P)^* : L^2_W \to L^2_Z$.[14] In particular, $\bar{\mathcal{R}}$ must equal the orthogonal complement to the null space of $\nabla m(W, h_P)^*$, and therefore[15]

$$\bar{\mathcal{R}} = L^2_W \text{ if and only if } \{0\} = \{s \in L^2_W : \nabla m(W, h_P)^*[s] = 0\} \; ; \tag{56}$$

i.e. $\bar{\mathcal{R}} = L^2_W$ if and only if $\nabla m(W, h_P)^*$ is injective. Interestingly, the requirement that the derivative $\nabla m(W, h_P)$ be injective is instrumental in ensuring local identification (Chen et al., 2014). Thus, (56) reflects a symmetry between local identification (injectivity of $\nabla m(W, h_P)$) and local just identification (injectivity of $\nabla m(W, h_p)^*$). ■

Remark 5.2. Imposing restrictions on h_P beyond it belonging to L^2_Z can reduce the tangent space $\bar{T}(P)$ and therefore affect the characterization of local overidentification. For instance, if h_P is instead assumed to belong to a vector subspace $\mathcal{H}$ of L^2_Z, then Proposition 5.1 continues to hold provided $\mathcal{R}$ is redefined as (compare to (49)):

$$\mathcal{R} \equiv \{f \in L^2_W : f = \nabla m(W, h_P)[s] \text{ for some } s \in \mathcal{H}\} \; . \tag{57}$$

Hence, restricting the parameter space from L^2_Z to $\mathcal{H}$ potentially reduces $\bar{\mathcal{R}}$ and yields local just identification ($\bar{\mathcal{R}} = L^2_W$) less tenable. The conclusion that semiparametric conditional moment restriction models are typically locally overidentified, for example, follows from Theorem 4.1(ii) and the availability of both efficient and inefficient estimators for such models (Chamberlain, 1992; Ai and Chen, 2003; Chen and Pouzo, 2009). In fact, the sufficient conditions in Bonhomme (2012) and Chen et al. (2014) for local identification of Euclidean parameters directly imply the local overidentification of P even in the presence of partially identified nuisance parameters.[16] ■

Exploiting Proposition 5.1 and Remark 5.1, we revisit Examples 5.1 and 5.2 to obtain simple characterizations of local overidentification in the NPIV and NPQIV models.

Example 5.1 (cont.): Recall that we showed the derivative $\nabla m(W, h_P) : L^2_Z \to L^2_W$ satisfies $\nabla m(W, h_P)[s] = -E[s(Z)|W]$ for all $s \in L^2_Z$. Hence, its adjoint is given by

$$\nabla m(W, h_P)^*[s] = -E[s(W)|Z] \tag{58}$$

for all $s \in L^2_W$, and by Proposition 5.1 and Remark 5.1 we conclude local just identifi-

[14] The adjoint $\nabla m(W, h_P)^*$ is the unique continuous linear map from L^2_W into L^2_Z satisfying $\int \{\nabla m(W, h_P)[h]\} s dP = \int h\{\nabla m(W, h_P)^*[s]\} dP$ for all $h \in L^2_Z$ and $s \in L^2_W$.

[15] See for instance Theorem 6.6.3(2) in Luenberger (1969).

[16] Examples for identification of Euclidean parameter without identification of an unknown function of endogenous variables include Santos (2011) (for NPIV), Florens et al. (2012) (for partially linear NPIV) and Chen et al. (2014) (for single-index IV).

cation is equivalent to the conditional expectation operator $E[\cdot|Z]$ being injective:

$$E[s(W)|Z] = 0 \text{ implies } s(W) = 0 \text{ for all } s \in L^2_W \ . \tag{59}$$

The property in (59) is known as the distribution of (Z, W) being L^2-complete with respect to Z (Andrews, 2011). The related requirement of L^2-completeness with respect to W is necessary for identification, and thus while identification needs W to be a "good" instrument for Z, local overidentification requires Z to be a "poor" instrument for W. Since examples of distributions exist for which L^2-completeness fails (Santos, 2012), we conclude P can be locally overidentified even if Z and W are of equal dimension. ■

Example 5.2 (cont): Given the formulation of the derivative $\nabla m(W, h_P)$ in (53), it is straightforward to characterize its adjoint $\nabla m(W, h_P)^*$ as satisfying for any $s \in L^2_W$

$$\nabla m(W, h_p)^*[s] = E[f_{Y|Z,W}(h_P(Z)|Z, W)s(W)|Z] \ . \tag{60}$$

Since by Proposition 5.1 and Remark 5.1 local just identification is equivalent to injectivity of the adjoint $\nabla m(W, h_P)^*$, in this context we obtain the characterization

$$E[f_{Y|Z,W}(h_P(Z)|Z, W)s(W)|Z] = 0 \text{ implies } s(W) = 0 \ \text{ for all } s \in L^2_W \ . \tag{61}$$

We note the similarity of the local just identification condition in the NPIV and NPQIV models (compare (59) and (61)), though in the latter the presence of the conditional density $f_{Y|Z,W}$ reflects the nonlinearity of the problem. ■

5.3 Special Case: Exogeneity

In distilling local overidentification to the condition $\bar{\mathcal{R}} \subsetneq L^2_W$, Proposition 5.1 and our previous results fully characterize whether locally nontrivial specification tests exist (Theorem 3.2) and whether efficiency considerations should be of concern (Theorem 4.1). While intuitive, the local overidentification requirement $\bar{\mathcal{R}} \subsetneq L^2_W$ is certainly harder to verify than comparing the dimension of the parameter to the number of restrictions, and may in fact be untestable without appropriate restrictions on P (Canay et al., 2013). Fortunately, as we next show, additional structure such as exogeneity of Z can help further simplify the characterization of local overidentification.

In what follows, we refer to Z as exogenous if it is part of the conditioning variable W – i.e. $W = (Z, V)$ for some possibly degenerate variable V.[17] Thus, we impose:

Assumption 5.2. *(i) $W = (Z, V)$ and $E[WW'] < \infty$ (ii) There exists a $d_0 : \mathbf{R}^{d_w} \to \mathbf{R}$ such that $d_0(W)$ is bounded P-a.s. and $\nabla m(W, h_P)[s] = d_0(W)s(Z)$ for all $s \in L^2_Z$.*

[17] We recognize this may not be a standard definition of "exogeneity". However, we employ it due to the definition reflecting the common practice of conditioning on exogenous variables and because whether Z is part of W or not is of key importance from a mathematical perspective.

While Assumption 5.2(i) formalizes the requirement of exogeneity, Assumption 5.2(ii) strengthens Assumption 5.1(iii) by imposing an additional requirement on the specific structure of the derivative $\nabla m(W, h_P) : L^2_Z \to L^2_W$ – namely that it be of the form $\nabla m(W, h_P)[s] = d_0(W)s(Z)$ for all $s \in L^2_Z$. The latter specification is easily verified whenever $\rho : \mathbf{R}^{d_y} \times \mathbf{R} \to \mathbf{R}$ is partially differentiable in its final argument, in which case Assumption 5.2(ii) holds under mild conditions with $d_0(W) = E[\partial_h \rho(Y, h_P(Z))|W]$. In the more general case where $\rho : \mathbf{R}^{d_y} \times \mathbf{R} \to \mathbf{R}$ is not differentiable but the map $m(W, \cdot) : L^2_Z \to L^2_W$ is, Assumption 5.2(ii) can still be verified on a case by case basis.

Under the additional structure provided by Assumption 5.2, the following Proposition simplifies the characterization of local just identification ($\bar{\mathcal{R}} = L^2_W$).

Proposition 5.2. *If Assumptions 5.1(i)-(ii), 5.2 hold, then $\bar{\mathcal{R}} = L^2_W$ if and only if*

$$P(E[V|Z] = V) = 1 \quad \textit{and} \quad P(d_0(W) \neq 0) = 1 \ . \tag{62}$$

Heuristically, Proposition 5.2 characterizes local just identification in terms of two key conditions. First, V must be a deterministic function of Z ($E[V|Z] = V$) thus preventing W from possessing variation that is unexplained by Z. Second, the derivative $\nabla m(W, h_P) : L^2_Z \to L^2_W$ must be injective ($d_0(W) \neq 0$), which is a condition often associated with local identification (Chen et al., 2014). In the present context, however, local just identification is also intrinsically linked to the rank condition on the derivative $\nabla m(W, h_P)$ due to the latter being self-adjoint when $Z = W$; see Remark 5.3.

Remark 5.3. Whenever $Z = W$ we have $L^2_Z = L^2_W$ and the derivative $\nabla m(W, h_P) : L^2_W \to L^2_W$ satisfying Assumption 5.2(ii) implies it is self-adjoint. Since, as argued in Remark 5.1, local just identification is equivalent to the adjoint $\nabla m(W, h_P)^*$ being injective, it follows that under self-adjointness local just identification is tantamount to injectivity of $\nabla m(W, h_P)$ – a requirement that reduces to $P(d_0(W) \neq 0) = 1$ under Assumption 5.2(ii). It is worth noting that injectivity of $\nabla m(W, h_P)$ is a strictly weakly requirement than its invertibility, which fails for instance if $E[\{d_0(W)\}^{-2}] = \infty$. ■

5.4 Discussion: Two Step Procedures

In conjunction with our previous results, Propositions 5.1 and 5.2 have strong implications for the estimation of regular parameters in nonparametric conditional moment restrictions models with or without endogeneity.

For the special case in which the conditioning variable (W) equals the argument of the nonparametric function (Z), an extensive literature has examined estimation of functionals of h_P such as the average derivative or consumer surplus (Powell et al., 1989; Newey and Stoker, 1993). A common feature of these estimation problems is that "plug-in" estimators are efficient and moreover that their asymptotic distribution is invariant to the choice of estimator for h_P (Newey, 1994). This classical result is in fact immediately implied by Theorem 4.1(i) whenever P is locally just identified, since we may always view regular functionals, such as the average derivatives of h_P, as a parameter $\theta : \mathbf{P} \to \mathbf{R}$. In particular, by Proposition 5.2 we conclude regular plug-in estimators must be efficient whenever condition (62) is satisfied.

Our results, however, further imply that plug-in estimators of functionals of h_P need not be efficient when P is locally overidentified by $\mathbf{P}$. For instance, even under exogeneity of Z, Proposition 5.2 and Theorem 4.1(ii) imply plug-in estimators may fail to be efficient whenever W contains variation unexplained by Z – i.e. condition (62) fails.[18] Analogously, we also conclude that plug-in estimators of regular parameters need not be efficient when Z is endogenous since, as illustrated by Examples 5.1 and 5.2, P may be locally overidentified in such problems; see also Remark 5.4. We note, however, that as shown in Ai and Chen (2012) efficient estimators of functionals of h_P may still be available when P is locally overidentified and, as implied by Theorem 4.2, employed to construct Hausman tests if desired.

Remark 5.4. Building on Newey and Powell (1999), Ackerberg et al. (2014) show that with an "exactly identified" nonparametric conditional moment model first step, semiparametric two step GMM estimation can be fully efficient. The key requirements they refer to as "exact identification" are that $Z = W$ in the conditional moment restriction (5.2) and that $d_0(W) \neq 0$ almost surely. Our analysis complements theirs, since Proposition 5.2 establishes their "exact identification" requirement is equivalent to P being locally just identified by the first stage nonparametric conditional moment restriction. We also conclude from our results that semiparametric two step estimation can be inefficient when (i) Z is exogenous but $Z \subsetneq W$, (ii) Further restrictions are imposed on h_P (Remark 5.2), or (iii) Z is endogenous as in NPIV. ■

6 Conclusion

This paper reinterprets the common practice of counting the number of restrictions and parameters in GMM to determine overidentification as an approach that implicitly examines the tangent space $\bar{T}(P)$ as a subset of L^2_0. This abstraction naturally leads to the notion of local overidentification, which we show is responsible for an intrinsic link between the semiparametric efficient estimation of regular parameters and the local testability of a model. When applied to nonparametric conditional moment restriction

[18]We emphasize this conclusion does not contradict the results of Newey (1994), who in establishing the first order equivalence of asymptotically linear and regular estimators imposes a condition that is tantamount to P being locally just identified.

models, we obtain a simple condition that determines both whether the model is locally testable and whether efficiency gains are available in estimating regular parameters.

This paper assumes that P is the distribution of a single observation from an i.i.d. sample for the sake of simplicity. Most of the results carry over to weakly dependent data. There is some work (such as Ploberger and Phillips (2012) and the references therein) on applying limits of experiments to specific models with nonstationary, strongly dependent data. We conjecture that many results in this paper could be extended to general semi/nonparametric models with temporal or/and spatial dependent processes by using limits of experiments theories for martingales and conditional scores. We leave such extension for future work.

Appendix A - Proof of Main Results

The following list includes notation and definitions that will be used in the appendix.

$[d]$	For an integer $d > 0$, the set $[d] \equiv \{1, 2, \cdots, d\}$.
$\ell^\infty(\mathbf{A})$	For a set $\mathbf{A}$, $\ell^\infty(\mathbf{A}) \equiv \{f : \mathbf{A} \to \mathbf{R} : \sup_{a\in A} |f(a)| < \infty\}$.
L_0^2	The set $L_0^2 \equiv \{g : \int g dP = 0$ and $\int g^2 dP < \infty\}$.
$\bar{T}(P)$	The closure of $T(P) \equiv \{g \in L_0^2 :$ (1) holds for some $t \mapsto P_{t,g} \in \mathbf{P}\}$.
$\bar{T}(P)^\perp$	The complement $\bar{T}(P)^\perp \equiv \{g \in L_0^2 : \int gf dP = 0$ for all $f \in \bar{T}(P)\}$.

Proof of Theorem 3.1: First note that for any $g \in L_0^2$ it is possible to construct a path $t \mapsto P_{t,g}$ whose score is indeed g; see Example 3.2.1 in Bickel et al. (1993) for a concrete construction. Moreover, further observe that any two paths $t \mapsto \tilde{P}_{t,g}$ and $t \mapsto P_{t,g}$ with the same score g satisfy by $0 \le \phi_n \le 1$ and Lemma B.1:

$$\lim_{n\to\infty} |\int \phi_n d\tilde{P}^n_{1/\sqrt{n},g} - \int \phi_n dP^n_{1/\sqrt{n},g}| \le \lim_{n\to\infty} \int |dP^n_{1/\sqrt{n},g} - d\tilde{P}^n_{1/\sqrt{n},g}| = 0 \ . \quad \text{(A.1)}$$

For each $g \in L_0^2$ we may therefore select an arbitrary path $t \mapsto P_{t,g}$ whose score is indeed g, and for $\mathcal{B}$ the Borel σ-algebra on $\mathbf{R}$ we consider the sequence of experiments

$$\mathcal{E}_n \equiv (\mathbf{R}^{dn}, \mathcal{B}^{dn}, \bigotimes_{i=1}^n P_{1/\sqrt{n},g} : g \in L_0^2) \ . \quad \text{(A.2)}$$

Next note that since $\{\psi_k^T\}_{k=1}^{d_T} \cup \{\psi_k^\perp\}_{k=1}^{d_{T^\perp}}$ forms an orthonormal basis for L_0^2, we obtain from Lemma B.3 that $\mathcal{E}_n$ converges weakly to the experiment $\mathcal{E}$ given by

$$\mathcal{E} \equiv (\mathbf{R}^{d_T} \times \mathbf{R}^{d_{T^\perp}}, \mathcal{B}^{d_T} \times \mathcal{B}^{d_{T^\perp}}, Q_g : g \in L_0^2) \ , \quad \text{(A.3)}$$

where we have exploited that for $d_P \equiv \dim\{L_0^2\}$ we have $\mathbf{R}^{d_T} \times \mathbf{R}^{d_{T^\perp}} = \mathbf{R}^{d_P}$ and $\mathcal{B}^{d_T} \times \mathcal{B}^{d_{T^\perp}} = \mathcal{B}^{d_P}$. The existence of a test function $\phi : \mathbf{R}^{d_T} \times \mathbf{R}^{d_{T^\perp}} \to [0,1]$ satisfying (17) then follows from Theorem 7.1 in van der Vaart (1991a).

To conclude the proof, it only remains to show that ϕ must controls size in (14). To this end, note that $\Pi_{T^\perp}(g) = 0$ if and only if $g \in \bar{T}(P)$. Fixing $\delta > 0$ then observe that for any $g \in \bar{T}(P)$ there exists a $\tilde{g} \in T(P)$ such that $\|g - \tilde{g}\|_{L^2} < \delta$. Moreover, since $\tilde{g} \in T(P)$, there exists a path $t \mapsto \tilde{P}_{t,\tilde{g}} \in \mathbf{P}$ with score $\tilde{g}$ and hence

$$\begin{aligned} \int \phi dQ_g &= \lim_{n\to\infty} \int \phi_n dP^n_{1/\sqrt{n},g} \\ &\le \lim_{n\to\infty} \int \phi_n d\tilde{P}^n_{1/\sqrt{n},\tilde{g}} + \limsup_{n\to\infty} \int |dP^n_{1/\sqrt{n},g} - d\tilde{P}^n_{1/\sqrt{n},\tilde{g}}| \\ &\le \alpha + 2\{1 - \exp\{-\frac{\delta^2}{4}\}\}^{1/2} \ , \end{aligned} \quad \text{(A.4)}$$

where the first inequality employed that $0 \leq \phi_n \leq 1$ and the second inequality exploited (16) and Lemma B.1. Since $\delta > 0$ was arbitrary, we conclude from (A.4) that $\int \phi dQ_g \leq \alpha$ whenever $g \in \bar{T}(P)$ and the Theorem follows. ■

Proof of Corollary 3.1: The proof proceeds by contradiction. First note that if (18) does not hold, then we may pass to a subsequence $\{n_k\}_{k=1}^{\infty}$ such that

$$\lim_{k\to\infty} \inf_{\{g\in L_0^2:\|\Pi_{T^\perp}(g)\|_{L^2}\geq B\}} \int \phi_{n_k} dP^{n_k}_{1/\sqrt{n_k},g} \geq M_0 \tag{A.5}$$

for some constant $M_0 > P(\mathcal{X}^2(B) \geq \chi^2_{1-\alpha})$. Further observe that by Theorem 3.1, there exists a level α test ϕ of (14) such that for every path $t \mapsto P_{t,g}$ we have

$$\lim_{k\to\infty} \int \phi_{n_k} dP^{n_k}_{1/\sqrt{n_k},g} = \int \phi dQ_g \ . \tag{A.6}$$

Moreover, by results (A.5), (A.6), $\bar{T}(P)^\perp \subseteq L_0^2$, and $g = \Pi_{T^\perp}(g)$ for all $g \in \bar{T}(P)^\perp$,

$$M_0 \leq \inf_{\{g\in L_0^2:\|\Pi_{T^\perp}(g)\|_{L^2}\geq B\}} \int \phi dQ_g \leq \inf_{\{g\in \bar{T}(P)^\perp:\|g\|_{L^2}\geq B\}} \int \phi dQ_g \ . \tag{A.7}$$

Let Φ denote the standard normal measure on $\mathbf{R}$, and note that for any $g \in \bar{T}(P)^\perp$, the measure Q_g is such that: (i) Y is independent of Z, (ii) $Y \sim \bigotimes_{k=1}^{d_T} \Phi$, and (iii) $Z \sim \bigotimes_{k=1}^{d_{T^\perp}} \Phi(\cdot - h_{g,k})$ where $h_{g,k} \equiv \int g\psi_k^\perp dP$. For each $z \in \mathbf{R}^{d_{T^\perp}}$ then define

$$\bar{\phi}(z) \equiv E[\phi(Y, z)] \ , \tag{A.8}$$

where the expectation is taken with respect to $Y \sim \bigotimes_{i=1}^{d_T} \Phi$. We then obtain that

$$\inf_{\{g\in \bar{T}(P)^\perp:\|g\|_{L^2}\geq B\}} \int \phi dQ_g = \inf_{\{h\in \mathbf{R}^{d_{T^\perp}}:\|h\|\geq B\}} \int \bar{\phi} d\{\bigotimes_{k=1}^{d_{T^\perp}} \Phi(\cdot - h_k)\} \ , \tag{A.9}$$

by (A.8), the law of iterated expectations, and noting $\bar{T}(P)^\perp$ is isometrically isomorphic to $\mathbf{R}^{d_{T^\perp}}$ via the map $\Upsilon(g) = (h_{g,1}, \ldots, h_{g,d_{T^\perp}})'$ due to $\{\psi_k^\perp\}_{k=1}^{d_{T^\perp}}$ being an orthonormal basis of $\bar{T}(P)^\perp$ and $d_{T^\perp} < \infty$. Finally, we observe that $\bar{\phi}$ also satisfies

$$\int \bar{\phi} d\{\bigotimes_{k=1}^{d_{T^\perp}} \Phi\} = \int \phi dQ_0 \leq \alpha \ , \tag{A.10}$$

by (A.8), ϕ being a level α test of (14), and 0 trivially satisfying $\Pi_{T^\perp}(0) = 0$. In particular, $\bar{\phi}$ is a level α test based on a single observation of Z of the null hypothesis $Z \sim \bigotimes_{k=1}^{d_{T^\perp}} \Phi$ against the alternative hypothesis that $Z \sim \bigotimes_{k=1}^{d_{T^\perp}} \Phi(\cdot - h_k)$ for some $(h_1, \ldots, h_{d_{T^\perp}})' = h \neq 0$. By Problem 8.29 in Lehmann and Romano (2005), the maximin power for such a hypothesis testing problem is given by

$$\inf_{\{h\in\mathbf{R}^{d_{T^\perp}}:\|h\|\geq B\}}\int\bar{\phi}d\{\bigotimes_{k=1}^{d_{T^\perp}}\Phi(\cdot-h_k)\}\leq P(\mathcal{X}^2(B)\geq\chi^2_{1-\alpha})\ . \quad \text{(A.11)}$$

Results (A.7) and (A.9), however, contradict (A.11) and the Corollary follows. ■

Proof of Corollary 3.2: By Theorem 3.1, there exists a level α test ϕ of (14) with

$$\lim_{n\to\infty}\int\phi_n dP^n_{1/\sqrt{n},g}=\int\phi dQ_g\ . \quad \text{(A.12)}$$

However, if $\mathbf{P}$ is just identified at P, then $\bar{T}(P)=L^2_0$, or equivalently $\bar{T}(P)^\perp=\{0\}$. Therefore, the null hypothesis in (14) holds for all $g\in L^2_0$, which implies $\int\phi dQ_g\leq\alpha$ for all $g\in L^2_0$, and the claim of the Corollary then follows from (A.12). ■

Proof of Lemma 3.1: First note that since $\{f_k\}_{k=1}^{d_F}$ is such that $\sum_{k=1}^{d_F}\int f_k^2dP<\infty$ by Assumption 3.1, Theorem 2.13.1 in van der Vaart and Wellner (1996) implies $\mathcal{F}$ is P-Donsker. Moreover, since $\mathcal{F}\subseteq\bar{T}(P)^\perp\subseteq L^2_0$, it also follows that $\int fdP=0$ for all $f\in\mathcal{F}$. Hence, Theorem 3.10.12 in van der Vaart and Wellner (1996) lets us conclude that for any path $t\mapsto P_{t,g}$, we have (as a process in $\ell^\infty([d_F])$) that

$$\mathbb{G}_n\overset{L_{n,g}}{\to}\mathbb{G}_0+\Delta_g\ . \quad \text{(A.13)}$$

The first claim of the Lemma then follows immediately since $t\mapsto P_{t,g}\in\mathbf{P}$ implies $g\in T(P)$ and hence $\int gfdP=0$ for all $f\in\mathcal{F}$ due to $\mathcal{F}\subseteq\bar{T}(P)^\perp$.

Next observe Lemma B.1 implies that if $t\mapsto P_{t,g}$ satisfies (26) then we must have

$$g\notin\bar{T}(P)\ . \quad \text{(A.14)}$$

Moreover, since $\bar{T}(P)$ is a linear space by Assumption 2.1(ii), Theorem 3.4.1 in Luenberger (1969) implies $g=\Pi_T(g)+\Pi_{T^\perp}(g)$. Thus, since $g\notin\bar{T}(P)$ by (A.14) we obtain

$$\int g\{\Pi_{T^\perp}(g)\}dP=\int\{\Pi_T(g)+\Pi_{T^\perp}(g)\}\{\Pi_{T^\perp}(g)\}dP=\int\{\Pi_{T^\perp}(g)\}^2dP>0\ . \quad \text{(A.15)}$$

Furthermore, since $\Pi_{T^\perp}(g)\in\bar{T}(P)^\perp$, it follows that if $\text{cl}\{\text{lin}\{\mathcal{F}\}\}=\bar{T}(P)^\perp$, then there exists an integer $K<\infty$ and a sequence of scalars $\{\alpha_k\}_{k=1}^K$ such that

$$\|\Pi_{T^\perp}(g)-\sum_{k=1}^K\alpha_kf_k\|_{L^2}<\frac{1}{2}\frac{\|\Pi_{T^\perp}(g)\|^2_{L^2}}{\|g\|_{L^2}}\ . \quad \text{(A.16)}$$

Therefore, from (A.15) and (A.16), and the Cauchy-Schwarz inequality we obtain that

$$\int g\{\sum_{k=1}^{K}\alpha_k f_k\}dP$$
$$\geq \int g\{\Pi_{T^\perp}(g)\}dP - |\int g\{\Pi_{T^\perp}(g) - \sum_{k=1}^{K}\alpha_k f_k\}dP| \geq \frac{1}{2}\|\Pi_{T^\perp}(g)\|_{L^2}^2 > 0 \ . \quad \text{(A.17)}$$

We conclude from (A.17) that $\int g f_k dP \neq 0$ for some $f_k \in \mathcal{F}$ and the Lemma follows. ■

Proof of Theorem 3.2: For the first claim, note that Assumption 3.1, Lemma 3.1, and the continuous mapping theorem imply that under any $t \mapsto P_{t,g} \in \mathbf{P}$

$$\Psi(\mathbb{G}_n) \overset{L_{n,g}}{\to} \Psi(\mathbb{G}_0) \ . \quad \text{(A.18)}$$

We further observe that $\mathbb{G}_0$ is tight in $\ell^\infty([d_F])$ by Lemma 3.1 and Radon by Lemma A.3.11 in Bogachev (1998). Next note for any $t > 0$, continuity of Ψ, $\Psi(0) = 0$, and $\Psi(b) \geq 0$ for all $b \in \ell^\infty([d_F])$ imply there is a neighborhood N_t of $0 \in \ell^\infty([d_F])$ such that $0 \leq \Psi(b) \leq t$ for all $b \in N_t$. Thus, we can conclude that

$$P(\Psi(\mathbb{G}_0) \leq t) \geq P(\mathbb{G}_0 \in N_t) > 0 \ , \quad \text{(A.19)}$$

where the final inequality follows from $0 \in \ell^\infty([d_F])$ being in the support of $\mathbb{G}_0$ by Theorem 3.6.1 in Bogachev (1998). We therefore obtain that

$$t_0 \equiv \inf\{t : P(\Psi(\mathbb{G}_0) \leq t) > 0\} = 0 \ , \quad \text{(A.20)}$$

and from (A.20) and Theorem 11.1 in Davydov et al. (1998) that the cdf of $\Psi(\mathbb{G}_0)$ is continuous everywhere except possibly at $t_0 = 0$. Since $c_{1-\alpha} > 0$ by hypothesis, the cdf of $\Psi(\mathbb{G}_0)$ is in fact continuous at $c_{1-\alpha}$, and thus we can conclude from (A.18) that

$$\lim_{n\to\infty} P^n_{1/\sqrt{n},g}(\Psi(\mathbb{G}_n) > c_{1-\alpha}) = P(\Psi(\mathbb{G}_0) > c_{1-\alpha}) = \alpha \ . \quad \text{(A.21)}$$

For the second claim of the Lemma, recall $\Delta_g \equiv \{\Delta_{g,k}\}_{k=1}^{d_F} \in \ell^\infty([d_F])$ is defined by

$$\Delta_{g,k} \equiv \int g f_k dP \ . \quad \text{(A.22)}$$

Lemma 3.1, $\{b \in \ell^\infty([d_F]) : \Psi(b) > c_{1-\alpha}\}$ being open by continuity of Ψ, and Theorem 1.3.4 in van der Vaart and Wellner (1996) then imply that

$$\liminf_{n\to\infty} P^n_{1/\sqrt{n},g}(\Psi(\mathbb{G}_n) > c_{1-\alpha}) \geq P(\Psi(\mathbb{G}_0 + \Delta_g) > c_{1-\alpha}) \ . \quad \text{(A.23)}$$

Moreover, note that since $\mathbb{G}_0$ is Radon, Lemma B.5 further implies that $-\Delta_g$ is in the support of $\mathbb{G}_0$. Hence, arguing as in (A.19) and (A.20) it follows that

$$\inf\{t : P(\Psi(\mathbb{G}_0 + \Delta_g) \leq t) > 0\} = 0 . \tag{A.24}$$

Since $\Psi(\cdot + \Delta_g) : \ell^\infty([d_F]) \to \mathbf{R}_+$ is convex, result (A.24) and a second application of Theorem 11.1 in Davydov et al. (1998) implies the cdf of $\Psi(\mathbb{G}_0 + \Delta_g)$ is continuous everywhere except possible at 0. In particular, since $c_{1-\alpha} > 0$, we obtain that

$$P(\Psi(\mathbb{G}_0 + \Delta_g) > c_{1-\alpha}) = 1 - P(\Psi(\mathbb{G}_0 + \Delta_g) < c_{1-\alpha}) . \tag{A.25}$$

Finally, we note that since $\mathbb{G}_0$ is centered and Radon, Theorem 3.6.1 in Bogachev (1998) implies its support is a separable vector subspace of $\ell^\infty([d_F])$, and hence a separable Banach subspace under $\|\cdot\|_\infty$. Since Δ_g is in the support of $\mathbb{G}_0$, we can finally conclude

$$P(\Psi(\mathbb{G}_0 + \Delta_g) < c_{1-\alpha}) < P(\Psi(\mathbb{G}_0) < c_{1-\alpha}) = 1 - \alpha \tag{A.26}$$

by Lemma B.4 and $0 \neq \Delta_g$ by Lemma 3.1. The second claim of the Theorem then follows from results (A.23), (A.25) and (A.26). ■

Proof of Theorem 4.1: We first note that since $T(P)$ is linear by Assumption 2.1(ii), and $\hat{\theta}_n$ is regular by Assumption 4.1(ii), Lemma B.6 and Theorem 5.2.3 in Bickel et al. (1993) imply θ is pathwise differentiable at P – i.e. there exists a bounded linear operator $\dot{\theta} : \bar{T}(P) \to \mathbf{B}$ such that for any $t \mapsto P_{t,g} \in \mathbf{P}$ it follows that

$$\lim_{t\to 0} \|t^{-1}\{\theta(P_{t,g}) - \theta(P)\} - \dot{\theta}(g)\|_{\mathbf{B}} = 0 . \tag{A.27}$$

Then note that for any $b^* \in \mathbf{B}^*$, $b^* \circ \dot{\theta} : \bar{T}(P) \to \mathbf{R}$ is a continuous linear functional. Hence, since $\bar{T}(P)$ is a Hilbert space under $\|\cdot\|_{L^2}$, the Riesz representation theorem implies there exists a $\dot{\theta}_{b^*} \in \bar{T}(P)$ such that for all $g \in \bar{T}(P)$ we have that

$$b^*(\dot{\theta}(g)) = \int \dot{\theta}_{b^*} g dP . \tag{A.28}$$

Moreover, since $\hat{\theta}_n$ is an asymptotically linear estimator of $\theta(P)$, it follows that $b^*(\hat{\theta}_n)$ is an asymptotically linear estimator of $b^*(\theta(P))$ with influence function $b^* \circ \nu$. Proposition 3.3.1 in Bickel et al. (1993) then implies that for all $g \in \bar{T}(P)$

$$\int \{\dot{\theta}_{b^*} - b^* \circ \nu\} g dP = 0 . \tag{A.29}$$

Analogously, if $\tilde{\theta}_n : \{X_i\}_{i=1}^n \to \mathbf{B}$ is an asymptotically linear regular estimator satisfying

$$\sqrt{n}\{\tilde{\theta}_n - \theta(P)\} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\nu}(X_i) + o_p(1) \tag{A.30}$$

for some $\tilde{\nu} : \mathbf{R}^{d_x} \to \mathbf{B}$, then it follows that (A.29) also holds with $\tilde{\nu}$ in place of ν, yielding

$$\int \{b^* \circ \nu - b^* \circ \tilde{\nu}\} g dP = 0 \tag{A.31}$$

for all $g \in \bar{T}(P)$. If P is just identified by $\mathbf{P}$, however, then $\bar{T}(P) = L_0^2$, which implies

$$P(b^*(\nu(X_i)) = b^*(\tilde{\nu}(X_i))) = 1 \ , \tag{A.32}$$

for all $b^* \in \mathbf{B}^*$ by result (A.31). Thus, since $\|b\|_{\mathbf{B}} = \sup_{\|b^*\|_{\mathbf{B}^*}=1} b^*(b)$, see for example Lemma 6.10 in Aliprantis and Border (2006), we can conclude by (A.32) that

$$P(\|\nu(X_i) - \tilde{\nu}(X_i)\|_{\mathbf{B}} = 0) = P(\sup_{\|b^*\|_{\mathbf{B}^*}=1} b^*(\nu(X_i) - \tilde{\nu}(X_i)) = 0) = 1 \ . \tag{A.33}$$

The first claim of the Theorem then follows from (A.30), (A.33), and Assumption 4.1(ii).

In order to establish the second claim of the Theorem, note that if P is overidentified by $\mathbf{P}$, then there exists a $0 \neq g^{\perp} \in \bar{T}(P)^{\perp}$. For an arbitrary $0 \neq \tilde{b} \in \mathbf{B}$ we then define

$$\tilde{\theta}_n \equiv \hat{\theta}_n + \tilde{b} \times \{\frac{1}{n}\sum_{i=1}^n g^{\perp}(X_i)\} \ . \tag{A.34}$$

Since $\hat{\theta}_n$ is asymptotically linear by Assumption 4.1(ii) we then immediately conclude

$$\sqrt{n}\{\tilde{\theta}_n - \theta(P)\} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \{\nu(X_i) + \tilde{b} \times g^{\perp}(X_i)\} + o_p(1) \ . \tag{A.35}$$

Setting $\tilde{\nu}(X_i) \equiv \nu(X_i) + \tilde{b} \times g^{\perp}(X_i)$, we obtain for any $b^* \in \mathbf{B}^*$ that $b^*(\tilde{\nu}) = \{b^*(\nu) + b^*(\tilde{b}) \times g^{\perp}\} \in L_0^2$ since $b^*(\nu) \in L_0^2$ by Assumption 4.1(ii) and $g^{\perp} \in \bar{T}(P)^{\perp} \subseteq L_0^2$. Hence, (A.35) implies $\tilde{\theta}_n$ is indeed asymptotically linear and its influence function equals $\tilde{\nu}$. Moreover, by Lemma B.6, $(\sqrt{n}\{\hat{\theta}_n - \theta(P)\}, \frac{1}{\sqrt{n}}\sum_{i=1}^n g^{\perp}(X_i))$ converge jointly in distribution in $\mathbf{B} \times \mathbf{R}$ under $\bigotimes_{i=1}^n P$ and hence by the continuous mapping theorem

$$\sqrt{n}\{\tilde{\theta}_n - \theta(P)\} = \sqrt{n}\{\hat{\theta}_n - \theta(P)\} + \tilde{b} \times \{\frac{1}{\sqrt{n}}\sum_{i=1}^n g^{\perp}(X_i)\} \xrightarrow{L} \mathbb{Z} \tag{A.36}$$

on $\mathbf{B}$ under $\bigotimes_{i=1}^n P$ for some tight Borel random variable $\mathbb{Z}$. In addition, we have

$$\sqrt{n}\{\hat{\theta}_n - \tilde{\theta}_n\} = \tilde{b} \times \{\frac{1}{\sqrt{n}}\sum_{i=1}^n g^{\perp}(X_i)\} \xrightarrow{L} \Delta \tag{A.37}$$

by the central limit and continuous mapping theorems. Further note that since $\tilde{b} \neq 0$, we trivially have $\Delta \neq 0$ in $\mathbf{B}$ because $b^*(\Delta) \sim N(0, \|b^*(\tilde{b})g^{\perp}\|_{L^2}^2)$.

Thus, to conclude the proof it only remains to show that $\tilde{\theta}_n$ is regular. To this end let $t \mapsto P_{t,g} \in \mathbf{P}$ and set $L_{n,g} \equiv \bigotimes_{i=1}^n P_{1/\sqrt{n},g}$. By Lemma 25.14 in van der Vaart (1998)

$$\sum_{i=1}^n \log(\frac{dP_{1/\sqrt{n},g}}{dP}(X_i)) = \frac{1}{\sqrt{n}}\sum_{i=1}^n g(X_i) - \frac{1}{2}\int g^2 dP + o_p(1) \tag{A.38}$$

under $\bigotimes_{i=1}^n P$, and thus Example 3.10.6 in van der Vaart and Wellner (1996) implies $\bigotimes_{i=1}^n P$ and $\bigotimes_{i=1}^n P_{1/\sqrt{n},g}$ are mutually contiguous. Moreover, since $\tilde{\theta}_n$ is asymptotically linear, Lemma B.6 implies $(\sqrt{n}\{\tilde{\theta}_n - \theta(P)\}, \frac{1}{\sqrt{n}}\sum_{i=1}^n g(X_i))$ converge jointly in $\mathbf{B} \times \mathbf{R}$. Thus, by (A.38) and Lemma A.8.6 in Bickel et al. (1993) we obtain that

$$\sqrt{n}\{\tilde{\theta}_n - \theta(P)\} \stackrel{L_{n,g}}{\to} \mathbb{Z}_g \tag{A.39}$$

for some tight Borel $\mathbb{Z}_g$ on $\mathbf{B}$. Hence, combining results (A.27) and (A.39) implies

$$\sqrt{n}\{\tilde{\theta}_n - \theta(P_{1/\sqrt{n},g})\} \stackrel{L_{n,g}}{\to} \mathbb{Z}_g + \dot{\theta}(g) \ . \tag{A.40}$$

Next note for any $b^* \in \mathbf{B}^*$, results (A.35), (A.38) and the central limit theorem yield

$$\left(\begin{array}{c} \sqrt{n}\{b^*(\tilde{\theta}_n) - b^*(\theta(P))\} \\ \sum_i \log(\frac{dP_{1/\sqrt{n},g}}{dP}(X_i)) \end{array} \right) \stackrel{L}{\to} N\Big(\Big[\begin{array}{c} 0 \\ -\frac{1}{2}\int g^2 dP \end{array} \Big], \Sigma\Big) \tag{A.41}$$

under $\bigotimes_{i=1}^n P$, where since $\int g g^{\perp} dP = 0$ due to $g \in T(P)$ and $g^{\perp} \in \bar{T}(P)^{\perp}$, we have

$$\Sigma = \Big[\begin{array}{cc} \int (b^*(\nu) + b^*(\tilde{b})g^{\perp})^2 dP & \int b^*(\nu) g dP \\ \int b^*(\nu) g dP & \int g^2 dP \end{array} \Big] \ . \tag{A.42}$$

Notice, however, that by results (A.28) and (A.29) it follows that $\int b^*(\nu) g dP = b^*(\dot{\theta}(g))$. Therefore, results (A.41), (A.42), and Lemma A.9.3 in Bickel et al. (1993) together imply

$$\sqrt{n}\{b^*(\tilde{\theta}_n) - b^*(\theta(P_{1/\sqrt{n},g}))\} \stackrel{L_{n,g}}{\to} N(0, \int (b^*(\nu) + b^*(\tilde{b})g^{\perp})^2 dP) \ . \tag{A.43}$$

Define $\zeta_{b^*}(X_i) \equiv \{b^*(\nu(X_i)) + b^*(\tilde{b})g^{\perp}(X_i)\}$, and for any finite collection $\{b_k^*\}_{k=1}^K \subset \mathbf{B}^*$ let $(\mathbb{W}_{b_1^*}, \ldots, \mathbb{W}_{b_K^*})$ denote a multivariate normal vector with $E[\mathbb{W}_{b_k^*}] = 0$ for all $1 \leq k \leq K$ and $E[\mathbb{W}_{b_k^*}\mathbb{W}_{b_j^*}] = E[\zeta_{b_k^*}(X_i)\zeta_{b_j^*}(X_i)]$ for any $1 \leq j \leq k \leq K$. Letting $C_b(\mathbf{R}^K)$ denote the set of continuous and bounded functions on $\mathbf{R}^K$, we then obtain from (A.39), (A.43), the Cramer-Wold device, and the continuous mapping theorem that

$$E[f(b_1^*(\mathbb{Z}_g + \dot{\theta}(g)), \ldots, b_K^*(\mathbb{Z}_g + \dot{\theta}(g)))] = E[f(b_1^*(\mathbb{W}_{b_1^*}), \ldots, b_K^*(\mathbb{W}_{b_K^*}))] \ , \tag{A.44}$$

for any $f \in C_b(\mathbf{R}^K)$. Since $\mathcal{G} \equiv \{f \circ (b_1^*, \ldots, b_K^*) : f \in C_b(\mathbf{R}^K), \{b_k^*\}_{k=1}^K \subset \mathbf{B}^*, 1 \leq K < \infty\}$ is a vector lattice that separates points in $\mathbf{B}$, it follows from Lemma 1.3.12 in van der Vaart and Wellner (1996) that there is a unique tight Borel measure $\mathbb{W}$ on $\mathbf{B}$

satisfying (A.44). In particular, since the right hand side of (A.44) does not depend on g, we conclude the law of $\mathbb{Z}_g + \dot{\theta}(g)$ is constant in g, establishing the regularity of $\tilde{\theta}_n$. ■

Proof of Theorem 4.2: Let $F : \mathbf{B} \to \ell^\infty(\mathbb{U})$ be given by $F(b) = b^* \mapsto b^*(b)$ for any $b \in \mathbf{B}$. Since $\mathbb{U} \subset \mathbf{B}^*$ is norm bounded, and in addition we have

$$\|F(b_1) - F(b_2)\|_\infty = \sup_{b^* \in \mathbb{U}} |b^*(b_1) - b^*(b_2)| \leq \sup_{b^* \in \mathbb{U}} \|b^*\|_{\mathbf{B}^*} \times \|b_1 - b_2\|_{\mathbf{B}} \tag{A.45}$$

it follows that F is continuous, in fact Lipschitz. Hence, Assumption 4.2(ii) and the continuous mapping theorem imply $\sqrt{n}\{\hat{\theta}_n - \tilde{\theta}_n\}$ converges in distribution in $\mathbf{B}$, while the continuity of F and a second application of the continuous mapping theorem yield

$$\hat{\mathbb{G}}_n \stackrel{L}{\to} \mathbb{G}_0 \tag{A.46}$$

under $\bigotimes_{i=1}^n P$ on $\ell^\infty(\mathbb{U})$ for some tight $\mathbb{G}_0$. Next, define a process $\bar{\mathbb{G}}_n$ on $\ell^\infty(\mathbb{U})$ by

$$\bar{\mathbb{G}}_n(b^*) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n b^*(\nu(X_i) - \tilde{\nu}(X_i)) \ , \tag{A.47}$$

and note that since $\hat{\theta}_n$ and $\tilde{\theta}_n$ are asymptotically linear by Assumption 4.2(i) we have

$$\sup_{b^* \in \mathbb{U}} |\hat{\mathbb{G}}_n(b^*) - \bar{\mathbb{G}}_n(b^*)|$$
$$\leq \sup_{b^* \in \mathbb{U}} \|b^*\|_{\mathbf{B}^*} \times \|\sqrt{n}\{\hat{\theta}_n - \tilde{\theta}_n\} - \frac{1}{\sqrt{n}} \sum_{i=1}^n \{\nu(X_i) - \tilde{\nu}(X_i)\}\|_{\mathbf{B}} = o_p(1) \ . \tag{A.48}$$

In particular, letting $\mathcal{F} \equiv \{f : \mathbf{R}^{d_x} \to \mathbf{R} : f = b^*(\nu - \tilde{\nu}) \text{ for some } b^* \in \mathbb{U}\}$, we conclude from (A.46) and (A.48) that $\bar{\mathbb{G}}_n \stackrel{L}{\to} \mathbb{G}_0$ under $\bigotimes_{i=1}^n P$, or equivalently, that $\mathcal{F}$ is P-Donsker. Moreover, since $\bigotimes_{i=1}^n P$ and $\bigotimes_{i=1}^n P_{1/\sqrt{n},g}$ are mutually contiguous by (A.38) and Corollary 12.3.1 in Lehmann and Romano (2005), it follows from (A.48) that $\hat{\mathbb{G}}_n = \bar{\mathbb{G}}_n + o_p(1)$ on $\ell^\infty(\mathbb{U})$ under $\bigotimes_{i=1}^n P_{1/\sqrt{n},g}$ as well. We therefore obtain that

$$\hat{\mathbb{G}}_n = \bar{\mathbb{G}}_n + o_p(1) \stackrel{L_{n,g}}{\to} \mathbb{G}_0 + \Delta_g \tag{A.49}$$

by Theorem 3.10.12 in van der Vaart and Wellner (1996), which verifies (41).

The claim that $b^*(\nu - \tilde{\nu}) \in \bar{T}(P)^\perp$ has already been shown in the proof of Theorem 4.1, see result (A.31). Since $g \in T(P)$ whenever $t \mapsto P_{t,g} \in \mathbf{P}$ it is then immediate that $\Delta_g(b^*) = \int b^*(\nu - \tilde{\nu}) g dP = 0$ for all $b^* \in \mathbf{B}^*$, and hence $\Delta_g = 0$ in $\ell^\infty(\mathbb{U})$. On the other hand, if $t \mapsto P_{t,g}$ satisfies (42), then Lemma B.1 implies $g \notin \bar{T}(P)$. The fact that $\Delta_g(b^*) \neq 0$ for some $b^* \in \mathbb{U}$ can then be established using the same arguments as in (A.15)-(A.17), which establishes the second claim of the Theorem. ■

Proof of Corollary 4.1: Let $\mathcal{F} \equiv \{b_k^* \circ (\nu - \tilde{\nu})\}_{k=1}^{d_F}$ and note that by Theorem

4.2 $\mathcal{F} \subset \bar{T}(P)^{\perp}$ while $\sum_{k=1}^{d_F} \int (b_k^* \circ (\nu - \tilde{\nu}))^2 dP < \infty$ by hypothesis. Thus, $\mathcal{F}$ satisfies Assumption 3.1 and moreover by (24), (A.47), and (A.48) it follows that $\hat{\mathbb{G}}_n = \mathbb{G}_n + o_p(1)$. The Corollary then follows by arguments identical to those in Theorem 3.2. ■

Proof of Proposition 5.1: Throughout, we let $L^\infty \equiv \{f : |f| \text{ is bounded } P\text{-.a.s.}\}$, set L^∞_W and L^∞_Z to denote the subsets of L^∞ depending only on W and Z respectively, and recall L^2_W and L^2_Z are analogously defined. In addition, defining the set $\mathcal{V}$ by

$$\mathcal{V} \equiv \{f \in L^2 : f(X) = \{\rho(Y, h_P(Z))v(W) + C\} \ P\text{-a.s. for some } v : \mathbf{R}^{d_w} \to \mathbf{R},\ C \in \mathbf{R}\} \ ,$$

we then set the desired subset $\mathcal{D}$ with which we will work to be given by $\mathcal{D} \equiv L^\infty \setminus \mathcal{V}$. It is then immediate that $\mathcal{D}$ is a subset of L^2 since $\mathcal{D} \subseteq L^\infty \subset L^2$. To establish that $\mathcal{D}$ is dense in L^2 with respect to $\|\cdot\|_{L^2}$, note that the fact that $\text{Var}\{\rho(Y, h_P(Z))|W\} > 0$ almost surely implies $\rho(Y, h_P(Z))$ is not a measurable function of W. Furthermore, also note that for any functions $g \in L^\infty_W$ and $f \in L^\infty$, and $\epsilon_n \downarrow 0$ it follows that

$$\lim_{n\to\infty} \|\{g\epsilon_n + f\} - f\|_{L^2} = 0 \ , \tag{A.50}$$

and moreover setting $g = 0$ if $f \notin \mathcal{V}$ and $g \in L^\infty_W$ nonconstant if $f \in \mathcal{V}$ we conclude that $g\epsilon_n + f \in L^\infty \setminus \mathcal{V}$ since $\mathcal{V}$ is closed under addition and as argued nonconstant $g \in L^2_W$ do not belong to $\mathcal{V}$. In particular, it follows from (A.50) that $\mathcal{D}$ is dense in L^∞ with respect to $\|\cdot\|_{L^2}$ and hence also in L^2 by denseness of L^∞ in L^2.

While we avoided stating an explicit formulation for Ω^*_f in the main text for ease of exposition, it is now necessary to characterize it for all $f \in \mathcal{D}$. To this end, we let $\Sigma_f \equiv \text{Var}\{f(X)\}$, and following Ai and Chen (2012) for any $f \in \mathcal{D}$ define

$$\Sigma_2(W) \equiv E[\{\rho(Y, h_P(Z))\}^2|W] \tag{A.51}$$

$$\Lambda(W) \equiv E[f(X)\rho(Y, h_P(Z))|W]\{\Sigma_2(W)\}^{-1} \tag{A.52}$$

$$\Sigma_1 \equiv \text{Var}\{f(X) - \Lambda(W)\rho(Y, h_P(Z))\} \ . \tag{A.53}$$

Further notice that: (i) $\{\Sigma_2(W)\}^{-1} \in L^\infty$ since $P(\Sigma_2(W) > \eta) = 1$ for some $\eta > 0$ by hypothesis, (ii) $\Lambda(W) \in L^2$ due to $f \in \mathcal{D} \subset L^\infty$, $\{\Sigma_2(W)\}^{-1} \in L^\infty$, Assumption 5.1(i), and Jensen's inequality, (iii) $\Sigma_1 > 0$ since $f \notin \mathcal{V}$, and (iv) by direct calculation:

$$\Sigma_1 = \Sigma_f - E[\{\Lambda(W)\}^2\Sigma_2(W)] \ . \tag{A.54}$$

Hence, in our context the Fisher norm of a $s \in L^2_Z$ is (see eq. (4) in Ai and Chen (2012)):

$$\|s\|^2_w \equiv E[\{\nabla m(W, h_P)[s]\}^2\{\Sigma_2(W)\}^{-1}] + \{\Sigma_1\}^{-1}\{E[\Lambda(W)\nabla m(W, h_P)[s]]\}^2 \ , \tag{A.55}$$

and note $\|s\|^2_w < \infty$ for any $s \in L^2_Z$ since $\{\Sigma_2(W)\}^{-1} \in L^\infty_W$, $\Lambda(W) \in L^2$, and $\nabla m(W, h_P)[s] \in L^2_W$ by Assumption 5.1(ii). Letting $\mathcal{W}$ denote the closure of L^2_Z under

$\|\cdot\|_w$, Theorem 2.1 in Ai and Chen (2012) then establishes that

$$\{\Omega_f^*\}^{-1} = \inf_{s\in\mathcal{W}}\Big\{\{\Sigma_1\}^{-1}\{1+E[\Lambda(W)\nabla m(W,h_P)[s]]\}^2 + E[\{\Sigma_2(W)\}^{-1}\{\nabla m(W,h_P)[s]\}^2]\Big\} . \quad (A.56)$$

It is convenient for our purposes, however, to exploit the structure of our problem to further simplify (A.56). To this end, note that by (A.55) and Cauchy-Schwarz

$$|E[\{\Sigma_2(W)\}^{-1}\{\nabla m(W,h_P)[s_1]\}^2] - E[\{\Sigma_2(W)\}^{-1}\{\nabla m(W,h_P)[s_2]\}^2]| \leq \{E[\{\Sigma_2(W)\}^{-1}\{\nabla m(W,h_P)[s_1+s_2]\}^2]\}^{\frac{1}{2}} \times \|s_1-s_2\|_w . \quad (A.57)$$

Similarly, by (A.52), (A.55), and the Cauchy Schwarz inequality we also obtain

$$|E[\Lambda(W)\nabla m(W,h_P)[s_1]] - E[\Lambda(W)\nabla m(W,h_P)[s_2]]| \leq \{E[\{\Sigma_2(W)\}^{-1}\{E[f(X)\rho(Y,h_P(Z))|W]\}^2]\}^{\frac{1}{2}} \times \|s_1-s_2\|_w , \quad (A.58)$$

where we note $E[f(X)\rho(Y,h_P(Z))|W] \in L^2$ due to Assumption 5.1(i) and $f \in L^\infty$. Thus, results (A.57) and (A.58) imply that the objective in (A.56) is continuous in $\|\cdot\|_w$. Therefore, since $\mathcal{W}$ is the completion of L_Z^2 under $\|\cdot\|_w$, it then follows that

$$\{\Omega_f^*\}^{-1} = \inf_{s\in L_Z^2}\Big\{\{\Sigma_1\}^{-1}\{1+E[\Lambda(W)\nabla m(W,h_P)[s]]\}^2 + E[\{\Sigma_2(W)\}^{-1}\{\nabla m(W,h_P)[s]\}^2]\Big\} . \quad (A.59)$$

Next, recall $\mathcal{R} = \{f \in L_W^2 : f(W) = \nabla m(W,h_P)[s] \text{ for some } s \in L_Z^2\}$, $\bar{\mathcal{R}}$ denotes its closure in L_W^2, and note $\bar{\mathcal{R}}$ is a subspace of L_W^2 by linearity of $\nabla m(W,h_P)$. By (A.59),

$$\begin{aligned}\{\Omega_f^*\}^{-1} &= \inf_{r\in\bar{\mathcal{R}}}\{\{\Sigma_1\}^{-1}\{1+E[\Lambda(W)r(W)]\}^2 + E[\{\Sigma_2(W)\}^{-1}\{r(W)\}^2]\} \\ &= \min_{r\in\bar{\mathcal{R}}}\{\{\Sigma_1\}^{-1}\{1+E[\Lambda(W)r(W)]\}^2 + E[\{\Sigma_2(W)\}^{-1}\{r(W)\}^2]\} , \end{aligned} \quad (A.60)$$

where attainment follows from $\bar{\mathcal{R}}$ being a vector space, the criterion being convex and diverging to infinity as $\|r\|_{L^2} \uparrow \infty$, and Proposition 38.15 in Zeidler (1984). In particular, note that if r_0 is the minimizer of (A.60), then for any $\delta \in \bar{\mathcal{R}}$ we must have

$$E[\delta(W)\{r_0(W)\{\Sigma_2(W)\}^{-1} + \Lambda(W)\{\Sigma_1\}^{-1}\{1+E[\Lambda(W)r_0(W)]\}\}] = 0 . \quad (A.61)$$

Next, we aim to solve the optimization problem in (A.60) under the hypothesis that $\bar{\mathcal{R}} = L_W^2$. In that case, (A.61) must hold for all $\delta \in L_W^2$, which implies

$$r_0(W) = -\{\Sigma_1\}^{-1}\{1+E[\Lambda(W)r_0(W)]\}\Lambda(W)\Sigma_2(W) . \quad (A.62)$$

It is evident from (A.62) that $r_0(W) = -\Lambda(W)\Sigma_2(W)C_0$ for some $C_0 \in \mathbf{R}$, and by plugging into (A.62) we can solve for C_0 and exploit (A.54) to obtain

$$r_0(W) = -\{\Sigma_f\}^{-1}\Lambda(W)\Sigma_2(W) . \tag{A.63}$$

Thus, combining (A.60) and (A.63), and repeatedly exploiting (A.54) we conclude

$$\begin{aligned}\{\Omega_f^*\}^{-1} &= \{\Sigma_1\}^{-1}\{1 - \{\Sigma_f\}^{-1}E[\{\Lambda(W)\}^2\Sigma_2(W)]\}^2 + \{\Sigma_f\}^{-2}E[\{\Lambda(W)\}^2\Sigma_2(W)] \\ &= \Sigma_1\{\Sigma_f\}^{-2} + \{\Sigma_f\}^{-2}E[\{\Lambda(W)\}^2\Sigma_2(W)] = \{\Sigma_f\}^{-1} ,\end{aligned} \tag{A.64}$$

or equivalently, that $\Omega_f^* = \Sigma_f$. While (A.64) was derived under the supposition that $\bar{\mathcal{R}} = L_W^2$, we note that since $\bar{\mathcal{R}} \subseteq L_W^2$, the minimum in (A.60) is attained, and $r_0(W) = -\{\Sigma_f\}^{-1}\Lambda(W)\Sigma_2(W)$ is the unique minimizer on L_W^2, we obtain from (A.63) that

$$\Omega_f^* = \Sigma_f \text{ if and only if } -\{\Sigma_f\}^{-1}\Lambda(W)\Sigma_2(W) \in \bar{\mathcal{R}} . \tag{A.65}$$

Since result (A.65) holds for all $f \in \mathcal{D}$ and $\bar{\mathcal{R}}$ is a vector space, (A.52) yields

$$\Omega^* = \Sigma_f \ \forall f \in \mathcal{D} \text{ if and only if } E[f(X)\rho(Y, h_P(Z))|W] \in \bar{\mathcal{R}} \ \forall f \in \mathcal{D} . \tag{A.66}$$

Also note that if $\|f_n - f\|_{L^2} = o(1)$, then by the Cauchy-Schwarz and Jensen's inequalities

$$\begin{aligned}\lim_{n\to\infty} &E[\{E[f_n(X)\rho(Y, h_P(Z))|W] - E[f(X)\rho(Y, h_P(Z))|W]\}^2] \\ &\leq \lim_{n\to\infty} E[\{f_n(X) - f(X)\}^2\Sigma_2(W)] = 0 ,\end{aligned} \tag{A.67}$$

where the final equality follows from $\Sigma_2(W) \in L^\infty$. Therefore, since as argued $\mathcal{D}$ is a dense subset of L^2 and in addition $\bar{\mathcal{R}}$ is closed under $\|\cdot\|_{L^2}$, (A.66) implies that in fact

$$\Omega^* = \Sigma_f \ \forall f \in \mathcal{D} \text{ if and only if } E[f(X)\rho(Y, h_P(Z))|W] \in \bar{\mathcal{R}} \ \forall f \in L^2 . \tag{A.68}$$

Moreover, for any $g_0 \in L_W^\infty$ we may set $f_0(X) \equiv g_0(W)\rho(Y, h_P(Z)\{\Sigma_2(W)\}^{-1}$ which we note satisfies $f_0 \in L^2$ by $\{\Sigma_2(W)\}^{-1} \in L^\infty$ and Assumption 5.1(i). In addition,

$$E[f_0(X)\rho(Y, h_P(Z))|W] = g_0(W)\{\Sigma_2(W)\}^{-1}E[\{\rho(Y, h_P(Z))\}^2|W] = g_0(W) , \tag{A.69}$$

and hence since $g_0 \in L_W^\infty$ was arbitrary, it follows that if $E[f(X)\rho(Y, h_P(Z))|W] \in \bar{\mathcal{R}}$ for all $f \in L^2$, then $L_W^\infty \subseteq \bar{\mathcal{R}}$. However, since $\bar{\mathcal{R}}$ is closed under $\|\cdot\|_{L^2}$, we can also conclude that if $L_W^\infty \subseteq \bar{\mathcal{R}}$, then $L_W^2 = \bar{\mathcal{R}}$ and therefore from result (A.68) finally obtain

$$\Omega^* = \Sigma_f \ \forall f \in \mathcal{D} \text{ if and only if } L_W^2 = \bar{\mathcal{R}} , \tag{A.70}$$

which establishes the claim of the Proposition. ■

Proof of Proposition 5.2: We first suppose that $\bar{\mathcal{R}} = L^2_W$, and define $f_0 \in L^2_W$ by

$$f_0(W) \equiv 1\{d_0(W) = 0\} \ . \tag{A.71}$$

Next observe that since $\bar{\mathcal{R}} = L^2_W$ by hypothesis, it follows that $f_0 \in \bar{\mathcal{R}}$ and therefore

$$0 = \inf_{s \in L^2_Z} E[\{\nabla m(W, h_P)[s] - f_0(W)\}^2] \geq E[\{f_0(W)\}^2] = P(d_0(W) = 0) \ , \tag{A.72}$$

where we exploited Assumption 5.2(ii) and the inequality applies to the infimum since it applies for any $s \in L^2_Z$, while the final equality follows from definition of f_0. Hence, we conclude that if $\bar{\mathcal{R}} = L^2_W$, then $P(d_0(W) \neq 0) = 1$. Furthermore, notice that for an arbitrary $f \in L^2_W$ we have $d_0 f \in L^2_W$ by Assumption 5.2(ii), and hence

$$0 = \inf_{s \in L^2_Z} E[\{\nabla m(W, h_P)[s] - d_0(W) f(W)\}^2] \tag{A.73}$$

$$= \min_{s \in L^2_Z} E[\{d_0(W)\}^2 \{s(Z) - f(W)\}^2] \ , \tag{A.74}$$

where the first equality follows from $\bar{\mathcal{R}} = L^2_W$, while attainment in (A.74) results from the criterion being convex and diverging to infinity as $\|s\|_{L^2} \uparrow \infty$, and Proposition 38.15 in Zeidler (1984). Thus, we conclude from (A.72) and (A.74) that for every $f \in L^2_W$ there exists and $r_f \in L^2_Z$ such that $P(r_f(Z) = f(W)) = 1$. In particular, writing $V = (V^{(1)}, \ldots, V^{(d_v)})$ it follows from Assumption 5.2(i) that each coordinate $V^{(j)} \in L^2_W$ and hence by (A.74) that $P(E[V^{(j)}|Z] = V^{(j)}) = 1$ for any $1 \leq j \leq d_v$. We thus conclude from results (A.72) and (A.74) that if $\bar{\mathcal{R}} = L^2_W$, then condition (62) must hold.

Next, suppose instead that condition (62) holds. Then note that for any $f \in L^2_W$

$$P(f((Z, V)) = f((Z, E[V|Z]))) = 1 \tag{A.75}$$

due to $P(E[V|Z] = V) = 1$, and thus we may identify L^2_W with L^2_Z. Hence, interpreting the domain of $\nabla m(W, h_P)$ as L^2_W in place of L^2_Z, it follows from Assumption 5.2(ii) that $\nabla m(W, h_P) : L^2_W \to L^2_W$ is self adjoint. Thus, Theorem 6.6.3 in Luenberger (1969) implies $\bar{\mathcal{R}} = L^2_W$ if and only if $\nabla m(W, h_P)$ is injective. However, injectivity of $\nabla m(W, h_P) : L^2_W \to L^2_W$ is equivalent to $P(d_0(W) \neq 0) = 1$, and therefore $\bar{\mathcal{R}} = L^2_W$. ■

Appendix B - Proof of Auxiliary Results

Lemma B.1. *If $t \mapsto P_{t,g_1}$ and $t \mapsto P_{t,g_2}$ are arbitrary paths, then it follows that:*

$$\limsup_{n \to \infty} \int |dP^n_{1/\sqrt{n}, g_1} - dP^n_{1/\sqrt{n}, g_2}| \leq 2\{1 - \exp\{-\frac{1}{4}\|g_1 - g_2\|^2_{L^2}\}\}^{1/2} \ . \tag{B.1}$$

PROOF: First observe that since $t \mapsto P_{t,g_1}$ and $t \mapsto P_{t,g_2}$ satisfy (1), we must have

$$\lim_{n\to\infty} n \int [dP^{1/2}_{1/\sqrt{n},g_1} - dP^{1/2}_{1/\sqrt{n},g_2}]^2 = \frac{1}{4}\int [g_1 dP^{1/2} - g_2 dP^{1/2}]^2 = \frac{1}{4}\|g_1 - g_2\|^2_{L^2} \ . \quad \text{(B.2)}$$

Moreover, by Theorem 13.1.2 in Lehmann and Romano (2005) we can also conclude

$$\frac{1}{2}\int |dP^n_{1/\sqrt{n},g_1} - dP^n_{1/\sqrt{n},g_2}| \leq \{1 - [\int \{dP^n_{1/\sqrt{n},g_1}\}^{1/2}\{dP^n_{1/\sqrt{n},g_2}\}^{1/2}]^2\}^{1/2}$$
$$= \{1 - [\int dP^{1/2}_{1/\sqrt{n},g_1} dP^{1/2}_{1/\sqrt{n},g_2}]^{2n}\}^{1/2} = \{1 - [1 - \frac{1}{2}\int [dP^{1/2}_{1/\sqrt{n},g_1} - dP^{1/2}_{1/\sqrt{n},g_2}]^2]^{2n}\}^{1/2} \ , \quad \text{(B.3)}$$

where in the first equality we exploited $P^n_{1/\sqrt{n},g_1}$ and $P^n_{1/\sqrt{n},g_2}$ are product measures, while the second equality follows from direct calculation. Thus, by (B.2) and (B.3)

$$\limsup_{n\to\infty} \frac{1}{2}\int |dP^n_{1/\sqrt{n},g_1} - dP^n_{1/\sqrt{n},g_2}|$$
$$\leq \limsup_{n\to\infty}\{1 - [1 - \frac{1}{2n}\int n[dP^{1/2}_{1/\sqrt{n},g_1} - dP^{1/2}_{1/\sqrt{n},g_2}]^2]^{2n}\}^{1/2}$$
$$= \{1 - \exp\{-\frac{1}{4}\|g_1 - g_2\|^2_{L^2}\}\}^{1/2} \ , \quad \text{(B.4)}$$

which establishes the claim of the Lemma. ■

Lemma B.2. *Let $\{P_n\}$, $\{Q_n\}$, $\{V_n\}$ be probability measures defined on a common space. If $\{dQ_n/dP_n\}$ is asymptotically tight under P_n and $\int |dP_n - dV_n| = o(1)$, then*

$$|\frac{dQ_n}{dP_n} - \frac{dQ_n}{dV_n}| \stackrel{P_n}{\to} 0 \ . \quad \text{(B.5)}$$

PROOF: Throughout let $\mu_n = P_n + Q_n + V_n$, note μ_n dominates P_n, Q_n, and V_n, and set $p_n \equiv dP_n/d\mu_n$, $q_n \equiv dQ_n/d\mu_n$, and $v_n \equiv dV_n/d\mu_n$. We then obtain

$$\int |\frac{dP_n}{dV_n} - 1| dV_n = \int |\frac{p_n}{v_n} - 1| v_n d\mu_n = \int_{v_n>0} |\frac{p_n}{v_n} - \frac{v_n}{v_n}| v_n d\mu_n$$
$$\leq \int |p_n - v_n| d\mu_n = \int |dP_n - dV_n| = o(1) \ , \quad \text{(B.6)}$$

where the second to last equality follows by definition, and the final equality by assumption. Hence, by (B.6) and Markov's inequality we obtain $dP_n/dV_n \stackrel{V_n}{\to} 1$. Moreover, since $\int |dV_n - dP_n| = o(1)$ implies $\{P_n\}$ and $\{V_n\}$ are mutually contiguous, we conclude

$$\frac{dP_n}{dV_n} \stackrel{P_n}{\to} 1 \ . \quad \text{(B.7)}$$

Next observe that for any continuous and bounded function $f : \mathbf{R} \to \mathbf{R}$ we have that

$$\int f(\frac{dQ_n}{dP_n} - \frac{dQ_n}{dV_n})dP_n = \int f(\frac{q_n}{p_n} - \frac{q_n}{v_n})p_n d\mu_n$$
$$= \int_{p_n>0} f(\frac{q_n}{p_n}(1 - \frac{p_n}{v_n}))p_n d\mu_n = \int f(\frac{dQ_n}{dP_n}(1 - \frac{dP_n}{dV_n}))dP_n \to f(0) \ , \quad \text{(B.8)}$$

where the final result follows from (B.7), dQ_n/dP_n being asymptotically tight under P_n and continuity and boundedness of f. Since (B.8) holds for any continuous and bounded f, we conclude $dQ_n/dP_n - dQ_n/dV_n$ converges in law (under P_n) to zero, and hence also in P_n probability, which establishes (B.5). ■

Lemma B.3. *Let $H \subseteq L_0^2$, assume that for each $g \in H$ there is a path $t \mapsto P_{t,g}$ such that* (1) *holds, and for $\mathcal{B}$ the Borel σ-algebra on $\mathbf{R}$ define the experiments*

$$\mathcal{E}_n \equiv (\mathbf{R}^{dn}, \mathcal{B}^{dn}, \bigotimes_{i=1}^{n} P_{1/\sqrt{n},g} : g \in H) \ . \quad \text{(B.9)}$$

If $0 \in H$, $\{\psi_k\}_{k=1}^{d_P}$ is an orthonormal basis for L_0^2, and Φ denotes the standard normal measure on $\mathbf{R}$, then $\mathcal{E}_n$ converges weakly to the dominated experiment $\mathcal{E}$

$$\mathcal{E} \equiv (\mathbf{R}^{d_P}, \mathcal{B}^{d_P}, Q_g : g \in H) \ , \quad \text{(B.10)}$$

where for each $g \in H$, $Q_g(\cdot) = Q_0(\cdot - T(g))$ for $T(g) \equiv \{\int g\psi_k dP\}_{k=1}^{d_P}$ and $Q_0 = \bigotimes_{k=1}^{d_P} \Phi$.

Proof: The conclusion of the Lemma is well known (see e.g. Subsection 8.2 in van der Vaart (1991a)), but we were unable to find a concrete reference and hence we include its proof for completeness. Since the Lemma is straightforward when the dimension of L_0^2 is finite ($d_P < \infty$) we focus on the case $d_P = \infty$. To analyze $\mathcal{E}$, let

$$\ell^2 \equiv \{\{c_k\}_{k=1}^{\infty} \in \mathbf{R}^{\infty} : \sum_{k=1}^{\infty} c_k^2 < \infty\} \ , \quad \text{(B.11)}$$

and note that by Example 2.3.5 in Bogachev (1998), ℓ^2 is the Cameron-Martin space of Q_0.[19] Hence, since for any $g \in L_0^2$ we have $\{\int g\psi_k dP\}_{k=1}^{\infty} \in \ell^2$ due to $\{\psi_k\}_{k=1}^{\infty}$ being an orthonormal basis for L_0^2, Theorem 2.4.5 in Bogachev (1998) implies

$$Q_g \equiv Q_0(\cdot - T(g)) \ll Q_0 \quad \text{(B.12)}$$

for all $g \in L_0^2$, and thus $\mathcal{E}$ is dominated by Q_0. Denoting an element of $\mathbf{R}^{\infty}$ by $\omega = \{\omega_k\}_{k=1}^{\infty}$, we then obtain from $\{\int g\psi_k dP\}_{k=1}^{\infty} \in \ell^2$ and the Martingale convergence

[19] See page 44 in Bogachev (1998) for a definition of a Cameron Martin space.

theorem, see for example Theorem 12.1.1 in Williams (1991), that

$$Q_0(\omega : \lim_{n\to\infty}\sum_{k=1}^{n}\omega_k\int\psi_k g dP \text{ exists}) = 1 \tag{B.13}$$

$$\lim_{n\to\infty}\int(\sum_{k=n+1}^{\infty}\omega_k\int g\psi_k dP)^2 dQ_0(\omega) = 0 \ . \tag{B.14}$$

Therefore, Example 2.3.5 and Corollary 2.4.3 in Bogachev (1998) yield for any $g \in L_0^2$

$$\begin{aligned}\log(\frac{dQ_g}{dQ_0}(\omega)) &= \sum_{k=1}^{\infty}\omega_k\int g\psi_k dP - \frac{1}{2}\int(\sum_{k=1}^{\infty}\omega_k\int g\psi_k dP)^2 dQ_0(\omega) \\ &= \sum_{k=1}^{\infty}\omega_k\int g\psi_k dP - \frac{1}{2}\int g^2 dP \ ,\end{aligned} \tag{B.15}$$

where the right hand side of the first equality is well defined Q_0 almost surely by (B.13), while the second equality follows from (B.14) and $\sum_{k=1}^{\infty}(\int g\psi_k dP)^2 = \int g^2 dP$ due to $\{\psi_k\}_{k=1}^{\infty}$ being an orthonormal basis for L_0^2.

Next, select an arbitrary finite subset $\{g_j\}_{j=1}^{J} \equiv I \subseteq H$ and vector $(\lambda_1, \ldots, \lambda_J)' \equiv \lambda \in \mathbf{R}^J$. From result (B.15) we then obtain Q_0 almost surely that

$$\sum_{j=1}^{J}\lambda_j\log(\frac{dQ_{g_j}}{dQ_0}(\omega)) = \sum_{k=1}^{\infty}\omega_k\int(\sum_{j=1}^{J}\lambda_j g_j)\psi_k dP - \sum_{j=1}^{J}\frac{\lambda_j}{2}\int g_j^2 dP \ . \tag{B.16}$$

In particular, we can conclude from Example 2.10.2 and Proposition 2.10.3 in Bogachev (1998) together with (B.14) and $\sum_{j=1}^{J}\lambda_j g_j \in L_0^2$ that, under Q_0, we have

$$\sum_{j=1}^{J}\lambda_j\log(\frac{dQ_j}{dQ_0}) \sim N(-\sum_{j=1}^{J}\frac{\lambda_j}{2}\int g_j^2 dP, \int(\sum_{j=1}^{J}\lambda_j g_j)^2 dP) \ . \tag{B.17}$$

Thus, for $\mu_I \equiv \frac{1}{2}(\int g_1^2 dP, \ldots, \int g_J^2 dP)'$ and $\Sigma_I \equiv \int(g_1, \ldots, g_J)'(g_1, \ldots, g_J)dP$, we have

$$(\log(\frac{dQ_{g_1}}{dQ_0}), \ldots, \log(\frac{dQ_{g_J}}{dQ_0}))' \sim N(-\mu_I, \Sigma_I) \ , \tag{B.18}$$

under Q_0 due to (B.17) holding for arbitrary $\lambda \in \mathbf{R}^J$.

To obtain an analogous result for the sequence of experiments $\mathcal{E}_n$, let $P^n \equiv \bigotimes_{i=1}^{n} P$ and $\{X_i\}_{i=1}^{n} \sim P^n$. From Lemma 25.14 in van der Vaart (1998) we obtain under P^n

$$\sum_{i=1}^{n}\log(\frac{dP_{1/\sqrt{n},g_j}}{dP}(X_i)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}g_j(X_i) - \frac{1}{2}\int g_j^2 dP + o_p(1) \tag{B.19}$$

for any $1 \leq j \leq J$. Thus, defining $P^n_{1/\sqrt{n},g_j} \equiv \bigotimes_{i=1}^n P_{1/\sqrt{n},g_j}$, we can conclude that

$$(\log(\frac{dP^n_{1/\sqrt{n},g_1}}{dP^n}),\ldots,\log(\frac{dP^n_{1/\sqrt{n},g_J}}{dP^n}))' \stackrel{L}{\to} N(-\mu_I,\Sigma_I)\ , \tag{B.20}$$

under P^n by (B.19), the central limit theorem, and the definitions of μ_I and Σ_I. Furthermore, also note Lemma B.1 implies $\int |dP^n - dP^n_{1/\sqrt{n},0}| = o(1)$ and hence

$$(\frac{dP^n_{1/\sqrt{n},g_1}}{dP^n},\ldots,\frac{dP^n_{1/\sqrt{n},g_J}}{dP^n})' = (\frac{dP^n_{1/\sqrt{n},g_1}}{dP^n_{1/\sqrt{n},0}},\ldots,\frac{dP^n_{1/\sqrt{n},g_J}}{dP^n_{1/\sqrt{n},0}})' + o_p(1) \tag{B.21}$$

under P^n by Lemma B.2 and result (B.20). Thus, by (B.20) and (B.21) we obtain

$$(\log(\frac{dP^n_{1/\sqrt{n},g_1}}{dP^n_{1/\sqrt{n},0}}),\ldots,\log(\frac{dP^n_{1/\sqrt{n},g_J}}{dP^n_{1/\sqrt{n},0}}))' \stackrel{L}{\to} N(-\mu_I,\Sigma_I)\ , \tag{B.22}$$

under P^n, and since $\int |dP^n - dP^n_{1/\sqrt{n},0}| = o(1)$ also under $P^n_{1/\sqrt{n},0}$. Hence, the Lemma follows from (i) (B.18), (ii) (B.22), and (iii) $\{P^n_{1/\sqrt{n},g}\}$ and $\{P^n_{1/\sqrt{n},0}\}$ being mutually contiguous for any $g \in H$ by (B.19) and Corollary 12.3.1 in Lehmann and Romano (2005), which together verify the conditions of Lemma 10.2.1 in LeCam (1986). ■

Lemma B.4. *Let $\mathbb{G}_0$ be a centered Gaussian measure on a separable Banach space $\mathbf{B}$ and $0 \neq \Delta \in \mathbf{B}$ belong to the support of $\mathbb{G}_0$. Further suppose $\Psi : \mathbf{B} \to \mathbf{R}_+$ is continuous, convex, and satisfies $\Psi(0) = 0$, $\Psi(b) = \Psi(-b)$ for all $b \in \mathbf{B}$, and $\{b \in \mathbf{B} : \Psi(b) \leq t\}$ is bounded for any $0 < t < \infty$. For any $t > 0$ it then follows that*

$$P(\Psi(\mathbb{G}_0 + \Delta) < t) < P(\Psi(\mathbb{G}_0) < t)\ .$$

Proof: Let $\|\cdot\|_{\mathbf{B}}$ denote the norm of $\mathbf{B}$, fix an arbitrary $t > 0$ and define a set C by

$$C \equiv \{b \in \mathbf{B} : \Psi(b) < t\}\ . \tag{B.23}$$

For $\mathbf{B}^*$ the dual space of $\mathbf{B}$ let $\|\cdot\|_{\mathbf{B}^*}$ denote its norm, and $\nu_C : \mathbf{B}^* \to \mathbf{R}$ be given by

$$\nu_C(b^*) = \sup_{b\in C} b^*(b)\ , \tag{B.24}$$

which constitutes the support functional of C. Then note for any $b^* \in \mathbf{B}^*$ we have

$$\nu_C(-b^*) = \sup_{b\in C} -b^*(b) = \sup_{b\in C} b^*(-b) = \sup_{b\in -C} b^*(b) = \nu_C(b^*)\ , \tag{B.25}$$

due to $C = -C$ since $\Psi(b) = \Psi(-b)$ for all $b \in \mathbf{B}$. Moreover, note that $0 \in C$ since $\Psi(0) = 0 < t$, and hence there exists a $M_0 > 0$ such that $\{b \in \mathbf{B} : \|b\|_{\mathbf{B}} \leq M_0\} \subseteq C$ by continuity of Ψ. Thus, by definition of $\|\cdot\|_{\mathbf{B}^*}$ we obtain for any $b^* \in \mathbf{B}^*$ that

$$\nu_C(b^*) = \sup_{b\in C} b^*(b) \geq \sup_{\|b\|_{\mathbf{B}}\leq M_0} b^*(b) = M_0 \times \sup_{\|b\|_{\mathbf{B}}\leq 1} |b^*(b)| = M_0\|b^*\|_{\mathbf{B}^*} \ . \tag{B.26}$$

Analogously, note that by assumption $M_1 \equiv \sup_{b\in C} \|b\|_{\mathbf{B}} < \infty$, and thus for any $b^* \in \mathbf{B}^*$

$$\nu_C(b^*) = \sup_{b\in C} b^*(b) \leq \|b^*\|_{\mathbf{B}^*} \times \sup_{b\in C} \|b\|_{\mathbf{B}} = M_1\|b^*\|_{\mathbf{B}^*} \ . \tag{B.27}$$

We next aim to define a norm on $\mathbf{B}$ under which C is the open unit sphere. To this end, recall that the original norm $\|\cdot\|_{\mathbf{B}}$ on $\mathbf{B}$ may be written as

$$\|b\|_{\mathbf{B}} = \sup_{\|b^*\|_{\mathbf{B}^*}=1} b^*(b) \ , \tag{B.28}$$

see for instance Lemma 6.10 in Aliprantis and Border (2006). Similarly, instead define

$$\|b\|_{\mathbf{B},C} \equiv \sup_{\|b^*\|_{\mathbf{B}^*}=1} \frac{b^*(b)}{\nu_C(b^*)} \ , \tag{B.29}$$

and note that: (i) $\|b_1 + b_2\|_{\mathbf{B},C} \leq \|b_1\|_{\mathbf{B},C} + \|b_2\|_{\mathbf{B},C}$ for any $b_1, b_2 \in \mathbf{B}$ by direct calculation, (ii) $\|\alpha b\|_{\mathbf{B},C} = |\alpha|\|b\|_{\mathbf{B},C}$ for any $\alpha \in \mathbf{R}$ and $b \in \mathbf{B}$ by (B.25) and (B.29), and (iii) results (B.26), (B.27), (B.28), and (B.29) imply that for any $b \in \mathbf{B}$ we have

$$M_0\|b\|_{\mathbf{B},C} \leq \|b\|_{\mathbf{B}} \leq M_1\|b\|_{\mathbf{B},C} \ , \tag{B.30}$$

which establishes $\|b\|_{\mathbf{B},C} = 0$ if an only if $b = 0$, and hence we conclude $\|\cdot\|_{\mathbf{B},C}$ is indeed a norm on $\mathbf{B}$. In fact, (B.30) implies that the norms $\|\cdot\|_{\mathbf{B}}$ and $\|\cdot\|_{\mathbf{B},C}$ are equivalent, and hence $\mathbf{B}$ remains a separable Banach space and its Borel σ-algebra unchanged when endowed with $\|\cdot\|_{\mathbf{B},C}$ in place of $\|\cdot\|_{\mathbf{B}}$.

Next, note that the continuity of Ψ implies C is open, and thus for any $b_0 \in C$ there is an $\epsilon > 0$ such that and $\{b : \|b - b_0\|_{\mathbf{B}} \leq \epsilon\} \subset C$. We then obtain

$$\nu_C(b^*) \geq \sup_{\|b-b_0\|_{\mathbf{B}}\leq\epsilon} b^*(b) = \sup_{\|b\|_{\mathbf{B}}\leq 1} \{b^*(b_0) + \epsilon b^*(b)\} = b^*(b_0) + \epsilon\|b^*\|_{\mathbf{B}^*} \ , \tag{B.31}$$

where the final equality follows as in (B.26). Thus, from (B.27) and (B.31) we conclude $1 - \epsilon/M_1 \geq b^*(b_0)/\nu_C(b^*)$ for all b^* with $\|b^*\|_{\mathbf{B}^*} = 1$, and hence we conclude

$$C \subseteq \{b \in \mathbf{B} : \|b\|_{\mathbf{B},C} < 1\} \ . \tag{B.32}$$

Suppose on the other hand that $\|b_0\|_{\mathbf{B},C} < 1$, and note (B.29) implies for some $\delta > 0$

$$b^*(b_0) < \nu_C(b^*)(1 - \delta) \tag{B.33}$$

for all $b^* \in \mathbf{B}^*$ with $\|b^*\|_{\mathbf{B}^*} = 1$. Setting $\eta \equiv \delta M_0$ and arguing as in (B.31) then yields

$$\sup_{\|b^*\|_{\mathbf{B}^*}=1} \sup_{\|b-b_0\|_{\mathbf{B}}\leq\eta} \{b^*(b) - \nu_C(b^*)\} = \sup_{\|b^*\|_{\mathbf{B}^*}=1} \{b^*(b_0) + \eta\|b^*\|_{\mathbf{B}^*} - \nu_C(b^*)\}$$
$$< \sup_{\|b^*\|_{\mathbf{B}^*}=1} \{\eta - \nu_C(b^*)\delta\} = \sup_{\|b^*\|_{\mathbf{B}^*}=1} \delta(M_0 - \nu_C(b^*)) \leq 0 \ , \quad \text{(B.34)}$$

where the first inequality follows from (B.33), the second equality by definition of η, and the final inequality follows from (B.26). Since C is convex by hypothesis, (B.34) and Theorem 5.12.5 in Luenberger (1969) imply $\{b : \|b - b_0\|_{\mathbf{B}} \leq \eta\} \subseteq \bar{C}$. We conclude b_0 is in the interior of $\bar{C}$, and since C is convex and open, Lemma 5.28 in Aliprantis and Border (2006) yields that $b_0 \in C$. Thus, we can conclude that

$$\{b \in \mathbf{B} : \|b\|_{\mathbf{B},C} < 1\} \subseteq C \ , \quad \text{(B.35)}$$

which together with (B.32) yields $C = \{b \in \mathbf{B} : \|b\|_{\mathbf{B},C} < 1\}$. Therefore, $\mathbf{B}$ being separable under $\|\cdot\|_{\mathbf{B},C}$, $0 \neq \Delta$ being in the support of $\mathbb{G}_0$ by hypothesis, and Corollary 2 in Lewandowski et al. (1995) finally enable us to derive

$$P(\Psi(\mathbb{G}_0 + \Delta) < t) = P(\mathbb{G}_0 + \Delta \in C) < P(\mathbb{G}_0 \in C) = P(\Psi(\mathbb{G}_0) < t) \ , \quad \text{(B.36)}$$

which establishes the claim of the Lemma. ■

Lemma B.5. *Let Assumptions 2.1(i), 3.1 hold, and* $\mathbb{G}_0 \equiv \{\mathbb{G}_{0,k}\}_{k=1}^{d_F}$ *be a Radon centered Gaussian measure on* $\ell^\infty([d_F])$ *satisfying* $E[\mathbb{G}_{0,k}\mathbb{G}_{0,j}] = \int f_k f_j dP$ *for any* $1 \leq k, j \leq d_F$. *If* $g \in L_0^2$, *then* $\Delta_g \equiv \{\int f_k g dP\}_{k=1}^{d_F}$ *belongs to the support of* $\mathbb{G}_0$.

Proof: Let $\Delta_{g,k} \equiv \int f_k g dP$ and note $\mathcal{F} \subset L_0^2$, $g \in L_0^2$, and the Cauchy-Schwarz inequality imply $|\Delta_{g,k}| \leq \|g\|_{L^2}\|f_k\|_{L^2}$. Moreover, since $\|f\|_{L^2}$ is uniformly bounded in $f \in \mathcal{F}$ by Assumption 3.1, we conclude $\Delta_g \in \ell^\infty([d_F])$. Letting $\ell^1([d_F]) \equiv \{\{c_k\}_{k=1}^{d_F} : \sum_{k=1}^{d_F} |c_k| < \infty\}$ and $\ell^\infty([d_F])^*$ denote the dual space of $\ell^\infty([d_F])$, we next aim to show

$$\sup\{b^*(\Delta_g) : b^* \in \ell^\infty([d_F])^* \text{ and } E[(b^*(\mathbb{G}_0))^2] \leq 1\}$$
$$= \sup\{b^*(\Delta_g) : b^* \in \ell^1([d_F]) \text{ and } E[(b^*(\mathbb{G}_0))^2] \leq 1\} \ , \quad \text{(B.37)}$$

where for each $\{b_k^*\}_{k=1}^{d_F} \equiv b^* \in \ell^1([d_F])$ and $\{b_k\}_{k=1}^{d_F} \equiv b \in \ell^\infty([d_F])$, $b^*(b) = \sum_{k=1}^{d_F} b_k^* b_k$. To this end, note that if $d_F < \infty$, then $\ell^\infty([d_F])^* = \ell^1([d_F])$ and (B.37) is immediate. For the case $d_F = \infty$, let $\ell^0 \equiv \{b \in \ell^\infty([d_F]) : \lim_{k\to\infty} b_k \text{ exists}\}$ and define

$$\ell_d \equiv \{b^* \in \ell^\infty([d_F])^* : \exists M \in \mathbf{R} \text{ such that } b^*(b) = M \lim_{k\to\infty} b_k \ \forall b \in \ell^0\} \ . \quad \text{(B.38)}$$

By Lemma 16.30 in Aliprantis and Border (2006), $\ell^\infty([d_F])^* = \ell^1([d_F]) \oplus \ell_d$, and hence

$$\begin{aligned}&\sup\{b^*(\Delta_g) : b^* \in \ell^\infty([d_F])^* \text{ and } E[(b^*(\mathbb{G}_0))^2] \leq 1\} \\ &\quad = \sup\{\{b_1^* + b_d^*\}(\Delta_g) : b_1^* \in \ell^1([d_F]), b_d^* \in \ell_d \text{ and } E[(\{b_1^* + b_d^*\}(\mathbb{G}_0))^2] \leq 1\} \ . \end{aligned} \tag{B.39}$$

However, note that the Cauchy-Schwarz inequality and $\sum_{k=1}^{d_F} \int f_k^2 dP < \infty$ imply that

$$\lim_{k\to\infty} |\Delta_{g,k}| \leq \|g\|_{L^2} \times \lim_{k\to\infty} \|f_k\|_{L^2} = 0 \ . \tag{B.40}$$

Similarly, by Markov's inequality, $E[\mathbb{G}_{0,k}^2] = \int f_k^2 dP$ if $1 \leq k \leq d_F$, and Assumption 3.1

$$\sum_{k=1}^{\infty} P(|\mathbb{G}_{0,k}| > \epsilon) \leq \sum_{k=1}^{\infty} \frac{1}{\epsilon^2} E[\mathbb{G}_{0,k}^2] = \sum_{k=1}^{\infty} \frac{1}{\epsilon^2} \int f_k^2 dP < \infty \ . \tag{B.41}$$

Thus, by the Borel-Cantelli Lemma $\lim_{k\to\infty} \mathbb{G}_{0,k} = 0$ almost surely, which implies $b_d^*(\mathbb{G}_0) = 0$ almost surely for any $b_d^* \in \ell_d$. Since similarly $b_d^*(\Delta_g) = 0$ for any $b_d^* \in \ell_d$ by result (B.40), we can conclude from (B.39) that (B.37) holds when $d_F = \infty$ as well.

Next, note that for any $\{b_k^*\}_{k=1}^{d_F} \equiv b^* \in \ell^1([d_F])$, we obtain from $\Delta_g \in \ell^\infty([d_F])$ that

$$\begin{aligned}|b^*(\Delta_g)| &= |\sum_{k=1}^{d_F} b_k^* \int g f_k dP| \\ &= |\int (\sum_{k=1}^{d_F} b_k^* f_k) g dP| \leq \|g\|_{L^2} \times \{\int (\sum_{k=1}^{d_F} b_k^* f_k)^2 dP\}^{\frac{1}{2}} \ , \end{aligned} \tag{B.42}$$

where the final result follows from the Cauchy-Schwarz inequality. Furthermore, since $\int f_k f_j dP = E[\mathbb{G}_{0,k}\mathbb{G}_{0,j}]$ for any $1 \leq k, j \leq d_F$, we obtain that for any finite $K \leq d_F$

$$\int (\sum_{k=1}^{K} b_k^* f_k)^2 dP = E[(\sum_{k=1}^{K} b_k^* \mathbb{G}_{0,k})^2] \ . \tag{B.43}$$

In particular, if $d_F < \infty$, then combining (B.42) and (B.43) implies for any $b^* \in \ell^1([d_F])$

$$b^*(\Delta_g) \leq \|g\|_{L^2} \times \{E[(b^*(\mathbb{G}_0))^2]\}^{\frac{1}{2}} \ . \tag{B.44}$$

In order to obtain an analogous result when $d_F = \infty$, we apply the dominated convergence theorem with the dominating functions $(\sum_{k=1}^{d_F} b_k^* f_k)^2 \leq (\sum_{k=1}^{d_F} |b_k^*|)^2 \sum_{k=1}^{d_F} f_k^2$ and

$(\sum_{k=1}^{d_F} b_k^* \mathbb{G}_{0,k})^2 \leq (\sum_{k=1}^{d_F} |b_k^*|)^2 \sum_{k=1}^{d_F} (\mathbb{G}_{0,k})^2$ together with (B.43) to conclude

$$\int (\sum_{k=1}^{\infty} b_k^* f_k)^2 dP = \lim_{K\to\infty} \int (\sum_{k=1}^{K} b_k^* f_k)^2 dP$$
$$= \lim_{K\to\infty} E[(\sum_{k=1}^{K} b_k^* \mathbb{G}_{0,k})^2] = E[(b^*(\mathbb{G}_0))^2] \ . \quad \text{(B.45)}$$

Thus, from (B.42) and (B.45) it follows that (B.44) holds for any $b^* \in \ell^1([d_F])$ for both finite and infinite d_F. Hence, combining (B.37) and (B.44) we finally obtain that

$$\sup\{b^*(\Delta_g) : b^* \in \ell^\infty([d_F])^* \text{ and } E[(b^*(\mathbb{G}_0))^2] \leq 1\} \leq \|g\|_{L^2} < \infty \ . \quad \text{(B.46)}$$

Since $\mathbb{G}_0$ is centered, it follows that Δ_g belongs to the Cameron-Martin space of $\mathbb{G}_0$, and hence we conclude from Theorem 3.6.1 in Bogachev (1998) and $\mathbb{G}_0$ being Radon by hypothesis that Δ_g belongs to the support of $\mathbb{G}_0$. ■

Lemma B.6. *Let Assumptions 2.1(i) and 4.1(i) hold, and suppose* $\tilde{\theta}_n : \{X_i\}_{i=1}^n \to \mathbf{B}$ *is an asymptotically linear estimator for* $\theta(P)$ *such that* $\sqrt{n}\{\tilde{\theta}_n - \theta(P)\} \xrightarrow{L} \mathbb{Z}$ *under* $\bigotimes_{i=1}^n P$ *on* $\mathbf{B}$ *for some tight Borel* $\mathbb{Z}$. *It then follows that for any function* $h \in L_0^2$, $(\sqrt{n}\{\tilde{\theta}_n - \theta(P)\}, \frac{1}{\sqrt{n}}\sum_{i=1}^n h(X_i))$ *converges in distribution under* $\bigotimes_{i=1}^n P$ *on* $\mathbf{B} \times \mathbf{R}$.

PROOF: For notational simplicity, let $\eta(P) \equiv (\theta(P), 0) \in \mathbf{B} \times \mathbf{R}$ and similarly define $\hat{\eta}_n \equiv (\tilde{\theta}_n, \frac{1}{n}\sum_{i=1}^n h(X_i)) \in \mathbf{B} \times \mathbf{R}$. Further let $(\mathbf{B} \times \mathbf{R})^*$ denote the dual space of $\mathbf{B} \times \mathbf{R}$ and note that for any $d^* \in (\mathbf{B} \times \mathbf{R})^*$ there are $b_{d^*}^* \in \mathbf{B}^*$ and $r_{d^*}^* \in \mathbf{R}$ such that $d^*((b, r)) = b_{d^*}^*(b) + r_{d^*}^*(r)$ for all $(b, r) \in \mathbf{B} \times \mathbf{R}$. For $\tilde{\nu}$ the influence function of $\tilde{\theta}_n$ then define $\zeta_{d^*}(X_i) \equiv \{b_{d^*}^*(\tilde{\nu}(X_i)) + r_{d^*}^*(h(X_i))\}$ to obtain that under $\bigotimes_{i=1}^n P$ we have

$$d^*(\sqrt{n}\{\hat{\eta}_n - \eta(P)\}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \zeta_{d^*}(X_i) + o_p(1) \quad \text{(B.47)}$$

by asymptotic linearity of $\tilde{\theta}_n$. Thus, for any finite set $\{d_k^*\}_{k=1}^K \subset (\mathbf{B} \times \mathbf{R})^*$, we have

$$(d_1^*(\sqrt{n}\{\hat{\eta}_n - \eta(P)\}), \ldots, d_K^*(\sqrt{n}\{\hat{\eta}_n - \eta(P)\})) \xrightarrow{L} (\mathbb{W}_{d_1^*}, \ldots, \mathbb{W}_{d_K^*}) \quad \text{(B.48)}$$

for $(\mathbb{W}_{d_1^*}, \ldots, \mathbb{W}_{d_K^*})$ a multivariate normal random variable satisfying $E[\mathbb{W}_{d_k^*}] = 0$ for all $1 \leq k \leq K$ and $E[\mathbb{W}_{d_j^*}\mathbb{W}_{d_k^*}] = E[\zeta_{d_j^*}(X_i)\zeta_{d_k^*}(X_i)]$ for all $1 \leq j \leq k \leq K$.

Next note that since $\sqrt{n}\{\tilde{\theta}_n - \theta(P)\}$ is asymptotically measurable and asymptotically tight by Lemma 1.3.8 in van der Vaart and Wellner (1996), it follows that $\sqrt{n}\{\hat{\eta}_n - \eta(P)\}$ is asymptotically measurable and asymptotically tight on $\mathbf{B} \times \mathbf{R}$ by Lemmas 1.4.3 and 1.4.4 in van der Vaart and Wellner (1996). Hence, we conclude by Theorem 1.3.9 in van der Vaart and Wellner (1996) that any sequence $\{n_k\}$ has a subsequence $\{n_{k_j}\}$ with

$$\sqrt{n}_{k_j}\{\hat{\eta}_{n_{k_j}} - \eta(P)\} \xrightarrow{L} \mathbb{W} \tag{B.49}$$

under $\bigotimes_{i=1}^{n_{k_j}} P$ for $\mathbb{W}$ some tight Borel Law on $\mathbf{B} \times \mathbf{R}$. However, letting $C_b(\mathbf{R}^K)$ denote the set of continuous and bounded functions on $\mathbf{R}^K$, we obtain from (B.48), (B.49), and the continuous mapping theorem that for any $\{d_k^*\}_{k=1}^K \subset (\mathbf{B} \times \mathbf{R})^*$ and $f \in C_b(\mathbf{R}^K)$

$$E[f((d_1^*(\mathbb{W}), \ldots, d_K^*(\mathbb{W})))] = E[f((\mathbb{W}_{d_1^*}, \ldots, \mathbb{W}_{d_K^*}))] \ . \tag{B.50}$$

In particular, since $\mathcal{G} \equiv \{f \circ (d_1^*, \ldots, d_K^*) : f \in C_b(\mathbf{R}^K), \{d_k^*\}_{k=1}^K \subset (\mathbf{B} \times \mathbf{R})^*, 1 \leq K < \infty\}$ is a vector lattice that separates points in $\mathbf{B} \times \mathbf{R}$, Lemma 1.3.12 in van der Vaart and Wellner (1996) implies there is a unique tight Borel measure $\mathbb{W}$ on $\mathbf{B} \times \mathbf{R}$ satisfying (B.50). Thus, since the original sequence $\{n_k\}$ was arbitrary, we conclude all limit points of the law of $\sqrt{n}\{\hat{\eta}_n - \eta(P)\}$ coincide, and the Lemma follows. ■

Lemma B.7. *Let Assumption 2.1 hold, $\mathcal{D}$ be a dense subset of L^2, and for any $f \in \mathcal{D}$ let Ω_f^* denote the semiparametric efficiency bound for estimating $\int f dP$. It then follows that $\Omega_f^* = Var\{f(X)\}$ for all $f \in \mathcal{D}$ if and only if P is just identified.*

PROOF: First note the parameter $\theta_f(P) \equiv \int f dP$ is pathwise differentiable at $\theta_f(P)$ relative to $T(P)$ with derivative $\dot{\theta}_f(g) \equiv \int \Pi_T(f) g dP$. Therefore, by Theorem 5.2.1 in Bickel et al. (1993) its efficiency bound is given by $\Omega_f^* = \|\Pi_T(f)\|_{L^2}^2$. For any $f \in L^2$ let $\Pi_{L_0^2}(f)$ denote its projection onto L_0^2 and note that $\Pi_{L_0^2}(f) = \{f - \int f dP\}$, and hence $\text{Var}\{f(X)\} = \|\Pi_{L_0^2}(f)\|_{L^2}$. By orthogonality of $\bar{T}(P)$ and $\bar{T}(P)^\perp$ we then obtain that

$$\begin{aligned} \text{Var}\{f(X)\} = \|\Pi_{L_0^2}(f)\|_{L^2}^2 &= \|\Pi_T(\Pi_{L_0^2}(f)) + \Pi_{T^\perp}(\Pi_{L_0^2}(f))\|_{L^2}^2 \\ &= \|\Pi_T(\Pi_{L_0^2}(f))\|_{L^2}^2 + \|\Pi_{T^\perp}(\Pi_{L_0^2}(f))\|_{L^2}^2 = \Omega_f^* + \|\Pi_{T^\perp}(f)\|_{L^2} \ , \end{aligned} \tag{B.51}$$

where in the final equality we exploited that $\Pi_T(\Pi_{L_0^2}(f)) = \Pi_T(f)$ and $\Pi_{T^\perp}(\Pi_{L_0^2}(f)) = \Pi_{T^\perp}(f)$ for any $f \in L^2$ due to $\bar{T}(P)$ and $\bar{T}(P)^\perp$ being subspaces of L_0^2. Thus, by (B.51) $\text{Var}\{f(X)\} = \Omega_f^*$ for all $f \in \mathcal{D}$ if and only if $\Pi_{T^\perp}(f) = 0$ for all $f \in \mathcal{D}$, which by denseness of $\mathcal{D}$ is equivalent to $T(P)^\perp = \{0\}$. ■

References

Ackerberg, D., Chen, X., Hahn, J. and Liao, Z. (2014). Asymptotic efficiency of semiparametric two-step gmm. *Review of Economic Studies*, **81(3)** 919–943.

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, **71** 1795–1844.

AI, C. and CHEN, X. (2012). The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, **170** 442–457.

ALIPRANTIS, C. D. and BORDER, K. C. (2006). *Infinite Dimensional Analysis – A Hitchhiker's Guide*. Springer-Verlag, Berlin.

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

ANDREWS, D. W. K. (1999). Estimation When a Parameter is on a Boundary. *Econometrica*, **67** pp. 1341–1383.

ANDREWS, D. W. K. (2011). Examples of l2-complete and boundedly-complete distributions. Cowles Foundation Discussion Paper No. 1801.

ARELLANO, M., HANSEN, L. P. and SENTANA, E. (2012). Underidentification? *Journal of Econometrics*, **170** 256–280.

BICKEL, P., KLAASSEN, C., RITOV, Y. and WELLNER, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press.

BOGACHEV, V. I. (1998). *Gaussian Measures*. American Mathematical Society, Providence.

BONHOMME, S. (2012). Functional differencing. *Econometrica*, **80** 1337–1385.

CANAY, I. A., SANTOS, A. and SHAIKH, A. M. (2013). On the testability of identification in some nonparametric models with endogeneity. *Econometrica*, **81** 2535–2559.

CHAMBERLAIN, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica*, **60** 567–596.

CHEN, X., CHERNOZHUKOV, V., LEE, S. and NEWEY, W. K. (2014). Local identification of nonparametric and semiparametric models. *Econometrica*, **82** 785–809.

CHEN, X. and POUZO, D. (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics*, **152** 46–60.

CHEN, X. and POUZO, D. (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, **80** 277–321.

CHERNOZHUKOV, V., IMBENS, G. W. and NEWEY, W. K. (2007). Instrumental variable estimation of nonseparable models. **139** 4–14.

DAROLLES, S., FAN, Y., FLORENS, J.-P. and RENAULT, E. (2011). Nonparametric instrumental regression. *Econometrica*, **79** 1541–1565.

DAVYDOV, Y. A., LIFSHITS, M. A. and SMORODINA, N. V. (1998). *Local Properties of Distributions of Stochastic Functionals.* American Mathematical Society, Providence.

FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A.*, **222** 309–369.

FLORENS, J., MARIMOUTOU, V. and PEGUIN-FEISSOLLE, A. (2007). *Econometric Modeling and Inference.* Cambridge University Press.

FLORENS, J. P., JOHANNES, J. and BELLEGEM, S. V. (2012). Instrumental regression in partially linear models. *The Econometrics Journal*, **15** 304–324.

HALL, P. and HOROWITZ, J. (2005). Nonparametric methods for inference in the presense of instrumental variables. *Annals of Statistics*, **33** 2904–2929.

HANSEN, L. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, **50** pp. 1029–1054.

HAUSMAN, J. A. (1978). Specification tests in econometrics. *Econometrica*, **46** 1251–1271.

HOROWITZ, J. L. and LEE, S. (2007). Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica*, **75** 1191–1208.

KOOPMANS, T. C. and RIERSOL, O. (1950). The identification of structural characteristics. *The Annals of Mathematical Statistics*, **21** 165–181.

LECAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, New York.

LEHMANN, E. and ROMANO, J. (2005). *Testing Statistical Hypotheses.* Springer Verlag.

LEHMANN, E. L. (1952). Testing multiparameter hypotheses. *The Annals of Mathematical Statistics*, **23** 541–552.

LEWANDOWSKI, M., RYZNAR, M. and ZAK, T. (1995). Anderson inequality is strict for gaussian and stable measures. *Proceedings of the American Mathematical Society*, **123** 3875–3880.

LUENBERGER, D. (1969). *Optimization by Vector Space Methods.* John Wiley & Sons.

MANSKI, C. F. (2003). *Partial Identification of Probability Distributions.* Springer-Verlag, New York.

MOREAU, J.-J. (1962). Décomposition orthogonale dun espace hilbertien selon deux cônes mutuellement polaires. *CR Acad. Sci. Paris*, **255** 238–240.

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

NEWEY, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, **62** 1349–1382.

NEWEY, W. K. and POWELL, J. (2003). Instrumental variables estimation of nonparametric models. *Econometrica*, **71** 1565–1578.

NEWEY, W. K. and POWELL, J. L. (1999). Two-step estimation, optimal moment conditions, and sample selection models. Working Paper, MIT.

NEWEY, W. K. and STOKER, T. M. (1993). Efficiency of weighted average derivative estiamtors and index models. *Econometrica*, **61** 1199–1223.

PLOBERGER, W. and PHILLIPS, P. C. (2012). Optimal estimation under nonstandard conditions. *Journal of Econometrics*, **169** 258–265.

POWELL, J. (1994). Estimation of semiparametric models. *Handbook of Econometrics, R.F. Engle III and D.F. McFadden (eds.)*, **4**.

POWELL, J. L., STOCK, J. H. and STOKER, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, **57** 1403–1430.

ROMANO, J. P. (2004). On non-parametric testing, the uniform behavior of the t-test and related problems. *Scandinavian Journal of Statistics*, **31** 567–584.

SANTOS, A. (2011). Instrumental variables methods for recovering continuous linear functionals. *Journal of Econometrics*, **161** 129–146.

SANTOS, A. (2012). Inference in nonparametric instrumental variables with partial identification. *Econometrica*, **80** 213–275.

SARGAN, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, **26** 393–415.

VAN DER VAART, A. (1989). On the asymptotic information bound. *The Annals of Statistics*, **17** 1487–1500.

VAN DER VAART, A. (1991a). An asymptotic representation theorem. *International Statistical Review*, **59** 97–121.

VAN DER VAART, A. (1991b). On Differentiable Functionals. *The Annals of Statistics*, **19** pp. 178–204.

van der Vaart, A. (1998). *Asymptotic Statistics.* Cambridge University Press.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes.* Springer Verlag.

van der Vaart, A. and Wellner, J. A. (1989). Prohorov and continuous mapping theorems in the hoffman-jorgensen weak convergence theory, with applications to convolution and asymptotic minimax theorems. Tech. Report 157. Dept. of Statistics, University of Washington, Seattle.

Williams, D. (1991). *Probability with Martingales.* Cambridge University Press, New York.

Yuan, A., Xu, J. and Zheng, G. (2014). On empirical likelihood statistical functions. *Journal of Econometrics*, **178** 613–623.

Zeidler, E. (1984). *Nonlinear Functional Analysis and its Applications III.* Springer-Verlag, New York.