THEORY OF WEAK IDENTIFICATION IN SEMIPARAMETRIC MODELS

Tetsuya Kaji

June 6, 2018

Abstract

We provide general formulation of weak identification in semiparametric models and a novel efficiency concept. Weak identification occurs when a parameter is *weakly regular*, i.e., when it depends on the score asymptotically. When this happens, consistent or equivariant estimation is shown to be impossible. We then show that behind every weakly regular parameter there exists an underlying parameter that is *regular* and fully characterizes the weakly regular parameter. While this parameter is not unique, concepts of *sufficiency* and *minimality* help pin down the desirable choice. If the estimation of minimal sufficient underlying parameters is inefficient, it introduces noise in the corresponding estimation of weakly regular parameters, whence we can improve the estimators by *local asymptotic Rao-Blackwellization*. We call an estimator *weakly efficient* if it attains an asymptotic distribution that does not admit such improvement. We demonstrate in heteroskedastic linear IV models that popular estimators can be improved under some conditions.

JEL Codes: C13, C14.

Keywords: weak identification, semiparametric efficiency.

1 INTRODUCTION

Weak identification arises in a wide range of empirical settings. A leading example is the linear instrumental variables (IV) model in which the instruments and endogenous regressors are barely correlated (Nelson and Startz, 1990; Bound et al., 1995). When this happens, even with a large sample, classical asymptotic theory is known to yield poor approximations to the behavior of familiar statistics (Staiger and Stock, 1997), causing problems in both estimation and inference. We encounter this problem in various other contexts: Stock and Wright (2000) analyze weak identification in generalized

tkaji@mit.edu. MIT Department of Economics. An online supplementary appendix is also attached. I thank Anna Mikusheva, Isaiah Andrews, Victor Chernozhukov, Whitney Newey, Hidehiko Ichimura, Kengo Kato, Rachael Meager, Peter Hull, Michal Kolesár, and seminar participants at the MIT Econometrics Lunch for helpful comments and suggestions.

method of moments (GMM) models; Guggenberger and Smith (2005, 2008) and Otsu (2006) in generalized empirical likelihood (GEL) models; Andrews and Cheng (2012), Han and McCloskey (2017), and Cox (2017) in extremum estimation models; Iskrev (2008), Ruge-Murcia (2007), and Canova and Sala (2009) in dynamic stochastic general equilibrium (DSGE) models; Armstrong (2016) in differentiated products demand estimation models. Many estimators of weakly identified parameters exhibit inconsistency and bias, and, as a consequence, standard inference procedures such as t- and Wald tests may have substantially distorted sizes (Phillips, 1984, 1989; Dufour, 1997; Hirano and Porter, 2015, as well as aforementioned papers). Following these practically challenging problems, a vast amount of theoretical work has been published.

The theoretical literature on weak identification is confined to specific estimation and inference procedures in specific models. Many papers consider particular asymptotic embeddings, find statistics that are well-behaved, and derive robust statistical procedures in various models, especially in the linear IV model. In contrast, many fundamental questions—such as what is the common cause of known instances of weak identification, what is a general guideline to look for well-behaved statistics, and what is the semiparametric efficiency in the presence of weak identification—have been largely left unanswered. Such exploration is essential, however, not only to facilitate unified understanding of the phenomenon but to measure performance of different procedures and develop general systematic construction methods for estimation and inference. This is more important than it has ever been, especially now that numerous inference procedures have been developed in many empirically relevant settings.

This paper studies weak identification from the perspective of semiparametric theory. We explore how weak identification emerges in the classical framework of Bickel et al. (1993), Van der Vaart (1998, Chapter 25), and Kosorok (2008, Part III). We find that weak identification occurs as a result of the parameter's asymptotic dependence on the score; we call such a parameter *weakly regular*. This is in stark contrast to the classical *regular* parameters, whose derivatives (local parameters) depend on the score, not the parameters themselves. As an immediate consequence of this observation, we derive—without reference to a specific estimation or inference procedure—that there exists neither a consistent estimator, a consistent test, nor an equivariant (hence pivotal) estimator when the parameter is weakly regular. The dependence on the score is homogeneous of degree zero and essentially nonlinear, and this nonlinearity is the root cause of many non-Gaussian nonpivotal asymptotic distributions witnessed throughout the literature (Staiger and Stock, 1997; Stock and Wright, 2000; Guggenberger and Smith, 2005; Andrews and Cheng, 2012; Cox, 2017). To circumvent the problem of almost arbitrary nonlinearity, we seek ways to explore weak regularity from the standpoint of *regular* parameters.

We show that behind every weakly regular parameter there exists an underlying parameter that is regular and controls the limit behavior of the weakly regular parameter. In other words, a weakly regular parameter can be represented as a nonlinear transformation of the local parameter of some regular parameter. Finding such a parameter allows us to reformulate the model so it consists only of regular parameters and thus provides a tractable foundation on which to discuss estimation and inference easily. This consorts with the repeated observation in the literature that reduction to regular parameters (usually referred to as "reduced-form parameters") can substantially simplify the problems (Staiger and Stock, 1997; Stock and Wright, 2000; Chernozhukov et al., 2009; Magnusson and Mavroeidis, 2010; Magnusson, 2010; Guerron-Quintana et al., 2013; Andrews and Mikusheva, 2016a,b; Andrews, 2016; Cox, 2017, among many others); we generalize this observation to arbitrary semiparametric models and show that there exists an underlying regular parameter for every weakly regular parameter. However, underlying regular parameters are not unique, and statistical analyses based on different underlying parameters may yield different performances. This gives rise to the need for criteria to choose which underlying parameter to use.

We consider desirable properties of underlying parameters from two perspectives. Intuitively, a good underlying regular parameter would exhaustively contain all information about the weakly regular parameter that can be inferred by the model, *and* it would contain no irrelevant information that may lead to noisy analyses; this intuition parallels that of efficient influence functions of classical nuisance parameter theory. In light of this, we define an underlying parameter to be *sufficient* if knowing the value of its local parameter reveals as much information as knowing the weakly regular parameter. The key is to understand that information about the weakly regular parameter comes from two sources: the value of the weakly regular parameter and the very fact that it is identified. A sufficient underlying parameter would contain *both* pieces of information. Next, we define an underlying parameter to be *minimal* if knowing its local parameter does not reveal more information than knowing the weakly regular parameter. If it does, its estimation would create additional noise in an effort to estimate its unnecessary "nuisance" component. In short, the best underlying regular parameter would be minimal and sufficient. We show existence of minimal sufficient underlying parameters, provide a way to assess their sufficiency and minimality in general setups, and present examples of minimal sufficient underlying parameters.

With these concepts, we define a new notion of efficiency for estimating weakly regular parameters. Efficiency of estimation under weak identification has received little treatment in the literature. This is because non-Gaussianity and nonpivotality of the asymptotic distributions render the classical efficiency concepts, the convolution and minimax theorems, inapplicable, at least in their direct forms. Our formulation enables us to decompose estimation of weakly regular parameters into estimation of the minimal sufficient underlying regular parameters and their transformation. As the underlying regular parameters admit the classical convolution theorem, efficiency of their estimation can be discussed through the classical theory. Moreover, if the estimators of the underlying parameters contain unnecessary noise, then their transformations would also contain unnecessary noise. Such noise can then be eliminated by taking expectation with respect to it since the noise and the asymptotic distributions of efficient estimators are asymptotically independent. Conceptually, this corresponds to applying the Rao-Blackwell theorem to the local asymptotic representations of the estimators, exploiting the fact that the efficient asymptotic distributions of regular parameters are "sufficient" in the local expansion. The resulting conditional expectation estimators are, as a consequence, more concentrated toward the same means without affecting the size of the biases. We formalize this idea as a theorem and name it *local asymptotic Rao-Blackwellization (LAR)*. If such improvement is impossible, we call the estimators *weakly efficient*. We put the qualifier "weakly" as weakly efficient estimators are not unique. We also discuss relationship between *weak efficiency* and classical *efficiency*.

Most estimators in the literature can be written as functions of estimators of some underlying regular parameters and thus fall within the class of estimators covered by our results. We apply our results to heteroskedastic linear IV models and present examples of weakly efficient estimators. Conventional estimators such as two-stage least squares (2SLS), GMM, and Fuller as well as the unbiased estimator of Andrews and Armstrong (2017) are shown to be inefficient in the presence of heteroskedasticity and, under the availability of an efficient estimator of the reduced-form coefficients, admit transformations into weakly efficient estimators by LAR. We carry out simulation to investigate how weakly efficient estimators outperform their original estimators.

There is a large body of literature that studies the optimality of statistical procedures under weak identification. Müller and Wang (2017) study estimation under weak identification that minimizes the weighted average risk when the asymptotic distribution of the statistics is known. Armstrong (2016) analyzes identification strength in demand estimation and prescribes diagnostics. Moreira (2003) and Andrews et al. (2006, 2007) develop optimal conditional likelihood ratio (CLR) tests in linear IV models with normal homoskedastic errors. Müller (2011) studies efficient inference under a weak convergence assumption. Cattaneo et al. (2012) consider estimation and discuss nearly optimal tests in weakly identified linear IV models with independent but possibly non-Gaussian errors. Elliott et al. (2015) develop the power envelope in models with nuisance parameters and apply it to weakly identified linear IV models. There are also numerous studies about inference procedures that are robust to weak identification and identification failure in many settings, including Zivot et al. (1998, 2006), Wang and Zivot (1998), Kleibergen (2002, 2004, 2005, 2007), Dufour (2003), Dufour and Taamouti (2005), Mikusheva (2010), Chaudhuri and Zivot (2011), Guggenberger et al. (2012), Andrews and Cheng (2013, 2014), Andrews and Mikusheva (2014, 2015, 2016a,b), Qu (2014), and Cheng (2015). Also, some degree of efficient estimation under semi-strong identification is investigated (Antoine and Renault, 2009, 2012; Antoine and Lavergne, 2014).

The rest of the paper is organized as follows. Section 2 provides examples of weak identification in economics. Section 3 defines weak identification in semiparametric models, gives impossibility results, and introduces the notion of underlying regular parameters. Section 4 introduces sufficiency and minimality of underlying regular parameters. Section 5 derives LAR for the estimation of weakly regular parameters, whence we define weak efficiency. Section 6 discusses application of LAR to heteroskedastic linear IV models and provides simulation results. Section 7 concludes. The Appendix contains proofs. The Online Supplementary Appendix contains supplementary results.

2 WEAK IDENTIFICATION IN ECONOMICS

We encounter the problem of weak identification in various places in economics, from micro- to macroeconomic contexts. This section describes prominent examples of weak identification. Among them, the linear IV model (Example 1) is the most important one; it branches into special cases that capture various aspects of weak identification.

Example 1 (Linear IV)**.** Consider the IV regression model:

$$\begin{cases} y_i = x_i'\beta + \varepsilon_i, & \mathbb{E}[\varepsilon_i \mid z_i] = 0, \\ x_i' = z_i'\pi + v_i', & \mathbb{E}[v_i \mid z_i] = 0, \end{cases}$$

where y_i and ε_i are scalars, x_i, β, and v_i are $d \times 1$ vectors, z_i is a $k \times 1$ vector, π is a $k \times d$ full column rank matrix, and $k \geq d$. The first equation is called the *second-stage equation* and the second the *first-stage equation*; they are collectively called the *structural equations*. On the other hand, the *reduced-form equations* are obtained by substituting the first-stage equation into the second-stage:

$$\begin{cases} y_i = z_i'\pi\beta + u_i, & \mathbb{E}[u_i \mid z_i] = 0, \\ x_i' = z_i'\pi + v_i', & \mathbb{E}[v_i \mid z_i] = 0. \end{cases}$$

The first equation of the reduced-form equations may sometimes be referred to as the *second-stage equation* when there is no confusion. The model is said to be *just-identified* if $k = d$ and *overidentified* if $k > d$. We are interested in the parameter β, which is called the *structural parameter*.

The conditional moment restrictions $\mathbb{E}[u_i \mid z_i] = 0$ and $\mathbb{E}[v_i \mid z_i] = 0$ are the key identifying assumptions; they are sometimes replaced by the weaker versions of the unconditional moment restrictions $\mathbb{E}[z_i u_i] = 0$ and $\mathbb{E}[z_i v_i'] = 0$. The theory developed in this paper subsumes both cases while the demonstration of our theory in later sections will be based on the conditional moment restrictions. The second moments of all variables are assumed to be finite. The model is called *homoskedastic* if $\mathbb{E}[u_i^2 \mid z_i]$, $\mathbb{E}[u_i v_i \mid z_i]$, and $\mathbb{E}[v_i v_i' \mid z_i]$ do not depend on z_i; otherwise, it is *heteroskedastic*.[1]

Weak identification of the structural parameter β occurs when the correlation between the endogenous regressors x_i and the instruments z_i is weak, which is captured by conventional asymptotic embedding (Staiger and Stock, 1997): $\pi_n = O(1/\sqrt{n})$. Another possibility of weak identification is the case in which π_n does not vanish but approaches a rank deficient matrix at root-n: $\pi_n = \pi_0 + O(1/\sqrt{n})$, where π_0 is nonzero but not of full column rank (Andrews and Guggenberger, 2017, call this *joint weak identification*). Through suitable reparametrization and normalization, this model reduces to one where π_0 is a diagonal matrix whose upper $\ell \times \ell$ submatrix equals identity for some $\ell < d$ and all other elements zero (Section S.2). In the main text, we further

[1] With unconditional moment restrictions, homoskedasticity means the following weaker version: $\mathbb{E}[u_i^2 z_i z_i']$, $\mathbb{E}[u_i v_i \otimes z_i z_i']$, and $\mathbb{E}[v_i v_i' \otimes z_i z_i']$ are proportional to $\mathbb{E}[z_i z_i']$ (up to the Kronecker structure).

assume that π_0 is a zero matrix as in Staiger and Stock (1997). Let

$$\pi_n = \frac{\dot{\pi}}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)$$

for a $k \times d$ matrix $\dot{\pi}$ that is of full rank.[2] This embedding induces the asymptotics to the second-stage coefficients, $\pi_n \beta_n = \dot{\pi}\beta/\sqrt{n} + o(1/\sqrt{n})$ for a $d \times 1$ vector β. When the model is weakly identified, the 2SLS estimator $\hat{\beta}$ is no longer consistent for β. Staiger and Stock (1997) show that the 2SLS converges in distribution to a Cauchy-like distribution that depends on the local parameter. It is important to understand that $\hat{\beta}$ itself is $O_P(1)$, not the inflated version $\sqrt{n}(\hat{\beta} - \beta)$.

Example 2 (Nonlinear GMM)**.** Many structural models in economics identify parameters of interest in the form of a nonlinear moment equation. In particular, the parameter of interest $\beta \in \mathbb{R}^d$ is identified as a unique solution to

$$\mathbb{E}[M_i(\beta)] = m(\beta) = 0$$

for some random process M_i (e.g., $Z_i h(X_i; \beta)$ for some X_i and Z_i), indexed by β. Weak identification of β happens when the moment function m converges uniformly to a function that has multiple zeros at rate $\sqrt{n}$. Under weak identification, the nonlinear GMM estimator $\hat{\beta}$ is not consistent and converges in distribution to a nonstandard nonpivotal distribution (Stock and Wright, 2000).

Example 3 (Extremum estimation)**.** Another popular specification that arises from structural models is the *extremum estimation* model. The parameter of interest $\beta \in \mathbb{R}^d$ is identified as the minimizer of an unknown function of which we have an observable random estimator; the minimizer of the random function yields the estimator of β. This model is closely related to Example 2 and some models (including linear IV models) admit both representations as GMM models and extremum estimation models. Weak identification of β occurs when the objective function flattens out (partially or fully) as the sample size tends to infinity. Andrews and Cheng (2012) consider cases when one of the parameters parameterizes the identification strength of other parameters; Han and McCloskey (2017) offer a way to reparameterize some extremum estimation models into the framework of Andrews and Cheng (2012) when the source of identification failure is known; and Cox (2017) considers models that are doubly parameterized by

[2] In fact, it is not even necessary that π_n approach zero no faster than root-n, in which case it will induce asymptotic partial identification; see Section S.3 for further discussion.

structural and reduced-form parameters where the reduced-form parameters are always identified.

Example 4 (Differentiated product demand estimation)**.** Endogeneity resulting from the simultaneous determination of prices and quantities poses a problem in demand estimation in industrial organization. In a situation in which one observes characteristics of many markets (but not individual purchasing behaviors), this endogeneity is often solved by using characteristics of other products as instruments for the endogenous prices (Berry et al., 1995). To invoke asymptotic approximation, the limit is often considered in the number of products tending to infinity, the so-called "large market asymptotics."

Armstrong (2016) shows, however, that the strength of these instruments is sensitive to how many products there are; in particular, when the number of products diverges along with the number of markets, the demand parameters may exhibit behaviors of strong identification, weak identification, or identification failure, depending on the relative rate of growth of the numbers of products and markets. A distinct feature of this example is that weak or non-identification asymptotics arises as a consequence of an equilibrium outcome, rather than as a purely statistical consideration of approximation.

Example 5 (Limited information macroeconomic models)**.** Modern DSGE models contain three key equations: a Phillips curve, an Euler equation, and a monetary policy rule. When we estimate parameters of a DSGE model, we often take one of the equations individually and estimate via GMM with a particular choice of instruments. As this makes use of only a part of the full structure of the DSGE model, it is called the *limited information approach* and allows one to conduct estimation with a minimal set of assumptions about the nature of macroeconomic dynamics.

Increasing attention is given to the fact that these instruments have the tendency to be weak. To name a few examples, Dufour et al. (2006) and Nason and Smith (2008) analyze the case of potential weak identification in the context of New Keynesian Phillips curve estimation; Yogo (2004) in the context of Euler equation estimation; and Mavroeidis (2010) in the context of monetary policy estimation.

Weak identification is thus widely observed in economics. There are also other instances of "weak identification" that actually entail partial identification in the limit. Our main focus in this paper is on "pure" weak identification; see Section S.3 for more discussions on this phenomenon.

Suppose we observe i.i.d. random variables $X_1, \dots, X_n$ from the sample space $(\mathcal{X}, \mathscr{A})$. The set of possible distributions of each X is denoted by $\mathcal{P}$ and is called the *model*. To obtain fruitful asymptotics around a distribution $P \in \mathcal{P}$, we consider a *path*[3] of distributions $Q_t \in \mathcal{P}$ indexed by a real number $t \in (0, 1]$ that is *differentiable in quadratic mean (DQM)* at P, that is, there exists a measurable function $g : \mathcal{X} \to \mathbb{R}$ such that[4]

$$\int_{\mathcal{X}} \left[\frac{dQ_t^{1/2} - dP^{1/2}}{t} - \frac{1}{2} g dP^{1/2} \right]^2 \longrightarrow 0 \qquad \text{as} \qquad t \to 0.$$

This convergence is denoted by $Q_t \to^{\text{DQM}} P$, and we call g the *(model) score* induced by the path $\{Q_t\}$.[5] The idea behind asymptotic approximation theory is that the path of "alternatives" $\{Q_t\}$ that approaches P at the same rate as the path of "samples" $\{\hat{P}_n\}$ is not deterministically discriminable in the limit and hence yields an approximation that reflects finite sample uncertainty; therefore, in many examples, it is suggestive to understand $t = 1/\sqrt{n}$ and in a minor abuse of notation to denote $Q_{1/\sqrt{n}}$ by Q_n.

We often do not consider every possible path in $\mathcal{P}$;[6] let $\mathscr{P}_P$ denote the set of paths we consider that tend to P in DQM. Since there is little chance of misunderstanding, we hereafter denote $\{Q_t\}$ simply by Q_t, for example, $Q_t \in \mathscr{P}_P$; therefore, Q_t can refer to the entire path $\{Q_t\}$ or an element Q_t of the path for a specific t, depending on the context. The set $\dot{\mathcal{P}}_P$ of scores g induced by the paths in $\mathscr{P}_P$ is called the *tangent set* at P. It is clear from the definition of scores that $\dot{\mathcal{P}}_P$ is a subset of $L_2(P)$.[7] Depending on the structure of $\mathscr{P}_P$, the tangent set might be a linear space, a cone,[8] or just a set without much structure; we assume that we can always augment $\mathscr{P}_P$ linearly so that the induced tangent set will be linear. For this reason, we call the tangent set the *tangent space*. The tangent space can be considered the local approximation of the model by a linear vector space. Finally, a parameter $\psi : \mathcal{P} \to \mathbb{D}$ is defined as a map from the model $\mathcal{P}$ to a Banach space $\mathbb{D}$.

If the parameter $\psi : \mathcal{P} \to \mathbb{D}$ is differentiable in a suitable sense, by the chain rule we may approximate the change in the parameter along any path by a linear map from

[3] A path is also called a *(parametric) submodel*.

[4] The integral is understood with respect to some σ-finite measure dominating P and Q_t, and dP and dQ_t are the Radon-Nikodym derivatives of P and Q_t with respect to it.

[5] Throughout the paper, dependence of g on $\{Q_t\}$ will be implied by the context.

[6] See, e.g., Bickel and Ritov (2000).

[7] In this sense, $\dot{\mathcal{P}}$ is the set of *equivalence classes* of scores, to be precise.

[8] A subset X of a linear space is called a *cone* if $x \in X$ implies $ax \in X$ for every $a > 0$.

the tangent space $\dot{\mathcal{P}}_P$ to the parameter space $\mathbb{D}$. Any infinitesimal perturbation of distribution P then leads to a linear perturbation of parameter ψ. Such a parameter is known to behave well and is said to be *regular*. This case is well studied in the literature (Groeneboom and Wellner, 1992; Bickel et al., 1993; Van der Vaart, 1988, 1998; Kosorok, 2008), and we will make good use of it in the study of weak identification. The appropriate notion of differentiability is given as follows.

Definition (Regular parameter)**.** A parameter $\psi : \mathcal{P} \to \mathbb{D}$ is *regular* (or *differentiable*) at P relative to $\mathscr{P}_P$ if there exists a continuous linear map $\dot{\psi}_P : \dot{\mathcal{P}}_P \to \mathbb{D}$ such that[9]

$$\frac{\psi(Q_t) - \psi(P)}{t} \longrightarrow \dot{\psi}_P g \qquad \text{for every} \qquad Q_t \in \mathscr{P}_P.$$

The derivative map $\dot{\psi}_P$ is called the *local parameter* of ψ. The adjoint map $\dot{\psi}_P^* : \mathbb{D}^* \to \overline{\dot{\mathcal{P}}_P}$ is called the *efficient influence map* of ψ, where $\mathbb{D}^*$ is the dual space of $\mathbb{D}$ and $\overline{\dot{\mathcal{P}}_P}$ the completion of $\dot{\mathcal{P}}_P$.[10]

Remark. In the classical context, the tangent set "represents" the set of paths (Kosorok, 2008, Section 18.1), so regularity (differentiability) is often defined "relative to the tangent set" (Van der Vaart, 1998, Chapter 25; Kosorok, 2008, Section 18.1). In the context of weak identification, however, the corresponding tangent set does not represent the set of paths (see the next section); therefore, we keep the original wording "relative to the set of paths" from Van der Vaart (1991b). The word "regular" is taken from Van der Vaart and Wellner (1996, Chapter 3.11).

3.1 Weakly Regular Parameters

Now we define a weakly identified parameter. When we talk about weak identification, often have we in mind a situation in which the sequence of distributions converges to a point of identification failure. In this respect, the weakly identified parameter may only be defined on a subset $\mathcal{P}_\beta$ of $\mathcal{P}$ where the difference $\mathcal{P} \setminus \mathcal{P}_\beta$ represents all points of identification failure. Accordingly, the path cannot fall outside of the submodel $\mathcal{P}_\beta$, and the tangent set must be restricted in order for the weakly identified parameter to

[9] If $\mathscr{P}_P$ is the set of *all* possible paths in $\mathcal{P}$, then regularity of ψ is equivalent to Hadamard differentiability of ψ.

[10] The function $\tilde{\psi}_P : \mathcal{X} \to \mathbb{D}$ such that $\dot{\psi}_P^* \delta^* = \delta^* \tilde{\psi}_P$ for every $\delta^* \in \mathbb{D}^*$ is called the *efficient influence function* of ψ (Bickel et al., 1993, Section 5.2). The qualifier *efficient* is justified in the context of the convolution theorem as remarked in Section 5. Kosorok (2008, Section 18.1) also gives alternative definitions (interpretations) of efficient influence functions in the context of functional parameters.

be well defined. Computing the corresponding tangent subset requires care, however, since too rapid an approach to the point of identification failure must be avoided. Somewhat counterintuitively, the tangent set pertinent to $\mathcal{P}_\beta$ cannot be defined as the set of all scores induced by the paths taking values in $\mathcal{P}_\beta$; it is the set of all scores that are *not* induced by the paths *not* taking values in $\mathcal{P}_\beta$.

Definition (Pertinent tangent cone)**.** The tangent set $\dot{\mathcal{P}}_{P,\beta} \subset \dot{\mathcal{P}}_P$ *pertinent to* the submodel $\mathcal{P}_\beta$ at $P \in \mathcal{P}$, possibly $P \in \mathcal{P} \setminus \mathcal{P}_\beta$, is the set of scores $g \in \dot{\mathcal{P}}_P$ such that there does not exist a path in $\mathscr{P}_P$ that takes values in $\mathcal{P} \setminus \mathcal{P}_\beta$ and induces g. Define $\mathscr{P}_{P,\beta}$ to be the set of paths in $\mathscr{P}_P$ that take values in $\mathcal{P}_\beta$ and induce scores in $\dot{\mathcal{P}}_{P,\beta}$.

The following example illustrates why we need this circuitous definition.

Example 1 (Linear IV, continued)**.** Consider a simple parametric linear IV model with $d = k = 1$ and $(u, v) \sim N(0, I_2)$. So $\mathcal{P}$ is the set of distributions of (x, y, z) such that $(y - z\pi\beta, x - z\pi) \sim N(0, I_2)$ for some $\pi, \beta \in \mathbb{R}$ and $\mathcal{P}_\beta$ is the subset of $\mathcal{P}$ such that $\pi \neq 0$. Consider the asymptotic embedding $\pi_n = \dot\pi/n$. The path of this embedding is $dQ_n = \frac{1}{2\pi}\exp\bigl(-\frac{(y-z\dot\pi\beta/n)^2+(x-z\dot\pi/n)^2}{2}\bigr)$, which converges in DQM to the point of identification failure $dP = \frac{1}{2\pi}\exp\bigl(-\frac{y^2+x^2}{2}\bigr)$ with the score $\sqrt{n}\frac{dQ_n-dP}{dP} \to 0$. Although Q_n takes values only on $\mathcal{P}_\beta$, its score can also be induced by the path $\tilde{Q}_n \equiv P$, which should be excluded. Therefore, $\dot{\mathcal{P}}_{P,\beta}$ cannot be taken as the set of all scores induced by paths in $\mathcal{P}_\beta$, but as the set of scores not induced by paths not in $\mathcal{P}_\beta$.

From the observation that P is not in $\mathcal{P}_\beta$, we see that $\dot{\mathcal{P}}_{P,\beta}$ is only a cone.

Lemma 1. *$\dot{\mathcal{P}}_{P,\beta}$ and $\dot{\mathcal{P}}_P \setminus \dot{\mathcal{P}}_{P,\beta}$ are cones.*

Remark. In classical asymptotic theory, the limit distribution P is often regarded as the "null hypothesis" and the path Q_t as a drifting sequence of "alternatives." When it comes to weak identification, in contrast, both the null and alternatives reside as paths in $\mathscr{P}_{P,\beta}$; P is merely a point of reference of identification failure.

If the set of paths $\mathscr{P}_P$ is much richer than $\mathscr{P}_{P,\beta}$ in a way that $\operatorname{Span}\dot{\mathcal{P}}_{P,\beta}$ is a strict subset of $\dot{\mathcal{P}}_P$, then there exists a superfluously rich side of the model on which β is not even defined. Since it is meaningless to consider such parts of the model when one's focus is on the parameter β, we assume innocuously that $\overline{\operatorname{Span}}\,\dot{\mathcal{P}}_{P,\beta} = \overline{\dot{\mathcal{P}}_P}$.[11]

[11] Later on we define the underlying regular parameter on the whole of $\mathcal{P}$, so it is actually harmful to require that that parameter be regular on the unconsidered realm of the model.

Now we define the weakly identified parameter under the name *weakly regular parameter*.[12] We henceforth shun the use of the qualifier "weakly identified" since weak identification in the literature may not always exclude cases of in fact *no* identification (e.g., Moreira, 2009; Andrews and Cheng, 2013, 2014; Han and McCloskey, 2017). In this paper, we assume that weakly regular parameters are identified at every fixed n in that there exists a unique value of the parameter for any given distribution Q_n belonging to $\mathcal{P}_\beta$. Moreover, we assume that the parameters remain identified in the limit in the sense that there exists a unique value of the parameter for each score g in $\dot{\mathcal{P}}_{P,\beta}$. See Section S.3 for cases that entail partial identification in the limit. Let $\mathbb{B}$ be another Banach space on which a weakly regular parameter will be defined.

Definition (Weakly regular parameter)**.** A parameter $\beta : \mathcal{P}_\beta \to \mathbb{B}$ is *weakly regular* at $P \in \mathcal{P}$, possibly $P \in \mathcal{P} \setminus \mathcal{P}_\beta$, relative to $\mathscr{P}_{P,\beta}$ if there exists a continuous map $\beta_P : \dot{\mathcal{P}}_{P,\beta} \to \mathbb{B}$ that is homogeneous of degree zero such that

$$\beta(Q_t) \longrightarrow \beta_P(g) \qquad \text{for every} \qquad Q_t \in \mathscr{P}_{P,\beta}.$$

The idea behind the definition is that a weakly identified parameter is necessarily accompanied by the fundamental uncertainty of the model summarized by the score. This captures our observation that the estimator of a weakly identified parameter does not converge to the true value but retains some randomness in the limit. In this sense, asymptotics of weakly regular parameters is "global" in nature, and nonlinearity often observed in the literature (e.g., Cox, 2017) arises compellingly from nonlinearity of the map β_P, resulting in a nonstandard limit distribution. The definition clarifies the extent of such randomness and nonlinearity that are inherent to the model (and hence may not be compensated for by the choice of estimation methods); see Theorem 2 for the consequential, fundamental impossibility results.

Remark. Since we regard the reference point of identification failure P and the sets of paths $\mathscr{P}_P$ and $\mathscr{P}_{P,\beta}$ as fixed, we often mention (weak) regularity of a parameter without referring to the point of evaluation and the relative collection of paths.

Remark. Being a continuous map, a regular parameter is trivially weakly regular; that is, if $\psi : \mathcal{P} \to \mathbb{D}$ is regular, then $\psi(Q_t) \to \psi_P(g)$ where $\psi_P(g) \equiv \psi(P)$. Also, if β is a nontrivial weakly regular parameter, i.e., β_P is nonconstant, then β_P cannot be linear since a linear function that is homogeneous of degree zero must be identically zero.

[12] Not to be confused with *weak regularity* of an estimator defined in Van der Vaart (1988, Section 2.2) or Bickel et al. (1993, Definition 5.2.6).

Remark. A weakly regular parameter is not only nondifferentiable at P; it is even discontinuous at P (unless β_P is trivially constant). This is in contrast to the literature on continuous but not (fully) differentiable parameters (Hirano and Porter, 2012; Fang, 2015, 2016; Fang and Santos, 2015; Hong and Li, 2017). This discontinuity plays a key role in one of the impossibility results in Theorem 2; we exploit the "continuity" of asymptotic distributions implied by Le Cam's third lemma.

Remark. Homogeneity of β_P is a natural consequence of dependence on g. Since $\beta(Q_{kt})$ for fixed $k > 0$ converges to the same limit as $\beta(Q_t)$, we have $\beta_P(kg) = \beta_P(g)$.[13] Continuity of β_P is required only on its domain $\dot{\mathcal{P}}_{P,\beta}$; it is not possible to extend β_P continuously to the whole of $\dot{\mathcal{P}}_P$ unless β_P is trivially constant.

Now we look at examples. We construct paths that encapsulate the asymptotic embeddings discussed in Section 2 and show that the parameters can be written as continuous and homogeneous functions of the model scores.

Example 1 (Linear IV, continued)**.** Let $\mathcal{P}_{uvz}$ be the set of probability distributions P_{uvz} on (u, v', z) with second moments such that $\mathbb{E}[u \mid z] = 0$, $\mathbb{E}[v \mid z] = 0$, $\mathbb{E}[zz']$ is invertible, and dP_{uvz} differentiable almost everywhere in (u, v').[14] The model $\mathcal{P}$ is the set of probability distributions P on the observable elements (x, y, z) such that

$$dP(x, y, z) = dP_{uvz}(y - z'\pi\beta, x' - z'\pi, z) \ \text{ for some } \ P_{uvz} \in \mathcal{P}_{uvz},\ \pi \in \mathbb{R}^{k\times d},\ \beta \in \mathbb{R}^d.$$

The distribution of z is characterized by P_{uvz}; that of y by P_{uvz} and $\pi\beta$; that of x by P_{uvz} and π; thus, we have "parameterized" the semiparametric model $\mathcal{P}$ by three parameters P_{uvz}, π, and β. The submodel $\mathcal{P}_\beta$ is the subset of $\mathcal{P}$ of all distributions with $\det(\pi'\pi) \neq 0$. Let $\pi(P) = 0$ at $P \in \mathcal{P} \setminus \mathcal{P}_\beta$. We consider a path Q_t toward P such that $[\pi(Q_t) - 0]/t$ converges to some element $\dot{\pi}$ in $\mathbb{R}^{k\times d}$. If $\det(\dot{\pi}'\dot{\pi}) = 0$, then there exists a path taking values in $\mathcal{P} \setminus \mathcal{P}_\beta$ that yields the same limit of π; this means that for every $Q_t \in \mathscr{P}_{P,\beta}$ we have $\det(\dot{\pi}'\dot{\pi}) \neq 0$. Such a path can be represented as $dQ_t(x, y, z) = dQ_{t,uvz}(y - z'(t\dot{\pi}_t\beta_t), x' - z'(t\dot{\pi}_t), z)$ for some path $Q_{t,uvz}$ in $\mathcal{P}_{uvz}$, and $\dot{\pi}_t \to \dot{\pi}$ and $\beta_t \to \beta$. Being a probability distribution by itself, $Q_{t,uvz}$ has its own "model score" g_{uvz}. To see what it is like, note that the only essential restriction of

[13] Fang and Santos (2015) observe a related fact that a directional derivative must be homogeneous of degree one.

[14] Differentiability is not necessary as long as each one-dimensional parametric submodel is differentiable in quadratic mean (see, e.g., Pollard, 1997; Van der Vaart, 1998, Section 7.2). Here we assume this for illustration of derivation of scores. See also Van der Vaart (1988, Section 1.2 and Appendix A.2).

$Q_{t,uvz}$ is $\int zudQ_{t,uvz} = 0$ and $\int zv'dQ_{t,uvz} = 0$. Therefore,

$$0 = \frac{1}{t}\left(\int zudQ_{t,uvz} - \int zudP_{uvz}\right) \longrightarrow \int zug_{uvz}dP_{uvz} = \mathbb{E}_P[zug_{uvz}].$$

Similarly, one sees that $\mathbb{E}_P[zv'g_{uvz}] = 0$. Therefore, the set of scores of the parameter P_{uvz} consists of all appropriate scores that satisfy these two restrictions.[15] Using this, the model score for the path of interest Q_t can be calculated as

$$\begin{aligned}\frac{dQ_t - dP}{tdP} &= \frac{dQ_{t,uvz}(y - z'(t\dot{\pi}_t\beta_t), x' - z'(t\dot{\pi}_t)) - dP_{uvz}(y - z'(t\dot{\pi}_t\beta_t), x' - z'(t\dot{\pi}_t))}{tdP} \\ &\quad + \frac{dP_{uvz}(y - z'(t\dot{\pi}_t\beta_t), x' - z'(t\dot{\pi}_t)) - dP_{uvz}(y, x)}{tdP} \\ &\longrightarrow g = g_{uvz} - z'\dot{\pi}\beta\frac{dP_{uvz,u}}{dP} - z'\dot{\pi}\frac{dP_{uvz,v}}{dP},\end{aligned} \tag{1}$$

where $P_{uvz,u}$ and $P_{uvz,v}$ represent the partial derivatives of P_{uvz} with respect to u and v. Observe that by integration by parts $\mathbb{E}_P[zug] = -\int zuz'\dot{\pi}\beta dP_{uvz,u} - \int zuz'\dot{\pi}dP_{uvz,v} = \int zz'\dot{\pi}\beta dP = \mathbb{E}_P[zz']\dot{\pi}\beta$. Similarly, $\mathbb{E}_P[zv'g] = \mathbb{E}_P[zz']\dot{\pi}$. Therefore, β_t converges to

$$\beta = (\mathbb{E}_P[zz']^{-1}\mathbb{E}_P[zv'g])^{\rightarrow}(\mathbb{E}_P[zz']^{-1}\mathbb{E}_P[zug]) =: \beta_P(g),$$

where $A^{\rightarrow}$ denotes the left inverse of A. This map is continuous on $\dot{\mathcal{P}}_{P,\beta}$ and homogeneous of degree zero but nonlinear.

Thus, we have shown that β defined on $\mathcal{P}_\beta$ converges to a continuous and homogeneous function β_P of a score along every path in $\mathscr{P}_{P,\beta}$, meaning that β is weakly regular at P relative to $\mathscr{P}_{P,\beta}$.

Example 2 (Nonlinear GMM, continued)**.** Let $\mathbb{D}$ be the space of bounded continuous functions from $\mathbb{R}^d$ to $\mathbb{R}$. Let $\mathcal{P}_m$ be the set of distributions P_m of zero-mean stochastic processes taking values in $\mathbb{D}$. The model $\mathcal{P}$ can be represented as the set of distributions P on M_i such that $dP(M_i) = dP_m(M_i - m)$ for some $m \in \mathbb{D}$. The submodel $\mathcal{P}_\beta$ is the subset of $\mathcal{P}$ whose elements have a mean function in the subset $\mathbb{D}_\beta$ of $\mathbb{D}$ of functions m that have unique zeros. Recall that we are interested in paths along which the moment functions vanish at rate $\sqrt{n}$, that is, $m_n(\cdot) = 0(\cdot) + \frac{\dot{m}(\cdot)}{\sqrt{n}}$. Letting $t = 1/\sqrt{n}$, write

$$\frac{m_t(\cdot) - 0(\cdot)}{t} \longrightarrow \dot{m}(\cdot) \in \mathbb{D}_\beta.$$

[15] See Van der Vaart (1998, Example 25.28) for related discussion.

In short, we are interested in the paths $Q_t \to^{\mathrm{DQM}} P$ that yield the moment function (as a parameter $m : \mathcal{P} \to \mathbb{D}$) to be regular. Thus, for some path $Q_{t,m}$ to P_m in $\mathcal{P}_m$, we can write the path Q_t to P in $\mathcal{P}_\beta$ as

$$dQ_t(M_i) = dQ_{t,m}(M_i - m_t) \qquad \text{where} \qquad \frac{m_t - 0}{t} \longrightarrow \dot{m} \in \mathbb{D}_\beta.$$

Meanwhile, since $\mathbb{E}_{Q_t}[M_i] = m_t$ and $\mathbb{E}_P[M_i] = 0$, we have

$$\frac{m_t - 0}{t} = \int M_i \frac{dQ_t - dP}{t} \longrightarrow \int M_i g dP = \mathbb{E}_P[M_i g] = \dot{m}.$$

Therefore, $\mathbb{E}_P[M_i g](\theta) = 0$, that is, $\theta = \theta_P(g)$ is defined as the zero of $\mathbb{E}_P[M_i g]$. This is a nonlinear and continuous map on $\dot{\mathcal{P}}_{P,\beta}$ that is homogeneous of degree zero. Again, we conclude that θ is weakly regular at P.

Example 6 (Testing local hypotheses)**.** For a regular parameter $\psi : \mathcal{P} \to \mathbb{D}$, consider the hypothesis: $H_0 : \psi(P) \in \mathbb{D}_0$ vs $H_1 : \psi(P) \in \mathbb{D}_1$, where $\mathbb{D}_0 \cap \mathbb{D}_1 = \varnothing$ and $\mathbb{D}_0 \cup \mathbb{D}_1 \subset \mathbb{D}$. This induces a local testing problem at a boundary P of the following form: $H_0 : \dot{\psi}_P g \in \mathbb{D}_{P,0}$ vs $H_1 : \dot{\psi}_P g \in \mathbb{D}_{P,1}$ with $\mathbb{D}_{P,0} \cap \mathbb{D}_{P,1} = \varnothing$ and $\mathbb{D}_{P,0} \cup \mathbb{D}_{P,1} \subset \mathbb{D}$. These testing problems can be represented by a weakly regular parameter $\beta : \mathcal{P}_\beta \to [0,1]$ such that $\beta(P) := \mathbb{1}\{\psi(P) \in \mathbb{D}_0\}$, where $\mathcal{P}_\beta := \psi^{-1}(\mathbb{D}_0 \cup \mathbb{D}_1)$, and its corresponding limit $\beta_P(g) := \mathbb{1}\{\dot{\psi}_P g \in \mathbb{D}_{P,0}\}$. If $\mathbb{D}_{P,0}$ and $\mathbb{D}_{P,1}$ are cones and their boundary is excluded, β can be considered weakly regular.

3.2 Fundamental Impossibility

The utility of our theoretical formalism can be readily harvested in the following theorem. It gives a formal proof to the conventional wisdom that a "weakly identified" parameter cannot be estimated consistently or pivotally—but not as a characteristic of a specific estimation method—as a direct consequence of the characteristic of the model (see, *inter alia*, Phillips, 1984, 1989; Staiger and Stock, 1997; Stock and Wright, 2000; Guggenberger and Smith, 2005; Andrews and Cheng, 2012; Cox, 2017).[16] This result can also be viewed as a generalized proof of nonexistence of a consistent test conjectured by Hahn et al. (2011).[17] Distinct but related are the impossibility results by

[16] Consistent estimation may be possible in linear IV models if the number of weak instruments tends to infinity and some other conditions are met (Chao and Swanson, 2005; Newey and Windmeijer, 2009). In this case, the structural parameter is not weakly regular.

[17] Their setup can be translated into ours by taking $\mathbb{B}$ to be the product space for two estimators compared in the Hausman test, observing that a regular parameter is trivially weakly regular.

Dufour (1997) and Hirano and Porter (2015); their setup is a generalization of the weak linear IV structure whereas our setup is a generalization of the weak identification phenomena. Indeed, Dufour (1997) shows nonexistence of bounded confidence sets (which is "stronger" than nonexistence of consistent estimators) while there exist weakly regular parameters that admit bounded confidence sets (Example 6); Hirano and Porter (2015) show the impossibility of unbiased estimation while there exist weakly regular parameters that admit unbiased estimation (Andrews and Armstrong, 2017).

Theorem 2 (Impossibility of consistent and equivariant estimation)**.** *There is no consistent sequence of estimators of a nontrivial weakly regular parameter; there is no consistent sequence of nontrivial tests of a nontrivial weakly regular parameter; there is no equivariant-in-law sequence of estimators of a nontrivial weakly regular parameter with a separable limit law.*

Remark. The assumption of separability of the limit law is without "great loss of generality;" we treat it as general impossibility of equivariant estimation in the main text. See discussions of Van der Vaart and Wellner (1996, Theorem 1.3.10).

Impossibility of equivariant estimation implies that the asymptotic distribution of any estimator of a weakly regular parameter, when centered at the true value, is nonpivotal and not consistently estimable. However, it does not preclude the possibility that there exist *test statistics* whose distributions are pivotal or consistently estimable (Kleibergen, 2002, 2005). In fact, almost any reasonable inference procedure would be based on statistics whose asymptotic distributions are known or at least estimable; hence, the problem of estimation and the problem of inference bear quite distinct aspects when it comes to weakly regular parameters.[18] This partly explains the specialty of current literature on inference problems pertaining to weak identification.

3.3 Underlying Regular Parameters

The idea on analyzing the weak regularity of a parameter is that in many cases there exists another parameter that is regular and whose local parameter controls the limit behavior of the weakly regular parameter. In the literature, such a parameter is known

[18] This is in stark contrast to the classical context of regular parameters, in which efficient estimation and "efficient" inference are closely related to each other. Van der Vaart (1998, Chapter 25) states that "[s]emiparametric theory has little more to offer than the comforting conclusion that tests based on efficient estimators are efficient."

as the "reduced-form parameter" and is considerably utilized in various robust inference procedures under weak identification (*inter alia*, Magnusson and Mavroeidis, 2010; Mavroeidis, 2010; Guerron-Quintana et al., 2013; Andrews and Mikusheva, 2016a; Armstrong, 2016; Cox, 2017).[19] Then, the weakly regular parameter acts by itself as (a transformation of) the local parameter of some "underlying" regular parameter; in other words, it is sufficient to know the value of (the local parameter of) the underlying regular parameter in order to infer the value of the weakly regular parameter in the local expansion around the point of identification failure. We now formalize this idea, starting with the following definition.

Definition (Underlying regular parameter)**.** Let $\beta : \mathcal{P}_\beta \to \mathbb{B}$ be weakly regular at $P \in \mathcal{P}$ relative to $\mathscr{P}_{P,\beta}$. The parameter $\psi : \mathcal{P} \to \mathbb{D}$ is an *underlying (regular) parameter* for β at P relative to $\mathscr{P}_P$ if it is regular at P relative to $\mathscr{P}_P$ and there exists a continuous map $\beta_{P,\psi} : \mathbb{D}_\beta \to \mathbb{B}$ that is homogeneous of degree zero such that

$$\beta(Q_t) \longrightarrow \beta_{P,\psi}(\dot{\psi}_P g) \qquad \text{for every} \qquad Q_t \in \mathscr{P}_{P,\beta},$$

where $\mathbb{D}_\beta$ is the subset of $\mathbb{D}$ on which the local parameter of ψ takes values, that is, $\{\delta \in \mathbb{D} : \delta = \dot{\psi}_P g \text{ for some } g \in \dot{\mathcal{P}}_{P,\beta}\}$.

Remark. There exists a map $\beta_\psi : \mathbb{D} \to \mathbb{B}$ such that $\beta(Q_t) = \beta_\psi(\psi(Q_t)) + o(1)$ and thus β_ψ admits approximation at $\psi(P)$ by a homogeneous function β_ψ; set, e.g., $\beta_\psi(\delta) = \beta_{P,\psi}(\delta - \psi(P))$. In many applications, moreover, there exists an exact direct representation $\beta(Q_t) = \beta_\psi(\psi(Q_t))$ with some function $\beta_\psi : \mathbb{D} \to \mathbb{B}$ that is "locally homogeneous" at $\psi(P)$. For instance, we show below in Example 1 that the structural parameter β in linear IV models has a direct representation by the underlying regular parameter taken to be the reduced-form coefficients.

Remark. An underlying parameter is regular and hence susceptible to various estimation and inference techniques developed in statistics (Bickel et al., 1993; Van der Vaart, 1998; Kosorok, 2008).

Remark. In the context of extremum estimation, Cox (2017) defines "reduced-form parameters" as functions of "structural parameters." We take the opposite route: a weakly regular parameter approaches a function of (the local parameter of) an underlying regular parameter.

[19] On the other hand, the weakly regular parameter is often referred to as the "structural parameter."

This definition requires that knowing the local parameter of the underlying regular parameter is enough to recover the value of the weakly regular parameter; the reduction of information from knowing g to knowing $\dot{\psi}_P g$ does not impair the ability to discern β in the limit. With this definition, several questions arise: Does an underlying parameter always exist? How do we find an underlying regular parameter? How can we check whether a particular parameter is an underlying regular parameter? Which underlying regular parameter is better than another? We answer the first two questions in the remainder of this section and the rest in the next section.

The first question turns out to be straightforward. If one takes the root likelihood ratio $Q \mapsto dQ^{1/2}/dP^{1/2}$ to be a parameter, one can trivially claim that there always exists an underlying regular parameter for any weakly regular parameter. However, whether there exists an underlying regular parameter that admits root-n consistent estimation is a different matter. For this, we need to search for a good underlying parameter in each model separately.

Lemma 3 (Existence of underlying regular parameter)**.** *Let $\beta : \mathcal{P}_\beta \to \mathbb{B}$ be weakly regular. Then, there exist a Banach space $\mathbb{D}$ and an underlying regular parameter $\psi : \mathcal{P} \to \mathbb{D}$ for β.*

Now we look into underlying regular parameters in examples. We see that the natural parameters that appear in each example above constitute underlying regular parameters; however, the linear IV case contains other interesting and equally natural underlying parameters that deserve attention.

Example 1 (Linear IV, continued)**.** Define $\psi := (\psi_1, \mathrm{vec}(\psi_2)) := (\pi\beta, \mathrm{vec}(\pi))$ to be the $(k + kd) \times 1$ parameter. This is the so-called "reduced-form coefficients" in linear IV models. Many papers on weak instruments start from considering reduced-form coefficients and their estimators. Let us verify that ψ is indeed an underlying regular parameter for β. Recall from Example 1 in the previous section that $\dot{\pi}\beta = \mathbb{E}_P[zz']^{-1}\mathbb{E}_P[zug]$ and $\dot{\pi} = \mathbb{E}_P[zz']^{-1}\mathbb{E}_P[zv'g]$, that is, the local parameter of $(\pi\beta, \mathrm{vec}(\pi))$ is a continuous linear functional of the score; therefore, ψ is regular with $\dot{\psi}_P g = (\dot{\psi}_1, \mathrm{vec}(\dot{\psi}_2)) = (\dot{\pi}\beta, \dot{\pi})$. The submodel $\mathcal{P}_\beta$ contains all distributions in $\mathcal{P}$ that satisfy $\det(\dot{\psi}_2'\dot{\psi}_2) \neq 0$ and $(I_k - \dot{\psi}_2\dot{\psi}_2^{\to})\dot{\psi}_1 = 0$. Since $\beta_P(g) = \dot{\psi}_2^{\to}\dot{\psi}_1$, ψ is an underlying regular parameter for β with $\beta_{P,\psi}(\dot{\psi}_1, \dot{\psi}_2) = \dot{\psi}_2^{\to}\dot{\psi}_1$ defined on $\mathbb{D}_\beta = \{(\dot{\psi}_1, \mathrm{vec}(\dot{\psi}_2)) \in \mathbb{R}^k \times \mathbb{R}^{k\times d} : \det(\dot{\psi}_2'\dot{\psi}_2) \neq 0,\, (I_k - \dot{\psi}_2(\dot{\psi}_2'\dot{\psi}_2)^{-1}\dot{\psi}_2')\dot{\psi}_1 = 0\}$. In fact, this underlying parameter admits the direct representation $\beta(Q_t) = \psi_2(Q_t)^{\to}\psi_1(Q_t)$.

There are other choices of an underlying regular parameter. Let π_d be the first $d \times d$ submatrix of the $k \times d$ matrix π. Then $\psi_{(d)} := (\pi_d \beta, \operatorname{vec}(\pi_d))$ is also an underlying regular parameter since $\beta_P(g) = \dot{\psi}_{\vec{(d)},2} \dot{\psi}_{(d),1}$ with an analogous definition of $\dot{\psi}_{(d)}$. The $d \times d$ matrix can in fact be any (nondegenerate) combination of coefficients on k instruments, as long as one can recover the value of β. This is to say that in overidentified linear IV models ($k > d$), there are many natural choices of underlying regular parameters.

Example 2 (Nonlinear GMM, continued)**.** By the definition of θ_P, we can guess that the moment function $m : \Theta \to \mathbb{R}$ is an underlying regular parameter for θ. As seen earlier, it is regular at 0 relative to the paths of interest since $\frac{tm(Q_t)-0}{t} = \frac{tm_t-0}{t} \to \dot{m} = \mathbb{E}_P[M_i g]$, which is a continuous linear functional of the score. Moreover, since $\theta_P(g)$ is the zero of $\mathbb{D}_P[M_i g]$, it can also be written as the zero of $\dot{m}$. Thus, by taking $\theta_{P,m}(\dot{m})$ to be the zero of $\dot{m}$ defined on the subset $\mathbb{D}_\beta \subset \mathbb{D}$ of functions with unique zeros, one sees that the moment function is an underlying regular parameter. This underlying parameter admits the direct representation $\theta(Q_t) = \theta_{P,m}(m(Q_t))$.

4 MINIMAL SUFFICIENT UNDERLYING REGULAR PARAMETERS

This section characterizes desirable properties of underlying regular parameters. We say that an underlying parameter is *sufficient* if it contains all the information captured by the weakly regular parameter: its identification and value. We say that an underlying parameter is *minimal* if it does not contain information that is irrelevant to the weakly regular parameter; in other words, a minimal underlying parameter does not contain a "nuisance parameter." Putting these together, once we find an underlying parameter that is sufficient and minimal, we can "forget" about the weakly regular parameter and concentrate on the model that consists only of regular parameters.

4.1 Nuisance Tangent Spaces

By definition, a weakly regular parameter satisfies $\beta(Q_t) \to \beta_P(g)$ for every path Q_t. To illuminate the idea in the coming definition, let us assume for the sake of argument that $\beta_P : \dot{\mathcal{P}}_{P,\beta} \to \mathbb{B}$ is "invertible" in the sense that the equation $\beta_P(g) = b$ can be written as $g = g_\eta + g_\beta(b)$. The first term g_η does not affect the value of β_P, so it is nuisance; the second term is the important component of the score that contains information about β_P. Thus, the tangent space consists of two subspaces: one spanned by $\{g_\eta\}$

and the other by $\{g_\beta(b) : b \in \mathbb{B}\}$. In the nuisance parameter literature, the "efficient tangent space" that represents the maximal amount of relevant information contained in the model is derived through the projection of $\{g_\beta(b)\}$ onto the orthocomplement of $\{g_\eta\}$.[20] Then, we naturally expect that the efficient tangent space for a good underlying regular parameter should contain and only contain the efficient tangent space for β, representing the right amount of information in the local expansion.

What makes the analysis nonstandard is the involvement of the nonlinear map β_P in the local expansion. In the classical semiparametric theory, the score and the local parameters are related linearly to each other, thereby leading to a very nice use of the theory of linear operators (Bickel et al., 1993). The following definition extends the key notions from this literature to a nonlinear map defined on a cone of a linear space.

Definition. Let $\mathscr{X}$ be a linear space and $\mathscr{Y}$ a set. For a map $f : A \to \mathscr{Y}$ defined on a cone A in $\mathscr{X}$, define the *range* R and *kernel* N by $R(f) := \{y \in \mathscr{Y} : y = f(x)$ for some $x \in A\}$ and $N(f) := \{\tilde{x} \in \mathscr{X} : x + \tilde{x} \in A$ and $f(x + \tilde{x}) = f(x)$ for every $x \in A\}$.

Remark. If f is linear and $A = \mathscr{X}$, $N(f)$ reduces to the standard definition of a kernel for linear maps.

Now we define the smallest tangent set for β, using the notion of the kernel instead of "invertibility." The key is that any score that does not affect identification or the value of β only contains information about the path that is irrelevant to β; such a score can be deemed nuisance. Since the tangent space $\dot{\mathcal{P}}_P$ is linear, we separate the space into the space spanned by nuisance scores and the residual space. That residual space will, by construction, only contain scores that are relevant to either identification or value of β. This is the minimal tangent set, which we will show shortly is a cone.

Definition (Nuisance tangent space)**.** Let $\beta : \mathcal{P}_\beta \to \mathbb{B}$ be weakly regular. Call the kernel $N(\beta_P) \subset \dot{\mathcal{P}}_P$ of β_P the *nuisance tangent space*. Denote by $\Pi_{-\beta}$ and Π_β the projection operators onto $N(\beta_P)$ and $N(\beta_P)^\perp$ in $L_2(P)$.

The definition of $N(\beta_P)$ tells that $\tilde{g} \in N(\beta_P)$ means $g + \tilde{g} \in \dot{\mathcal{P}}_{P,\beta}$ and $\beta_P(g + \tilde{g}) = \beta_P(g)$ for every $g \in \dot{\mathcal{P}}_{P,\beta}$; the first condition is the preservation of identification and the second the preservation of the value of β. The flip side of this is that if $\tilde{g} \notin \dot{\mathcal{P}}_{P,\beta}$, then there exists $g \in \dot{\mathcal{P}}_{P,\beta}$ such that either $g + \tilde{g} \notin \dot{\mathcal{P}}_{P,\beta}$ or $\beta_P(g + \tilde{g}) \neq \beta_P(g)$ is true. Therefore, such $\tilde{g}$ can be considered to hold information of either identification or distinction of β.

[20] For more intuition on this point, see Van der Vaart (1998, Section 25.4).

Lemma S.1 verifies that the minimal tangent cone is indeed a cone and the nuisance tangent space is a linear space. Now we clarify these concepts in linear IV models.

Example 1 (Linear IV, continued)**.** The score formula (1) implies that the score space spanned by $g \in \dot{\mathcal{P}}_P$ such that $\mathbb{E}_P[zug] = 0$ and $\mathbb{E}_P[zv'g] = 0$ must be contained in $N(\beta_P)$. On the other hand, any other score will change the value of either $\dot\pi$ or β. Suppose that for $g_1 \in \dot{\mathcal{P}}_{P,\beta}$, adding the score $\tilde g \in \dot{\mathcal{P}}_P$ would change the value of $\dot\pi$ but not β. So one may write $g_1 = g_{uvz} - z'\dot\pi\beta\frac{dP_{uvz,u}}{dP} - z'\dot\pi\frac{dP_{uvz,v}}{dP}$ and $\tilde g = \tilde g_{uvz} - z'\tilde\pi\beta\frac{dP_{uvz,u}}{dP} - z'\tilde\pi\frac{dP_{uvz,v}}{dP}$. Then take g_2 to be such that $g_2 = -z'\dot\pi(2\beta)\frac{dP_{uvz,u}}{dP} - z'\dot\pi\frac{dP_{uvz,v}}{dP}$ so that $\beta_P(g_2) = 2\beta$. Then adding $\tilde g$ to g_2 will change the value of β since $(\dot\pi+\tilde\pi)^{\rightarrow}(2\dot\pi+\tilde\pi)\beta \neq \beta$; it can even be that $(2\dot\pi + \tilde\pi)\beta$ falls outside of the column space of $\dot\pi + \tilde\pi$. Therefore, such $\tilde g$ cannot be in $N(\beta_P)$. Thus we see that $N(\beta_P)$ equals the set of scores g such that $\mathbb{E}_P[zug] = 0$ and $\mathbb{E}_P[zv'g] = 0$, or, the set of scores induced by $\mathcal{P}_{uvz}$.

Example 7 (Regular parameter)**.** A regular parameter $\psi : \mathcal{P} \to \mathbb{D}$ can be considered trivially weakly regular. Since ψ is defined on the whole of $\mathcal{P}$, $\mathcal{P}_\beta = \mathcal{P}$ and $\dot{\mathcal{P}}_{P,\beta} = \dot{\mathcal{P}}_P$. Since ψ is differentiable at P, it is continuous at P, that is, $\psi(Q_t) \to \psi(P)$ for every $Q_t \in \mathscr{P}_{P,\beta}$. Therefore, the limit of $\psi(Q_t)$ can be trivially written as a constant function $\psi_P : \dot{\mathcal{P}}_P \to \mathbb{D}$ such that $\psi_P(g) \equiv \psi(P)$ for every $g \in \dot{\mathcal{P}}_P$. Any change in the score cannot affect the value of ψ_P, so $N(\psi_P) = \dot{\mathcal{P}}_P$ and $\dot{\mathcal{P}}_{P,\psi} = \{0\}$.

4.2 Sufficiency and Minimality of Underlying Regular Parameters

The underlying regular parameters are characterized by the span of their "scores," or equivalently, of their efficient influence maps.[21] The first property we want in the underlying regular parameter is that it contain all relevant information about β. Here, "information" is captured by the ability to discern distinct scores in the limit.

Definition (Sufficiency of underlying regular parameter)**.** Let $\beta : \mathcal{P}_\beta \to \mathbb{B}$ be weakly regular. An underlying regular parameter $\psi : \mathcal{P} \to \mathbb{D}$ for β is *sufficient* if $N(\dot\psi_P) \subset N(\beta_P)$, or equivalently, $N(\beta_P)^\perp \subset R(\dot\psi_P^*)$.[22]

The efficient influence map $\dot\psi_P^*$ of an underlying parameter ψ summarizes the set of scores that the local parameter of ψ can distinguish. If ψ is sufficient, then knowing the local parameter of ψ gives a sufficient amount of information that a score contains

[21] See Van der Vaart (1991b) and Bickel et al. (1993, Section 5.4) for equivalence.

[22] Use the property of an adjoint operator: $N(\dot\psi_P)^\perp = R(\dot\psi_P^*)$ (Kosorok, 2008, Equation 17.3).

about the identification or distinction of β. The equivalent formulation says that the score that ψ cannot distinguish is never used in identification or distinction of β. The following example shows that an underlying regular parameter need not be sufficient.

Example 1 (Insufficiency in linear IV, continued)**.** Let $d = 1$ and $k > 1$. Consider the underlying regular parameter $\psi(Q) = (\pi_1\beta, \pi_1)$ that induces $\dot{\psi}_P g = (\dot{\pi}_1\beta, \dot{\pi}_1)$. This parameter only uses the first instrument and abandons information from all other instruments available in the model. Therefore, $N(\dot{\psi}_P)$ contains elements g that change the value of $\dot{\pi}_2$. However, changing the value of $\dot{\pi}_2$ without adjusting for the values of $\dot{\pi}_1\beta$ and $\dot{\pi}_1$ will make β undefined and push the score outside of $\dot{\mathcal{P}}_{P,\beta}$, so we have $g \notin N(\beta_P)$. Hence, ψ is not sufficient.

Not surprisingly, a sufficient underlying regular parameter contains information of all instruments.

Example 1 (Sufficiency in linear IV, continued)**.** The underlying regular parameter $\psi(Q) = (\pi\beta, \operatorname{vec}(\pi))$ is sufficient. To see this, note that $\dot{\psi}_P g = (\dot{\pi}\beta, \operatorname{vec}(\dot{\pi}))$ and take $g_\eta \in N(\dot{\psi}_P)$. The values of $\dot{\pi}\beta$ and $\dot{\pi}$ do not change by adding g_η to the score. Recalling the score formula (1), one sees that $g + g_\eta \in \dot{\mathcal{P}}_{P,\beta}$ whenever $g \in \dot{\mathcal{P}}_{P,\beta}$ and $\beta_P(g + g_\eta) = \beta_P(g)$, that is, $g_\eta \in N(\beta_P)$. Therefore, ψ is sufficient.

The next property we want in an underlying regular parameter is that it has only relevant information for the weakly regular parameter. Otherwise, the underlying parameter contains some information of a "nuisance parameter" and estimating it may capture unwanted noise that is irrelevant to estimation of the weakly regular parameter.

Definition (Minimality of underlying regular parameter)**.** Let $\beta : \mathcal{P}_\beta \to \mathbb{B}$ be weakly regular. An underlying regular parameter $\psi : \mathcal{P} \to \mathbb{D}$ for β is *minimal* if $N(\beta_P) \subset N(\dot{\psi}_P)$, or equivalently, $R(\dot{\psi}_P^*) \subset N(\beta_P)^\perp$.

Minimality of ψ requires the opposite inclusion between $N(\beta_P)$ and $N(\dot{\psi}_P)$. This is to say that the score irrelevant to identification or distinction of β is also irrelevant to distinction of the local parameter of ψ. Equivalently, the range of the efficient influence map of ψ does not contain a score that is unrelated to β. In this sense, a minimal underlying parameter is "free" of potential nuisance parameters.

Example 1 (Minimality in linear IV, continued)**.** As seen in Example 1 in the previous subsection, $N(\beta_P)$ is the set of scores g_{uvz} induced by $\mathcal{P}_{uvz}$. Again, recalling the score formula, one sees that adding such scores does not change the values of $\dot{\pi}\beta$ and $\dot{\pi}$, which implies $N(\beta_P) \subset N(\dot{\psi}_P)$ for both choices of the underlying regular parameter we discussed: $(\pi\beta, \mathrm{vec}(\pi))$ and $(\pi_1\beta, \pi_1)$.[23]

Remark. Minimal sufficiency in our definition is of a parameter, while minimal sufficiency in the context of sufficient statistics is of a statistic.

This theorem ensures that a minimal sufficient underlying parameter exists.

Theorem 4 (Existence of minimal sufficient underlying regular parameter)**.** *For a weakly regular parameter $\beta : \mathcal{P}_\beta \to \mathbb{B}$, there exists a minimal sufficient underlying regular parameter.*

Minimal sufficiency *per se* is not strong enough to pin down the underlying parameter uniquely. However, underlying parameters that are both minimal and sufficient are almost equivalent in terms of the information they contain. Theorem S.2 characterizes minimal sufficient underlying parameters and establishes this "equivalence."

Let us look at examples of minimal sufficient underlying parameters.

Example 1 (Linear IV, continued)**.** As seen earlier, $\psi = (\pi\beta, \mathrm{vec}(\pi))$ is a natural choice of an underlying regular parameter that is minimal and sufficient.

Example 2 (Nonlinear GMM, continued)**.** The moment function m is sufficient and, in many cases, minimal. In particular, if there exists $g \in \dot{\mathcal{P}}_{P,\beta}$ such that $\dot{m}_P g = \mathbb{E}_P[M_i g]$ has a zero at θ, then any $\tilde{g} \in \dot{\mathcal{P}}_P$ for which $\dot{m}_P \tilde{g}$ is nonzero at θ is not in $N(\theta_P)$. This is obvious since $\dot{m}_P(g + \tilde{g})$ does not have a zero at θ by the linearity of expectations. In short, if for any value θ of $\mathbb{R}^d$ there exists $g \in \dot{\mathcal{P}}_{P,\beta}$ such that θ is the (unique) zero of $\dot{m}_P g$, then the entire moment function m is minimal.

Given the minimal sufficient underlying regular parameter, the problem of estimation or inference of a weakly regular parameter can be translated into a problem of estimation or inference of the local parameter of the minimal sufficient underlying parameter. Being a local parameter of a regular parameter, it provides workable grounds for many statistical analyses.

One caveat: Unlike classical theory for regular parameters, we "know" that the local parameter $\dot{\psi}_P g$ lies in the strict subset $\mathbb{D}_\beta$ of $\mathbb{D}$. This constraint, next to nonlinearity of $\beta_{P,\psi}$, stands as a major source of complication in estimation and inference.

[23] Note that $\psi = (\pi\beta, \mathrm{vec}(\pi))$ is still minimal even in the homoskedastic model. Homoskedasticity helps simplify efficient estimation, but does not help simplify the semiparametric structure itself.

5 WEAK EFFICIENCY FOR WEAKLY REGULAR PARAMETERS

This section defines a novel notion of efficiency of estimators of a weakly regular parameter. The difficulty in formulating a reasonable goodness criterion for estimators of weakly regular parameters lies in the fact that their asymptotic distribution is nonstandard. The classical convolution theorem requires symmetry of the distribution while many natural estimators of a weakly regular parameter do not lead to a symmetric distribution. In our context, the limit of a weakly regular parameter is a nonlinear transformation of the local parameter of an underlying regular parameter. As the estimator of a local parameter often leads to Gaussian distributions, one can anticipate that the asymptotic distribution of an estimator of a weakly regular parameter is some nonlinear transformation of a Gaussian distribution. Moreover, a nonlinear transformation of a Gaussian distribution can, in general, be *anything*. Our idea of defining efficiency lies in that the consequence of the convolution theorem—inefficient limit distribution involves an irrelevant noise—carries over after a nonlinear transformation. In light of Jensen's inequality, the involvement of noise must increase convex loss.

Assuming that $\beta_{P,\psi}$ is smooth enough, the problem of accurately estimating β translates into a problem of accurately estimating ψ. If the natural estimator of ψ takes values so that the estimated local parameter falls into $\mathbb{D}_\beta$ (the range of $\dot{\psi}_P$ on the pertinent tangent cone $\dot{\mathcal{P}}_{P,\beta}$) with probability one, then there is nothing that needs to be done (and indeed this is the case in some applications). If not, we need to accommodate the constraint $\dot{\psi}_P g \in \mathbb{D}_\beta$ in either of the following ways.

1. Estimate ψ with the constraint, so we have $\widehat{\dot{\psi}_P g} \in \mathbb{D}_\beta$ almost surely.

2. Estimate ψ without the constraint; then deal with values outside of $\mathbb{D}_\beta$.

Noting that the first approach can be viewed as estimating ψ without the constraint and then reconstructing another estimator that satisfies the constraint, the two approaches are essentially equivalent. In this section, therefore, we employ the latter interpretation.

In order to make use of the convolution theorem, we restrict attention to the estimators of β that are transformations of regular estimators of ψ. Throughout this section, we assume that ψ can be estimated regularly and efficiently at root-n.[24] First, recall the definition of a regular estimator for a regular parameter.

[24] Not all regular parameters admit root-n consistent estimation, especially infinite-dimensional ones (Wellner et al., 2006; Giné and Nickl, 2015). We make this assumption here since the focus of this paper is to demonstrate the power of reduction to regular parameters in the context of weak identification and many interesting examples in economics indeed admit root-n consistent estimation.

Definition (Regular estimator for regular parameter)**.** A sequence of estimators $\hat{\psi}_n$ for a regular parameter $\psi : \mathcal{P} \to \mathbb{D}$ is called *regular* at $P \in \mathcal{P}$ relative to $\mathscr{P}_P$ if there exists a tight Borel random element L in $\mathbb{D}$ such that

$$\sqrt{n}(\hat{\psi}_n - \psi(Q_n)) \overset{Q_n}{\rightsquigarrow} L \qquad \text{for every} \qquad Q_n \in \mathscr{P}_P.$$

This sequence is called *(semiparametric) efficient* at P relative to $\mathscr{P}_P$ if it attains the distributional lower bound (denote it by L_ψ) of the convolution theorem (Van der Vaart, 1991a, Theorem 2.1; Kosorok, 2008, Theorem 18.3).

Remark. The convolution theorem states that $L = L_\psi + L_\eta$ where L_ψ and L_η are independent tight Borel random elements in $\mathbb{D}$ such that $\Pr(L_\psi \in \overline{R(\dot{\psi}_P)}) = 1$ and $\delta^* L_\psi \sim N\big(0, \|\dot{\psi}_P^* \delta^*\|_{L_2(P)}^2\big)$ for every $\delta^* \in \mathbb{D}^*$. This is to say, the asymptotic distribution of any regular estimator of a regular parameter is the sum of a Gaussian variable with covariance being the "L_2 norm" of the efficient influence map and an independent noise. It is efficient when no noise is involved, that is, when $L_\eta \equiv 0$.

Remark. If we center $\hat{\psi}_n$ at $\psi(P)$, convergence becomes $\sqrt{n}(\hat{\psi}_n - \psi(P)) \rightsquigarrow^{Q_n} \dot{\psi}_P g + L$.

Now we define the class of estimators we consider. We focus on estimators of a weakly regular parameter that can be represented as transformations of estimators of a minimal underlying regular parameter. Many estimators in the literature fall within this class. Asymptotic randomness of such estimators must come from estimators of the local parameter of the underlying parameter $\sqrt{n}(\hat{\psi}_n - \psi(P))$ and possibly some irrelevant noise. Toward this end, we define functions that admit approximation by a root-n normalization around the point of identification failure.

Definition (Local continuous approximability)**.** Let $\mathbb{D}_n$ be indexed subsets of $\mathbb{D}$. A sequence of maps $T_n : \mathbb{D}_n \times [0,1] \to \mathbb{B}$ is *locally continuously root-n approximable at $\delta \in \mathbb{D}$ tangentially to $\mathbb{D}_\infty \subset \mathbb{D}$* if there exists a measurable map $T_\delta : \mathbb{D}_\infty \times [0,1] \to \mathbb{B}$ such that for every $u \in [0,1]$, $\delta + \delta_n/\sqrt{n} \in \mathbb{D}_n$, $\delta_\infty \in \mathbb{D}_\infty$, if a subsequence $\delta_{n'}$ satisfies $\delta_{n'} \to \delta_\infty$, then $T_{n'}\big(\delta + \frac{\delta_{n'}}{\sqrt{n'}}, u\big) \to T_\delta(\delta_\infty, u)$ as $n \to \infty$. Denote this by $T_n \to^\delta T_\delta$ and call T_δ the *approximating function* of T_n at δ. We use these definitions even when T_n and T_δ do not depend on the second argument since they can be considered trivially dependent on (i.e., constant with respect to) $u \in [0,1]$.

Remark. By construction T_δ is continuous on $\mathbb{D}_\infty$.

Definition (Regular estimator for weakly regular parameter)**.** A sequence of estimators $\hat{\beta}_n$ for a weakly regular parameter $\beta : \mathcal{P}_\beta \to \mathbb{B}$ is called *regular* at $P \in \mathcal{P}$ relative to $\mathscr{P}_{P,\beta}$ if there exist a minimal underlying regular parameter $\psi : \mathcal{P} \to \mathbb{D}$ for β, a regular sequence of estimators $\hat{\psi}_n$ of ψ with $\sqrt{n}(\hat{\psi}_n - \psi(Q_n)) \rightsquigarrow^{Q_n} L$, a sequence of nonrandom maps $T_n : \mathbb{D} \times [0,1] \to \mathbb{B}$ that is locally continuously root-n approximable at $\psi(P)$ tangentially to the range of $\dot{\psi}_P g + L$ for $g \in \dot{\mathcal{P}}_{P,\beta}$, and an independent noise $U \sim U[0,1]$ such that $T_{\psi(P)}(L, U)$ is Borel measurable and

$$\hat{\beta}_n = T_n(\hat{\psi}_n, U) + o_P(1) \qquad \text{under every} \qquad Q_n \in \mathscr{P}_{P,\beta}.$$

Remark. We may without loss of generality take ψ as sufficient; for otherwise one can augment ψ and have T_n ignore the augmented part.

The asymptotic distribution of a regular estimator is a transformation of the asymptotic distribution of the underlying regular parameter.

Proposition 5. *Let $\hat{\beta}_n = T_n(\hat{\psi}_n, U) + o_P(1)$ be a regular sequence of estimators for a weakly regular parameter $\beta : \mathcal{P}_\beta \to \mathbb{B}$ and $\sqrt{n}(\hat{\psi}_n - \psi(Q_n)) \rightsquigarrow^{Q_n} L$. Then,*

$$\hat{\beta}_n \overset{Q_n}{\rightsquigarrow} T_{\psi(P)}(\dot{\psi}_P g + L, U).$$

Remark. More generally, a regular estimator $\hat{\beta}_n$ can be viewed as a possibly random transformation $\hat{T}_n(\hat{\psi}_n, U)$ of $\hat{\psi}_n$ and U, and randomness of $\hat{T}_n$ vanishes in an appropriate sense. In the linear IV example below, we show directly that there exists a nonrandom transformation T_n such that the difference $\hat{T}_n(\hat{\psi}_n, U) - T_n(\hat{\psi}_n, U)$ is $o_P(1)$.

Example 1 (Linear IV, continued)**.** Popular estimators of the linear IV model are regular. Consider the 2SLS. Observe that the reduced-form coefficients $(\pi\beta, \pi)$ are regular and the 2SLS can be written as a function of their estimators $\tilde{\pi}_n = (Z'Z)^{-1}Z'X$ and $\widetilde{\pi\beta}_n = (Z'Z)^{-1}Z'Y$:

$$\tilde{\beta}_{\text{2SLS}} = (\tilde{\pi}_n'(Z'Z)\tilde{\pi}_n)^{-1}\tilde{\pi}_n'(Z'Z)\widetilde{\pi\beta}_n = (\tilde{\pi}_n'\mathbb{E}[zz']\tilde{\pi}_n)^{-1}\tilde{\pi}_n'\mathbb{E}[zz']\widetilde{\pi\beta}_n + o_P(1).$$

The residual is $o_P(1)$ since $(Z'Z)/n$ converges to $\mathbb{E}[zz']$ in probability under every path. This shows regularity of the 2SLS with the homogeneous approximating function $T : \mathbb{R}^k \times \mathbb{R}^{k\times d} \to \mathbb{R}^d$, $T(\pi\beta, \pi) = (\pi'\mathbb{E}[zz']\pi)^{-1}\pi'\mathbb{E}[zz']\pi\beta$. Recall that a subset of the instruments is not sufficient (Example 1). However, the 2SLS estimator that uses only a part of the instruments is also regular. To see this, let the subscript (d) denote the selection of d instruments. The partial 2SLS estimator can be seen as a function of an estimator for the entire reduced-form coefficients as

$$\tilde{\beta}_{\text{2SLS},(d)} = T_n(\tilde{\pi}_n, \widetilde{\pi\beta}_n, U) = (\tilde{\pi}'_{(d)}\mathbb{E}[z_{(d)}z'_{(d)}]\tilde{\pi}_{(d)})^{-1}\tilde{\pi}'_{(d)}\mathbb{E}[z_{(d)}z'_{(d)}]\widetilde{\pi\beta}_{(d)} + o_P(1),$$

where $\tilde{\pi}_{(d)}$ and $\widetilde{\pi\beta}_{(d)}$ are an estimator of the entire reduced-form coefficients using only the (d) subset of instruments, and the remaining parts of $\tilde{\pi}$ and $\widetilde{\pi\beta}$ can be anything (as long as $(\widetilde{\pi\beta}, \tilde{\pi})$ is regular).

Similarly, GMM can be shown to be regular. Denote by W the weighting matrix. The GMM estimator $\tilde{\beta}_{\text{GMM}}$ with weighting W solves $\min_b \big[\frac{Z'(Y-Xb)}{n}\big]' W \big[\frac{Z'(Y-Xb)}{n}\big]$. Write the objective function as

$$\frac{1}{n}\sqrt{n}(\widetilde{\pi\beta}_n - \tilde{\pi}_n b)'\frac{Z'Z}{n}W\frac{Z'Z}{n}\sqrt{n}(\widetilde{\pi\beta}_n - \tilde{\pi}_n b).$$

The oracle weighting matrix for the efficient GMM is $W = \mathbb{E}[(y - x'\beta)^2 zz']^{-1}$, while it is estimated in case of the feasible GMM. In particular, the two-step GMM estimates W by plugging in the 2SLS estimator for β and taking its sample counterpart, i.e., $\hat{W}_{\text{2SGMM}} = \mathbb{E}_n[(y - x'\tilde{\beta}_{\text{2SLS}})^2 zz']^{-1}$. The expectation involved in $\hat{W}$ (other than the 2SLS estimator) can be consistently estimated. Moreover, minimization is invariant to scaling, so if one sees minimization as a function of the 2SLS estimators, it is homogeneous of degree zero (and continuous). Thus, the two-step GMM is regular.

The Fuller estimator proposed by Fuller (1977) is regular while we suspect that the heteroskedasticity-robust Fuller (HFUL) estimator proposed by Hausman et al. (2012) is not. Let $P := Z(Z'Z)^{-1}Z'$. For a constant C, let $\tilde{P}_{\text{Fuller}} := P + (C/n)(I - P)$. The Fuller estimator is given by

$$\begin{aligned}\hat{\beta}_{\text{Fuller}} &= (X'\tilde{P}_{\text{Fuller}}X)^{-1}(X'\tilde{P}_{\text{Fuller}}Y) \\ &= \Big(C\mathbb{E}[xx'] + \sqrt{n}\tilde{\pi}'_n\mathbb{E}[zz']\sqrt{n}\tilde{\pi}_n\Big)^{-1}\Big(C\mathbb{E}[xy] + \sqrt{n}\tilde{\pi}'_n\mathbb{E}[zz']\sqrt{n}\widetilde{\pi\beta}_n\Big) + o_P(1).\end{aligned}$$

Thus, under weak identification, the Fuller estimator can be thought of as a "weighted combination" of OLS ($C = \infty$) and 2SLS ($C = 0$). On the other hand, due to its jackknife form, the HFUL estimator requires calculation of the off-diagonal matrix of P (Hausman et al., 2012). While this is their source of robustness to heteroskedasticity (under different asymptotics), this makes it challenging, possibly infeasible, to represent HFUL only as a function of the OLS estimator.

The unbiased estimator by Andrews and Armstrong (2017) is also regular. For simplicity, let $d = 1$ and $k = 1$ and assume that $\pi > 0$ and $\tilde{\pi}_n$ and $\widetilde{\pi\beta}_n$ are asymptotically uncorrelated. Then the unbiased estimator of β is

$$\hat{\beta}_{\text{unbiased}} = \frac{\sqrt{n}\widehat{\pi\beta}_n}{\hat{\sigma}_{\pi\beta,n}} \frac{1 - \Phi(\sqrt{n}\hat{\pi}_n/\hat{\sigma}_{\pi,n})}{\hat{\sigma}_{\pi,n}\phi(\sqrt{n}\hat{\pi}_n/\hat{\sigma}_{\pi,n})} = \frac{\sqrt{n}\widehat{\pi\beta}_n}{\sigma_{\pi\beta}} \frac{1 - \Phi(\sqrt{n}\hat{\pi}_n/\sigma_\pi)}{\sigma_\pi\phi(\sqrt{n}\hat{\pi}_n/\sigma_\pi)} + o_P(1),$$

which is regular with

$$T_n(\hat{\pi}_n, \widehat{\pi\beta}_n, U) = \frac{\sqrt{n}\widehat{\pi\beta}_n}{\sigma_{\pi\beta}} \frac{1 - \Phi(\sqrt{n}\hat{\pi}_n/\sigma_\pi)}{\sigma_\pi\phi(\sqrt{n}\hat{\pi}_n/\sigma_\pi)}, \quad T_{(0,0)}(\hat{\pi}, \hat{\pi}\beta, U) = \frac{\hat{\pi}\beta}{\sigma_{\pi\beta}} \frac{1 - \Phi(\hat{\pi}/\sigma_\pi)}{\sigma_\pi\phi(\hat{\pi}/\sigma_\pi)}.$$

5.1 Local Asymptotic Rao-Blackwellization

We show that for any regular estimator of a weakly regular parameter, there exists another regular estimator that is weakly better in terms of convex loss. A strict improvement is always possible unless our estimator is already a nonrandom transformation of an efficient estimator of the underlying parameter. In other words, whenever an estimator contains "noise" irrelevant to the efficient estimation of ψ, one can always construct another estimator that shares the same expectation and is more concentrated around it. We demonstrate the power of this improvement in Section 6.

Theorem 6 (Local asymptotic Rao-Blackwellization)**.** *Let $\beta : \mathcal{P}_\beta \to \mathbb{B}$ be weakly regular and $\psi : \mathcal{P} \to \mathbb{D}$ a minimal underlying regular parameter for β. Let $\tilde{\psi}_n$ be a regular sequence of estimators of ψ and $\tilde{\beta}_n = T_n(\tilde{\psi}_n, U) + o_P(1)$ be a regular sequence of estimators of β with noise $U \sim U[0, 1]$. Suppose that an efficient regular sequence of estimators $\hat{\psi}_n$ of ψ exists and $\bar{T}_n(\delta) := \mathbb{E}[T_n(\delta + L_\eta/\sqrt{n}, U)]$ exists as a Bochner integral.*[25] *Then $\bar{T}_n(\hat{\psi}_n)$ is a better regular estimator than $\tilde{\beta}_n$ in the sense that for every convex continuous loss function $\ell : \mathbb{B} \to \mathbb{R}$ such that $\ell(\tilde{\beta}_n - \beta(Q_n))$ and $\ell(\bar{T}_n(\hat{\psi}_n) - \beta(Q_n))$ are asymptotically equiintegrable under $Q_n \in \mathscr{P}_{P,\beta}$,*[26]

$$\liminf_{n\to\infty} \mathbb{E}_{Q_n,*}[\ell(\tilde{\beta}_n - \beta(Q_n))] - \mathbb{E}^*_{Q_n}[\ell(\bar{T}_n(\hat{\psi}_n) - \beta(Q_n))] \geq 0.$$

Remark. Theorem 6 is a kind of admissibility requirement for a convex loss. Unlike popular discussion of inadmissibility, however, it confines attention to the class of regular estimators while providing an improvement method (Rao-Blackwellization) to achieve "admissibility." If $\mathbb{B} = \mathbb{R}$, $\bar{T}_n$ first-order stochastically dominates T_n.

[25] See Bharucha-Reid (1972) for a discussion of Bochner integrals.

[26] X_n is *asymptotically equiintegrable* if $\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{E}^*[|X_n|\mathbb{1}\{|X_n| > M\}] = 0$ (Van der Vaart and Wellner, 1996, p. 421).

Remark. Efficiency is usually justified for *subconvex* loss functions (Kosorok, 2008, Theorem 18.4; Van der Vaart and Wellner, 1996, Theorem 3.11.5). Theorem 6 is in the same spirit but restricts us to *convex* functions.[27] This difference comes from the fact that our best asymptotic distribution is a nonlinear transformation of Gaussian; there is no symmetry of the distribution we can exploit to accommodate subconvexity.

Example 1 (Linear IV, continued)**.** Suppose that the reduced-form errors are heteroskedastic, and the feasible GLS estimator $(\widehat{\pi\beta}_n, \hat{\pi}_n)$ is available. As seen in Example 1 in the previous section, all of the 2SLS, GMM, Fuller, and unbiased estimators are functions of the OLS estimator of the reduced-form coefficients. Then, Theorem 6 suggests in such cases that the use of the 2SLS estimator is asymptotically suboptimal in terms of the concentration of asymptotic distributions measured by convex loss functions. In the case of 2SLS, one can construct a better estimator $\bar{T}_n(\widehat{\pi\beta}_n, \hat{\pi}_n)$ by

$$\bar{T}_n(\pi\beta, \pi) := \mathbb{E}\left[\left(\left[\pi + \frac{U_\pi}{\sqrt{n}}\right]' \mathbb{E}[zz'] \left[\pi + \frac{U_\pi}{\sqrt{n}}\right]\right)^{-1} \left(\left[\pi + \frac{U_\pi}{\sqrt{n}}\right]' \mathbb{E}[zz'] \left[\pi\beta + \frac{U_{\pi\beta}}{\sqrt{n}}\right]\right)\right],$$

where

$$\begin{pmatrix} \sqrt{n}(\widetilde{\pi\beta}_n - \widehat{\pi\beta}_n) \\ \sqrt{n}(\tilde{\pi}_n - \hat{\pi}_n) \end{pmatrix} \rightsquigarrow \begin{pmatrix} U_{\pi\beta} \\ U_\pi \end{pmatrix}.$$

In Section 6, we compare the performance of 2SLS and its improvement. Interestingly, even with the oracle weighting matrix, GMM contains noise that can be removed if an efficient estimator of the reduced-form coefficients is available.

Note that the limited information maximum likelihood (LIML) estimator is known to have no moment (Chao et al., 2012), being outside the direct scope of Theorem 6, but it can be said to be regular by definition. LIML estimates $\hat{W}(b)$ assuming homoskedasticity, that is, $\hat{W}_{\text{LIML}}(b) = n(Z'Z)^{-1}/\hat{\sigma}^2(b)$ where $\hat{\sigma}^2(b) = \mathbb{E}_n[(y - x'b)^2]$ (Andrews, 2017). Since the second and cross moments of y and x can be consistently estimated, LIML is asymptotically only a function of the OLS estimators of the reduced-form coefficients. Similarly, although the continuously updating GMM is suspected to have no moment (Guggenberger, 2005), it is regular as it uses $\hat{W}_{\text{CUGMM}}(b) = \mathbb{E}_n[(y - x'b)^2 zz']^{-1}$, which, again, admits consistent estimation.

[27] Technically, there is no implication between convexity and subconvexity of a function. In this context, subconvexity can be thought of as roughly weaker.

5.2 Weakly Efficient Estimators of Weakly Regular Parameters

Backed up by this result, we define an efficiency concept for estimating a weakly regular parameter. The idea is that when an estimator of a weakly regular parameter does not admit an improvement by Theorem 6, we want to call such an estimator "efficient." The condition under which an estimator does not allow improvement is that it is already a nonrandom transformation of an efficient estimator of the minimal sufficient underlying regular parameter.

Definition (Weak efficiency for weakly regular parameter)**.** A regular sequence of estimators $\hat{\beta}_n$ for a weakly regular parameter β is *weakly (semiparametric) efficient* at $P \in \mathcal{P}$ relative to $\mathscr{P}_{P,\beta}$ if there exist a minimal sufficient underlying regular parameter $\psi : \mathcal{P} \to \mathbb{D}$, its efficient sequence of estimators $\hat{\psi}_n$, and a sequence of nonrandom measurable maps $T_n : \mathbb{D} \to \mathbb{B}$ that is locally continuously root-n approximable at $\psi(P)$ tangentially to the range of $\dot{\psi}_P g + L_\psi$ such that $\hat{\beta}_n = T_n(\hat{\psi}_n) + o_P(1)$ under every $Q_n \in \mathscr{P}_{P,\beta}$.

Weak efficiency is related to classical efficiency. Consider two *regular* parameters, ψ_1 and ψ_2, related to each other through a Hadamard differentiable function f, $\psi_2 = f(\psi_1)$. Van der Vaart (1998, Theorem 25.47) shows that efficiency of $\hat{\psi}_1$ for ψ_1 implies efficiency of $f(\hat{\psi}_1)$ for ψ_2. A weakly regular parameter β is often directly related to an underlying regular parameter ψ by $\beta = \beta_\psi(\psi)$. We define $\hat{\beta}$ to be *weakly efficient* if it is a nonrandom transformation of efficient $\hat{\psi}$. The reason we do not require $\hat{\beta}$ to be specifically $\beta_\psi(\hat{\psi})$ is for the impossibility of consistent and nonpivotal estimation (Theorem 2). In principle, there is no necessity to treat uncertainty of $\hat{\psi}$ the same way we treat unknownness of ψ for the sake of a "good" estimator, if the relationship is nonlinear. Thus, weakly efficient estimators are not unique, for which we call it *weak*.

A simple but useful byproduct of our definition is that, if at some $Q \in \mathcal{P}_\beta$ the transformation locally reduces to β_ψ and β_ψ is Hadamard differentiable, then a weakly efficient estimator constructed for $P \in \mathcal{P} \setminus \mathcal{P}_\beta$ becomes *efficient* in the classical sense under the "strong identification asymptotics" at $Q \in \mathcal{P}_\beta$, provided that $\hat{\psi}_n$ remains efficient in both asymptotics. This is a direct consequence of the delta method (Van der Vaart, 1991a, 1998, Section 25.7). Therefore, weak efficiency can also be regarded as a generalization of efficiency to discontinuous but locally continuously approximable transformations of efficient estimators.

Remark. Our definition of weak efficiency does not assume existence of moments. We should note, however, that Theorem 6 does not apply to estimators with no moments.

Example 1 (Linear IV, continued)**.** As explained earlier, the 2SLS, GMM, Fuller, and unbiased estimators are inefficient in the presence of heteroskedasticity. If an efficient estimator of the reduced-form coefficients is available, then the Rao-Blackwellized 2SLS, GMM, Fuller, and unbiased estimators conditional on this efficient estimator are weakly efficient. See Section 6 for the comparison of the these estimators in simulation.

Example 2 (Nonlinear GMM, continued)**.** If the entire moment function is the minimal underlying parameter, the efficiency of an estimator in a nonlinear GMM model boils down to the efficiency of the estimator of the moment function.

Remark. A shortcoming of our efficiency concept is that the constraint $\dot{\psi}_P g \in \mathbb{D}_\beta$ is not explicitly accounted for. The convolution theorem requires the tangent set be a convex cone. If it is a mere cone, the convolution theorem can only provide a lower bound on variance (Van der Vaart, 1998, Theorem 25.20), no longer allowing Jensen's inequality. In our setup, the tangent set $\dot{\mathcal{P}}_{P,\beta}$ is a cone but not necessarily convex (Lemma S.1).

6 SIMULATION OF WEAK EFFICIENCY IN LINEAR IV MODELS

To illustrate weak efficiency, we conduct simulation studies of a linear IV model with heteroskedasticity. In Example 1, we consider discrete instruments that collectively take up only on four distinct values; since z_i has a finite support, we can estimate the heteroskedastic covariance matrix without imposing any parametric assumption. This enables us to compute the feasible GLS estimator of the reduced-form coefficients without further restricting the model and use it to construct the Rao-Blackwellized (RB) improvements suggested by Theorem 6. We carry out simulation of four estimators: the 2SLS, GMM, Fuller, and unbiased estimators.

We let $n = 10{,}000$, $d = 1$, $k = 2$, and $z_i = (z_{i1}, z_{i2})$ where z_{ij} are independent Bernoulli variables. We set $\beta = 0$ and $\pi = (1, 1)/\sqrt{n}$. For each of the 4 values of z_i, the covariance matrix of (u_i, v_i') is drawn randomly and fixed at the beginning of the simulation. We call this the *model A* and iterate it for 5,000 times. We compare three estimators with their Rao-Blackwellization: 2SLS, two-step GMM, and Fuller. We also compute HFUL for comparison, while we conjecture HFUL is non-regular.

Since unbiased estimators in overidentified models take complicated forms, we employ just-identified models for their performance evaluation. We take $n = 10{,}000$ and

$d = k = 1$ and let z_i distribute uniformly on $\{0, 1, 2\}$; then we compare unbiased and Fuller to RB unbiased and RB Fuller, assuming that the sign of the first-stage coefficient is known (Andrews and Armstrong, 2017). We also provide 2SLS for comparison, but note that 2SLS has no first moment in just-identified models. We set $\beta = 1$ and $\pi = 1/\sqrt{n}$. Similarly as before, the covariance matrix of (u_i, v_i) is randomly determined for each of the 3 values of z_i and fixed afterwards. We call this the *model B* and iterate it for 5,000 times for two specifications of heteroskedasticity.

To compute the RB versions of estimators, we derive the feasible GLS estimator of the reduced-form coefficients. As classical GLS is considered within a single-equation framework, we transform the two stages into one equation:

$$\underbrace{\begin{bmatrix} Y \\ \mathrm{vec}(X) \end{bmatrix}}_{\tilde{Y}} = \begin{bmatrix} Z \\ 0 \end{bmatrix} \pi\beta + \begin{bmatrix} 0 \\ \mathbb{1}_d \otimes Z \end{bmatrix} \pi + \begin{bmatrix} u \\ \mathrm{vec}(v) \end{bmatrix} = \underbrace{\begin{bmatrix} Z & 0 \\ 0 & \mathbb{1}_d \otimes Z \end{bmatrix}}_{\tilde{Z}} \underbrace{\begin{bmatrix} \pi\beta \\ \pi \end{bmatrix}}_{\psi} + \underbrace{\begin{bmatrix} u \\ \mathrm{vec}(v) \end{bmatrix}}_{e}.$$

Consequently, the conditional covariance matrix Ω of the error terms has nonzero off-diagonal elements. We estimate it with the OLS coefficients and compute the feasible GLS estimator for $(\pi\beta, \pi)$. Note that variances of OLS and GLS are given by $\mathrm{Var}(\hat{\psi}_{\mathrm{OLS}} \mid Z) = \tilde{Z}'\tilde{Z}\big(\sum_{i=1}^n e_i^2 \tilde{z}_i \tilde{z}_i'\big)^{-1} \tilde{Z}'\tilde{Z}$ and $\mathrm{Var}(\hat{\psi}_{\mathrm{GLS}} \mid Z) = (\tilde{Z}'\Omega^{-1}\tilde{Z})^{-1}$. Since GLS is efficient, by orthogonality $\mathrm{Var}(\hat{\psi}_{\mathrm{OLS}} - \hat{\psi}_{\mathrm{GLS}} \mid Z) = \mathrm{Var}(\hat{\psi}_{\mathrm{OLS}} \mid Z) - \mathrm{Var}(\hat{\psi}_{\mathrm{GLS}} \mid Z)$. With this, we compute the conditional expectations of 2SLS and GMM conditional on GLS using 100,000 draws from $\left(\begin{smallmatrix} U_{\pi\beta} \\ U_{\pi} \end{smallmatrix}\right) \sim N\Big(\left[\begin{smallmatrix} \widehat{\pi\beta}_{\mathrm{FGLS},n} \\ \hat{\pi}_{\mathrm{FGLS},n} \end{smallmatrix}\right], \mathrm{Var}(\hat{\psi}_{\mathrm{OLS}} - \hat{\psi}_{\mathrm{GLS}} \mid Z)\Big)$.

Given this, the RB 2SLS estimator of β is given by[28]

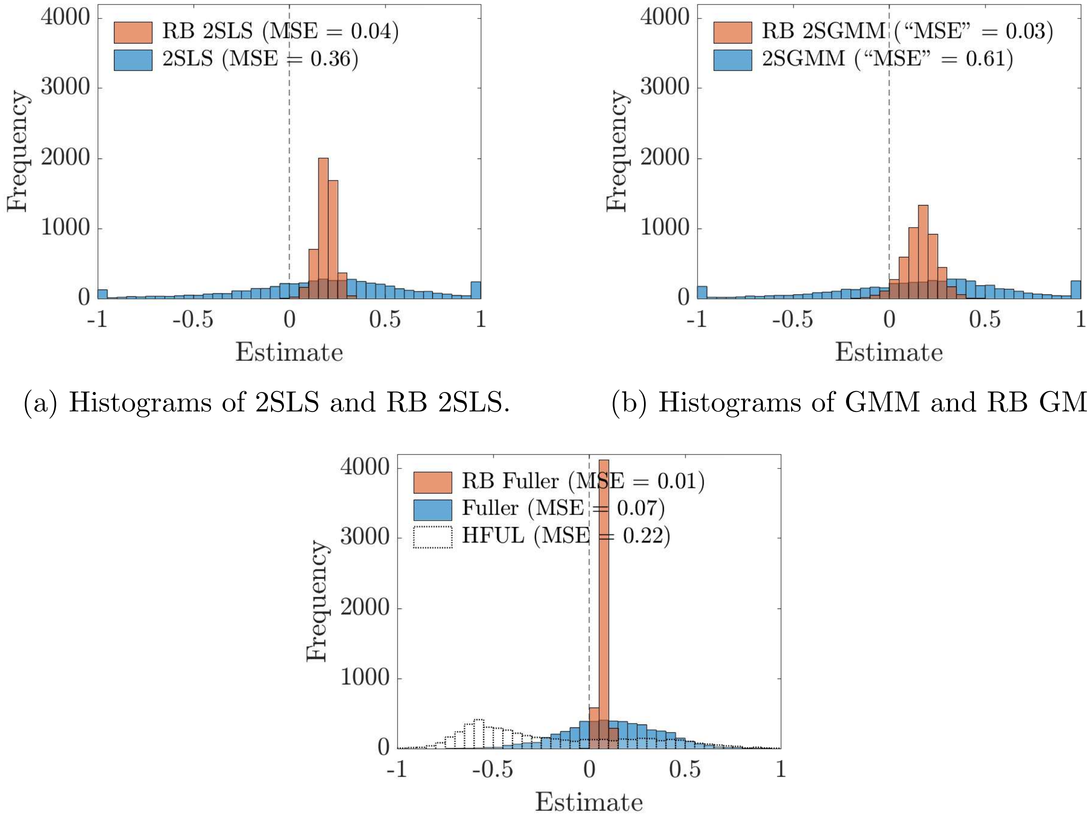
$$\mathbb{E}_U[((\hat{\pi}_{\mathrm{FGLS},n} + U_\pi)'(Z'Z)(\hat{\pi}_{\mathrm{FGLS},n} + U_\pi))^{-1}(\hat{\pi}_{\mathrm{FGLS},n} + U_\pi)'(Z'Z)(\widehat{\pi\beta}_{\mathrm{FGLS},n} + U_{\pi\beta})],$$

where $\mathbb{E}_U$ denotes expectation with respect to $(U_{\pi\beta}, U_\pi)$. The RB two-step GMM estimator of β is given by

$$\begin{aligned} \mathbb{E}_U[((\hat{\pi}_{\mathrm{FGLS},n} + U_\pi)'(Z'Z)\hat{W}(U_{\pi\beta}, U_\pi)(Z'Z)(\hat{\pi}_{\mathrm{FGLS},n} + U_\pi))^{-1} \\ ((\hat{\pi}_{\mathrm{FGLS},n} + U_\pi)'(Z'Z)\hat{W}(U_{\pi\beta}, U_\pi)(Z'Z)(\widehat{\pi\beta}_{\mathrm{FGLS},n} + U_{\pi\beta}))], \end{aligned}$$

where $\hat{W}(U_{\pi\beta}, U_\pi) = \mathbb{E}_n[(y - x'(\hat{\pi}_{\mathrm{FGLS},n} + U_\pi)^{\rightarrow}(\widehat{\pi\beta}_{\mathrm{FGLS},n} + U_{\pi\beta}))^2 zz']$. The RB Fuller estimator of β is given by

[28] Note that $U_{\pi\beta}$ and U_π are already denormalized by $\sqrt{n}$.



(a) Histograms of 2SLS and RB 2SLS.

(b) Histograms of GMM and RB GMM.

(c) Histograms of Fuller and RB Fuller.

Figure 1: Distributions of 2SLS, GMM, and Fuller, and their Rao-Blackwellization under heteroskedasticity (model A). Simulated with 10,000 observations and 5,000 iterations. Clusters at the boundaries indicate observations outside of the range.

$$\mathbb{E}_U\left[\left(\frac{C}{n}X'X+\left[1-\frac{C}{n}\right](\hat{\pi}_{\mathrm{FGLS},n}+U_\pi)'(Z'Z)(\hat{\pi}_{\mathrm{FGLS},n}+U_\pi)\right)^{-1}\right.$$
$$\left.\left(\frac{C}{n}X'y+\left[1-\frac{C}{n}\right](\hat{\pi}_{\mathrm{FGLS},n}+U_\pi)'(Z'Z)(\widehat{\pi\beta}_{\mathrm{FGLS},n}+U_{\pi\beta})\right)\right].$$

The RB unbiased estimator of β does not require Monte Carlo computation of expectations. Andrews and Armstrong (2017) show uniqueness of the unbiased estimator; the RB version of the unbiased estimator constructed with OLS equals the one constructed with GLS, $\frac{\sqrt{n}\widehat{\pi\beta}_{\mathrm{FGLS},n}}{\hat{\sigma}_{\pi\beta,\mathrm{FGLS},n}}\frac{1-\Phi(\sqrt{n}\hat{\pi}_{\mathrm{FGLS},n}/\hat{\sigma}_{\pi,\mathrm{FGLS},n})}{\hat{\sigma}_{\pi,\mathrm{FGLS},n}\phi(\sqrt{n}\hat{\pi}_{\mathrm{FGLS},n}/\hat{\sigma}_{\pi,\mathrm{FGLS},n})}$.

Figure 1a is the histogram of 2SLS and RB 2SLS in model A. The vertical dotted line indicates the true value, $\beta = 0$. It shows that the distribution of RB 2SLS is more concentrated than 2SLS, illustrating the power of LAR. Note that since Rao-Blackwellization does not affect its mean, both estimators have the same bias. Figure 1b is the histogram of GMM and RB GMM for the same run as Figure 1a. Their distributions are very close to their 2SLS counterparts. Figure 1c is the histogram of Fuller, RB Fuller, and HFUL for the same run. Due to the shrinkage property, RB Fuller is very concentrated.

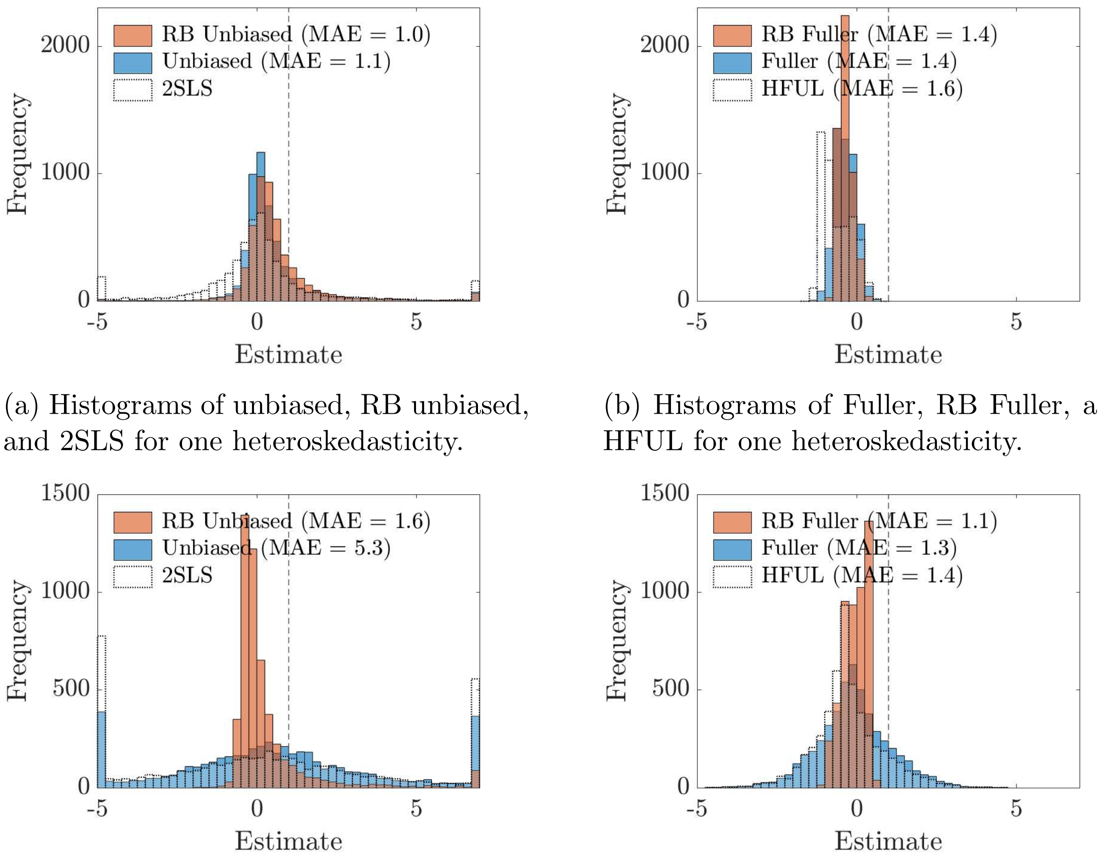
To connect to Theorem 6, consider two loss functions $\ell : \mathbb{R} \to \mathbb{R}$ such that $\ell(x) = x^2$ and $\ell(x) = |x|$. That is, the expected losses are measured by the mean squared error (MSE) and mean absolute error (MAE).[29] The MSE of 2SLS is 0.36 while that of RB 2SLS is 0.04; the "MSE" of GMM is 0.61 while that of RB GMM is 0.03; the MSE of Fuller is 0.07 while that of RB Fuller is 0.01 and that of HFUL is 0.22. The MAE of 2SLS is 0.42 while that of RB 2SLS is 0.19; the "MAE" of GMM is 0.46 while that of RB GMM is 0.17; the MAE of Fuller, RB Fuller, and HFUL are 0.21, 0.07, and 0.42. LAR (Theorem 6) guarantees that the MSE of the RB versions never exceeds that of the original ones, at least asymptotically. It is, therefore, preferable to use a weakly efficient estimator whenever available.

The second simulation is the comparison of unbiased and Fuller with RB unbiased and RB Fuller. We use a just-identified model since the closed-form expression is available in Andrews and Armstrong (2017).[30] We also present 2SLS for comparison, although 2SLS is not subject to LAR due to the lack of the first moment;[31] GMM coincides with 2SLS as it is just-identified. Figures 2a and 2b are the histograms of unbiased and RB unbiased, and of Fuller and RB Fuller in model B with one type of heteroskedasticity, which show slight improvement by LAR; Figures 2c and 2d are the histograms with another type of heteroskedasticity. Figures 2c and 2d show that improvement of RB unbiased estimators can vary. Note that Fuller is biased under the employed weak identification asymptotics. The vertical dotted lines represent the true value, $\beta = 1$. With $d = k = 1$, the unbiased estimator does not have a second moment (Andrews and Armstrong, 2017), so we use MAE as the measure of dispersion. Although in Figure 2a the distribution of RB unbiased does not necessarily *look* more concentrated, its MAE is 1.0, improving from 1.1 of unbiased. Table 1 summarizes losses; it confirms Theorem 6, indicating reduction in convex losses from LAR.

[29] Although our simulation suggests that two-step GMM has as many moments as 2SLS, it is not known theoretically. Therefore, we place double quotes around *MSE* and *MAE* of GMM estimators.

[30] Unbiased estimation itself is possible in overidentified models (Andrews and Armstrong, 2017).

[31] Technically, 2SLS may have moments conditional on GLS, in which case Rao-Blackwellization makes some sense. Numerical Rao-Blackwellization indicates numerical losses can go either upward or downward.



(a) Histograms of unbiased, RB unbiased, and 2SLS for one heteroskedasticity.

(b) Histograms of Fuller, RB Fuller, and HFUL for one heteroskedasticity.

(c) Histograms of unbiased, RB unbiased, and 2SLS for another heteroskedasticity.

(d) Histograms of Fuller, RB Fuller, and HFUL for another heteroskedasticity.

Figure 2: Distributions of unbiased, RB unbiased, 2SLS, Fuller, RB Fuller, and HFUL under heteroskedasticity (model B). Simulated with 10,000 observations and 5,000 iterations. Clusters at the boundaries indicate observations outside of the range.

Note that the conditional moment assumption, $\mathbb{E}[u_i \mid z_i] = 0$ and $\mathbb{E}[v_i \mid z_i] = 0$, plays a crucial role in this exercise. OLS is inefficient because this assumption holds; if we only have unconditional moment restrictions, $\mathbb{E}[u_i z_i'] = 0$ and $\mathbb{E}[v_i z_i'] = 0$, then GLS is not consistent to what they define. Another important assumption is the availability of the efficient estimator, GLS. A notable example in which the form of heteroskedasticity is known *a priori* is when y_i is binary and one has a conditional moment restriction, $\mathbb{E}[y_i \mid x_i] = f(x_i)$; the form of heteroskedasticity is uniquely determined by f as $\mathbb{E}[(y_i - f(x_i))^2 \mid x_i] = f(x_i) - f(x_i)^2$. If f can be estimated, for example for being linear, one may use feasible GLS with no additional loss of generality. In other linear models with an unknown form of heteroskedasticity, feasible GLS with a

Table 1: Mean absolute errors (MAEs) and mean squared errors (MSEs).

	Model A		Model B			
	(1)		(2)		(3)	
	MAE	MSE	MAE	MSE	MAE	MSE
2SLS	0.42	0.36	∞	∞	∞	∞
RB 2SLS	0.19	0.04	∞	∞	∞	∞
GMM	"0.46"	"0.61"	∞	∞	∞	∞
RB GMM	"0.17"	"0.03"	∞	∞	∞	∞
Unbiased	—	∞	1.15	∞	5.28	∞
RB unbiased	—	∞	1.04	∞	1.59	∞
Fuller	0.21	0.07	1.357	1.94	1.31	2.47
RB Fuller	0.07	0.01	1.356	1.88	1.06	1.27
HFUL	0.42	0.22	1.62	2.85	1.41	2.64
Observations	10,000	10,000	10,000	10,000	10,000	10,000
Iterations	5,000	5,000	5,000	5,000	5,000	5,000

* (1) Randomly generated heteroskedasticity for Figures 1a to 1c; (2) for Figures 2a and 2b; (3) for Figures 2c and 2d. Quotes indicate that their finiteness is not known.

nonparametric estimator is available under various assumptions (Carroll, 1982; Robinson, 1987; Newey, 1994). See also Romano and Wolf (2017) for recent reinvestigation of the use of GLS in practice.

7 CONCLUSION

This paper studies weak identification in semiparametric models and investigates efficient estimation. First, we show that weak identification is captured by the notion of *weak regularity* with which the parameter value depends on the score asymptotically. This dependence is homogeneous of degree zero and nonlinear, leading to the impossibility of consistent estimation and inference and equivariant estimation. Then, we show that for each weakly regular parameter there exists an underlying parameter that is regular and fully characterizes the weakly regular parameter locally. As underlying regular parameters are not unique, we stipulate two desirable properties of an underlying parameter, *sufficiency* and *minimality*. The minimal sufficient underlying regular parameter contains the necessary and sufficient amount of information of the weakly regular parameter in the tangent space.

Regarding the estimation of a weakly regular parameter as the estimation of the minimal sufficient underlying parameter plus its transformation, we argue that the "efficiency" of the final estimator of a weakly regular parameter can be cast in terms of the involvement of noise in the estimator of the underlying parameter. When such noise exists, we can construct its improvement by taking a conditional expectation of the estimator, shown as *local asymptotic Rao-Blackwellization*. Intuitively, this exploits the property that an efficient estimator of a regular parameter is "asymptotically sufficient" and applies the Rao-Blackwell theorem to the asymptotic representations of the local expansion.[32] Simulation is carried out in a linear IV model, demonstrating that the 2SLS, GMM, Fuller, and unbiased estimators can be made more concentrated given the availability of a (feasible) GLS estimator of the reduced-form coefficients.

APPENDIX

Proof of Lemma 1. Since $\dot{\mathcal{P}}_P$ is assumed to be linear, if $g \in \dot{\mathcal{P}}_P$ then $ag \in \dot{\mathcal{P}}_P$ for every $a \in \mathbb{R}$. If g is induced by a path $t \mapsto Q_t$ and $a > 0$, then ag can be induced by the path $t \mapsto Q_{at}$, which is the same path up to a scaled index. Therefore, if $Q_t \in \mathscr{P}_P \setminus \mathscr{P}_{P,\beta}$ then $Q_{at} \in \mathscr{P}_P \setminus \mathscr{P}_{P,\beta}$, implying that if $g \in \dot{\mathcal{P}}_P \setminus \dot{\mathcal{P}}_{P,\beta}$ then $ag \in \dot{\mathcal{P}}_P \setminus \dot{\mathcal{P}}_{P,\beta}$. Being defined as a difference between a linear space and a cone, $\dot{\mathcal{P}}_{P,\beta}$ is a cone. ■

Proof of Theorem 2. Let $\beta : \mathcal{P}_\beta \to \mathbb{B}$ be weakly regular and β_P nonconstant.

The first assertion. Suppose that $\hat{\beta}_n : \mathcal{X}^n \to \mathbb{B}$ is a consistent sequence of estimators, or even weaker, that there exist two paths $Q_{n1}, Q_{n2} \in \mathscr{P}_{P,\beta}$ inducing $g_1, g_2 \in \dot{\mathcal{P}}_{P,\beta}$ such that $\beta_P(g_1) \neq \beta_P(g_2)$ and $\hat{\beta}_n \rightarrow^{Q_{nj}*} \beta_P(g_j)$ under each $Q_{nj} \in \{Q_{n1}, Q_{n2}\}$. Define $2\varepsilon := \|\beta_P(g_1) - \beta_P(g_2)\|_{\mathbb{B}}$. Denote by Q_{nj}^n the product measure of Q_{nj} on the product sample space $\mathcal{X}^n$. By the portmanteau theorem (Van der Vaart and Wellner, 1996, Theorem 1.3.4) and the assumption of convergence in outer probability, $\limsup_{n\to\infty} Q_{n1}^n(\|\hat{\beta}_n - \beta_P(g_1)\|_{\mathbb{B}}^* \geq \varepsilon) \leq 0$ while $\liminf_{n\to\infty} Q_{n2}^n(\|\hat{\beta}_n - \beta_P(g_1)\|_{\mathbb{B}}^* \geq \varepsilon) \geq \liminf_{n\to\infty} Q_{n2}^n(\|\hat{\beta}_n - \beta_P(g_1)\|_{\mathbb{B},*} > \varepsilon) \geq 1$. Therefore, Q_{n2}^n is not contiguous to Q_{n1}^n. Being paths, however, Q_{n2}^n must be contiguous to P^n and P^n to Q_{n1}^n (Van der Vaart and Wellner, 1996, Lemma 3.10.11 and Theorem 3.10.9), hence a contradiction.

The second assertion. Let $H_0 : \beta \in \mathbb{B}_0$ and $H_1 : \beta \in \mathbb{B}_1$ be the null and alternative hypotheses such that $\mathbb{B}_0$ and $\mathbb{B}_1$ are nonempty. Suppose that $\phi_n : \mathcal{X}^n \to [0, 1]$ is a consistent sequence of tests of H_0 of level $\alpha < 1$ so that there exist two paths

[32] Cattaneo et al. (2012) also exploits "asymptotic sufficiency" of efficient estimators in semiparametric models. See also Le Cam and Yang (2000) and Van der Vaart (2002) for related discussion of "asymptotic sufficiency" in parametric models.

$Q_{n0}, Q_{n1} \in \mathscr{P}_{P,\beta}$ with $\beta_P(g_0) \in \mathbb{B}_0$ and $\beta_P(g_1) \in \mathbb{B}_1$ such that $\phi_n \to^{Q_{n0}*} \alpha$ and $\phi_n \to^{Q_{n1}*} 1$. Then by the same reasoning a contradiction follows.

The third assertion. Let $\hat{\beta}_n$ be an equivariant-in-law sequence of estimators of β with a separable limit law, that is, there exists a fixed separable Borel probability measure L on $\mathbb{B}$ such that $\hat{\beta}_n - \beta(Q_n) \overset{Q_n}{\rightsquigarrow} L$ for every $Q_n \in \mathscr{P}_{P,\beta}$. Pick $g_1, g_2 \in \dot{\mathcal{P}}_{P,\beta}$ such that $\beta_P(g_1) \neq \beta_P(g_2)$. Denote $\beta_1 := \beta_P(g_1)$ and $\beta_2 := \beta_P(g_2)$. Since $\dot{\mathcal{P}}_{P,\beta}$ is a cone (Lemma 1), ag_1 and ag_2 are also in $\dot{\mathcal{P}}_{P,\beta}$ for every $a > 0$ and by homogeneity we have $\beta_P(ag_j) = \beta_j$. For each positive integer k, take $Q_{nk1}, Q_{nk2} \in \mathscr{P}_{P,\beta}$ to be paths that induce scores g_1/k and g_2/k. Let d_{Q_n} denote the metric that metrizes weak topology on $\mathbb{B}$ under Q_n toward separable limits (Van der Vaart and Wellner, 1996, p. 73). For each k, let n_k be such that for every $n \geq n_k$,

$$\int_{\mathcal{X}} \biggl[\frac{dQ_{nk1}^{1/2} - dP^{1/2}}{1/\sqrt{n}} - \frac{1}{2}\frac{g_1}{k} dP^{1/2}\biggr]^2 \vee \int_{\mathcal{X}} \biggl[\frac{dQ_{nk2}^{1/2} - dP^{1/2}}{1/\sqrt{n}} - \frac{1}{2}\frac{g_2}{k} dP^{1/2}\biggr]^2 < \frac{1}{k},$$
$$d_{Q_{nk1}}\bigl(\hat{\beta}_n - \beta(Q_{nk1}), L\bigr) \vee d_{Q_{nk2}}\bigl(\hat{\beta}_n - \beta(Q_{nk2}), L\bigr) < \frac{1}{k}.$$

Then one can take n'_k so that $n'_k \geq n_k$ and $n'_{k+1} > n'_k$ for every k. Construct two paths Q'_{n1} and Q'_{n2} by $Q'_{nj} = Q_{nk_nj}$ where k_n satisfies $n'_{k_n} \leq n < n'_{k_n+1}$. Then $Q'_{nj} \to^{\text{DQM}} P$ with scores equal to zero and $\hat{\beta}_n - \beta(Q'_{nj})$ converges weakly to L under Q'_{nj}. Now we want to show that dQ'^n_{n2}/dQ'^n_{n1} converges to 1 and invoke Le Cam's third lemma. For this, we adopt the same proof strategy as Van der Vaart (1998, Theorem 7.2). Observe that $\mathbb{E}_{Q'_{n1}}\bigl[n\bigl(1 - \frac{dQ'^{1/2}_{n2}}{dQ'^{1/2}_{n1}}\bigr)^2\bigr] \leq \int_{\mathcal{X}}\bigl[\frac{dQ'^{1/2}_{n1} - dQ'^{1/2}_{n2}}{1/\sqrt{n}}\bigr]^2 \to 0$. By Taylor's theorem, $\log x^2 = -2(1-x) - (1-x)^2 + (1-x)^2 R(1-x)$ for some function $R : \mathbb{R} \to \mathbb{R}$ such that $R(1-x) \to 0$ as $x \to 1$. Then, $\log \frac{dQ'^n_{n2}}{dQ'^n_{n1}}(X_1, \dots, X_n) = \log\bigl(\frac{dQ'_{n2}}{dQ'_{n1}}(X_1) \cdots \frac{dQ'_{n2}}{dQ'_{n1}}(X_n)\bigr) = \sum_{i=1}^n \log \frac{dQ'_{n2}}{dQ'_{n1}} = -2\sum_{i=1}^n W_{ni} - \sum_{i=1}^n W_{ni}^2 + \sum_{i=1}^n W_{ni}^2 R(W_{ni})$, where $W_{ni} := 1 - dQ'^{1/2}_{n2}/dQ'^{1/2}_{n1}$. We argue that all three terms converge to zero in probability. Under Q'_{n1},

$$\biggl|\mathbb{E}\sum_{i=1}^n W_{ni}\biggr| = n\biggl|1 - \int \frac{dQ'^{1/2}_{n2}}{dQ'^{1/2}_{n1}} dQ'_{n1}\biggr| \leq \frac{1}{2}\int \biggl[\frac{dQ'^{1/2}_{n1} - dQ'^{1/2}_{n2}}{1/\sqrt{n}}\biggr]^2 \longrightarrow 0,$$
$$\operatorname{Var}\biggl(\sum_{i=1}^n W_{ni}\biggr) \leq \mathbb{E}[nW_{ni}^2] = \mathbb{E}\biggl[n\biggl(1 - \frac{dQ'^{1/2}_{n2}}{dQ'^{1/2}_{n1}}\biggr)^2\biggr] \longrightarrow 0.$$

These results imply that the expectation and variance of $\sum W_{ni}$ converge to zero; hence it converges to zero in probability. The second result implies that nW_{ni}^2 converges to zero in mean; by the law of large numbers $\sum W_{ni}^2$ converges to zero in probability. By

Markov's inequality, $\Pr\bigl(\max_{1\leq i\leq n}|W_{ni}| > \varepsilon\bigr) \leq n\Pr(|W_{ni}| > \varepsilon) \leq n\Pr(nW_{ni}^2 > n\varepsilon^2) \leq \frac{\mathbb{E}[nW_{ni}^2]}{\varepsilon^2} \to 0$ for every $\varepsilon > 0$. Thus, $\max_{1\leq i\leq n}|W_{ni}|$ converges to zero in probability, meaning that $\max_{1\leq i\leq n}|R(W_{ni})|$ converges to zero in probability as well. Therefore, the third term $\sum W_{ni}^2 R(W_{ni})$ converges to zero in probability. We conclude that dQ'^n_{n2}/dQ'^n_{n1} converges to 1 in probability under Q'_{n1}. Since L is separable, by Slutsky's lemma (Van der Vaart and Wellner, 1996, Example 1.4.7), $\bigl(\hat{\beta}_n, \frac{dQ'^n_{n2}}{dQ'^n_{n1}}\bigr) \overset{Q'_{n1}}{\rightsquigarrow} (\beta_1 + L, 1)$. By Le Cam's third lemma (Van der Vaart and Wellner, 1996, Theorem 3.10.7), $(\beta_2 + L)(B) = \mathbb{E}\mathbb{1}\{\beta_1 + L \in B\}1 = (\beta_1 + L)(B)$ for every Borel $B \subset \mathbb{B}$, which contradicts $\beta_1 \neq \beta_2$. ■

Proof of Lemma 3. Denote by $\mathbb{D}$ the Banach space of P-square integrable functions on $\mathcal{X}$ and define $\psi : \mathcal{P} \to \mathbb{D}$ by $\psi(Q) = dQ^{1/2}/dP^{1/2}$. Note that ψ is regular with derivative $\dot{\psi}_P : \dot{\mathcal{P}}_P \to \mathbb{D}$, $\dot{\psi}_P g = g$. Thus, we have $\beta_{P,\psi} = \beta_P$. ■

Proof of Theorem 4. Let $\mathbb{D} = L_2(P)$ and define $\psi : \mathcal{P} \to \mathbb{D}$ by $\psi(Q) = 2\Pi_\beta dQ^{1/2}/dP^{1/2}$. Then ψ is regular with the derivative $\dot{\psi}_P : \dot{\mathcal{P}}_P \to \mathbb{D}$, $\dot{\psi}_P g = \Pi_\beta g$. Note that $\beta_P(g) = \beta_P(\Pi_\beta g)$. This implies that ψ is an underlying regular parameter for β and that $N(\dot{\psi}_P) = N(\beta_P)$, which implies minimal sufficiency of ψ. ■

Proof of Proposition 5. Define $T_{\delta,n} : \mathbb{D} \times [0,1] \to \mathbb{B}$ by $T_{\delta,n}(\delta, u) := T_n\bigl(\delta + \frac{\delta}{\sqrt{n}}, u\bigr)$. Then, $\hat{\beta}_n = T_{\psi(P),n}(\sqrt{n}(\hat{\psi}_n - \psi(P)), U) + o_P(1)$. By the extended continuous mapping theorem (Van der Vaart and Wellner, 1996, Theorem 1.11.1 and Problem 1.11.1), the claim follows. ■

Proof of Theorem 6. Observe that $\bar{T}_n$ is locally continuously root-n approximable at $\psi(P)$ tangentially to the range of $\dot{\psi}_P g + L_\psi$ with the approximating function $\bar{T}_{\psi(P)}(\delta) := \mathbb{E}[T_{\psi(P)}(\delta + L_\eta, U)]$; hence $\bar{T}_n(\hat{\psi}_n)$ is regular. For $Q_n \in \mathscr{P}_{P,\beta}$, write

$$\begin{aligned}\mathbb{E}_*[\ell(\tilde{\beta}_n - \beta)] - \mathbb{E}^*[\ell(\bar{T}_n(\hat{\psi}_n) - \beta)] &= \mathbb{E}_*[\ell(\tilde{\beta}_n - \beta)] - \mathbb{E}[\ell(T_{\psi(P)}(\dot{\psi}_P g + L_\psi + L_\eta, U) - \beta)]\\ &\quad + \mathbb{E}[\mathbb{E}[\ell(T_{\psi(P)}(\dot{\psi}_P g + L_\psi + L_\eta, U) - \beta) - \ell(\bar{T}_{\psi(P)}(\dot{\psi}_P g + L_\psi) - \beta) \mid L_\psi]]\\ &\quad + \mathbb{E}[\ell(\bar{T}_{\psi(P)}(\dot{\psi}_P g + L_\psi) - \beta)] - \mathbb{E}^*[\ell(\bar{T}_n(\hat{\psi}_n) - \beta)].\end{aligned}$$

The first difference converges to zero by Proposition 5 and Van der Vaart and Wellner (1996, Theorem 1.11.3); the second difference is nonnegative since the inner conditional expectation is nonnegative by a generalized Jensen's inequality (To and Yip, 1975); the third difference converges to zero by approximability of $\bar{T}_n$, the extended continuous mapping theorem (Van der Vaart and Wellner, 1996, Theorem 1.11.1 and Problem 1.11.1), and Van der Vaart and Wellner (1996, Theorem 1.11.3). ■

REFERENCES

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

——— (2013): "Maximum Likelihood Estimation and Uniform Inference with Sporadic Identification Failure," *Journal of Econometrics*, 173, 36–56.

——— (2014): "GMM Estimation and Uniform Subvector Inference with Possible Identification Failure," *Econometric Theory*, 30, 287–333.

Andrews, D. W. K. and P. Guggenberger (2017): "Asymptotic Size of Kleibergen's LM and Conditional LR Tests for Moment Condition Models," *Econometric Theory*, 33, 1046–1080.

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

——— (2007): "Performance of Conditional Wald Tests in IV Regression with Weak Instruments," *Journal of Econometrics*, 139, 116–132.

Andrews, I. (2016): "Conditional Linear Combination Tests for Weakly Identified Models," *Econometrica*, 84, 2155–2182.

——— (2017): "On the Structure of IV Estimands," *Journal of Econometrics*, forthcoming.

Andrews, I. and T. B. Armstrong (2017): "Unbiased Instrumental Variables Estimation Under Known First-Stage Sign," *Quantitative Economics*, 8, 479–503.

Andrews, I. and A. Mikusheva (2014): "Weak Identification in Maximum Likelihood: A Question of Information," *American Economic Review: Papers and Proceedings*, 104, 195–199.

——— (2015): "Maximum Likelihood Inference in Weakly Identified Dynamic Stochastic General Equilibrium Models," *Quantitative Economics*, 6, 123–152.

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

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

Antoine, B. and P. Lavergne (2014): "Conditional Moment Models under Semi-Strong Identification," *Journal of Econometrics*, 182, 59–69.

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

——— (2012): "Efficient Minimum Distance Estimation with Multiple Rates of Convergence," *Journal of Econometrics*, 170, 350–367.

Armstrong, T. B. (2016): "Large Market Asymptotics for Differentiated Product Demand Estimators With Economic Models of Supply," *Econometrica*, 84, 1961–1980.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–890.

Bharucha-Reid, A. T. (1972): *Random Integral Equations*, New York and London: Academic Press.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore and London: Johns Hopkins University Press.

Bickel, P. J. and Y. Ritov (2000): "Non- and Semiparametric Statistics: Compared and Contrasted," *Journal of Statistical Planning and Inference*, 91, 209–228.

Bound, J., D. A. Jaeger, and R. M. Baker (1995): "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak," *Journal of the American Statistical Association*, 90, 443–450.

Canova, F. and L. Sala (2009): "Back to square one: Identification issues in DSGE models," *Journal of Monetary Economics*, 56, 431–449.

Carroll, R. J. (1982): "Adapting for Heteroscedasticity in Linear Models," *Annals of Statistics*, 10, 1224–1233.

Cattaneo, M. D., R. K. Crump, and M. Jansson (2012): "Optimal Inference for Instrumental Variables Regression with Non-Gaussian Errors," *Journal of Econometrics*, 167, 1–15.

Chao, J. C., J. A. Hausman, W. K. Newey, N. R. Swanson, and T. Woutersen (2012): "An Expository Note on the Existence of Moments of Fuller and HFUL Estimators," in *Essays in Honor of Jerry Hausman*, ed. by B. H. Baltagi, R. C. Hill, W. K. Newey, and H. L. White, Emerald Group Publishing Limited, vol. 29 of *Advances in Econometrics*, 87–106.

Chao, J. C. and N. R. Swanson (2005): "Consistent Estimation with a Large Number of Weak Instruments," *Econometrica*, 73, 1673–1692.

Chaudhuri, S. and E. Zivot (2011): "A New Method of Projection-Based Inference in GMM with Weakly Identified Nuisance Parameters," *Journal of Econometrics*, 164, 239–251.

Cheng, X. (2015): "Robust Inference in Nonlinear Models with Mixed Identification Strength," *Journal of Econometrics*, 189, 207–228.

Chernozhukov, V., C. Hansen, and M. Jansson (2009): "Admissible Invariant Similar Tests for Instrumental Variables Regression," *Econometric Theory*, 25, 806–818.

Cox, G. (2017): "Weak Identification in a Class of Generically Identified Models with an Application to Factor Models," Ph.D. thesis, Yale University.

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

——— (2003): "Identification, Weak Instruments, and Statistical Inference in Econometrics," *Canadian Journal of Economics*, 36, 767–808.

Dufour, J.-M., L. Khalaf, and M. Kichian (2006): "Inflation Dynamics and the New Keynesian Phillips Curve: An Identification Robust Econometric Analysis," *Journal of Economic Dynamics & Control*, 30, 1707–1727.

Dufour, J.-M. and M. Taamouti (2005): "Projection-Based Statistical Inference in Linear Structural Models with Possibly Weak Instruments," *Econometrica*, 73, 1351–1365.

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

Fang, Z. (2015): "Estimation and Inference of Directionally Differentiable Functions: Theory and Applications," Ph.D. thesis, University of California, San Diego.

——— (2016): "Optimal Plug-in Estimators of Directionally Differentiable Functionals," Working paper.

Fang, Z. and A. Santos (2015): "Inference on Directionally Differentiable Functions," Working paper.

Fuller, W. A. (1977): "Some Properties of a Modification of the Limited Information Estimator," *Econometrica*, 45, 939–953.

Giné, E. and R. Nickl (2015): *Mathematical Foundations of Infinite-Dimensional Statistical Models*, Cambridge: Cambridge University Press.

Groeneboom, P. and J. A. Wellner (1992): *Information Bounds and Nonparametric Maximum Likelihood Estimation*, Basel: Birkhäuser.

Guerron-Quintana, P., A. Inoue, and L. Kilian (2013): "Frequentist Inference in Weakly Identified Dynamic Stochastic General Equilibrium Models," *Quantitative Economics*, 4, 197–229.

Guggenberger, P. (2005): "Monte-Carlo Evidence Suggesting a No Moment Problem of the Continuous Updating Estimator," *Economics Bulletin*, 3, 1–6.

Guggenberger, P., F. Kleibergen, S. Mavroeidis, and L. Chen (2012): "On the Asymptotic Sizes of Subset Anderson-Rubin and Lagrange Multiplier Tests in Linear Instrumental Variables Regression," *Econometrica*, 80, 2649–2666.

Guggenberger, P. and R. J. Smith (2005): "Generalized Empirical Likelihood Estimators and Tests under Partial, Weak, and Strong Identification," *Econometric Theory*, 21, 667–709.

——— (2008): "Generalized Empirical Likelihood Tests in Time Series Models with Potential Identification Failure," *Journal of Econometrics*, 142, 134–161.

Hahn, J., J. C. Ham, and H. R. Moon (2011): "The Hausman Test and Weak Instruments," *Journal of Econometrics*, 160, 289–299.

Han, S. and A. McCloskey (2017): "Estimation and Inference with a (Nearly) Singular Jacobian," Working paper.

Hausman, J. A., W. K. Newey, T. Woutersen, J. C. Chao, and N. R. Swanson (2012): "Instrumental Variable Estimation with Heteroskedasticity and Many Instruments," *Quantitative Economics*, 3, 211–255.

Hirano, K. and J. R. Porter (2012): "Impossibility Results for Nondifferentiable Functionals," *Econometrica*, 80, 1769–1790.

——— (2015): "Location Properties of Point Estimators in Linear Instrumental Variables and Related Models," *Econometric Reviews*, 34, 719–732.

Hong, H. and J. Li (2017): "The Numerical Delta Method and Bootstrap," Working paper.

Iskrev, N. I. (2008): "How Much Do We Learn from the Estimation of DSGE Models? A Case Study of Identification Issues in a New Keynesian Business Cycle Model," in *Essays On Identification And Estimation of Dynamic Stochastic General Equilibrium Models*, Ph.D. thesis, University of Michigan.

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

——— (2004): "Testing Subsets of Structural Parameters in the Instrumental Variables Regression Model," *Review of Economics and Statistics*, 86, 418–423.

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

——— (2007): "Generalizing Weak Instrument Robust IV Statistics towards Multiple Parameters, Unrestricted Covariance Matrices and Identification Statistics," *Journal of Econometrics*, 139, 181–216.

Kosorok, M. R. (2008): *Introduction to Empirical Processes and Semiparametric Inference*, New York: Springer.

Le Cam, L. and G. L. Yang (2000): *Asymptotics in Statistics*, New York: Springer, second ed.

Magnusson, L. M. (2010): "Inference in Limited Dependent Variable Models Robust to Weak Identification," *Econometrics Journal*, 13, S56–S79.

Magnusson, L. M. and S. Mavroeidis (2010): "Identification-Robust Minimum Distance Estimation of the New Keynesian Phillips Curve," *Journal of Money, Credit and Banking*, 42, 465–481.

Mavroeidis, S. (2010): "Monetary Policy Rules and Macroeconomic Stability: Some New Evidence," *American Economic Review*, 100, 491–503.

Mikusheva, A. (2010): "Robust Confidence Sets in the Presence of Weak Instruments," *Journal of Econometrics*, 157, 236–247.

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

——— (2009): "Tests with Correct Size When Instruments can be Arbitrarily Weak," *Journal of Econometrics*, 152, 131–140.

Müller, U. K. (2011): "Efficient Tests Under a Weak Convergence Assumption," *Econometrica*, 79, 395–435.

Müller, U. K. and Y. Wang (2017): "Nearly Weighted Risk Minimal Unbiased Estimation," Working paper.

Nason, J. M. and G. W. Smith (2008): "Identifying the New Keynesian Phillips Curve," *Journal of Applied Econometrics*, 23, 525–551.

Nelson, C. R. and R. Startz (1990): "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator," *Econometrica*, 58, 967–976.

Newey, W. K. (1994): "Series Estimation of Regression Functionals," *Econometric Theory*, 10, 1–28.

Newey, W. K. and F. Windmeijer (2009): "Generalized Method of Moments With Many Weak Moment Conditions," *Econometrica*, 77, 687–719.

Otsu, T. (2006): "Generalized Empirical Likelihood Inference for Nonlinear and Time Series Models under Weak Identification," *Econometric Theory*, 22, 513–527.

Phillips, P. C. B. (1984): "The Exact Distribution of LIML: I," *International Economic Review*, 25, 249–261.

——— (1989): "Partially Identified Econometric Models," *Econometric Theory*, 5, 181–240.

Pollard, D. (1997): "Another Look at Differentiability in Quadratic Mean," in *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, ed. by D. Pollard, E. Torgersen, and G. L. Yang, New York: Springer, chap. 19, 305–314.

Qu, Z. (2014): "Inference in Dynamic Stochastic General Equilibrium Models with Possible Weak Identification," *Quantitative Economics*, 5, 457–494.

Robinson, P. M. (1987): "Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form," *Econometrica*, 55, 875–891.

Romano, J. P. and M. Wolf (2017): "Resurrecting Weighted Least Squares," *Journal of Econometrics*, 197, 1–19.

Ruge-Murcia, F. J. (2007): "Methods to Estimate Dynamic Stochastic General Equilibrium Models," *Journal of Economic Dynamics & Control*, 31, 2599–2636.

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

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

To, T.-O. and K. W. Yip (1975): "A Generalized Jensen's Inequality," *Pacific Journal of Mathematics*, 58, 255–259.

van der Vaart, A. W. (1988): *Statistical Estimation in Large Parameter Spaces*, Amsterdam: Centrum voor Wiskunde en Informatica.

——— (1991a): "Efficiency and Hadamard Differentiability," *Scandinavian Journal of Statistics*, 18, 63–75.

——— (1991b): "On Differentiable Functionals," *Annals of Statistics*, 19, 178–204.

——— (1998): *Asymptotic Statistics*, Cambridge: Cambridge University Press.

——— (2002): "The Statistical Work of Lucien Le Cam," *Annals of Statistics*, 30, 631–682.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics*, New York: Springer.

Wang, J. and E. Zivot (1998): "Inference on Structural Parameters in Instrumental Variables Regression with Weak Instruments," *Econometrica*, 66, 1389–1404.

Wellner, J. A., C. A. J. Klaassen, and Y. Ritov (2006): "Semiparametric Models: A Review of Progress since BKRW," in *Frontiers in Statistics*, ed. by J. Fan and H. L. Koul, London: Imperial College Press.

Yogo, M. (2004): "Estimating the Elasticity of Intertemporal Substitution When Instruments Are Weak," *Review of Economics and Statistics*, 86, 797–810.

Zivot, E., R. Startz, and C. R. Nelson (1998): "Valid Confidence Intervals and Inference in the Presence of Weak Instruments," *International Economic Review*, 39, 1119–1144.

——— (2006): "Improved Inference in Weakly Identified Instrumental Variables Regression," in *Econometric Theory and Practice: Frontiers of Analysis and Applied Research*, ed. by D. Corbae, S. N. Durlauf, and B. E. Hansen, New York: Cambridge University Press.

THEORY OF WEAK IDENTIFICATION IN SEMIPARAMETRIC MODELS
Online Appendix

Tetsuya Kaji

June 6, 2018

Abstract

This is the online supplementary appendix to "Theory of Weak Identification in Semiparametric Models."

S.1 SUPPORTIVE RESULTS

This section provides supportive results. The following lemma is characterization of the minimal tangent cone.

Lemma S.1. *The following hold.*

i. $N(\beta_P)$ *is a linear space.*

ii. *If* $P \in \mathcal{P} \setminus \mathcal{P}_\beta$, *then* $N(\beta_P) \subset \dot{\mathcal{P}}_P \setminus \dot{\mathcal{P}}_{P,\beta}$.

iii. *If* $P \in \mathcal{P} \setminus \mathcal{P}_\beta$, *then* $g \in \dot{\mathcal{P}}_{P,\beta}$ *implies* $\Pi_\beta g \neq 0$.

Proof. (i) Trivially, $0 \in N(\beta_P)$. The definition of the kernel implies that if $\tilde{g} \in N(\beta_P)$, then $-\tilde{g} \in N(\beta_P)$. Take $\tilde{g} \in N(\beta_P)$ and $a > 0$. Since $\dot{\mathcal{P}}_{P,\beta}$ is a cone (Lemma 1) and β_P is homogeneous of degree zero, $\beta_P(g) = \beta_P(g/a) = \beta_P(g/a + \tilde{g}) = \beta_P(g + a\tilde{g})$ for every $g \in \dot{\mathcal{P}}_{P,\beta}$. This means $a\tilde{g} \in N(\beta_P)$. Therefore, $N(\beta_P)$ is linear.

(ii) If $P \in \mathcal{P} \setminus \mathcal{P}_\beta$, then $0 \notin \dot{\mathcal{P}}_{P,\beta}$. Since $g \in N(\beta_P) \cap \dot{\mathcal{P}}_{P,\beta}$ implies $\beta_P(g) = \beta_P(g - g) = \beta_P(0)$, $N(\beta_P) \cap \dot{\mathcal{P}}_{P,\beta}$ must be empty.

(iii) If $\Pi_\beta g = 0$ then $g \in N(\beta_P)$, which implies $g \notin \dot{\mathcal{P}}_{P,\beta}$ by (ii). ■

The following theorem characterizes minimal sufficient underlying regular parameters.

Theorem S.2 (Characterization of minimal sufficient underlying regular parameter)**.** *Let* $\beta : \mathcal{P}_\beta \to \mathbb{B}$ *be weakly regular and* $\psi : \mathcal{P} \to \mathbb{D}$ *a sufficient underlying regular parameter for* β. *Then* ψ *is minimal if and only if for any sufficient underlying regular parameter* $\phi : \mathcal{P} \to \mathbb{E}$ *for* β *on a Banach space* $\mathbb{E}$ *there exists a linear map* $\tau : \mathbb{E} \to \mathbb{D}$ *such that*

$$\tau(\dot{\phi}_P g) = \dot{\psi}_P g \qquad \text{for every} \qquad g \in \dot{\mathcal{P}}_P.$$

Remark. Theorem S.2 can be understood as "almost uniqueness" of observational information regarding minimal sufficient underlying regular parameters. If the linear map τ between two minimal sufficient underlying regular parameters is bicontinuous, efficiency in one parameterization implies efficiency in the other (see, e.g., Van der Vaart, 1991, 1998, Section 25.7).

Proof. Sufficiency. Assume that for any sufficient underlying regular parameter $\phi : \mathcal{P} \to \mathbb{E}$ for β there exists a map $\tau : \mathbb{E} \to \mathbb{D}$ such that $\tau(\dot{\phi}_P g) = \dot{\psi}_P g$ for every $g \in \dot{\mathcal{P}}_{P,\beta}$. This means that $N(\dot{\phi}_P) \subset N(\dot{\psi}_P)$. Take ϕ to be minimal; then $N(\beta_P) = N(\dot{\phi}_P) \subset N(\dot{\psi}_P)$. On the other hand, since ψ is assumed to be a sufficient underlying parameter, we have $N(\beta_P) \supset N(\dot{\psi}_P)$.

Necessity. Assume that $\psi : \mathcal{P} \to \mathbb{D}$ is a minimal sufficient underlying regular parameter for β. Take $\phi : \mathcal{P} \to \mathbb{E}$ to be another sufficient underlying regular parameter for β. Then $\beta_{P,\psi}(\dot{\psi}_P g) = \beta_{P,\phi}(\dot{\phi}_P g)$ for every $g \in \dot{\mathcal{P}}_{P,\beta}$ and $N(\dot{\psi}_P) = N(\beta_P) \supset N(\dot{\phi}_P)$. The first property implies $\dot{\psi}_P g \in \beta_{P,\psi}^{-1}\beta_{P,\phi}(\dot{\phi}_P g)$ for every $g \in \dot{\mathcal{P}}_{P,\beta}$. The second property implies that if $\dot{\phi}_P g_1 = \dot{\phi}_P g_2$ then $\dot{\psi}_P g_1 = \dot{\psi}_P g_2$. Conclude that there exists a linear map $\tau : \mathbb{E}_0 \to \mathbb{D}$ such that $\dot{\psi}_P g = \tau(\dot{\phi}_P g)$ for $g \in \dot{\mathcal{P}}_0$ where $\mathbb{E}_0 := \dot{\phi}_P(\dot{\mathcal{P}}_{P,\beta})$. One can extend τ on the whole of $\mathbb{E}$ by letting $\tau(e) := \tau(\Pi_{\mathbb{E}_0} e)$. ■

S.2 GENERAL WEAK LINEAR IV MODELS

This section analyzes the general linear IV model from Example 1 in which π approaches a rank deficient matrix instead of zero. Recall

$$\begin{cases} y_i = z_i'\psi_1 + u_i, & \mathbb{E}[z_i u_i] = 0, \\ x_i' = z_i'\psi_2 + v_i', & \mathbb{E}[z_i v_i'] = 0, \end{cases}$$

where $\beta = \psi_2^{\rightarrow}\psi_1$. We are interested in a path Q_n that approaches a point of identification failure P such that

$$\psi_2(Q_n) = \pi + \frac{\dot{\pi}}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right), \qquad \beta(Q_n) = \beta + \frac{\dot{\beta}}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right),$$
$$\psi_1(Q_n) = \psi_2(Q_n)\beta(Q_n) = \pi\beta + \frac{\dot{\pi}\beta + \pi\dot{\beta}}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right).$$

Apply the singular value decomposition to π to write $\pi = USV'$ for a $k \times k$ orthogonal matrix U, a $k \times d$ nonnegative diagonal matrix S, and a $d \times d$ orthogonal matrix V. If β is weakly regular, then the number ℓ of positive elements of S is strictly less than d.

The regression equations can then be written as

$$\begin{cases} y_i = \tilde{z}_i' \tilde{\psi}_1 + u_i, \\ \tilde{x}_i' = \tilde{z}_i' \tilde{\psi}_2 + \tilde{v}_i', \end{cases}$$

for $\tilde{\psi}_1 = \Lambda^{-1} U' \psi_1$, $\tilde{\psi}_2 = \Lambda^{-1} U' \psi_2 V$, $\tilde{\beta} = V' \beta$, $\tilde{z}_i = \Lambda U' z_i$, $\tilde{x}_i = V' x_i$, $\tilde{v}_i = V' v_i$, and Λ^{-1} is a positive diagonal matrix such that the first ℓ diagonal entries are inverses of positive entries of S and all other diagonal entries are one. Since $\psi_2(P) = \pi$ is known in the limit, we may assume without loss of generality that π is a diagonal matrix whose first ℓ elements are one and all others zero.

Due to this structure of π, the first ℓ elements of β are uniquely determined by $\psi_1(P) = \pi\beta$ and $\psi_2(P) = \pi$ regardless of the score. On the other hand, the remaining elements of β must be determined by the local parameters $\dot{\psi}_{1,P}(g) = \dot{\pi}\beta + \pi\dot{\beta}$ and $\dot{\psi}_{2,P}(g) = \dot{\pi}$. For β to be weakly regular, the remaining components of β must be uniquely determined by the score. Since $A\pi = 0$ for a $k \times k$ diagonal matrix A whose first ℓ diagonal elements are zero and remaining elements one, if the last $d - \ell$ elements of β are uniquely determined by $\dot{\psi}_{2,P}(g) = \dot{\pi}$ and $A\dot{\psi}_{1,P}(g) = A\dot{\pi}\beta$, then β is weakly regular; this is the case when the lower bottom $(k - \ell) \times (d - \ell)$ matrix of $\dot{\pi}$ is of full column rank. We henceforth make this assumption.

In order to see when the first ℓ elements of β become regular, we aim to represent the first ℓ elements of $\dot{\beta}$ as a continuous linear map of the score. Similarly as Example 1, the score is given by

$$g = g_{uvz} - z'(\dot{\pi}\beta + \pi\dot{\beta}) \frac{dP_{uvz,u}}{dP} - z'\dot{\pi} \frac{dP_{uvz,v}}{dP}$$

and we have $\mathbb{E}_P[zug] = \mathbb{E}_P[zz'](\dot{\pi}\beta + \pi\dot{\beta})$ and $\mathbb{E}_P[zv'g] = \mathbb{E}_P[zz']\dot{\pi}$. Denote by β_ℓ and $\beta_{-\ell}$ the first ℓ and last $d - \ell$ elements of β and $\dot{\pi} =: \left[\begin{smallmatrix} \dot{\pi}_1 & \dot{\pi}_3 \\ \dot{\pi}_2 & \dot{\pi}_4 \end{smallmatrix}\right]$ where $\dot{\pi}_1$ is $\ell \times \ell$. Since β_ℓ does not depend on the score, we may take its value as granted. Then

$$\begin{aligned} \mathbb{E}_P\left[z\left(u - v' \begin{bmatrix} \beta_\ell \\ 0 \end{bmatrix}\right) g\right] &= \mathbb{E}_P[zz'] \left(\dot{\pi} \begin{bmatrix} 0 \\ \beta_{-\ell} \end{bmatrix} + \pi\dot{\beta}\right) \\ &= \mathbb{E}_P[zz'] \left(\begin{bmatrix} 0 & \dot{\pi}_3 \\ 0 & \dot{\pi}_4 \end{bmatrix} \begin{bmatrix} 0 \\ \beta_{-\ell} \end{bmatrix} + \begin{bmatrix} I_\ell & 0 \\ 0 & 0 \end{bmatrix} \dot{\beta}\right). \end{aligned}$$

Therefore, if $\dot{\pi}_3$ is zero, then with the $k \times k$ diagonal matrix B whose first ℓ diagonal elements are one and all others zero, one obtains

$$B\mathbb{E}_P[zz']^{-1}\mathbb{E}_P\left[z\left(u - v'\begin{bmatrix}\beta_\ell\\ 0\end{bmatrix}\right)g\right] = B\mathbb{E}_P[zz']^{-1}\pi\dot{\beta}.$$

Thus, one can represent the first ℓ components of $\dot{\beta}$ as a linear functional of the score. Otherwise, one needs to know $\beta_{-\ell}$ in order to eliminate $\dot{\pi}\beta$, suggesting that β_ℓ might not be regular.

Now we investigate the tangent spaces. With these notations, we have

$$\beta_P(g) = \begin{bmatrix}\beta_\ell\\ \dot{\pi}_4^{\to}(\dot{\pi}_4\beta_{-\ell})\end{bmatrix},$$

where $\dot{\pi}_4$ and $\dot{\pi}_4\beta_{-\ell}$ are calculated from

$$\dot{\pi} = \mathbb{E}_P[zz']^{-1}\mathbb{E}_P[zv'g], \qquad \begin{bmatrix}0\\ \dot{\pi}_4\beta_{-\ell}\end{bmatrix} = A\mathbb{E}_P[zz']^{-1}\mathbb{E}_P\left[z\left(u - v'\begin{bmatrix}\beta_\ell\\ 0\end{bmatrix}\right)g\right],$$

whence we may guess the minimal sufficient underlying parameter. Denote by $\psi_{1,\ell}$ and $\psi_{1,-\ell}$ the first ℓ and last $k-\ell$ components of ψ_1 and $\psi_2 =: \left[\begin{smallmatrix}\psi_{2,1} & \psi_{2,2}\\ \psi_{2,3} & \psi_{2,4}\end{smallmatrix}\right]$ for an $\ell \times \ell$ submatrix $\psi_{2,1}$. Then the parameters $\psi_{1,-\ell}$ and $\psi_{2,4}$ that induce local parameters $\dot{\pi}_4\beta_{-\ell}$ and $\dot{\pi}_4$ can be guessed to constitute the minimal sufficient underlying parameter. To compute the nuisance tangent space, recall the score formula

$$\begin{aligned} g = g_{uvz} &- z_\ell'(\dot{\pi}_1\beta_\ell + \dot{\pi}_3\beta_{-\ell})\frac{dP_{uvz,u}}{dP} - z_{-\ell}'(\dot{\pi}_2\beta_\ell + \dot{\pi}_4\beta_{-\ell})\frac{dP_{uvz,u}}{dP} \\ &- z'\pi\dot{\beta}\frac{dP_{uvz,u}}{dP} - (z_\ell'\dot{\pi}_1 + z_{-\ell}'\dot{\pi}_2)\frac{dP_{uvz,v,\ell}}{dP} - (z_\ell'\dot{\pi}_3 + z_{-\ell}'\dot{\pi}_4)\frac{dP_{uvz,v,-\ell}}{dP}, \end{aligned}$$

where z_ℓ and $z_{-\ell}$ denote the first ℓ and last $k-\ell$ components of z. It is clear that adding the scores

$$g_{uvz}, \quad z'\pi C\frac{dP_{uvz,u}}{dP}, \quad z_\ell' D\beta_\ell\frac{dP_{uvz,u}}{dP} + z_\ell' D\frac{dP_{uvz,v,\ell}}{dP}, \quad z_{-\ell}' E\beta_\ell\frac{dP_{uvz,u}}{dP} + z_{-\ell}' E\frac{dP_{uvz,v,\ell}}{dP}$$

for any $d \times 1$ vector C, $\ell \times \ell$ matrix D, and $(k-\ell) \times \ell$ matrix E does not change the value of $\beta_{-\ell}$; therefore, they are in $N(\beta_P)$. Also, adding the score

$$z_\ell' F\frac{dP_{uvz,v,-\ell}}{dP}$$

for any $\ell \times (d-\ell)$ matrix F would change the values of $\dot{\pi}_3$ and $\dot{\beta}$ (depending on the value of $\beta_{-\ell}$) but does not change the value of $\beta_{-\ell}$ itself; therefore, it is in $N(\beta_P)$ as well.

S.3 PARTIAL WEAK IDENTIFICATION

It has been observed that weak identification and partial identification in linear IV models arise from similar sources (Poskitt and Skeels, 2013). In the context of weak instruments, if the first-stage coefficients approach degeneracy too quickly, then even if the structural parameter β is identified at each n, it suffers partial identification in the corresponding local expansion (Section S.3). In extremum estimation models, Cox (2017) uses a higher-order Taylor expansion to characterize the asymptotic behavior of the extremum estimator under the asymptotic embedding he calls “super-weak sequences of parameters” which cause similar phenomena. Another example of partial weak identification appears in DSGE models; Andrews and Mikusheva (2016) note in their supplementary material that a simple DSGE model approaches a limit at which only four out of six parameters are identified.

In this section, we explain how these cases involve partial identification in the local expansion. Partial identification in the local expansion means that given the model score, which summarizes all information of the first-order approximation of the model, the value of the parameter of interest is not pinned down uniquely.

The first example is when the first-stage coefficients in the linear IV model approach zero faster than root-n. The second example is the DSGE model considered in Andrews and Mikusheva (2016).

Example 1 (Linear IV with partial weak identification, continued)**.** If the first-stage coefficients converge to degeneracy faster than root-n, the structural parameter β suffers partial identification in the limit. Let $d = k = 2$ and $(u, v) \sim N(0, I_3)$, and consider the embedding

$$\pi_n = \begin{pmatrix} \frac{1}{\sqrt{n}} + \frac{1}{n} & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} + \frac{1}{n} \end{pmatrix}.$$

The first-stage coefficients π_n are of full rank at each n and approach zero with rate $\sqrt{n}$, and $\sqrt{n}\pi$ approaches a degenerate matrix $\pi_0 := \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ with rate n. The corresponding path is given by

$$dQ_n = \frac{1}{(2\pi)^{3/2}} \exp\biggl(-\frac{(y - z'\pi_n\beta)^2 + (x' - z'\pi_n)(x' - z'\pi_n)'}{2}\biggr).$$

From here, the score can be calculated as

$$\sqrt{n}\,\frac{dQ_n - dP}{dP} \longrightarrow g = z'\pi_0\beta y + z'\pi_0 x = z'\begin{pmatrix}\beta_1+\beta_2\\ \beta_1+\beta_2\end{pmatrix} y + z'\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix} x.$$

Therefore, only the combination $\beta_1+\beta_2$ is identified as a function of the score, but not β_1 and β_2 separately.

Example S.1 (Full information DSGE model)**.** Consider the simple DSGE model in Andrews and Mikusheva (2014) and Andrews and Mikusheva (2016, Supplementary Appendix):

$$\begin{aligned} b\mathbb{E}_t\pi_{t+1} + \kappa x_t - \pi_t &= 0, && \text{(Phillips curve)}\\ r_t - \mathbb{E}_t\pi_{t+1} - \rho\Delta a_t &= \mathbb{E}_t x_{t+1} - x_t, && \text{(Euler equation)}\\ b^{-1}\pi_t + u_t &= r_t, && \text{(Monetary policy)}\end{aligned}$$

where the shock processes follow

$$\begin{cases}\Delta a_t = \rho\Delta a_{t-1} + \varepsilon_{a,t},\\ u_t = \delta u_{t-1} + \varepsilon_{u,t},\end{cases} \qquad \begin{pmatrix}\varepsilon_{a,t}\\ \varepsilon_{u,t}\end{pmatrix} \sim N\left(0, \begin{bmatrix}\sigma_a^2 & 0\\ 0 & \sigma_u^2\end{bmatrix}\right) \text{ i.i.d.}$$

These equations reduce to

$$\begin{aligned} m_{t1} &:= \frac{\varepsilon_{a,t}}{\sigma_a} = \frac{1}{\sigma_a}\frac{(1-\rho b)(1-\delta b)}{b^2\kappa}\frac{(b+\kappa-\rho b)}{\rho(\rho-\delta)}\left[\pi_t - \rho\pi_{t-1} - \frac{\kappa}{1-\delta b}(x_t - \rho x_{t-1})\right],\\ m_{t2} &:= \frac{\varepsilon_{u,t}}{\sigma_u} = \frac{1}{\sigma_u}\frac{(1-\rho b)(1-\delta b)}{b^2\kappa}\frac{(b+\kappa-\delta b)}{(\rho-\delta)}\left[\pi_t - \delta\pi_{t-1} - \frac{\kappa}{1-\rho b}(x_t - \delta x_{t-1})\right],\end{aligned}$$

which are i.i.d. We are concerned about identification when ρ is close to δ, that is, $\rho_T = \delta + h/\sqrt{T}$. Other parameters are left unspecified; thus, we consider $b_T = b + h_b/\sqrt{T}$, $\kappa_T = \kappa + h_\kappa/\sqrt{T}$, $\sigma_{a,T} = \sigma_a + h_a/\sqrt{T}$, and $\sigma_{u,T} = \sigma_u + h_u/\sqrt{T}$.

Since (m_{t1}, m_{t2}) is $O_P(1)$, the terms in the square brackets stay $O_P(1/\sqrt{T})$. In light of this, define

$$\frac{z_t}{\sqrt{T}} := \pi_t - \delta\pi_{t-1} - \frac{\kappa}{1-\delta b}(x_t - \delta x_{t-1}).$$

Then the asymptotic representation of the bracket terms can be given by

$$\pi_t - \rho_T\pi_{t-1} - \frac{\kappa_T(x_t - \rho_T x_{t-1})}{1-\delta b_T} = \frac{z_t}{\sqrt{T}} - \frac{1}{\sqrt{T}}\left[\frac{h_\kappa}{\kappa} + \frac{\delta h_b}{1-\delta b}\right]\frac{\kappa(x_t - \delta x_{t-1})}{1-\delta b}$$

$$
\begin{aligned}
&\quad - \frac{h}{\sqrt{T}}\left[\pi_{t-1} - \frac{\kappa x_{t-1}}{1-\delta b}\right] + o_P\left(\frac{1}{\sqrt{T}}\right), \\
\pi_t - \delta\pi_{t-1} - \frac{\kappa_T(x_t - \delta x_{t-1})}{1-\rho_T b_T} &= \frac{z_t}{\sqrt{T}} - \frac{1}{\sqrt{T}}\frac{hb+\delta h_b}{1-\delta b}\frac{\kappa(x_t-\delta x_{t-1})}{1-\delta b} \\
&\quad - \frac{1}{\sqrt{T}}\frac{h_\kappa}{\kappa}\frac{\kappa(x_t-\delta x_{t-1})}{1-\delta b} + o_P\left(\frac{1}{\sqrt{T}}\right).
\end{aligned}
$$

The distribution of z_t will be uniquely determined by that of $(m_{t1}, m_{t2}) \sim Q_T$. Let P be the limit distribution that yields $\rho = \delta \in (0,1)$ (but satisfies $1-\delta b \neq 0$, $b+\kappa-\delta b \neq 0$, $b \neq 0$, $\kappa \neq 0$, $\sigma_a > 0$, $\sigma_u > 0$). Denote the weak limit of (m_{t1}, m_{t2}) by

$$
\begin{aligned}
m_1 &:= \underbrace{\frac{1}{\sigma_a}\frac{(1-\delta b)^2}{b^2\kappa}\frac{b+\kappa-\delta b}{\delta}}_{A}\left[\frac{z}{h} - \pi_{-1} + \frac{\kappa x_{-1}}{1-\delta b}\right], \\
m_2 &:= \underbrace{\frac{1}{\sigma_u}\frac{(1-\delta b)^2}{b^2\kappa}(b+\kappa-\delta b)}_{B}\left[\frac{z}{h} - \frac{b\kappa(x-\delta x_{-1})}{(1-\delta b)^2}\right].
\end{aligned}
$$

Finally, define the sequence of moments by

$$
\begin{aligned}
m_{t1,T} &:= \frac{(1-\rho_T b_T)(1-\delta b_T)}{\sigma_{a,T} b_T^2 \kappa_T}\frac{(b_T+\kappa_T-\rho_T b_T)}{\rho_T(\rho_T-\delta)}\left[\pi_t - \rho_T\pi_{t-1} - \frac{\kappa_T(x_t-\rho_T x_{t-1})}{1-\delta b_T}\right], \\
m_{t2,T} &:= \frac{(1-\rho_T b_T)(1-\delta b_T)}{\sigma_{u,T} b_T^2 \kappa_T}\frac{(b_T+\kappa_T-\delta b_T)}{(\rho_T-\delta)}\left[\pi_t - \delta\pi_{t-1} - \frac{\kappa_T(x_t-\delta x_{t-1})}{1-\rho_T b_T}\right],
\end{aligned}
$$

After tedious algebra, one can compute the score for this general path as

$$
\begin{aligned}
&\sqrt{T}\big[dQ_T(m_{t1,T}, m_{t2,T}) - dP(m_1, m_2)\big] \\
&\quad \longrightarrow g = g_Q + \frac{dP_1}{dP}m_1\left[-\frac{h_a}{\sigma_a} - \frac{bh+2\delta h_b}{1-\delta b} - \frac{2h_b}{b} - \frac{h_\kappa}{\kappa} + \frac{h_b+h_\kappa-bh-\delta h_b}{b+\kappa-\delta b} - \frac{h}{\delta}\right] \\
&\qquad\qquad + \frac{dP_1}{dP}A\left[-\frac{1}{h}\left(\frac{h_\kappa}{\kappa} + \frac{\delta h_b}{1-\delta b}\right)\frac{\kappa(x-\delta x_{-1})}{1-\delta b}\right] \\
&\qquad\qquad + \frac{dP_2}{dP}m_2\left[-\frac{h_u}{\sigma_u} - \frac{bh+2\delta h_b}{1-\delta b} - \frac{2h_b}{b} - \frac{h_\kappa}{\kappa} + \frac{h_b+h_\kappa-\delta h_b}{b+\kappa-\delta b}\right] \\
&\qquad\qquad + \frac{dP_2}{dP}B\left[-\frac{1}{h}\left(\frac{h_\kappa}{\kappa} + \frac{\delta h_b}{1-\delta b}\right)\frac{\kappa(x-\delta x_{-1})}{1-\delta b}\right] \\
&\qquad =: g_Q + \frac{dP_1}{dP}m_1 H_1 + \frac{dP_1}{dP}(x-\delta x_{-1})H_A + \frac{dP_2}{dP}m_2 H_2 + \frac{dP_2}{dP}(x-\delta x_{-1})H_B.
\end{aligned}
$$

Thus, by knowing the model score g we can recover only up to four local parameters H_1, H_2, H_A, and H_B but not all of six parameters; there is no map $\beta_P : \dot{\mathcal{P}}_{P,\beta} \to \mathbb{R}^6$

that recovers six parameters.

REFERENCES

Andrews, I. and A. Mikusheva (2014): “Weak Identification in Maximum Likelihood: A Question of Information,” *American Economic Review: Papers and Proceedings*, 104, 195–199.

——— (2016): “A Geometric Approach to Nonlinear Econometric Models,” *Econometrica*, 84, 1249–1264.

Cox, G. (2017): “Weak Identification in a Class of Generically Identified Models with an Application to Factor Models,” Ph.D. thesis, Yale University.

Poskitt, D. S. and C. L. Skeels (2013): “Inference in the Presence of Weak Instruments: A Selected Survey,” *Foundations and Trends® in Econometrics*, 6, 1–99.

van der Vaart, A. W. (1991): “Efficiency and Hadamard Differentiability,” *Scandinavian Journal of Statistics*, 18, 63–75.

——— (1998): *Asymptotic Statistics*, Cambridge: Cambridge University Press.