Forchini, Giovanni; Hillier, Grant H.

Working Paper

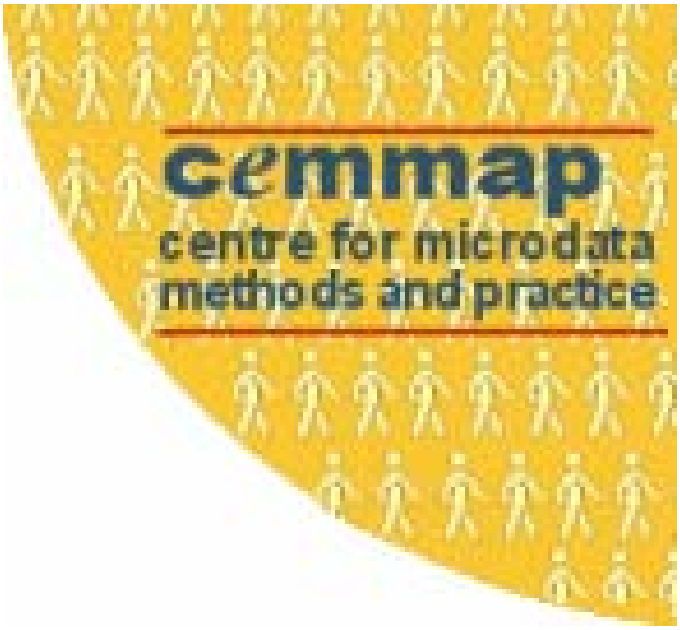
Ill-conditioned problems, Fisher information, and weak instruments

cemmap working paper, No. CWP04/05

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Forchini, Giovanni; Hillier, Grant H. (2005) : Ill-conditioned problems, Fisher information, and weak instruments, cemmap working paper, No. CWP04/05, Centre for Microdata Methods and Practice (cemmap), London, http://dx.doi.org/10.1920/wp.cem.2005.0405

This Version is available at:
http://hdl.handle.net/10419/79381



ILL-CONDITIONED PROBLEMS, FISHER INFORMATION, AND WEAK INSTRUMENTS

Giovanni Forchini
Grant Hillier

THE INSTITUTE FOR FISCAL STUDIES
DEPARTMENT OF ECONOMICS, UCL
cemmap working paper CWP04/05

Ill-Conditioned Problems, Fisher Information, and Weak Instruments

Giovanni Forchini* and Grant Hillier**

April 5 2005

Abstract

The existence of a uniformly consistent estimator for a particular parameter is well-known to depend on the uniform continuity of the functional that defines the parameter in terms of the model. Recently, Pötscher (*Econometrica*, 70, pp 1035 - 1065) showed that estimator risk may be bounded below by a term that depends on the *oscillation* (*osc*) of the functional, thus making the connection between continuity and risk quite explicit. However, *osc* has no direct statistical interpretation. In this paper we slightly modify the definition of *osc* so that it reflects a (generalized) *derivative* (*der*) of the functional. We show that *der* can be directly related to the familiar statistical concepts of Fisher information and identification, and also to the condition numbers that are used to measure 'distance from an ill-posed problem' in other branches of applied mathematics. We begin the analysis assuming a fully parametric setting, but then generalize to the nonparametric case, where the inverse of the Fisher information matrix is replaced by the covariance matrix of the efficient influence function. The results are applied to a number of examples, including the structural equation model, spectral density estimation, and estimation of variance and precision.

KEYWORDS: Continuity, Derivative, Divergence, Fisher Information, Ill-conditioned problem, Ill-posed problem, Interest-functional, Oscillation, Precision.

---

*Department of Econometrics and Business Statistics, Faculty of Economics Monash University, Clayton, Vic. 3800 Australia; Corresponding Author.
e-mail: Giovanni.Forchini@BusEco.monash.edu.au

**CEMMAP and Economics Division, School of Social Sciences, University of Southampton, Highfield SO17 1BJ, Southampton, U.K.;
e-mail: ghh@soton.ac.uk

1 Introduction

It has been well understood for many years that the possibility of 'successful inference' on a parameter of interest requires that the mapping defining the interest parameter as a functional on the space of distributions should be sufficiently smooth. Early contributors to this understanding were (Bahadur and Savage 1956), (Singh 1963), and (LeCam and Schwartz 1960). These papers showed that the existence of uniformly consistent estimators and bounded confidence sets both require that (with some appropriate topology on the space of distributions) the map $\{set\ of\ distributions\} \to \{interest\ parameter\}$ be uniformly continuous. The subject continues to attract attention from both statisticians generally, and econometricians - see (Koschat 1987), (Gleser and Hwang 1987), (Pfanzagl 1998), (Dufour 1997), (Pötscher 2002), for instance.

To be more precise, if η is a parameter of interest, defined as a functional on the family ($\mathcal{P}$) of model densities under consideration, and we endow $\mathcal{P}$ with the topology induced by, say, the *total variation distance* (see equation (1) below), the continuity or otherwise of the map defining η essentially determines the properties of inferential procedures for η. For uniformly consistent estimators of η, and confidence sets of bounded size for η, to exist, the map $\eta : \mathcal{P} \to \mathbb{R}^q$ must be uniformly continuous, or must be rendered so by either imposing primitive conditions on the parameter space for $\mathcal{P}$, or otherwise restricting the set of probability measures $\mathcal{P}$ on which η is defined. In parametric models, this is usually done by imposing *identification* conditions (e.g., in instrumental variable models), or *nonredundancy* conditions (in time series models). In nonparametric models, further conditions are usually required (see Sections 4 and 5 of (Pötscher 2002) and Section 5 below). (Pötscher 2002) calls a problem *ill-posed* if the usual indicator of continuity, the *oscillation* of the functional (*osc*), does not vanish everywhere on the parameter space. In that case, no uniformly consistent estimator for η can exist.

The problem is intimately related to the identification problem familiar to econometricians, and (Rothenberg 1971), and (Sargan 1983) both discuss the identification problem from a closely related point of view. Note, though, that in contrast to much of the discussion that has taken place in the econometric literature, ill-posedness is a *property of the functional of interest,* not just of the model, and thus depends on the properties of both η and $\mathcal{P}$, and the way they interact. This aspect of the problem will become quite clear in what follows.

Although (Pötscher 2002) does succeed in relating the statistical properties of the problem (e.g., estimator *risk*) to the properties of the functional (*osc*), there are two difficulties that arise with those results. First, *osc* is an 'all-or-nothing' property of a function, whereas one suspects from the recent "weak instruments" literature in econometrics that being close to an ill-posed problem might lead to inferential difficulties, albeit not to the impossibility results that characterise the

extreme case. The other is that *osc* seems to have no direct relation to other familiar statistical concepts.

In this paper we attempt to remedy both of those deficiencies. We propose a measure of the properties of the functional η that is essentially a generalisation of the derivative - and we therefore denoted it by *der*. This is a measure of the sensitivity of the interest parameter η to perturbations in the underlying density. It can take any positive value, and takes the value $+\infty$ when the problem is ill-posed. Smaller values indicate that inference should be relatively easy. It turns out (see Theorem 3 and Corollary 1 below) that *der* is essentially determined by the familiar Fisher Information matrix - in particular, the minimum eigenvalue of that matrix (modified by terms reflecting the properties of η). As we shall see, the measure we propose also has close connections with both numerical analysis and convex optimisation theory, where *condition numbers* are used to measure distance from a critical point, in the first case, and problem difficulty in the second. An *ill-posed problem* in these contexts is the extreme case, and the condition number measures "distance from ill-posedness" (see, e.g., (Zolezzi 2002)). In fact, our *der* could well be called instead the *condition number for the problem of inference on* η. In an instrumental variables context it provides a direct measure of instrument weakness (see Section 5.1 below), but also extends that idea to much more general inference problems.

The plan of the paper is as follows. In Section 2 we discuss the definition of *der* and derive its properties in the context of a fully parametric family of densities. In this case η is regarded as an ordinary function of the model parameters. In Section 4 we extend the analysis to a nonparametric setting, arriving at analogous results that are expressed in terms of the covariance matrix of the efficient influence function - the analogue of Fisher information for nonparametric inference.

The phrase 'successful inference' is used above to reflect the fact that every inferential procedure has two components: the procedure itself (e.g., construction of a point estimate, or a confidence set), and a measure of the precision that can be attached to that procedure (e.g., risk, mean squared error, asymptotic variance, or expected size of the confidence set). We shall show (Theorem 2) that *der* provides a direct measure of how difficult it is to assess the accuracy of inferences on η. Since measures of precision are themselves typically functionals on the space of distributions, in Section 3 we apply the earlier results to the problem of assessing the variance and/or precision of an estimator, again in a fully parametric setting. Section 5 contains a number of examples that illustrate the usefulness of the earlier analysis.

2 Fully Parametric problems

Let $\mathcal{P}$ be a family of probability measures on a common measurable space $(\mathcal{X}, \mathcal{A})$. Throughout the paper we assume that the distributions in $\mathcal{P}$ possess densities (with respect to some common dominating measure), and largely ignore technical measurability issues. Further, we assume that the family of densities $\mathcal{P}$ is parameterised by $\theta \in \Omega$, with Ω an open subset of $\mathbb{R}^q$. We denote the densities in $\mathcal{P}$ by p_θ, for the most part suppressing the data x, say, from the notation. Thus, $\mathcal{P} = \{p_\theta : \theta \in \Omega\}$ becomes a $q-$dimensional manifold.

To measure distance in $\mathcal{P}$ we use the notion of *divergence* between distributions ((Amari 1985) p. 84). This is a function $\delta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}^+$, denoted by $\delta(p_0, p_1)$ for $p_0 = p_{\theta_0}, p_1 = p_{\theta_1} \in \mathcal{P}$, with the following properties:

(1) $\delta(p_0, p_1) \geq 0$, and $\delta(p_0, p_1) = 0$ if and only if $p_0 = p_1$;

(2) $\mathrm{D}_{\theta_0} \delta(p_0, p_1)|_{\theta_0=\theta_1} = \mathrm{D}_{\theta_1} \delta(p_0, p_1)|_{\theta_1=\theta_0} = 0$, and

(3) $\mathrm{D}^2_{\theta_1} \delta(p_0, p_1)\big|_{\theta_1=\theta_0} = G(\theta_0)$, a positive (semi)-definite matrix,

where D_θ denotes differentiation with respect to θ.

Remark 1 *We assume throughout that the matrix $G(\theta)$ is non-singular for all θ. If that is not the case the statements and proofs of some of the results to follow must be modified. If there were a θ_0 for which rank$[G(\theta_0)] < q$, the parameter θ itself (and therefore also any one-to-one function of it) would be unidentified. This would not rule out the existence of identified functions of θ, in the manner of (Phillips 1989) concept of "partially identified models", but would considerably complicate the discussion that follows. An example occurs in Section 5.2, where we outline a possible procedure to deal with such situations.*

The choice of δ determines the matrix G, and the discussion that follows can be cast in terms of any relevant divergence function. However, for purposes of exposition we assume that G is the Fisher Information matrix $G(\theta_0) = -E_{\theta_0}\left[\mathrm{D}^2_\theta \ln(p_\theta)|_{\theta=\theta_0}\right]$. Thus, we assume that the divergence of interest is what (Blyth 1994) calls "locally Rao". See also (Gibbs and Su 2002) for a survey of divergence measures and the relationships among them. The total variation distance

$$\delta(p_0, p_1) = \sup_{A \in \mathcal{A}} \{|P_0(A) - P_1(A)|\}, P_0, P_1 \in \mathcal{P}. \tag{1}$$

used in the statistical literature mentioned in the Introduction, and the Hellinger distance, an example of a divergence that is locally Rao, induce the same topology

on the set of probability measures (see, for example, (LeCam and Yang 1990) and (Gibbs and Su 2002)). Here, $P_0\,(P_1)$ denotes the distribution corresponding to the density p_{θ_0} (p_{θ_1}). Note that, in general, G would depend on the sample size n, but again we suppress this in the notation.

A fundamental property of the divergence is that it behaves locally as half the square of a Euclidean distance, as shown by the following Lemma.

Lemma 1 *(i) The divergence $\delta(p_0, p_1)$ between two neighbouring points $p_0 = p_{\theta_0}$ and $p_1 = p_{\theta_1}$ has the form:*

$$\delta(p_0, p_1) = \frac{1}{2}(\theta_1 - \theta_0)' G(\theta_0)(\theta_1 - \theta_0) + O\left(|\theta_1 - \theta_0|^3\right).$$

(ii) (Morse's Lemma) There is a neighbourhood $\mathcal{N}(\theta_0)$ of θ_0, and a diffeomorphism ψ, such that $\psi(0) = 0$ and for every $\theta \in \mathcal{N}(\theta_0)$,

$$\delta\left(p_{\theta_0}, p_{\theta_0+\psi(\tau)}\right) = \tau'\tau.$$

For a discussion of, and background on, Morse's lemma see (Milnor 1963) and (Castrigiano and Hayes 1993).

Next, an *interest parameter* is, in general, defined as a map $\eta : \mathcal{P} \to \mathbb{R}^m$ $(m \leq q)$. However, we shall here confine attention to interest parameters that are ordinary functions of θ, so that $\eta = \eta(\theta)$ can also be regarded as a function from $\mathbb{R}^q \to \mathbb{R}^m$. Nevertheless, it is still the properties of η as a function in the first sense that will prove decisive for inference. We assume that $\eta(\cdot)$ is differentiable everywhere in Ω, and denote the differential of η, $D_\theta\eta(\theta)$ (an $m \times q$ matrix), by $\dot{\eta}(\theta)$. In addition, we assume throughout that $rank[\dot{\eta}(\theta)] = m$ for all θ. This assumption on the rank of $\dot{\eta}(\theta)$ is not essential, but simplifies the presentation of the results.

In the statements of the results to follow we use the following notation: a ball in $\mathbb{R}^q$ of radius ε and centre at the point θ_0 in $\mathbb{R}^q$ will be the set

$$\mathcal{B}_{\mathbb{R}^q}(\theta_0, \varepsilon) = \left\{\theta \in \mathbb{R}^q : (\theta - \theta_0)' G(\theta_0)(\theta - \theta_0) < \varepsilon^2\right\}.$$

By definition, this depends on the matrix G, so is adapted to a particular divergence measure (or family of measures sharing the same G). Similarly, a ball of radius ε centred at the point $p_0 \in \mathcal{P}$ in the manifold of densities is the set $\mathcal{B}_{\mathcal{P}}(p_0, \varepsilon) = \{p \in \mathcal{P} : \delta(p_0, p) < \varepsilon^2\}$. In view of Lemma 1, for sufficiently small ε we may identify $\mathcal{B}_{\mathcal{P}}(p_0, \varepsilon)$ with the set

$$\mathcal{B}_{\mathcal{P}}(\theta_0, \varepsilon) = \{\theta \in \mathcal{B}_{\mathbb{R}^q}(\theta_0, \varepsilon)\} \subset \mathbb{R}^q. \tag{2}$$

Now, just as for ordinary functions (cf. (Spivak 1965, p. 13)), the continuity properties of the map $\eta : \mathcal{P} \to \mathbb{R}^m$ can be captured by the *oscillation* of η at p_0, defined as follows:

Definition 1 *For $p_0 \in \mathcal{P}$, $\eta : \mathcal{P} \to \mathbb{R}^m$, and $\mathcal{B}_{\mathcal{P}}(\theta_0, \varepsilon)$ defined as above,*

$$osc(\eta, p_0) = \lim_{\varepsilon \to 0} \sup_{\theta \in \mathcal{B}_{\mathcal{P}}(\theta_0, \varepsilon)} |\eta(\theta) - \eta(\theta_0)|. \tag{3}$$

Then, if $\hat{\eta}$ is an estimator of $\eta(\theta)$, we have the following result from (Pötscher 2002). (Strictly speaking (Pötscher 2002) considers more general risk functions than we do, and a topology on the set of probability measures based on the total variation distance. As noted above, this induces the same topology on the set of probability measures as the Hellinger distance.)

Theorem 1 *((Pötscher 2002)) The risk of any estimator $\hat{\eta}$ of $\eta(\theta_0)$ satisfies the inequality*

$$\lim_{\varepsilon \to 0} \sup_{\theta \in \mathcal{B}_{\mathcal{P}}(\theta_0, \varepsilon)} E_\theta\left(|\hat{\eta} - \eta(\theta)|^2\right) \geq \frac{1}{4}\left[osc(\eta, p_0)\right]^2,$$

where E_θ denotes the expectation taken with respect to the density p_θ.

Since $osc(\eta, p_0) = 0$ if and only if the function η is continuous at θ_0, Theorem 1 establishes the link between the risk of an estimator and the continuity of η: if η is discontinuous at some point θ_0, then the risk of any estimator for η must be positive (and may be infinite) in a neighbourhood of p_0, even when the sample size tends to infinity. It follows from this result that there can be no uniformly consistent estimator for η if there is a $\theta_0 \in \Omega$ for which $osc(\eta, p_0) > 0$. Intuitively, this is because a small perturbation of the density (i.e, θ) can, if *osc* is positive, induce an arbitrarily large change in η.

If we say that an interest parameter $\eta(\theta_0)$ is *identified* if and only if $p_\theta \to p_{\theta_0}$ implies that $\eta(\theta) \to \eta(\theta_0)$, Lemma 1 implies that (provided $G(\theta)$ has full rank q), $\eta(\theta_0)$ is identified if and only if $osc(\eta, p_0) = 0$. So, $osc(\eta, p_0) > 0$, implies that $\eta(\theta_0)$ is *unidentified.* Note here that θ_0 is not (necessarily) the true value of θ, merely some fixed value.

Now, although Theorem 1 does provide a key connection between the properties of the interest parameter η and the possibility of inference for it, the result, like $osc(\cdot, \cdot)$ itself, is essentially "all or nothing": either a uniformly consistent estimator for η exists, or not. Thus, situations where $osc(\eta, p_0) > 0$ are usually ruled out by assumption, in the sense that conditions are imposed on either $\mathcal{P}$ or η to guarantee that the map $\eta : \mathbb{R}^q \to \mathbb{R}^m$ is continuous. However, recent literature on "weak identification" (e.g., (Staiger and Stock 1997), (Dufour 1997)) has made it clear that there can be problems for inference about an interest parameter that is technically identified ($osc = 0$ everywhere), but for which the map $\eta : \mathcal{P} \to \mathbb{R}^m$ is *close to* being discontinuous. In particular, there can be real difficulty in assessing the precision that can be assigned to inferential procedures for such a parameter (see (Staiger and Stock 1997), and (Forchini and Hillier 2003)).

The situation may be illustrated by the function $\eta(\theta) = 1/(1+\exp\{-\kappa\theta\})$. It is easy to see that η is everywhere continuous for all finite κ (i.e., $osc(\eta,\theta) = 0$ for all θ), but tends to the step function $\eta^*(\theta) = 0$ for $\theta < 0$, $\eta^*(\theta) = 1$ for $\theta > 0$ (with a discontinuity at the origin), as $\kappa \to \infty$. On the other hand, for κ large, η is "almost" a step function at the origin, and small changes in θ lead to huge changes in η. Of course, what we are now talking about is the *derivative* of η at the origin ($= \kappa/4$). It seems clear intuitively that for interest functionals exhibiting this kind of behaviour inference is likely to be extremely difficult, but not impossible.

Our purpose now will be to formalise this intuition, to show that it is essentially correct, and to explore its connections with the more familiar arguments outlined above. It turns out that the measure we define as the analogue of the derivative has a second advantage: it is intimately related to the properties of the matrix G implicit in the divergence measure used, and therefore, in a wide variety of cases, to the Fisher information matrix. It is also closely related to the definition of the *condition number* ($cond(\cdot,\cdot)$) that arises in other branches of applied mathematics (see (Zolezzi 2002), and (Demmel 1987)). In that context, an extreme (infinite) value of $cond(\cdot,\cdot)$ indicates an ill-posed problem, and finite values calibrate the difficulty of the problem. Because of our more statistical context, and the link with the derivative of the interest-functional, we shall use a different notation, $der(\eta, p_0)$, defined as follows:

Definition 2 *For fixed $p_0 \in \mathcal{P}$ and $\eta : \mathcal{P} \to \mathbb{R}^m$, and $\mathcal{B}_{\mathcal{P}}(\theta_0, \varepsilon)$ defined as above, define*

$$der(\eta, p_0) = \lim_{\varepsilon \to 0} \sup_{\theta \in \mathcal{B}_{\mathcal{P}}(\theta_0,\varepsilon)} \frac{|\eta(\theta) - \eta(\theta_0)|}{\varepsilon}, \tag{4}$$

Correspondingly, we now consider not the risk $E_\theta\left(|\hat{\eta} - \eta(\theta)|^2\right)$ itself, but how the risk behaves in a shrinking neighbourhood $\mathcal{B}_{\mathcal{P}}(\theta_0, \varepsilon)$ of θ_0, *relative to* the size ε of that neighbourhood. The following Theorem gives the result analogous to Theorem 1 for $der(\eta, p_0)$:

Theorem 2 *For any estimator $\hat{\eta}$ of $\eta(\theta_0)$*

$$\lim_{\varepsilon \to 0} \sup_{\theta \in \mathcal{B}_{\mathcal{P}}(\theta_0,\varepsilon)} E_\theta\left(\left[\frac{|\hat{\eta} - \eta(\theta)|}{\varepsilon}\right]^2\right) \geq \frac{1}{4}\left[der(\eta, p_0)\right]^2$$

where E_θ denotes the expectation taken with respect to the density p_θ.

Remark 2 *If $der(\eta, p_0)$ is large, the Theorem says that the risk of any estimator $\hat{\eta}$ for η can be changing rapidly in some regions of the parameter space. This implies that it will be difficult to accurately assess the precision of any estimator*

for η. Thus, whereas Theorem 1 concerns the possible accuracy with which η can be estimated, Theorem 2 concerns the possible accuracy with which ***that accuracy*** *can be assessed. This theme is developed further in Section 3 below.*

As noted above, one of the advantages of $der(\eta, p_0)$ over $osc(\eta, p_0)$ is that $der(\eta, p_0)$ is directly related to the properties of the model since it depends on the Fisher information for η, a quantity having a familiar statistical interpretation. This relationship is given in the following result:

Theorem 3

$$der(\eta, p_0) = \left[\lambda_M\left([\dot{\eta}(\theta_0)]G^{-1}(\theta_0)[\dot{\eta}(\theta_0)]'\right)\right]^{\frac{1}{2}}, \tag{5}$$

where $\lambda_M(A)$ denotes the largest eigenvalue of the matrix A, and $\dot{\eta}(\theta_0) = \mathrm{D}_\theta \eta(\theta)|_{\theta=\theta_0}$. Moreover, for a fixed interest parameter η, $der(\eta, p_0)$ does not depend on the parameterization of $\mathcal{P}$.

Remark 3 *The expression for $der(\eta, p_0)$ in Theorem 2 involves both the information on θ itself (through $G(\cdot)$), the properties of the function $\eta(\cdot)$ (through $\dot{\eta}(\cdot)$), and how they interact to produce the matrix $[\dot{\eta}(\theta_0)]G^{-1}(\theta_0)[\dot{\eta}(\theta_0)]'$. This makes sense since the data provides information on η only indirectly though θ.*

Remark 4 *The matrix $[\dot{\eta}(\theta_0)]G^{-1}(\theta_0)[\dot{\eta}(\theta_0)]'$ is the (asymptotic) covariance matrix for a consistent and efficient estimator of η, so that $\left([\dot{\eta}(\theta_0)]G^{-1}(\theta_0)[\dot{\eta}(\theta_0)]'\right)^{-1}$ is the (asymptotic) Fisher information for the parameter η. Evidently, $der(\eta, p_0)$ will be large when the Fisher information matrix is 'almost singular'.*

Corollary 1 *Let $\lambda_M(A)$ and $\lambda_m(A)$ denote the largest and the smallest eigenvalues of the matrix A.*
(i) For $\eta = \theta$,

$$der(\theta, p_0) = [\lambda_M\left(G(\theta_0)^{-1}\right)]^{\frac{1}{2}} = [\lambda_m(G(\theta_0))]^{-\frac{1}{2}}. \tag{6}$$

(ii) For $\eta = \theta_1$, where θ is partitioned as $\theta = (\theta_1', \theta_2')'$,

$$der(\theta_1, p_0) = [\lambda_M\left(G_{11.2}(\theta_0)^{-1}\right)]^{\frac{1}{2}} = [\lambda_m(G_{11.2}(\theta_0))]^{-\frac{1}{2}}, \tag{7}$$

where, with G partitioned conformably to θ as

$$G(\theta_0) = \begin{pmatrix} G_{11}(\theta_0) & G_{12}(\theta_0) \\ G_{21}(\theta_0) & G_{22}(\theta_0) \end{pmatrix},$$

$G_{11.2}(\theta_0)$ denotes the matrix

$$G_{11.2}(\theta_0) = G_{11}(\theta_0) - G_{12}(\theta_0)[G_{22}(\theta_0)]^{-1}G_{21}(\theta_0). \tag{8}$$

Moreover, $der(\theta_1, p_0)$ does not depend on the parameterization of the submanifold involving θ_2.

Remark 5 *In some situations the parameter of interest may be only implicitly defined by a relationship of the form* $f(\theta, \eta) = 0$, *where* $f : \mathbb{R}^q \times \mathbb{R}^m \to \mathbb{R}^m$ *is a smooth function with smooth first derivatives. In such a case, the implicit function theorem allows one to write* $\dot{\eta}(\theta_0)$ *in terms of the partial derivatives of* f *(provided the matrix of partial derivatives of* f *with respect to* η *has full rank). Thus, the formulae above can be applied to functions* η *that are only implicitly defined.*

The relationship between $osc(\eta, p_0)$ and $der(\eta, p_0)$ is given by the following Theorem, and the link between $der(\eta, p_0)$ and uniform consistency is given in the Corollary that follows it.

Theorem 4 *Assume that* $G(\theta)$ *has rank* q *for all* $\theta \in \Omega$. *Then: (i)* $der(\eta, p_0) > osc(\eta, p_0)$ *for every* η *and* p_0*; (ii) if* $der(\eta, p_0) < +\infty$ *then* $osc(\eta, p_0) = 0$*; (iii) if* $osc(\eta, p_0) > 0$ *then* $der(\eta, p_0) = +\infty$.

Remark 6 *When, as we have done,* η *is assumed to be differentiable at* θ_0, $der(\eta, p_0) < +\infty$, *and, since differentiability at* θ_0 *implies continuity at* θ_0, $osc(\eta, p_0) = 0$. *Hence, if* η *is differentiable in a compact subset* $\Theta \subset \mathbb{R}^q$, *then it is uniformly continuous in* Θ, *and also uniformly continuous in the set* $\{p_\theta : \theta \in \Theta \subset \mathbb{R}^q, \Theta \text{ compact}\} \subset \mathcal{P}$. *We may state:*

Corollary 2 *If* $der(\eta, p_\theta) < +\infty$ *for all* $\theta \in \Theta$, *where* Θ *is a compact subset of* $\mathbb{R}^q$, *then uniformly consistent estimators for* $\eta(\theta)$ *exist for* $\theta \in \Theta$.

Although we have ruled this out by assumption, part (iii) of Theorem 4 says that if the inference problem is ill-posed in Pötscher's ((Pötscher 2002)) sense ($osc(\eta, p_0) > 0$), then $der(\eta, p_0) = +\infty$. However, since non-differentiability ($der(\eta, p_0) = +\infty$) does not imply non-continuity, the converse is not true. Let us therefore call a problem for which $der(\eta, p_0) = +\infty$ for at least some point(s) in the parameter space an *ill-conditioned problem*, to distinguish this case from that of an *ill-posed* problem in Pötscher's sense. In practice, a problem that is ill-conditioned will typically also be ill-posed - see section 5 below for examples.

If $der(\eta, p_\theta)$ is everywhere near zero, η is insensitive to perturbations of the underlying density, and one does not need to know p_θ very precisely to learn η: inference on η is easy in that case (the problem is *well-conditioned*). On the other hand, if $der(\eta, p_0)$ can be large at some points in the parameter space, small perturbations of θ can generate large changes in $\eta(\theta)$, and one needs to learn p_θ very precisely to learn the value of η (the problem is *ill-conditioned*).

The next two results establish more concretely the relationship between $der(\eta, p_0)$ as we define it, and condition numbers used elsewhere:

Theorem 5 *Let* $B_{\mathbb{R}^m}(\eta_0, \varepsilon) = \{\eta \in \mathbb{R}^m : (\eta - \eta_0)'(\eta - \eta_0) < \varepsilon^2\}$ *denote a ball of radius* ε *in* $\mathbb{R}^m$ *(with the ordinary Euclidean metric). Then:*

$$der\left(\eta, p_0\right) = \lim_{\varepsilon \to 0} \inf \left\{\phi : \eta\left(\mathcal{B}_{\mathcal{P}}\left(\theta_0, \varepsilon\right)\right) \subset B_{\mathbb{R}^m}\left(\eta\left(\theta_0\right), \phi\varepsilon\right)\right\}$$

This Theorem confirms that *der* actually reflects the behaviour it is designed for. The set $\eta\left(\mathcal{B}_{\mathcal{P}}\left(\theta_0, \varepsilon\right)\right)$ is the image of the set of perturbed densities $\mathcal{B}_{\mathcal{P}}\left(\theta_0, \varepsilon\right)$, a ball of radius ε. Roughly speaking, the Theorem says that, for small ε, the radius of the smallest ball around $\eta\left(\theta_0\right)$ that contains $\eta\left(\mathcal{B}_{\mathcal{P}}\left(\theta_0, \varepsilon\right)\right)$ is $\varepsilon \times der\left(\eta, p_0\right)$. Thus, $der\left(\eta, p_0\right)$ measures how seriously a perturbation of p_0 can affect $\eta\left(\theta_0\right)$.

Theorem 6 *Let* $\|A\|_2$ *denote the spectral norm of the square matrix* A*. Then*

$$\min_{A:|A|=0} \left\| \left([\dot{\eta}\left(\theta_0\right)] G^{-1}\left(\theta_0\right) [\dot{\eta}\left(\theta_0\right)]' \right)^{-1} - A \right\|_2 = \left(der\left(\eta, p_0\right)\right)^{-1}$$

Thus, $\left(der\left(\eta, p_0\right)\right)^{-1}$ measures how far the problem of inference on η is from being from an ill-conditioned problem. And, Theorem 6 links ill-posed problems to lack of local identification: it is well known that a sufficient (although not necessary) condition for identification is that the Fisher information matrix is non-singular (see, for instance, Theorem 1 in (Rothenberg 1971), Section 3 of (Bowden 1973), or (Sargan 1983)).

For reasons explained earlier, an issue that has been attracting attention in the econometric literature recently is that of measuring how 'strongly a parameter is identified', or, what amounts to the same thing in an instrumental variables setting, measuring the 'strength of the instruments' used or available. A little reflection shows that what is at issue here is the question of how to capture those features of the inferential problem that affect one's ability to accurately assess the actual (as opposed to asymptotic) inferential precision that can be claimed for a procedure. In view of the results above, we suggest that $\left(\mathrm{der}\left(\eta, p_0\right)\right)^{-1}$ provides just such a measure: a small value of $\left(der\left(\eta, p_0\right)\right)^{-1}$ indicates that inference for η is close to being an ill-conditioned problem, and it will be impossible to know whether *any* reported precision measure is accurate. A large value indicates that inference for η is likely to be straightforward, and assessments of precision should be reasonably reliable. The results in the next section, and the examples discussed in Section 5 below, both support this suggestion.

Although we shall not discuss confidence sets in detail here, in constructing a confidence set, say $C\left(x; \eta\right)$, for η, one would like $C\left(x; \eta\right)$ to be bounded with probability one. If $osc\left(\eta, p_0\right) > 0$ at some point θ_0 (so $der\left(\eta, p_0\right) = +\infty$) this is not possible. Heuristically, by perturbing θ_0 slightly to $\theta_0 + \Delta$, the change in $\eta\left(\theta_0\right)$ is so large that no bounded confidence set can contain $\eta\left(\theta_0 + \Delta\right)$. In this case, to be able to make precise inference about η one needs to exclude neighbourhoods of

points like θ_0 from the admissible parameter space. This is the heuristic behind the impossibility theorems of (Bahadur and Savage 1956), (Singh 1963), (Koschat 1987), (Gleser and Hwang 1987), (Dufour 1997), (Pfanzagl 1998) and (Pötscher 2002).

3 Variance and Precision

Given an asymptotically efficient estimator $\hat{\eta}$ for η, its (usually, asymptotic) variance $v(\theta)$, and its precision $v(\theta)^{-1}$, become new interest parameters, and the question arises, can *these* be estimated accurately? For simplicity we consider only the case $\eta = \theta_1$ (scalar), with $\hat{\theta}_1$ an asymptotically efficient estimator for θ_1. The asymptotic Fisher Information (precision) matrix for $\hat{\theta}_1$ is $G_{11.2}(\theta)$, and its asymptotic covariance matrix is $G_{11.2}^{-1}(\theta)$ (Amari (1985), Theorem 8.1). From our earlier results we have the following:
For interest parameter $\eta = G_{11.2}(\theta)$ (precision),

$$der(G_{11.2}, p_\theta) = [\dot{G}_{11.2}(\theta) G(\theta)^{-1} \dot{G}_{11.2}(\theta)']^{\frac{1}{2}},$$

and, for interest parameter $\eta = G_{11.2}^{-1}(\theta)$ (variance),

$$\begin{aligned} der(G_{11.2}^{-1}, p_\theta) &= der(G_{11.2}, p_\theta)/G_{11.2}(\theta). \\ &= der(G_{11.2}, p_\theta)[der(\theta_1, p_\theta)]^2. \end{aligned} \tag{9}$$

where $\dot{G}_{11.2}(\theta)$ is the $1 \times q$ vector $D_\theta G_{11.2}(\theta)$.

This, of course, is simply a version of the familiar formula for the derivative of an inverse. However, it does have the interesting consequences stated in:

Proposition 1 *(i) If $der(\theta_1, p_\theta)$ is large, so that inference on θ_1 is close to being an ill-conditioned problem, estimation of the (asymptotic) variance of an efficient estimator for θ_1 will also be an ill-conditioned problem, but inference on its precision will be well-conditioned. (ii) Conversely, when $der(\theta_1, p_\theta)$ is small, estimation of the variance will be well-conditioned, but estimation of precision will be close to ill-conditioned.*

The first part of this proposition says that, if a parameter is very sensitive to the underlying density, the variance of an asymptotically efficient estimator for it will also be sensitive, but estimation precision - which will necessarily be near zero, of course - will not be. In the opposite situation, estimation of variance will be well-conditioned, but - precisely because the parameter is insensitive to the underlying density - inference on precision will be difficult. These results evidently generalise in the obvious way.

4 Semiparametric ill-posed problems

We now extend the analysis above to a semi/non-parametric set-up. We consider a random sample $X_1, X_2, ..., X_n$ from a distribution P belonging to a family $\mathcal{P}$ of probability measures on a common measurable space $(\mathcal{X}, \mathcal{A})$. We assume that there exists a "carrier" measure μ on $(\mathcal{X}, \mathcal{A})$ such that all the measures in $\mathcal{P}$ are absolutely continuous with respect to it. It follows from the Radon-Nikodym theorem that each probability measure in $\mathcal{P}$ has a density. The density of P is denoted by p.

Let $\mathcal{P}_0$ be a one-dimensional family of probability measures on $(\mathcal{X}, \mathcal{A})$ which pass through $P_0 \in \mathcal{P}$ (with probability density function p_0) and are differentiable in quadratic mean at P_0. That is, one considers *differentiable paths* $t \to P_t$ from $[0, \varepsilon)$, for some $\varepsilon > 0$, to $\mathcal{P}$ such that, as $t \to 0$,

$$\int_X \left[\frac{p_t^{1/2} - p_0^{1/2}}{t} - \frac{1}{2} g p_0^{1/2} \right]^2 d\mu \to 0 \tag{10}$$

for some measurable function $g : \mathcal{X} \to \mathbb{R}$. The function g can be called the *score function* for the model P_t. By letting $t \to P_t$ range over a collection of submodels, we obtain a collection of score functions called the *tangent set* to $\mathcal{P}$ at P_0, and denoted by $\dot{\mathcal{P}}_{P_0}$. In the sequel we will make enough assumptions to guarantee that the tangent set is a Hilbert space, as is common in the literature on semi/non-parametric efficiency bounds (e.g. (Newey 1990), (Severini and Tripathi 2001), (Van der Vaart 2000)).

Define an inner product in $\dot{\mathcal{P}}_{P_0}$ by

$$\langle h, g \rangle_{P_0} = 4 \int_X \left(h p_0^{1/2}/2 \right) \left(g p_0^{1/2}/2 \right) d\mu = \int_X h g dP_0 = \mathrm{E}_{P_0}(hg),$$

i.e. $\langle h, g \rangle_{P_0}$ is the covariance between h and g. The covariance matrix of g is the analogue of the Fisher information matrix in a parametric problem.

The distance of P_t from P_0 (along the differentiable path $t \to P_t$) is defined to be $t \langle g, g \rangle_{P_0}^{1/2}$ where $\langle g, g \rangle_{P_0}^{1/2}$ is the norm induced by the inner product on $\dot{\mathcal{P}}_{P_0}$. So, for example, a ball of radius ε and centre at P_0 is the set of all one dimensional probability measures passing through P_0 :

$$B_{\mathcal{P}}(P_0, \varepsilon) = \left\{ P_t \in \mathcal{P}_0 : g \in \overline{\operatorname{lin} \dot{\mathcal{P}}_{P_0}}, \langle g, g \rangle_{P_0}^{1/2} < \varepsilon \right\}. \tag{11}$$

where $\overline{\operatorname{lin} \dot{\mathcal{P}}_{P_0}}$ is the closure of the linear span of the tangent set $\dot{\mathcal{P}}_{P_0}$. The reason for the use of $\overline{\operatorname{lin} \dot{\mathcal{P}}_{P_0}}$ rather than $\dot{\mathcal{P}}_{P_0}$, is that $\dot{\mathcal{P}}_{P_0}$ is a subset of the set of square integrable functions but does not span the whole space, and may not even be a

subspace. Note that t does not appear in (11) because $\overline{\operatorname{lin}\dot{\mathcal{P}}_{P_0}}$ is a linear space, so that $t\langle g,g\rangle_{P_0}^{1/2}=\langle g',g'\rangle_{P_0}^{1/2}$ for $g'=tg$.

Consider a map $\eta:\mathcal{P}\to\mathbb{R}^m$ defining the parameter of interest as before, and assume that η is differentiable at P_0 relative to the tangent set $\dot{\mathcal{P}}_{P_0}$. That is, there exists a continuous linear map $\dot{\eta}:L^2(\mathcal{X})\to\mathbb{R}^m$ such that, for every $g\in\dot{\mathcal{P}}_{P_0}$ and submodel $t\to P_t$ with score function g, one has

$$\frac{\eta(P_t)-\eta(P_0)}{t}=\dot{\eta}_{P_0}g+o(t) \tag{12}$$

as $t\to 0$. Note that this requires the derivative of $t\to\eta(P_t)$ to exist, and also to have the special form of equation (12). Moreover, since $\dot{\eta}$ is continuous, it follows from the Riesz-Fréchet Theorem (e.g. (Severini and Tripathi 2001) Theorem A.1), applied to each component of $\dot{\eta}_{P_0}g$ separately, that one can write the derivative of $t\to\eta(P_t)$ as

$$\dot{\eta}_{P_0}g=\int_X\tilde{\eta}_{P_0}gdP_0=\left\langle\tilde{\eta}_{P_0},g\right\rangle_{P_0}. \tag{13}$$

The function $\tilde{\eta}_{P_0}$ is uniquely defined in $\overline{\operatorname{lin}\dot{\mathcal{P}}_{P_0}}$ (but is not unique in $\dot{\mathcal{P}}_{P_0}$), and is called the *efficient influence function* ((Van der Vaart 2000) p. 363). Note that here we have used $\dot{\eta}_{P_0}$ instead of $\dot{\eta}(P_0)$, which would be closer to our notation for the parametric case, to keep the notation consistent with that currently used in the literature on non/semi-parametric efficiency bounds.

We may define, in analogy with the parametric case considered in Definition 2, the following measure:

Definition 3 *For fixed $P_0\in\mathcal{P}$ and $\eta:\mathcal{P}\to\mathbb{R}^m$, and with $B_{\mathcal{P}}(P_0,\varepsilon)$ defined as in (11), define*

$$\widetilde{der}(\eta,P_0)=\sup_{c'c=1}\lim_{\varepsilon\to 0}\sup_{P\in B_{\mathcal{P}}(P_0,\varepsilon)}\frac{|c'(\eta(P)-\eta(P_0))|}{\varepsilon}. \tag{14}$$

Remark 7 *The definition here differs from the earlier Definition 2 in that we have introduced the vector c, and the corresponding operation $\sup_{c'c=1}$, into the definition of $\widetilde{der}(\eta,P_0)$. This is necessary in order to be able to use the Riesz-Fréchet Theorem in Theorem 8.*

Theorem 7 *For any estimator $\hat{\eta}$ of $\eta(P_0)$*

$$\lim_{\varepsilon\to 0}\sup_{P\in B_{\mathcal{P}}(P_0,\varepsilon)}\mathrm{E}_P\left(\left[\frac{|c'(\hat{\eta}-\eta(P))|}{\varepsilon}\right]^2\right)\geq\frac{1}{4}\left[\widetilde{\mathrm{der}}(\eta,P_0)\right]^2$$

where E_P denotes the expectation taken with respect to the probability measure P.

Results similar to Theorems 3, 5 and 6 follow:

Theorem 8

$$\widetilde{der}\left(\eta, P_0\right) = \lambda_M^{1/2}\left(\mathrm{E}_{P_0}\left[\tilde{\eta}_{P_0}\tilde{\eta}'_{P_0}\right]\right) \tag{15}$$

where $\mathrm{E}_{P_0}\left[\tilde{\eta}_{P_0}\tilde{\eta}'_{P_0}\right]$ *is the covariance matrix of the efficient influence function* $\tilde{\eta}_{P_0}$.

Theorem 9

$$\widetilde{der}\left(\eta, P_0\right) = \sup_{c'c=1}\lim_{\varepsilon\to 0}\inf\left\{\phi : c'\eta\left(\mathcal{B}_{\mathcal{P}}\left(P_0, \varepsilon\right)\right) \subset B'_{\mathbb{R}}\left(c'\eta\left(p_0\right), \phi t\right), t < \varepsilon\right\}$$

where $B'_{\mathbb{R}}\left(c'\eta\left(P_0\right), \phi t\right)$ *is an interval of the real line,*

$$B'_{\mathbb{R}}\left(c'\eta\left(p_0\right), \phi t\right) = \left\{b \in \mathbb{R} : \left|b - c'\eta\left(P_0\right)\right| < \phi t\right\}.$$

Theorem 10 *Let* $\left\|A\right\|_2$ *denote the spectral norm of the square matrix* A. *Then*

$$\min_{A:|A|=0}\left\|\left(\mathrm{E}_{P_0}\left[\tilde{\eta}_{P_0}\tilde{\eta}'_{P_0}\right]\right)^{-1} - A\right\|_2 = \left(\widetilde{\mathrm{der}}\left(\eta, P_0\right)\right)^{-1}.$$

An important special case in the class of models considered above is a semi-parametric model for which each member of $\mathcal{P}$ can be written as $P_{\eta,\phi}$ where $\eta \in \mathbb{R}^q$ and ϕ belongs to an arbitrary set H (a subset of a Hilbert space). The map of interest is $\eta : \mathcal{P} \to \mathbb{R}^m$ defined by $\eta\left(P\right) = \eta$. In this case the covariance matrix of the efficient influence function (and thus $\widetilde{der}\left(\eta, P_0\right)$) takes on a particular form.

Consider a one-dimensional parameterization of the form

$$t \to P_{\eta+ta,\phi_t}$$

where $t \to \phi_t$ is in H. In this case one shows ((Van der Vaart 2000) Section 25.4) that

$$\left.\frac{\partial \ln dP_{\eta+ta,\phi_t}}{\partial t}\right|_{t=0} = a'\dot{l}_{\eta,\phi_0} + g \tag{16}$$

where $\dot{l}_{\eta,\phi_0}$ is the score function for η in the model for which ϕ is fixed at ϕ_0. The function g is the score function for ϕ if η is fixed. The infinite-dimensional set over which g runs is called the tangent set for ϕ and is denoted by ${}_{\phi}\dot{\mathcal{P}}_{P_{\eta,\phi_0}}$.

The efficient influence function $\tilde{\eta}_{P_0}$ must be orthogonal to the tangent set for the nuisance parameters, ${}_{\phi}\dot{\mathcal{P}}_{P_{\eta,\phi_0}}$, so one can let Π_{η,ϕ_0} be the the orthogonal projection onto the closure of the linear span of ${}_{\phi}\dot{\mathcal{P}}_{P_{\eta,\phi_0}}$, and define

$$\tilde{l}_{\eta,\phi_0} = \dot{l}_{\eta,\phi_0} - \Pi_{\eta,\phi_0}\dot{l}_{\eta,\phi_0}. \tag{17}$$

The quantity $\tilde{l}_{\eta,\phi_0}$ is called the *efficient score function for* η, and its covariance matrix

$$\Sigma_{\eta\eta} = E_{P_{\eta,\phi_0}}\left[\tilde{l}_{\eta,\phi_0}\tilde{l}'_{\eta,\phi_0}\right] \tag{18}$$

is called the *efficient information matrix*. Lemma 25.25 in (Van der Vaart 2000) says that if $\Sigma_{\eta\eta}$ is nonsingular, then η has *efficient influence function*

$$\tilde{\eta}_{P_0} = \Sigma_{\eta\eta}^{-1}\tilde{l}_{\eta,\phi_0}. \tag{19}$$

Thus, for this case we have the following corollary to Theorem 8:

Corollary 3 *For* $\eta(P)$ *defined as above,* $\widetilde{der}(\eta, P_0) = \lambda_M^{1/2}\left(\Sigma_{\eta\eta}^{-1}\right)$, *where* $\Sigma_{\eta\eta}$ *is the covariance matrix of the efficient score function* $\tilde{l}_{\eta,\phi_0}$ *for* η.

A few remarks are in order. Firstly, the results given in this section generalize those for parametric models given earlier. The structure is unchanged, and the key factor is the covariance matrix of the efficient influence function. Secondly, the analysis of this section is limited to the i.i.d. case, but it may easily be extended to the non i.i.d. case by considering P as the probability of the sample rather than that of a single observation. All the results would go through, but the terminology and the available theory to calculate "efficient influence functions" does not seem to have been developed.

Finally, we note that $\widetilde{der}(\eta, P_0)$ for non/semi-parametric models cannot be smaller than the value of $der(\eta, p_0)$ for any parametric model that it embeds. That is, a non/semi-parametric inference problem must be at least as ill-conditioned as any parametric version of itself. Hence, the results for parametric models given in Section 2 can identify classes of ill-conditioned, or poorly-conditioned, nonparametric problems as well.

5 Examples

5.1 Structural equation models and weak instruments

Consider a single structural equation

$$y_1 = Y_2\beta + Z_1\gamma + u, \tag{20}$$

where y_1 and Y_2 are respectively a $T \times 1$ vector and a $T \times n$ matrix of endogenous variables, Z_1 is a $T \times k_1$ matrix of exogenous variables, and β and γ are, respectively, $n \times 1$ and $k_1 \times 1$ vectors of parameters. The reduced form - the process generating the data - corresponding to (20) is:

$$(y_1, Y_2) = Z_1(\phi_1, \Phi_2) + Z_2(\pi_1, \Pi_2) + (v_1, V_2), \tag{21}$$

where Z_2 is a $T \times k_2$ matrix of exogenous variables not included in the structural equation, (ϕ_1, Φ_2) and (π_1, Π_2) are matrices of parameters of dimension $k_1 \times (n+1)$, $k_2 \times (n+1)$ respectively. We assume throughout that $k_2 \geq n$. The rows of $V = (v_1, V_2)$ are assumed to be independent normal vectors with mean zero and common $(n+1) \times (n+1)$ covariance matrix

$$\Omega = \begin{pmatrix} \omega_{11} & \omega_{21}' \\ \omega_{21} & \Omega_{22} \end{pmatrix},$$

where ω_{11}, ω_{21} and Ω_{22} are respectively (1×1), $(n \times 1)$ and $(n \times n)$ matrices of parameters (i.e., $V \sim N(0, I_T \otimes \Omega)$). The structural equation (20) is embedded in the reduced form (21) through the relations $\gamma = \phi_1 - \Phi_2\beta, u = v_1 - V_2\beta$, and

$$\pi_1 - \Pi_2\beta = 0. \tag{22}$$

This last equation (implicitly) defines β (one of the interest parameters) in terms of (π_1, Π_2), and the equation $\gamma = \phi_1 - \Phi_2\beta$ then defines γ (the other) in terms of β and (ϕ_1, Φ_2). We have:

Theorem 11 *Let* $\omega^2 = \omega_{11} - \omega_{21}'\Omega_{22}^{-1}\omega_{21}$, $\tilde{\beta} = \Omega_{22}^{\frac{1}{2}}\left(\beta - \Omega_{22}^{-1}\omega_{21}\right)/\omega$, *and*

$$\sigma^2 = \mathrm{var}(u_t) = \omega^2\left(1 + \tilde{\beta}'\tilde{\beta}\right).$$

Then

$$der(\beta, \gamma, p_0) = \sigma\lambda_m^{-1/2}\begin{pmatrix} \Pi_2'Z_2'Z_2\Pi_2 & \Pi_2'Z_2'Z_1 \\ Z_1'Z_2\Pi_2 & Z_1'Z_1 \end{pmatrix},$$

and

$$der(\beta, p_0) = \sigma\lambda_m^{-1/2}\left(\Pi_2'Z_2'M_{Z_1}Z_2\Pi_2\right).$$

If Π_2 *can have reduced rank the problem of inference about* β *and/or* γ *is ill-conditioned since* $\sup_{p\in\mathcal{P}}\{der(\eta, p)\} = +\infty$. *Moreover, in this case the problem is also ill-posed and no uniformly consistent estimator for* β *and/or* γ *exists.*

The density p_0 here is that induced by the reduced form (21) with fixed parameters. One may note the similarity between $der(\beta, p_0)$ and the concentration parameter considered by (Stock, Wright, and Yogo 2002). The difference is that $der(\beta, p_0)$ depends on $\sigma^2 = \mathrm{var}(u_t)$ (rather than on the covariance matrix of the rows of V), and therefore on the degree of endogeneity. In fact, if we let ρ^2 denote the multiple correlation coefficient between u_t and V_{2t} - an obvious measure of the degree of endogeneity - it easy to verify that $\rho^2 = \tilde{\beta}'\tilde{\beta}/\left(1 + \tilde{\beta}'\tilde{\beta}\right)$, so that $\sigma^2 = \omega^2/(1-\rho^2)$, and we have the important result that:

Proposition 2 *Since*

$$der\left(\beta, p_0\right) = \omega[\left(1-\rho^2\right)\lambda_m\left(\Pi_2' Z_2' M_{Z_1} Z_2 \Pi_2\right)]^{-\frac{1}{2}},$$

inferential problems arise when either Π_2 *is close to a rank-deficient matrix, or when* ρ^2 *is close to one.*

If $T^{-1}Z'Z \xrightarrow{P} Q$, and Π_2 is fixed and of full rank n, then it is easy to see that $der\left(\beta, p_0\right) \to 0$ as $T \to \infty$. If the instruments are weak in the sense of (Staiger and Stock 1997), i.e. $\Pi_2 = T^{-1/2}\bar{\Pi}$ for a fixed $\bar{\Pi}$, then $der\left(\beta, p_0\right)$ converges to a constant that depends on $\bar{\Pi}$, which is large when $\bar{\Pi}$ is close to zero.

5.2 Spectral density estimation

Consider an $ARMA(p,q)$ process

$$\begin{aligned} \phi\left(L\right) X_t &= \theta\left(L\right)\varepsilon_t \\ \varepsilon_t &\sim NID\left(0, \sigma^2\right) \end{aligned} \tag{23}$$

$t = 1, 2, ..., T$, where $\phi\left(L\right) = 1 - \sum_{i=1}^{p}\phi_i L^i$, $\theta\left(L\right) = 1 - \sum_{j=1}^{q}\theta_j L^j$ and L is the lag operator. It is assumed that the process is not redundant (i.e. $\phi\left(L\right)$ and $\theta\left(L\right)$ have no common zeros), and that $\phi\left(L\right)$ has all roots outside the unit circle. Then $\{X_t\}$ has spectral density

$$f\left(\lambda\right) = \frac{\sigma^2}{2\pi}\frac{\left|\theta\left(e^{-i\lambda}\right)\right|^2}{\left|\phi\left(e^{-i\lambda}\right)\right|^2}$$

for $-\pi \leq \lambda \leq \pi$. Any interest parameter η will be defined in terms of the underlying parameters $(\phi_i, \theta_j, \sigma^2)$. For example, one may be interested in the spectral density at frequency zero, $f\left(0\right)$, i.e.,

$$\eta = f\left(0\right) = \frac{\sigma^2}{2\pi}\frac{\left|1 - \sum_{i=1}^{p}\phi_i\right|^2}{\left|1 - \sum_{j=1}^{q}\theta_j\right|^2},$$

This is related to the problem of the estimation of the persistence of the process (e.g., in macroeconomics, see (Pötscher 2002)). In general, for any $\eta : \mathcal{P} \to \mathbb{R}^q$ defined on the set of $ARMA(p,q)$ processes $\mathcal{P}$ with values in $\mathbb{R}^q$, we have:

Theorem 12 *Let* $I_T\left(\phi, \theta, \sigma^2\right)$ *be the information matrix of an* $ARMA(p,q)$ *process, with* $p \geq 1$ *and* $q \geq 1$. *Then*

$$der\left(\eta, p_0\right) = \lambda_M^{1/2}\left(\left[\dot{\eta}\left(\phi, \theta, \sigma^2\right)\right] I_T^{-1}\left(\phi, \theta, \sigma^2\right)\left[\dot{\eta}\left(\phi, \theta, \sigma^2\right)\right]'\right)$$

may be unbounded ($\sup_{p\in\mathcal{P}}\{der(\eta,p)\}=+\infty$*) if the model can be arbitrarily close to one which is redundant. Moreover if* $\sup_{p\in\mathcal{P}}\{der(\eta,p)\}=+\infty$ *the problem is ill-posed and no uniformly consistent estimator of* η *exists.*

The information matrix $I_T(\phi,\theta,\sigma^2)$ is known in very special cases, but numerical algorithms are needed to evaluate it in general (e.g. (Klein, Mélard, and Zahaf 1998)).

Theorem 12 applies to any function $\eta:\mathcal{P}\to\mathbb{R}^q$ for which the partial derivatives exist. The behaviour $der(\eta,p_0)$ in a region close to the points where $I_T(\phi,\theta,\sigma^2)$ is rank deficient depends in a complicated way on $\dot{\eta}(\phi,\theta,\sigma^2)$. When I_T is rank-deficient it may be possible to reparameterize the model in terms of parameters ξ_1 and ξ_2 in such a way that the information matrix, in this new coordinates system, has the form

$$\begin{pmatrix} I_T(\xi_1,\xi_2) & 0 \\ 0 & 0 \end{pmatrix}$$

where $I_T(\xi_1,\xi_2)$ is a square matrix having as many rows as the components of ξ_1. If so, and if η depends only on ξ_1, then $der(\eta,p_0)$ will be finite, and uniformly consistent estimators of η will exist. On the other hand, whenever η depends on both ξ_1 and ξ_2, $der(\eta,p_0)=+\infty$ and the inferential problem will be ill-conditioned. Once again, since ξ_2 is not identified, there will be no uniformly consistent estimator for such a parameter η.

Since the spectral density at zero, $f(0)$, depends on all ϕ's and θ's, we may state:

Corollary 4 *The problem of estimating* $f(0)$ *is ill-conditioned as well as ill-posed when* $\mathcal{P}$ *is the family of* $ARMA(p,q)$ *with* $p\geq 1$ *and* $q\geq 1$.

This corresponds to Theorem 4.2 (d) of (Pötscher 2002).

5.3 Estimation of long memory parameter

Consider a Gaussian stationary process with zero mean and spectral density function

$$f(\lambda)=\lambda^{-2d}g(\lambda)$$

where $g(\lambda)$ is the spectral density for an $ARMA(p,q)$ process as in equation (23). (Li and McLeod 1986) have shown that the large-sample Fisher information matrix per observation for a fractional time series model in which $\sigma^2=1$ is

$$\begin{pmatrix} I_{p,q} & J \\ J' & c \end{pmatrix}$$

where $c = \pi^2/6$ is a constant, $I_{p,q}$ is the information matrix for the $ARMA(p,q)$ process with $\sigma^2 = 1$,

$$J = \left[\gamma_{Xd}(0), ..., \gamma_{Xd}(p-1), \gamma_{\varepsilon d}(0), ..., \gamma_{\varepsilon d}(q-1)\right]'$$

and

$$\begin{aligned} \gamma_{Xd}(s) &= \sum_{i=0}^{\infty} \frac{\phi_i'}{s+i+1} \\ \gamma_{\varepsilon d}(s) &= \sum_{i=0}^{\infty} \frac{\theta_i'}{s+i+1} \end{aligned}$$

where $\phi^{-1}(L) = \sum_{i=0}^{\infty} \phi_i' L^i$ and $\theta^{-1}(L) = \sum_{i=0}^{\infty} \theta_i' L^i$. We then have, for interest-parameter $\eta = d$:

Proposition 3 *Let the $ARMA(p,q)$ component be nonredundant. Then,*

$$der(d, p_0) = \left(c - J' I_{p,q}^{-1} J\right)^{-\frac{1}{2}}, \tag{24}$$

and this is finite and positive for all processes with finite p and q. So, for processes with finite $ARMA(p,q)$ component, the problem is well-conditioned, and uniformly consistent estimators of d exist.

Note that, even though $G_{11.2} = (c - J' I_{p,q}^{-1} J)$ is positive, it can become very small when p and/or q are large, because an $ARMA(p,q)$ process of sufficiently high order can approximate any stationary time series. For example, $der(d, p_0)$ can be as large as 7.38342 in an $AR(1)$ model, and can reach 50.3292 in an $ARMA(1,1)$ model. The following Theorem formalizes this

Theorem 13 *Suppose that the $ARMA(p,q)$ component is not redundant (i.e. it has nonsingular Fisher information matrix). Then, the estimation problem of estimating d is ill-conditioned (i.e. $\sup_{p \in \mathcal{P}} \{der(\eta, p)\} = +\infty$) if p or q tend to infinity. The estimation problem is also ill-posed and no uniformly consistent estimator of d exists if the model allows either the MA or AR component to be infinite.*

5.4 Data Reduction

We return now to the fully parametric setting of Section 2. Consider a statistic defined on the sample space $\mathcal{X}$, $y = t(x)$, say, taking values in $\mathbb{R}^p$, $p \leq n$. Let $\mathcal{P}$ be the family of underlying model densities, as in Section 2, and assume that the

matrix $\mathrm{D}_x\, t\,(x)$ is everywhere of full rank p. For example, y may be an estimator for some parameter of interest η, or a test statistic for testing an hypothesis about θ.

Any such function $t : \mathbb{R}^n \to \mathbb{R}^p$ induces a mapping from $\mathcal{P}$, the space of densities p_θ for x, to $\mathcal{Q}$, a space of densities f_ψ, say, for y, via the formula:

$$f_\psi = \int_{t^{-1}(y)} (\det\left[\left[D_x t\,(x)\right]\left[D_x t\,(x)\right]'\right])^{-\frac{1}{2}} p_\theta(dx), \tag{25}$$

where $t^{-1}\,(y) = \{x \in \mathcal{X} : t\,(x) = y\}$ denotes the manifold in $\mathcal{X}$ on which t is fixed at y (see, for instance, Proposition 8.1.2 in (Tjur 1980), or (Hillier and Armstrong 1999)), and $\psi = \psi(\theta)$ denotes a parameter indexing the densities in $\mathcal{Q}$. Denote this map by $t(p_\theta)$, the image of p_θ induced by t.

Hence, if $\psi = \psi(\theta) \in \Psi$ has dimension m ($m \leq q$), t also induces a map $\psi : \Omega \to \Psi \subset \mathbb{R}^m$, as in the following diagram:

$$\begin{array}{ccc} \mathcal{P} & \overset{f_\psi = t(p_\theta)}{\to} & \mathcal{Q} \\ \updownarrow & & \updownarrow \\ \mathbb{R}^q & \overset{\psi = \psi(\theta)}{\to} & \mathbb{R}^m \end{array}$$

In general, ψ will not be the identity map, and m will be less than q. That is, the manifold $\{f_\psi : \psi \in \Psi\}$ will be of lower dimension than that of $\{p_\theta : \theta \in \Omega\}$. When $m < q$ the statistic t may not preserve all of the information in x about θ, but may convey information about certain characteristics of θ. [For example, if $x \sim N(\mu, I_q)$, and $y = t(x) = x\prime x \sim \chi'^2(q, \mu'\mu)$, so y can convey information about the length of μ, but not its direction.] Thus, after reduction to t, we must, of necessity, focus on those functions of θ - represented here by ψ - for which y is informative.

As before, let $G\,(\theta) = -E_\theta\left[D^2_\theta p_\theta\right]$ be the information matrix for p_θ, and $H\,(\psi) = -E_\psi\left[D^2_\psi f_\psi\right]$ be the information matrix for f_ψ. It is of interest to measure the sensitivity of the induced density f_ψ to perturbations of the underlying density p_θ. To measure this, define:

$$der\,(f, p_0) = \lim_{\varepsilon \to 0} \sup_{\theta \in \mathcal{B}_\mathcal{P}(\theta_0, \varepsilon)} \frac{\left[\left(\psi\,(\theta) - \psi\,(\theta_0)\right)' H\,(\psi(\theta_0))\left(\psi\,(\theta) - \psi\,(\theta_0)\right)\right]^{\frac{1}{2}}}{\varepsilon}. \tag{26}$$

A result analogous to Theorem 3 holds.

Theorem 14 *If $\psi\,(\theta)$ is differentiable at $\theta = \theta_0$, then*

$$der\,(f, p_0) = \left[\lambda_M\left(\left[\dot{\psi}\,(\theta_0)\right] G\,(\theta_0)^{-1}\left[\dot{\psi}\,(\theta_0)'\right] H(\psi_0)\right)\right]^{\frac{1}{2}}. \tag{27}$$

Moreover, this does not depend on the parameterization of $\mathcal{P}$ or $\mathcal{Q}$.

The matrix $[\dot{\psi}(\theta_0)]G(\theta_0)^{-1}[\dot{\psi}(\theta_0)']$ is the covariance matrix of a consistent and efficient estimator for ψ (when θ_0 is the true value of θ). So, $der(f, p_0)$ compares the Fisher information for ψ_0 based on the statistic y with the asymptotic Fisher information of a consistent and efficient estimator for ψ_0 based on the full data x. That is:

Proposition 4 *$der(f, p_0)$ is a measure of the (asymptotic) relative efficiency of the statistic y - relative to the data x - in providing information about ψ_0.*

Since $[H(\psi_0) - ([\dot{\psi}(\theta_0)]G(\theta_0)^{-1}[\dot{\psi}(\theta_0)'])^{-1}]$ must be positive semi-definite (the information on ψ cannot increase when we reduce x to y), $der(f, p_0) \geq 1$ for all θ. The reduction to y is justified if ψ is a parameter of interest, and $\text{der}(f, p_0)$ is everywhere close to one. If y is a sufficient statistic (for θ), $der(f, p_0) = 1$ for all θ. Otherwise, if $der(f, p_0)$ can be close to 1 for some values of θ, y is *locally sufficient for* ψ in at least some regions of the parameter space.

6 Discussion and Conclusions

The possibility that a statistical problem may be ill-posed has been known for many years from the work of (Bahadur and Savage 1956), (Singh 1963), (LeCam and Schwartz 1960), and others. More recently it has come to be realised - particularly in the econometrics literature (e.g. (Dufour 1997), (Staiger and Stock 1997), and (Pötscher 2002)) - that problems that are close to being ill-posed may also present serious inferential difficulties. In particular, there are difficulties in reliably assessing the precision that can be attributed to inferential procedures in such cases. Thus, to assess how reliable statistical inference is, one needs to find ways to measure how far from ill-posed the problem under consideration actually is. That has been our purpose in this paper.

Although uniform continuity of the interest functional - the criterion discussed in most of the literature to date - is useful for identifying the extreme version of the problem, it does not seem to capture the "distance from ill-posed" quality of the problem. We have suggested a measure more akin to the derivative of the functional, and shown that it does indeed have the properties one wants. Our suggested measure has much in common with the condition number used in numerical analysis and optimisation theory to measure "distance from ill-posedness". In parametric models it is determined the Fisher information matrix, and in non-parametric models by the covariance matrix of the efficient influence function. The examples considered in Section 5 demonstrate the usefulness of our suggested measure.

Measures of the variance and precision of estimators are also of interest in any inference problem. We have shown that there is a direct link between the

properties of the three interest functionals {parameter, precision, variance}. This has been discussed briefly here, but deserves further investigation. Finally, we show that our suggested measure provides useful information on the sensitivity of the induced properties of a statistic, such as an estimator or test statistic, to perturbations of the underlying model density. This too deserves further study.

References

Amari, S.-I. (1985): *Differential-Geometrical Methods in Statistics*. Springer-Verlag, Heidelberg.

Bahadur, R. R., and L. J. Savage (1956): "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems," *Annals of Mathematical Statistics*, 27, 1115–1122.

Blyth, S. (1994): "Local Divergence and Association," *Biometrika*, 81, 579–584.

Bowden, R. (1973): "The Theory of Parametric Identification," *Econometrica*, 41, 1069–1074.

Castrigiano, D. P. L., and S. A. Hayes (1993): *Catastrophe Theory*. Addison-Wesley Publishing Company, New York.

Demmel, J. W. (1987): "On Condition Numbers and the Distance to the Nearest Ill-Posed Problem," *Numerische Mathematik*, 51, 251–289.

Dufour, J.-M. (1997): "Some Impossibility Theorems in Econometrics with Applications to Instrumental Variables and Dynamic Models," *Econometrica*, 65, 1365–1388.

Forchini, G., and G. Hillier (2003): "Conditional Inference for Possibly Unidentified Structural Equations," *Econometric Theory*, 19, 707–743.

Gibbs, A. L., and F. E. Su (2002): "On Choosing and Bounding Probability Metrics," *International Statistical Review*, 70, 419–436.

Gleser, L. J., and J. T. Hwang (1987): "The Nonexistence of 100(1-α)," *The Annals of Statistics*, 15, 1351–1362.

Hillier, G., and M. Armstrong (1999): "The Density of the Maximum Likelihood Estimator," *Econometrica*, 67, 1459–1470.

Kahan, W. (1966): "Numerical Linear Algebra," *Canadian Mathematical Bulletin*, 9, 757–801.

Klein, A., G. Mélard, and T. Zahaf (1998): "Computation of the Exact Information Matrix of Gaussian Dynamic Regression Time Series Models," *The Annals of Statistics*, 26, 1636–1650.

Koschat, M. (1987): "A Characterization of the Fieller Solution," *The Annals of Statistics*, 15, 462–468.

LeCam, L., and L. Schwartz (1960): "A Necessary and Sufficient Condition for the Existence of Consistent Estimates," *Annals of Mathematical Statistics*, 31, 140–150.

LeCam, L., and G. L. Yang (1990): *Asymptotics in Statistics*. Springer-Verlag, New York.

Li, W. K., and A. I. McLeod (1986): "Fractional Time Series Modelling," *Biometrika*, 73, 217–221.

McLeod, A. I. (1999): "Necessary and Sufficient Condition for Nonsingular Fisher Information Matrix in ARMA Models," *The American Statistician*, 53, 71–72.

Milnor, J. (1963): *Morse Theory*. Princeton University Press, Princeton.

Muirhead, R. J. (1982): *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, Inc., New York.

Newey, W. K. (1990): "Semiparametric Efficiency Bounds," *Jounal of Applied Econometrics*, 5, 99–135.

Pfanzagl, J. (1998): "The Nonexistence of Confidence Sets for Discontinuous Functionals," *Journal of Statistical Planning and Inference*, 75, 9–20.

Phillips, P. C. B. (1989): "Partially Identified Econometric Models," *Econometric Theory*, 5, 181–240.

Pötscher, B. M. (2002): "Lower Risk Bounds and Properties of Confidence Sets for Ill-Posed Estimation Problems with Applications to Spectral Density and Persistence Estimation, Unit Roots, and Estimation of Long Memory Parameters," *Econometrica*, 70, 1035–1065.

Rothenberg, T. J. (1971): "Identification in Parametric Models," *Econometrica*, 39, 577–591.

Sargan, J. D. (1983): "Identification and Lack of Identification," *Econometrica*, 51, 1605–1633.

Severini, T. A., and G. Tripathi (2001): "A Simplified Approach to Computing Efficiency Bounds in Semiparametric Models," *Journal of Econometrics*, 102, 23–66.

Singh, R. (1963): "Existence of Bounded Length Confidence Intervals," *Annals of Mathematical Statistics*, 34, 1474–1485.

Spivak, M. (1965): *Calculus on Manifolds. A Modern Approach to Classical Theorems of Advanced Calculus*. W. A. Benjamin, Inc, New York-Amsterdam.

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

Stock, J. H., J. H. Wright, and M. Yogo (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business and Economic Statistics*, 20, 518–529.

Tjur, T. (1980): *Probability Based on Radon Measures*. John Wiley & Sons, Inc., New York.

Van der Vaart, A. W. (2000): *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Zolezzi, T. (2002): "On the Distance Theorem in Quadratic Optimization," *Journal of Convex Analysis*, 9, 693–700.

A Appendix: Proofs

A.1 Proof of Lemma 1

Part (i) is a standard result (see (Blyth 1994)). To prove part (ii) regard $\delta(p_{\theta_0}, p_\theta)$ as a function of θ, and note that it has a critical point at $\theta = \theta_0$ (from (2) of the definition of divergence) where the matrix of second derivatives is nonsingular (from (3) of the definition of divergence). Then apply Morse's Lemma (e.g. (Milnor 1963) and (Castrigiano and Hayes 1993)).

A.2 Proof of Theorem 1

The argument follows the proof of Theorem 2.1 of (Pötscher 2002). We just need to note that (i) the sets $\mathcal{B}_{\mathbb{R}^m}(\theta_0, 1/i)$ for $i \in \mathbb{N}$ sufficiently large generate a neighbourhood basis as in Theorem 2.1 of (Pötscher 2002); (ii) $|\cdot|^2$ is a proper loss function in the sense of (Pötscher 2002); the map $p \to P_p = \int p d\mu$ is continous when the set of all probability

measures is endowed with the total variation distance (this follows from the fact that locally the divergence and the total variation distance induce the same topology). Thus Corollary 2.2 of (Pötscher 2002) applies.

A.3 Proof of Theorem 2

It follows from Theorem 1 that for sufficiently small $\varepsilon > 0$ one has

$$\sup_{\theta \in \mathcal{B}_{\mathbb{R}^{\shortparallel}}(\theta_0,\varepsilon)} \mathrm{E}_p\left(\left|\hat{\eta} - \eta(p)\right|^2\right) \geq 2^{-2}\left[\sup_{\theta \in \mathcal{B}_{\mathbb{R}^{\shortparallel}}(\theta_0,\varepsilon)} \left|\eta(p) - \eta(p_0)\right|\right]^2.$$

Thereom 2 follows by divinding by ε^2 both sides of the above inequality and taking the limit as ε goes to zero.

A.4 Proof of Theorem 3

Let

$$der_\varepsilon(\eta, p_0) = \sup_{\theta \in \mathcal{B}_{\mathbb{R}^{\shortparallel}}(\theta_0,\varepsilon)} \frac{\left|\eta(p) - \eta(p_0)\right|}{\varepsilon}$$

so that $der(\eta, p_0) = \lim_{\varepsilon \to 0} der_\varepsilon(\eta, p_0)$. Expand $\eta(\theta)$ as a Taylor series,

$$\eta(\theta) - \eta(\theta_0) = \dot{\eta}(\theta_0)[\theta - \theta_0] + O\left(\varepsilon^2\right)$$

in $\mathcal{B}_{\mathbb{R}^{\shortparallel}}(\theta_0, \varepsilon)$, and note that

$$[\eta(\theta) - \eta(\theta_0)]'[\eta(\theta) - \eta(\theta_0)] = [\theta - \theta_0]'\dot{\eta}(\theta_0)'\dot{\eta}(\theta_0)[\theta - \theta_0] + O\left(\varepsilon^3\right).$$

Thus one can write

$$der_\varepsilon(\eta, p_0) = \sup_{\theta \in \mathcal{B}_{\mathbb{R}^{\shortparallel}}(\theta_0,\varepsilon)} \sqrt{\frac{[\theta - \theta_0]'\dot{\eta}(\theta_0)'\dot{\eta}(\theta_0)[\theta - \theta_0]}{\varepsilon^2}} + O(\varepsilon).$$

Since $\dot{\eta}(\theta_0)'\dot{\eta}(\theta_0)$ is positive semidefinite, the supremum must occur at the boundary as a maximum, so

$$\begin{aligned} der_\varepsilon(\eta, p_0) &= \sup_{[\theta-\theta_0]'G(\theta)[\theta-\theta_0]=\varepsilon^2} \sqrt{\frac{[\theta - \theta_0]'\dot{\eta}(\theta)'\dot{\eta}(\theta)[\theta - \theta_0]}{\varepsilon^2}} + O(\varepsilon) \\ &= \sup_{v'v=1} \sqrt{v' G(\theta)^{-1/2}\dot{\eta}(\theta)'\dot{\eta}(\theta) G(\theta)^{-1/2} v} + O(\varepsilon) \\ &= \sqrt{\lambda_M\left[\dot{\eta}(\theta)'\dot{\eta}(\theta) G(\theta)^{-1}\right]} + O(\varepsilon). \end{aligned}$$

The first part of the theorem follows by noting that $\lambda_M\left[\dot{\eta}(\theta)'\dot{\eta}(\theta) G(\theta)^{-1}\right] = \lambda_M\left[\dot{\eta}(\theta) G(\theta)^{-1}\dot{\eta}(\theta)'\right]$, and by taking the limit at ε goes to zero.

To show invariance to reparameterizations of P, define $\theta = \phi(\tau)$ so that $\theta_0 = \phi(\tau_0)$ and note that

$$\begin{aligned} G(\theta_0) &\rightarrow \dot{\phi}(\tau_0)' G(\theta_0) \dot{\phi}(\tau_0) \\ \dot{\eta}(\theta) &\rightarrow D_\tau \eta(\phi(\tau))|_{\tau=\tau_0} = \dot{\eta}(\theta_0) \dot{\phi}(\tau_0) \end{aligned}$$

where $\dot{\phi}(\tau_0) = D_\tau \phi(\tau)|_{\tau=\tau_0}$ is a $q \times q$ non-singular matrix.

A.5 Proof of Corollary 1

This is a simple consequence of Theorem 3.

A.6 Proof of Theorem 4

(i) Choose any $0 < \varepsilon < 1$, then

$$\sup_{\theta \in \mathcal{B}_{\mathbb{R}^{\text{II}}}(\theta_0, \varepsilon)} \frac{|\eta(p) - \eta(p_0)|}{\varepsilon} \geq \sup_{\theta \in \mathcal{B}_{\mathbb{R}^{\text{II}}}(\theta_0, \varepsilon)} |\eta(p) - \eta(p_0)|$$

The first part is proved by taking the limit as $\varepsilon \rightarrow 0$.

(ii) If $der(\eta, p_0) \leq M < +\infty$, then for $\varepsilon > 0$ sufficiently small,

$$\sup_{\theta \in \mathcal{B}_{\mathbb{R}^{\text{II}}}(\theta_0, \varepsilon)} \frac{|\eta(p) - \eta(p_0)|}{\varepsilon} \leq M.$$

Multiplying right and left hand sides by ε one has

$$\sup_{\theta \in \mathcal{B}_{\mathbb{R}^{\text{II}}}(\theta_0, \varepsilon)} |\eta(p) - \eta(p_0)| \leq \varepsilon M$$

so the result follows by taking the limit as ε goes to zero.

Part (iii) If $osc(\eta, p_0) \geq \delta > 0$, then for $\varepsilon > 0$ sufficiently small one has $\sup_{\theta \in \mathcal{B}_{\mathbb{R}^{\text{II}}}(\theta_0, \varepsilon)} |\eta(p) - \eta(p_0)| \geq \delta$. Dividing both sides by ε and taking the limit as ε goes to zero we obtain $der(\eta, p_0) = +\infty$.

A.7 Proof of Corollary 2

It follows from Remark 6 that η is continuous in $\mathbb{R}^q$. Let $\mathcal{F}$ denote the family of compact subsets of $\mathbb{R}^q$. (LeCam and Schwartz 1960) p. 148 show that η is $\mathcal{F}$-consistently estimable if and only if η is uniformly continuous in $\mathbb{R}^q$. This implies that η can be uniformly consistently estimated in any compact subset of $\mathbb{R}^q$.

A.8 Proof of Theorem 5

This result follows directly from the definition of $der(\eta, p_0)$.

A.9 Proof of Theorem 6

This result follows from the Gastinel Theorem ((Kahan 1966), p 775) and Theorem A5.3 of (Muirhead 1982).

A.10 Proof of Theorem 7

As in the proof of Theorems 1 and 2 one can show that

$$\sup_{P\in B_{\mathcal{P}}(P_0,\varepsilon)} \mathrm{E}_P\left(\left[\frac{|\hat{\eta}-\eta(p)|}{\varepsilon}\right]^2\right) \geq 2^{-2}\left(\sup_{P\in B_{\mathcal{P}}(P_0,\varepsilon)} \frac{|\eta(P)-\eta(P_0)|}{\varepsilon}\right)^2. \tag{28}$$

In order to check this (i) the sets $B_{\mathcal{P}}(P_0, 1/i)$ for $i \in \mathbb{N}$ sufficiently large generate a neighbourhood basis as in Theorem 2.1 of (Pötscher 2002); (ii) $|\cdot|^2$ is a proper loss function in the sense of (Pötscher 2002); (iii) follows from the fact that we have restricted our attention to the set of one-dimensional probability measures which are differentiable in quadratic mean at P_0 (so that a sequence of densities p_t converges to p_0 as $t \to 0$ along a differentiable path in the Hellinger distance, which induces the same topology as the total variation distance).

Note that

$$\begin{aligned} |\eta(P)-\eta(P_0)| &= \sqrt{[\eta(P)-\eta(P_0)]'[\eta(P)-\eta(P_0)]} \\ &\geq \sqrt{[\eta(P)-\eta(P_0)]' cc' [\eta(P)-\eta(P_0)]} \end{aligned}$$

for any $m \times 1$ vector c such that $c'c = 1$. Thus,

$$|\eta(P)-\eta(P_0)| \geq \max_{c'c=1} \left|c'\eta(P) - c'\eta(P_0)\right|. \tag{29}$$

The result follows from the inequalities (28) and (29).

A.11 Proof of Theorem 8

For any $P \in B_{\mathcal{P}}(P_0, \varepsilon)$ we have

$$c'(\eta(P)-\eta(P_0)) = tc'\dot{\eta}_{P_0} g + O\left(\varepsilon^2\right)$$

and

$$\left[c'(\eta(P)-\eta(P_0))\right]^2 = t^2\left(c'\dot{\eta}_{P_0} g\right)^2 + O\left(\varepsilon^3\right).$$

So

$$\begin{aligned}
\widetilde{der}\left(\eta, P_0\right) &= \sup_{c'c=1} \lim_{\varepsilon \to 0} \sup_{P \in B_{\mathcal{P}}(P_0, \varepsilon)} \sqrt{\frac{\left(c' \dot{\eta}_{P_0} g\right)^2}{\varepsilon^2} + O\left(\varepsilon\right)} \\
&= \sup_{c'c=1} \lim_{\varepsilon \to 0} \sup_{g \in \overline{\text{lin}\, \dot{\mathcal{P}}_{P_0}} : \langle g, g \rangle_{P_0}^{1/2} \leq \varepsilon} \sqrt{\frac{\left(c' \dot{\eta}_{P_0} g\right)^2}{\varepsilon^2}} + O\left(\varepsilon\right) \\
&= \sup_{c'c=1} \lim_{\varepsilon \to 0} \sup_{g \in \overline{\text{lin}\, \dot{\mathcal{P}}_{P_0}} : \langle g, g \rangle_{P_0}^{1/2} \leq 1} \sqrt{\left(c' \dot{\eta}_{P_0} g\right)^2} + O\left(\varepsilon\right)
\end{aligned}$$

The last line follows from the linearity of $\langle g, g \rangle_{P_0}$ in both arguments, and the fact that $\dot{\mathcal{P}}_{P_0}$ is a linear space. Then,

$$\widetilde{der}\left(\eta, P_0\right) = \sup_{c'c=1} \sup_{g \in \text{lin}\, \dot{\mathcal{P}}_{P_0} : \langle g, g \rangle_{P_0}^{1/2} \leq 1} \sqrt{\left(c' \dot{\eta}_{P_0} g\right)^2}.$$

The Riesz-Fréchet Theorem (e.g. (Severini and Tripathi 2001) Theorem A.1) allows us to write

$$c' \dot{\eta}_{P_0} g = \int_X c' \tilde{\eta}_{P_0} g dP_0$$

with $\tilde{\eta}_{P_0}$ in $\overline{\text{lin}\, \dot{\mathcal{P}}_{P_0}}$, and the Cauchy-Schwarz inequality implies that

$$\left(c' \dot{\eta}_{P_0} g\right)^2 \leq \left[\int_X \left[c' \tilde{\eta}_{P_0}\right]^2 dP_0\right] \left[\int_X g^2 dP_0\right].$$

Using these two results we can rewrite

$$\widetilde{der}\left(\eta, P_0\right) = \sup_{c'c=1} \sqrt{\int_X \left[c' \tilde{\eta}_{P_0}\right]^2 dP_0}.$$

Finally, note that

$$\int_X \left[c' \tilde{\eta}_{P_0}\right]^2 dP_0 = \text{E}_{P_0}\left(c' \tilde{\eta}_{P_0}\right)^2 = \text{E}_{P_0}\left[c' \tilde{\eta}_{P_0} \tilde{\eta}'_{P_0} c\right] = c' \left(\text{E}_{P_0}\left[\tilde{\eta}_{P_0} \tilde{\eta}'_{P_0}\right]\right) c$$

so that

$$\begin{aligned}
\widetilde{der}\left(\eta, P_0\right) &= \sup_{c'c=1} \sqrt{c' \left(\text{E}_{P_0}\left[\tilde{\eta}_{P_0} \tilde{\eta}'_{P_0}\right]\right) c} \\
&= \lambda_M^{1/2}\left(\text{E}_{P_0}\left[\tilde{\eta}_{P_0} \tilde{\eta}'_{P_0}\right]\right),
\end{aligned}$$

and the theorem is proved.

A.12 Proof of Theorem 9

The result follows directly from the definition of $\widetilde{\text{der}}\left(\eta, P_0\right)$.

A.13 Proof of Theorem 10

The proof is the same as that of Theorem 6 with the obvious change of notation.

A.14 Proof of Corollary 3

The result follows from Theorem 9 above and Lemma 25.25 of (Van der Vaart 2000).

A.15 Proof of Theorem 11

The partial information matrix for the structural equation model specified in the text is $\Pi_2' Z_2' M_{Z_1} Z_2 \Pi_2 / \sigma^2$ for β and

$$\sigma^{-2} \begin{pmatrix} \Pi_2' Z_2 Z_2 \Pi_2 & \Pi_2' Z_2' Z_1 \\ Z_1' Z_2 \Pi_2 & Z_1 Z_1 \end{pmatrix}$$

for (β, γ). The fact the problem is ill-conditioned follows easily. The fact that the problem is ill-posed follows from the fact that $der(\beta, p_0) = +\infty$ implies that Π_2 is rank deficient and thus β and (β, γ) are unidentified.

A.16 Proof of Theorem 12

The result follows from Theorem 1 of (McLeod 1999) according to which the information matrix of an ARMA(p, q) model is singular if and only if the model is redundant. Since redundancy implies lack of identification, the nonexistence of uniformly consistent estimators follows.

A.17 Proof of Corollary 4

This result follows from Theorem 12 and the fact that $\eta = f(0)$ has continuous partial derivatives. This implies that η is differentiable.

A.18 Proof of Proposition 3

The result follows from the discussion preceding the proposition.

A.19 Proof of Theorem 13

It will suffice to prove the theorem for $p = 0$ and q tending to infinity. In this case the model (23) has the form

$$X_t = \theta(L)\varepsilon_t, \;\; \varepsilon_t \sim NID\left(0, \sigma^2\right).$$

The asymptotic Fisher information matrix is $I(d, \theta) = E(A_t A_t')$ where $A_t = (w_t, u_{t-1}, u_{t-2}, ..., u_{t-q})$, $\theta(L) u_{t-1} = \varepsilon_{t-1}$ and $w_{t-1} = -\sum_{j=1}^{\infty} j^{-1} \varepsilon_{t-j}$

$= \left[-\sum_{j=0}^{\infty}(j+1)^{-1}L^j\right]\varepsilon_{t-1}$ (e.g. (Li and McLeod 1986) and (McLeod 1999)). Following (McLeod 1999), we observe that the matrix $I(d,\theta)$ is singular if and only if $l'I(d,\theta)\,l = 0$ for some vector $l = \left(\delta, \beta_0, ..., \beta_{q-1}\right)$. Moreover,

$$l'I(d,\theta)\,l = \text{var}\left(\delta w_{t-1} + \sum_{i=0}^{q-1}\beta_i u_{t-1-i}\right).$$

Since the information matrix corresponding to the MA component is nonsingular we have $\delta \neq 0$, and without loss of generality we can set δ equal to 1. Let $\theta^{-1}(L) = \sum_{i=0}^{\infty}\theta_i' L^i$, and write u_{t-1} as $u_{t-1} = \left(\sum_{i=0}^{\infty}\theta_i' L^i\right)\varepsilon_{t-1}$ and $\sum_{i=0}^{q-1}\beta_i u_{t-1-i} = \left(\sum_{i=0}^{q-1}\beta_i L^i\right)u_{t-1}$ so that

$$\begin{aligned} l'I(d,\theta)\,l &= \text{var}\left(\left[-\sum_{j=0}^{\infty}(j+1)^{-1}L^j\right]\varepsilon_{t-1} + \left(\sum_{i=0}^{q-1}\beta_i L^i\right)\left(\sum_{i=0}^{\infty}\theta_i' L^i\right)\varepsilon_{t-1}\right) \\ &= \text{var}\left(\left[\sum_{j=0}^{\infty}\left[-(j+1)^{-1} + \sum_{k=0}^{\min\{q-1,j\}}\beta_k\theta_{j-k}'\right]L^j\right]\varepsilon_{t-1}\right). \end{aligned}$$

One may note that $l'I(d,\theta)\,l$ vanishes only if all the coefficients of

$$\sum_{j=0}^{\infty}\left[-(j+1)^{-1} + \sum_{k=0}^{\min\{q-1,j\}}\beta_k\theta_{j-k}'\right]L^j$$

are zero. We can certainly find q values of $\beta_0, ..., \beta_{q-1}$ to make the first q coefficients of the series above vanish. However, to make all of them vanish we need to allow q to be infinity. The theorem follows from the fact that the determinant of

$$I(d,\theta) = \begin{pmatrix} I_q & J \\ J' & \pi^2/6 \end{pmatrix},$$

where J is as defined in Section 5.3, can be written as

$$|I(d,\theta)| = |I_q|\left|\pi^2/6 - J'I_q^{-1}J\right|.$$

As q goes to infinity the left-hand side goes to zero, but $|I_q| \neq 0$ because the MA component cannot be redundant. Thus $\left|\pi^2/6 - J'I_q^{-1}J\right| \to 0$ as $q \to \infty$. Thus the problem is ill-conditioned.

It follows from (McLeod 1999) Theorem 2 that the information matrix of the fractional ARIMA model is nonsingular if and only if the model is non redundant. Since redundancy implies lack of identification, it follows that no uniformly consistent estimator of d exists.

A.20 Proof of Theorem 14

As in the proof of Theorem 3, we let

$$der_{\varepsilon}\left(f,p_0\right)=\sup_{(\theta-\theta_0)'G(\theta)(\theta-\theta_0)\leq\varepsilon^2}\sqrt{\frac{\left(\psi\left(\theta\right)-\psi\left(\theta_0\right)\right)'H\left(\psi\left(\theta_0\right)\right)\left(\psi\left(\theta\right)-\psi\left(\theta_0\right)\right)}{\varepsilon^2}}.$$

Since $\theta\in\mathcal{B}_{\mathcal{P}}\left(\theta_0,\varepsilon\right)$, and ψ is differentiable

$$\begin{aligned}\psi\left(\theta\right)-\psi\left(\theta_0\right) &= \left[\mathrm{D}_{\theta}\,\psi\left(\theta\right)\right]_{\theta=\theta_0}\left(\theta-\theta_0\right)+O\left(\varepsilon^2\right)\\ &= \dot{\psi}\left(\theta_0\right)\left(\theta-\theta_0\right)+O\left(\varepsilon^2\right),\end{aligned}$$

so that

$$\begin{aligned}der_{\varepsilon}\left(f,p_0\right) &= \sup_{(\theta-\theta_0)'G(\theta_0)(\theta-\theta_0)\leq\varepsilon^2}\sqrt{\frac{\left(\theta-\theta_0\right)'\dot{\psi}\left(\theta_0\right)'H\left(\psi_0\right)\dot{\psi}\left(\theta_0\right)\left(\theta-\theta_0\right)}{\varepsilon^2}}+O\left(\varepsilon\right)\\ &= \max_{\tau'\tau=1}\sqrt{\tau' G\left(\theta_0\right)^{-1/2}\dot{\psi}\left(\theta_0\right)'H\left(\psi\right)\dot{\psi}\left(\theta_0\right)G\left(\theta_0\right)^{-1/2}\tau}+O\left(\varepsilon\right)\end{aligned}$$

where $\psi_0=\psi\left(\theta_0\right)$ and we again use the fact that the supremum occurs on the boundary. Thus,

$$\begin{aligned}der_{\varepsilon}\left(f,p_0\right) &= \left[\lambda_M\left(G\left(\theta_0\right)^{-1/2}\dot{\psi}\left(\theta_0\right)'H\left(\psi_0\right)\dot{\psi}\left(\theta_0\right)G\left(\theta_0\right)^{-1/2}\right)\right]^{1/2}+O\left(\varepsilon\right)\\ &= \left[\lambda_M\left(G\left(\theta_0\right)^{-1}\dot{\psi}\left(\theta_0\right)'H\left(\psi_0\right)\dot{\psi}\left(\theta_0\right)\right)\right]^{1/2}+O\left(\varepsilon\right)\end{aligned}$$

The result follows by taking the limit as ε goes to zero. Invariance to reparameterizations follows from arguments similar to those of Theorem 3.