Ill-Posed Problems and Instruments' Weakness*

G. Forchini †

University of York

G. H. Hillier‡

University of Southampton

May 2004
PRELIMINARY AND (STILL) INCOMPLETE

Abstract

Pötscher (2002, *Econometrica*, 70, 1035-1065) has pointed out that several estimation problems in econometrics are ill-posed. This paper further studies the nature of ill-posed problems in parametric models. The starting point is that both parameters of interest and many estimators of the parameters of interest in parametric models are maps from the manifold of density functions (parameterized by a m-dimensional parameter) to the q-dimensional Euclidean space ($q \leq m$). We are able to measure how ill-posed a problem is at each point on the manifold of probability density functions by focusing on the properties that these maps have to transmit small perturbations. We also argue that these measures should be reported by practitioners as they indicate how reliable their inference is. Applications of the measures of transmission of perturbations to some interesting situations are given. The case of structural equations models is particularly important since the indices of transmission of perturbations proposed form coherent measures of instruments' weakness.

*This is a substantially revised and expanded version of "Assessment for Sensitivity for Parameterizations and Estimators" which was presented at the Brussels - York meeting in statistics (21 June 2002) and the ESRC Econometric Study Group Annual Conference in Bristol (10-12 July 2003).

†Address for correspondence: Giovanni Forchini, Department of Economics and Related Studies, University of York, York YO1 5DD, UK. E-mail: gf7@york.ac.uk.

‡Address for correspondence: Grant Hillier, Economics Division, School of Social Sciences, University of Southampton, Southampton SO17 1BJ, U.K., e-mail: G.H.Hillier@soton.ac.uk

1 Introduction

The problems associated to total lack of identification of structural equations models have been well know for several years (e.g. Sims (1980) and Sargan (1983)). Phillips (1983), Phillips (1989) and Hillier (1985) show that the standard estimators are completely uninformative about the structural parameters if these parameters are totally unidentified. The standard asymptotic theory fails even in intermediate situations such as the partially identified models of Phillips (1989) and Choi and Phillips (1992) and in the weakly identified models of Staiger and Stock (1997). The surprising result, is that the standard asymptotic theory breaks down, and confidence sets can be unbounded with positive probability even if the model is identified but can be arbitrarily close to being unidentified (Dufour (1997), Staiger and Stock (1997)). This discovery has had a huge impact on applied econometric work because evidence of identification of the structural parameters is often very weak (Staiger and Stock (1997)).

Although relatively new in econometrics, the history of these kind of problems goes back a long way in the statistics literature. Let $\mathcal{P}$ be a family of probability measures on a common measurable space $(X, \mathcal{A})$, and $\eta : \mathcal{P} \to \mathbb{R}^q$ be a map. Let $\delta(P_1, P_2) = \sup_{A \in \mathcal{A}} \{|P_1(A) - P_2(A)|\}$ be the total variation distance between the probability measures $P_1, P_2 \in \mathcal{P}$. Bahadur and Savage (1956) and Singh (1963) show that if there is a bounded length confidence interval for η based on a sample of fixed size, then the map η is uniformly continuous on $(\mathcal{P}, \delta)$ (for more recent results see Koschat (1987), Gleser and Hwang (1987), Dufour (1997), Pfanzagl (1998)). LeCam and Schwartz (1960) notice that if the map η is discontinuous at $P_0 \in \mathcal{P}$ then there can be no uniformly consistent estimator of η. Uniform consistency excludes estimators with disturbing local behaviour (such as Hodges' superefficient estimator) having unbounded local asymptotic risk (see also Pötscher (2002)).

A simple example helps to understand the nature of the problems which may arise even in a very basic context. Consider a sample of n independent observations $(x_1, x_2)_i$ from a bivariate normal distribution with mean vector (μ_1, μ_2) and identity covariance matrix I_2. The sample means $(\bar{x}_1, \bar{x}_2)$ contain all the sample information

about the parameter (μ_1, μ_2). Inference about this parameter does not involve any problem whatsoever. However, if the mean is reparameterized as $\mu_1 = \psi\mu_2$, $\mu_2 = \mu_2$, then it is well known that inference about ψ is problematic, although it is difficult to state clearly what the problem is. In very loose terms, there are two interrelated sources for the difficulty of inference about ψ. The first one concerns the parameterization employed because ψ is not well defined for all values of μ_2. The second one pertains to the estimator: the natural estimator of ψ, $\hat{\psi} = \bar{x}_1/\bar{x}_2$, is imprecise when $\bar{x}_2$ is close to zero.

In this paper we connect the statistics and the econometrics literature and try to make clear what goes wrong in situations like the one described above. The starting point is the observation that a parameterization and many estimators define maps from the manifold of probability density functions (PDFs) to the parameter space (usually a submanifold of the q-dimensional Euclidean space). The properties of this maps are fundamental as the work of Bahadur and Savage (1956), LeCam and Schwartz (1960), Singh (1963), and Pfanzagl (1998) shows. We study how these maps transmit perturbations. Intuitively, problems arise when either the perturbations are not transmitted at all or when they are enormously amplified. In the latter case, the manifold of PDFs is *too rich* and this gives rise to the problems emphasised by the econometric and statistic literature referred to above. The lack of transmission of perturbations may reflect the fact that very restrictive conditions have been imposed on the manifold of PDFs. These may be relaxed without compromising statistical inference.

An important by-product of this analysis is the derivation of measures of weakness of instruments for structural equations models, which can be easily calculated and interpreted. We regard weak instruments as a very sensitive relationship between the manifold of PDFs and the parameters of interest, and embed the analysis in a general set-up which includes parametric ill-posed problems. This provide both a definition and a measure for weak instruments, which complement those of Shea (1997), Godfrey (1999), Stock, Wright, and Yogo (2002), Hahn and Hausman (2002a), Hahn and Hausman (2002b), Poskitt and Skeels (2002) and Stock and Yogo (2003).

By observing that the weak instruments problem is linked to the fact that the manifold of PDFs is *too rich*, our analysis may suggest that a solution may consist in restricting the manifold of PDFs in suitable ways. It is clear from the analysis of Dufour (1997) that imposing identification does not restrict the manifold of PDFs enough, and it is difficult to think of agreeable ways of further restricting such manifold. The model could be changed as done for example by Chamberlain and Imbens (2004) and Han and Phillips (2003), however, if we do not want to do so, we need to account for points in the manifold of PDFs, in a neighbourhood of which inference is difficult. This can be done by using the measures of weakness of the instruments as *post-data measures of precision* (Goutis and Casella (1995)). Forchini and Hillier (2003) argued for conditioning on an identification test statistic which measures the distance of the observed point in the manifold of PDFs from the point where identification does not hold. The measures suggested below may be certainly used in this way since they are less ad hoc than the statistic used in the previous paper. Alternatively, one could report estimates of a loss function as post-experimental measures of precision (see for instance Goutis and Casella (1995), Lindsay and Li (1997) and Sundberg (2003)). For structural equation models, however, instead of reporting an estimate of the loss we report an estimate of the "sharpness" of the loss function (see also Bowden (1973)).

An application the problem of estimating return to schooling using the data set of Angrist and Krueger (1991) is given. There is clear evidence that inference about return to schooling is weak and that changes on the manifold of PDFs are transformed into large variations of the parameter of interest (see Bound, Jaeger, and Baker (1995) for a related conclusion). The results also show how differently the TSLS and the OLS estimator are affected by identification.

The remaining part of the paper is organized as follows. Section 2 describe the set-up considered and gives bound on the transmission of perturbation from the data generation process to the parameter of interest. Section 3 deals with the case of singular information metric. Section 4 gives some examples of applications of the measured proposed. These include the classical linear regression model, the linear regression model with time-trend and autocorrelated errors, and some

variations of the Fieller-Creasy problem. Section 5 discusses structural equations models and considers an empirical application. The conclusions end the paper.

2 The model and the main results

Consider the manifold of PDFs, $P = \{p(x;\theta)\}$, parameterized by a q dimensional vector of parameters $\theta \in \Theta \subset \mathbb{R}^q$, and assume that P is a differential manifold (Amari (1985)). The manifold of PDFs is often described in econometrics as the set of all data-generation processes (DGPs). The variable x is an $n \times k$ matrix containing n observations on k variables.

This paper studies the map $\eta : P \to \Theta$ and develops measures for the "sensitivity of this map". With the expression "sensitivity of a map" we mean the property that a map has of transmitting small perturbations of $p \in P$ to $\eta(p) \in \Theta$. Intuitively, if the map $\eta(p)$ amplifies perturbations of p, then it is very difficult to discover where $\eta(p)$ is when p is only imprecisely located. However, if perturbations of p are not transmitted at all, then all the points in P are mapped to the same value $\eta(p)$, indicating that the manifold of PDFs has been considerably restricted by the researcher's assumptions. To capture these ideas we will measure the largest possible change of $\eta(p) \in \Theta$ which can be achieved by changing p slightly, and the smallest variation of p necessary to produce a fixed (small) change of $\eta(p)$. We will be more precise below.

Before starting our analysis we need to impose some structure on the problem at hand. In order to define a neighbourhood of $p(x;\theta_0)$ in P we use the notion of divergence. A *divergence* (Amari (1985) p. 84) is a function $\delta(p_1, p_2)$, $p_1 = p(x;\theta_1)$, $p_2 = p(x;\theta_2) \in P$, such that:

(1) $\delta(p_1, p_2) \geq 0$, and $\delta(p_1, p_2) = 0$ if and only if $p_1 = p_2$;

(2) $D_{\theta_1}\delta(p_1, p_2)|_{\theta_1=\theta_2} = D_{\theta_2}\delta(p_1, p_2)|_{\theta_2=\theta_1} = 0$ at p_1, p_2 and

(3) $D^2_{\theta_2}\delta(p_1, p_2)\big|_{\theta_1=\theta_2} = G(\theta_1)$ and $G(\theta_1)$ is a positive definite matrix,

where D_θ denotes differentiation with respect to θ. The following well known result is fundamental for the results to follow:

Lemma 1 *(i) The divergence $\delta(\theta_0, \theta_0+\theta)$ between two neighbouring points θ_0 and $\theta_0+\theta$ is*

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

(ii) (Morse's Lemma) There is a neighbourhood $\mathcal{B}_{\mathbb{R}^q}(\theta_0, \varepsilon^*)$ of θ_0 and a diffeomorphism ϕ such that $\phi(0)=0$ and for every $\theta \in \mathcal{B}_{\mathbb{R}^q}(\theta_0, \varepsilon^*)$

$$\delta(\theta_0, \theta_0+\phi(\tau)) = \tau'\tau.$$

For a discussion and background on the Morse's lemma see Milnor (1963) and Castrigiano and Hayes (1993). Lemma 1 implies that $\delta(\theta_0, \theta_0+\theta)$ behaves locally as half the square of an Euclidean distance. Several measures of statistical distance have this form (see Blyth (1994) who calls them "Rao divergences"): the Kullback-Leibler divergence, the α-divergences of Amari (1985) and the Hellinger distance (see Gibbs and Su (2002) for a survey of the relationships among them). The total variation distance used in the statistical literature mentioned in the introduction and the Hellinger distance (which is a special case of divergence) induce the same topology and uniformity on the set of probability measures (see for example LeCam and Yang (1990) and Gibbs and Su (2002)).

In the rest of the paper we assume that the matrix G used to define the divergence is the Fisher information matrix, but other choices are possible (e.g. the observed information). In $\mathbb{R}^q$ a ball of radius ε and centre $\theta_0 \in \mathbb{R}^q$ is the set

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

where $|.|$ denotes the usual Euclidean distance. Analogously, using this notation introduced above, a ball of radius ε centred at $p_0 \in P$ is the set

$$\mathcal{B}_P(p_0, \varepsilon) = \left\{p \in P : \delta(p_0, p) < \frac{1}{2}\varepsilon^2\right\}.$$

To simplify the presentation we will write $\delta(\theta_1, \theta_2)$ for $\delta(p_1, p_2)$, when $p_1 = p(x;\theta_1), p_2 = p(x;\theta_2)$.

Consider a map $\eta : P \to \mathbb{R}^m$, $m \leq q$. A (small) perturbation of p_0 will induce a change of $\eta(p_0)$. The largest amplification factor for a perturbation of p_0 on $\eta(p_0)$

is

$$M^{\eta}(p_0) = \sqrt{\liminf_{\varepsilon \to 0} \left\{ \phi : \eta\left(\mathcal{B}_P(p_0, \varepsilon)\right) \subset \mathcal{B}_{\mathbb{R}^m}\left(\eta(p_0), \frac{\phi\varepsilon}{\sqrt{n}}\right) \right\}}.$$

This definition takes explicitly into account the sample size since it considers a neighbourhood $\phi\varepsilon/\sqrt{n}$ of $\eta(p_0)$ which shrinks with the sample size and, thus, acknowledges the fact that information about the interest parameter is usually of order $O(n)$. This simplifies the interpretation of the results of the paper when using standard asymptotic arguments, but it may be modified in some cases (see also Sections 3 and 4.2 below).

Heuristically, if we let $p_0 = p(x; \theta_0)$, and with a slight abuse of notation, denote the perturbed density by $p_0 + \varepsilon$, then by changing p_0 to $p_0 + \varepsilon$, the largest change in $\eta(p_0)$ is at most $M^{\eta}(p_0)(\varepsilon/\sqrt{n})$.

The following theorem and two corollaries are proved in Appendix A.

Theorem 1 *The quantity $M^{\eta}(p_0)$ is well defined and*

$$M^{\eta}(p_0) = \lambda_M^{1/2}\left(\dot{\eta}(\theta_0)\left[n^{-1}G(\theta_0)\right]^{-1}\left[\dot{\eta}(\theta_0)\right]'\right)$$

where $\lambda_M(A)$ denotes the largest eigenvalue of the matrix A, $\eta(\theta) = \eta(p(x;\theta))$ and $\dot{\eta}(\theta_0) = D_\theta \eta(\theta)|_{\theta=\theta_0}$. The rank of $\dot{\eta}(\theta_0)$ is less or equal to m.

Corollary 1 *For a fixed map η, $M^{\eta}(p_0)$ is invariant to reparameterizations of P.*

Corollary 2 *Let $\lambda_M(A)$ and $\lambda_m(A)$ denote the largest and the smallest eigenvalues of the matrix A. If $\eta = \theta$,*

$$M^{\theta}(p_0) = \lambda_M^{1/2}\left(\left[n^{-1}G(\theta_0)\right]^{-1}\right) = \lambda_m^{-1/2}\left(n^{-1}G(\theta_0)\right).$$

If $\eta = \theta_1$ where $\theta = (\theta_1', \theta_2')$, then

$$M^{\theta_1}(p_0) = \lambda_M^{1/2}\left(\left[n^{-1}\bar{G}_{11.2}(\theta_0)\right]^{-1}\right) = \lambda_m^{-1/2}\left(n^{-1}\bar{G}_{11.2}(\theta_0)\right)$$

where $G_{11.2}(\theta_0)$ denotes the orthogonalised metric

$$\bar{G}_{11.2}(\theta_0) = G_{11}(\theta_0) - G_{12}(\theta_0)\left[G_{22}(\theta_0)\right]^{-1} G_{21}(\theta_0)$$

and

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

is partitioned conformably to θ*. Moreover,* $M^{\theta_1}(p_0)$ *is invariant to the reparameterizations of* θ_2*.*

The results above show that $M^{\eta}(p_0)$ is well defined for the class of models under consideration, and is unaffected by the way the manifold of PDFs is parameterized. Theorem 1 requires the function $\eta(p)$ to be differentiable at p_0, but does not impose any restriction on the rank of $\dot{\eta}(\theta_0)$. If the map of interest is the parameterization itself, the largest change in the parameter of interest induced by the move from p_0 to $p_0 + \varepsilon$ is the square root of the reciprocal of the smallest eigenvalue of the standardized information matrix.

We will now impose restrictions on the map $\eta : P \to \mathbb{R}^m$ and look at the problem from a slightly different perspective. Suppose that the map $\eta : P \to \mathbb{R}^m$ is a submersion (i.e. $\dot{\eta}(\theta_0) = D_{\theta}\eta(p(x;\theta))|_{\theta=\theta_0}$ has rank $m \leq q$). Then, we can reparameterise the manifold P in term of η and other $q-m$ parameters orthogonal to η, ϕ say, $\theta = \psi(\eta, \phi)$. The metric changes to

$$\begin{aligned} G(\eta_0, \phi_0) &= \left[\dot{\psi}(\eta_0, \phi_0)\right]' G(\theta) \dot{\psi}(\eta_0, \phi_0) \\ &= \begin{pmatrix} G_{11}(\eta_0, \phi_0) & G_{12}(\eta_0, \phi_0) \\ G_{21}(\eta_0, \phi_0) & G_{22}(\eta_0, \phi_0) \end{pmatrix} \end{aligned}$$

where $\dot{\psi}(\eta_0, \phi_0) = D_{(\eta,\phi)}\psi\big|_{(\eta,\phi)=(\eta_0,\phi_0)}$. Let $|G_{22}(\eta_0, \phi_0)| > 0$, then the orthogonalised metric $G_{11}(\tilde{\eta}, \phi)$ is

$$\bar{G}_{11.2}(\eta_0, \phi_0) = G_{11}(\eta_0, \phi_0) - G_{12}(\eta_0, \phi_0)' G_{22}(\eta_0, \phi_0)^{-1} G_{21}(\eta_0, \phi_0).$$

and it is invariant to reparameterizations of ϕ.

We are now interested in looking at changes along the η coordinates keeping ϕ fixed at ϕ_0. To do this we need to define the divergence $\bar{\delta}$ along the η coordinates with respect to $\bar{G}_{11}(\eta_0, \phi_0)$ in the same way as above, and let

$$\bar{\mathcal{B}}_P(p_0, \varepsilon) = \left\{ p \in P : \bar{\delta}(p_0, p) < \frac{1}{2}\varepsilon^2 \right\}.$$

Then we have the following result:

Theorem 2 *The quantity*

$$\bar{M}^{\eta}(p_0) = \sqrt{\liminf_{\varepsilon \to 0} \left\{ \phi : \eta\left(\bar{\mathcal{B}}_P(p_0, \varepsilon)\right) \subset \mathcal{B}_{\mathbb{R}^m}\left(\eta(p_0), \frac{\phi\varepsilon}{\sqrt{n}}\right) \right\}}$$

is well defined and equals $M^{\eta}(p_0) = \lambda_m^{-1/2}\left(n^{-1}\bar{G}_{11.2}(\theta_0)\right)$ *given in Corollary 2. Moreover, the direction in the* η *coordinates in which this is achieved is given by the eigenvectors of* $n^{-1}\bar{G}_{11.2}(\tilde{\eta}_0, \phi_0)$ *associated to* $\lambda_m(n^{-1}G_{11.2}(\theta_0))$.

Given that $\bar{M}^{\eta}(p_0) = M^{\eta}(p_0)$, in the rest of the paper we will denote them with the same symbol $M^{\eta}(p_0)$. Heuristically, by changing p_0 to $p_0 + \varepsilon$ along the η coordinates, the largest change in $\eta(p_0)$ is at most $\bar{M}^{\eta}(p_0)(\varepsilon/\sqrt{n})$.

Also define

$$\bar{\mu}^{\eta}(p_0) = \sqrt{\liminf_{\varepsilon \to 0} \left\{ \phi : \eta^{-1}\left(\mathcal{B}_{\mathbb{R}^m}\left(\eta(p_0), \frac{\varepsilon}{\sqrt{n}}\right)\right) \subset \bar{\mathcal{B}}_P(p_0, \phi\varepsilon) \right\}}.$$

This is the change in p_0 (along the η coordinates) required to change η by an amount equal to ε.

Theorem 3 *The quantity* $\bar{\mu}^{\eta}(p_0)$ *is well defined and*

$$\bar{\mu}^{\eta}(p_0) = \lambda_M^{1/2}\left(n^{-1}\bar{G}_{11.2}(\eta_0, \phi_0)\right).$$

Moreover, the direction in the η *coordinates in which this is achieved is given by the eigenvectors of* $n^{-1}\bar{G}_{11.2}(p_0)$ *associated to* $\lambda_M(n^{-1}G_{11.2}(\theta_0))$.

Corollary 3 *Let* $\|A\|_2$ *denote the spectral norm of the matrix* A*. Then, the distance of* $n^{-1}\bar{G}_{11.2}(\eta_0, \phi_0)$ *from the nearest point of where the matrix* G *is rank deficient (measured in terms of the norm* $\|.\|_2$*) equals* $1/M^{\eta}(p_0) = 1/\bar{M}^{\eta}(p_0)$.

2.1 Discussion

Both $M^{\eta}(p_0)$ and $\bar{\mu}^{\eta}(p_0)$ measure how perturbations of p_0 along the η coordinates are mapped into changes in $\eta(p_0)$, and they quantify two related but different ideas. The quantity $M^{\eta}(p_0)$ is the largest change in $\eta(p_0)$ that we can achieve

when we change p_0 along the η coordinates. On the other hand $\bar{\mu}^{\eta}(p_0)$ is the smallest change of p_0 (along the η coordinates) which yields a given change in $\eta(p_0)$. Consider the case where $\bar{\mu}^{\eta}(p_0)$ is infinite: when p_0 is slightly perturbed $\eta(p_0)$ is unaffected, i.e. $\eta(p_0)$ is locally constant. Inference about the point $\eta(p_0)$ can be done using $\eta(\hat{p})$ where $\hat{p}$ denotes any DGP sufficiently close to p_0. In this situation we have imposed very tight restrictions on the manifold of PDFs and, in a neighbourhood of p_0, all PDFs are mapped into the same point $\eta(p_0)$. We are using a model which is too restrictive, and we could relax some of these restrictions without compromising the statistical inference about $\eta(p_0)$.

The other extreme situation is the one where $M^{\eta}(p_0)$ is infinite: small perturbations of p_0 generate infinitely large changes of $\eta(p_0)$. The point $\eta(p_0)$ is not exactly defined and inference about it can be extremely difficult and imprecise. For example, if we are interested in constructing a confidence set $C(x)$ for $\eta(p_0)$, then we would like

$$\sup_{\varepsilon>0} \inf_{p\in\mathcal{B}_P(p_0,\varepsilon)} \int_{\{x:\eta(p)\in C(x)\}} p(x)\,dx = 1-\alpha$$

for a fixed $0<\alpha<1$ and also we would like $C(x)$ to be bounded with probability one. However, if $M^{\eta}(p_0)=\infty$ this is not possible. Heuristically, by perturbing p_0 slightly, the change of $\eta(p_0)$ is so large that no bounded confidence set can contain $\eta(p_0+\varepsilon)$. In this case, to be able to make precise inference about $\eta(p_0)$ we need to restrict the manifold of PDFs. This is analogous to the impossibility results of Bahadur and Savage (1956), Singh (1963), Koschat (1987), Gleser and Hwang (1987), Dufour (1997), Pfanzagl (1998) and Pötscher (2002). Formally, Pötscher (2002) has shown that, given some regularity conditions, for any proper loss function, the minimax risk for estimating $\eta(p)$ is bounded below by $2^{-2}osc(\eta,p_0)$ where $osc(\eta,p_0)$ denotes the oscillation of the function $\eta: P\to R^m$ at the point $p_0\in P$. Using the notation introduced above

$$\begin{aligned} osc(\eta,p_0) &= \lim_{\varepsilon\to 0} \sup_{\{p\in P:\delta(p_0,p)\leq\varepsilon^2/2\}} \sqrt{\frac{(\eta(p)-\eta(p_0))'(\eta(p)-\eta(p_0))}{(\varepsilon/\sqrt{n})^2}} \\ &= M^{\eta}(p_0) \end{aligned}$$

where the second line follows from the first because the square root is a continuos

function. Thus if the matrix $G(\theta_0)$ is close to be singular the minimax risk for estimating $\eta(p)$ can be arbitrarily large. Inferential problems in this case have been documented by Bottai (2003).

If G denotes the Fisher information matrix, it is well known that a sufficient (although not necessary) condition for local identification is that G is non singular (see for instance Theorem 1 p 579 of Rothenberg (1971) and Section 3 of Bowden (1973)). For models for which identification is determined by the non-singularity of the Fisher information matrix, Corollary 1 establishes a link between $M^{\eta}(p_0)$ and the set where the parameter of interest η is not identified (a discussion of the singular information matrix case is in Section 3). As such, it is a measure of identification (i.e. of instruments' weakness): small values of $M^{\eta}(p_0)$ indicate that η is identified, and large values of $M^{\eta}(p_0)$ suggest that the parameter η is close to being unidentified.

This interpretation is in accordance with Bowden (1973) who suggests using the rate of change of the Kullback-Leibler divergence to measure the "sharpness" of identification. Furthermore, for a general M-estimator the divergence $\theta \rightarrow M(\theta) = \delta(\theta_0, \theta_0 + \theta)$ represent an "asymptotic criterion function", and Lemma 1 gives approximations for it in a neighbourhood of θ_0. It is well known that if this map changes quickly as θ moves away from 0 then the estimator $\hat{\theta}_n$ maximizing the sample equivalent of the "asymptotic criterion function" has a high rate of convergence given some regularity conditions (see for example Van der Vaart (2000) Section 5.8).

A further interpretation of the measures suggested above hinges on identifying the divergence with a loss function having a minimum at the true DGP $p_0 \in P$. Looking at the proof of Theorem 3 we deduce that $\bar{\mu}^{\eta}(p_0)$ is the largest relative increase in the loss function which can be caused by a small change of the parameters of interest. Analogously, $M^{\eta}(p_0)$ is the relative change in the interest parameters necessary to increase the loss function by a small amount. Both quantities measure how quickly the loss function (i.e. the divergence) changes as the parameters of interest change.

An *estimator* of the point $\eta(p)$ is a function

$$\hat{\eta} : P \to \mathbb{R}^q.$$

This is an abstract, although useful, definition of estimator when looking at the problem under consideration. It is different from the most common definition for which an estimator is a function from the sample space to the parameter space, but many common estimators have this form. For example the MLE is the function $P \to \mathbb{R}^q$ which associates to a point $p \in P$ the quantity $\arg\max_\theta \{\ln[p(x;\theta)]\}$. Note that even if $\eta(P) = \hat{\eta}(P)$, the estimator and the parameterization are not necessarily the same function, although in most cases they do coincide. Since an estimator is a map defined on the manifold of PDFs, the quantities $M^{\hat{\eta}}(p_0)$, $\bar{M}^{\hat{\eta}}(p_0)$ and $\bar{\mu}^{\hat{\eta}}(p_0)$ are well defined.

Note that the rule according to which an element of P is chosen gives rise to a map $\hat{p} : P \to P$. For example if P is parameterised as $p(x;\theta)$ and $\hat{\theta}$ is the MLE of θ, then $\hat{p} = p\left(x;\hat{\theta}\right)$ is the image of such a map. In this case if we take this rule as given we can regard a function $\eta : P \to \mathbb{R}^p$ as the map $\hat{p} \to \eta(\hat{p})$. For example, if P is a *full exponential family*, i.e., its elements are of the form

$$p(x;\theta) = \exp\{\theta' x - K(\theta)\}, \theta \in \mathbb{R}^q \tag{1}$$

with respect to a certain dominating measure, then we can introduce the expectation coordinates as in Amari (1985)

$$\eta_i(\theta) = \frac{\partial K(\theta)}{\partial \theta_i} = E_{p(x;\theta)}(x_i)$$

and regard η as a function of the observed x (the sufficient statistics). In particular, the MLE of η is just $\hat{\eta} = x$, and $E(\hat{\eta}) = \eta$. If we take the rule $\hat{p}$ as given, we can also focus on the observed quantities $M^{\hat{\eta}}(\hat{p})$ and $\bar{\mu}^{\hat{\eta}}(\hat{p})$.

By evaluating an estimator $\hat{\eta}$ at the observed $\hat{p}$ we establish a link with the work of Van Garderen (1996) who investigates how curvature affects the precision of the maximum likelihood estimator (MLE) in curved exponential models. Van Garderen (1996) shows with a few examples that, in the presence of a significant curvature, small changes of the observations (sufficient statistics) lead to

large changes in the estimates. Our measure differ from the idea of model curvature of Amari (1985), because, for example, the Fieller-Creasy model is flat but none of our measures of sensitivity is constant over the parameter space or over the manifold of PDFs.

Although our measures $\bar{\mu}^{\eta}(p_0)$ and $M^{\eta}(p_0)$ differ from the idea of model curvature of Amari (1985), they are closely related to it. In the case where $q = 2$, $\bar{\mu}^{\eta}(p_0)$ and $M^{\eta}(p_0)$ are the maximal and minimal Euler curvature respectively. Their product, the Gaussian curvature, may be constant but $\bar{\mu}^{\eta}(p_0)$ and $M^{\eta}(p_0)$ may vary considerably. This is the case, for instance, in the Fieller-Creasy set-up.

3 The singular information matrix case

The fact that the information matrix is positive definite is a sufficient but by no means necessary condition for identification of a parametric model. We can extend this observation to the set-up considered in this paper by noting that we do not need G to be nonsingular to define a divergence. We can generalize the notion of divergence by replacing condition (3) above with

(3') $D^{i}_{\theta_2}\delta(p_1,p_2)\big|_{\theta_1=\theta_2} = 0$ for $i = 1,2,...,2m+1$ $(m \geq 2)$ and $t_{i_1 i_2 .... i_{2m+2}}\theta_2^{i_1}\theta_2^{i_2}...\theta_2^{i_{2m+2}} > 0$ for all $\theta_2 = \left(\theta_2^1,\theta_2^2,...,\theta_2^q\right) \in \mathbb{R}^q$, $\theta \neq 0$, where we use the summation convention whereby summation is implied when an index is repeated on the upper and lower level, and we define

$$t_{i_1 i_2 ... i_{2m+2}} = \left.\frac{\partial^j \delta(p_1,p_2)}{\partial\theta_2^{i_1}\partial\theta_2^{i_2}...\partial\theta_2^{i_j}}\right|_{\theta_1=\theta_2}$$

where $\theta_2 = \left(\theta_2^1,\theta_2^2,...,\theta_2^q\right)$ and $j = i_1 + i_2 + ... + i_{2m+2}$.

Theorems 3.2 and 3.3 of Schaffler (1992) guarantee the existence of a minimum for the divergence at $p_1 = p_2$. If rank of G equal $q_1 < q$, the Reduction Lemma (Castrigiano and Hayes (1993) p 64) implies that there is a local diffeomorphism such that

$$\delta(\theta_0, \theta_0 + \psi(\tau,\phi)) = \tau'\tau + g(\phi)$$

where $\tau \in \mathbb{R}^{q_1}$ and $\phi \in \mathbb{R}^{q-q_1}$, and g is a smooth function such that $D^2_\phi g(\phi)\big|_{\phi=0} = 0$ for which the origin is a critical point and $g(0) = 0$. Note that under assumption (3), Morse Lemma allows us to reduce the divergence in a neighbourhood of a particular point to the square of the Euclidean distance. If, on the other hand, the matrix G is rank deficient the Reduction Lemma yields a decomposition of the divergence in a neighbourhood of a given point into two components: one is the square of the Euclidean distance ($\tau'\tau$) and one is not an Euclidean distance ($g(\phi)$).

The function g has a minimum at $\phi = 0$ (see Corollary 2.2 of Schaffler (1992)) and for a sufficiently small neighbourhood of zero $g(\phi) \leq \phi'\phi$. By expanding $g(\phi)$ as a Taylor series around $\phi = 0$ we have

$$g(\phi) = \frac{1}{i_1!i_2!...i_{2m+2}!} g_{i_1 i_2 .... i_{2m+2}} \phi^{i_1}\phi^{i_2}...\phi^{i_{2m+2}} + O\left(\|\phi\|^{2m+3}\right)$$

where

$$g_{i_1 i_2 ... i_{2m+2}} = \left.\frac{\partial^j g(\phi)}{\partial\phi^{i_1}\partial\phi^{i_2}...\partial\phi^{i_{2m+2}}}\right|_{\phi=0}.$$

This has very important implications. Let $\tau = 0$, $\phi = 0$ then $\delta(\theta_0, \theta_0 + \psi(0,0)) = 0$. Suppose we change one coordinate at a time by a quantity ε leaving all other coordinates unchanged. If we increase $\tau^i = 0$ to $\tau^i = \varepsilon$ $(i = 1, 2, ..., p_1)$ the increase in the divergence is ε^2. However, for a change in the ϕ coordinates from $\phi^i = 0$ to $\phi^i = \varepsilon$, the change in the divergence equals $\varepsilon^{2m+2} + O(\varepsilon^{2m+3})$. Given that the order of magnitude of changes in the τ and the ϕ coordinates are so different, we need to consider changes in the two coordinates separately.

To simplify the notation let

$$\delta(\theta_0, \theta_0 + \psi(\tau, \phi)) = \delta_{\theta_0}(\tau, \phi).$$

Also assume that there are no nuisance parameters (if there are, as before, we assume that η is a submersion and that P is parameterized in terms of η and some other $q - m$ parameters orthogonal to η, which are kept fixed and we consider changes along the η coordinates). Moreover, define

$$\mathcal{B}^{\tau}_P(p_0, \varepsilon) = \left\{ p = p(x; \theta_0 + \psi(\tau, \phi)) : \delta_{\theta_0}(\tau, 0) < \frac{1}{2}\varepsilon^2 \right\},$$

and

$$\mathcal{B}_P^{\phi}(p_0,\varepsilon)=\left\{p=p(x;\theta_0+\psi(\tau,\phi)):\delta_{\theta_0}(0,\phi)<\varepsilon^{2m+2}\right\}.$$

With this notation we can write

$$\begin{aligned}\bar{M}_{\tau}^{\eta}(p_0) &= \sqrt{\liminf_{\varepsilon\to 0}\left\{\phi:\eta\left(\mathcal{B}_P^{\tau}(p_0,\varepsilon)\right)\subset\mathcal{B}_{\mathbb{R}^m}\left(\eta(p_0),\frac{\phi\varepsilon}{\sqrt{n}}\right)\right\}}\\ \bar{M}_{\phi}^{\eta}(p_0) &= \sqrt{\liminf_{\varepsilon\to 0}\left\{\phi:\eta\left(\mathcal{B}_P^{\phi}(p_0,\varepsilon)\right)\subset\mathcal{B}_{\mathbb{R}^m}\left(\eta(p_0),\frac{\phi\varepsilon}{n^{m+1}}\right)\right\}}\end{aligned}$$

and

$$\begin{aligned}\bar{\mu}_{\tau}^{\eta}(p_0) &= \sqrt{\liminf_{\varepsilon\to 0}\left\{\phi:\eta^{-1}\left(\mathcal{B}_{\mathbb{R}^m}(\eta(p_0),\varepsilon)\right)\subset\mathcal{B}_P^{\tau}\left(p_0,\frac{\phi\varepsilon}{\sqrt{n}}\right)\right\}}\\ \bar{\mu}_{\phi}^{\eta}(p_0) &= \sqrt{\liminf_{\varepsilon\to 0}\left\{\phi:\eta^{-1}\left(\mathcal{B}_{\mathbb{R}^m}(\eta(p_0),\varepsilon)\right)\subset\mathcal{B}_P^{\phi}\left(p_0,\frac{\phi\varepsilon}{n^{m+1}}\right)\right\}}.\end{aligned}$$

Theorem 4 *The four measures of sensitivity defined above are well defined and*

$$\begin{aligned}\bar{M}_{\tau}^{\eta}(p_0) &= \lambda_M^{1/2}(nA_{11})\\ \bar{M}_{\phi}^{\eta}(p_0) &= \sqrt{n^{2(m+1)}\max_{\frac{1}{i_1!i_2!...i_{2m+2}!}g_{i_1i_2....i_{2m+2}}v^{i_1}v^{i_2}...v^{i_{2m+2}}=1} v'A_{22}v}\\ \bar{\mu}_{\tau}^{\eta}(p_0) &= \lambda_M^{1/2}\left(\frac{1}{n}A_{11}^{-1}\right)\\ \bar{\mu}_{\phi}^{\eta}(p_0) &= \sqrt{\frac{1}{n2^{(m+1)}}\max_{v'A_{22}v=1}\frac{1}{i_1!i_2!...i_{2m+2}!}g_{i_1i_2....i_{2m+2}}v^{i_1}v^{i_2}...v^{i_{2m+2}}}\end{aligned}$$

where

$$\begin{pmatrix} A_{11} & A_{12}\\ A_{12}' & A_{22}\end{pmatrix}=\left[\dot{\psi}(0,0)\right]'\left[\dot{\eta}(\theta)\right]'\dot{\eta}(\theta)\dot{\psi}(0,0)$$

and

$$\begin{aligned}\dot{\eta}(\theta) &= D_{\theta}\eta(\theta)|_{\theta=\theta_0}\\ \dot{\psi}(0,0) &= D_{\tau,\phi}\psi(\tau,\phi)|_{(\tau,\phi)=0}\end{aligned}$$

4 Examples

Examples of applications of the measures of sensitivity derived in Sections 2 and 3 are now given. A further application to structural equations models is given in Section 5.

4.1 The linear regression model

Consider a Gaussian linear regression model

$$y = X\beta + u$$

where β is an unknown $k \times 1$ vector of parameters, X is a nonstochastic $n \times k$ matrix and $u \sim N(0, \sigma^2 I_n)$. The Fisher partial information about β is $(1/\sigma^2) X'X$. So

$$\begin{aligned} M^{\beta}(p_0) &= \bar{M}^{\beta}(p_0) = \frac{\sigma}{\lambda_m^{1/2}(n^{-1}X'X)} \\ \bar{\mu}^{\beta}(p_0) &= \frac{\lambda_M^{1/2}(n^{-1}X'X)}{\sigma}. \end{aligned}$$

Note that $M^{\beta}(p_0) < \infty$ is, in this case, a necessary condition for identification of β. Note that the condition number suggested by Besley, Kuh, and Welsh (1980) as a measures of multicollinearity is

$$\mathrm{K}(X'X) = \bar{\mu}^{\beta}(p_0)\,\bar{M}^{\beta}(p_0) = \sqrt{\frac{\lambda_M(n^{-1}X'X)}{\lambda_m(n^{-1}X'X)}}.$$

The condition number has been criticised as a measure of multicollinearity because it depends on the units of measurement of the columns of X. However, if we declare that β defined by $y \sim N(X\beta, \sigma^2 I_n)$ is the parameter of interest, then X must be taken as given. By rescaling the columns of X, the parameter of interest is changed, and the measures of sensitivity reflect this.

4.2 Trend and autoregression

Consider a normal linear regression model of the form

$$y = x\beta + u$$

where β is a scalar parameter, x is the $n \times 1$ vector $x = (1, 2, ..., n)'$ and u is an $n \times 1$ random vector such that $u_i = \rho u_{i-1} + NID(0, \sigma^2)$ and $u_0 = 0$. The Fisher information matrix for the parameters (β, σ^2, ρ) is

$$\begin{pmatrix} \frac{n(n+1)(1+2n)}{6\sigma^2} - \frac{1}{3\sigma^2} n(n-1)(n+1)\rho + \frac{1}{6\sigma^2} n(n-1)(2n-1)\rho^2 & 0 & 0 \\ 0 & \frac{n}{2\sigma^4} & 0 \\ 0 & 0 & \frac{1}{\sigma^2} \sum_{j=1}^{n-1} \sum_{i=0}^{j-1} \rho^{2i} \end{pmatrix}$$

So taking into account that in this case the information about β is of order $O(n^3)$,

$$M^{\beta}(\beta, \sigma^2, \rho) = \sqrt{\frac{n^3}{\frac{n(n+1)(1+2n)}{6\sigma^2} - \frac{1}{3\sigma^2} n(n-1)(n+1)\rho + \frac{1}{6\sigma^2} n(n-1)(2n-1)\rho^2}}$$

and, as n grows, this equals

$$\lim_{n\to\infty} M^{\beta}(\beta, \sigma^2, \rho) = \sqrt{\frac{3\sigma^2}{1 - \rho + \rho^2}}$$

When the sample size is large, the sensitivity of the map from the DGP to $\mathbb{R}$ given by the parameter of interest β depends on both the error variance σ^2 and the error autocorrelation coefficient ρ. The quantity $\lim_{n\to\infty} M^{\beta}(\beta, \sigma^2, \rho)$ has a maximum at $\rho = 1/2$. This decreases rapidly as ρ moves away from $1/2$. Note that $\lim_{n\to\infty} M^{\beta}(\beta, \sigma^2, 0) = \sqrt{3\varepsilon^3} = \lim_{n\to\infty} M^{\beta}(\beta, \sigma^2, 1)$, so that the parameter of interest β is equally sensitive when $\rho = 0$ and when $\rho = 1$.

4.3 The Fieller-Creasy problem

Let $(x_{1i}, x_{2i})'$ be a sequence of n pairs of independent observations from a bivariate normal distribution with covariance matrix $\sigma^2 I_2$, and σ^2 is unknown. In this case we can reduce to problem by sufficiency to one for which

$$\begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \frac{\sigma^2}{n} I_2 \right)$$

$$s^2/\sigma^2 \sim \chi^2(2(n-1)),$$

and $(\bar{x}_1, \bar{x}_2)'$ and $s^2 = \sum_{i=1}^{n} \left((x_{1i} - \bar{x}_1)^2 + (x_{2i} - \bar{x}_2)^2 \right)$ are independent. The Fisher information matrix is

$$\frac{1}{n} G(\mu_1, \mu_2, \sigma) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 & 0 \\ 0 & \frac{1}{\sigma^2} & 0 \\ 0 & 0 & \frac{1}{\sigma\sqrt{n}} \end{pmatrix}.$$

We consider the effects of a reparameterization of the mean function, and regard σ as a nuisance parameter throughout this subsection.

Suppose the mean vector is written in polar coordinates as $\mu_1 = \rho \cos\phi$, $\mu_2 = \rho \sin\phi$ and that ρ is the parameter of interest. Then, defining $p_0 = p(x; \rho, \phi, \sigma)$

$$\begin{aligned} M^{\rho}(p_0) &= \bar{M}^{\rho}(p_0) = \frac{\sigma}{|\rho|} \\ \bar{\mu}^{\rho}(p_0) &= \frac{|\rho|}{\sigma}. \end{aligned}$$

Similarly if we write $\mu_1 = \mu_2 \psi$ and $\mu_2 = \mu_2$, $\sigma = \sigma$ (the Fieller-Creasy model), and focus on the parameter ψ. Then

$$\begin{aligned} M^{\psi}(p_0) &= \bar{M}^{\psi}(p_0) = \frac{\sigma\sqrt{1+\psi^2}}{|\mu_2|} \\ \bar{\mu}^{\psi}(p_0) &= \frac{|\mu_2|}{\sigma\sqrt{1+\psi^2}}, \end{aligned}$$

where $p_0 = p(x; \psi, \mu_2, \sigma)$. Estimators of the parameters of interest involve the same functions as the parameters evaluated at $(\mu_1, \mu_2) = (\bar{x}_1, \bar{x}_2)$.

Both formulations show that a small change of the DGP may create huge changes in the parameter of interest (ρ and ψ respectively) when (μ_1, μ_2) is in particular areas of $\mathbb{R}^2$ (see also James, Wilkinson, and Venables (1974), and Wallace (1980)). Such areas are defined in different ways depending on the parameter of interest (see Forchini and Hillier (2003) for further discussion), and are close to the region (of measure zero) in the (μ_1, μ_2)-space where the parameterizations $(\mu_1, \mu_2) \to (\rho, \phi)$ and $(\mu_1, \mu_2) \to (\psi, \mu_2)$ are not defined. It appears that as we get close to the edge of the regions where the parameterizations are well defined inference about the parameter of interest becomes more difficult because small changes of the DGP imply large changes for the parameter of interest, while changes of the parameter of interest only marginally affect the DGP.

5 Structural equations models

In this section we apply the measures derived above to structural equations models, and argue that they can be used to measure instruments weakness. We consider a single structural equation

$$y_1 = Y_2\beta + Z_1\gamma + u, \tag{2}$$

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 corresponding to (2) is

$$(y_1, Y_2) = Z_1 (\phi_1, \Phi_2) + Z_2 (\pi_1, \Pi_2) + (v_1, V_2), \tag{3}$$

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 $k_1 \times (1+n)$, $k_2 \times (1+n)$ 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 (2) is embedded in the reduced form (3).

A structural equations model is a multivariate linear regression model subject to the restrictions

$$\begin{aligned} \pi_1 - \Pi_2\beta &= 0 \\ \phi_1 + \Phi_2\beta &= \gamma \\ v_1 + V_2\beta &= u \end{aligned}$$

where β is a $n \times 1$ vector of parameters of interest (the coefficients of the endogenous variables). Forchini (2003) gives a detailed discussion of the role of these

compatibility conditions in determining identification of the structural parameters. Here we focus on the coefficients of the endogenous variable and ignore the last two restrictions. There is no loss of generality in doing this because the partial information about β is independent of the parameterization of the nuisance parameters.

We focus on the case in which the matrix G defining the divergence is the (partial) Fisher information about β. This is given by the following result.

Lemma 2 *The partial information matrix for a multivariate linear regression model subject to the restriction $\pi_1 = \Pi_2\beta$ is*

$$G\left(\beta : \Pi_2, \Phi, \Omega\right) = \frac{1}{\text{var}\left(u\right)}\Pi_2' Z_2 M_{Z_1} Z_2 \Pi_2$$

where

$$\begin{aligned} \beta^* &= \omega_{11.2}^{-1/2}\left(\Omega_{22}^{1/2}\beta - \Omega_{22}^{-1/2}\omega_{21}\right) \\ \text{var}\left(u\right) &= \left(1 + \beta^{*\prime}\beta^*\right)\omega_{11.2} \end{aligned}$$

and

$$\omega_{11.2} = \omega_{11} - \omega_{21}'\Omega_{22}^{-1}\omega_{21}.$$

Note that $G\left(\beta : \Pi_2, \Phi, \Omega\right)$ is similar to the concentration parameter considered by Stock, Wright, and Yogo (2002), but the partial information depends on $\text{var}\left(u\right)$ rather than on the covariance matrix of the rows of V.

If follows from the result above that

$$M^{\beta}\left(p_0\right) = \bar{M}^{\beta}\left(p_0\right) = \sqrt{\frac{T\,\text{var}\left(u\right)}{\lambda_m\left(\Pi_2' Z_2 M_{Z_1} Z_2 \Pi_2\right)}}$$

and

$$\bar{\mu}^{\beta}\left(p_0\right) = \sqrt{\frac{\lambda_M\left(\Pi_2' Z_2 M_{Z_1} Z_2 \Pi_2\right)}{T\,\text{var}\left(u\right)}}.$$

Therefore, $M^{\beta}\left(p_0\right)$ is arbitrarily large when the structural equation can be arbitrarily close to be unidentifiable (i.e. $rank\left(\Pi_2\right) < n$). In this case a small change

in $p \in P$ significantly affects β because some components of β are not identified (Phillips (1989)) and can thus take on any value.

The quantity $\bar{\mu}^{\beta}(p_0)$ is arbitrarily close to zero when the structural equation can be arbitrarily close to be totally unidentifiable (i.e. $\Pi_2 = 0$). If β is only partially identified, a change of these parameters affects the manifold of PDFs through the identified parameters.

If $T^{-1}Z'Z \xrightarrow{P} Q$, where $Z = [Z_1, Z_2]$ and Q is a finite positive definite matrix we have that

$$\bar{\mu}^{\beta}(p_0) \xrightarrow{P} \sqrt{\frac{\lambda_M\left(\Pi_2' Q_{11.2} \Pi_2\right)}{\operatorname{var}(u)}}$$

and

$$M^{\beta}(p_0) \xrightarrow{P} \sqrt{\frac{\operatorname{var}(u)}{\lambda_m\left(\Pi_2' Q_{11.2} \Pi_2\right)}},$$

where $Q_{11.2} = Q_{11} - Q_{12}Q_{22}^{-1}Q_{21}$ and Q is partitioned conformingly to Z. In the case where the instruments are weak, i.e. $\Pi_2 = O\left(T^{-1/2}\right)$, we have $\bar{\mu}^{\beta}(p_0) \xrightarrow{P} 0$ and $M^{\beta}(p_0) \xrightarrow{P} \infty$.

In structural models, the common estimators define maps which are different from the components of the parameterizations corresponding to the parameters of interest. Here we will consider the sensitivity of the TSLS and the OLS estimators of β.

Note that both TSLS and OLS are defined in terms of the sufficient statistics

$$\begin{aligned}
\hat{\Phi} &= \left(Z_1'Z_1\right)^{-1} Z_1'Y \sim N\left(\Phi + \left(Z_1'Z_1\right)^{-1} Z_1'Z_2\Pi, \left(Z_1'Z_1\right)^{-1} \otimes \Omega\right) \\
\hat{\Pi} &= \left(Z_2'M_{Z_1}Z_2\right)^{-1} Z_2'M_{Z_1}Y \sim N\left(\Pi, \left(Z_2'M_{Z_1}Z_2\right)^{-1} \otimes \Omega\right) \\
S &= Y'M_Z Y \sim W_{n+1}\left(\upsilon - k_2, \Omega\right),
\end{aligned}$$

(note that they also are independent of each other) where $\Phi = (\phi_1, \Phi_2)$ and $\Pi = (\pi_1, \Pi_2)$. These are the observed values of the Φ, Π and Ω for a given sample. In terms of this parameterization, the TSLS and the OLS estimators are

$$\begin{aligned}
\hat{\beta}_{TSLS}(\pi_1, \Pi_2) &= \left(\Pi_2'Z_2'M_{Z_1}Z_2\Pi_2\right)^{-1} \Pi_2'Z_2'M_{Z_1}Z_2\pi_1 \\
\hat{\beta}_{OLS}(\pi_1, \Pi_2, \Omega) &= \left(T\Omega_{22} + \Pi_2'Z_2'M_{Z_1}Z_2\Pi_2\right)^{-1} \left(T\omega_{21} + \Pi_2'Z_2'M_{Z_1}Z_2\pi_1\right).
\end{aligned}$$

For the case $n = 1$ we have the following result.

Theorem 5 *Let $A = T^{-1}Z_2'M_{Z_1}Z_2$. For the TSLS estimator we have*

$$TG^*_{TSLS} = \frac{\omega_{11} - 2\omega_{12}\hat{\beta}_{TSLS}}{\Pi_2'A\Pi_2} + \Omega_{22}\frac{\pi_1'A\pi_1}{(\Pi_2'A\Pi_2)^2},$$

and for the OLS estimator we have

$$\begin{aligned} TG^*_{OLS} &= \frac{\omega_{11} - 2\omega_{12}\hat{\beta}_{OLS}}{\Omega_{22} + \Pi_2'A\Pi_2} + \Omega_{22}\frac{\pi_1'A\pi_1}{(\Omega_{22} + \Pi_2'A\Pi_2)^2} \\ &\quad - \left(\frac{\omega_{12} - 2\Omega_{22}\hat{\beta}_{OLS}}{\Omega_{22} + \Pi_2'A\Pi_2}\right)^2 + \frac{2\beta^2_{OLS}\Omega^2_{22}}{(\Omega_{22} + \Pi_2'A\Pi_2)^2}. \end{aligned}$$

Since for this simple case $\mu^{\hat{\beta}}(p_0) = \sqrt{TG^*} = 1/M^{\hat{\beta}}(p_0)$, and since $\Omega_{22} + \Pi_2'Z_2'M_{Z_1}Z_2\Pi_2 > \Pi_2'Z_2'M_{Z_1}Z_2\Pi_2$, we can conclude that the OLS estimator is affected less than the TSLS estimator by the weak instruments problem. This is in accordance with the results of Forchini and Hillier (2003) who show that the density on the TSLS estimator conditional on an identification test statistic is more sensitive to identification than the OLS estimator.

If we consider a sequence of $\{\Pi_2\}_{p=1}^{\infty}$ converging to a zero vector, we have that $TG^*_{TSLS} \to \infty$ while $TG^*_{OLS} \to |\Omega|/\Omega^2_{22}$. Thus, even if the model is very close to be unidentified, the effect of a change of the DGP on the OLS estimator is finite, and depends only on Ω.

5.1 An application to return to schooling

We now consider an application of the measures discussed above to the estimation of return to schooling. Angrist and Krueger (1991) estimate the return to schooling for men born in the U.S. in 1930-1939 using Public-Use Microdata Samples (PUMS) for 1980 by using quarter of birth, and quarter of birth interacted with other variables as instrumental variables (see Angrist and Krueger (1991) for the construction of the dataset). Bound, Jaeger, and Baker (1995) point out that the instruments used by Angrist and Krueger (1991) are weak, and, as a consequence, that inference based on them is unreliable.

We use the same dataset as Angrist and Krueger (1991), and calculate the measures of transmission of perturbations. They depend on the parameters of the

model, and they can be estimated consistently under very general conditions. In Table 1 estimates of the parameterization and estimator sensitivity are reported for various model specifications as in Table V of Angrist and Krueger (1991). If we consider for example column (1) the parameter β is estimated using OLS, and this estimate (0.0710) is used in all measures of sensitivity in such column. $M^{\beta}(\hat{p}_0)$ indicates that a small perturbation ε in the manifold of PDFs may change the parameter of the endogenous variable (education) by as much as $9.2017 \times \varepsilon$. The smallest change in the manifold of PDFs needed to produce a change ε in the parameter of education is $0.1087 \times \varepsilon$. One may consider 0.1087 as the distance from the nearest point where β is not identified. The next quantities in column (1) refer to the estimator (OLS) and indicate that OLS is insensitive to changes in the manifold of PDFs.

Table 1 shows that identification may be an issue in this dataset as already pointed out by serval authors before. It also show that OLS and TSLS may be affected by identification in very different ways even though the estimates obtained are very close. Precisely, TSLS appears to be very sensitive. OLS seem to be insensitive to small changes of the DGP. OLS also take values approximately equal to $\left|\hat{\Omega}\right| / \hat{\Omega}_{22}^2$ as we would expect if the model would be unidentified.

Table 1 approximately here

6 Concluding remarks

The relationship between the parameters of interest and the manifold of PDFs can be very weak in some ill-posed problems. In this paper we have argued that it is possible to quantify the weakness of this relationship in a parametric set-up by measuring the how it transmits perturbations from the space of DGP to a multidimensional Euclidean space (e.g. the parameter space). We have provided several interpretations of these measures, and have argued that by evaluating these measures at the observed PDF we obtain information about the post-data precision with which this map can be located.

An application to structural equations models have been considered in detail. We have shown how coherent, simple and easily interpretable measures of weakness of instruments can be obtained and interpreted. Their empirical relevance has been illustrated with the estimation of the returns to schooling with the dataset of Angrist and Krueger (1991).

References

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

Angrist, J. D., and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?," *Quarterly Journal of Economics*, CVI, 979–1014.

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

Besley, D., E. Kuh, and R. Welsh (1980): *Regression Diagnostics*. Wiley, New York.

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

Bottai, M. (2003): "Confidence Regions When the Fisher Information is Zero," *Biometrika*, 90, 73–84.

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.

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.

Chamberlain, G., and G. Imbens (2004): "Random Effects Estimators with Many Instrumental Variables," *Econometrica*, 72 (1), 295 – 306.

Choi, I., and P. C. Phillips (1992): "Asymptotic and Finite Sample Distribution Theory for IV Estimators and Tests in Partially Identified Structural Equations," *Journal of Econometrics*, 51, 113–150.

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

Forchini, G. (2003): "Testing the Relevance of Instrumental Variables in Structural Estimation," Mimeo, University of York.

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.

Godfrey, L. G. (1999): "Instrument Relevance in Multivariate Linear Models," *Review of Economics and Statistics*, LXXXI, 550–552.

Goutis, C., and G. Casella (1995): "Frequentist Post-Data Inference," *International Statistical Review*, 63, 325–344.

Hahn, J., and J. Hausman (2002a): "A New Specification Test for the Validity of Instrumental Variables," *Econometrica*, 70, 163–189.

——— (2002b): "Weak Instruments: Diagnosis and Cures in Empirical Econometrics," mimeo, MIT.

Han, C., and P. C. Phillips (2003): "GMM with Many Moment Conditions," mimeo.

Hillier, G. H. (1985): "On the Joint and Marginal Densities of Instrumental Variable Estimators in a General Structural Equation," *Econometric Theory*, 1, 53–72.

James, A., G. Wilkinson, and W. Venables (1974): "Interval Estimates for a Ratio of Means," *Sankhya, Series A*, 36, 177–183.

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

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.

Lindsay, B. G., and B. Li (1997): "On Second-Order Optimality of the Observed Fisher Information," *Annals of Statistics*, 25, 2172–2199.

Magnus, J. R., and H. Neudecker (1988): *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley and Sons Ltd., New York.

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

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

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

Phillips, P. C. B. (1983): "Exact Small Sample Theory in the Simultaneous Equation Model," in *Handbook of Econometrics*, ed. by M. D. Intriligator, and Z. Griliches, pp. 449–516. North Holland, Amsterdam.

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

Poskitt, D. S., and C. L. Skeels (2002): "Assessing Instrumental Variable Relevance: An Alternative Measure and some Exact Finite Sample Theory," Working paper No. 862, Department of Economics, University of Melbourne.

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.

Schaffler, S. (1992): "Classification of Critical Stationary Points in Unconstrained Optimization," *Siam Journal of Optimization*, 2, 1–6.

Shea, J. (1997): "Instrument Relevance in Multivariate Linear Models: A Simple Measure," *The Review of Economics and Statistics*, LXXIX, 348–352.

Sims, C. A. (1980): "Macroeconomics and Reality," *Econometrica*, 48, 1–48.

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

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.

Stock, J. H., and M. Yogo (2003): "Testing for Weak Instruments in Linear IV Regression," Mimeo, Harvard University.

Sundberg, R. (2003): "Conditional Statistical Inference and Quantification of Relevance," *Journal of the Royal Statistical Society B*, 65, 299–315.

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

Van Garderen, K. J. (1996): "Variance Inflation in Curved Exponential Models," Discussion Paper 9522, Department of Economics, University of Southampton.

Wallace, D. L. (1980): "The Behrens-Fisher and the Fieller-Creasy Problems," in *R.A. Fisher: An Appreciation*, ed. by S. Fienberg, and D. Hinkley, Lecture Notes in Statistics. Springer-Verlag, New York.

A Proofs of results in Section 2

A.1 Proof of Theorem 1

Let

$$M_{\varepsilon}^{\eta}(p_0) = \sup_{\{p \in \mathcal{B}_P(p_0,\varepsilon)\}} \frac{(\eta(p) - \eta(p_0))'(\eta(p) - \eta(p_0))}{(\varepsilon/\sqrt{n})^2}$$

so that $M^{\eta}(p_0) = \sqrt{\lim_{\varepsilon \to 0} M_{\varepsilon}^{\eta}(p_0)}$. According to Lemma Morse's lemma (lemma 1 (ii)) it is possible to find a local diffeomorphism ψ such that $\delta(\theta_0, \theta_0 + \theta) = \delta_{\theta_0}(\theta)$ is locally equal to

$$\delta_{\theta_0}(\psi(\tau)) = \tau'\tau.$$

Now, let $\eta(\theta) = \eta(p(x;\theta))$, and expand $\eta(\theta_0 + \theta) - \eta(\theta_0) = \eta(\theta_0 + \psi(\tau)) - \eta(\theta_0)$ around $\tau = 0$ in a Taylor series

$$\eta(\theta_0 + \theta) - \eta(\theta_0) = \dot{\eta}(\theta_0)\, D_{\tau}\psi(\tau)|_{\tau=0}\, \tau + O\left(|\tau|^2\right)$$

so that

$$\begin{aligned} & (\eta(\theta_0 + \psi(\tau)) - \eta(\theta_0))'(\eta(\theta_0 + \psi(\tau)) - \eta(\theta_0)) \\ = \;& \tau' \left[D_{\tau}\psi(\tau)|_{\tau=0}\right]' \left[\dot{\eta}(\theta_0)\right]' \dot{\eta}(\theta_0)\, D_{\tau}\psi(\tau)|_{\tau=0}\, \tau + O\left(|\tau|^3\right) \\ = \;& \tau' \left[D_{\tau}\psi(\tau)|_{\tau=0}\right]' \left[\dot{\eta}(\theta_0)\right]' \dot{\eta}(\theta_0)\, D_{\tau}\psi(\tau)|_{\tau=0}\, \tau + O\left(\varepsilon^3\right). \end{aligned}$$

Thus $M_{\varepsilon}^{\eta}(p_0)$ equals

$$M_{\varepsilon}^{\eta}(p_0) = n \sup_{\{\tau : \tau'\tau \leq \varepsilon^2\}} \frac{\tau' \left[D_{\tau}\psi(\tau)|_{\tau=0}\right]' \left[\dot{\eta}(\theta_0)\right]' \dot{\eta}(\theta_0)\, D_{\tau}\psi(\tau)|_{\tau=0}\, \tau + O\left(\delta^3\right)}{\varepsilon^2}$$

Since $\left[D_{\tau}\psi(\tau)|_{\tau=0}\right]' \left[\dot{\eta}(\theta_0)\right]' \dot{\eta}(\theta_0)\, D_{\tau}\psi(\tau)|_{\tau=0}$ is positive semidefinite, the supremum must occur at the boundary as a maximum,

$$\begin{aligned} M_{\varepsilon}^{\eta}(p_0) &= n \sup_{\{\tau : \tau'\tau = \varepsilon^2\}} \frac{\tau' \left[D_{\tau}\psi(\tau)|_{\tau=0}\right]' \left[\dot{\eta}(\theta_0)\right]' \dot{\eta}(\theta_0)\, D_{\tau}\psi(\tau)|_{\tau=0}\, \tau + O\left(\delta^3\right)}{\varepsilon^2} \\ &= n \max_{\tau'\tau = 1} \tau' \left[D_{\tau}\psi(\tau)|_{\tau=0}\right]' \left[\dot{\eta}(\theta_0)\right]' \dot{\eta}(\theta_0)\, D_{\tau}\psi(\tau)|_{\tau=0}\, \tau + \frac{O\left(\varepsilon^3\right)}{\varepsilon^2} \\ &= n\lambda_M \left[\left[D_{\tau}\psi(\tau)|_{\tau=0}\right]' \left[\dot{\eta}(\theta_0)\right]' \dot{\eta}(\theta_0)\, D_{\tau}\psi(\tau)|_{\tau=0}\right] + \frac{O\left(\varepsilon^3\right)}{\varepsilon^2} \\ &= n\lambda_M \left[\dot{\eta}(\theta_0)\, D_{\tau}\psi(\tau)|_{\tau=0} \left[D_{\tau}\psi(\tau)|_{\tau=0}\right]' \left[\dot{\eta}(\theta_0)\right]'\right] + \frac{O\left(\varepsilon^3\right)}{\varepsilon^2} \end{aligned}$$

Note that

$$\begin{aligned} D^2\delta_{\theta_0}(\psi(\tau))\big|_{\tau=0} &= [D\psi(\tau)|_{\tau=0}]' D^2\delta_{\theta_0}(\theta)\big|_{\theta=0} [D\psi(\tau)|_{\tau=0}] \\ &= [D\psi(\tau)|_{\tau=0}]' G(\theta_0) [D\psi(\tau)|_{\tau=0}] \\ &= I_p \end{aligned}$$

so that

$$G(\theta_0)^{-1} = D\psi(\tau)|_{\tau=0} [D\psi(\tau)|_{\tau=0}]'$$

and

$$M_\varepsilon^\eta(p_0) = n\lambda_M \left[\dot{\eta}(\theta_0) G(\theta_0)^{-1} [\dot{\eta}(\theta_0)]'\right] + \frac{O(\varepsilon^3)}{\varepsilon^2}.$$

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

A.2 Proof of Corollary 1

Suppose we define $\theta = \phi(\tau)$ so that $\theta_0 = \phi(\tau_0)$. The metric $G(\theta_0)$ changes to $\left[D_\tau\phi(\tau)|_{\tau=\tau_0}\right]' G(\theta_0) D_\tau\phi(\tau)|_{\tau=\tau_0}$ where $D_\tau\phi(\tau)|_{\tau=\tau_0}$ is a $p \times p$ nonsingular matrix. Moreover, $D_\tau\eta(p(x;\phi(\tau)))|_{\tau=\tau_0} = D_\theta\eta(p(x;\theta))|_{\theta=\theta_0} D_\tau\theta(\tau)|_{\tau=\tau_0}$.

A.3 Proof of Corollary 2

This is a special case of the results in Corollary 1

A.4 Proof of Theorem 2

The proof is very similar to the proof of Theorem 1 and is thus omitted

A.5 Proof of Theorem 3

Let

$$\begin{aligned}
\bar{\mu}_{\varepsilon}^{\eta}(p_0) &= \sup_{\tilde{\eta}'\tilde{\eta}\leq\left(\varepsilon/\sqrt{n}\right)^2} \frac{2\bar{\delta}(\eta_0,\eta_0+\eta)}{\varepsilon^2} \\
&= \sup_{\tilde{\eta}'\tilde{\eta}\leq\left(\varepsilon/\sqrt{n}\right)^2} \frac{\eta'\bar{G}_{11}(\eta_0,\phi_0)\eta + O\left(\delta^3\right)}{\varepsilon^2} \\
&= \frac{1}{n}\max_{v'v=1} v'\bar{G}_{11}(\eta_0,\phi_0)v + \frac{O(\varepsilon^3)}{n\varepsilon^2} \\
&= \frac{1}{n}\lambda_M\left(\bar{G}_{11}(\eta_0,\phi_0)\right) + \frac{O(\varepsilon^3)}{n\varepsilon^2}
\end{aligned}$$

and the result follows by taking the limit as ε goes to zero and taking the square root.

A.6 Proof of Corollary 3

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

B Proofs of results in Section 3

B.1 Proof of Theorem 4

Note that locally

$$\begin{aligned}
&(\eta(\theta_0+\psi(\tau,\phi))-\eta(\theta_0))'(\eta(\theta_0+\psi(\tau,\phi))-\eta(\theta_0)) \\
=\ &\tau'A_{11}\tau + 2\tau'(A_{12}+A_{12}')\phi + \phi'A_{22}\phi
\end{aligned}$$

so

$$\begin{aligned}
M_{\tau\varepsilon}^{\eta}(p_0) &= \sup_{\tau'\tau\leq\varepsilon^2} \frac{\tau' A_{11}\tau + O\left(\delta^3\right)}{\left(\varepsilon/\sqrt{n}\right)^2} = \lambda_M\left(nA_{11}\right) \\
M_{\phi\varepsilon}^{\eta}(p_0) &= \sup_{g(\phi)\leq\varepsilon^{2m+2}} \frac{\phi' A_{22}\phi + O\left(\varepsilon^3\right)}{\left(\varepsilon/n^{m+1}\right)^2} \\
&= n^{(m+1)2} \sup_{g(\varepsilon v)\leq\frac{1}{(2m+2)!}\varepsilon^{2m+2}} v' A_{22} v + \frac{O\left(\varepsilon^3\right)}{\varepsilon^2} \\
&= n^{(m+1)2} \sup_{\frac{1}{i_1!i_2!\ldots i_{2m+2}!}\varepsilon^{2m+2} g_{i_1 i_2 \ldots i_{2m+2}} v^{i_1} v^{i_2}\ldots v^{i_{2m+2}} + O(\varepsilon^{2m+3}) \leq \frac{1}{(2m+2)!}\varepsilon^{2m+2}} v' A_{22} v + \frac{O\left(\varepsilon^3\right)}{\varepsilon^2} \\
&= n^{(m+1)2} \sup_{\frac{1}{i_1!i_2!\ldots i_{2m+2}!} g_{i_1 i_2 \ldots i_{2m+2}} v^{i_1} v^{i_2}\ldots v^{i_{2m+2}} \leq 1 + O(\varepsilon)} v' A_{22} v + \frac{O\left(\delta^3\right)}{\delta^2} \\
&= n^{m+1} \max_{\frac{1}{i_1!i_2!\ldots i_{2m+2}!} g_{i_1 i_2 \ldots i_{2m+2}} v^{i_1} v^{i_2}\ldots v^{i_{2m+2}} = 1 + O(\varepsilon)} v' A_{22} v + \frac{O\left(\varepsilon^3\right)}{\varepsilon^2}.
\end{aligned}$$

Moreover

$$\begin{aligned}
\bar{\mu}_{\tau\varepsilon}^{\eta}(p_0) &= \sup_{\tau' A_{11}\tau\leq\left(\varepsilon/\sqrt{n}\right)^2} \frac{\tau'\tau}{\varepsilon^2} = \frac{1}{n} \max_{\tau' A_{11}\tau = 1} \tau'\tau = \frac{1}{n}\lambda_M\left(A_{11}^{-1}\right) \\
\bar{\mu}_{\phi\varepsilon}^{\eta}(p_0) &= \sup_{\phi' A_{22}\phi\leq\left(\varepsilon/n^{m+1}\right)^2} \frac{g(\phi)}{\varepsilon^{2m+2}} \\
&= \frac{1}{n^{(m+1)2}} \max_{v' A_{22} v = 1} \frac{g(\varepsilon v)}{\varepsilon^{2m+2}} \\
&= \frac{1}{n^{(m+1)2}} \max_{v' A_{22} v = 1} \frac{1}{i_1! i_2! \ldots i_{2m+2}!} g_{i_1 i_2 \ldots i_{2m+2}} v^{i_1} v^{i_2}\ldots v^{i_{2m+2}} + O(\varepsilon)
\end{aligned}$$

All results follow by taking the squares root of the limits of the quantities obtained above.

C Proofs of results in Section 5

C.1 Proof of Lemma 2

The likelihood for the simultaneous equation model can be written as

$$\begin{aligned}
l &= -\frac{(n+1)T}{2}\ln(2\pi) - \frac{T}{2}\ln|\Omega| - \frac{1}{2}\operatorname{tr}\left\{\Omega^{-1} V' V\right\} \\
V &= \Theta - Z_1\Phi + Z_2\Pi_2\left(\beta, I_n\right).
\end{aligned}$$

Minus the expected value of the second differential is

$$\begin{aligned} -E\left(d^2 l\right) &= \frac{T}{2}\left(dv\left(\Omega\right)\right)' D_{n+1}'\left(\Omega^{-1}\otimes\Omega^{-1}\right) D_{n+1} dv\left(\Omega\right) \\ &\quad + \operatorname{vec}\left(dV\right)'\left(\Omega^{-1}\otimes I_T\right)\operatorname{vec}\left(\left(dV\right)'\right). \end{aligned}$$

Now

$$-\operatorname{vec}\left(dV\right) = \left(\begin{array}{ccc} e_{n+1}\otimes Z_2\Pi_2 & \left(\beta, I_n\right)'\otimes Z_2 & I_{n+1}\otimes Z_1 \end{array}\right)\left(\begin{array}{c} d\beta \\ d\operatorname{vec}\left[\Pi_2\right] \\ d\operatorname{vec}\left[\Phi\right] \end{array}\right)$$

where e_{n+1} is a $n+1\times 1$ vector for which all components are zero apart from the element in position 1 which is 1, i.e. $e_{n+1} = (1,0,...,0,0)'$. So the Fisher information matrix

$$G = \left(\begin{array}{ccc} e_{n+1}'\Omega^{-1}e_{n+1}\Pi_2' Z_2 Z_2 \Pi_2 & B' & 0 \\ B & A & 0 \\ 0 & 0 & \frac{T}{2} D_{n+1}'\left(\Omega^{-1}\otimes\Omega^{-1}\right) D_{n+1} \end{array}\right).$$

where

$$\begin{aligned} B &= \left(\begin{array}{c} \left(\beta, I_n\right)\Omega^{-1}e_{n+1}\otimes Z_2' Z_2\Pi_2 \\ \Omega^{-1}e_{n+1}\otimes Z_1' Z_2\Pi_2 \end{array}\right) \\ A &= \left(\begin{array}{cc} \left(\beta, I_n\right)\Omega^{-1}\left(\beta, I_n\right)'\otimes Z_2' Z_2 & \left(\beta, I_n\right)\Omega^{-1}\otimes Z_2' Z_1 \\ \Omega^{-1}\left(\beta, I_n\right)'\otimes Z_1' Z_2 & \Omega^{-1}\otimes Z_1' Z_1 \end{array}\right). \end{aligned}$$

The partial Fisher information about β is

$$\begin{aligned} G\left(\beta : \Pi_2, \Phi, \Omega\right) &= e_{n+1}'\Omega^{-1}e_{n+1}\Pi_2' Z_2 Z_2\Pi_2 \\ &\quad - \left(\begin{array}{cc} B' & 0 \end{array}\right)\left(\begin{array}{cc} A & 0 \\ 0 & \frac{T}{2} D_{n+1}'\left(\Omega^{-1}\otimes\Omega^{-1}\right) D_{n+1} \end{array}\right)^{-1}\left(\begin{array}{c} B \\ 0 \end{array}\right) \\ &= e_{n+1}'\Omega^{-1}e_{n+1}\Pi_2' Z_2 Z_2\Pi_2 - B' A^{-1} B. \end{aligned}$$

To simplify this expression, let

$$C = \left(\begin{array}{cc} I_n & -\left(\beta, I_n\right)\otimes Z_2' Z_1\left(Z_1' Z_1\right)^{-1} \\ 0 & I_{n+1} \end{array}\right)$$

and note that

$$CAC' = \left(\begin{array}{cc} \left(\beta, I_n\right)\Omega^{-1}\left(\beta, I_n\right)'\otimes Z_2' M_{Z_1} Z_2 & 0 \\ 0 & \Omega^{-1}\otimes Z_1' Z_1 \end{array}\right)$$

so that

$$A^{-1} = C' \begin{pmatrix} \left[(\beta, I_n)\,\Omega^{-1}\,(\beta, I_n)'\right]^{-1} \otimes (Z_2' M_{Z_1} Z_2)^{-1} & 0 \\ 0 & \Omega \otimes (Z_1' Z_1)^{-1} \end{pmatrix} C.$$

It is easy to check that

$$CB = \begin{pmatrix} (\beta, I_n)\,\Omega^{-1} e_{n+1} \otimes Z_2' M_{Z_1} Z_2 \Pi_2 \\ \Omega^{-1} e_{n+1} \otimes Z_1' Z_2 \Pi_2 \end{pmatrix}$$

and that

$$\begin{aligned} & B'C' \begin{pmatrix} \left[(\beta, I_n)\,\Omega^{-1}\,(\beta, I_n)'\right]^{-1} \otimes (Z_2' M_{Z_1} Z_2)^{-1} & 0 \\ 0 & \Omega \otimes (Z_1' Z_1)^{-1} \end{pmatrix} CB \\ = & \left\{ e_{n+1}' \Omega^{-1}\,(\beta, I_n)' \left[(\beta, I_n)\,\Omega^{-1}\,(\beta, I_n)'\right]^{-1} (\beta, I_n)\,\Omega^{-1} e_{n+1} \right\} \Pi_2' Z_2' M_{Z_1} Z_2 \Pi_2 \\ & + e_{n+1}' \Omega^{-1} e_{n+1} \Pi_2 Z_2' Z_1 \left(Z_1' Z_1\right)^{-1} Z_1' Z_2 \Pi_2 \end{aligned}$$

So $G(\beta : \Pi_2, \Phi, \Omega)$ equals

$$\left(e_{n+1}' \Omega^{-1} e_{n+1} - e_{n+1}' \Omega^{-1}\,(\beta, I_n)' \left[(\beta, I_n)\,\Omega^{-1}\,(\beta, I_n)'\right]^{-1} (\beta, I_n)\,\Omega^{-1} e_{n+1} \right) \Pi_2' Z_2 M_{Z_1} Z_2 \Pi_2.$$

Now note that, after some tedious calculation, we have

$$\begin{aligned} (\beta, I_n)\,\Omega^{-1} e_{n+1} &= \omega_{11.2}^{-1/2} \Omega_{22}^{-1/2} \beta^*, \\ e_{n+1}' \Omega^{-1} e_{n+1} &= \omega_{11.2}^{-1}, \\ (\beta, I_n)\,\Omega^{-1}\,(\beta, I_n)' &= \Omega_{22}^{-1/2}\left(I_n + \beta^* \beta^{*\prime}\right) \Omega_{22}^{-1/2}, \\ 1 - \beta^{*\prime}\left(I_n + \beta^* \beta^{*\prime}\right)^{-1} \beta^* &= \frac{1}{1 + \beta^{*\prime} \beta^*}, \end{aligned}$$

so that $G(\beta : \Pi_2, \Phi, \Omega)$ can be simplified to the form given in the statement of the Lemma.

C.2 Proof of Theorem 5

Minus the expected value of the second differential of the likelihood is as in Theorem 3, with $V = (y_1, Y_2) - Z_1\Phi - Z_2(\pi_1, \Pi_2)$, so that $dV = -Z_1(d\Phi) - Z_2(d\pi_1, d\Pi_2)$ and

$$\operatorname{vec}(dV) = -\left(I_{n+1} \otimes Z_1, I_{n+1} \otimes Z_2\right) \begin{pmatrix} \operatorname{vec}(d\Phi) \\ \operatorname{vec}(d\Pi) \end{pmatrix}.$$

The Fisher information matrix for the multivariate linear model (3) is

$$T^{-1}G = \begin{pmatrix} \Omega^{-1} \otimes T^{-1}Z_1'Z_1 & \Omega^{-1} \otimes T^{-1}Z_1'Z_2 & 0 \\ \Omega^{-1} \otimes T^{-1}Z_2'Z_1 & \Omega^{-1} \otimes T^{-1}Z_2'Z_2 & 0 \\ 0 & 0 & \frac{1}{2}T^{-1}D_{n+1}'\left(\Omega^{-1} \otimes \Omega^{-1}\right)D_{n+1} \end{pmatrix}$$

where D_{n+1} is the duplication matrix (Magnus and Neudecker (1988), p 48-50). We will now focus on the case $n = 1$. To simplify the notation let $A = T^{-1}Z_2'M_{Z_1}Z_2$. Then, in order to find $D\left(\hat{\beta}_{TSLS}\right)$ we need to evaluate the differential.

$$\begin{aligned} d\hat{\beta}_{TSLS} &= \left[\frac{\hat{\Pi}_2'A}{\hat{\Pi}_2'A\hat{\Pi}_2}, \frac{\left(\hat{\pi}_1' - 2\hat{\Pi}_2'\hat{\beta}_{TSLS}\right)A}{\hat{\Pi}_2'A\hat{\Pi}_2}\right]\begin{pmatrix} d\hat{\pi}_1 \\ d\hat{\Pi}_2 \end{pmatrix} \\ &= \left[\frac{\hat{\Pi}_2'A}{\hat{\Pi}_2'A\hat{\Pi}_2}, \frac{\left(\hat{\pi}_1' - 2\hat{\Pi}_2'\hat{\beta}_{TSLS}\right)A}{\hat{\Pi}_2'A\hat{\Pi}_2}\right]\mathrm{vec}\left(d\Pi\right), \end{aligned}$$

and TG^*_{TSLS} follows from tedious but straightforward simplification. For the OLS estimator we have

$$\hat{\beta}_{OLS} = \left(\Omega_{22} + \Pi_2'A\Pi_2\right)^{-1}\left(\omega_{21} + \Pi_2'A\pi_1\right),$$

and the differential is

$$\begin{aligned} d\hat{\beta}_{OLS} &= \left[\frac{\Pi_2'A}{\Omega_{22} + \Pi_2'A\Pi_2}, \frac{\left(\pi_1' - 2\hat{\beta}_{OLS}\Pi_2'\right)A}{\Omega_{22} + \Pi_2'A\Pi_2}, \frac{1}{\Omega_{22} + \Pi_2'A\Pi_2}, \frac{-\hat{\beta}_{OLS}}{\Omega_{22} + \Pi_2'A\Pi_2}\right]\begin{pmatrix} d\pi_1 \\ d\Pi_2 \\ d\omega_{21} \\ d\Omega_{22} \end{pmatrix} \\ &= \left(\Omega_{22} + \Pi_2'A\Pi_2\right)^{-1}\left[\Pi_2'A, \left(\pi_1' - 2\hat{\beta}_{OLS}\Pi_2'\right)A, 1, -\hat{\beta}_{OLS}\right]\begin{pmatrix} d\pi_1 \\ d\Pi_2 \\ d\omega_{21} \\ d\Omega_{22} \end{pmatrix}. \end{aligned}$$

Note that since we are assuming $n = 1$ we have

$$vec\left(\Omega\right) = \begin{pmatrix} \omega_{11} \\ \omega_{12} \\ \omega_{12} \\ \Omega_{22} \end{pmatrix}, v\left(A\right) = \begin{pmatrix} \omega_{11} \\ \omega_{12} \\ \Omega_{22} \end{pmatrix}, D_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

so that

$$\left(\frac{1}{2}D_2'\left(\left(\Omega^{-1} \otimes \Omega^{-1}\right)\right)D_2\right)^{-1} = \begin{pmatrix} 2\omega_{11}^2 & 2\omega_{11}\omega_{12} & 2\omega_{12}^2 \\ 2\omega_{11}\omega_{12} & \omega_{11}\Omega_{22} + \omega_{12}^2 & 2\omega_{12}\Omega_{22} \\ 2\omega_{12}^2 & 2\omega_{12}\Omega_{22} & 2\Omega_{22}^2 \end{pmatrix}.$$

The components of G^{-1} that we need are

$$\begin{pmatrix} \Omega \otimes A^{-1} & 0 & 0 \\ 0 & \omega_{11}\Omega_{22} + \omega_{12}^2 & 2\omega_{12}\Omega_{22} \\ 0 & 2\omega_{12}\Omega_{22} & 2\Omega_{22}^2 \end{pmatrix}$$

Then G^*_{OLS} follows from a tedious but straightforward simplification.

Table 1: Measures of sensitivity for the coefficient of return to education for men born 1930-1939

	(1) OLS	(2) TSLS	LIML	(3) OLS	(4) TSLS	LIML	(5) OLS	(6) TSLS	LIML	(7) OLS	(8) TSLS	LIML
Years of Education	0.0710	0.0891	0.0929	0.0711	0.0760	0.0811	0.0632	0.0806	0.0838	0.0632	0.0600	0.0574
$M^{\beta}(\hat{p}_0)$	9.2017	9.2471	9.2653	16.6208	16.6262	16.6431	9.3680	9.7079	9.4200	16.6375	16.6351	16.6425
$\mu^{\beta}(\hat{p}_0) = 1/M^{\beta}(\hat{p}_0)$	0.1087	0.1081	0.1079	0.0602	0.0601	0.0601	0.1067	0.1030	0.1062	0.0601	0.06011	0.0601
$M^{\hat{\beta}}$	0.2045	10.0143	-	0.2046	20.4401	-	0.2019	10.1201	-	0.2019	19.9175	-
$\mu^{\hat{\beta}}(\hat{p}_0) = 1/M^{\hat{\beta}}(\hat{p}_0)$	4.8900	0.0999	-	4.8876	0.0489	-	4.9529	0.0988	-	4.9529	0.0502	-
Race	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
SMSA	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Married	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
9 Year-of-birth dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
8 Region-of-residence dummies	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Age	No	No	No	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes
Age-squared	No	No	No	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes

Note: Calculated from the 5% PUMS of the 1980 U.S. census. The sample size is 329509. Age and age-squared are measured in quarters of years. Each equation includes an intercept. The dependent variable is the log of weekly earnings. The instruments are quarter-of-birth dummies and and quarter-of-birth times year-of-birth interactions. The dataset is described in Angrist and Krueger (1991).