Econometric Theory, **18**, 2002, 140–168. Printed in the United States of America.

OPTIMAL INFERENCE WITH MANY INSTRUMENTS

JINYONG HAHN
Brown University

In this paper, I derive the efficiency bound of the structural parameter in a linear simultaneous equations model with many instruments. The bound is derived by applying a convolution theorem to Bekker's (1994, *Econometrica* 62, 657–681) asymptotic approximation, where the number of instruments grows to infinity at the same rate as the sample size. Usual instrumental variables estimators with a small number of instruments are heuristically argued to be efficient estimators in the sense that their asymptotic distribution is minimal. Bayesian estimators based on parameter orthogonalization are heuristically argued to be inefficient.

1. INTRODUCTION

Recently, a renewed interest in instrumental variables (IV) has taken place in the literature. The modern interest can be partly ascribed to the "natural experiment" based empirical analysis by Angrist and Krueger (1991), who use a two stage least squares estimator (2SLS) based on 180 quarter-of-birth and state-of-birth variables as instruments, and Bound, Jaeger, and Baker's (1995) finding that it is possible to come up with similar numbers as in Angrist and Krueger (1991) even if artificial random variables generated by the computer are used as instruments. It can also be ascribed to the theoretical finding by Nelson and Startz (1990) and Maddala and Jeong (1992) that approximation of sampling distributions of conventional IV estimators by asymptotic normal distributions can be quite poor. Such findings led researchers to investigate nonstandard approximations to sampling distributions. As Chamberlain and Imbens (1996) point out, recent theoretical econometric literature on the IV estimator can be categorized into two strands. The first one focuses on the statistical properties of the IV estimator when the instrument is "weak." Staiger and Stock (1997) propose an asymptotic approximation where the correlation between instrument and endogenous regressor vanishes as a function of sample size, based on which they explain sampling properties of IV estimators. See also Dufour (1997), Hall, Rudebusch, and Wilcox (1996), Quintos (1997), Zivot, Startz, and Nelson (1997), and Wang and Zivot (1997). The second strand focuses on the case where the

Helpful comments by Donald Andrews, Joshua Angrist, Gary Chamberlain, Jack Porter, two anonymous referees, and seminar participants of Brown University, Michigan State University, University of Pennsylvania, and Princeton University are appreciated. Financial support was provided by the Research Foundation of the University of Pennsylvania. Address correspondence to: Jinyong Hahn, Department of Economics, Brown University, Box B, Providence, RI 02912, USA.

 © 2002 Cambridge University Press 0266-4666/02 $9.50

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

number of instruments is "large." Bekker (1994) proposes an asymptotic approximation where the number of instruments grows at the same rate as the sample size, and he finds evidence that such an approximation provides a good explanation of the sampling properties of IV estimators even when the number of instruments is relatively small. He finds that 2SLS is inconsistent but limited information maximum likelihood (LIML) is consistent under the nontraditional asymptotic approximation.

Despite numerous research projects on the question of the quality of asymptotic approximations to sampling distributions, corresponding research on the question of efficiency has been conspicuously absent in the literature. Most IV estimators in the literature have originally been developed to be "efficient" under traditional asymptotic approximation. Under traditional asymptotic approximation, their finite sample distributions should approximate a certain target distribution, say, a normal distribution with mean zero and variance equal to some efficiency bound. If the quality of traditional approximation is poor as argued by the recent literature, it is not clear whether the traditional efficiency bound provides a relevant point of reference, either.

The purpose of this paper is rather narrow in scope. I derive the efficiency bound when an increasing number of instruments is available. For this purpose, I adopt Bekker's (1994) alternative asymptotic approximation and examine the notion of asymptotic optimality when the number of instruments is "large." I assume that the number of instruments grows to infinity as a linear function of sample size. I then establish a Hajék-type convolution theorem and asymptotic minimax theorem. This is not a standard argument, which would be based on the sample size growing to infinity with the data generating process (DGP) fixed. The traditional (parametric, semiparametric, and nonparametric) approach assumes that data consist of independent and identically distributed (i.i.d.) observations, whose common underlying density is fixed even when sample size grows to infinity. The triangular nature of the DGP makes it impossible to apply traditional arguments. Instead, I utilize the notion developed by Le Cam (1986), Le Cam and Yang (1990), and van der Vaart (1994). See Phillips (1989) or Jeganathan (1995) for recent econometric applications.

The efficiency bound developed in this paper turns out to be isomorphic to that in the traditional parametric model, which, in turn, is isomorphic to the semiparametric efficiency bound developed by Chamberlain (1987). The parametric asymptotic bound is equal to the Cramér–Rao bound in linear simultaneous equations models with normal errors, whereas a semiparametric asymptotic bound with infinite number of instruments can be found in Chamberlain's paper (1987). Intuition suggests that a Chamberlain-type bound should be applicable when the number of instruments is finite but large relative to the sample size, but a formal result for this situation has been absent from the literature. This paper provides a link between the traditional parametric approach and the recent semiparametric approach by considering a case where the number of instruments is finite yet large.

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

The efficiency bound developed in this paper is invariant to the rate at which the number of instruments grows. Such intuition suggests that any estimator whose asymptotic distribution depends on the asymptotic ratio of the number of instruments to the sample size cannot be asymptotically optimal. Bekker (1994) observes that, although a LIML that uses all available instruments is consistent and asymptotically normal, its asymptotic distribution depends on the ratio. Therefore, it follows that LIML is inefficient.

The efficiency bound also reveals that an important Bayesian procedure cannot produce an asymptotically efficient estimator. Recently, a Bayesian innovation emerged for inference in models with a large number of parameters. Lancaster (1997) considers a panel model and points out that a diffuse prior on suitably reparametrized fixed effects eliminates the incidental parameter problem. His technique can be applied to the simultaneous equations model with many instruments. It can then be asked whether the posterior distribution is fully informative. Instead of investigating a particular Bayes estimator based on such a posterior, we may compare the asymptotic variance bound derived in this paper with the smallest possible asymptotic variance of estimators that can be expressed as functions of the sufficient statistic in the posterior. It turns out that the latter bound is strictly larger than the bound derived in this paper: Bayes estimator based on diffuse prior on reparametrized model cannot be asymptotically efficient.

In Section 2, I consider the likelihood ratio process of the simultaneous equations model of interest and derive the convolution theorem. In Section 3, I heuristically argue that a Bayesian estimator based on parameter orthogonalization cannot be optimal. Section 4 concludes. All technical arguments, including background discussion of the convolution theorem, are contained in Appendixes A–E.

2. OPTIMAL INFERENCE WITH MANY INSTRUMENTS

In this section, I formally discuss the simple simultaneous equations model of interest and develop a corresponding efficiency bound. To simplify notation, I only consider a model with one endogenous regressor and no exogenous regressors. As discussed later in this section, the model with many instruments provides a link between the traditional model with a finite number of instruments and a semiparametric model with an infinite number of instruments. Therefore, it is entirely plausible that the efficiency bound of such a model can be computed by a strategy similar to that used in computation of semiparametric efficiency bounds. Theorem 4 in Appendix A provides such a strategy.

2.1. Convolution Theorem

Consider a linear simultaneous equations model with one endogenous regressor and no exogenous regressors. Suppose that

$$y_i^n = x_i^n \beta + u_i^n, \qquad i = 1, \ldots, n,$$

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

where we assume $\dim(x_i^n) = 1$ for simplicity. We assume that x_i^n may be random. Now assume that we are given a deterministic sequence of instruments $z_i^n \equiv (z_{1i}^n, \ldots, z_{Ki}^n)'$ such that $E[x_i^n] = z_i^{n'} \Pi^n$. We assume that $E[u_i^n] = 0$ and $\mathrm{Var}(u_i^n) = \sigma^2$. We also assume that $\dim(z_i^n) = \dim(\Pi^n) = K$. We further assume that $(y_i^n, x_i^n)'$ $i = 1, \ldots, n$ are normal such that

$$(y_i^n, x_i^n) = z_i^{n'} \Pi^n(\beta, 1) + (v_{1i}^n, V_{2i}^n)$$

with $(v_{1i}^n, V_{2i}^n)'$ $i = 1, \ldots, n$ i.i.d. normal with zero mean and covariance matrix Ω:

$$v_i^n \equiv \begin{pmatrix} v_{1i}^n \\ V_{2i}^n \end{pmatrix} \sim \mathcal{N}(0, \Omega) = \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{bmatrix} \right).$$

I will assume that the number of instruments K is large. To make this statement precise, I will assume that $K = K^n \to \infty$ as $n \to \infty$. This asymptotic approximation is expected to be helpful when the number of instruments is truly large. But there is another reason why we may want to entertain this alternative asymptotic approximation. Bekker (1994) argues that $K \to \infty$ such that $K/n \to \alpha$ for some $0 < \alpha < 1$ asymptotics provides a superior approximation of the sampling properties of various IV estimators even when K is reasonably small. As in Bekker (1994), I thus assume that the parameter sequence is such that β and Ω are kept fixed, $(1/n)\sum_{i=1}^n (z_i^{n'} \Pi^n)^2$ is convergent but $\lim_{n\to\infty} K/n = \alpha$. Observe that we are now dealing with a triangular system of DGP. As a technical condition, I also require that

$$\frac{1}{n} \max_{1 \le i \le n} z_i^{n'} z_i^n \to 0.$$

Remark 1. The assumption that $(1/n)\sum_{i=1}^n (z_i^{n'} \Pi^n)^2$ is convergent is a high-level assumption. One obvious primitive condition is a series-expansion-type condition where Z_i is a random variable bounded by one in absolute value, $\sum_{k=0}^\infty \theta_k Z_i^k$ converges to $\vartheta(Z_i)$ sufficiently fast in the mean square sense, and $z_i^{n'} \Pi^n$ denotes $\sum_{k=0}^K \theta_k Z_i^k$.

THEOREM 1. *For every regular sequence of estimators $\{T_n\}$ of β, we have*

$$\sqrt{n}(T_n - \beta(n)) \Rightarrow \mathcal{N}\left(0, \frac{\sigma^2}{\Theta(\Pi, \Pi)}\right) \oplus \mathcal{W}$$

for some distribution $\mathcal{W}$ on the real line, where $\Theta(\Pi, \Pi) \equiv \lim(1/n) \times \sum_{i=1}^n (z_i^{n'} \Pi^n)^2$ and $\Pi \equiv (\Pi^1, \Pi^2, \ldots)$ denotes the sequence of Π^n's.

Proof. See Appendix C.

Theorem 1 suggests that the asymptotic variance of a sequence of efficient estimators is $\sigma^2/\Theta(\Pi, \Pi)$,[1] which does not depend on α. This has an important

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

practical implication. Bekker (1994) shows that LIML has a desirable property compared to 2SLS when the number of instruments is large by arguing that LIML is consistent whereas 2SLS is not under the alternative approximation. He also shows that the asymptotic distribution of LIML depends on α. We may therefore come to the practical conjecture that even LIML cannot be an optimal estimator when the number of instruments is large.[2]

2.2. Discussion

To put the asymptotic variance bound

$$\frac{\sigma^2}{\Theta(\Pi,\Pi)} = \frac{\mathrm{Var}(u_i^n)}{\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(x_i^n - V_{2i}^n)^2} \tag{1}$$

in Theorem 1 into perspective, consider two extreme cases.

In the first case, K is kept fixed so that the model is reduced to the usual finite-dimensional parametric model. Observe that the right side of the expression (1) is equal to that implied by the Cramér–Rao lower bound: an optimal estimator in our particular parametric model with many nuisance parameters should have the same asymptotic property as in a similar parametric model with a finite, fixed number of nuisance parameters.

In the second case, K is kept equal to ∞. One particular situation when this is valid is the case

$$x_i^n = f(Z_i) + V_{2i}^n,$$

where Z_i is a finite-dimensional random variable, and

$$f(Z_i) = \sum_{k=1}^{\infty} \pi_k \rho_k(Z_i) \equiv \sum_{k=1}^{\infty} \pi_k z_{ki}, \tag{2}$$

where $\{\rho_1(Z_i), \rho_2(Z_i), \ldots\}$ is a linearly independent sequence of functions that forms a complete basis for the $L_2(Z_i)$-space and the infinite series converges in an L_2 sense. Under such circumstances, Chamberlain (1987)[3] has shown that the semiparametric efficiency bound for β is equal to

$$\frac{\mathrm{Var}(u_i^n)}{E[\,f(Z_i)^2\,]} = \frac{\mathrm{Var}(u_i^n)}{E[(x_i^n - V_{2i}^n)^2]} = \frac{\mathrm{Var}(u_i^n)}{\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(x_i^n - V_{2i}^n)^2}.$$

Because Bekker's (1994) asymptotics provide the link between the finite-dimensional parametric model and a semiparametric model, and because both models have the same asymptotic variance bound, it is quite intuitive that the

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

bound (1) takes such a form. The reason why the bound does not depend on α can be understood by the same reasoning.

2.3. Is Optimal Inference Feasible?

Theorem 1 is not interesting unless we can construct an estimator that achieves the bound (1). Here I heuristically argue that it is possible to construct an efficient estimator under some high-level assumptions. I argue that efficient estimation is possible by a 2SLS estimation based on a smaller set of instruments than is available. It is based on the same intuition as that of Newey (1990b), who considers a model given by conditional moment restrictions. He considers a sequence of IV estimators based on an increasing number of instruments and shows that such an estimator would be asymptotically efficient under a certain rate of increase of the number of instruments. It should be noted that the number of instruments used in his procedure grows quite slowly as a function of the sample size. Therefore, it can be conjectured that 2SLS based on a small number of instruments would achieve the bound (1) in Bekker's (1994) model.

In this section and thereafter, I will write

$$X^n = \begin{bmatrix} x_1^n \\ \vdots \\ x_n^n \end{bmatrix}, \qquad Y^n = \begin{bmatrix} y_1^n \\ \vdots \\ y_n^n \end{bmatrix}, \qquad Z^n = \begin{bmatrix} z_1^{n\prime} \\ \vdots \\ z_n^{n\prime} \end{bmatrix},$$

$$U^n = [y^n, X^n], \qquad M^n = Z^n \Pi^n(\beta, 1),$$

so that we may rewrite the model as

$$U^n = M^n + v^n.$$

I will also assume, as Bekker (1994) does, that $(\Pi^{n\prime} Z^{n\prime} Z^n \Pi^n)/n$ is actually fixed at $\Theta(\Pi, \Pi)$.

Subsequently, I will argue that a 2SLS-type estimator is efficient under some high-level regularity condition.

Condition 1. There exists a sequence $\{P^n\}$ of symmetric idempotent $n \times n$ matrices such that

$$\sqrt{n}\left(\frac{\Pi^{n\prime} Z^{n\prime} P^n Z^n \Pi^n}{n} - \Theta(\Pi, \Pi)\right) \to 0, \quad \frac{\operatorname{rank}(P^n)}{\sqrt{n}} \to 0.$$

Condition 1 requires that the projection described by P^n is very close to the projection $\mathbf{P}^n \equiv Z^n(Z^{n\prime} Z^n)^{-1} Z^{n\prime}$. It also requires that the range of the projection P^n is a low-dimensional linear space. An example would be $P^n = Z_r^n (Z_r^{n\prime} Z_r^n)^{-1} Z_r^{n\prime}$, where Z_r^n consists of $r_n = o(\sqrt{n})$ columns of Z^n.

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

PROPOSITION 1. *Under Condition 1, we have*

$$\sqrt{n}\left(\frac{U^{n\prime}P^nU^n_{(1,2)}}{U^{n\prime}P^nU^n_{(2,2)}}-\beta\right)\Rightarrow\mathcal{N}\left(0,\frac{\sigma^2}{\Theta(\Pi,\Pi)}\right). \tag{3}$$

Here, $U^{n\prime}P^nU^n_{(1,2)}$ *and* $U^{n\prime}P^nU^n_{(2,2)}$ *denote* (1,2) *and* (2,2) *elements of the* 2×2 *matrix* $U^{n\prime}P^nU^n$.

Proof. See Appendix D.

It was heuristically argued that IV estimators based on a small number of instruments are efficient under Bekker's (1994) alternative asymptotic approximation. This has an interesting practical implication because it suggests that we may not want to use every instrument even when many instruments are potentially available. This is somewhat counterintuitive because the usual method-of-moments intuition as found in Hansen (1982) or Manski (1988) suggests that an IV estimator based on more instruments should be more efficient. But such intuition may be misleading because the generalized method of moments (GMM) estimators are efficient only for models with finite numbers of unconditional moment restrictions. In other words, the GMM intuition applies only to the case where the number of instruments is small. Bekker's (1994) argument may indicate that the GMM intuition is applicable only when the number of instruments is *very small*.

To understand why we do not want to use a large number of instruments, it is instructive to consider the model with conditional moment restrictions as considered by Chamberlain (1987). It is easy to see that a model with a finite number of conditional moment restrictions can be rewritten as a model with an infinite number of unconditional moment restrictions using every square integrable function of the conditioning variable as instruments. For such models, Newey (1990b) shows that an efficient estimator takes a form of an IV estimator where the set of instruments, i.e., the polynomial functions of conditioning variables, increases *very slowly* as a function of the sample size. See Chamberlain (1992) and Hahn (1997) for similar arguments in panel context.

3. BAYES ESTIMATOR WITH MANY NUISANCE PARAMETERS: PARAMETER ORTHOGONALIZATION

In this section I examine whether an important Bayes procedure can produce an asymptotically efficient estimator. Lancaster (1997) notes that the incidental parameter problem in panel models with fixed effects may be eliminated by reparameterizing the fixed effects parameters in such a way that the cross derivative of the log likelihood with respect to parameter of interest and reparameterized fixed effects would have a zero expectation. Although Lancaster (1997) does not explicitly apply this idea to simultaneous equation models, it is of interest to examine whether such a strategy would result in an efficient estima-

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

tor. I show that any Bayesian estimator based on parameter orthogonalization cannot be asymptotically efficient in Bekker's (1994) model.

3.1. Parameter Orthogonalization

I show that any Bayesian estimator based on parameter orthogonalization is a function of $(\bar{S}, S^{\perp})$, where

$$\bar{S} \equiv U^{n\prime} Z^n (Z^{n\prime} Z^n)^{-1} Z^{n\prime} U^n, \qquad S^{\perp} \equiv U^{n\prime} U^n - \bar{S}.$$

To emphasize the Bayesian nature of the procedure, I will omit the n superscript throughout the rest of this section.

Let π_k and z_{ki} denote the kth components of Π and z_i. Ignoring irrelevant constant, the log likelihood is equal to

$$\begin{aligned} L = {} & \frac{n}{2} \log\det \Psi - \frac{1}{2} \psi_{11} \sum_{i=1}^{n} \left(y_i - \beta \sum_{k=1}^{K} \pi_k z_{ki} \right)^2 \\ & - \psi_{12} \sum_{i=1}^{n} \left(y_i - \beta \sum_{k=1}^{K} \pi_k z_{ki} \right) \left(x_i - \sum_{k=1}^{K} \pi_k z_{ki} \right) \\ & - \frac{1}{2} \psi_{22} \sum_{i=1}^{n} \left(x_i - \sum_{k=1}^{K} \pi_k z_{ki} \right)^2, \end{aligned} \tag{4}$$

where I write

$$\Psi = \Omega^{-1} = \begin{bmatrix} \psi_{11} & \psi_{12} \\ \psi_{12} & \psi_{22} \end{bmatrix}$$

for simplicity. We want to find parameter transformations $\pi_k = \pi_k(\beta, \xi_k)$ such that

$$E\left[\frac{d^2 L}{d\beta d\xi_k} \right] = 0. \tag{5}$$

With some algebra, it can be shown that the reparametrization

$$\pi_k = \Upsilon \xi_k \quad \forall k, \tag{6}$$

where

$$\Upsilon \equiv \frac{1}{\sqrt{\psi_{11}\beta^2 + 2\psi_{12}\beta + \psi_{22}}},$$

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

has the desired property (5). It can be shown that, with a diffuse prior on ξ's, the posterior likelihood for (β, Ψ) is proportional to

$$(\sqrt{\det(\Psi)})^n \exp\left\{-\frac{1}{2}\left[\operatorname{trace}(\Psi(\bar{S}+S^{\perp})) - \frac{(\beta,1)\Psi\bar{S}\Psi(\beta,1)'}{\psi_{11}\beta^2+2\psi_{12}\beta+\psi_{22}}\right]\right\}. \quad (7)$$

See Lemma 11 in Appendix E. Therefore, it follows that $(\bar{S}, S^{\perp})$ is the sufficient statistic of the posterior likelihood.

Remark 2. Parameter orthogonalization may provide some intuition as to why LIML is consistent under Bekker's (1994) asymptotics. It is well known that the incidental parameter problem can cause the maximum likelihood estimator to be inconsistent. As a result of potential incidental parameter problems related to the number of π's, maximum likelihood estimation as a principle cannot explain the consistency of LIML. It can be shown that LIML maximizes (7). This coincidence may explain the consistency of LIML even though the number of nuisance parameters is large.

3.2. Information for β in $(\bar{S}, S^{\perp})$

I now investigate the information content of the sufficient statistic $(\bar{S}, S^{\perp})$ in the posterior likelihood. For this purpose, I borrow an important tool from van der Ploeg and Bekker (1996). Suppose that we are given a finite-dimensional statistic s such that

$$\sqrt{n}(s-\psi) \Rightarrow N(0, \mathbf{\Lambda}), \quad (8)$$

where Λ is invertible. Also suppose that $\psi = \psi(\theta, \eta)$, where θ is a (one-dimensional) parameter of interest and η is a finite-dimensional nuisance parameter. I will define the asymptotic variance bound for θ in s as the inverse of

$$\min_x (1 \quad -x')(F'\mathbf{\Lambda}^{-1}F)(1 \quad -x')', \quad (9)$$

where $F = (\partial\psi/\partial\theta, \partial\psi/\partial\eta')$. The definition can be justified by considering a differentiable function $\hat{\delta}(s)$ of s, such that $\sqrt{n}(\hat{\delta}(s)-\theta)$ is asymptotically normal. Here, $\hat{\delta}(s)$ denotes an arbitrary estimator of θ using only s. It is of interest to ask what the minimum of the asymptotic variance of $\hat{\delta}(s)$ is. Such a minimum would be indicative of the partial information for θ contained in s. Van der Ploeg and Bekker (1996, Theorem 1) show that the minimum asymptotic variance is equal to the inverse of (9).

I now consider the information for β in $\bar{S}$ and $S^{\perp}$. Using Lemma 10 in Appendix E, which derives the asymptotic distribution of $\left((1/n)\bar{S}_{11}, (1/n)\bar{S}_{12}, (1/n)\bar{S}_{22}, (1/n)S^{\perp}_{11}, (1/n)S^{\perp}_{12}, (1/n)aS^{\perp}_{22}\right)'$, we can compute the asymptotic variance bound for β in $(\bar{S}, S^{\perp})$ as follows:

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

THEOREM 2. *The asymptotic variance bound for* β *in* $\left((1/n)\bar{S}_{11},(1/n)\bar{S}_{12},(1/n)\bar{S}_{22},(1/n)S_{11}^{\perp},(1/n)S_{12}^{\perp},(1/n)S_{22}^{\perp}\right)'$ *is equal to*

$$\frac{\alpha \det(\Omega) + (1-\alpha)(\omega_{11} - 2\beta\omega_{12} + \beta^2\omega_{22})\Theta(\Pi,\Pi)}{(1-\alpha)\Theta(\Pi,\Pi)^2}. \tag{10}$$

Comparing the information content in $(\bar{S},S^{\perp})$ with Theorem 1, it is easy to see that an efficient estimator cannot be a function of $(\bar{S},S^{\perp})$: the information content in $\bar{S}$ and $S^{\perp}$ is strictly smaller than the information implicit in the convolution theorem. Van der Ploeg and Bekker (1996) recently calculated the bound of the model with many instruments by focusing on estimators that can be expressed as a function of $(\bar{S},S^{\perp})$. Although their bound is of interest in its own right, Theorem 2 implies that any estimator that can be expressed as a function of $(\bar{S},S^{\perp})$ is necessarily inefficient. Theorem 2 also implies that any Bayesian estimator based on parameter orthogonalization is necessarily inefficient. This is because parameter orthogonalization leads to a posterior, for which $(\bar{S},S^{\perp})$ is sufficient as discussed in Section 3.1, and therefore, any Bayesian estimator is necessarily a function of $(\bar{S},S^{\perp})$.

Remark 3. Although the asymptotic variance bound

$$\frac{\alpha}{(1-\alpha)}\frac{\det(\Omega)}{\Theta(\Pi,\Pi)^2} + \frac{\mathrm{Var}(u_i^n)}{\Theta(\Pi,\Pi)}$$

in $(\bar{S},S^{\perp})$ clearly exceeds the bound $\mathrm{Var}(u_i^n)/\Theta$ in the convolution theorem, it coincides with the asymptotic variance of LIML as computed by Bekker (1994). This suggests that LIML is efficient in the limited sense that it exhausts information contained in $(\bar{S},S^{\perp})$.

4. SUMMARY

In this paper, I derived the efficiency bound of a structural parameter in a simultaneous equations model with many instruments. IV estimators based on a small number of instruments were heuristically argued to be asymptotically optimal. One deficiency of this paper is that it does not provide a way of optimally choosing a set of instruments in the finite sample. It is a problem that this paper shares with Newey's (1990b) series approximation based efficient IV estimation and Hahn's (1997) similar method applied to models with sequential moment restrictions.

Recently, Donald and Newey (1997) have considered a model where the set of available instruments consists of a countably infinite number of functions of a conditioning variable and suggested a concrete methodology of choosing instruments that would approximately achieve optimality in the mean square error criterion. It is useful to note that their assumed DGP is fixed even when the

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

sample size grows to infinity. Although it is beyond the scope of this paper, it would be interesting to examine the performance of their methodology under Bekker's (1994) alternative asymptotics and to compare it with the asymptotic variance bound developed in this paper.

NOTES

1. This implicitly assumes that there exists a sequence of estimators with asymptotic variance $\sigma^2/\Theta(\Pi,\Pi)$. Otherwise, $\sigma^2/\Theta(\Pi,\Pi)$ would be an infeasible bound. See Section 2.3 for related discussion.

2. Because it is not yet clear whether the bound $\sigma^2/\Theta(\Pi,\Pi)$ can be attained, such a claim can only be a "conjecture."

3. Chamberlain (1987) derives bounds for models with a finite number of conditional moment restrictions. Formally, a finite number of conditional moment restrictions can be equivalently written as a countably infinite number of unconditional moment restrictions. Therefore, Chamberlain's (1987) result may be reinterpreted as the semiparametric efficiency bound when the number of instruments is countably infinite.

REFERENCES

Angrist, J. & A. Krueger (1991) Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics* 106, 979–1014.

Begun, J.M., W.J. Hall, W.M. Huang, & J.A. Wellner (1983) Information and asymptotic efficiency in parametric-nonparametric models. *Annals of Statistics* 11, 432–452.

Bekker, P. (1994) Alternative approximations to the distributions of instrumental variable estimators. *Econometrica* 62, 657–681.

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

Bound, J., D. Jaeger, & R. Baker (1995) Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variables is weak. *Journal of the American Statistical Association* 90, 443–450.

Chamberlain, G. (1987) Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* 34, 305–334.

Chamberlain, G. (1992) Comment: Sequential moment restrictions in panel data. *Journal of Business and Economic Statistics* 10, 20–26.

Chamberlain, G. & G. Imbens (1996) Hierarchical Bayes Models with Many Instrumental Variables. Harvard Institute of Economic Research, Discussion paper 1781.

Donald, S. & W. Newey (1997) Choosing the Number of Instruments. Manuscript.

Dufour, J.-M. (1997) Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica* 65, 1365–1388.

Hahn, J. (1997) Efficient estimation of panel data models with sequential moment restrictions. *Journal of Econometrics* 79, 1–21.

Hall, A.R., D. Rudebusch, & D.W. Wilcox (1996) Judging instrument relevance in instrumental variables estimation. *International Economic Review* 37, 283–298.

Hansen, L. (1982) Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029–1054.

Jeganathan, P. (1995) Some aspects of asymptotic theory with applications to time series models. *Econometric Theory* 11, 818–887.

Lancaster, T. (1997) Orthogonal Parameters and Panel Data. Manuscript.

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

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

Le Cam, L. & G.L. Yang (1990) *Asymptotics in Statistics: Some Basic Concepts*. New York: Springer-Verlag.
Maddala, G.S. & J. Jeong (1992) On the exact small sample distribution of the instrumental variable estimator. *Econometrica* 60, 181–183.
Magnus, J. & H. Neudecker (1988) *Matrix Differential Calculus with Applications in Statistics and Econometrics*. New York: Wiley.
Manski, C. (1988) *Analog Estimation Methods in Econometrics*. New York: Chapman and Hall.
Nelson, C. & R. Startz (1990) Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica* 58, 967–976.
Newey, W.K. (1990a) Semiparametric efficiency bounds. *Journal of Applied Econometrics* 5, 99–135.
Newey, W.K. (1990b) Efficient instrumental variables estimation of nonlinear models. *Econometrica* 58, 809–837.
Phillips, P.C.B. (1989) Partially identified econometric models. *Econometric Theory* 5, 181–240.
Quintos, C. (1997) Weak Instruments and Spurious Regressions. Manuscript.
Staiger, D. & J. Stock (1997) Instrumental variables regression with weak instruments. *Econometrica* 65, 556–586.
Stein, C. (1956) Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. Berkeley: University of California Press.
van der Ploeg, J. & P. Bekker (1996) Efficiency Bounds for Instrumental Variable Estimators under Group-Asymptotics. Manuscript.
van der Vaart, A.W. (1994) Limits of Experiments. Lecture notes.
van der Vaart, A.W. & J.A. Wellner (1996) *Weak Convergence and Empirical Processes with Applications to Statistics*. New York: Springer-Verlag.
Wang, J. & E. Zivot (1997) Inference on Structural Parameters in Instrumental Variables Regression with Weak Instruments. Manuscript.
Zivot, E., R. Startz, & C. Nelson (1997) Valid Confidence Intervals and Inference in the Presence of Weak Instruments. Institute for Economic Research Discussion paper 97-17, University of Washington.

APPENDIX A: CONVOLUTION THEOREM

In this section, I briefly examine the notion of statistical experiments and related asymptotic results as discussed in Le Cam (1986), Le Cam and Yang (1990), van der Vaart (1994), and van der Vaart and Wellner (1996). Whereas Bekker's (1994) attention was focused on finding a model such that properties of various estimators are well approximated, I will use the statistical experiment concept to approximate Bekker's (1994) model itself.

An *experiment* $\mathcal{E} = \{\mathcal{X}, \mathcal{A}, P_h : h \in H\}$ is a collection of probability measures $\{P_h : h \in H\}$ on a measurable space $\{\mathcal{X}, \mathcal{A}\}$, the *sample space*. The parameter set H may be arbitrary. A researcher observes a value $x \in \mathcal{X}$. He does not know the value of h and must select a decision $\rho(x) \in \mathcal{Z}$. (For simplicity of notation, I assume that the decision is not randomized.) If he has chosen z when the true distribution is P_h, he suffers a *loss* $W(z, h)$. His average loss, or the risk, when x is picked according to P_h is the integral $\int W(\rho(x), h) P_h(dx)$. For simplicity of notation, I will simply write an experiment $\mathcal{E} = (P_h : h \in H)$.

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

Now consider a researcher who is contemplating two possible experiments, say, $\mathcal{E} = (P_h : h \in H)$ and $\mathcal{F} = (Q_h : h \in H)$. How can one compare the two experiments? For this purpose, we need two definitions.

DEFINITION 1 (Le Cam and Yang, 1990, p. 6). *$\mathcal{R}(\mathcal{E}, W)$ denotes the set of functions $R(h)$ of h such that $R(\cdot) \in \mathcal{R}(\mathcal{E}, W)$ if and only if there exists some decision procedure $\rho(\cdot)$ with $\int W(\rho(x), h) P_h(dx) \leq R(h)$ for all $h \in H$.*

DEFINITION 2 (Le Cam and Yang, 1990, p. 7). *The deficiency $\delta(\mathcal{E}, \mathcal{F})$ of $\mathcal{E}$ with respect to $\mathcal{F}$ is the smallest number $\epsilon \in [0,1]$ such that, for every loss function W with $0 \leq W(z,h) \leq 1$ and every $R_2 \in \mathcal{R}(\mathcal{F}, W)$, there exists an $R_1 \in \mathcal{R}(\mathcal{E}, W)$ with $R_1(h) \leq R_2(h) + \epsilon$.*

Equipped with the two definitions, we define the (pseudo)distance $\Delta(\mathcal{E}, \mathcal{F})$ between $\mathcal{E}$ and $\mathcal{F}$ as follows.

DEFINITION 3 (Le Cam and Yang, 1990, p. 7). $\Delta(\mathcal{E}, \mathcal{F}) = \max[\delta(\mathcal{E}, \mathcal{F}), \delta(\mathcal{F}, \mathcal{E})]$.

The distance $\Delta(\mathcal{E}, \mathcal{F})$ is defined in such a way that, if only loss functions W with $0 \leq W(z,h) \leq 1$ are used, any decision we can make in one of the two experiments can be made in the other within $\Delta(\mathcal{E}, \mathcal{F})$. We say that two experiments $\mathcal{E}$ and $\mathcal{F}$ are *equivalent* if $\Delta(\mathcal{E}, \mathcal{F}) = 0$. Now, suppose that we are given a sequence of experiments $\{\mathcal{E}_n : n = 1,2,\ldots\} = \{(P_{n,h} : h \in H) : n = 1,2,\ldots\}$. If there exists an experiment $\mathcal{E}$ such that $\Delta(\mathcal{E}_n, \mathcal{E}) \to 0$, we can call $\mathcal{E}$ the *limit* of experiments $\{\mathcal{E}_n\}$ and write $\mathcal{E}_n \to \mathcal{E}$. Roughly speaking, the limit experiment $\mathcal{E}$ is almost equivalent to $\mathcal{E}_n$ when n is sufficiently large.

PROPOSITION 2 (van der Vaart, 1994, Lemma 2.6). *Suppose that $\mathcal{E}_n \equiv (P_{n,h} : h \in H)$ and $\mathcal{E} \equiv (P_h : h \in H)$ are experiments such that $P_{n,h}$ is dominated by P_{n,h_0} for some h_0. Then $\mathcal{E}_n \to \mathcal{E}$ if and only if the distribution of the random vector*

$$\left(\frac{dP_{n,h}}{dP_{n,h_0}} \right)_{I \in H}$$

under P_{n,h_0} converges weakly to the distribution of

$$\left(\frac{dP_h}{dP_{h_0}} \right)_{I \in H}$$

for every finite subset $I \subset H$. Here, dP_h/dP_{h_0} denotes the density of P_h with respect to P_{h_0}.

I will now make a connection with the well-known parametric inference. Let $\mathcal{E}$ denote a one-dimensional *shift normal* experiment with known variance $\mathcal{I}^{-1}$ and unknown mean, i.e.,

$$\mathcal{E} = \{\mathcal{N}(h, \mathcal{I}^{-1}), h \in \mathbb{R}\}.$$

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

Observe that $H = \mathbb{R}$. Also observe that the sample size is equal to one. Taking $h_0 = 0$, we can see that the log of the likelihood ratio process is given by

$$\log \frac{\sqrt{\mathcal{I}} \exp\left(-\frac{1}{2\mathcal{I}^{-1}} (X-h)^2\right)}{\sqrt{\mathcal{I}} \exp\left(-\frac{1}{2\mathcal{I}^{-1}} X^2\right)} = \mathcal{I}X \cdot h - \frac{\mathcal{I}}{2} h^2,$$

where the distribution of $\mathcal{I}X$ under h_0 is $\mathcal{N}(0, \mathcal{I})$. Now, consider an experiment based on n i.i.d. observations from a sufficiently smooth density $f(x, \theta)$, whose information at θ_0 is equal to $\mathcal{I}$, i.e., $\mathcal{I} = -E[\partial^2 \log f(Y, \theta_0)/\partial \theta^2]$, where $Y \sim f(\cdot, \theta_0)$. Reparametrize θ such that $\theta = \theta_0 + (1/\sqrt{n})h$, and let $\mathcal{E}_n$ denote the corresponding experiment, i.e., $\mathcal{E}_n = \{f(\cdot, \theta_0 + (1/\sqrt{n})h), h \in \mathbb{R}\}$. Taking $h_0 = 0$, it can be seen that the log of the likelihood ratio process is given by

$$\begin{aligned}
\log \frac{\prod_{i=1}^n f\left(Y_i, \theta_0 + \frac{1}{\sqrt{n}} h\right)}{\prod_{i=1}^n f(Y_i, \theta_0)} &\approx \sum_{i=1}^n \frac{\partial \log f(Y_i, \theta_0)}{\partial \theta} \frac{1}{\sqrt{n}} h + \frac{1}{2} \sum_{i=1}^n \frac{\partial^2 \log f(Y_i, \theta_0)}{\partial \theta^2} \frac{1}{n} h^2 \\
&\approx \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \log f(Y_i, \theta_0)}{\partial \theta}\right) h \\
&\quad + \frac{1}{2} \left(\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \log f(Y_i, \theta_0)}{\partial \theta^2}\right) h^2 \\
&\approx \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \log f(Y_i, \theta_0)}{\partial \theta}\right) h - \frac{\mathcal{I}}{2} h^2,
\end{aligned}$$

where Y_i is i.i.d. $f(Y_i, \theta_0)$ under $h_0 = 0$ and $\approx$ denotes that the difference between left-hand side and right-hand side converges to zero in probability as $n \to \infty$. Now, observe that $(1/\sqrt{n}) \sum_{i=1}^n (\partial \log f(Y_i, \theta_0)/\partial \theta)$ converges in distribution to $N(0, \mathcal{I})$ as $n \to \infty$. According to Proposition 2, we can see that the two experiments are approximately equivalent if n is sufficiently large: any estimator for h that can be constructed in $\mathcal{E}$ has a property that can be well approximated by that of an estimator for h constructed in $\mathcal{E}_n$. In fact, the estimator X has a distribution $\mathcal{N}(0, \mathcal{I}^{-1})$, whereas the maximum likelihood estimator $\hat{\theta}$, which implicitly estimates h via $\sqrt{n}(\hat{\theta} - \theta_0)$, has an asymptotic distribution $\mathcal{N}(0, \mathcal{I}^{-1})$.

Observe that it is not easy to come up with an approximate optimality criterion based on Proposition 2 alone unless it is easy to find an optimal inference in the limit experiment $\mathcal{E}$. Therefore, approximation by an arbitrary $\mathcal{E}$ does not necessarily simplify statistical analysis. On the other hand, if the limit experiment is a shift normal experiment as in the preceding example, we have some interesting consequences: we can obtain a Hajék-type convolution theorem. For the preceding example, the maximum likelihood estimator can be justified to be optimal by such a convolution theorem. What follows are the necessary ingredients for the convolution theorem when the experiment of interest is asymptotically shift normal.

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

DEFINITION 4 (van der Vaart and Wellner, 1996, p. 412). *Let* $\{\Delta_h : h \in H\}$ *be a real-valued iso-Gaussian process with zero mean and covariance function* $E[\Delta_{h_1}\Delta_{h_2}] \equiv \langle h_1, h_2 \rangle$. *A sequence of experiments is called* asymptotically shift normal *if*

$$\log \frac{dP_{n,h}}{dP_{n,h_0}} = \Delta_{n,h} - \frac{1}{2}\|h\|^2$$

for stochastic processes $\{\Delta_{n,h} : h \in H\}$ *such that* $(\Delta_{n,h})_{I \in H}$ *converges weakly to* $(\Delta_h)_{I \in H}$ *under* P_{n,h_0} *for every finite subset* $I \subset H$. *Here,* $\|h\|^2 \equiv \langle h, h \rangle$.

DEFINITION 5 (van der Vaart and Wellner, 1996, p. 413). *A sequence of parameters* $\kappa_n(h)$ *belonging to a Banach space B is* regular *if*

$$\mathbf{r}_n(\kappa_n(h) - \kappa_n(h_0)) \to \dot{\kappa}(h)$$

for a continuous linear map $\dot{\kappa} : H \to \mathbf{B}$ *and certain linear maps* $\mathbf{r}_n : \mathbf{B} \to \mathbf{B}$. *A sequence of estimators* T_n *is* regular *with respect to the norming operators* $\mathbf{r}_n$ *if* $\mathbf{r}_n(T_n - \kappa_n(h))$ *under h converges weakly to some fixed tight Borel probability measure* $\mathcal{L}$ *on* $\mathbf{B}$ *for every h.*

Denote the adjoint map of $\dot{\kappa}$ by $\dot{\kappa}^*$, which is obviously defined by

$$\langle \dot{\kappa}^* b^*, h \rangle = b^* \dot{\kappa}(h)$$

for every b^* in the dual $\mathbf{B}^*$ of $\mathbf{B}$. Van der Vaart and Wellner (1996) establish the following Hajék-type convolution theorem for a regular sequence of parameters $\{\kappa_n(h)\}$ and estimators $\{T_n\}$ when $\{\mathcal{E}_n\}$ is asymptotically shift normal.

THEOREM 3 (van der Vaart and Wellner, 1996, Theorem 3.11.2). *Assume that* $\{P_{n,h} : h \in H\}$ *is asymptotically shift normal. Also assume that the sequence of parameters* $\kappa_n(h)$ *and estimators* T_n *is regular. Then, the limit distribution* $\mathcal{L}$ *of* $\mathbf{r}_n(T_n - \kappa_n(h))$ *equals the sum* $G + W$ *of independent, tight, Borel measurable random elements in* $\mathbf{B}$ *such that*

$$b^* G \sim \mathcal{N}(0, \|\dot{\kappa}^* b^*\|^2)$$

for every b^* *in the dual* $\mathbf{B}^*$.

I will now present a characterization of Theorem 3 for the special case where a component of h is real valued, so that we can write $h = (\delta, \Xi)$, where δ denotes the real-valued component. The result is deceptively similar to the characterization of the semiparametric efficiency bound introduced by Stein (1956) and developed by Begun, Hall, Huang, and Wellner (1983) and Bickel, Klaassen, Ritov, and Wellner (1993). See Newey (1990a), e.g., for a review on this subject.

THEOREM 4. *Assume that* $(P_{n,h} : h \in H)$ *is asymptotically shift normal. (As discussed in Definition 4, this implicitly assumes that H is a Hilbert space.) Also, suppose that (i)* $h = (\delta, \Xi)$, $\delta \in \mathbb{R}$, $h_0 = (0,0)$; *(ii)* $\kappa_n(h) \equiv \beta_n \equiv \beta_0 + (1/r_n)\delta$ *for some* $\beta_0 \in \mathbb{R}$; *and (iii)* $\Delta_h \equiv \Delta_1 \delta + \Delta_2(\Xi)$, *where* Δ_1 *and* $\Delta_2(\Xi)$ *are real valued. Further suppose that,*

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

with respect to the norm $\|\cdot\|$, (iv) the mapping $\dot{\kappa}:(\delta,\Xi)\to\delta$ is continuous; and (v) H is complete. Then, for every regular sequence of estimators $\{T_n\}$, we have

$$r_n(T_n-\kappa_n(h))\Rightarrow\mathcal{N}(0,E[\tilde{\Delta}_1^2]^{-1})\oplus\mathcal{W}$$

for some distribution $\mathcal{W}$ on the real line, where $\oplus$ denotes convolution and $\tilde{\Delta}_1$ is the residual in the projection of Δ_1 on $\{\Delta_2(\Xi)$: for some $(\delta,\Xi)\in H\}$. (Here, $\Rightarrow$ denotes weak convergence and $\oplus$ denotes convolution.)

Proof. Observe that $\kappa_n(h)$ is *regular* (see Definition 5) because $\mathbf{r}_n(\kappa_n(h)-\kappa_n(h_0))\to\delta$, where $\mathbf{r}_n(\theta)\equiv r_n\times\theta$. (The definition of regularity does not necessarily require that $r_n\nearrow\infty$, although most econometric examples would exhibit such property.) Also observe that $\dot{\kappa}:(\delta,\Xi)\to\delta$ is linear and continuous. Write $\dot{\kappa}^*:d\mapsto(\dot{\kappa}_1^*(d),\dot{\kappa}_2^*(d))\in H$. (Note that $\mathbf{B}=\mathbb{R}$. Obviously, the dual of $\mathbb{R}$ is $\mathbb{R}$ itself.) It is defined by $\langle(\dot{\kappa}_1^*(d),\dot{\kappa}_2^*(d)),(\delta,\Xi)\rangle=d\delta$ for $d,\delta\in\mathbb{R}$. Linearity of the adjoint map allows us to write $(\dot{\kappa}_1^*(d),\dot{\kappa}_2^*(d))=d(\dot{\kappa}_1^*,\dot{\kappa}_2^*)$ for some $(\dot{\kappa}_1^*,\dot{\kappa}_2^*)$. Therefore, we can rewrite $d\langle(\dot{\kappa}_1^*,\dot{\kappa}_2^*),(\delta,\Xi)\rangle=d\delta$, which can happen if and only if

$$\langle(\dot{\kappa}_1^*,\dot{\kappa}_2^*),(\delta,\Xi)\rangle=\delta\quad\forall(\delta,\Xi). \tag{A.1}$$

In particular, (A.1) should hold for $\delta=1$, which can be rewritten using the bilinearity of the inner product $\langle\cdot,\cdot\rangle$ as

$$\dot{\kappa}_1^*\left\langle\left(1,\frac{1}{\dot{\kappa}_1^*}\dot{\kappa}_2^*\right),(1,\Xi)\right\rangle=1\quad\forall\Xi.$$

In other words, we have

$$\dot{\kappa}_1^*E[(\Delta_1+\Delta_2(\Xi^*))(\Delta_1+\Delta_2(\Xi))]=1, \tag{A.2}$$

where I write $\Xi^*=(1/\dot{\kappa}_1^*)\dot{\kappa}_2^*$ for simplicity. Letting $(\delta,\Xi)=(\dot{\kappa}_1^*,\dot{\kappa}_2^*)$ in (A.1), we obtain

$$\|(\dot{\kappa}_1^*,\dot{\kappa}_2^*)\|^2=\langle(\dot{\kappa}_1^*,\dot{\kappa}_2^*),(\dot{\kappa}_1^*,\dot{\kappa}_2^*)\rangle=\dot{\kappa}_1^*,$$

which, combined with (A.2), implies that

$$\|(\dot{\kappa}_1^*,\dot{\kappa}_2^*)\|^2=\frac{1}{E[(\Delta_1+\Delta_2(\Xi^*))(\Delta_1+\Delta_2(\Xi))]}. \tag{A.3}$$

We now characterize $E[(\Delta_1+\Delta_2(\Xi^*))(\Delta_1+\Delta_2(\Xi))]$, the denominator on the right-hand side of (A.3). Because (A.2) holds for every Ξ, it follows that $E[(\Delta_1+\Delta_2(\Xi^*))(\Delta_2(\Xi))]=0$ for every Ξ. Write $\Delta_1=\hat{\Delta}_1+\tilde{\Delta}_1$, where $\hat{\Delta}_1$ is the projection of Δ_1 on $\{\Delta_2(\Xi)$: for some $(\delta,\Xi)\in H\}$. By definition, we should have $E[\tilde{\Delta}_1\Delta_2(\Xi)]=0$ for every Ξ. Therefore, we should have $E[(\hat{\Delta}_1+\Delta_2(\Xi^*))\Delta_2(\Xi)]=0$. Observe that both $\hat{\Delta}_1$ and $\Delta_2(\Xi^*)$ are in $\{\Delta_2(\Xi)$: for some $(\delta,\Xi)\in H\}$, a linear space. It follows

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

that there exists some $\ddot{\Xi}$ such that $\hat{\Delta}_1 + \Delta_2(\Xi^*) = \Delta_2(\ddot{\Xi})$. Therefore, we should have $E[(\hat{\Delta}_1 + \Delta_2(\Xi^*))^2] = 0$, which can happen if and only if

$$\Delta_2(\Xi^*) = -\hat{\Delta}_1. \tag{A.4}$$

Plugging (A.4) into (A.3), we obtain $\|(\dot{\kappa}_1^*, \dot{\kappa}_2^*)\|^2 = 1/E[\tilde{\Delta}_1^2]$. ■

APPENDIX B: TECHNICAL LEMMAS

DEFINITION 6. *Let*

$$\Delta_n(\delta, \Xi, \Lambda) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i^{n\prime} \Psi_0 W_i^n(\delta, \Xi) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{1}{2} v_i^{n\prime} \Lambda v_i^n - \frac{1}{2} E[v_i^{n\prime} \Lambda v_i^n] \right),$$

where

$$W_i^n(\delta, \Xi) \equiv (z_i^{n\prime} \Pi_0^n \delta + z_i^{n\prime} \Xi^n \beta_0, z_i^{n\prime} \Xi^n)'.$$

LEMMA 1. $\mathrm{Cov}(v_i^{n\prime} \Lambda_1 v_i^n, v_i^{n\prime} \Lambda_2 v_i^n) = 2\,\mathrm{trace}(\Omega_0 \Lambda_1 \Omega_0 \Lambda_2)$.

Proof. Write

$$(\Upsilon_0^{-1})' \Lambda_j \Upsilon_0^{-1} \equiv \begin{bmatrix} \varpi_{11}^j & & \varpi_{1,K+1}^j \\ & \ddots & \\ \varpi_{K+1,1}^j & & \varpi_{K+1,K+1}^j \end{bmatrix}, \qquad j = 1,2,$$

where Υ_0 is such that $\Upsilon_0 \Omega_0 \Upsilon_0' = I$. Observe that $u_i^n \equiv \Upsilon_0 v_i^n \sim \mathcal{N}(0, I)$. We then have

$$E[(v_i^{n\prime} \Lambda_1 v_i^n)(v_i^{n\prime} \Lambda_2 v_i^n)] = \mathrm{trace}(\Omega_0 \Lambda_1)\mathrm{trace}(\Omega_0 \Lambda_2) + 2\,\mathrm{trace}(\Omega_0 \Lambda_1 \Omega_0 \Lambda_2).$$

The conclusion follows from $E[v_i^{n\prime} \Lambda_j v_i^n] = \mathrm{trace}(\Omega_0 \Lambda_j)$. ■

LEMMA 2. $v_i^{n\prime} \Psi_0 W_i^n$ *and* $\frac{1}{2} v_i^{n\prime} \Lambda v_i^n - \frac{1}{2} E[v_i^{n\prime} \Lambda v_i^n]$ *are uncorrelated.*

Proof. Write

$$(\Upsilon_0^{-1})' \Lambda \Upsilon_0^{-1} \equiv \begin{bmatrix} \varpi_{11} & & \varpi_{1,K+1} \\ & \ddots & \\ \varpi_{K+1,1} & & \varpi_{K+1,K+1} \end{bmatrix},$$

where Υ_0 is such that $\Upsilon_0 \Omega_0 \Upsilon_0' = I$. Observe that $u_i^n \equiv \Upsilon_0 v_i^n \sim \mathcal{N}(0, I)$. It suffices to show that u_i^n and $v_i^{n\prime} \Lambda v_i^n = u_i^{n\prime} (\Upsilon_0^{-1})' \Lambda \Upsilon_0^{-1} u_i^n$ are uncorrelated. Observe that the jth element of the vector $E[u^{n\prime} (\Upsilon_0^{-1})' \Lambda \Upsilon_0^{-1} u^n \cdot u^n]$ equals $E[u^{n\prime} (\Upsilon_0^{-1})' \Lambda \Upsilon_0^{-1} u^n \cdot u_j^n] = \sum_{k,l} \varpi_{kl} E[u_k^n u_l^n u_j^n]$, where I omitted the i subscript for notational simplicity. Because $E[u_k^n u_l^n u_j^n] = 0$ for every j, k, l combination, we obtain the desired conclusion. ■

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

LEMMA 3. $\langle h_1, h_2 \rangle = \lim_{n\to\infty} \mathrm{Cov}(\Delta_n(h_1), \Delta_n(h_2))$.

Proof. As a result of Lemmas 1 and 2, we have

$$\mathrm{Cov}(\Delta_n(h_1), \Delta_n(h_1)) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Cov}(v_i^{n\prime}\Psi_0 W_i^n(\delta_1, \Xi_1), v_i^{n\prime}\Psi_0 W_i^n(\delta_2, \Xi_2)) + \frac{1}{2}\,\mathrm{trace}(\Omega_0\Lambda_1\Omega_0\Lambda_2).$$

The conclusion follows from

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \mathrm{Cov}(v_i^{n\prime}\Psi_0 W_i^n(\delta_1, \Xi_1), v_i^{n\prime}\Psi_0 W_i^n(\delta_2, \Xi_2)) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} (z_i^{n\prime}\Pi_0^n\delta_1 + z_i^{n\prime}\Xi_1^n\beta_0, z_i'\Xi_1)\Psi_0 \begin{bmatrix} \delta_2'\Pi_0^{n\prime} z_i^n + \beta_0'\Xi_2^{n\prime} z_i^n \\ \Xi_2^{n\prime} z_i^n \end{bmatrix} \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} \mathrm{trace}\left\{ \Psi_0 \begin{bmatrix} \delta_2'\Pi_0^{n\prime} z_i^n + \beta_0'\Xi_2^{n\prime} z_i^n \\ \Xi_2^{n\prime} z_i^n \end{bmatrix} (z_i^{n\prime}\Pi_0^n\delta_1 + z_i^{n\prime}\Xi_1^n\beta_0, z_i^{n\prime}\Xi_1^n) \right\} \\
&\quad = \mathrm{trace}[\Psi_0\Phi(h_1, h_2)] + o(1). \qquad \blacksquare
\end{aligned}$$

LEMMA 4. *For every finite set I, the vector* $(\Delta_n(h))_{h\subset I}$ *is asymptotically normal.*

Proof. For simplicity, I only establish a one-dimensional asymptotic distribution. It is obvious that $(1/\sqrt{n})\sum_{i=1}^n v_i^{n\prime}\Psi_0 W_i^n(\delta, \Xi)$ is asymptotically normal; its finite sample distribution is normal by the normality of v_i^n. Thus, by using the Cramér–Wold device, it suffices to establish the asymptotic normality for

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \{\theta v_i^{n\prime}\Psi_0 W_i^n(\delta, \Xi) - (v_i^{n\prime}\Lambda v_i^n - E[v_i^{n\prime}\Lambda v_i^n])\} \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{n} q_i^n$$

for every nonzero $\theta \in \mathbb{R}$. For this purpose, I simply verify the Lindeberg condition. Note that $(1/\sqrt{n})\sum_{i=1}^n q_i^n$ is a sum of zero-mean independent random variables with variance equal to

$$s_n^2 \equiv \frac{\theta^2}{n}\sum_{i=1}^{n} W_i^n(\delta, \Xi)'\Psi_0 W_i^n(\delta, \Xi) + 2\,\mathrm{trace}[(\Omega_0\Lambda)^2].$$

We need to establish

$$\lim_{n\to\infty} \frac{1}{n s_n^2}\sum_{i=1}^{n} E[1(q_i^2 \geq n\epsilon s_n^2) q_i^2] = 0$$

for every $\epsilon > 0$. Here, $1(\cdot)$ denotes the usual indicator function. Observe that $s_n^2 = O(1)$. Further, unless $(\delta, \Xi, \Lambda) = (0,0,0)$, in which case $\Delta_n(h)$ is asymptotically normal with variance equal to zero, s_n^2 is bounded away from zero when n is sufficiently large. Without loss of generality, assume that $s_n^2 \geq \varkappa > 0$. Thus, it suffices to show that

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}E[1(q_i^2\geq n\epsilon)(q_i^n)^2]=0. \tag{B.1}$$

Because

$$(q_i^n)^2\leq 2\theta^2(v_i^{n\prime}\Psi_0 W_i^n(\delta,\Xi))^2+2(v_i^{n\prime}\Lambda v_i^n-E[v_i^{n\prime}\Lambda v_i^n])^2,$$

we may establish (B.1) by showing

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}E[1((v_i^{n\prime}\Psi_0 W_i^n(\delta,\Xi))^2\geq n\epsilon)(v_i^{n\prime}\Psi_0 W_i^n(\delta,\Xi))^2]=0, \tag{B.2}$$

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}E[1((v_i^{n\prime}\Lambda v_i^n-E[v_i^{n\prime}\Lambda v_i^n])^2\geq n\epsilon)(v_i^{n\prime}\Psi_0 W_i^n(\delta,\Xi))^2]=0, \tag{B.3}$$

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}E[1((v_i^{n\prime}\Psi_0 W_i^n(\delta,\Xi))^2\geq n\epsilon)(v_i^{n\prime}\Lambda v_i^n-E[v_i^{n\prime}\Lambda v_i^n])^2]=0, \tag{B.4}$$

and

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}E[1((v_i^{n\prime}\Lambda v_i^n-E[v_i^{n\prime}\Lambda v_i^n])^2\geq n\epsilon)(v_i^{n\prime}\Lambda v_i^n-E[v_i^{n\prime}\Lambda v_i^n])^2]=0. \tag{B.5}$$

For (B.3) and (B.5), we note that $1((v_i^{n\prime}\Lambda v_i^n-E[v_i^{n\prime}\Lambda v_i^n])^2\geq n\epsilon)\to 0$ almost surely, and hence, the desired conclusions follow from the dominated convergence theorem. As for (B.2) and (B.4), we note that

$$1((v_i^{n\prime}\Psi_0 W_i^n(\delta,\Xi))^2\geq n\epsilon)\leq 1(M_n(v_i^{n\prime}\Psi_0)^2\geq n\epsilon),$$

where

$$M_n=\max_{1\leq i\leq n}W_i^n(\delta,\Xi)'W_i^n(\delta,\Xi)=o(n).$$

Thus, $1((v_i^{n\prime}\Psi_0 W_i^n(\delta,\Xi))^2\geq n\epsilon)\to 0$ almost surely, and the dominated convergence theorem again delivers the desired conclusions. ■

APPENDIX C: PROOF OF THEOREM 1

Observe that the joint likelihood is proportional to

$$\prod_{i=1}^{n}\sqrt{\det(\Psi)}\exp\left\{-\frac{1}{2}(y_i^n-z_i^{n\prime}\Pi^n\beta,x_i^n-z_i^{n\prime}\Pi^n)\Psi(y_i^n-z_i^{n\prime}\Pi^n\beta,x_i^n-z_i^{n\prime}\Pi^n)'\right\}, \tag{C.1}$$

where we write $\Psi=\Omega^{-1}$ for simplicity of notation. I will consider parameter values around (β_0,Π_0,Ψ_0). Specifically, I consider the sequence of *likelihood processes* $\{P_{n,h}\}$

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

indexed by $h \in H$, where H is a linear space consisting of $h = (\delta, \Xi, \Lambda)$ and where $P_{n,h}$ has a density equal to (C.1) with respect to the Lesbesgue measure on $\mathbb{R}^{2n}$ with

$$\beta = \beta(n) = \beta_0 + \frac{1}{\sqrt{n}}\,\delta, \qquad \Psi = \Psi(n) = \Psi_0 + \frac{1}{\sqrt{n}}\,\Lambda$$

and

$$\Pi^n = \Pi^n(n) = \Pi_0^n + \frac{1}{\sqrt{n}}\,\Xi^n.$$

Here, $\Pi(n)$ and $\Pi_0 + (1/\sqrt{n})\Xi$ denote sequences $(\Pi^1(n), \Pi^2(n), \ldots)$ and $(\Pi_0^1 + (1/\sqrt{n})\Xi^1, \Pi_0^2 + (1/\sqrt{n})\Xi^2, \ldots)$. Therefore, we have

$$\begin{aligned}\log \frac{dP_{n,h}}{dP_{n,h_0}} = {} & \frac{n}{2} \log \frac{\det(\Psi(n))}{\det(\Psi_0)} \\ & - \frac{1}{2} \sum_{i=1}^{n} (y_i^n - z_i^{n\prime}\Pi^n(n)\beta(n), x_i^n - z_i^{n\prime}\Pi^n(n)) \\ & \times \Psi(n)(y_i^n - z_i^{n\prime}\Pi^n(n)\beta(n), x_i^n - z_i^{n\prime}\Pi^n(n))' \\ & + \frac{1}{2} \sum_{i=1}^{n} .(y_i^n - z_i^{n\prime}\Pi_0^n\beta_0, x_i^n - z_i^{n\prime}\Pi_0^n)\Psi_0(y_i^n - z_i^{n\prime}\Pi_0^n\beta_0, x_i^n - z_i^{n\prime}\Pi_0^n)'.\end{aligned}$$

I will make a high-level assumption that H is such that $(1/n)\sum_{i=1}^n \Xi^{n\prime} z_i^n z_i^{n\prime} \Xi^n$ is convergent. Therefore, if $(\delta_1, \Xi_1, \Lambda_1), (\delta_2, \Xi_2, \Lambda_2) \in H$, the bilinear form

$$\Theta(\Xi_1, \Xi_2) \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Xi_1^{n\prime} z_i^n z_i^{n\prime} \Xi_2^n$$

is well defined.

DEFINITION 7. *Let*

$$\langle h_1, h_2 \rangle \equiv \operatorname{trace}[\Psi_0 \Phi(h_1, h_2)] + \frac{1}{2} \operatorname{trace}(\Omega_0 \Lambda_1 \Omega_0 \Lambda_2),$$

where

$$\Phi(h_1, h_2) \equiv \begin{bmatrix} \Phi_{11}(h_1, h_2) & \Phi_{12}(h_1, h_2) \\ \Phi_{21}(h_1, h_2) & \Phi_{22}(h_1, h_2) \end{bmatrix},$$

with

$$\Phi_{11}(h_1, h_2) \equiv \delta_1\delta_2\Theta(\Pi_0, \Pi_0) + \delta_1\beta_0\Theta(\Xi_2, \Pi_0) + \delta_2\beta_0\Theta(\Pi_0, \Xi_1) + \beta_0^2\Theta(\Xi_2, \Xi_1),$$

$$\Phi_{12}(h_1, h_2) \equiv \delta_2\Theta(\Pi_0, \Xi_1) + \beta_0\Theta(\Xi_2, \Xi_1),$$

$$\Phi_{21}(h_1, h_2) \equiv \delta_1\Theta(\Xi_2, \Pi_0) + \beta_0\Theta(\Xi_2, \Xi_1),$$

$$\Phi_{22}(h_1, h_2) \equiv \Theta(\Xi_2, \Xi_1).$$

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

THEOREM 5. *Under* (β_0, Π_0, Ψ_0), *the log likelihood ratio process is equal to*

$$\Delta_n(h) - \frac{1}{2}\|h\|^2 + o_p(1),$$

where $\|h\|^2 = \langle h, h\rangle$ *and*

$$\Delta_n(h) \Rightarrow \Delta(h) \sim \mathcal{N}(0, \|h\|^2).$$

Furthermore, $\mathrm{Cov}(\Delta(h_1), \Delta(h_2)) = \langle h_1, h_2\rangle$.

Proof. Let

$$w_i^n \equiv \left(z_i^{n\prime} \Pi_0^n \delta + z_i^{n\prime} \Xi^n(n) \beta_0 + \frac{1}{\sqrt{n}} z_i^{n\prime} \Xi^n(n) \delta, z_i^{n\prime} \Xi^n(n) \right)'.$$

Because $y_i^n = z_i^{n\prime} \Pi_0^n \beta_0 + v_{1i}^n$, and $x_i^n = \Pi_0^{n\prime} z_i^n + V_{2i}^n$ under (β_0, Π_0, Ψ_0), we may write

$$(y_i^n - z_i^{n\prime} \Pi^n(n) \beta(n), x_i^n - z_i^{n\prime} \Pi^n(n))' = v_i^n - \frac{1}{\sqrt{n}} w_i^n.$$

It follows that

$$\begin{aligned}
&-\frac{1}{2} \sum_{i=1}^n (y_i^n - \beta(n) z_i^{n\prime} \Pi^n(n), x_i^n - z_i^{n\prime} \Pi^n(n)) \\
&\qquad \times \Psi(n) \left(y_i - \beta(n)\, \Pi^{n\prime}(n)\, z_i^n, x_i^n - z_i^{n\prime} \Pi^n(n)\right)' \\
&\qquad + \frac{1}{2} \sum_{i=1}^n (y_i^n - \beta_0 z_i^{n\prime} \Pi_0^n, x_i^n - z_i^{n\prime} \Pi_0^n)\, \Psi_0 (y_i^n - \beta_0 z_i^{n\prime} \Pi_0^n, x_i^n - z_i^{n\prime} \Pi_0^n) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n v_i^{n\prime} \Psi_0 w_i^n - \frac{1}{2n} \sum_{i=1}^n w_i^{n\prime} \Psi_0 w_i^n - \frac{1}{2\sqrt{n}} \sum_{i=1}^n (v_i^{n\prime} \Lambda v_i^n - E[v_i^{n\prime} \Lambda v_i^n]) \\
&\qquad + \frac{1}{n} \sum_{i=1}^n v_i^{n\prime} \Lambda w_i^n - \frac{1}{2n\sqrt{n}} \sum_{i=1}^n w_i^{n\prime} \Lambda w_i^n - \frac{\sqrt{n}}{2} E[v_i^{n\prime} \Lambda v_i^n].
\end{aligned}$$

Now, observe that

$$\frac{n}{2} \log \frac{\det\left(\Psi_0 + \frac{1}{\sqrt{n}} \Lambda\right)}{\det(\Psi_0)} = \frac{\sqrt{n}}{2} \mathrm{trace}(\Omega_0 \Lambda) - \frac{1}{4} \mathrm{trace}[(\Omega_0 \Lambda)^2] + o(1),$$

which can be established by Magnus and Neudecker's (1988) Theorem 2 and Exercise 1 on pages 150 and 152, e.g. Further, note that

$$E[v_i^{n\prime} \Lambda v_i^n] = E[\mathrm{trace}(\Lambda v_i^n v_i^{n\prime})] = \mathrm{trace}(\Lambda \Omega_0).$$

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

We may thus write

$$\log \frac{dP_{n,h}}{dP_{n,h_0}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i^{n\prime} \Psi_0 w_i^n - \frac{1}{2n} \sum_{i=1}^{n} w_i^{n\prime} \Psi_0 w_i^n$$

$$- \frac{1}{2\sqrt{n}} \sum_{i=1}^{n} (v_i^{n\prime} \Lambda v_i^n - E[v_i^{n\prime} \Lambda v_i^n]) - \frac{1}{4} \operatorname{trace}[(\Omega_0 \Lambda)^2]$$

$$+ \frac{1}{n} \sum_{i=1}^{n} v_i^{n\prime} \Lambda w_i^n - \frac{1}{2n\sqrt{n}} \sum_{i=1}^{n} w_i^{n\prime} \Lambda w_i^n + o(1). \tag{C.2}$$

It is easy to see that

$$\frac{1}{2n\sqrt{n}} \sum_{i=1}^{n} w_i^{n\prime} \Lambda w_i^n = o(1). \tag{C.3}$$

Further, because $(1/n)\sum_{i=1}^{n} v_i^{n\prime} \Lambda w_i^n$ has mean zero and variance $(1/n^2) \times \sum_{i=1}^{n} \operatorname{trace}(\Lambda \Omega_0 \Lambda w_i^n w_i^{n\prime}) = o(1)$, we have

$$\frac{1}{n} \sum_{i=1}^{n} v_i^{n\prime} \Lambda w_i^n = o_p(1). \tag{C.4}$$

Combining (C.2)–(C.4), we can see that, under (β_0, Π_0, Ψ_0), the log likelihood ratio equals

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i^{n\prime} \Psi_0 w_i^n - \frac{1}{2n} \sum_{i=1}^{n} w_i^{n\prime} \Psi_0 w_i^n - \frac{1}{2\sqrt{n}} \sum_{i=1}^{n} (v_i^{n\prime} \Lambda v_i^n - E[v_i^{n\prime} \Lambda v_i^n])$$

$$- \frac{1}{4} \operatorname{trace}[(\Omega_0 \Lambda)^2] + o_p(1).$$

Because

$$\frac{1}{n} \sum_{i=1}^{n} v_i^{n\prime} \Psi_0 (z_i^{n\prime} \Xi^n(n)\delta, 0)' = o_p(1),$$

$$\frac{1}{n\sqrt{n}} \sum_{i=1}^{n} w_i^n \Psi_0 (z_i^{n\prime} \Xi^n(n)\delta, 0)' = o(1), \quad \text{and}$$

$$\frac{1}{n^2} \sum_{i=1}^{n} (z_i^{n\prime} \Xi^n(n)\delta, 0) \Psi_0 (z_i^{n\prime} \Xi^n(n)\delta, 0)' = o(1),$$

we may further simplify the log likelihood ratio under (β_0, Π_0, Ψ_0) as

$$\Delta_n(\delta, \Xi, \Lambda) - \frac{1}{2n} \sum_{i=1}^{n} W_i^{n\prime}(\delta, \Xi) \Psi_0 W_i^n(\delta, \Xi) - \frac{1}{4} \operatorname{trace}[(\Omega_0 \Lambda)^2] + o_p(1).$$

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

We now note that $\Delta_n(\delta, \Xi, \Lambda)$ is asymptotically normal with asymptotic variance given by

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} W_i^{n\prime}(\delta, \Xi) \Psi_0 W_i^n(\delta, \Xi) + \frac{1}{2} \operatorname{trace}[(\Omega_0 \Lambda)^2].$$

Also, from Lemma 3, we can see that $\langle \cdot, \cdot \rangle$ indeed is the desired asymptotic covariance kernel. ■

Consider a sequence of parameters $\kappa_n(\delta, \Xi, \Lambda) = \beta(n)$. Observe that it is indeed regular (see Definition 5) because

$$\sqrt{n}(\kappa_n(\delta, \Xi, \Lambda) - \kappa_n(0,0,0)) = \delta$$

for every (δ, Ξ, Λ).

I now invoke Theorem 4. Let

$$\Delta_{1n} \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i^{n\prime} \Psi_0 W_{1i}^n \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i^{n\prime} \Psi_0 (z_i^{n\prime} \Pi_0^n, 0)'$$

and

$$\begin{aligned}\Delta_{2n}(\Xi, \Lambda) &\equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i^{n\prime} \Psi_0 W_{2i}^n(\Xi, \Lambda) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{1}{2} v_i^{n\prime} \Lambda v_i^n - \frac{1}{2} E[v_i^{n\prime} \Lambda v_i^n] \right) \\ &\equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i^{n\prime} \Psi_0 (z_i^{n\prime} \Xi^n \beta_0, z_i^{n\prime} \Xi^n)' - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{1}{2} v_i^{n\prime} \Lambda v_i^n - \frac{1}{2} E[v_i^{n\prime} \Lambda v_i^n] \right).\end{aligned}$$

It can be shown that $\Delta_n(\delta, \Xi, \Lambda) = \Delta_{1n}\delta + \Delta_{2n}(\Xi, \Lambda)$ and the limit distribution of $(\Delta_{1n}\delta, \Delta_{2n}(\Xi, \Lambda))$ under P_{n,h_0} equals that of the iso-Gaussian process $(\Delta_1\delta, \Delta_2(\Xi, \Lambda))$ such that

$$\begin{aligned}&\operatorname{Cov}[(\Delta_1\delta_1, \Delta_2(\Xi_1, \Lambda_1)), (\Delta_1\delta_2, \Delta_2(\Xi_2, \Lambda_2))] \\ &\quad = \lim_{n\to\infty} \operatorname{Cov}[(\Delta_{1n}\delta_1, \Delta_{2n}(\Xi_1, \Lambda_1)), (\Delta_{1n}\delta_2, \Delta_{2n}(\Xi_2, \Lambda_2))],\end{aligned}$$

from which we can show that

$$\begin{aligned}\operatorname{Var}[\Delta_1] &= \lim_{n\to\infty} \operatorname{Var}[\Delta_{1n}] = \operatorname{trace}\left(\Psi_0 \begin{bmatrix} \Theta(\Pi_0, \Pi_0) & 0 \\ 0 & 0 \end{bmatrix} \right) \\ &= \frac{\Theta(\Pi_0, \Pi_0)}{\omega_{11}\omega_{22} - \omega_{12}\omega_{21}} \omega_{22}, \\ \operatorname{Cov}[\Delta_1, \Delta_2(\Xi, \Lambda)] &= \lim_{n\to\infty} \operatorname{Cov}[\Delta_{1n}, \Delta_{2n}(\Xi, \Lambda)] = \operatorname{trace}\left(\Psi_0 \begin{bmatrix} \Theta(\Xi, \Pi_0)\beta & 0 \\ \Theta(\Xi, \Pi_0) & 0 \end{bmatrix} \right) \\ &= \frac{\Theta(\Xi, \Pi_0)}{\omega_{11}\omega_{22} - \omega_{12}\omega_{21}} (\omega_{22}\beta_0 - \omega_{12}),\end{aligned}$$

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

$$\begin{aligned}\mathrm{Var}[\Delta_2(\Xi,\Lambda)] &= \lim_{n\to\infty} \mathrm{Var}[\Delta_{2n}(\Xi,\Lambda)] \\ &= \mathrm{trace}\left(\Psi_0\begin{bmatrix}\Theta(\Xi,\Xi)\beta^2 & \Theta(\Xi,\Xi)\beta \\ \Theta(\Xi,\Xi)\beta & \Theta(\Xi,\Xi)\end{bmatrix}\right) + \frac{1}{2}\mathrm{trace}[(\Omega_0\Lambda)^2] \\ &= \frac{\Theta(\Xi,\Xi)}{\omega_{11}\omega_{22}-\omega_{12}\omega_{21}}(\omega_{22}\beta_0^2 - 2\omega_{12}\beta_0 + \omega_{11}) + \frac{1}{2}\mathrm{trace}[(\Omega_0\Lambda)^2].\end{aligned}$$

I will now compute $E[\tilde{\Delta}_1^2]$, where $\tilde{\Delta}_1$ is the residual in the projection of Δ_1 on the space $\{\Delta_2(\Xi,\Lambda)\colon \text{for some } (\delta,\Xi,\Lambda)\in H\}$. Because the space $\{\Delta_2(\Xi,\Lambda)\colon(\delta,\Xi,\Lambda)\in H\}$ is linear, we can write

$$E[\tilde{\Delta}_1^2] = \min_{\Xi,\Lambda} E[(\Delta_1(1) - \Delta_2(\Xi,\Lambda))^2].$$

Observe that

$$\begin{aligned}E[(\Delta_1 - \Delta_2(\Xi,\Lambda))^2] &= \frac{\Theta(\Pi_0,\Pi_0)}{\omega_{11}\omega_{22}-\omega_{12}\omega_{21}}\omega_{22} \\ &\quad - 2\frac{\Theta(\Xi,\Pi_0)}{\omega_{11}\omega_{22}-\omega_{12}\omega_{21}}(\omega_{22}\beta_0 - \omega_{12}) \\ &\quad + \frac{\Theta(\Xi,\Xi)}{\omega_{11}\omega_{22}-\omega_{12}\omega_{21}}(\omega_{22}\beta_0^2 - 2\omega_{12}\beta_0 + \omega_{11}) \\ &\quad + \frac{1}{2}\mathrm{trace}[(\Omega_0\Lambda)^2] \\ &= \frac{\mathbb{N}(\Xi)}{\omega_{11}\omega_{22}-\omega_{12}\omega_{21}} + \frac{1}{2}\mathrm{trace}[(\Omega_0\Lambda)^2],\end{aligned}$$

where

$$\mathbb{N}(\Xi) \equiv \Theta(\Pi_0,\Pi_0)\omega_{22} - 2\Theta(\Xi,\Pi_0)(\omega_{22}\beta_0 - \omega_{12}) + \Theta(\Xi,\Xi)\sigma^2$$

and

$$\sigma^2 \equiv \mathrm{Var}(y_i^n - x_i^n\beta_0) = \omega_{22}\beta_0^2 - 2\omega_{12}\beta_0 + \omega_{11} > 0.$$

Therefore, $E[(\Delta_1 - \Delta_2(\Xi,\Lambda))^2]$ can be minimized by separately minimizing

$$\frac{\mathbb{N}(\Xi)}{\omega_{11}\omega_{22}-\omega_{12}\omega_{21}} \quad \text{and} \quad \frac{1}{2}\mathrm{trace}[(\Omega_0\Lambda)^2].$$

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

Minimization of $\frac{1}{2}\,\text{trace}[(\Omega_0\Lambda)^2]$ can be easily accomplished by putting $\Lambda = 0$. As for $\mathbb{N}(\Xi)$, observe that

$$\mathbb{N}(\Xi) = \Theta\left[\sigma\Xi - \frac{\omega_{22}\beta_0 - \omega_{12}}{\sigma}\,\Pi_0, \sigma\Xi - \frac{\omega_{22}\beta_0 - \omega_{12}}{\sigma}\right] + \Theta(\Pi_0, \Pi_0)\left(\omega_{22} - \frac{(\omega_{22}\beta_0 - \omega_{12})^2}{\sigma^2}\right).$$

Therefore, $\mathbb{N}(\Xi)$ is minimized by putting

$$\Xi = \frac{\omega_{22}\beta_0 - \omega_{12}}{\sigma^2}\,\Pi_0,$$

and we obtain

$$E[\tilde{\Delta}_1^2] = \frac{1}{\omega_{11}\omega_{22} - \omega_{12}\omega_{21}}\left(\omega_{22} - \frac{(\omega_{22}\beta_0 - \omega_{12})^2}{\sigma^2}\right)\Theta(\Pi_0, \Pi_0) = \frac{\Theta(\Pi_0, \Pi_0)}{\sigma^2}.$$

The conclusion follows.

APPENDIX D: PROOF OF PROPOSITION 1

Note that, for $a \equiv (0,1)'$,

$$E[U^{n\prime}P^nU^na] = M^{n\prime}P^nM^na + r_n\Omega a$$

and

$$\frac{1}{n}\,\text{Var}[U^{n\prime}P^nU^na] = \frac{1}{n}\,a'\Omega aM^{n\prime}P^nM^n + \frac{1}{n}\,a'M^{n\prime}P^nM^na\Omega + \frac{1}{n}\,\Omega aa'M^{n\prime}P^nM^n + \frac{1}{n}\,M^{n\prime}P^nM^naa'\Omega + \frac{1}{n}\,r_n a'\Omega a\Omega + \frac{1}{n}\,r_n\Omega aa'\Omega \to \Theta(\Pi,\Pi)\begin{bmatrix}\omega_{11} + \beta^2\omega_{22} + 2\omega_{12}\beta & 2\omega_{22}\beta + 2\omega_{12} \\ 2\omega_{22}\beta + 2\omega_{12} & 4\omega_{22}\end{bmatrix},$$

which follows easily from Bekker (1994, Lemma 2). Therefore, by Bekker (1994, Lemma 2), we would have

$$\frac{1}{\sqrt{n}}\,(U^{n\prime}P^nU^na - E[U^{n\prime}P^nU^na]) \Rightarrow \mathcal{N}\left(0, \Theta(\Pi,\Pi)\begin{bmatrix}\omega_{11} + \beta^2\omega_{22} + 2\omega_{12}\beta & 2\omega_{22}\beta + 2\omega_{12} \\ 2\omega_{22}\beta + 2\omega_{12} & 4\omega_{22}\end{bmatrix}\right).$$

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

Recall that condition (1) implies that

$$\sqrt{n}\left(\frac{1}{n}E[U^{n\prime}P^nU^na]-\begin{pmatrix}\beta\Theta(\Pi,\Pi)\\ \Theta(\Pi,\Pi)\end{pmatrix}\right)=o(1).$$

Therefore, by the delta method, we obtain (3).

APPENDIX E: TECHNICAL LEMMAS FOR SECTION 3

LEMMA 5.

$$E\begin{pmatrix}\bar{S}_{11}\\ \bar{S}_{12}\\ \bar{S}_{22}\\ S_{11}^{\perp}\\ S_{12}^{\perp}\\ S_{22}^{\perp}\end{pmatrix}=\begin{pmatrix}\sum_{i=1}^{n}(z_i^{n\prime}\Pi^n)^2\beta^2+K\omega_{11}\\ \sum_{i=1}^{n}(z_i^{n\prime}\Pi^n)^2\beta+K\omega_{12}\\ \sum_{i=1}^{n}(z_i^{n\prime}\Pi^n)^2+K\omega_{22}\\ (n-K)\omega_{11}\\ (n-K)\omega_{12}\\ (n-K)\omega_{22}\end{pmatrix}.$$

Proof. This proof follows easily from Bekker (1994, Lemma 1) and is omitted. ■

LEMMA 6. *$\bar{S}$ and $S^{\perp}$ are independent of each other.*

Proof. This proof follows easily from normality. ■

LEMMA 7.

$$\mathrm{Var}(\bar{S}_{11})=4\omega_{11}\beta^2\sum_{i=1}^{n}(z_i^{n\prime}\Pi^n)^2+2K\omega_{11}^2,$$

$$\mathrm{Var}(\bar{S}_{12})=(\omega_{11}+\beta^2\omega_{22}+2\omega_{12}\beta)\sum_{i=1}^{n}(z_i^{n\prime}\Pi^n)^2+K(\omega_{11}\omega_{22}+\omega_{12}^2),$$

$$\mathrm{Var}(\bar{S}_{22})=4\omega_{22}\sum_{i=1}^{n}(z_i^{n\prime}\Pi^n)^2+2K\omega_{22}^2,$$

$$\mathrm{Cov}(\bar{S}_{11},\bar{S}_{12})=2(\omega_{11}\beta+\omega_{12}\beta^2)\sum_{i=1}^{n}(z_i^{n\prime}\Pi^n)^2+2K\omega_{11}\omega_{12},$$

$$\mathrm{Cov}(\bar{S}_{11},\bar{S}_{22})=4\omega_{12}\beta\sum_{i=1}^{n}(z_i^{n\prime}\Pi^n)^2+2K\omega_{12}^2,$$

$$\mathrm{Cov}(\bar{S}_{12},\bar{S}_{22})=2(\omega_{22}\beta+\omega_{12})\sum_{i=1}^{n}(z_i^{n\prime}\Pi^n)^2+2K\omega_{22}\omega_{12},$$

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

and

$$\mathrm{Var}(S_{11}^{\perp}) = 2(n-K)\omega_{11}^2,$$

$$\mathrm{Var}(S_{12}^{\perp}) = (n-K)(\omega_{11}\omega_{22}+\omega_{12}^2),$$

$$\mathrm{Var}(S_{22}^{\perp}) = 2(n-K)\omega_{22}^2,$$

$$\mathrm{Cov}(S_{11}^{\perp}, S_{12}^{\perp}) = 2(n-K)\omega_{11}\omega_{12},$$

$$\mathrm{Cov}(S_{11}^{\perp}, S_{22}^{\perp}) = 2(n-K)\omega_{12}^2,$$

$$\mathrm{Cov}(S_{12}^{\perp}, S_{22}^{\perp}) = 2(n-K)\omega_{22}\omega_{12}.$$

Proof. This proof follows as a consequence of Bekker (1994, Lemma 2) and is omitted. ■

LEMMA 8. *Let* $\bar{\Lambda}$ *denote a symmetric* 3×3 *matrix such that*

$$\bar{\Lambda}_{1,1} = 4\omega_{11}\Theta\beta^2 + 2\alpha\omega_{11}^2,$$

$$\bar{\Lambda}_{1,2} = 2\omega_{11}\Theta\beta + 2\beta^2\Theta\omega_{12} + 2\alpha\omega_{11}\omega_{12},$$

$$\bar{\Lambda}_{1,3} = 4\beta\Theta\omega_{12} + 2\alpha\omega_{12}^2,$$

$$\bar{\Lambda}_{2,2} = \omega_{11}\Theta + \beta^2\Theta\omega_{22} + 2\Theta\omega_{12}\beta + \alpha\omega_{11}\omega_{22} + \alpha\omega_{12}^2,$$

$$\bar{\Lambda}_{2,3} = 2\omega_{22}\Theta\beta + 2\Theta\omega_{12} + 2\alpha\omega_{22}\omega_{12},$$

$$\bar{\Lambda}_{3,3} = 4\omega_{22}\Theta + 2\alpha\omega_{22}^2.$$

Then

$$\frac{1}{\sqrt{n}}\left(\begin{pmatrix}\bar{S}_{11}\\ \bar{S}_{12}\\ \bar{S}_{22}\end{pmatrix} - E\begin{pmatrix}\bar{S}_{11}\\ \bar{S}_{12}\\ \bar{S}_{22}\end{pmatrix}\right) \Rightarrow \mathcal{N}(0, \bar{\Lambda}).$$

Proof. This proof follows from Bekker (1994, Lemma 2) and Lemma 7. ■

LEMMA 9. *Let* $\Lambda^{\perp}$ *denote a symmetric* 3×3 *matrix such that*

$$\Lambda_{1,1}^{\perp} = 2(1-\alpha)\omega_{11}^2,$$

$$\Lambda_{1,2}^{\perp} = 2(1-\alpha)\omega_{11}\omega_{12},$$

$$\Lambda_{1,3}^{\perp} = 2(1-\alpha)\omega_{12}^2,$$

$$\Lambda_{2,2}^{\perp} = (1-\alpha)\omega_{11}\omega_{22} + (1-\alpha)\omega_{12}^2,$$

$$\Lambda_{2,3}^{\perp} = 2(1-\alpha)\omega_{22}\omega_{12},$$

$$\Lambda_{3,3}^{\perp} = 2(1-\alpha)\omega_{22}^2.$$

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

Then,

$$\frac{1}{\sqrt{n}}\left(\begin{pmatrix} S_{11}^{\perp} \\ S_{12}^{\perp} \\ S_{22}^{\perp} \end{pmatrix} - E\begin{pmatrix} S_{11}^{\perp} \\ S_{12}^{\perp} \\ S_{22}^{\perp} \end{pmatrix}\right) \Rightarrow N(0, \Lambda^{\perp}).$$

Proof. This proof follows from Bekker (1994, Lemma 2) and Lemma 7. ■

LEMMA 10. *Assume that* $k/n \to \alpha + o(n^{-1/2})$ *and that* $(1/n)\sum_{i=1}^{n}(z_i^{n\prime}\Pi^n)^2$ *is fixed at* $\Theta(\Pi, \Pi)$. *We then have*

$$\sqrt{n}\left(\begin{pmatrix} n^{-1}\bar{S}_{11} \\ n^{-1}\bar{S}_{12} \\ n^{-1}\bar{S}_{22} \\ n^{-1}S_{11}^{\perp} \\ n^{-1}S_{12}^{\perp} \\ n^{-1}S_{22}^{\perp} \end{pmatrix} - \begin{pmatrix} \Theta(\Pi,\Pi)\beta^2 + \alpha\omega_{11} \\ \Theta(\Pi,\Pi)\beta + \alpha\omega_{12} \\ \Theta(\Pi,\Pi) + \alpha\omega_{22} \\ (1-\alpha)\omega_{11} \\ (1-\alpha)\omega_{12} \\ (1-\alpha)\omega_{22} \end{pmatrix}\right) \Rightarrow \mathcal{N}\left(0, \begin{bmatrix} \bar{\Lambda} & 0 \\ 0 & \Lambda^{\perp} \end{bmatrix}\right),$$

where $\bar{\Lambda}$ *and* $\Lambda^{\perp}$ *are defined in Lemmas 8 and 9.*

Proof. This proof follows from Lemmas 5, 6, 8, and 9. ■

LEMMA 11. *Suppose that the log (joint) likelihood is given by (4). With a diffuse prior on* ξ*'s, the posterior likelihood for* (β, Ψ) *is proportional to*

$$(\sqrt{\det(\Psi)})^n \exp\left\{-\frac{1}{2}\left[\operatorname{trace}(\Psi(\bar{S} + S^{\perp})) - \frac{(\beta, 1)\Psi\bar{S}\Psi(\beta, 1)'}{\psi_{11}\beta^2 + 2\psi_{12}\beta + \psi_{22}}\right]\right\}.$$

Here, ξ*'s are defined in (6).*

Proof. With ξ's as defined in (6), we can see that the individual likelihood of the transformed model is proportional to

$$\sqrt{\det(\Psi)}\exp\left(-\frac{1}{2}\eta_i\right),$$

where

$$\eta_i = \left(y_i - \beta\Upsilon\sum_{k=1}^{K}\xi_k z_{ki}, x_i - \Upsilon\sum_{k=1}^{K}\xi_k z_{ki}\right)\Psi\begin{pmatrix} y_i - \beta\Upsilon\sum_{k=1}^{K}\xi_k z_{ki} \\ x_i - \Upsilon\sum_{k=1}^{K}\xi_k z_{ki} \end{pmatrix}$$

$$= (y_i, x_i)\Psi\begin{pmatrix} y_i \\ x_i \end{pmatrix} - 2\Upsilon(z_i'\xi)(\beta, 1)\Psi\begin{pmatrix} y_i \\ x_i \end{pmatrix} + \Upsilon^2(z_i'\xi)^2(\beta, 1)\Psi\begin{pmatrix} \beta \\ 1 \end{pmatrix}.$$

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press

Letting

$$\Phi_1 = \Upsilon^2(\beta, 1)\Psi \begin{pmatrix} \beta \\ 1 \end{pmatrix} \sum_{i=1}^{n} z_i z_i' = \sum_{i=1}^{n} z_i z_i',$$

$$\Phi_2 = \Upsilon(\beta, 1)\Psi \sum_{i=1}^{n} \begin{pmatrix} y_i \\ x_i \end{pmatrix} z_i',$$

$$\Phi_3 = \sum_{i=1}^{n} (y_i, x_i)\Psi \begin{pmatrix} y_i \\ x_i \end{pmatrix},$$

we can then write

$$\sum_{i=1}^{n} \eta_i = \xi'\Phi_1\xi - 2\Phi_2\xi + \Phi_3 = (\xi - \phi)'\Phi_1(\xi - \phi) + \Phi_3 - \phi'\Phi_1\phi,$$

where $\phi' \equiv \Phi_2\Phi_1^{-1}$. The joint likelihood is therefore proportional to

$$\mathcal{L} \equiv (\sqrt{\det(\Psi)})^n \exp\left(-\frac{1}{2}[\Phi_3 - \Phi_2\Phi_1^{-1}\Phi_2']\right) \exp\left(-\frac{1}{2}(\xi - \phi)'\Phi_1(\xi - \phi)\right).$$

Integrating $\mathcal{L}$ with respect to a diffuse prior on ξ's, we obtain the posterior proportional to

$$\mathbb{L} \equiv \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathcal{L} d\xi_1 \cdots d\xi_k \propto \frac{(\sqrt{\det(\Psi)})^n}{\sqrt{\det(\Phi_1)}} \exp\left(-\frac{1}{2}[\Phi_3 - \Phi_2\Phi_1^{-1}\Phi_2']\right).$$

Because

$$\Phi_3 - \Phi_2\Phi_1^{-1}\Phi_2' = \operatorname{trace}[\Psi(\bar{S} + S^{\perp})] - \left[(\beta, 1)\Psi \begin{pmatrix} \beta \\ 1 \end{pmatrix}\right]^{-1} (\beta, 1)\Psi\bar{S}\Psi \begin{pmatrix} \beta \\ 1 \end{pmatrix}$$

and

$$\det \Phi_1 = \det\left[\sum_{i=1}^{n} z_i z_i'\right],$$

we obtain the desired conclusion. ■

https://doi.org/10.1017/S0266466602181084 Published online by Cambridge University Press