Econometric Theory, **32**, 2016, 30–70.
doi:10.1017/S0266466614000851

AVERAGING OF AN INCREASING NUMBER OF MOMENT CONDITION ESTIMATORS

XIAOHONG CHEN
Yale University

DAVID T. JACHO-CHÁVEZ
Emory University

OLIVER LINTON
University of Cambridge

We establish the consistency and asymptotic normality for a class of estimators that are linear combinations of a set of $\sqrt{n}$-consistent nonlinear estimators whose cardinality increases with sample size. The method can be compared with the usual approaches of combining the moment conditions (GMM) and combining the instruments (IV), and achieves similar objectives of aggregating the available information. One advantage of aggregating the estimators rather than the moment conditions is that it yields robustness to certain types of parameter heterogeneity in the sense that it delivers consistent estimates of the mean effect in that case. We discuss the question of optimal weighting of the estimators.

1. INTRODUCTION

In this paper we derive the properties of an estimator formed by taking linear combinations of an increasing number of $\sqrt{n}$-consistent estimators obtained from a sequence of moment restrictions. The usual approach here is to combine the moment restrictions into a single objective function, either by combining the instruments or by stacking all the moment conditions together (see Han and Phillips, 2006, for a recent contribution); our proposal involves estimating the parameters several times from subsets of the moment conditions and then combining the resulting estimators in a linear fashion. The proposed methodology has the advantage that one can see how much variation there is in the parameter estimates

Earlier versions of this paper circulated under the titles "An Alternative Way of Computing Efficient Instrumental Variable Estimators" and "Averaging of Moment Condition Estimators." We would like to thank the co-editor Guido Kuersteiner, as well as four anonymous referees for valuable comments and suggestions. We thank STICERD, the NSF, and the ESRC for financial support. This paper was partly written while Oliver Linton was at Universidad Carlos III de Madrid-Banco Santander Chair of Excellence, and Oliver Linton thanks them as well as the ERC for financial support. Address correspondence to Oliver Linton, Department of Economics, University of Cambridge, Austin Robinson Building, Sidgwick Avenue, Cambridge CB3 9DD, United Kingdom; e-mail: obl20@cam.ac.uk.

(which are presumably in an interpretable scale unlike moment conditions, say), and how much weight an optimal combination would place on them. In cases where there is truly little variation, the practitioner can presumably do with very simple inference rules. This estimator is also possibly useful in high-dimensional models where there is a larger set of instruments than sample observations. A final advantage of the method of combining estimators rather than combining moment conditions arises under random coefficient type parameter heterogeneity. In that case, combining estimators can provide consistent estimates of the mean effect, whereas combining moment conditions produces an estimator of a parameter that is not so easy to interpret.

The idea of combining estimates is not new, and has been used to improve finite sample properties of estimators and forecasts. Granger and Jeon (2004) provide an useful discussion. For example, Sawa (1973) considered combining k-class estimators in simultaneous equations systems, for the reason of reducing bias. Breiman (1996, 1999) introduced the idea of bagging, which is based on using bootstrap resamples to compute a large(ish) sample of subsample estimators and then combining them. Watson (2003) and Stock and Watson (1999) propose various methods for combining large numbers of predictors to improve forecasting performance. More recent developments in these settings can be found in Hansen (2007, 2008, 2009, 2010) and Hansen and Racine (2012). In the nonparametric literature, Gray and Schucany (1972) and Bierens (1987) have proposed jackknife estimators that combine different kernel smoothers in order to reduce bias. Similarly, Kotlyarova and Zinde-Walsh (2006, 2007) and Schafgans and Zinde-Walsh (2010) have proposed combining kernel smoothers calculated with different bandwidths and kernel functions to construct robust estimators of densities and density-weighted average derivatives respectively. In additive nonparametric regression, integration, or averaging, has been shown to improve rates of convergence and to eliminate nuisance parameters (see e.g., Linton and Nielsen, 1995). Similarly, in high frequency econometrics, the TSRV (Two Scales Realized Volatility) estimator combines linearly a large number of subsample based estimators (i.e., Zhang, Mykland, and Ait-Sahalia, 2005).

Our method is in effect a generalization of the classical method of minimum chi-squared or minimum distance discussed in Malinvaud (1966) and Rothenberg (1973), which was conceived as a way of imposing equality restrictions in estimation via first estimating an unrestricted model and then finding the best combination of the unrestricted estimators that imposes the restrictions, to the case where the number of estimators and the number of restrictions increase at the same rate. In a number of cases this strategy is preferable to solving the constrained estimation problem directly. In our case, the best combination is linear with weights that add up to one.

There is a vast literature on estimating models defined through conditional moment restrictions. We just mention one paper that is particularly relevant to our study, Koenker and Machado (1999). They considered a similar problem albeit restricted to certain linear models and to a rather specific estimator. They proved

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

that a sufficient condition for the usual asymptotics for generalized method of moments estimation (GMM) to be valid when the number of unconditional moment equations τ increases with n is that $\tau^3/n \to 0$.[1] Their results can be interpreted as a warning not to include too many moment conditions in GMM: that the consequences of doing so are not just that no improvement is made, but that the distributional approximation can potentially break down. Our objective is quite different, and we deal with nonlinear models.

In linear models and with efficiency in mind, the proposed method can also be viewed as an alternative to choosing a subset of instruments among a large class of valid instruments (see e.g., Donald and Newey, 2001; Kuersteiner and Okui, 2010). For example, consider the case where an unknown but *fixed* number of instruments yields nonidentified (Lobato and Dominguez, 2004) or weakly identified (Stock and Wright, 2000) unconditional moment restrictions, then a simple averaging of the resulting estimates would make their contributions to the resulting average bias and variance vanish with sample size. On the other hand, if efficiency is not of primary importance, knowledge of the quality of instruments can be readily incorporated into the proposed estimator via the weighting scheme.

We first establish consistency and $\sqrt{n}$-asymptotic normality of a class of estimators that involve finite linear combinations of an infinite dimensional set of estimators, where the cardinality of the linear combinations increases with sample size. The class of estimators considered is allowed to include those computed from discontinuous criterion functions that are nonlinear in the parameters and data. We also establish that a member of our class of estimators achieves the semiparametric efficiency bound for the conditional moment model. We propose a scheme for estimating the optimal weights and show that this is consistent. We conclude by presenting results of two Monte Carlo experiments showing how our procedure works in practice.

2. THE MODEL FRAMEWORK AND ESTIMATION

We observe an independent and identically distributed (i.i.d.) sample $\{Z_i\}_{i=1}^n \in \mathbb{R}^d$. We suppose that there are a set of moment conditions g_j for $j \in \mathcal{J}$, where $\mathcal{J}$ is some set (of possibly infinite cardinality) with $g_j \in \mathbb{R}^q$ such that there is a unique $\theta_0 \in \Theta$, where Θ is a compact subset of $\mathbb{R}^p$, for which

$$E[g_j(Z_i, \theta_0)] = 0. \tag{2.1}$$

For simplicity, we shall assume that $q = p$ so that each considered moment condition yields exact identification. Each moment condition itself can be used for estimation through the sample equivalent

$$G_{nj}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g_j(Z_i, \theta).$$

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

There are three ways of aggregating the information in the moment conditions (2.1).

First, consider the objective function

$$Q_n(\theta) = \sum_{j \in \mathcal{J}} G_{nj}(\theta)^\top W_{nj} G_{nj}(\theta), \tag{2.2}$$

where W_{nj} are weights. This corresponds to a GMM objective function with block-diagonal weighting matrix. Then minimize $Q_n(\theta)$ with respect to θ. Second, combine the moment functions before averaging over the sample, i.e., let

$$M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j \in \mathcal{J}} W_{nj} g_j(Z_i, \theta) \right), \tag{2.3}$$

where W_{nj} are weights. Then find a zero of the function $M_n(\theta)$.

Both approaches are widely used and studied when $\mathcal{J}$ has finite cardinality (see e.g., Hansen, 1982; Newey, 1990; Chamberlain, 1992). In each case there is a question about the optimal choice of weighting sequence W_{nj}, and in each case this issue is resolved, and one can broadly achieve the same results by either approach.

We consider a third approach based on combining individual estimators. We define the estimators $\widehat{\theta}_j$, $j \in \mathcal{J}$ as any sequence that satisfies

$$G_{nj}\left(\widehat{\theta}_j\right) = \frac{1}{n} \sum_{i=1}^{n} g_j\left(Z_i, \widehat{\theta}_j\right) = o_p\left(n^{-1/2}\right). \tag{2.4}$$

For each $j \in \mathcal{J}$, this problem is parametric and will result in a $\sqrt{n}$-consistent and asymptotically normal estimator $\widehat{\theta}_j$ (under standard conditions). We combine these estimators in a linear fashion to produce a new estimator

$$\widehat{\theta} = \sum_{j \in \mathcal{J}} W_{nj} \widehat{\theta}_j, \tag{2.5}$$

where W_{nj} are some matrix weights, possibly stochastic, that sum to the identity. This defines a class of estimators $\mathcal{E}$ indexed by the weighting matrices $\{W_{nj}, j \in \mathcal{J}\}$. In the finite dimensional $\mathcal{J}$ case, the estimator (2.5) is a member of the class of minimum distance estimators (see Rothenberg, 1973; Newey and McFadden, 1994). Specifically, suppose that $j = 1, \ldots \tau$, and consider the optimization problem

$$\left[\begin{pmatrix} \widehat{\theta}_1 \\ \vdots \\ \widehat{\theta}_\tau \end{pmatrix} - R\theta \right]^\top \Xi \left[\begin{pmatrix} \widehat{\theta}_1 \\ \vdots \\ \widehat{\theta}_\tau \end{pmatrix} - R\theta \right] \tag{2.6}$$

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

for some $\tau p \times \tau p$ weighting matrix Ξ and $\tau p \times \tau$ known vector R (e.g., $R = i_\tau \otimes I_p$, where i_τ is a $\tau \times 1$ vector of ones). This yields an estimator of the form (2.5), where $W_{nj} = [R^\top \Xi R]^{-1}[R^\top \Xi]_j$ and $[R^\top \Xi]_j$ represents the jth block of $R^\top \Xi$. In this case, the optimal weighting is known, Ξ should be proportional to the inverse of the asymptotic variance matrix of the unrestricted estimators. The paradigm based on combining estimators can also benefit from the interpretation of (2.5) as a portfolio of estimators, and the large amount of work that is currently being done on estimation of inverse covariance matrices may be useful in finding optimal combinations.

We make one further remark comparing the combining estimators approach with the combining moment conditions approach. Suppose that there is parameter heterogeneity so that for each j there is a unique $\theta_j \in \Theta \subseteq \mathbb{R}^p$ for which $E[g_j(Z_i, \theta_j)] = 0$ and

$$\theta_j = \theta_0 + \upsilon_j, \tag{2.7}$$

where υ_j are i.i.d. mean zero random variables (and independent of the data) (e.g., Chesher, 1984). In this case, the average of the estimators consistently estimates θ_0 (see, for example, Pesaran, 2006), whereas first combining the moment conditions will not generally yield consistent estimators of the parameter θ_0, except when g is linear in θ. So our method is robust to parameter heterogeneity in this sense (although the asymptotic variance of our estimator is different under those conditions). We do not focus on this case, but rather mention it as a possible further motivation for why averaging estimators may be preferable to averaging moment conditions.

We suppose that $\mathcal{J}$ may have infinite cardinality, or it may have a cardinality that is increasing with sample size n. Han and Phillips (2006) consider a similar setup except that they combine the moment conditions in the classical GMM way, i.e., (2.2), using identity weighting; they also allow for the possibility that some of their moment conditions provide only weak identification. Lee (2010) considers the case where τ is fixed but each estimator may have a different rate of convergence (see also Antoine and Renault, 2012). We work with a situation where each estimator is $\sqrt{n}$-consistent, however, in our case the asymptotic variance V_{jj} of the jth estimator may increase to infinity with j, which essentially reflects the same phenomenon of different precision across the estimators.

The key issue we address here is to determine the asymptotic distribution of the estimators in $\mathcal{E}$. We also determine the optimal weighting and show that we can achieve efficiency within this class of estimators. In some special cases we can show that the resulting estimator will achieve a semiparametric efficiency bound where the moment conditions themselves completely reflect all the model information.

Even though each criterion function G_{nj} is a nonlinear function of θ, the overall computational costs of (2.5) may not be so great, since one can use the estimates $\widehat{\theta}_j$ for some j as starting values for computing other estimates $\widehat{\theta}_k$ for $k \neq j$.

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

Additional computational issues arise in connection with the weights W_{nj} but these are discussed below.

3. EXAMPLES

We give a number of examples. Some are very well studied, but they are included here just to clarify some concepts.

Example 1 (Classical two stage least squares in simultaneous equations) Suppose that[2]

$$y_{1i} = \theta y_{2i} + \varepsilon_i; \qquad y_{2i} = \pi_2^\top X_i + u_i,$$

where $(\varepsilon_i, u_i)^\top$ are i.i.d. error terms, $E[\varepsilon_i|X_i] = 0$, $E[u_i|X_i] = 0$, and $X_i \in \mathbb{R}^k$. Consider the moment conditions

$$g_j(Z_i, \theta) = (y_{1i} - \theta y_{2i}) X_{ji},$$

for $j = 1, \ldots, k$. Then let $\widehat{\theta}_j$ be the corresponding estimator that solves (2.4). In fact

$$\widehat{\theta}_j = \frac{\sum_{i=1}^n X_{ji} y_{1i}}{\sum_{i=1}^n X_{ji} y_{2i}} = \frac{\sum_{i=1}^n \widehat{y}_{2i}^j y_{1i}}{\sum_{i=1}^n \widehat{y}_{2i}^j y_{2i}}, \tag{3.1}$$

where $\widehat{y}_{2i}^j = \widehat{\pi}_{2j} X_{ji}$, and $\widehat{\pi}_{2j}$ are the least squares estimates obtained from the reduced form regression of y_{2i} on the single instrument X_{ji} for $j = 1, \ldots, k$. Our estimator is

$$\widehat{\theta} = \sum_{j=1}^k W_{nj} \widehat{\theta}_j, \tag{3.2}$$

where W_{nj} are scalar weights that satisfy $\sum_{j=1}^k W_{nj} = 1$.

For comparison, the two stage least squares (2SLS) estimator is

$$\widetilde{\theta} = \frac{\sum_{i=1}^n \widehat{y}_{2i} y_{1i}}{\sum_{i=1}^n (\widehat{y}_{2i})^2} = \frac{\sum_{i=1}^n \widehat{y}_{2i} y_{1i}}{\sum_{i=1}^n \widehat{y}_{2i} y_{2i}}, \tag{3.3}$$

where $\widehat{y}_{2i} = \widehat{\pi}_2^\top X_i$ and $\widehat{\pi}_2$ is the vector of least squares estimates obtained from the reduced form regression of y_{2i} on all the instruments $X_i = (X_{1i}, \ldots, X_{ki})^\top$. We note that there is a choice of W_{nj} that makes $\widehat{\theta}$ asymptotically equivalent to the 2SLS estimator $\widetilde{\theta}$ (see below). It is worth noting that a related result can be found in Swamy (1970) for random coefficient panel data models, namely a particular weighted average of the individual specific slope estimators is equivalent to the optimal GLS estimator.

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

The classical minimum distance estimator (generalized indirect least squares) exploits the relationship between the reduced form coefficients and the structural parameter, i.e., $\pi_{1j}/\pi_{2j} = \theta$, where $\pi_{\ell j} = E(y_{\ell i} X_{ji})/E(X_{ji}^2)$ are the parameters of the reduced form of $y_{\ell i}$ on X_{ji} for $\ell = 1, 2$ and $j = 1, \ldots, k$ (the estimator is a linear combination of $\widehat{\pi}_{1j}/\widehat{\pi}_{2j}$, where $\widehat{\pi}_{\ell j}$ are the corresponding reduced form estimators) (see Rothenberg, 1973).[3]

Now suppose that $k = \infty$, i.e.,

$$y_{2i} = \sum_{j=1}^{\infty} \pi_{2j} X_{ji} + u_i. \tag{3.4}$$

This allows for the possibility that all of the infinite number of instruments matter in the reduced form. In order for the right hand side of (3.4) to be well defined, we may either assume that: (a) $\sum_{j=1}^{\infty} \pi_{2j}^2 < \infty$ and $\sup_{1 \geq j} E X_{ji}^2 < \infty$, or (b) $\sup_{1 \geq j} |\pi_{2j}| < \infty$ and $\sum_{k=1}^{\infty} E X_{ki}^2 < \infty$. Let $\sigma_j^2 = E(X_{ji}^2)$, then (b) requires that σ_j^2 goes to zero at a rate faster than j^{-1} as $j \to \infty$. By changing variables to X_{ji}/σ_j the parameters become $\pi_{2j} \cdot \sigma_j$ and we are in case (a). In the sequel we shall restrict attention to case (a) as this seems more common in applications. This case is consistent with the process where $X_{ji} = \varphi_j(X_i)$ where X_i is some observed common covariate of fixed dimension and φ_j are known basis functions. It is also consistent with the case that X_{ji} are separate covariates. We do not impose a sparsity property (see e.g., Belloni, Chen, Chernozhukov, and Hansen, 2012), i.e., we allow all the π_{2j} to be nonzero.

The (approximate) 2SLS estimator is computed by truncating the sum to length τ and then regressing y_{2i} on the covariates $X_{1i}, \ldots, X_{\tau i}$. This may run into practical issues when τ is even moderately large. Instead, our procedure involves computing (3.1) for all $j = 1, 2, \ldots$, and compute (3.2) with k replaced by the truncation parameter τ. The individual regressions are very easy to compute. Furthermore, τ can be taken to be (much) larger than sample size (provided the weights are chosen appropriately). An issue arises only when trying to combine many estimators in an optimal way where the optimal weighting scheme may be hard to estimate (in the same way that the optimal instrument is hard to estimate).

Example 1P (Pathological case)
This example was suggested by a referee. Suppose that

$$g_j(Z_i, \theta) = (y_i - \theta X_i) X_i^{j-1},$$

where $E[\varepsilon_i | X_i] = 0$ with $\varepsilon_i = y_i - \theta_0 X_i$. In this case,

$$E G_{nj}(\theta) = (\theta_0 - \theta) E(X_i^j). \tag{3.5}$$

If X_i are symmetrically distributed around zero with well defined moment generating function, then $E(X_i^{2j-1}) = 0$ for all $j = 1, \ldots$, that is, $E G_{nj}(\theta) = 0$ for

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

all θ for odd j, although $EG_{nj}(\theta) = 0$ if and only if $\theta = \theta_0$ for even j. In this pathological case, $\widehat{\theta}_{2j-1},\ j = 1, 2, \ldots$ are inconsistent while $\widehat{\theta}_{2j},\ j = 1, 2, \ldots$ are consistent.

Example 2 (An infinite number of instruments)
This is a straightforward generalization of the linear case considered above. Specifically, suppose that $Y_i \in \mathbb{R}^D$, $X_i \in \mathbb{R}^\infty$, and $A_j(X_i) \in \mathbb{R}^p,\ j = 1, 2, \ldots$. Then suppose that

$$E[A_j(X_i)\rho(Y_i, \theta_0)] = 0,$$

where $\rho(y, \theta)$ is a potentially nonlinear "residual" vector, and $\theta \in \mathbb{R}^p$. This model includes many others as special cases. For estimation we may solve

$$G_{nj}(\widehat{\theta}_j) = o_p(n^{-1/2}),$$

$$G_{nj}(\theta) = \frac{1}{n}\sum_{i=1}^{n} A_j(X_i)\rho(Y_i, \theta), \tag{3.6}$$

for each j to yield estimators $\widehat{\theta}_j$ that are $\sqrt{n}$-consistent under some conditions. In general, this is a nonlinear system of equations for each j, but we may use $\widehat{\theta}_1$ as starting values for subsequent j, and even compute linearized estimators. We propose to combine these estimators using some weighting sequence W_{nj} as in (2.5).

There are two generally different settings where this arises. In the first case, X_i is genuinely large dimensional. In the second case, X_i is actually finite dimensional but the functions A_j vary.

Example 3 (Semiparametric instrumental variables)
Suppose that $Z_i^\top = (Y_i^\top, X_i^\top)$, and that there is a unique $\theta_0 \in \Theta \subseteq \mathbb{R}^p$ satisfying the conditional moment conditions

$$E[\rho(Z_i, \theta_0)\,|X_i\,] = 0, \tag{3.7}$$

with probability one, where $\rho(z, \theta)$ is a scalar residual function. This implies the unconditional moment conditions

$$E[A_j(X_i)\rho(Z_i, \theta_0)] = 0, \tag{3.8}$$

for any $p \times 1$ measurable vector $A_j(X_i)$ (for which the expectation exists).

Suppose that $E[\rho(Z_i, \theta_0)^2\,|X_i\,] = \sigma^2(X_i)$ is positive with probability one, and that $D_0(X_i) = (\partial E\,[\rho(Z_i, \theta)|X_i]\,/\partial\theta)_{\theta=\theta_0}$ exists with probability one. In this case, the optimal (instrumental variables) matrix is proportional to $A_{\text{oiv}}(X_i) = D_0(X_i)\sigma^{-2}(X_i)$, and the resulting optimal instrumental variables (oiv)—or optimal GMM—estimator $\widetilde{\theta}_{\text{oiv}}$ has asymptotic variance $\Sigma_{\text{oiv}} = \{E[\sigma^{-2}(X_i)D_0(X_i) \times$

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

$D_0(X_i)^\top]\}^{-1}$—see, for example, Hansen (1985), Chamberlain (1987), and Newey (1990, 1993) for smooth ρ and Chen and Pouzo (2009) for nonsmooth ρ. Suppose that the optimal matrix $A_{\text{oiv}}(\cdot)$ can be represented, in an L_2 sense, by the series expansion $A_{\text{oiv}}(x) = \sum_{j=1}^{\infty} \beta_{j0} A_j(x)$, where $A_j(\cdot)$ are known basis functions chosen by the practitioner, while β_{j0} are unknown coefficients determined uniquely by the basis. For notational convenience we shall allow A_j to be $p \times 1$ vectors; in general, β_{j0} depends on θ_0 and is a $p \times p$ matrix. A common approach here is to estimate the coefficients β_{j0} (by say series approximation, see e.g., Newey, 1990) and then to let $\widehat{A}_\theta(x) = \sum_{j=1}^{\tau(n)} \widehat{\beta}_j(\theta) A_j(x)$, where $\tau(n)$ is some truncation sequence that goes to infinity with sample size but at a slow rate. Then let $\widetilde{\theta}_{\text{oiv}}$ be any sequence that satisfies $n^{-1} \sum_{i=1}^{n} \widehat{A}_{\widetilde{\theta}_{\text{oiv}}}(X_i) \rho(Z_i, \widetilde{\theta}_{\text{oiv}}) = o_p(n^{-1/2})$. In current parlance this would be called a continuously updated oiv estimator. An alternative method is to use some preliminary consistent estimator of θ_0 to first construct a consistent estimator of A_{oiv}, and then to solve a similar first order condition with the estimated instrument. Newey (1990, 1993) worked with linearized two-step estimators that approximate such solutions. He showed that such estimators are asymptotically equivalent to the instrumental variable procedure based on knowing the optimal instrument function A_{oiv} and computing solutions $\widetilde{\theta}_{\text{oiv}}$ to

$$\frac{1}{n} \sum_{i=1}^{n} A_{\text{oiv}}(X_i) \rho(Z_i, \widetilde{\theta}_{\text{oiv}}) = o_p(n^{-1/2}).$$

See Newey and McFadden (1994) for discussion. There have been a number of alternative suggestions made more recently with a view to improving small sample performance. Newey and Smith (2004) contains an excellent review of this literature.

We can approach this estimation problem from our perspective: instead of combining instruments by preliminary estimation, we combine the estimators.

Example 4 (Maximum likelihood)
Suppose that $X_i \sim F(.; \theta_0)$, where θ_0 is an unknown scalar parameter. When F has a density f, one can use maximum likelihood estimation (MLE); specifically, choose θ to maximize the log likelihood function

$$\sum_{i=1}^{n} \log f(X_i; \theta).$$

An alternative approach to estimation of θ can be based on the cdf F such as the minimum distance estimator (e.g., Koul and Stute, 1999), which involves minimizing $\int [F_n(x) - F(x; \theta)]^2 \, dw(x)$ with respect to θ, where w is a weighting function. Instead consider the estimators that solve

$$F_n(x_j) - F(x_j; \widehat{\theta}_j) = o_p(n^{-1/2})$$

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

for points x_j, $j = 1, 2, \ldots$, which can be cast in the form of (2.4). We then combine these estimators linearly according to (2.5). In this case, it is natural to think of $\mathcal{J}$ as being isomorphic to the real line but the weights can only pick out a finite number of estimators. In some special cases of interest these estimators can be expressed in closed form. For example, suppose that θ is a location parameter and $F(x;\theta) = F(x-\theta)$, where F is strictly monotonic. Then, $\widehat{\theta}_j = x_j - F^{-1}(F_n(x_j))$, which can then be plugged into (2.5). Suppose that we take some $\tau < \infty$ and let $x_j = F_n^{-1}(j/\tau)$, $j = 1, \ldots, \tau$, then $\widehat{\theta}_j = F_n^{-1}(j/\tau) - F^{-1}(j/\tau)$. Then, let $\widehat{\theta} = \sum_{j=1}^{\tau} W_{nj}\widehat{\theta}_j$ for some weighting sequence $\{W_{nj}\}$. Zhao and Xiao (forthcoming) have pursued this estimation scheme in quantile regression.

4. LARGE SAMPLE PROPERTIES

We begin by defining the sample and population first order conditions. For $j \in \mathcal{J}$, let

$$G_{nj}(\theta) \equiv \frac{1}{n}\sum_{i=1}^{n} g_j(Z_i, \theta) \text{ and } G_j(\theta) \equiv EG_{nj}(\theta). \tag{4.1}$$

We do not assume that the functions $G_{nj}(\theta)$ are differentiable or even continuous, although smoothness conditions are imposed on the expectation $G_j(\theta)$. On the one hand, this level of generality allows the cases of quantile regression estimators (e.g., Koenker and Gilbert, 1978), Huber's (1967) M-estimators, and simulation-based estimators (e.g., McFadden, 1989; Pakes and Pollard, 1989) to be covered by our theory, and on the other hand, for some of the arguments, we are only able to provide high level conditions on the sample and population first order conditions. In this sense, our results can apply more generally to any linear combination of estimators that have appropriate expansions. We will restrict our attention to the case where $\mathcal{J}$ is isomorphic to the set of positive integers, and we shall further consider the sample moment conditions to be from a finite subset $\mathcal{J}_n$ whose cardinality, $\tau = \tau(n)$, is allowed to grow with sample size. We will take pathwise asymptotics throughout so that $\tau(n)$ as $n \to \infty$, but we will be working with cases where the sequential limits (first $n \to \infty$, then $\tau \to \infty$) will yield the same limits (see e.g., Phillips and Moon, 1999). These sequential limit arguments are easier to understand.

4.1. Consistency

In this subsection we give our consistency result for the estimator (2.5). For simplicity, we let $\mathcal{J} = \mathcal{J}_n = \{1, 2, \ldots, \tau(n)\}$. We shall accommodate the case where some or even many of the moment conditions are not identified at all (i.e., the instruments are irrelevant), since in practice it may be hard to know whether a given instrument is relevant or not. We also allow a more mundane type of weak identification that occurs naturally as $\tau \to \infty$. To this end, let $\mathcal{J}_n^* = \{1, 2, \ldots, \tau^*(n)\}$

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

with $\tau^*(n) \leq \tau(n)$. Implicit here is that there is an ordering of the estimating equations or at least a classification into two categories: leading ones (strong) and nonleading ones (weak or even irrelevant). We do not require this classification to be known a priori, although the knowledge of this classification would allow for a weakening of other conditions. In practice, this classification could be based on individual t or F statistics associated with each estimator sequence. See Fan and Lv (2008) and Fan, Feng, and Song (2011) for alternative screening approaches in related problems. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalues of a real symmetric matrix A and $\|A\| = (\mathrm{tr}(A^\top A))^{1/2}$ for any matrix A.

Assumption A. Let $\theta_0 \in \Theta$ satisfy model (2.1).

(A1) The triangular array $\{W_{nj}\}_{j \in \mathcal{J}_n}$, $n = 1, \ldots$, satisfies

$$\sum_{j \in \mathcal{J}_n} W_{nj} = I_p \text{ and } \sup_{n \geq 1} \sum_{j \in \mathcal{J}_n} \|W_{nj}\| < \infty \text{ w.p.1.} \tag{4.2}$$

Here, $\tau(n)$ satisfies $\tau(n) \to \infty$ as $n \to \infty$. Furthermore, for $\tau^*(n) \to \infty$ with $\tau^*(n) \leq \tau(n)$

$$\sup_{n \geq 1} \sum_{j=\tau^*(n)+1}^{\tau(n)} ||W_{nj}|| \to 0. \tag{4.3}$$

(A2) For all $\delta > 0$ and $n \geq 1$, there is an $\epsilon_n(\delta) > 0$ (with possibly $\epsilon_n(\delta) \to 0$ as $n \to \infty$) such that

$$\min_{j \in \mathcal{J}_n^*} \inf_{\|\theta - \theta_0\| > \delta} \|G_j(\theta)\| \geq \epsilon_n(\delta) > 0.$$

(A3) For the sequences $\epsilon_n(\delta)$, $\tau(n)$, and $\tau(n)^*$ defined above, there exists a positive sequence $\epsilon_{1n} = o(1)$ with $\sup_n(\epsilon_{1n}/\epsilon_n(\delta)) < \infty$ such that

$$\max_{j \in \mathcal{J}_n^*} \left(\|G_{nj}(\widehat{\theta_j})\| - \inf_{\theta \in \Theta} \|G_{nj}(\theta)\| \right) = o_p(\epsilon_{1n}).$$

(A4) For the sequences $\epsilon_n(\delta)$, $\tau(n)$, and $\tau(n)^*$ defined above, there exists a positive sequence $\epsilon_{2n} = o(1)$ with $\sup_n(\epsilon_{2n}/\epsilon_n(\delta)) < \infty$ such that

$$\max_{j \in \mathcal{J}_n^*} \sup_{\theta \in \Theta} \|G_{nj}(\theta) - G_j(\theta)\| = o_p(\epsilon_{2n}).$$

The assumptions on the weights are quite weak and are satisfied by many suitable weighting sequences both random and nonrandom. For example, equal weighting $W_{nj} = 1/\tau(n) I_p$ satisfies Assumption 1, where I_p represents a $p \times p$ identity matrix. There are no explicit conditions on the truncation sequences $\tau(n)$

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

and $\tau^*(n)$ here, but Assumptions 2–4 may implicitly require some restrictions on the rate at which $\tau^*(n)$ increases with n. Condition (4.3) is a sparseness condition on the weights that requires most weight to be put on a relatively small set of estimators. Conditions A2–A4 are natural extensions of Pakes and Pollard (1989) to the case where the number of moment conditions is allowed to grow. Assumption 3 is just a definition of the estimators $\widehat{\theta}_j$ with error controlled uniformly over j; as usual when the objective function is smooth and the parameter space is compact, this condition is redundant. Assumption 2 ensures that each estimator from the 'strong' set of moment conditions is identified, but the strength of even that identification is allowed to decrease. The rate at which $\epsilon_n(\delta)$ declines is determined by the sequence $\tau^*(n)$. For example, in the IV case, this is determined by the sequence A_j, in particular the rate at which $\|E[A_j(X)]\|$ decreases. By choosing $\tau^*(n)$ to grow very slowly we can compensate for a rapid decline in the moments of the instruments.

The uniform convergence Assumption 4 is easy to verify, although it requires one to extend standard arguments to accommodate the maximum over j. This factor usually costs little extra, see Lemma 1 in the Appendix and the subsequent discussion. Since we must have $\epsilon_n(\delta)$ of larger order than $n^{-1/2}$ in the case of i.i.d. data this puts an upper limit on the rate at which $\tau^*(n)$ can grow, but no lower limit. If $\tau^*(n)$ only increases very slowly, say like $\log n$, the stated rate is easy to achieve. The sequences ϵ_{1n} and ϵ_{2n} are needed because ϵ_n tends to zero and we require the error on the right hand side of A3 and A4 to be of smaller order than ϵ_n. Our conditions require that each member $\widehat{\theta}_j$ of the class indexed by $\mathcal{J}_n^*$ be consistent and further imply that $\max_{j\in\mathcal{J}_n^*}|\widehat{\theta}_j-\theta_0| = o_p(1)$, which is a stronger condition.

For clarification we consider two cases. First, there is no knowledge of the classification into strong and weak moment conditions, that is, we have included in our basket estimators with some potential problems, but we do not know which ones they are. In this situation, we might take equal weighting $W_{nj} = 1/\tau(n)I_p$ and we will need that $\tau^*(n)/\tau(n) \to 1$. That is, we can include an increasing number of inconsistent estimators, but we must have a larger fraction of consistent ones. This rules out the pathological Example 1P, where there are as many inconsistent estimators as consistent ones. We acknowledge that the usual diagonal weighted GMM will manage this situation better. In that case, we minimize the objective function $Q_n(\theta) = \sum_{j=1}^{\tau} G_{nj}^2(\theta)/\tau$ with respect to θ. In this setting, $EQ_n(\theta) = \sum_{j=1}^{\tau} EG_{nj}^2(\theta)/\tau$, and we can allow $EG_{nj}^2(\theta) \equiv 0$ for j with cardinality $\lambda\tau$ for any λ with $\lambda < 1$. In the second case, we suppose there is knowledge of the classification into strong and weak moment conditions, as would be the case in, say Example 3, but also perhaps in Example 1P after some preliminary screening method has been applied. In that case, we can effectively take $\tau(n)$ as large as we like and just ensure that the weighting on the weak moments is small. The optimal weighting method discussed below effectively achieves this.

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

THEOREM 1 (i). *Suppose that Assumptions 1–4 hold. Then* $\widehat{\theta} - \theta_0 = o_p(1)$.

For the purpose of obtaining $\sqrt{n}$-asymptotic normality of $\widehat{\theta}$ in the next subsection, we need to first establish that $\widehat{\theta} - \theta_0 = o_p(n^{-1/4})$ under the following stronger version of Assumption A:

Assumption A*. Let $\theta_0 \in \Theta$ satisfy model (2.1).

(A*1) A1 holds with

$$\sup_{n \geq 1} n^{1/4} \sum_{j=\tau^*(n)+1}^{\tau(n)} ||W_{nj}|| \to 0.$$

(A*2) For some $\delta > 0$ and all $\theta \in \Theta$ such that $\|\theta - \theta_0\| < \delta$, there is a positive sequence $\{\gamma_j, j \in \mathcal{J}_n^*\}$ such that

$$\|G_j(\theta)\| \geq \gamma_j ||\theta - \theta_0||,$$

where $\min_{j \in \mathcal{J}_n^*} \gamma_j \geq \epsilon_n > 0$ with possibly $\epsilon_n \to 0$ as $n \to \infty$.

(A*3) For all $\delta_n = o(1)$ and $n \geq 1$,

$$\max_{j \in \mathcal{J}_n^*} \left(\|G_{nj}(\widehat{\theta}_j)\| - \inf_{\|\theta - \theta_0\| \leq \delta_n} \|G_{nj}(\theta)\| \right) = o_p\left(\epsilon_n n^{-1/4}\right).$$

(A*4) For all $\delta_n = o(1)$ and $n \geq 1$,

$$\max_{j \in \mathcal{J}_n^*} \sup_{\|\theta - \theta_0\| \leq \delta_n} \|G_{nj}(\theta) - G_j(\theta)\| = o_p\left(\epsilon_n n^{-1/4}\right).$$

Assumption A*2 is standard as in Pakes and Pollard (1989), except that we require the lower bounds to decay at a rate under our control. The sequence ϵ_n depends on the sequence of moment conditions but also on the set $\mathcal{J}_n^*$: If this set contains few elements, then it is possible to make ϵ_n decay very slowly. Assumption A*3 again defines the estimators $\widehat{\theta}_j$ and is not needed in the case where the objective function is smooth and the parameter space is compact. Assumption A*4 strengthens the uniform convergence rate in A4. Both assumptions are similar to those often found in the estimation literature with nonsmooth objective functions (see Newey and McFadden, 1994, Sect. 7), with the exception that we are taking a maximum over an increasing number of first order conditions. However, these conditions can be verified in most problems. The uniformity across θ is usually satisfied, indeed we can expect in many cases that $\sup_{\theta \in \Theta} \|G_{nj}(\theta) - G_j(\theta)\| = O_p(1/\sqrt{n})$ for any compact parameter set Θ. In the Appendix we provide a lemma (namely Lemma 1) that can be used to verify the uniform convergence across j and may be useful elsewhere.

THEOREM 1 (ii). *Suppose that Assumptions A*1–A*4 hold. Then* $\widehat{\theta} - \theta_0 = o_p(n^{-1/4})$.

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

4.2. Asymptotic Normality

In this subsection we derive the asymptotic distribution of our estimator $\widehat{\theta}$, under additional conditions. We strengthen the conditions of Pakes and Pollard (1989) and Newey and McFadden (1994) to accommodate our more general set-up, but again we do not require smoothness conditions on the moment conditions $g_j(Z_i, \theta)$. Define

$$\Gamma_j = \frac{\partial}{\partial \theta^\top} G_j(\theta_0) = \frac{\partial}{\partial \theta^\top} E[g_j(Z_i, \theta)] \mid_{\theta=\theta_0} \tag{4.4}$$

for each j. In the IV example, if $D_0(X_i) = \{\partial E[\rho(Z_i, \theta)|X_i]/\partial\theta\}|_{\theta=\theta_0}$ exists with probability one, then we have $\Gamma_j = E[A_j(X_i)D_0(X_i)^\top]$. Further define

$$\Omega_{jl}(\theta) = E[g_j(Z_i, \theta)g_l(Z_i, \theta)^\top], \tag{4.5}$$

$$V_{jl} = \Gamma_j^{-1}\Omega_{jl}\Gamma_l^{-1\top}, \tag{4.6}$$

where $\Omega_{jl} = \Omega_{jl}(\theta_0)$. Under our conditions below, we have for each fixed j as $n \to \infty$,

$$\sqrt{n}(\widehat{\theta}_j - \theta_0) \Longrightarrow N(0, V_{jj}),$$

where V_{jj} is nonsingular and finite for each j. In our setting, we may have $V_{jj} \to \infty$ as $j \to \infty$. This could arise because $\Gamma_j \to 0$ or $\Omega_{jj} \to \infty$ or both. In many cases we have examined the source of asymptotic *over-variability* is from declining Γ_j (while Ω_{jj} stays bounded). This is another way of describing the 'different rate' phenomenon in Lee (2010) and Han and Phillips (2006). We may have V_{jj} bounded as $j \to \infty$, but in that case the $\tau^* p \times \tau^* p$ matrix $V = (V_{jl})$, which represents the joint asymptotic covariance matrix between the leading estimators, may be close to being singular (if this were not the case, then our estimator $\widehat{\theta}$ could obtain a rate improvement, i.e., converge faster than $\sqrt{n}$, an interesting case, but one not treated here)

Assumption B. Let $\theta_0 \in \Theta$ satisfy model (2.1). A1 holds with

$$\sup_{n\geq 1} n^{1/2} \sum_{j=\tau^*(n)+1}^{\tau(n)} ||W_{nj}|| \to 0. \tag{4.7}$$

(B1) $\max_{j\in\mathcal{J}_n^*}(\|G_{nj}(\widehat{\theta}_j)\| - \inf_{\|\theta-\theta_0\|\leq\delta_n}\|G_{nj}(\theta)\|) = o_p(1/\sqrt{n})$ for any $\delta_n = o(n^{-1/4})$.

(B2) The matrix Γ_j exists and is of full column rank for each $j \in \mathcal{J}_n^*$, i.e., $\gamma_n = \min_{j\in\mathcal{J}_n^*}\lambda_{\min}(\Gamma_j) > 0$. Furthermore, there exists a finite constant C such that for any θ such that $\|\theta - \theta_0\| \leq \delta_n$, where $\delta_n = o(n^{-1/4})$, we have

$$\max_{j\in\mathcal{J}_n^*} \left\| G_j(\theta) - \Gamma_j(\theta - \theta_0) \right\| \leq C\|\theta - \theta_0\|^2.$$

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

(B3) (a) $\max_{j\in\mathcal{J}_n^*} \|\sqrt{n}[G_{nj}(\theta_0) - G_j(\theta_0)]\| = O_p(1)$.
(b) For any $\delta_n = o(n^{-1/4})$,

$$\max_{j\in\mathcal{J}_n^*} \sup_{\|\theta-\theta_0\|\le\delta_n} \|[G_{nj}(\theta) - G_j(\theta)] - [G_{nj}(\theta_0) - G_j(\theta_0)]\| = o_p(1/\sqrt{n}).$$

(B4) There exists a deterministic sequence of matrices W_{nj}^0 satisfying: (a) $\sum_{j\in\mathcal{J}_n^*} ||(W_{nj} - W_{nj}^0)\Gamma_j^{-1}|| = o_p(1)$; (b) $\limsup_n \sum_{j\in\mathcal{J}_n^*} \left\| W_{nj}^0 \Gamma_j^{-1} \right\| < \infty$.

(B5) (a) The matrix $\Sigma_n = \sum_{j\in\mathcal{J}_n^*} \sum_{l\in\mathcal{J}_n^*} W_{nj}^0 V_{jl} W_{nl}^{0\top}$ has a finite positive definite limit Σ; (b) The triangular array of random variables $f_n(Z_i) = n^{-1/2} \sum_{j\in\mathcal{J}_n^*} c^\top W_{nj}^0 \Gamma_j^{-1} g_j(Z_i, \theta_0)$ satisfies $nE|f_n(Z_i)|^{2+\kappa} \to 0$ for all $c \in \mathbb{R}^p$ and some $\kappa > 0$.

(B6) θ_0 is in the interior of Θ.

(B7) $\max_{j\in\mathcal{J}_n^*} ||\widehat{\theta}_j - \theta_0|| = o_p(n^{-1/4})$.

We next discuss the assumptions. Assumption B1 again defines the estimators $\widehat{\theta}_j$ and is not needed in the case where the objective function is smooth and the parameter space is compact. Assumption B2 requires essentially two uniformly continuous derivatives for the population moment function at $\theta = \theta_0$, and that the first derivative matrix be of full rank uniformly over $j \le \tau^*(n)$.

Although Assumption B3(a) looks a bit strong, we show that it generically holds in Example 1 under certain conditions. Specifically, in that case

$$\sqrt{n} G_{nj}(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_{ji}\varepsilon_i. \tag{4.8}$$

Suppose that X_{ji} are independent of ε_i and that ε_i is standard normal (the argument holds more generally but is longer without these properties). Then $\sqrt{n} G_{nj}(\theta_0)$ is normally distributed (conditional on $X_{j1}, \ldots, X_{jn}$) with mean zero and variance $\sum_{i=1}^n X_{ji}^2/n$, that is,

$$\begin{aligned} \max_{j\in\mathcal{J}_n^*} \left\| \sqrt{n} G_{nj}(\theta_0)] \right\| &= \max_{j\in\mathcal{J}_n^*} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n X_{ji}\varepsilon_i \right| \\ &\le |Z| \times \max_{j\in\mathcal{J}_n^*} \left| \frac{1}{n} \sum_{i=1}^n X_{ji}^2 \right|, \end{aligned} \tag{4.9}$$

where Z is a standard normal random variable. Provided the second moment of the covariates is uniformly bounded (which we have assumed anyway), the right hand side of (4.9) is bounded in probability and the condition is satisfied. The essential reason for this is that ε_i are not varying with j and so the maximum over $j \in \mathcal{J}_n^*$ does not penalize (4.8) in terms of rate.

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

Regarding Assumption B3(b), consider Example 1. In that case we have

$$\max_{j\in\mathcal{J}_n^*}\sup_{\|\theta-\theta_0\|\le\delta_n}\|[G_{nj}(\theta)-G_j(\theta)]-[G_{nj}(\theta_0)-G_j(\theta_0)]\|$$
$$\le\delta_n\max_{j\in\mathcal{J}_n^*}\left|\frac{1}{n}\sum_{i=1}^{n}\left[X_{ji}y_{2i}-E(X_{ji}y_{2i})\right]\right|,$$

and we may apply Lemma 1 (in the Appendix) to the right hand side to establish the result.

In B4, we require that if the weights are random that they can be well approximated by some nonrandom sequence with certain summability properties (specifically, that they decline sufficiently quickly to counteract the growth of Γ_j^{-1}). This condition entails some restrictions on the rate of growth of τ^*, and these restrictions can be as much as requiring that $\tau^{*3}/n\to 0$ (see Koenker and Machado, 1999). The restrictions are not so stringent in special cases and really arise out of the nonlinearity of the estimating equation rather combined with the large number of parameters. Consider (4.8) in Example 1, and for simplicity, suppose that the regressors are mutually orthogonal and have mean zero and unit variance, then $\Gamma_j=\pi_{2j}\to 0$ as $j\to\infty$. This makes it clear how the weights should decline with j.

Assumption B5 allows us to apply the Liapunov's central limit theorem for triangular arrays to the leading term. This condition is satisfied for a variety of problems, and it implicitly imposes restrictions on how fast $\tau^*(n)$ could grow with sample size n and some restrictions on the weighting sequence. B5(a) is the requirement that the weighted average of the elements of V_{jl} is positive and finite. In the scalar parameter case we have by crude bounding $\Sigma_n\le\lambda_{\max}(V)\sum_{j=1}^{\tau}W_{nj}^2$ for large n and under equal weighting we could allow $\lambda_{\max}(V)$ to grow but no faster than τ, while if $\lambda_{\max}(V)$ were uniformly bounded we could allow more down-weighting such that $\sum_{j=1}^{\tau^*}W_{nj}^2$ remains bounded away from zero. In the special case of Example 1 where the covariates are mutually independent we have $V_{jj}=1/\pi_{2j}^2\to\infty$ and $V_{jl}=0$, so that provided $W_{nj}\le C|\pi_{2j}|$ for some constant C, the conditions will be satisfied. Suppose now that the covariates in Example 1 are not mutually independent, then in this more general case,

$$V_{jl}=K\frac{\operatorname{corr}(X_j,X_l)}{\operatorname{corr}(X_j,y_2)\operatorname{corr}(X_l,y_2)}, \tag{4.10}$$

where the generic constant K does not depend on j,l. We next try to understand how the matrix (V_{jl}) behaves for large n in this case. There are of course many different models that describe the behavior of such large matrices (see e.g., Bickel and Levina, 2008), we here just consider one simple case. Suppose that $R_{ni}=(y_{2i},X_{1i},\ldots,X_{\tau^*(n)i})\in\mathbb{R}^{\tau^*+1}$ are normally distributed with mean zero and correlation matrix Ψ. For example, Ψ could be the correlation matrix of a Gaussian process with y_2 taking the role of the 0th observation. Then $\Psi_{rs}=\psi(|r-s|)$ for some decreasing function ψ, which implies that $V_{jl}\propto\psi(|j-l|)/\psi(j)\psi(l)$.

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

For example, $\psi(u) = u^{-\kappa}$ for some $\kappa > 0$ leads to $V_{jl} \propto j^{\kappa} l^{\kappa}/|j-l|^{\kappa}$ and $V_{jj} \propto j^{2\kappa}$. In this case, for B5(a) it suffices that $W_{nj} = o\left(j^{-\kappa'}\right)$ for some $\kappa' > \kappa$. We give more discussion of this issue below.

Notice that Assumption B5(b) is simply: for some $\kappa > 0$ and for all c,

$$E\left(\left|\sum_{j\in\mathcal{J}_n^*} c^\top W_{nj}^0 \Gamma_j^{-1} g_j(Z_i,\theta_0)\right|^{2+\kappa}\right) = o\left(n^{\kappa/2}\right).$$

For example, suppose we only require that $g_j(Z_i,\theta_0)$ have uniformly bounded fourth moments. In B2 we defined the sequence γ_n. It follows by the Cauchy–Schwarz inequality that

$$nE\left[f_n(Z_i)^4\right] = \frac{1}{n}\sum_{j,k,l,m\in\mathcal{J}_n^*} E[\varphi_{ji}\varphi_{ki}\varphi_{li}\varphi_{mi}] \le \frac{1}{n\gamma_n^4}\left(\sup_n \sum_{j\in\mathcal{J}_n^*} \left\|W_{nj}^0\right\|\right)^4,$$

where $\varphi_{ji} = c^\top W_{nj}^0 \Gamma_j^{-1} g_j(Z_i,\theta_0)$. It suffices in this case that $n\gamma_n^4 \to \infty$. Now suppose that in fact, the scalar $g_j(Z_i,\theta_0)$ are normally distributed with mean zero and variance Γ_j and mutually independent, and that the weights are equal, i.e., $W_{nj}^0 = 1/\tau^*(n) I_p$ for each j. Then

$$nE\left[f_n(Z_i)^4\right] = \frac{1}{n\tau^4}\left(\sum_{j\in\mathcal{J}_n^*} 3\Gamma_j^{-2} + 3\sum_{j\neq k\in\mathcal{J}_n^*} \Gamma_j^{-1}\Gamma_k^{-1}\right) \le \frac{3}{n\tau^2\gamma_n^2},$$

which goes to zero provided $n\tau^2\gamma_n^2 \to \infty$. These conditions can be weakened considerably in special cases.

Notice that we can replace Assumptions B3(a) and B5 by the condition that $\left\{G_{nj}(\theta_0) - G_j(\theta_0) : j \in \mathcal{J}_n^*\right\}$ is a Donsker class, i.e., it satisfies the uniform central limit theorem. This kind of assumption has been used in Portnoy (1984) for example.

The condition B7 that $\max_{j\in\mathcal{J}_n^*} ||\widehat{\theta}_j - \theta_0|| = o_p(n^{-1/4})$ follows from our Theorem 1(ii). It is not necessary in for example linear cases. It may be possible to prove our general result below without a sup-norm convergence result like this, although we have not been able to find a proof based on other convergence criterions like L_p. The usual proofs in other semiparametric estimation problems typically make use of similar results about the convergence of nuisance parameters (see e.g., Newey and McFadden, 1994).

THEOREM 2. *Suppose that Assumptions 1 and B1–B7 hold. Then* $\sqrt{n}(\widehat{\theta} - \theta_0) \Longrightarrow N(0, \Sigma)$.

The asymptotic variance matrix Σ depends on the weighting scheme and on the class of estimators considered and, of course, on the underlying distribution of the data. We discuss the nature of the asymptotic variance more in the next section.

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

To construct consistent estimates of Σ, we compute

$$\widehat{\Sigma} = \sum_{j \in \mathcal{J}_n^*} \sum_{l \in \mathcal{J}_n^*} W_{nj} \widehat{V}_{jl} W_{nl}^\top, \tag{4.11}$$

$$\widehat{V}_{jl} = \widehat{\Gamma}_j^{-1} \frac{1}{n} \sum_{i=1}^{n} g_j(Z_i, \widehat{\theta}) g_l(Z_i, \widehat{\theta})^\top \widehat{\Gamma}_l^{-1\top}. \tag{4.12}$$

Note that there we do not impose any ordering on the covariance matrices V_{jl}, and we have no need of a bandwidth parameter here since the cardinality of $\mathcal{J}_n^*$ is small compared with n. The estimation of Γ_j is straightforward when G_{nj} are differentiable. In this case, for each j

$$\widehat{\Gamma}_j = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial g_j(Z_i, \widehat{\theta})}{\partial \theta} \to^p \Gamma_j \tag{4.13}$$

under some mild regularity conditions. When G_{nj} are not differentiable, as for example in the Least Absolute Deviation (LAD) case, this method is not feasible. In some cases, one might be able to estimate directly the quantity Γ_j. For example, in the LAD case (with errors independent of covariates), Γ_j is proportional to the density of the errors evaluated at their median. This quantity can be estimated by a variety of nonparametric methods. A general strategy for estimating Γ_j is to use 'numerical derivatives', that is, let

$$\widehat{\Gamma}_{j;lk} = \frac{1}{n} \sum_{i=1}^{n} \frac{g_{jl}(Z_i, \widehat{\theta} + \delta e_k) - g_{jl}(Z_i, \widehat{\theta})}{\delta}, \tag{4.14}$$

where e_k is a vector of zeros with one in the kth position, while δ is a small constant. If we let $\delta(n)$ go to zero at a certain rate as sample size increases, we can show that $\widehat{\Gamma}_{j;lk} \to^p \Gamma_{j;lk}$ (see, for example, Pakes and Pollard, 1989). The actual derivative (4.13) makes δ go to zero before n, but our modified estimator (4.14) allows δ to go to zero with n and indeed slower than n. Under stronger conditions, including $\max_{j \in \mathcal{J}_n^*} \| \widehat{\Gamma}_j - \Gamma_j \| \to^p 0$ and $\max_{j,l \in \mathcal{J}_n^*} \| \widehat{V}_{jl} - V_{jl} \| \to^p 0$, we can obtain $\widehat{\Sigma} \to^p \Sigma$. Provided that $\tau(n) \to \infty$ slowly as $n \to \infty$, the additional conditions are not particularly onerous. The estimation of optimal weights also requires estimation of V and we discuss this further below.

5. OPTIMAL WEIGHTS

We now discuss the question of optimal weights in the sense of minimizing asymptotic variance within our class of estimators. We also comment on the more general question about optimality given the information expressed through the moment conditions (2.1), which is a more difficult question. For simplicity, we restrict attention to a simple leading case where $\tau = \tau^*$ and $\mathcal{J}_n = \mathcal{J}_n^*$. We first consider the case where τ is fixed and then turn to the case where it is increasing.

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

5.1. Case 1: Fixed τ

We can consider the optimal weights to be those that minimize the asymptotic variance matrix of Theorem 2 in the special case where τ is fixed, i.e., minimize

$$\Sigma^{\tau} = \sum_{j=1}^{\tau}\sum_{l=1}^{\tau} W_{nj} V_{jl} W_{nl}^{\top}$$

with respect to $p \times p$ matrices $W_{n1}, \ldots, W_{n\tau}$ subject to the restriction that $\sum_{j=1}^{\tau} W_{nj} = I_p$. The solution to this can be found explicitly. Following Rothenberg (1973), we take $\Xi = V^{-1}$ in (2.6), where $V = \left[V_{j,l}\right]$, in which case

$$W_{0j}^{\text{opt}} = [(i_\tau \otimes I_p)^{\top} V^{-1} (i_\tau \otimes I_p)]^{-1} [(i_\tau \otimes I_p)^{\top} V^{-1}]_j. \tag{5.1}$$

The corresponding estimator $\widehat{\theta}^{\text{opt}} = \sum_{j \in \mathcal{J}_n^*} W_{0j}^{\text{opt}} \widehat{\theta}_j$ has asymptotic (as $n \to \infty$ and τ fixed) variance

$$\Sigma_{\text{opt}}^{\tau} = \sum_{j=1}^{\tau}\sum_{l=1}^{\tau} W_{0j}^{\text{opt}} V_{jl} W_l^{\text{opt}\top} = \left[(i_\tau \otimes I_p)^{\top} V^{-1} (i_\tau \otimes I_p)\right]^{-1},$$

which is the smallest amongst our class of estimators. We require here that the matrix V be nonsingular, which seems like a reasonable requirement for fixed τ. In the scalar case, $W_{0j}^{\text{opt}} = (i_k^{\top} V^{-1} i_k)^{-1} [i_k^{\top} V^{-1}]_j$ and $\Sigma_{\text{opt}}^{\tau} = (i_k^{\top} V^{-1} i_k)^{-1}$, and W_{0j}^{opt} are known as the global minimum variance portfolio weights (see e.g., Campbell, Lo, and Mackinlay, 1997).

Example 1 (Classical two stage least squares in simultaneous equations)
Recall the optimal GMM estimator in this model (i.e., under homoskedasticity, etc.) is simply the two stage least squares estimator

$$\widetilde{\theta} = (Y_2^{\top} P_X Y_2)^{-1} Y_2^{\top} P_X Y_1,$$

where $P_X = X(X^{\top}X)^{-1}X^{\top}$, $Y_1 = (y_{11}, \ldots, y_{1n})^{\top}$, $Y_2 = (y_{21}, \ldots, y_{2n})^{\top}$, $X = (X_1^{\top}, \ldots, X_n^{\top})$, $X_i = (X_{1i}, \ldots, X_{ki})^{\top}$. Within our class of estimators $\mathcal{E}$, the optimal estimator is

$$\widehat{\theta} = \sum_{j=1}^{k} W_{nj}^{\text{opt}} \widehat{\theta}_j = (i_k^{\top} V^{-1} i_k)^{-1} i_k^{\top} V^{-1} \begin{bmatrix} \widehat{\theta}_1 \\ \vdots \\ \widehat{\theta}_k \end{bmatrix},$$

where $\widehat{\theta}_j = (Y_2^{\top} P_j Y_2)^{-1} Y_2^{\top} P_j Y_1$ for $j = 1, \ldots, k$, where $P_j = X_j (X_j^{\top} X_j)^{-1} X_j^{\top}$ and V is the $k \times k$ covariance matrix defined for general estimators in (4.6). In the homoskedastic error case (and treating the regressors as fixed for simplicity),

$$V_{jl} = \sigma_{\varepsilon}^2 \left(\pi_2^{\top} X^{\top} X_j X_j^{\top} X \pi_2 / n\right)^{-1} \times \left(\pi_2^{\top} X^{\top} X_j X_j^{\top} X_l X_l^{\top} X \pi_2 / n\right) \left(\pi_2^{\intercal} X^{\top} X_l X_l^{\top} X \pi_2 / n\right)^{-1} = \sigma_{\varepsilon}^2 \frac{M_{jl}}{M_{\circ j} M_{\circ l}},$$

where $M_{jl} = X_j^{\top} X_l / n$ and $M_{\circ j} = \pi_2^{\top} X^{\top} X_j / n$.

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

Suppose that the instruments are mutually orthogonal, then it is easy to see that $\widehat{\theta}$ is identically equal to $\widetilde{\theta}$.[4] This gives yet another interpretation to 2SLS as being, in this case, the optimal combination of exactly identified instrumental variables estimators.[5] In general, the two estimators are asymptotically equivalent, but not identical.

Example 2 (Semiparametric instrumental variable)
Suppose that the moment conditions are of the form

$$E\left[A_j(X_i)\rho(Z_i,\theta_0)\right]=0, \quad j=1,\ldots,\tau, \tag{5.2}$$

where τ is fixed, and $A_j \in \mathbb{R}^p$. This is a standard unconditional moments estimation problem, and the optimal estimator (smallest variance) can be arrived at by several routes: (1) through the optimal combination of the sample moment conditions (GMM); (2) or through the optimal combination of the instruments into a single estimating equation. The asymptotic variance of the efficient estimator is given by

$$\begin{aligned}\Sigma^{\tau}_{\text{oiv}} &= \left(E\left[A^{\tau}(X)D_0(X)^{\top}\right]^{\top}\left[E\left(\sigma_0^2(X)A^{\tau}(X)A^{\tau}(X)^{\top}\right)\right]^{-1}\right.\\ &\qquad \left.\times\, E\left[A^{\tau}(X)D_0(X)^{\top}\right]\right)^{-1}\\ &\equiv \left(\Gamma^{\tau\top}\Psi_{\tau}^{-1}\Gamma^{\tau}\right)^{-1},\end{aligned} \tag{5.3}$$

where $A^{\tau}=(A_1^{\top},\ldots,A_{\tau}^{\top})^{\top}\in\mathbb{R}^{\tau p}$.

We can show the equivalence between the optimal minimum distance estimator, as defined above, and the optimal instrumental variable estimator in the following proposition.

PROPOSITION 1. *For each fixed* τ, $\widehat{\theta}^{\text{opt}}$ *is asymptotically efficient for (5.2) with* $\Sigma^{\tau}_{\text{opt}}=\Sigma^{\tau}_{\text{oiv}}$. *Moreover the optimal weighting is*

$$\begin{aligned}W_{0j}^{\text{opt}} &= -\left(\sum_{j=1}^{\tau}\alpha_j\Gamma_j^{\top}\right)^{-1}\alpha_j\Gamma_j^{\top} \text{ for } j=1,\ldots,\tau, \text{ with } (\alpha_1,\ldots,\alpha_{\tau})\\ &= \Gamma^{\tau\top}\Psi_{\tau}^{-1}.\end{aligned} \tag{5.4}$$

This says that the optimal IV estimator has the same asymptotic expansion to order $n^{-1/2}$ as the member of our class $\mathcal{E}$ that has weights given in (5.4). That is, our best estimator achieves the same efficiency as optimal GMM in the case with a finite number of unconditional moments.

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

Chamberlain (1987) considered a sequence of models with unconditional moment restrictions (where the distribution of X is multinomial with τ support points). He showed that for each τ

$$\Sigma_{\text{oiv}}^{\tau} = \left(E[\sigma_0^{-2}(X)D_0(X)D_0(X)^\top]\right)^{-1}.$$

5.2. Case 2: Increasing τ

We now consider the case where τ increases with sample size. A key issue here is around the behavior of the $\tau p \times \tau p$ covariance matrix V as its dimensions increase. If V were diagonal with bounded elements, then equal weighted-like averaging could improve the rate of convergence. The latter setting applies to panel data where there is an additional source of variation, but not when this additional variation is not present. We focus on the case where this rate improvement is not possible, either because the largest diagonal element of V is increasing with the dimensions and/or the covariance matrix V is becoming singular.

We show that our efficiency properties carry over to the increasing τ case. In our proofs of Theorem 2 we obtain the stochastic expansion (uniformly in $\tau \leq \tau(n)$)

$$\sqrt{n}(\widehat{\theta}(W(\tau)) - \theta_0) = L(W(\tau)) + o_p(1), \tag{5.5}$$

where the random variable L is a linear term with mean zero and finite variance, i.e., $L = n^{-1/2}\sum_{i=1}^{n}\xi_i$ with $E[\xi_i] = 0$ and $E[\xi_i\xi_i^\top] < \infty$; we use the notation $W(\tau)$ to denote the weighting sequence $\{W_{nj}\}$ for $\tau(n)$ and we include $W(\tau)$ as an argument of $\widehat{\theta}$ and L to emphasize their dependency on these matrices. Provided V is nonsingular for each τ, we can define the optimal weights (5.1) for each τ and the expansion (5.5) holds for the optimal weighting sequence. We have $\text{var}[L(W^{\text{opt}}(\tau))] \leq \text{var}[L(W(\tau))]$ for any other weighting sequence for all τ and in this sense the estimator $\widehat{\theta}(W^{\text{opt}}(\tau_n))$ is efficient.

There are two questions we address. First, whether

$$\lim_{\tau\to\infty}\left[(i_\tau \otimes I_p)^\top V^{-1}(i_\tau \otimes I_p)\right]^{-1} = \Sigma_{\text{opt}}^{\infty}, \tag{5.6}$$

where $\Sigma_{\text{opt}}^{\infty}$ is finite and positive definite. Second, whether there exists an estimator of θ with a smaller asymptotic variance.

If B5 is satisfied for the optimal weighting sequence, then (5.6) is satisfied. We next discuss how to verify Assumption B5 for the optimal weighting sequence. This condition is clearly not always satisfied: it depends on the underlying estimator sequences and their relationship. It is a common assumption that V (or matrices equivalent to V) is finite and invertible for finite τ, the question is whether the implications of this property are maintained as $\tau \to \infty$.

In Example 1, when the errors are independent of the instruments, we have (4.10), and there are a variety of schemes for the covariance matrix of $(y_2, X_1, \ldots, X_{\tau(n)})$ that would support the assumption (i.e., (5.6) holds) for the optimal weighting sequence.

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

We look at the issue in the general scalar parameter case, in which case $\Sigma_{\text{opt}}^{\tau} = (i^\top V^{-1} i)^{-1}$. By crude bounding we obtain

$$i^\top V^{-1} i = \left(i^\top i\right) \frac{i^\top V^{-1} i}{i^\top i} \in \left[\tau \lambda_{\min}\left(V^{-1}\right), \tau \lambda_{\max}\left(V^{-1}\right)\right]$$
$$= [\tau/\lambda_{\max}(V), \tau/\lambda_{\min}(V)],$$

which implies that $(i^\top V^{-1} i)^{-1} \leq \lambda_{\max}(V)/\tau$. Provided $\limsup_{\tau\to\infty} \lambda_{\max}(V)/\tau < \infty$, we have $\limsup_{\tau\to\infty} (i^\top V^{-1} i)^{-1} < \infty$ and we can expect (5.6) to hold. However, this condition is much stronger than necessary, as can be seen from the diagonal case where a variety of rates can be assumed on the elements of $V = \text{diag}\{v_1, \ldots, v_\tau\}$ to ensure that $i^\top V^{-1} i = \sum_{j=1}^{\tau} v_j^{-1}$ converges to a finite positive number. For example, suppose that $v_j = cj^\alpha$ for $\alpha > 1$, then $\lambda_{\max}(V)/\tau = c\tau^{\alpha-1} \to \infty$ but $\sum_{j=1}^{\tau} v_j^{-1}$ converges to a finite positive number. This can also hold in nondiagonal cases. For example, consider the equicorrelated case where

$$V = \Delta^{1/2}(I + \rho_\tau i i^\top)\Delta^{1/2}, \tag{5.7}$$

where $\Delta = \text{diag}\{\sigma_1^2, \ldots, \sigma_\tau^2\}$ and $\rho_\tau > 0$. This corresponds to a situation where each estimator is correlated with each other by the same positive amount ρ_τ. In this case,

$$V^{-1} = \Delta^{-1/2}(I - c_\tau i i^\top)\Delta^{-1/2},$$

where $c_\tau = \rho_\tau/(\rho_\tau \tau - 1)$. Then

$$i^\top V^{-1} i = \sum_{j=1}^{\tau} \sigma_j^{-2} - \frac{\rho_\tau}{\rho_\tau \tau - 1}\left(\sum_{j=1}^{\tau} \sigma_j^{-1}\right)^2.$$

We can suppose that $\sigma_j \to \infty$ as $j \to \infty$, so that if $\sum_{j=1}^{\infty} \sigma_j^{-1} < \infty$, then $\sum_{j=1}^{\infty} \sigma_j^{-2} < \infty$. Provided $\rho_\tau >> 1/\tau$, $i^\top V^{-1} i \to \sum_{j=1}^{\infty} \sigma_j^{-2} \in (0, \infty)$. This case corresponds to the single factor model used in the portfolio choice problem with many assets; see e.g., Fan, Li, and Yu (2012) for more general structures that also broadly fit into this framework.

Although we can provide a number of situations where $\Sigma_{\text{opt}}^{\infty}$ is finite and positive definite, without further structure we are unable to address the second question, namely whether the optimal estimator within our class is efficient amongst all regular estimators, i.e., we do not generally have a semiparametric efficiency bound to compare with. So we next turn to two special cases where a semiparametric efficiency standard is known against which we may compare our procedure.

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

Example 3 (Semiparametric instrumental variables)
Let Σ_{oiv} be the asymptotic variance of the optimal instrumental variable (oiv) estimator, which is the semiparametric efficiency bound (defined above). Then, for certain choices of $A_1, \ldots, A_\tau$, we have

$$\lim_{\tau \to \infty} \Sigma_{\text{oiv}}^{\tau} = \Sigma_{\text{oiv}} \tag{5.8}$$

and the GMM estimator based on the moment conditions (5.2) achieves the semiparametric efficiency bound. For example, Chamberlain's (1987) estimator sequence achieves this efficiency bound. Since our optimal estimator (combining the estimators from each moment condition A_j) is at least as efficient as the optimal GMM estimator for each τ, in such cases where (5.8) holds our estimator will also achieve the efficiency bound (the information contained in $A_1, \ldots, A_\tau$ is equivalently expressed through the optimal GMM procedure or through our minimum distance method).

Example 4 (Maximum likelihood)
For the estimators $\widehat{\theta}_j = F_n^{-1}(j/\tau) - F^{-1}(j/\tau)$, we have

$$V_{j,l} = \frac{\min\{j/\tau, l/\tau\} - (j/\tau)(l/\tau)}{f(F^{-1}(j/\tau)) f(F^{-1}(l/\tau))},$$

where f is the density function. In the standard uniform case ($F(x) = x$) it is known that the eigenvalues of the $\tau \times \tau$ matrix V lie between a/τ and $c\tau$ for positive finite constants a, c, (Shorack and Wellner, 2009, p. 222). In this case, the optimal weighting is going to put positive weight only on the first estimator (in this ordering) and will result in a superconsistency. If f is standard Gaussian, then $f(F^{-1}(1/\tau)) \propto 1/\tau$ and a more regular averaging of the estimators is expected. In the Gaussian case, the MLE is the sample mean and this can be expressed as the average of all quantiles. Therefore, if we take $\tau = n$ and let $W_{nj} = 1/n$, then $\widehat{\theta}$ is exactly the sample mean. Hence, full efficiency can be achieved by our method in this special case.

6. ESTIMATION OF OPTIMAL WEIGHTS

In this section, we consider estimation of the optimal weights and construction of a feasible asymptotically optimal estimator (in the sense of minimizing asymptotic variance within our class of procedures). In particular, we shall estimate the optimal weights defined in (5.1). Recall that V is the $\tau p \times \tau p$ asymptotic (as $n \to \infty$ with τ fixed) covariance matrix of the vector of estimators $\widehat{\theta}_j$, $j \in \mathcal{J}_n^*$. We estimate the optimal weights as follows

$$\widehat{W}_{0j}^{\text{opt}} = \left(\sum_{l=1}^{\tau(n)} \widehat{B}_l \right)^{-1} \widehat{B}_j = [(i_\tau \otimes I_p)^\top \widehat{V}^{-1} (i_\tau \otimes I_p)]^{-1} [(i_\tau \otimes I_p)^\top \widehat{V}^{-1}]_j, \tag{6.1}$$

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

where $(\widehat{B}_1, \ldots, \widehat{B}_\tau) = (i_\tau \otimes I_p)^\top \widehat{V}^{-1}$, and $\widehat{V}$ has (j,l) submatrix calculated using formulas (4.12) and (4.13) for $j, l = 2, \ldots, \tau$ based on a preliminary $\sqrt{n}$-consistent estimator of θ, denoted $\widehat{\theta}$. The resulting feasible estimator is then defined as

$$\widehat{\theta}^\dagger = \sum_{j \in \mathcal{J}_n} \widehat{W}_{0j}^{\text{opt}} \widehat{\theta}_j. \tag{6.2}$$

We now provide a consistency result for this feasible estimator defined in (6.2). The strategy is to verify condition B4(a) of Theorem 2 for the estimated weights; we are implicitly assuming that B4(b) and B5 hold, so that the infeasible optimal weights are well defined. If the other conditions of Theorem 2 are satisfied, which are about the moment conditions, then the estimator based on (6.1) is consistent with Theorem 2. We shall restrict attention to the case where g_j are all differentiable. Define for each θ in a neighborhood of θ_0:

$$\Gamma_{nj}(\theta) = \frac{1}{n}\sum_{i=1}^n \frac{\partial g_j}{\partial \theta}(Z_i, \theta); \quad \Omega_{njl}(\theta) = \frac{1}{n}\sum_{i=1}^n g_j(Z_i, \theta) g_l(Z_i, \theta)^\top,$$

where $\Gamma_j(\theta) = E\Gamma_{nj}(\theta)$, $\Omega_{jl}(\theta) = E\Omega_{njl}(\theta)$, $V_{jl} = \Gamma_j^{-1}(\theta_0)\Omega_{jl}(\theta_0)\Gamma_l^{-1}(\theta_0)^\top$ were defined above. We shall assume the following high level conditions.

Assumption D.

(D1) The $\tau p \times \tau p$ matrix V satisfies $\lambda_{\min}(V) = O(\tau^{-\rho})$ for some $\rho \geq 0$.

(D2) For some $\rho_1 \geq 0$, $\min_{j \in \mathcal{J}_n} \lambda_{\min}(\Gamma_j) = o(\tau^{-\rho_1})$.

(D3) For some sequence $\delta_n \to 0$ and some $\eta > 0$:

$$\max_{j \in \mathcal{J}_n} \sup_{\|\theta - \theta_0\| \leq \delta_n \sqrt{n}} \|\Gamma_{nj}(\theta) - \Gamma_j(\theta)\| = o_p(n^{-\eta});$$

$$\max_{j,l \in \mathcal{J}_n} \sup_{\|\theta - \theta_0\| \leq \delta_n \sqrt{n}} \|\Omega_{njl}(\theta) - \Omega_{jl}(\theta)\| = o_p(n^{-\eta}).$$

As we discussed above Assumption D1 can be verified in a number of different cases. It allows the $\tau p \times \tau p$ matrix V to become asymptotically singular, but at a rate that is controlled by ρ. Similar comments apply to D2 in the sense that it is easy to find a variety of examples consistent with this assumption for some small ρ_1.

Note that we do not have a lower bound on the rate that $\tau(n) \to \infty$—this is because we are combining estimators that are already root-n consistent and so we do not need to average a lot of them so as to achieve rate improvement. The uniform laws of large numbers in D3 can be verified under some primitive conditions, along the lines of the discussion around Theorem 2. These conditions are extensions of standard conditions for standard error estimation to the case where $\tau(n) \to \infty$. We will need the estimation error in $\widehat{V}$ to be small relative to the dimensions of τ.

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

THEOREM 3. *Suppose that* $\Sigma_{\text{opt}}^{\infty}$ *exists and is positive definite. Suppose that Assumptions D holds and that* $\tau(n) = n^c$ *for some c such that* $(2+\rho+2\rho_1)c < \eta$. *Then conditions B4(a) of Theorem 2 hold for the weighting sequence (6.1). Therefore, provided the other conditions of Theorem 2 are satisfied, then* $\sqrt{n}(\widehat{\theta}^{\dagger} - \theta_0)$ *is asymptotically normal with mean zero and variance matrix* $\Sigma_{\text{opt}}^{\infty}$.

For further issues surrounding estimating the optimal weights for similar estimation problems, we refer the reader to Newey (1990) and Koenker and Machado (1999). We note that the problem of estimating large covariance matrices is very well studied as it arises naturally in for example portfolio choice problems (see e.g., Ledoit and Wolf, 2004). Cai, Liu, and Zhou (forthcoming) consider estimation of functionals of the inverse of a large covariance matrix and give optimal rates of convergence for this endeavor for certain classes of sparsity structure. We have not imposed sparsity in our estimation strategy for V, which is why we may allow only a relatively slow rate of expansion for τ.

7. MONTE CARLO

As to demonstrate how the proposed alternative procedure of combining estimators works in practice, we consider two data generating processes (DGPs). The first one (DGP 1) corresponds to the framework of Example 3, and it is adapted from Newey (1990) who consider an endogenous dummy variable model with the following specification:

$$
\begin{aligned}
&Y_i = \beta_{10} + \beta_{20} s_i + \varepsilon_i;\\
\text{DGP1: } &s_i = 1(\alpha_{10} + \alpha_{20} X_i + \eta_i > 0),\\
&X_i \sim N(0,1); \qquad \alpha_{10} = \alpha_{20} = \beta_{10} = \beta_{20} = 1,
\end{aligned}
$$

where the errors ε_i and η_i are generated as

$$
\begin{bmatrix} \varepsilon_i \\ \eta_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \varphi \\ \varphi & 1 \end{bmatrix} \right), \tag{7.1}
$$

in which $\varphi \in \{0.2, 0.5, 0.8\}$ indicate weak, medium, and strong endogeneity respectively. The optimal instrument for s is $\pi(x) = \Pr[s = 1 | X = x]$, which makes $D(x) = (1, \pi(x))^{\top}$.

Tables 1 and 2 report results for two estimators of β_{20}. The first estimator corresponds to Newey's (1990) and the second is ours. To obtain both estimators, we follow Newey (1990) and use the polynomials $A_j(x) = x^{j-1}$ and $A_j(x) = [x/(1+|x|)]^{j-1}$ as basis. Newey estimator becomes

$$
\widetilde{\beta} = \begin{pmatrix} \widetilde{\beta}_{10} \\ \widetilde{\beta}_{20} \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n s_i \\ \sum_{i=1}^n \widehat{\pi}(X_i) & \sum_{i=1}^n \widehat{\pi}(X_i) s_i \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n \widehat{\pi}(X_i) Y_i \end{pmatrix},
$$

$$
\widehat{\pi}(x) = \sum_{j=1}^{\tau} \widehat{\gamma}_j A_j(x).
$$

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

TABLE 1. DGP 1: $A_j(x) = x^{j-1}$, $j = 2, \ldots, 5$

	(A)			(B)			(A)			(B)		
τ	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
	$\varphi = 0.2, n = 100$						$\varphi = 0.2, n = 200$					
2	0.004	0.489	0.488	0.004	0.489	0.488	−0.014	0.319	0.319	−0.014	0.319	0.319
3	0.008	0.451	0.451	0.002	0.487	0.487	−0.009	0.300	0.300	−0.022	0.314	0.314
4	0.007	0.442	0.442	0.019	0.523	0.524	−0.009	0.299	0.299	0.012	0.361	0.361
5	0.008	0.438	0.437	0.012	0.669	0.669	−0.003	0.297	0.297	−0.020	0.449	0.449
	$\varphi = 0.5, n = 100$						$\varphi = 0.5, n = 200$					
2	0.014	0.470	0.470	0.014	0.470	0.470	0.000	0.327	0.327	0.000	0.327	0.327
3	0.027	0.442	0.441	0.011	0.465	0.465	−0.001	0.309	0.309	0.001	0.324	0.324
4	0.042	0.431	0.431	0.033	0.515	0.514	0.003	0.302	0.302	0.000	0.371	0.371
5	0.057	0.422	0.423	0.010	0.643	0.644	0.006	0.296	0.296	0.009	0.471	0.471
	$\varphi = 0.8, n = 100$						$\varphi = 0.8, n = 200$					
2	−0.014	0.503	0.504	−0.014	0.503	0.504	0.001	0.354	0.355	0.001	0.354	0.355
3	0.006	0.473	0.473	0.028	0.493	0.493	0.007	0.337	0.337	0.019	0.317	0.317
4	0.025	0.460	0.460	0.026	0.532	0.533	0.013	0.331	0.331	0.026	0.374	0.374
5	0.047	0.443	0.444	0.027	0.669	0.670	0.028	0.326	0.326	0.040	0.459	0.459

Note: Monte Carlo bias (Bias), standard deviation (Std. Dev.), and Root Mean Square Error (RMSE) based on 5,000 replications. (A) = $\widetilde{\beta}_{20}$ and (B) = $\widehat{\beta}_{20}$.

TABLE 2. DGP 1: $A_j(x) = [x/(1+|x|)]^{j-1}$, $j = 2, \ldots, 5$

	(A)			(B)			(A)			(B)		
τ	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
	$\varphi = 0.2, n = 100$						$\varphi = 0.2, n = 200$					
2	−0.014	0.493	0.494	−0.014	0.493	0.494	−0.017	0.328	0.328	−0.017	0.328	0.328
3	0.005	0.458	0.458	−0.005	0.383	0.383	−0.015	0.304	0.304	−0.012	0.252	0.252
4	0.002	0.448	0.448	0.028	0.384	0.384	−0.002	0.297	0.297	−0.001	0.244	0.244
5	0.022	0.440	0.440	0.023	0.405	0.405	−0.001	0.298	0.298	0.001	0.246	0.246
	$\varphi = 0.5, n = 100$						$\varphi = 0.5, n = 200$					
2	0.029	0.485	0.485	0.029	0.485	0.485	−0.001	0.333	0.333	−0.001	0.333	0.333
3	0.034	0.451	0.451	0.052	0.375	0.376	−0.003	0.309	0.309	−0.001	0.258	0.258
4	0.047	0.427	0.427	0.040	0.365	0.365	0.012	0.299	0.299	−0.002	0.252	0.252
5	0.074	0.417	0.419	0.026	0.394	0.394	0.013	0.296	0.296	−0.009	0.259	0.259
	$\varphi = 0.8, n = 100$						$\varphi = 0.8, n = 200$					
2	−0.009	0.511	0.513	−0.009	0.511	0.513	−0.001	0.358	0.359	−0.001	0.358	0.359
3	0.014	0.474	0.474	0.023	0.384	0.384	0.004	0.337	0.338	0.012	0.263	0.263
4	0.048	0.454	0.454	0.023	0.386	0.387	0.026	0.330	0.329	0.014	0.258	0.258
5	0.058	0.442	0.443	0.017	0.410	0.410	0.037	0.325	0.325	0.008	0.271	0.271

Note: Monte Carlo bias (Bias), standard deviation (Std. Dev.), and Root Mean Square Error (RMSE) based on 5,000 replications. (A) = $\widetilde{\beta}_{20}$ and (B) = $\widehat{\beta}_{20}$.

for series-based estimated weights $\widehat{\gamma}_j$. Using the same bases, our estimator becomes

$$\widehat{\beta} = \sum_{j=2}^{\tau} \widehat{W}_{0j}^{\text{opt}} \widehat{\beta}_j, \text{ where} \tag{7.2}$$

$$\widehat{\beta}_j = \begin{pmatrix} \widehat{\beta}_{10;\,j} \\ \widehat{\beta}_{20;\,j} \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n s_i \\ \sum_{i=1}^n A_j(X_i) & \sum_{i=1}^n A_j(X_i) s_i \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n A_j(X_i) Y_i \end{pmatrix} \tag{7.3}$$

and $\widehat{W}_{0j}^{\text{opt}}$ calculated as in (6.1).[6] We consider two samples sizes: $n = 100$, 200, and 5, 000 replications. Simulated bias, standard deviation (Std. Dev.), and Root Mean Squared Error (RMSE) are reported for each estimator in Tables 1 and 2.

We observe that for each sample size and endogeneity parameter value (φ) under consideration, the RMSE associated with the proposed estimator of β_{20} is roughly comparable to that of Newey (1990) for low values of τ. While biases are small for both sets of estimates, their variance behave quite differently with relation to τ. In particular, the precision of $\widetilde{\beta}_{20}$ increases with τ, while that of $\widehat{\beta}_{20}$ actually decreases. This might be caused by the estimation error in $\widehat{W}_{0j}^{\text{opt}}$. Table 2 shows that the proposed estimator outperforms Newey's (1990) for $\tau = 2$ and 3 when $n = 100$ and 200 respectively.

The next scenario is adapted from Example 1 in the main text where $k > n$ and the Monte Carlo design in Okui (2011). In this data generating process (DGP 2), we consider a high-dimensional problem involving a two-equation system with the following specification:

$$\begin{aligned} &Y_i = \beta_{10} + \beta_{20} s_i + \varepsilon_i; \\ \text{DGP2: } &s_i = \sum_{l=1}^{k} \alpha_{l;0} X_{li} + \eta_i, \\ &X_i \sim N(0, I_k); \qquad \beta_{10} = \beta_{20} = 1, \end{aligned}$$

where $X_i = (X_{1i}, \ldots X_{ki})^\top$ and $(\varepsilon_i, \eta_i)^\top$ are generated as in (7.1). We try three different specifications of $\alpha_0 = (\alpha_{1;0}, \ldots, \alpha_{k;0})^\top$ in Okui (2011), namely Model (a): $\alpha_{l;0} = \sqrt{R_f^2/[k(1-R_f^2)]}$, $\forall l$; Model (b): $\alpha_{1;0} = c(k)$, $\alpha_{l;0} = c(k)/\sqrt{k-1}$, $\forall l = 2, \ldots, k$; and Model (c): $\alpha_{l;0} = c(k)[1 - l/(k+1)]$[4], where the constant $c(k)$ is chosen to satisfy $\alpha_0^\top \alpha_0 = R_f^2/(1-R_f^2)$ with R_f^2 representing the theoretical R^2 of the first stage regression. In Model (a), all instruments are equally important but they are also weak, while in Model (b), the first instrument is strong but others are weak. Finally, in Model (c), the strength of the instruments decreases moderately in $l = 1, \ldots, K$. We set $k = 30$ and assess the performance of the 'optimal' estimator here in another set of 5,000 replications, i.e., weights determined by (6.1) and $\tau = k$, with undersized samples of $n = 15$ and 25. Tables 3 and 4 show the results for $R_f^2 = 0.01$ and $R_f^2 = 0.1$ respectively. These tables report the Monte Carlo bias (Bias), standard deviations (Std. Dev), and Root Mean Squared Error (RMSE) of

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

TABLE 3. DGP 2: High-dimensional inference with $R_f^2 = 0.01$

	$\varphi = 0.2$			$\varphi = 0.5$			$\varphi = 0.8$		
n	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
	Model (a)								
15	0.2011	0.3077	0.3676	0.4890	0.2783	0.5626	0.7898	0.1918	0.8128
	0.2014	0.2772	0.3426	0.4920	0.2495	0.5517	0.7927	0.1714	0.8110
25	0.2035	0.2441	0.3178	0.4858	0.2240	0.5349	0.7871	0.1548	0.8022
	0.2030	0.2062	0.2893	0.4925	0.1829	0.5253	0.7920	0.1259	0.8020
50	0.2011	0.1980	0.2822	0.4851	0.1768	0.5163	0.7817	0.1253	0.7916
	0.2010	0.1838	0.2724	0.4889	0.1634	0.5155	0.7832	0.1147	0.7915
	Model (b)								
15	0.2002	0.3074	0.3668	0.4892	0.2792	0.5633	0.7900	0.1918	0.8129
	0.2008	0.2770	0.3421	0.4919	0.2496	0.5516	0.7929	0.1713	0.8112
25	0.2032	0.2442	0.3177	0.4861	0.2240	0.5352	0.7869	0.1547	0.8020
	0.2028	0.2063	0.2893	0.4925	0.1831	0.5254	0.7918	0.1258	0.8018
50	0.2004	0.1982	0.2818	0.4852	0.1762	0.5162	0.7816	0.1253	0.7915
	0.2006	0.1840	0.2722	0.4890	0.1632	0.5155	0.7831	0.1146	0.7915
	Model (c)								
15	0.2003	0.3060	0.3657	0.4892	0.2783	0.5628	0.7898	0.1921	0.8129
	0.2009	0.2763	0.3416	0.4918	0.2493	0.5514	0.7926	0.1715	0.8110
25	0.2038	0.2444	0.3183	0.4860	0.2245	0.5353	0.7871	0.1552	0.8022
	0.2030	0.2062	0.2893	0.4926	0.1832	0.5255	0.7917	0.1259	0.8017
50	0.2006	0.1985	0.2822	0.4852	0.1760	0.5161	0.7813	0.1261	0.7914
	0.2006	0.1842	0.2723	0.4890	0.1632	0.5155	0.7829	0.1150	0.7913

Note: Monte Carlo bias (Bias), standard deviation (Std. Dev.), and Root Mean Square Error (RMSE) of the averaging estimator are reported in the first row for each sample size. Results for the OLS estimator are displayed in the successive second rows for sample sizes $n = 15$ and 25, while results for the 2SLS estimator are shown in the successive second row for sample size $n = 50$.

various estimators of β_{20}. In particular, the new estimator constructed with feasible optimal weights is compared against generic 2SLS (with a fixed number of instruments). Notice that for these sample sizes generic 2SLS with as many instruments as the sample size is equivalent to Ordinary Least Squares (OLS), i.e., for $n = 15$ and $n = 25$.[7]

The results are qualitatively the same across models for $R_f^2 = 0.01$ and $R_f^2 = 0.1$ (even with weak instruments, i.e., Model (a)). For a low first stage R^2, the proposed estimator has roughly the same bias but considerably larger variance than the OLS and 2SLS estimators when the level of endogeneity is low, resulting in larger RMSE in these cases. Although the proposed estimator displays smaller biases than the OLS and 2SLS estimators when $\varphi = 0.5$ and $\varphi = 0.8$, their RMSE are larger than the OLS estimator, but they are similar to the 2SLS ones. On the other hand, all estimators perform better when $R_f^2 = 0.1$. In particular, Table 4 shows that the proposed estimator has better bias performance across models,

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

TABLE 4. DGP 2: High-dimensional inference with $R_f^2 = 0.1$

	$\varphi = 0.2$			$\varphi = 0.5$			$\varphi = 0.8$		
n	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
	Model (a)								
15	0.1805	0.2933	0.3444	0.4320	0.2723	0.5106	0.7011	0.2024	0.7298
	0.1843	0.2652	0.3230	0.4465	0.2419	0.5078	0.7205	0.1764	0.7418
25	0.1782	0.2341	0.2942	0.4220	0.2160	0.4740	0.6883	0.1647	0.7078
	0.1853	0.1969	0.2704	0.4472	0.1768	0.4808	0.7214	0.1303	0.7331
50	0.1685	0.1870	0.2517	0.4044	0.1693	0.4384	0.6534	0.1327	0.6668
	0.1731	0.1718	0.2439	0.4202	0.1559	0.4482	0.6748	0.1178	0.6850
	Model (b)								
15	0.1782	0.2935	0.3433	0.4322	0.2736	0.5115	0.7015	0.2023	0.7301
	0.1831	0.2643	0.3215	0.4462	0.2424	0.5078	0.7213	0.1761	0.7425
25	0.1768	0.2354	0.2944	0.4217	0.2160	0.4738	0.6880	0.1649	0.7075
	0.1848	0.1974	0.2704	0.4472	0.1776	0.4812	0.7211	0.1306	0.7328
50	0.1650	0.1875	0.2498	0.4018	0.1684	0.4357	0.6538	0.1341	0.6674
	0.1719	0.1722	0.2433	0.4203	0.1553	0.4481	0.6746	0.1175	0.6847
	Model (c)								
15	0.1785	0.2906	0.3410	0.4325	0.2725	0.5111	0.7013	0.2037	0.7303
	0.1833	0.2626	0.3203	0.4465	0.2416	0.5077	0.7204	0.1764	0.7417
25	0.1789	0.2336	0.2942	0.4220	0.2162	0.4741	0.6877	0.1649	0.7072
	0.1850	0.1970	0.2702	0.4476	0.1775	0.4815	0.7206	0.1304	0.7323
50	0.1665	0.1874	0.2507	0.4032	0.1681	0.4369	0.6525	0.1355	0.6665
	0.1716	0.1723	0.2432	0.4203	0.1553	0.4481	0.6739	0.1181	0.6842

Note: Monte Carlo bias (Bias), standard deviation (Std. Dev.), and Root Mean Square Error (RMSE) of the averaging estimator are reported in the first row for each sample size. Results for the OLS estimator are displayed in the successive second rows for sample sizes $n = 15$ and 25, while results for the 2SLS estimator are shown in the successive second row for sample size $n = 50$.

endogeneity parameter and sample sizes, with smaller RMSE than the OLS and the 2SLS counterparts for $\varphi = 0.8$.

Practical Choice of Weights, $\mathcal{J}_n$ and τ

We have shown how to compute an optimal estimator for a given choice of $\mathcal{J}_n$. The theoretical analysis of methods for determining $\mathcal{J}_n$ is quite complex and would justify a separate paper, since it involves a higher order theory. In theory, the larger is $\mathcal{J}_n$ the better in terms of variance, but in practice there is a tradeoff. Let $\widehat{\theta}(\mathcal{J}_n)$ be the feasible optimal estimator as computed in the previous section, and let $\widehat{\Sigma}_{\text{opt}}(\mathcal{J}_n)$ be a consistent estimator of its asymptotic variance. Along the lines of Politis and Romano (1992), one could choose $\mathcal{J}_n$ to be the place where the standard errors (a scalar function of the covariance matrix) are relatively stable and do not vary wildly as $\mathcal{J}_n$ varies close by. In practice, we have found it useful to plot the input estimators against their marginal standard errors as an informal device to select reasonable estimators.

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

We now conclude by describing how one would select the set $\mathcal{J}_n$ in a given application with finite sample n. There are a number of informal methods a practitioner can use. For example, for given τ one could compute the t-statistic (in the scalar case, otherwise a chi-squared statistic) for each estimator and retain only those estimators with the largest τ such values. Alternatively, one could choose to retain only those estimators with t-statistic, say, that exceeds some predetermined level like one. Alternatively, if G_{nj} is continuously differentiable with respect to θ, one could also use results in Rilstone, Srivastava, and Ullah (1996); Rilstone and Ullah (2005) to estimate $\mathcal{J}_n$ and the weights by means of minimizing an estimate of the proposed estimator's mean square error (see e.g., Schafgans and Zinde-Walsh, 2010). The theoretical justification of the latter is left for future research.

8. CONCLUSIONS AND FINAL REMARKS

This paper provides the asymptotic theory of an estimator obtained by taking linear combinations of $\sqrt{n}$-consistent estimators, where the cardinality of the linear combination increases with sample size. The principle of averaging estimators is very general and has found applications in nonparametric estimation (i.e., Linton and Nielsen, 1995) and semiparametric estimation (i.e., Härdle and Stoker, 1989), where averaging can improve convergence rates. In random coefficient panel data models, estimation of average values usually proceeds in this way by averagingindividual specific or time specific based estimators (i.e., Swamy, 1970). In the parametric case, the main purpose is to improve efficiency. The traditional approach of combining moment conditions before estimation can achieve this purpose, but the approach of averaging estimators can achieve broadly the same benefits. There are two main advantages of this latter approach. First, it has some additional benefits in graphical diagnostics since one has a 'distribution' of estimators of the same quantity and one can view the range of values that the estimators take, which is more interpretable than the corresponding set of moment conditions. Second, it is robust in a certain sense to parameter heterogeneity. In that case, our average estimator can be interpreted as an estimator of the average parameter, while the average the moment condition procedure would generally be estimating some nonlinear functional of the parameter distribution (see e.g., Pesaran, 2006). The 'estimator selection' issue is important in cases where information about identification strength is limited, and in that case without some effective screening method our method is less robust than the corresponding GMM estimator. We do not wish to oversell our approach, but just to point out it is a feasible alternative with some modest benefits and costs.

NOTES

1. Also related is work by Donald, Imbens, and Newey (2003) who transform conditional moment restriction into increasing number of unconditional moment equations, and obtain efficiency and consistent asymptotic variance estimation under $\tau^2/n \to 0$ instead.

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

2. For this particular linear model, a closely related paper to ours is Lee and Zhou (2011).

3. It may be that the moments of $\widehat{\theta}_j$ defined in this way do not exist in finite samples (see e.g., Phillips, 1983). To avoid this issue, one could divide the instruments into groups with two or more members, estimate the individual 2SLS within the group, and then average as before.

4. We are grateful to Tom Rothenberg for pointing this out to us.

5. Interpreting 2SLS in various ways has a long history in econometrics; see Rothenberg (1974) for an early example.

6. Monte Carlo results based on other bases such as Hermite, Laguerre, or Legendre polynomials and different weighting schemes can be found in the working paper version of this paper (i.e., Chen, Jacho-Chávez, and Linton, 2009).

7. Although alternative methods that rely on choosing a subset of instruments are readily available, see e.g., Donald and Newey (2001) and Kuersteiner and Okui (2010).

REFERENCES

Antoine, B. & E. Renault (2012) Efficient minimum distance estimation with multiple rates of convergence. *Journal of Econometrics* 170(2), 350–367.

Belloni, A., D. Chen, V. Chernozhukov, & C. Hansen (2012) Spare models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80(6), 2369–2429.

Bickel, P.J. & E. Levina (2008) Covariance regularization by thresholding. *Annals of Statistics* 36(6), 2577–2604.

Bierens, H.J. (1987) Kernel estimators of regression functions. In T.F. Bewley (ed.), *Advances in Econometrics – Fifth World Congress of the Econometric Society, vol. I of Econometric Society Monographs*, 1st ed., chap. 3, pp. 99–144. Cambridge University Press.

Breiman, L. (1996) Bagging predictors. *Machine Learning* 24(2), 123–140.

Breiman, L. (1999) Using Adaptive Bagging to Debias Regressions. Technical report 547, Department of Statistics, University of California Berkeley.

Cai, T.T., W. Liu, & H.H. Zhou (forthcoming) Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Annals of Statistics*.

Campbell, J.Y., A. Lo, & A.C. Mackinlay (1997) *The Econometrics of Financial Markets*, 1st ed. Princeton University Press.

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

Chamberlain, G. (1992) Efficiency bounds for semiparametric regression. *Econometrica* 60(3), 567–596.

Chen, X., D.T. Jacho-Chávez, & O.B. Linton (2009) An Alternative Way of Computing Efficient Instrumental Variable Estimators. Sticerd Working paper EM/2009/536, STICERD.

Chen, X. & D. Pouzo (2009) Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics* 152(1), 46–60.

Chesher, A. (1984) Testing for neglected heterogeneity. *Econometrica* 52(4), 865–872.

Donald, S.G., G.W. Imbens, & W.K. Newey (2003) Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics* 117(1), 55–93.

Donald, S.G. & W.K. Newey (2001) Choosing the number of instruments. *Econometrica* 69(5), 1161–1191.

Fan, J., Y. Feng, & R. Song (2011) Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association* 106(494), 544–557.

Fan, J., Y. Li, & K. Yu (2012) Vast volatility matrix estimation using high-frequency data for portfolio selection. *Journal of the American Statistical Association* 107(497), 412–428.

Fan, J. & J. Lv (2008) Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B* 70, 849–911.

Granger, C.W.J. & Y. Jeon (2004) Thick modeling. *Economic Modelling* 21(2), 323–343.

Gray, H.L. & W.R. Schucany (1972) *The Generalized Jackknife Statistic*, 1st ed. Marcel Dekker.

Han, C. & P.C.B. Phillips (2006) GMM with many moment conditions. *Econometrica* 74(1), 147–192.
Hansen, L.P. (1982) Large sample properties of generalized methods of moments estimators. *Econometrica* 50, 1029–1054.
Hansen, L.P. (1985) A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators. *Journal of Econometrics* 30(1–2), 203–238.
Hansen, B.E. (2007) Least squares model averaging. *Econometrica* 75(4), 1175–1189.
Hansen, B.E. (2008) Least-squares forecast averaging. *Journal of Econometrics* 146(2), 342–350.
Hansen, B.E. (2009) Averaging estimators for regressions with a possible structural break. *Econometric Theory* 25(06), 1498–1514.
Hansen, B.E. (2010) Averaging estimators for autoregressions with a near unit root. *Journal of Econometrics* 158(1), 142–155.
Hansen, B.E. & J.S. Racine (2012) Jackknife model averaging. *Journal of Econometrics* 167(1), 38–46.
Härdle, W. & T.M. Stoker (1989) Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association* 84(408), 986–995.
Huber, P.J. (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In L.M.L. Cam & J. Neyman (eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pp. 221–233. Statistical Laboratory of the University of California, Berkeley. University of California Press.
Koenker, R.W. & G. Bassett. (1978) Regression quantiles. *Econometrica* 46(1), 33–50.
Koenker, R. & J.A.F. Machado (1999) GMM inference when the number of moment conditions is large. *Journal of Econometrics* 93(2), 327–344.
Kotlyarova, Y. & V. Zinde-Walsh (2006) Non- and semi-parametric estimation in models with unknown smoothness. *Economics Letters* 93(3), 379–386.
Kotlyarova, Y. & V. Zinde-Walsh (2007) Robust kernel estimator for densities of unknown smoothness. *Journal of Nonparametric Statistics* 19(2), 89–101.
Koul, H.L. & W. Stute (1999) Nonparametric model checks for time series. *Annals of Statistics* 27(1), 204–236.
Kuersteiner, G. & R. Okui (2010) Constructing optimal instruments by first-stage prediction averaging. *Econometrica* 78(2), 697–718.
Ledoit, O. & M. Wolf (2004) A well-conditioned estimator for large dimensional covariance matrices. *Journal of Multivariate Analysis* 88(2), 365–411.
Lee, L. (2010) Pooling estimates with different rates of convergence – the minimum χ^2 approach with an emphasis on a social interactions model. *Econometric Theory* 26, 260–299.
Lee, Y. & Y. Zhou (2011) Averaged Instrumental Variables Estimator. Unpublished manuscript.
Linton, O.B. & J.P. Nielsen (1995) A kernel model of estimating structured nonparametric regression based on marginal integration. *Biometrika* 82, 93–100.
Lobato, I.N. & M.A. Dominguez (2004) Consistent estimation of models defined by conditional moment restrictions. *Econometrica* 72(5), 1601–1615.
Malinvaud, E. (1966) *Statistical Methods of Econometrics*, 1st ed. Studies in Mathematical and Managerial Economics. Rand McNally.
McFadden, D. (1989) A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica* 57(5), 995–1026.
Newey, W.K. (1990) Efficient instrumental variables estimation of nonlinear models. *Econometrica* 58(4), 809–837.
Newey, W.K. (1993) Efficient estimation of models with conditional moment restrictions. In G.S. Maddala, C.R. Rao, & H.D. Vinod (eds.), *Handbook of Statistics*, 1st ed., vol. 11, pp. 419–454. Elsevier.
Newey, W.K. & D. McFadden (1994) Large sample estimation and hypothesis testing. In D. McFadden & R.F. Engle (eds.), *Handbook of Econometrics*, vol. IV, pp. 2111–2245. Elsevier.
Newey, W.K. & R.J. Smith (2004) Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* 72(1), 219–255.

Okui, R. (2011) Instrumental variable estimation in the presence of many moment conditions. *Journal of Econometrics* 165, 70–86.

Pakes, A. & D. Pollard (1989) Simulation and the asymptotics of optimization estimators. *Econometrica* 57(5), 1027–1057.

Pesaran, M.H. (2006) Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica* 74(4), 967–1012.

Phillips, P.C.B. (1983) Exact small sample theory in the simultaneous equations model. In Z. Griliches & M.D. Intriligator (eds.), *Handbook of Econometrics*, 1st ed., vol. 1, chap. 8, pp. 449–516. Elsevier.

Phillips, P.C.B. & H.R. Moon (1999) Linear regression limit theory for nonstationary panel data. *Econometrica* 67(5), 1057–1112.

Politis, D.N. & J.P. Romano (1992) A general resampling scheme for triangular arrays of α-mixing random variables with application to the problem of spectral density estimation. *Annals of Statistics* 20(4), 1985–2007.

Portnoy, S. (1984) Asymptotic behavior of M-estimators of p regression parameters when p2/n is large. I. Consistency. *Annals of Statistics* 12(4), 1298–1309.

Rilstone, P., V.K. Srivastava, & A. Ullah (1996) The second-order bias and mean squared error of nonlinear estimators. *Journal of Econometrics* 75(2), 369–395.

Rilstone, P. & A. Ullah (2005) Corrigendum to 'The second-order bias and mean squared error of nonlinear estimators'; [Journal of Econometrics 75(2) (1996) 369-395]. *Journal of Econometrics* 124(1), 203–204.

Rothenberg, T.J. (1973) *Efficient Estimation with A Priori Information*, 1st ed. Yale University Press.

Rothenberg, T.J. (1974) A Note on Two-Stage Least Squares. Unpublished manuscript.

Sawa, T. (1973) Almost unbiased estimator in simultaneous equations systems. *International Economic Review* 14(1), 97–106.

Schafgans, M.M.A. & V. Zinde-Walsh (2010) Smoothness adaptive average derivative estimation. *Econometrics Journal* 13(1), 40–62.

Shorack, G.R. & J.A. Wellner (2009) *Empirical Processes with Applications to Statistics*, 1st ed. Classics in Applied Mathematics. Society for Industrial & Applied Mathematics.

Stock, J.H. & M.W. Watson (1999) Forecasting inflation. *Journal of Monetary Economics* 44(2), 293–335.

Stock, J.H. & J. Wright (2000) GMM with weak identification. *Econometrica* 68(5), 1055–1096.

Swamy, P.A.V.B. (1970) Efficient inference in a random coefficient regression model. *Econometrica* 38(2), 311–323.

Watson, M.W. (2003) Macroeconomic forecasting using many predictors. In M. Dewatripont, L.P. Hansen, & S.J. Turnovsky (eds.), *Advances in Economics and Econometrics: Theory and Applications – Eighth World Congress of the Econometric Society, vol. III of Econometric Society Monographs*, 1st ed., chap. 3, pp. 87–115. Cambridge University Press.

Zhang, L., P.A. Mykland, & Y. Ait-Sahalia (2005) A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association* 100, 1394–1411.

Zhao, Z. & Z. Xiao (2014) Efficient regression via optimally combining quantile information. *Econometric Theory* 30(6), 1272–1314.

APPENDIX A: Proofs

We start with a result that is useful to aid the discussion of Assumption A*4.

LEMMA A.1. *Let U_{ji} be a triangular array of random variables, $i = 1, \ldots, n$, $j = 1, \ldots, \tau(n)$, i.i.d. across i for each j with $E(U_{ji}) = 0$ and $E[|U_{ji}|^{\kappa}] = c_j < \infty$ for*

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

some $\kappa \geq 2$. *Let* $s_{nj}^2 = \sum_{i=1}^n \mathrm{var}(U_{ji}) = n\sigma_j^2$, *where* $\sigma_j^2 \to \infty$ *as* $j \to \infty$, *and let*

$$a_n = \left(\max_{1 \leq j \leq \tau(n)} \sigma_j^2 \right) \log \tau(n) + \left(\sum_{j=1}^{\tau(n)} \frac{c_j^2}{\sigma_j^{2\kappa}} \right)^{1/\kappa}. \tag{A.1}$$

Then we have for $\delta_n = a_n \varrho_n$ *for any increasing sequence* ϱ_n *that*

$$\max_{1 \leq j \leq \tau(n)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n U_{ji} \right| = o_p(\delta_n).$$

For example, if we take $\kappa = 2$, then $a_n = (\max_{1 \leq j \leq \tau(n)} \sigma_j^2) \log \tau(n) + \sqrt{\tau(n)}$. One application of this lemma is when $n^{-1/2} \sum_{i=1}^n U_{ji}$ is the leading term of the estimator $\widehat{\theta}_j$, in which case, σ_j^2 would be Γ_j^{-1} (under homoskedasticity) as defined in (4.4) below. Therefore, the corresponding a_n is of order $\Gamma_{\tau(n)}^{-1} \log \tau(n) + \sqrt{\tau(n)}$. Provided $\tau(n)$ does not increase too rapidly, this is less than $n^{1/4}$ as would be required by Assumption A*4. Furthermore, it implies that $\max_{1 \leq j \leq \tau(n)} \left\| \widehat{\theta}_j - \theta_0 \right\|$ goes to zero no slower in probability than $\left(\Gamma_{\tau(n)}^{-1} \log \tau(n) + \sqrt{\tau(n)} \right) / \sqrt{n}$.

Proof of Lemma A.1. We show that

$$\Pr \left[\max_{1 \leq j \leq \tau(n)} \left| \sum_{i=1}^n U_{ji} \right| \geq \lambda_n \right] \to 0$$

for any $\lambda_n = \delta_n \sqrt{n}$. For an array $\chi_{nj} \to \infty$ as $n \to \infty$ for each j, write

$$\begin{aligned} U_{ji} &= U_{ji} 1(|U_{ji}| \leq \chi_{nj}) + U_{ji} 1(|U_{ji}| > \chi_{nj}) \\ &= \widetilde{U}_{ji} + \widetilde{\widetilde{U}}_{ji}. \end{aligned}$$

We shall assume for simplicity that U_{ji} is symmetric about zero so that $E(\widetilde{U}_{ji}) = 0$. Therefore, $\widetilde{U}_{ji}$ are i.i.d. for each j with mean zero and are bounded from above by χ_{nj}. By the Bonferroni and Bernstein inequalities

$$\begin{aligned} \Pr \left[\max_{1 \leq j \leq \tau(n)} \left| \sum_{i=1}^n \widetilde{U}_{ji} \right| \geq \lambda_n \right] &\leq \sum_{j=1}^{\tau(n)} \Pr \left[\left| \sum_{i=1}^n \widetilde{U}_{ji} \right| \geq \lambda_n \right] \\ &\leq \sum_{j=1}^{\tau(n)} \exp \left(\frac{-\lambda_n^2}{s_{nj}^2 + 2\lambda_n \chi_{nj}} \right). \end{aligned} \tag{A.2}$$

We shall choose λ_n and χ_{nj} below to make this term vanish.

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

By the Bonferroni and Markov inequalities

$$\begin{aligned}\Pr\left[\max_{1\le j\le\tau(n)}\left|\sum_{i=1}^{n}\widetilde{\widetilde{U}}_{ji}\right|\ge\lambda_n\right] &\le \sum_{j=1}^{\tau(n)}\Pr\left[\left|\sum_{i=1}^{n}\widetilde{\widetilde{U}}_{ji}\right|\ge\lambda_n\right]\\ &\le \sum_{j=1}^{\tau(n)}\frac{E\left(\left|\sum_{i=1}^{n}\widetilde{\widetilde{U}}_{ji}\right|^{\kappa}\right)}{\lambda_n^{\kappa}}\\ &\le \sum_{j=1}^{\tau(n)}\frac{n^{\kappa}E\left(\left|U_{ji}\right|^{\kappa}\right)\Pr\left[|U_{ji}|>\chi_{nj}\right]}{\lambda_n^{\kappa}}\\ &\le \sum_{j=1}^{\tau(n)}\frac{n^{\kappa}[E\left(\left|U_{ji}\right|^{\kappa}\right)]^2}{\lambda_n^{\kappa}\chi_{nj}^{\kappa}}=o(1)\end{aligned}$$

provided $\sum_{j=1}^{\tau(n)} n^{\kappa}\chi_{nj}^{-\kappa}\lambda_n^{-\kappa}c_j^2\to 0$.

Letting $\lambda_n=\delta_n\sqrt{n}$ and $\chi_{nj}=\sigma_j^2\sqrt{n}$ we need to show that:

$$\sum_{j=1}^{\tau(n)}\exp\left(\frac{-\delta_n}{\sigma_j^2}\right)\to 0 \quad\text{and}\quad \frac{1}{\delta_n^{\kappa}}\sum_{j=1}^{\tau(n)}\frac{c_j^2}{\sigma_j^{2\kappa}}\to 0.$$

For the first condition it suffices that

$$\frac{\delta_n}{\max_{1\le j\le\tau(n)}\sigma_j^2\log\tau(n)}\to\infty.$$

For the second condition it certainly suffices if

$$\frac{\delta_n}{\left(\sum_{j=1}^{\tau(n)}c_j^2\sigma_j^{-2\kappa}\right)^{1/\kappa}}\to\infty.$$

■

Proof of Theorem 1 (i). By the triangle inequality

$$\left\|\widehat{\theta}-\theta_0\right\|\le\Delta\sum_{j\in\mathcal{J}_n/\mathcal{J}_n^*}\left\|W_{nj}\right\|+\sum_{j\in\mathcal{J}_n^*}\left\|W_{nj}\right\|\max_{j\in\mathcal{J}_n^*}\left\|\widehat{\theta}_j-\theta_0\right\|, \tag{A.3}$$

where Δ is the finite radius of Θ. Therefore, it suffices that (4.2) holds, that $\widehat{\theta}_j$ are uniformly consistent over the class $\mathcal{J}_n^*$, and that $\sum_{j\in\mathcal{J}_n/\mathcal{J}_n^*}\left\|W_{nj}\right\|\to 0$, which is condition (4.3).

From A2, if $\max_{j\in\mathcal{J}_n^*}\left\|\widehat{\theta}_j-\theta_0\right\|>\delta$, then $\left\|G_j(\widehat{\theta}_j)\right\|\ge\epsilon_n(\delta)$ for some j. Consequently

$$\Pr\left(\max_{j\in\mathcal{J}_n^*}||\widehat{\theta}_j-\theta_0||>\delta\right)\le\Pr\left(\max_{j\in\mathcal{J}_n^*}\|G_j(\widehat{\theta}_j)\|\ge\epsilon_n(\delta)\right), \tag{A.4}$$

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

and it is sufficient to prove that for the given $\epsilon_n(\delta) > 0$, the latter probability goes to zero. But

$$\begin{aligned}
\max_{j\in\mathcal{J}_n^*}\|G_j(\widehat{\theta}_j)\| &\le \max_{j\in\mathcal{J}_n^*}\|G_j(\widehat{\theta}_j)-G_{nj}(\widehat{\theta}_j)\|+\max_{j\in\mathcal{J}_n^*}\|G_{nj}(\widehat{\theta}_j)\| \text{ by the Triangle Inequality,}\\
&\le \max_{j\in\mathcal{J}_n^*}\sup_{\theta\in\Theta}\|G_j(\theta)-G_{nj}(\theta)\|+\max_{j\in\mathcal{J}_n^*}\|G_{nj}(\widehat{\theta}_j)\| \text{ by set inclusion,}\\
&= o_p(\epsilon_{2n})+\max_{j\in\mathcal{J}_n^*}\|G_{nj}(\widehat{\theta}_j)\| \text{ by A4,}\\
&\le o_p(\epsilon_{2n})+\max_{j\in\mathcal{J}_n^*}\left(\|G_{nj}(\widehat{\theta}_j)\|-\inf_{\theta\in\Theta}\|G_{nj}(\theta)\|\right)+\max_{j\in\mathcal{J}_n^*}\inf_{\theta\in\Theta}\|G_{nj}(\theta)\|,\\
&\le o_p(\epsilon_{2n})+\max_{j\in\mathcal{J}_n^*}\left(\|G_{nj}(\widehat{\theta}_j)\|-\inf_{\theta\in\Theta}\|G_{nj}(\theta)\|\right)+\max_{j\in\mathcal{J}_n^*}\|G_{nj}(\theta_0)\|,\\
&= o_p(\epsilon_{2n})+o_p(\epsilon_{1n}) = o_p(\epsilon_n(\delta))
\end{aligned}$$

by A3, A4, and the definition of θ_0. We conclude that $\max_{j\in\mathcal{J}_n^*}||\widehat{\theta}_j-\theta_0|| = o_p(1)$ by A1. ■

Proof of Theorem 1 (ii). By (A.3) and condition A*1, we need only consider the set $\mathcal{J}_n^*$. Consistency Theorem 1 (i) implies that for every $\epsilon > 0$ there exists a sequence $\{\delta_n\}$ with $\delta_n \to 0$, and an N such that for all $n \ge N$, $\Pr\{\max_{j\in\mathcal{J}_n^*}||\widehat{\theta}_j-\theta_0|| > \delta_n\} \le \epsilon$. Using the same proof as that of Theorem 1(i), we have under our stronger Assumption A*4 that with probability approaching 1 (wpa1)

$$\begin{aligned}
\max_{j\in\mathcal{J}_n^*}\|G_j(\widehat{\theta}_j)\| &\le \max_{j\in\mathcal{J}_n^*}\|G_j(\widehat{\theta}_j)-G_{nj}(\widehat{\theta}_j)\|+\max_{j\in\mathcal{J}_n^*}\|G_{nj}(\widehat{\theta}_j)\|\\
&\le \max_{j\in\mathcal{J}_n^*}\sup_{\|\theta-\theta_0\|\le\delta_n}\|G_j(\theta)-G_{nj}(\theta)\|+\max_{j\in\mathcal{J}_n^*}\|G_{nj}(\widehat{\theta}_j)\|\\
&= o_p(\epsilon_n n^{-1/4})+\max_{j\in\mathcal{J}_n^*}\|G_{nj}(\widehat{\theta}_j)\| \text{ by A*4,}\\
&= o_p(\epsilon_n n^{-1/4}) \text{ by A*3, A*4, and the definition of } \theta_0.
\end{aligned}$$

Therefore, by A*2

$$\max_{j\in\mathcal{J}_n^*}||\widehat{\theta}_j-\theta_0|| \le \frac{1}{\min_{j\in\mathcal{J}_n}\gamma_j}\max_{j\in\mathcal{J}_n^*}\|G_j(\widehat{\theta}_j)\| = o(n^{-1/4}).$$

Hence

$$\max_{j\in\mathcal{J}_n^*}||\widehat{\theta}_j-\theta_0|| = o_p(n^{-1/4}),$$

which implies that

$$||\widehat{\theta}-\theta_0|| \le \Delta\sum_{j\in\mathcal{J}_n/\mathcal{J}_n^*}\|W_{nj}\|+\sum_{j\in\mathcal{J}_n^*}\|W_{nj}\|\times\max_{j\in\mathcal{J}_n^*}||\widehat{\theta}_j-\theta_0|| = o_p(n^{-1/4})$$

as required, since $\sum_{j\in\mathcal{J}_n^*}\|W_{nj}\|$ is uniformly bounded by A1. ■

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

Proof of Theorem 2. By (A.3) and condition (4.7), we need only consider the set $\mathcal{J}_n^*$. Let

$$L_{nj}(\theta) = G_{nj}(\theta_0) + \Gamma_j(\theta - \theta_0)$$

for each $j = 1, 2, \ldots$ Then define θ_j^* as the minimizer of $\|L_{nj}(\theta)\|$ over $\theta \in \mathbb{R}^p$. (Note that θ_j^* minimizes over $\mathbb{R}^p$, and not over Θ. We ignore this difference below because θ_j^* will eventually be in Θ w.p.1.) The solution satisfies

$$\sqrt{n}(\theta_j^* - \theta_0) = -\Gamma_j^{-1}\sqrt{n}G_{nj}(\theta_0) \tag{A.5}$$

for each j. Therefore,

$$\begin{aligned}\sqrt{n}\sum_{j\in\mathcal{J}_n^*} W_{nj}(\theta_j^* - \theta_0) &= \sqrt{n}\sum_{j\in\mathcal{J}_n^*} W_{nj}^0(\theta_j^* - \theta_0) + \sqrt{n}\sum_{j\in\mathcal{J}_n^*}(W_{nj} - W_{nj}^0)(\theta_j^* - \theta_0)\\ &= \sum_{i=1}^n T_{in} + R_n,\end{aligned}$$

where $R_n = \sqrt{n}\sum_{j\in\mathcal{J}_n^*}(W_{nj} - W_{nj}^0)(\theta_j^* - \theta_0)$ and $T_{in} = \frac{-1}{\sqrt{n}}\sum_{j\in\mathcal{J}_n^*} W_{nj}^0\Gamma_j^{-1}g_j(Z_i, \theta_0)$. The result follows after we establish:

(i) $\sum_{i=1}^n c^\top T_{in} \Longrightarrow N(0, c^\top \Sigma c)$ for any $c \in \mathbb{R}^p$ with $||c|| = 1$;
(ii) The remainder term $R_n = o_p(1)$;
(iii) $\sqrt{n}\sum_{j\in\mathcal{J}_n^*} W_{nj}(\theta_j^* - \widehat{\theta}_j) = o_p(1)$.

For (i), the triangular array of random variables $c^\top T_{in}$ is mean zero and independent across i for each n. By B5(a) we have:

$$\begin{aligned}\sum_{i=1}^n E[c^\top T_{in}]^2 &= E\left[\left(\sum_{j\in\mathcal{J}_n^*} c^\top W_{nj}^0\Gamma_j^{-1}g_j(Z_i,\theta_0)\right)^2\right]\\ &= \sum_{j\in\mathcal{J}_n^*}\sum_{l\in\mathcal{J}_n^*} c^\top W_{nj}^0\Gamma_j^{-1}E\left[g_j(Z_i,\theta_0)g_l(Z_i,\theta_0)^\top\right]\Gamma_l^{-1\top}W_{nl}^{0\top}c\\ &\to c^\top\Sigma c\,.\end{aligned}$$

Similarly, by B5(b) we have for some $\kappa > 0$,

$$\sum_{i=1}^n E|c^\top T_{in}|^{2+\kappa} \to 0.$$

Hence we obtain (i) by applying the Liapunov's triangular array central limit theorem.

For (ii), notice that Assumption B3(a) and (A.5) imply that $\max_{j\in\mathcal{J}_n^*}\|\Gamma_j\sqrt{n}(\theta_j^* - \theta_0)\| = O_p(1)$. This together with Assumption B4(a) implies (ii) because

$$\begin{aligned}\left\|\sqrt{n}\sum_{j\in\mathcal{J}_n^*}(W_{nj} - W_{nj}^0)(\theta_j^* - \theta_0)\right\| &\le \sqrt{n}\max_{j\in\mathcal{J}_n^*}\left\|\Gamma_j(\theta_j^* - \theta_0)\right\|\sum_{j\in\mathcal{J}_n^*}\left\|(W_{nj} - W_{nj}^0)\Gamma_j^{-1}\right\|\\ &= O_p(1)\times o_p(1).\end{aligned}$$

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

For (iii), by the $n^{1/4}$-consistency result, there exists a positive sequence $\eta_n \to 0$ such that $\Pr[n^{1/4}||\widehat{\theta}-\theta_0|| > \eta_n] \to 0$. For each j we have

$$\begin{aligned} G_{nj}(\theta) &= G_{nj}(\theta_0) + G_j(\theta) + G_{nj}(\theta) - G_j(\theta) - G_{nj}(\theta_0) \\ &= L_{nj}(\theta) + O(||\theta-\theta_0||^2) + [G_{nj}(\theta) - G_j(\theta)] - G_{nj}(\theta_0) \text{ by B2.} \end{aligned}$$

Therefore, for the above η_n and constants a and C we have

$$\begin{aligned} &\max_{j\in\mathcal{J}_n^*} \sup_{||\theta-\theta_0||\le a\eta_n/n^{1/4}} \sqrt{n}\|G_{nj}(\theta) - L_{nj}(\theta)\| \\ &\qquad \le C \times \eta_n^2 a^2 + \max_{j\in\mathcal{J}_n^*} \sup_{||\theta-\theta_0||\le a\eta_n/n^{1/4}} \sqrt{n}\|[G_{nj}(\theta) - G_j(\theta)] - G_{nj}(\theta_0)\| \\ &\qquad = O_p(\eta_n^2) + o_p(1) = o_p(1) \text{ by B3(b).} \end{aligned}$$

Therefore,

$$\max_{j\in\mathcal{J}_n^*} \|\sqrt{n}[L_{nj}(\theta_j^*) - G_{nj}(\theta_j^*)]\| = o_p(1), \text{ and } \max_{j\in\mathcal{J}_n^*} \|\sqrt{n}[L_{nj}(\widehat{\theta}_j) - G_{nj}(\widehat{\theta}_j)]\| = o_p(1)$$

because θ_j^* is $\sqrt{n}$-consistent and $\widehat{\theta}_j$ is $o(n^{-1/4})$-consistent. It now follows from the definition of θ_j^* and Assumption B1 and the triangular inequality that

$$\max_{j\in\mathcal{J}_n^*} \left| \sqrt{n}\|L_{nj}(\theta_j^*)\| - \sqrt{n}\|L_{nj}(\widehat{\theta}_j)\| \right| = o_p(1). \tag{A.6}$$

This implies that $\max_{j\in\mathcal{J}_n^*} \|\Gamma_j \sqrt{n}(\theta_j^* - \widehat{\theta}_j)\| = o_p(1)$, because of the properties of least squares residuals. Then we have

$$\begin{aligned} \sqrt{n} \sum_{j\in\mathcal{J}_n^*} W_{nj}(\theta_j^* - \widehat{\theta}_j) &\le \sum_{j\in\mathcal{J}_n^*} \left\| W_{nj}\Gamma_j^{-1} \right\| \times \max_{j\in\mathcal{J}_n^*} \|\Gamma_j \sqrt{n}(\theta_j^* - \widehat{\theta}_j)\| \\ &\le O_p(1) \times o_p(1) = o_p(1), \end{aligned}$$

where the last inequality is due to Assumption B4(a) and (b) since

$$\begin{aligned} \sum_{j\in\mathcal{J}_n^*} \left\| W_{nj}\Gamma_j^{-1} \right\| &\le \sum_{j\in\mathcal{J}_n^*} \left\| W_{nj}^0\Gamma_j^{-1} \right\| + \sum_{j\in\mathcal{J}_n^*} \left\| (W_{nj} - W_{nj}^0)\Gamma_j^{-1} \right\| \\ &= O(1) + o_p(1) = O_p(1), \end{aligned}$$

the result (iii) follows. ■

Proof of Proposition 1. On the one hand, by the results in Hansen (1982), the optimal GMM (oiv) estimator is asymptotically efficient among all regular $\sqrt{n}$-asymptotic normal estimators for the moment restrictions (5.2), hence $\Sigma_{\text{oiv}}^{\tau} \le \Sigma_{\text{opt}}^{\tau}$ in the positive semidefinite matrix sense. On the other hand, we notice that the oiv (optimal GMM) estimator has the expansion

$$\sqrt{n}(\widetilde{\theta}_{\text{oiv}}^{\tau} - \theta_0) = -(\Gamma^{\tau\top}\Psi_{\tau}^{-1}\Gamma^{\tau})^{-1}\Gamma^{\tau\top}\Psi_{\tau}^{-1}\sqrt{n}G_n^{\tau}(\theta_0) + o_p(1),$$

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

which can be rewritten as

$$\sqrt{n}(\widetilde{\theta}^{\tau}_{\text{oiv}}-\theta_0)=-\left(\sum_{j=1}^{\tau}\alpha_j\Gamma_j^{\top}\right)^{-1}\sum_{j=1}^{\tau}\alpha_j\sqrt{n}G_{nj}(\theta_0)+o_p(1), \tag{A.7}$$

where $\Gamma^{\tau\top}\Psi_{\tau}^{-1}=(\alpha_1,\ldots,\alpha_\tau)$ with $\alpha_j\in R^{p\times p}$, and $\Gamma^{\tau}=(\Gamma_1^{\top},\ldots,\Gamma_\tau^{\top})^{\top}$ with $\Gamma_j=E[A_j(X)D_0(X)^{\top}]$, and $G_{nj}(\theta)=\frac{1}{n}\sum_{i=1}^{n}A_j(X_i)\rho(Z_i,\theta)$ for $j=1,\ldots,\tau$. That is, the optimal GMM (oiv) estimator $\widetilde{\theta}^{\tau}_{\text{oiv}}$ belongs to the class of linear combinations of the $\widehat{\theta}_j$, $j=1,\ldots,\tau$ with

$$\widetilde{\theta}^{\tau}_{\text{oiv}}=\sum_{j=1}^{\tau}W_{0j}^{\text{oiv}}\widehat{\theta}_j+o_p(n^{-1/2}),$$

and

$$W_{0j}^{\text{oiv}}=-\left(\sum_{j=1}^{\tau}\alpha_j\Gamma_j^{\top}\right)^{-1}\alpha_j\Gamma_j^{\top}\quad\text{for } j=1,\ldots,\tau.$$

However, by the results in Rothenberg (1973), $\widehat{\theta}^{\tau}_{\text{opt}}=\sum_{j=1}^{\tau}W_{0j}^{\text{opt}}\widehat{\theta}_j$ is asymptotically efficient among the regular class of estimators of the form $\sum_{j=1}^{\tau}W_{0j}\widehat{\theta}_j$ with $\sum_{j=1}^{\tau}W_{0j}=I_p$, hence $\Sigma^{\tau}_{\text{opt}}\le\Sigma^{\tau}_{\text{oiv}}$ in the positive semidefinite matrix sense. Therefore $\Sigma^{\tau}_{\text{opt}}=\Sigma^{\tau}_{\text{oiv}}$ in (5.3). ■

Proof of Theorem 3. We have

$$\sum_{j\in\mathcal{J}_n}||(\widehat{W}_{0j}^{\text{opt}}-W_{0j}^{\text{opt}})\Gamma_j^{-1}||\le\tau(n)^{1+\rho_1}\max_{j\in\mathcal{J}_n}||(\widehat{W}_{0j}^{\text{opt}}-W_{0j}^{\text{opt}})||,$$

where

$$\begin{aligned}\widehat{W}_{0j}^{\text{opt}}-W_{0j}^{\text{opt}}&=\left[(i_\tau\otimes I_p)^{\top}\widehat{V}^{-1}(i_\tau\otimes I_p)\right]^{-1}\widehat{B}_j-\left[(i_\tau\otimes I_p)^{\top}V^{-1}(i_\tau\otimes I_p)\right]^{-1}B_j\\&=\left[(i_\tau\otimes I_p)^{\top}V^{-1}(i_\tau\otimes I_p)\right]^{-1}\left[\widehat{B}_j-B_j\right]\\&\quad+\left\{\left[(i_\tau\otimes I_p)^{\top}\widehat{V}^{-1}(i_\tau\otimes I_p)\right]^{-1}-\left[(i_\tau\otimes I_p)^{\top}V^{-1}(i_\tau\otimes I_p)\right]^{-1}\right\}B_j\\&\quad+\left\{\left[(i_\tau\otimes I_p)^{\top}\widehat{V}^{-1}(i_\tau\otimes I_p)\right]^{-1}-\left[(i_\tau\otimes I_p)^{\top}V^{-1}(i_\tau\otimes I_p)\right]^{-1}\right\}\left[\widehat{B}_j-B_j\right].\end{aligned}$$

Therefore, it suffices to prove that $||\widehat{V}^{-1}-V^{-1}||=o_p(n^{-\gamma})$, for some $\gamma>0$. Since the preliminary estimator is $\sqrt{n}$-consistent we can restrict our attention to the set $\{\theta:\|\theta-\theta_0\|\le\delta_n\sqrt{n}\}$ for some sequence $\delta_n\to0$. It follows that

$$\begin{aligned}\max_{j,l\in\mathcal{J}_n}\left\|\widehat{V}_{jl}-V_{jl}\right\|&\le\left(\max_{j,l\in\mathcal{J}_n}\left\|\widehat{\Omega}_{jl}-\Omega_{jl}\right\|+\max_{j\in\mathcal{J}_n}\left\|\widehat{\Gamma}_j-\Gamma_j\right\|\right)\left(\min_{j\in\mathcal{J}_n}\lambda_{\min}(\Gamma_j)\right)^{-2}\\&=O_p\left(n^{-\eta}\tau^{2\rho_1}\right)\end{aligned}$$

by D3.

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press

We use the following standard matrix results. Suppose that A is a nonsingular m by m matrix, and E is an arbitrary matrix of the same dimensions. If $\lambda_{\max}(A^{-1}E) \equiv \rho < 1$, then $A+E$ is nonsingular and

$$\lambda_{\max}\left((A+E)^{-1} - A^{-1}\right) \leq \frac{\lambda_{\max}(E) * \lambda^2_{\max}(A^{-1})}{1-\rho}.$$

We take $A = V$ and $E = \widehat{V} - V$. We further use the relation $\max|a_{i,j}| \leq \lambda_{\max}(A) \leq m \max|a_{i,j}|$ for any square m by m matrix A.

It then follows that

$$\begin{aligned}
\left\|\widehat{V}^{-1} - V^{-1}\right\| &\leq \tau(n)\lambda_{\max}(\widehat{V}^{-1} - V^{-1}) \\
&\leq \tau(n)\frac{\lambda_{\max}(\widehat{V}-V)\lambda_{\max}\left(V^{-1}\right)}{1-\lambda_{\max}(V^{-1}(\widehat{V}-V))} \\
&\leq \tau^2(n)\max_{j,l\in\mathcal{J}_n}\left\|\widehat{V}_{jl} - V_{jl}\right\| \times \frac{1}{\lambda_{\min}(V)} \\
&\quad \times \frac{1}{1-\tau\lambda^{-1}_{\min}(V)\max_{j,l\in\mathcal{J}_n}\left\|\widehat{V}_{jl} - V_{jl}\right\|} \\
&= o_p(n^{-\eta}\tau^{2+\rho+2\rho_1}).
\end{aligned}$$

Provided $(2+\rho+2\rho_1)c < \eta$ this will be $o_p(1)$. ■

https://doi.org/10.1017/S0266466614000851 Published online by Cambridge University Press