Frazier, David T.; Renault, Eric

Article

Indirect inference: Which moments to match?

Econometrics

Provided in Cooperation with:
MDPI – Multidisciplinary Digital Publishing Institute, Basel

Suggested Citation: Frazier, David T.; Renault, Eric (2019) : Indirect inference: Which moments to match?, Econometrics, ISSN 2225-1146, MDPI, Basel, Vol. 7, Iss. 1, pp. 1-17, https://doi.org/10.3390/econometrics7010014

econometrics

MDPI

Article

Indirect Inference: Which Moments to Match?

David T. Frazier [1,*] and Eric Renault [2,†]

[1] Department of Econometrics and Business Statistics, Monash University, Melbourne 3800, Australia
[2] Department of Economics, University of Warwick, Coventry CV4 7AL, UK; Eric_Renault@brown.edu
* Correspondence: david.frazier@monash.edu; Tel.: +61-9905-2973
† We thank Geert Dhaene for helpful comments and discussions.

Received: 19 December 2018; Accepted: 7 March 2019; Published: 19 March 2019

Abstract: The standard approach to indirect inference estimation considers that the auxiliary parameters, which carry the identifying information about the structural parameters of interest, are obtained from some recently identified vector of estimating equations. In contrast to this standard interpretation, we demonstrate that the case of overidentified auxiliary parameters is both possible, and, indeed, more commonly encountered than one may initially realize. We then revisit the "moment matching" and "parameter matching" versions of indirect inference in this context and devise efficient estimation strategies in this more general framework. Perhaps surprisingly, we demonstrate that if one were to consider the naive choice of an efficient Generalized Method of Moments (GMM)-based estimator for the auxiliary parameters, the resulting indirect inference estimators would be inefficient. In this general context, we demonstrate that efficient indirect inference estimation actually requires a two-step estimation procedure, whereby the goal of the first step is to obtain an efficient version of the auxiliary model. These two-step estimators are presented both within the context of moment matching and parameter matching.

Keywords: indirect inference; auxiliary models; overidentification

JEL Classification: C10; C14; C15

1. Introduction

Twenty-five years ago, with the publication of their manuscript on "Efficient Method of Moments" (hereafter, EMM), Gallant and Tauchen (1996) (hereafter, GT) made a seminal contribution to the field of simulation-based estimation and inference. The EMM estimation approach proposed by GT estimates parameters of the underlying structural model by "matching moments" defined through a score generator, namely the score function of some hypothesized auxiliary model. In the EMM approach, the efficiency of the resulting structural parameter estimators will occur when the score function for the chosen auxiliary model asymptotically spans the score function of the well-specified parametric model that has generated the data.

The efficiency argument underlying EMM estimation was further developed in Gallant and Long (1997), where it was argued that the score function of the SNP (SemiNonParametric) density of Gallant and Nychka (1987) spans the score of most relevant distributions, at least when the number of terms, K, in the SNP expansion diverges to infinity as the sample size diverges. Similarly, and concurrent with the development of EMM, following Smith (1993), Gourieroux et al. (1993) (hereafter, GMR) demonstrate that, for well-chosen weighting matrices used for moment matching, indirect inference estimators based on score matching will be asymptotically equivalent to indirect inference estimators based on the direct matching of auxiliary parameters.

Therefore, the question of "Which Moments to Match" is actually independent of the efficient indirect inference estimation strategy employed, which we dub, by analogy with the trinity of

asymptotic tests, the score approach (i.e., matching the score function of the auxiliary model) and the Wald approach (i.e., matching directly the estimators of auxiliary parameters). In contrast, GMR have shown that a likelihood ratio type of approach (also proposed by Smith (1993)) can lead to inefficient indirect inference estimators. Given the asymptotic equivalence between the score and Wald indirect inference approaches, the only pending issue for satisfactory application of indirect inference is then the selection of the auxiliary model.

Several authors, including Gallant and Long (1997), Andersen and Lund (1997) (hereafter, AL), and Gallant et al. (1997), have carefully discussed the choice of auxiliary model in the context of EMM, namely through the use of some SNP score generator. Not surprisingly they find that, as in any moment matching exercise, to achieve good finite-sample performance of the indirect inference estimator "it is important to conserve on the number of elements in the score generator" (AL), that is, on the number of moments to match within estimation. While this concern for parsimony is obviously sensible, one may consider different ways that it can be achieved. Basically, the aforementioned studies, as well as other EMM studies, put forward two principles.

Principle 1: Following Eastwood (1991), implementation of the SNP approach requires choosing the truncation degree in the expansion in an adaptive (i.e., random, data-dependent) manner. While AL interpret the results of Eastwood (1991) to suggest that AIC is the optimal model choice strategy for this adaptive truncation, they eventually decide to elicit a "choice of score generator(s) (...) guided by the more conservative HQC and BIC criteria". In addition, Gallant and Long (1997) and Gallant et al. (1997) also use the BIC in determining the choice of the auxiliary model, while the latter stresses that "to implement the EMM estimator we require a score generator that fits these data well", leading to the use of the BIC to measure the trade-off between parsimony vs goodness-of-fit. **Principle 2:** For a given number of terms in the SNP expansion, the score generator can be interpreted as the score of an unconstrained parametric model, which ensures that, by definition, we end up with a just-identified set of moment conditions to match: the number of auxiliary parameters to estimate is exactly the number of components in the score vector. For instance, and in contrast to Gallant and Long (1997), who allow for conditional heterogeneity in the innovation density, AL (see p. 364) "find no evidence that such an extension is required", and, by the same token, AL eliminate the additional heterogeneity parameter introduced by Gallant and Long (1997) and the corresponding moments to match. However, one may realize that an alternative approach would have been adding moment conditions aimed at utilizing the knowledge that this kind of heterogeneity is not present in the data, and which would then lead to an overidentified set of moments to match.

The purpose of the present paper is to revisit the issue of selecting an auxiliary model that is optimal for the purpose of indirect inference, in terms of efficiency, without tying our hands by having to adhere to Principles 1 and 2 above. We contend that these principles are stricter than necessary for the following reasons.

(i) We argue that the first principle puts too much emphasis on the idea that "to implement the EMM estimator we require a score generator that fits these data well". Gallant and Long (1997) show that (see their Lemma 1, p. 135) under convenient regularity conditions, asymptotic efficiency is reached if and only if the linear span of a "true score" (i.e., the score of a well-specified parametric model for the structural model) is asymptotically included (at the true value of the structural parameters) in the linear span of the score of the auxiliary model. This does not require in any way that the auxiliary model is (even asymptotically for an arbitrary large number of parameters) a well-specified model, in the sense that it is consistent with the Data Generating Process (DGP). Of course, as emphasized by GT, a sufficient condition for this score spanning property is the so-called smooth embedding of the score generator, which means that there is a one-to-one and twice continuously differentiable mapping between the two parametrizations (i.e., the auxiliary and structural). Then score spanning is just a consequence of computing compounded derivatives. However, this sufficient condition for score spanning is definitely not necessary and, thus, there is no logical argument to impose a model selection criterion, like AIC or BIC, to select an auxiliary

model. We remind the reader that the purpose of the auxiliary model is not to describe the DGP, but to provide informative estimating equations. After all, it may be possible to satisfy the linear spanning condition by using a vector of moments that define well-suited auxiliary parameters but have no interpretation as a score function of a quasi-likelihood.

(ii) The next point is the realization that what determines the efficiency of indirect inference estimators is the moments they match, and not necessarily the auxiliary parameters. Hence, and in contrast to the example given in Principle 2 above, one may well contemplate using a set of moment conditions that overidentify the vector of unknown auxiliary parameters. Of course, moment estimation of the auxiliary parameters will eventually resort to a just-identified set of moment conditions, through the choice of a particular linear combination of the (possibly) overidentified moment conditions. However, we argue that the choice of this just-identified set of moment conditions should *not* be guided by efficiency of the resulting estimator of auxiliary parameters (as would be the case with an efficient two-step Generalized Method of Moments (GMM) estimator) but, on the contrary, by our goal of obtaining an asymptotically efficient indirect estimator of the structural parameters. We demonstrate that this new focus of interest produces a novel way to devise a two-step GMM estimator of the auxiliary parameters that is, in general, different from standard efficient two-step GMM estimators.

The fact that we wish to relax Principles 1 and 2, with our main focus on asymptotically efficiency of the elicited indirect inference estimator, does not mean that we overlook the need "to conserve on the number" of moments to match, for the purpose of finite-sample performance. However, our claim is that the trade-off between parsimony and asymptotic efficiency should set the focus on indirect estimation of structural parameters and not on goodness-of-fit of the auxiliary model. Ideally, one should devise a procedure to select the valid and relevant moments to match, for instance by resorting to "an information-based LASSO for GMM with many moments" as recently proposed by Cheng and Liao (2015). The key idea of the Cheng and Liao (2015) approach is to define a new adaptive penalty that ensures the valid and relevant moment conditions are consistently selected in the GMM shrinkage estimation. This adaptive penalty depends on a measure of the information content of the moment conditions. However, while the measure of information used by Cheng and Liao (2015) is naturally based on the asymptotic variance of the resulting GMM estimator, we do not really care about the asymptotic variance of the GMM estimator of our auxiliary parameters. As explained above, our focus of interest is the asymptotic variance of the resulting indirect inference estimator of structural parameters. To this end, the present paper demonstrates the relevant way to measure the information content of the moments we wish to match. While we only consider here a given finite set of moments, the use of this new measure of information for the purpose of moment selection in a possibly infinite set, by a well-tuned procedure in the spirit of Cheng and Liao (2015), is left for future research.

The remainder of the paper is organized as follows. In Section 2, we present a very general framework for indirect inference consisting of a set of moment conditions that are able to identify both the structural parameters (given a value of the auxiliary parameters) and the auxiliary parameters (given a value of the structural parameters). This duality leads us to revisit the issue of model choice criteria for eliciting an informative auxiliary model. In Section 3, we demonstrate that there exists an efficient choice of the auxiliary model that can be used to construct asymptotically efficient indirect inference estimators. The equivalence between moment matching and parameter matching is maintained in this general setting, and, in both cases, we devise an efficient two-step procedure to construct efficient indirect inference estimators, where the goal of the first-step estimator is to obtain an efficient choice of the auxiliary model. We conclude the paper in Section 4 by paving the way for these tools to be used in the context of several popular auxiliary models, including moment models, and vector auto-regressive models, for which there is a clear trade-off between asymptotic efficiency and parsimony. Proofs of certain results are detailed in the Appendix A.

2. Auxiliary versus Structural Models

We argue that the most general framework for Indirect Inference can be accommodated through a set of k moment conditions whose information content is two-fold identification:

- **Identification** of the true unknown value θ^0 of the structural parameters $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$, for a given value (the true unknown one) of the auxiliary parameters.
- **Identification** of the true unknown value β^0 of the auxiliary parameters $\beta \in B \subset \mathbb{R}^{d_\beta}$, for a given value (the true unknown one θ^0) of the structural parameters.

While the former identification will be encapsulated in the definition of the structural model, including a parametric model that will allow us to replace analytically intractable moment conditions by their Monte Carlo counterparts, evaluated at any possible value of $\theta \in \Theta$, the latter identification will remain true to the standard GMM setting, where the true unknown value θ^0 is implicitly contained in the data generating process (hereafter, DGP). Note that both identification schemes maintain that the moment conditions are numerous enough to identify, or overidentify, the structural (resp., auxiliary) parameters when the true values of the auxiliary (resp., structural) parameters are given; i.e., it must be that

$$k \geq \min\left\{d_\beta, d_\theta\right\}.$$

2.1. Auxiliary Model

We consider an auxiliary model characterized by a finite number of moment restrictions. We set our focus on a vector of auxiliary moment functions $g(y, \beta)$ taking values in $\mathbb{R}^k$, which (possibly) depends on a random vector y and unknown auxiliary parameters $\beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta}$. This model admits a true unknown value β^0 of the auxiliary parameters, which satisfy the moments conditions

$$E\left[g(y, \beta^0)\right] = 0, \tag{1}$$

and where $E[.]$ denotes expectation taken with respect to the distribution of y. In other words, the unknown DGP is assumed to fulfill a vector of moment restrictions that are known up to a vector of auxiliary parameters β that must be consistently estimated from data.

For this purpose, we have at our disposal a sequence of stationary observations $\{y_t\}_{t=1}^T$ on the random vector y, which allows us to compute sample counterparts of the moment functions

$$\bar{g}_T(\beta) = \frac{1}{T}\sum_{t=1}^{T} g(y_t, \beta).$$

In this section, we maintain the classical assumptions of Generalized Method of Moments (hereafter, GMM), implying that the true unknown value β^0 is identified, both globally and locally at first-order, and the moments are not redundant at the true value.

Assumption 1. *The following assumptions are satisfied.*

(i) $E[g(y, \beta)] = 0 \iff \beta = \beta^0$.
(ii) $J = \frac{\partial E[g(y,\beta)]}{\partial \beta'}\big|_{\beta=\beta^0}$ *is full column rank.*
(iii) $\Omega = \text{plim}_{T=\infty}\left[\sqrt{T}\bar{g}_T\left(\beta^0\right)\right]$ *is a positive definite matrix.*

Any consistent estimator of β^0 based on these moment conditions, and denoted by $\hat{\beta}_T(K)$, can be seen, at least asymptotically, as the solution of the just identified set of equations

$$K\bar{g}_T\left(\hat{\beta}_T(K)\right) = \frac{1}{T}\sum_{t=1}^{T} Kg\left(y_t, \hat{\beta}_T(K)\right) = 0, \tag{2}$$

for some $(d_\beta \times k)$-dimensional selection matrix K, with rank d_β. For example, Equation (2) may be defined from the first-order conditions of the minimization problem

$$\min_{\beta \in \mathcal{B}} \|\bar{g}_T(\beta)\|_W^2, \tag{3}$$

where $\|x\|_W^2 := x'Wx$ denotes the weighted Euclidean norm. In such an example, the selection matrix K would actually be a random matrix depending on the observed sample and where the Jacobian matrix J and possibly also the weighting matrix in the squared norm are replaced by sample counterparts.

However, recalling that our ultimate goal is efficient estimation of θ^0 and not efficient estimation of β^0, the results of Bates and White (1993) (see their Section 3.2), which implies that among all the discrepancy functions allowing us to consistently estimate β^0 the optimal one is a quadratic form of $\bar{g}_T(\beta)$, are not applicable in this context. Therefore, Equation (2) must be viewed as general estimating equations where the selection matrix K may be replaced by a random sample counterpart $\hat{K}_T$. However, as far as first-order asymptotic distributional theory is concerned, the choice of a consistent estimator $\hat{K}_T$ of a selection matrix K is immaterial. For this reason, we simplify the notations by overlooking the (possible) dependence of K on the observed sample. Since $\hat{\beta}_T(K)$ is defined only through the estimating Equation (2), and not necessarily via a minimization problem like (3), we will require a slight strengthening of Assumption 1 (ii).

Assumption 2. *The matrix KJ is non-singular.*

We note that Assumption 2 would be implied by Assumption 1 in the case where the estimating Equations in (2) are obtained from the minimization of a quadratic form. However, since we do not wish to impose such a restriction, we must explicitly maintain Assumption 2 in the general case. Assumption Assumption 2 implies that the selection matrix K is of full row rank and also further restricts it with respect to the Jacobian matrix J.

Note that it is quite natural in practice to expect that Assumption 2 is fulfilled. For example, if a subset $\tilde{g}(y, \beta)$ of the components of $g(y, \beta)$ just identifies β (with a Jacobian matrix $\tilde{J}$ conformable to Assumption 1), then a selection matrix K built by combining $(k - d_\beta)$ zero columns with a non-singular $(d_\beta \times d_\beta)$-dimensional matrix $\tilde{K}$ will satisfy Assumption 2 so long as the zero columns of K are such that they only elicit the components of $\tilde{g}(y, \beta)$ in the sense that:

$$KJ = \tilde{K}\tilde{J}.$$

Under standard regularity conditions, Assumptions 1 and 2 together with a Taylor expansion of Equation (2) will allow us to view $\hat{\beta}_T(K)$ as a "linear estimator" with asymptotic expansion

$$\sqrt{T}\left[\hat{\beta}_T(K) - \beta^0\right] = -(KJ)^{-1}K\sqrt{T}\bar{g}_T(\beta^0) + o_P(1). \tag{4}$$

In particular, $\sqrt{T}\left[\hat{\beta}_T(K) - \beta^0\right]$ will be asymptotically normal with variance

$$\Sigma_K(\beta^0) = [(KJ)]^{-1}K\Omega K'[(KJ)']^{-1}.$$

While efficient GMM estimation of β^0 would minimize this variance matrix, by eliciting a selection matrix $K = J'\Omega^{-1}$ (to obtain the asymptotic variance $\left[J'\Omega^{-1}J\right]^{-1}$), it is not the purpose of the present paper.

More generally, we note that, for any weighting matrix W, minimization of $\|\bar{g}_T(\beta)\|_W^2$ would amount to a selection matrix $K = J'W$. However, as we will see in Section 3, the optimal selection matrix, for the purpose of efficient indirect inference estimation, does not belong to this general family.

2.2. The Structural Model

We assume the parametric structural model is characterized by a transition density function $p(y_t|\mathbf{Y}_{t-1};\theta), \theta \in \Theta \subset \mathbb{R}^{d_\theta}$ and $\mathbf{Y}_{t-1} = \{y_1, y_2, ..., y_{t-1}\}$. We denote, for $h = 1, ..., H$, $\{y_t^{(h)}(\theta)\}_{t=1}^T$ a simulated sample path from $p(.|.;\theta)$. Then, we end up with $(H+1)$ mutually independent paths consisting of T stationary observations on y, H simulated paths $\{y_t^{(h)}(\theta)\}_{t=1}^T$ $(h = 1, ..., H)$, which can be computed for any possible value $\theta \in \Theta$ and always using the same random seeds, and the observed path $\{y_t\}_{t=1}^T$. Throughout the remainder we let $E_\theta[\cdot]$ denote the expectation taken with respect to the distribution of $y_t^{(h)}(\theta)$.

As announced at the beginning of this section, we require identification assumptions about the structural parameters (for given β^0) similar to the identification assumptions maintained about β in Assumption 1, and where θ^0 is implicitly given by the DGP. Note that, by definition of the true value θ^0, it corresponds to the DGP in the sense that, for all $\beta \in \mathcal{B}$,

$$E_{\theta^0}\left[g(y,\beta)\right] = E\left[g(y,\beta)\right].$$

To ensure identification of θ^0, we maintain the following global and local identification assumptions.

Assumption 3. *The following assumptions are satisfied.*

(i) $E_\theta\left[g(y,\beta^0)\right] = 0 \iff \theta = \theta^0$.
(ii) $\Gamma = \frac{\partial E_\theta\left[g(y,\beta^0)\right]}{\partial \theta'}\Big|_{\theta=\theta^0}$ *is full column rank.*

Assumption 3 is required to define a consistent and asymptotically normal indirect inference (hereafter, II) estimator of θ^0. More precisely, we consider II estimators defined through the following minimization program:

$$\hat{\theta}_{T,H}(K,W) = \arg\min_{\theta\in\Theta}\left[\frac{1}{TH}\sum_{h=1}^{H}\sum_{t=1}^{T} g\left(y_t^{(h)}(\theta), \hat{\beta}_T(K)\right)\right]' W \left[\frac{1}{TH}\sum_{h=1}^{H}\sum_{t=1}^{T} g\left(y_t^{(h)}(\theta), \hat{\beta}_T(K)\right)\right], \quad (5)$$

where W is a given positive definite weighting matrix.[1]

The II estimator $\hat{\theta}_{T,H}(K,W)$ is a generalization of the score-matching one initially proposed by GT. In particular, GT consider a similar minimization program to (5), but specifically require that $g(\cdot)$ is the score vector of some parametric auxiliary model with parameters β, which ensures that the dimension of $g(\cdot)$ and the dimension of β coincide. As a result, no selection matrix K is required in the GT approach.

In particular, our setting includes the case where, as in GT, the function $g(\cdot)$ is the score vector of some auxiliary model, but where this model is subject to a set of exclusion restrictions, such that the number of free parameters d_β is much smaller than the dimension k of the score vector $g(\cdot)$. A typical example of this situation would be a Vector Auto-Regressions (VAR) auxiliary model (see, e.g., Smith (1993)), which, for the sake of parsimony, is subject to some exclusion restrictions. Note that when AL eliminates the additional heterogeneity parameters introduced by Tauchen (1997), we would again be in a case where $k > d_\beta$ if some components of the complete score vector had not been arbitrarily eliminated.

In contrast to the GT approach, the II estimator (5) only requires $k \geq d_\beta$. This estimator is our focus of interest in this paper. We first note that we can describe this estimator through a standard

[1] Similar to the selection matrix K, the weighting matrix W can be replaced by a consistent, data-dependent estimator, say W_T such that $W_T \to_p W$ as $T \to \infty$, without altering the first-order asymptotic theory of the II estimator.

linear representation, similarly to common II estimators: under standard regularity conditions, a Taylor expansion of the first-order conditions for (5) and the linear expansion of $\sqrt{T}\left[\hat{\beta}_T(K) - \beta^0\right]$ in (4) yield[2]

$$\sqrt{T}\left[\hat{\theta}_{T,H}(K,W) - \theta^0\right] = \left[\Gamma' W \Gamma\right]^{-1} \Gamma' W \times \tag{6}$$
$$\left\{ J\left(KJ\right)^{-1} K \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left\{ g\left(y_t, \beta^0\right)\right\} - \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left\{ \frac{1}{H} \sum_{h=1}^{H} g\left(y_t^{(h)}(\theta^0), \beta^0\right)\right\}\right\} + o_P(1).$$

From Equation (6), we can conclude that $\sqrt{T}\left[\hat{\theta}_{T,H}(K,W) - \theta^0\right]$ is an asymptotically normal vector whose asymptotic variance may be reduced by the fact that our minimization (5) involves the entire vector of moment functions $g(y,\beta)$ and *not* just the subset $g_K(y,\beta) = Kg(y,\beta)$.

The reason for this possibility of variance reduction is two-fold. First, the component of (6) that is computed from simulated data uses the entire set of estimating functions $g(y,\beta)$ and not only the subset $g_K(y,\beta)$. This should be beneficial in terms of asymptotic variance in the same way that in standard GMM (see, e.g., the linear representation (4)), adding valid moment functions can only decrease (weakly) the asymptotic variance of the efficient GMM estimator. However, in the case of (6), this efficiency gain would vanish as the number simulated paths, H, diverges to infinity.

Second, and more importantly, the multiplicative factor in the asymptotic expansion for $\hat{\theta}_T(K,W)$ depends on the matrix $\left[\Gamma' W \Gamma\right]^{-1}$, which is determined by the Jacobian matrix for the entire set of estimating functions $g(\cdot,\beta)$ and not the subset of moments $g_K(\cdot,\beta)$. As such, and for Ω as defined in Assumption 1, we know from the theory of efficient GMM estimation that an asymptotic variance $\left[\Gamma'\Omega^{-1}\Gamma\right]^{-1}$ is smaller (or equal) to the asymptotic variance $\left[\Gamma' K' \left(K\Omega K'\right)^{-1} K\Gamma\right]^{-1}$, where the latter is obtained by considering only the subset $g_K(y,\beta)$ of estimating functions.

The intuition behind this possible efficiency gain will be formally confirmed in Section 3. For this purpose, it is worth shedding first more light on the issue of model choice that, as explained in the introduction, has been often used in the extant literature for the selection of the auxiliary model.

2.3. Moment Selection Criterion

As already explained, the standard strategy for II based on a score generator amounts to picking a parsimonious auxiliary model, with the implication being that certain components within a possibly large set of estimating functions are eliminated (jointly with some auxiliary parameters) to arrive at a vector of just identified auxiliary parameters β. We can nest this particular case within our general setup by examining the case where, all along the II estimation strategy, including the minimization (5), we decide to use only a given just identified subset of $g(y,\beta)$, namely

$$g_K(y,\beta) = Kg(y,\beta).$$

When considering the just identified subset of moments $g_K(y,\beta)$, we must redefine the Jacobian matrices J and Γ by $J_K = KJ$ and $\Gamma_K = K\Gamma$, respectively, as well as the variance matrix $\Omega_K = K\Omega K'$. Denoting W_K as the weighting matrix used in the minimization program (5) (with $g(y,\beta)$ replaced by $g_K(y,\beta)$), we can deduce that the corresponding II estimator has the following linear asymptotic expansion:

$$\begin{aligned}\sqrt{T}\left[\hat{\theta}_{T,H}^{(K)}(K,W) - \theta^0\right] &= \left[\Gamma_K' W_K \Gamma_K\right]^{-1} \Gamma_K' W_K \\ &\left\{ \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left\{ g_K\left(y_t, \beta^0\right)\right\} - \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left\{ \frac{1}{H} \sum_{h=1}^{H} g_K\left(y_t^{(h)}(\theta^0), \beta^0\right)\right\}\right\} + o_P(1),\end{aligned}$$

[2] See the Appendix A for more detailed derivations.

and the asymptotic variance of this II estimator is given by

$$\Sigma_H^{(K)}(K, W_K) = \left(1 + \frac{1}{H}\right) \left[\Gamma_K' W_K \Gamma_K\right]^{-1} \Gamma_K' W_K \Omega_K W_K \Gamma_K \left[\Gamma_K' W_K \Gamma_K\right]^{-1}.$$

In this case, the optimal choice of the weighting matrix is $W_K = \Omega_K^{-1}$, which would lead to an allegedly optimal asymptotic variance given by estimator

$$\Sigma_H^{*(K)}(K) = \left(1 + \frac{1}{H}\right) \left[\Gamma_K' \Omega_K^{-1} \Gamma_K\right]^{-1}. \tag{7}$$

However, the optimality of this II estimator is questionable. In particular, a more efficient II estimator could be obtained if we could overcome the information loss due to the selection of moments through the matrix K, which would allow us to obtain the asymptotic variance

$$\Sigma_H^{*} = \left(1 + \frac{1}{H}\right) \left[\Gamma' \Omega^{-1} \Gamma\right]^{-1}. \tag{8}$$

Surprisingly enough, we will show in the next section that (8) is an efficiency bound and that it can be feasibly reached in the standard context of II where $d_\beta \geq d_\theta$, i.e., when the vector of auxiliary parameters is large enough to possibly identify the structural parameters. However, it will turn out that obtaining such an efficiency bound will require a two-step II estimation procedure, where an optimal selection matrix K^* will be estimated in the first step, and the second step will use this estimator to deduce an II estimator that is capable of reaching the efficiency bound (8).

As one would suspect, two-step II estimators that reach the efficiency bound in (8) can be obtained under both the "moment matching" and "parameter matching" approaches to II. We note here that the reference to "moment matching" and "parameter matching" is explicit, and from hereon we will no longer refer to such approaches as "score matching" and "Wald approach", since the auxiliary model may not be a parametric model endowed with a score function. In particular, in this generalization it may not be possible to interpret the estimator of the auxiliary parameters as a quasi-maximum likelihood estimator, as these auxiliary estimators will be based on a preliminary estimator of an optimal selection matrix, K^*, the solution of the equation

$$\Gamma_K' \Omega_K^{-1} \Gamma_K = \Gamma' \Omega^{-1} \Gamma. \tag{9}$$

In this way, the optimal choice of the selection matrix K must be related to the Jacobian matrix Γ of the structural model and is definitely not what would be elicited by a model selection criterion blindly applied to the auxiliary model. As will be made clear later, the optimal selection matrix K^*, defined through (9), is not in general consistent with GMM estimation of the auxiliary parameters.

3. Which Moments to Match?

3.1. Optimal Selection Matrix K

The goal of this subsection is to deduce a selection matrix K that solves Equation (9). For this purpose, it is worth revisiting Equation (7) in terms of orthogonal projections. More precisely, let us associate to any given selection matrix K, of dimension $d_\beta \times k$ and rank d_β, a full column rank matrix X, defined as

$$X = (\Omega^{1/2})' K',$$

where $\Omega^{1/2}$ is any matrix such that

$$\Omega = (\Omega^{1/2})(\Omega^{1/2})'.$$

We can then write

$$\begin{aligned}\Gamma_K'\Omega_K^{-1}\Gamma_K &= \Gamma'K'\left(K\Omega K'\right)^{-1}K\Gamma \\ &= \Gamma'\left[(\Omega^{1/2})'\right]^{-1}X\left(X'X\right)^{-1}X'\left[\Omega^{1/2}\right]^{-1}\Gamma \\ &= \tilde{\Gamma}'P_X\tilde{\Gamma},\end{aligned}$$

where

$$\tilde{\Gamma} = \left[\Omega^{1/2}\right]^{-1}\Gamma, \text{ and } P_X = X\left(X'X\right)^{-1}X'.$$

Taking $A \succeq B$ to mean that the difference $(A - B)$ is positive semi-definite, we obviously deduce

$$\left[\Gamma_K'\Omega_K^{-1}\Gamma_K\right]^{-1} = \left[(P_X\tilde{\Gamma}')(P_X\tilde{\Gamma})\right]^{-1} \succeq \left[\tilde{\Gamma}'\tilde{\Gamma}\right]^{-1} = \left[\Gamma'\Omega^{-1}\Gamma\right]^{-1},$$

with equality if

$$P_X\tilde{\Gamma} = \tilde{\Gamma}. \tag{10}$$

The above confirms that the asymptotic variance of the II estimator $\hat{\theta}_{T,H}(K,\Omega_K^{-1})$ does not, in general, reach the efficiency bound (8) but can reach this bound when condition (10) is fulfilled, which requires that the columns of $\tilde{\Gamma}$ are in the range of the matrix X; i.e., this requires that

$$\left[\Omega^{1/2}\right]^{-1}\Gamma = X\Lambda = (\Omega^{1/2})'K'\Lambda$$

for some matrix Λ. In other words, a matrix K will satisfy (9) if there exists some matrix Λ such that

$$K'\Lambda = \Omega^{-1}\Gamma. \tag{11}$$

If the selection matrix K fulfills Equation (11), not only will the corresponding II estimator $\hat{\theta}_{T,H}(K,\Omega_K^{-1})$ reach the efficiency bound in (8), but it will also be asymptotically equivalent to the optimal II estimator $\hat{\theta}_{T,H}(K,W)$ defined in (5). To see this, we note from Equation (6)

$$\begin{aligned}\operatorname*{plim}_{T\to\infty}\operatorname{Var}\left\{\sqrt{T}\left[\hat{\theta}_{T,H}(K,W)-\theta^0\right]\right\} &= \left[\Gamma'W\Gamma\right]^{-1}\Gamma'WJ\left(KJ\right)^{-1}K\Omega K'\left(J'K'\right)^{-1}J'W\Gamma\left[\Gamma'W\Gamma\right]^{-1} \\ &+ \frac{1}{H}\left[\Gamma'W\Gamma\right]^{-1}\Gamma'W\Omega W\Gamma\left[\Gamma'W\Gamma\right]^{-1}.\end{aligned}$$

The second term is obviously minimized by choosing $W = \Omega^{-1}$. Interestingly enough, this choice is also optimal for the minimization of the first term, at least when condition (11) is fulfilled. To see this note that

$$\begin{aligned}\Gamma'WJ\left(KJ\right)^{-1}K\Omega K'\left(J'K'\right)^{-1}W\Gamma &= \Gamma'\Omega^{-1}J\left(KJ\right)^{-1}K\Omega K'\left(J'K'\right)^{-1}J'\Omega^{-1}\Gamma \\ &= \Lambda'KJ\left(KJ\right)^{-1}K\Omega K'\left(J'K'\right)^{-1}J'K'\Lambda \\ &= \Lambda'K\Omega K'\Lambda \\ &= \Gamma'\Omega^{-1}\Omega\Omega^{-1}\Gamma \\ &= \Gamma'\Omega^{-1}\Gamma.\end{aligned}$$

Hence, when K satisfies Equation (11), the asymptotic variance of $\hat{\theta}_{T,H}(K,W)$ is minimized at $W = \Omega^{-1}$, and this asymptotic variance achieves the efficiency bound in (8). Moreover, we note that when $d_\beta \geq d_\theta$, it is always possible to construct a choice of K conformable to (11); a selection matrix satisfying Equation (11) is given by

$$K^{*\prime} = \begin{bmatrix} \Omega^{-1}\Gamma & C \end{bmatrix}, \tag{12}$$

for C an arbitrary $k \times (d_\beta - d_\theta)$-dimensional matrix, with rank $(d_\beta - d_\theta)$ and whose columns do not belong to the space spanned by the columns of $\Omega^{-1}\Gamma$. In this case, Equation (11) is fulfilled for

$$\Lambda = \begin{bmatrix} Id_{d_\theta} \\ 0 \end{bmatrix},$$

where the zero lower block of Λ has dimension $(d_\beta - d_\theta) \times d_\theta$.

In order to summarize, three comments are in order. First, Σ_H^* defined by (8) is obviously an asymptotic efficiency bound for estimation of θ^0, at least when the number H of simulation goes to infinity. The asymptotic variance $\lim_{H\to\infty} \Sigma_H^*$ corresponds to the asymptotic variance of the efficient GMM estimator based on the infeasible moment conditions for θ:

$$E_\theta \left[g\left(y_t^{(h)}(\theta), \beta^0 \right) \right] = 0,$$

which are infeasible since they depend on the unknown value β^0 of the auxiliary parameters.

Second, importantly this efficiency bound is feasible, insofar as $d_\beta \geq d_\theta$ and, of course, up to consistent estimation of a selection matrix K^* satisfying (12) (the following two subsections describe the construction of such a consistent estimator).[3] It must be stressed that the identities in (7) and (8) imply that, when $K = K^*$, there is no additional efficiency gain when working with the whole vector $g\left(y_t^{(h)}(\theta), \hat{\beta}_T(K) \right)$ of moment functions in the moment matching estimator (5).

Third, and in contrary to the above point, when $d_\beta < d_\theta$ we cannot reach this efficiency bound in general and it would be more efficient to use the entire vector $g\left(y_t^{(h)}(\theta), \hat{\beta}_T(K) \right)$ of moment functions in the moment matching estimator (5). Otherwise, any selection matrix K used in the computation $\hat{\theta}_{T,H}^{(K)}(K, W)$ will likely lead to some information loss, since it cannot be chosen in an optimal manner (according to minimal asymptotic variance).

3.2. Efficient Two-Step Moment Matching

When $d_\beta \geq d_\theta$, an efficient two-step II estimation procedure for θ^0 can proceed as follows. Let $\tilde{\beta}_T = \hat{\beta}_T(K)$ be associated to some arbitrary $(d_\beta \times k)$-dimensional selection matrix K with rank d_β. As with standard two-step GMM, this will allow us to compute a consistent estimator $\hat{\Omega}_T$ of the matrix Ω. In turn, this allows us to compute a consistent estimator $\tilde{\theta}_T$ of θ^0 as

$$\tilde{\theta}_T = \arg\min_{\theta\in\Theta} \left[\frac{1}{T} \sum_{t=1}^{T} g\left(y_t^{(1)}(\theta), \tilde{\beta}_T \right) \right]' \hat{\Omega}_T^{-1} \left[\frac{1}{T} \sum_{t=1}^{T} g\left(y_t^{(1)}(\theta), \tilde{\beta}_T \right) \right]. \tag{13}$$

If $\tilde{\beta}_T$ were almost surely equal to the true unknown value β^0, the estimator of θ^0 defined by (13) would actually reach the efficiency bound (8). To see this, recall that, as already mentioned, the efficiency bound actually coincides with the asymptotic variance of an efficient GMM estimator of θ^0 based on the moment conditions $E_\theta \left[g(y, \beta^0) \right] = 0$. Unfortunately, these moment conditions are not feasible in general and, thus, $\tilde{\theta}_T$ incurs some efficiency loss because we have used a first-step estimator $\tilde{\beta}_T = \hat{\beta}_T(K)$, of β^0, that is sub-optimal (as far as estimation of θ^0 is concerned). Therefore, $\tilde{\theta}_T$ is nothing but a possibly inefficient "first-step" consistent estimator of θ^0. Indeed, there is no compelling reason to consider the first-step estimator $\tilde{\theta}_T$, defined by (13), over the more naive first-step estimator

$$\tilde{\theta}_T = \arg\min_{\theta\in\Theta} \left\| \frac{1}{T} \sum_{t=1}^{T} g\left(y_t^{(1)}(\theta), \tilde{\beta}_T \right) \right\|^2,$$

[3] A discussion and interpretation of this surprising feasibility (in spite of the fact that β^0 is unknown) is given in Section 3.4.

since both estimators are inefficient.

So long as $\tilde{\theta}_T$ is a consistent estimator of θ^0 (irrespective of the weighting matrix used for its computation), we can deduce a consistent estimator $\hat{\Gamma}_{T,H^*}$, of $\Gamma(\theta^0)$, using[4]

$$\hat{\Gamma}_{T,H^*} = \frac{\partial}{\partial \theta'} \left\{ \frac{1}{TH^*} \sum_{h=1}^{H^*} \sum_{t=1}^{T} g\left(y_t^{(h)}(\theta), \tilde{\beta}_T\right) \right\}_{\theta=\tilde{\theta}_T}.$$

Using this consistent estimator of $\Gamma(\theta^0)$, define the selection matrix $\hat{K}_T$ as

$$\hat{K}_T' = \left[\begin{array}{cc} \hat{\Omega}_T^{-1}\hat{\Gamma}_{T,H^*} & C_T \end{array} \right]$$

where C_T is an arbitrary $k \times \left(d_\beta - d_\theta\right)$-dimensional matrix with rank $\left(d_\beta - d_\theta\right)$, and whose columns do not belong to the space spanned by the columns of $\hat{\Omega}_T^{-1}\hat{\Gamma}_{T,H^*}$.

For $C_T \to_p C$, where C is an arbitrary $k \times \left(d_\beta - d_\theta\right)$-dimensional matrix, with rank $\left(d_\beta - d_\theta\right)$, and whose columns do not belong to the space spanned by the columns of $\Gamma'(\theta^0)\Omega^{-1}$, we can conclude that

$$\operatorname*{plim}_{T\to\infty} \left[\begin{array}{c} \hat{\Gamma}_{T,H^*}'\hat{\Omega}_T^{-1} \\ C_T' \end{array} \right] = \left[\begin{array}{c} \Gamma'(\theta^0)\Omega^{-1} \\ C' \end{array} \right] = K^*$$

Now, letting $\hat{\beta}_T = \hat{\beta}_T(\hat{K}_T)$ be defined as the solution of

$$\hat{K}_T \bar{g}_T(\hat{\beta}_T) = \frac{1}{T} \sum_{t=1}^{T} \hat{K}_T g(y_t, \hat{\beta}_T) = 0,$$

it directly follows, from the above arguments, that the feasible two-step II estimator

$$\hat{\theta}_{T,H}\left(\hat{K}_T, \hat{\Omega}_T^{-1}\right) = \arg\min_{\theta\in\Theta} \left[\frac{1}{TH} \sum_{h=1}^{H} \sum_{t=1}^{T} g\left(y_t^{(h)}(\theta), \hat{\beta}_T\left(\hat{K}_T\right)\right) \right]' \hat{\Omega}_T^{-1} \left[\frac{1}{TH} \sum_{h=1}^{H} \sum_{t=1}^{T} g\left(y_t^{(h)}(\theta), \hat{\beta}_T\left(\hat{K}_T\right)\right) \right]$$

will reach the efficiency bound in Equation (8).

3.3. Efficient Two-Step Parameter Matching

We now examine a version of the above II estimator based on parameter matching, where we explicitly work under the assumption that the "auxiliary" parameters β "identify" the structural parameters θ. Note that this assumption was not maintained for the general moment matching considered above. Assumption 4 thus complements Assumption 3 in this respect.

Assumption 4. *The selection matrix K is a $(d_\beta \times k)$-dimensional matrix with rank d_β, which fulfills the following conditions.*

(i) There exists a function $b_K(.)$ such that, for all $\theta \in \Theta$,

$$E_\theta\left[Kg(y, b_K(\theta))\right] = 0.$$

(ii) $b_K(\theta^0) = \beta^0 = \operatorname{plim}_{T\to\infty} \hat{\beta}_T(K)$ and $b_K(\theta) = \operatorname{plim}_{T\to\infty} \tilde{\beta}_{T,H}(\theta, K)$, where $\tilde{\beta}_{T,H}(\theta, K)$ is the solution of

$$\frac{1}{TH} \sum_{t=1}^{T} \sum_{h=1}^{H} Kg\left(y_t^{(h)}(\theta), \tilde{\beta}_{T,H}(\theta, K)\right) = 0.$$

---

[4] Note that one may want to compute this derivative numerically and to choose H^* very large (and possibly different from H in the rest of the procedure) to take advantage of the smoothness properties of the population moments.

(iii) $b_K(\theta) = \beta^0 \iff \theta = \theta^0$.
(iv) *The Jacobian matrix* $\partial b_K(\theta^0)/\partial\theta'$ *has rank* d_θ.

In this context, we can define a parameter matching II estimator of θ^0 as

$$\hat{\theta}^P_{T,H}(K,W) = \arg\min_{\theta\in\Theta} \left[\hat{\beta}_T(K) - \tilde{\beta}_{T,H}(\theta,K)\right]' W_T \left[\hat{\beta}_T(K) - \tilde{\beta}_{T,H}(\theta,K)\right], \tag{14}$$

where $W = \text{plim}_{T\to\infty} W_T$ is a $(d_\beta \times d_\beta)$-dimensional positive definite matrix. GMR show that (for given K), asymptotically efficient estimation of θ^0 is delivered by an optimal choice $W(K)$ of W that is the inverse of the asymptotic variance of $\hat{\beta}_T(K)$. Therefore, from the expansion in Equation (4), the optimal $W(K)$ is given by

$$W(K) = \left\{(KJ)^{-1}(K\Omega K')(J'K')^{-1}\right\}^{-1}.$$

Hence, by using the result of GMR, the asymptotic variance of the II estimator in Equation (14) is given by

$$\operatorname*{plim}_{T\to\infty} \text{Var}\left[\sqrt{T}\left[\hat{\theta}^P_{T,H}[K,W(K)] - \theta^0\right]\right] = \left(1+\frac{1}{H}\right)\left\{\frac{\partial b_K'(\theta^0)}{\partial\theta}(KJ)'(K\Omega K')^{-1}(KJ)\frac{\partial b_K(\theta^0)}{\partial\theta'}\right\}^{-1}. \tag{15}$$

Moreover, by differentiating the identity

$$KE_\theta\left[g\left(y_t^{(h)}(\theta), b_K(\theta)\right)\right] = 0,$$

we obtain

$$K\Gamma(\theta^0) + KJ\frac{\partial b_K(\theta^0)}{\partial\theta'} = 0 \tag{16}$$

so that Equation (15) can be rewritten as

$$\operatorname*{plim}_{T\to\infty} \text{Var}\left[\sqrt{T}\left[\hat{\theta}^P_{T,H}[K,W(K)] - \theta^0\right]\right] = \left(1+\frac{1}{H}\right)\left\{\Gamma'(\theta^0)K'(K\Omega K')^{-1}K\Gamma(\theta^0)\right\}^{-1}.$$

Not surprisingly, we find that the optimal parameter matching II estimator $\hat{\theta}^P_{T,H}[K,W(K)]$ based on a selection matrix K is asymptotically equivalent to the optimal moment matching II estimator $\hat{\theta}_{T,H}(K,\Omega^{-1})$ based on the same selection matrix. As a consequence, an optimal choice K^* satisfying Equation (11) will allow us to reach the efficiency bound (8). Moreover, this optimal selection matrix is the same for the optimal II estimator based on moment matching and the optimal II estimator based on parameter matching, with the two estimators being asymptotically equivalent. Indeed, an efficient two-step parameter matching estimator can be devised in a similar fashion to the efficient two-step moment matching estimator described in the former subsection.

However, it is important to note the distinction between the required identification conditions underpinning the two estimators, i.e., the moment matching estimator, $\hat{\theta}_{T,H}(K,W)$, and the parameter matching estimator, $\hat{\theta}^P_{T,H}(K,W)$. In particular, the identification conditions underpinning the parameter matching approach is much more restrictive than the conditions required for the moment matching estimator. In particular, Assumption **3** simply assumes, for given β^0, identification of the moment conditions at θ^0, as well as the existence of a consistent and asymptotically normal estimator $\hat{\beta}_T(K)$. However, in addition to the existence of a consistent and asymptotically normal estimator $\hat{\beta}_T(K)$, Assumption 4 assumes that we consider only selection matrices K for which there exists both a continuously differentiable limit map, $b_K(\theta)$, with full column-rank Jacobian, as well as a consistent estimator.

3.4. Interpretation of Results and Discussion

As alluded to in Section 3.1, the optimal asymptotic variance (8) demonstrates that we do not pay a price for ignoring the value of β^0 in terms of optimal II estimation of θ^0, which means that $\lim_{H\to\infty} \Sigma_H^*$ corresponds to the asymptotic variance of an efficient GMM estimator for the infeasible moment conditions about θ,

$$E_\theta \left[g\left(y_t^{(h)}(\theta), \beta^0 \right) \right] = 0.$$

The intuition behind this result is relatively simple and can be expressed as follows: when the binding function $b(.)$ is known, which is precisely the case for an infinite number of simulations, we can estimate β^0 through additional estimating equations $\beta - b(\theta) = 0$ that just identify β. This result echoes the following well-known result in GMM estimation theory (see, e.g., (Breusch et al. 1999)): when additional moment restrictions just identify the additional nuisance parameters that they introduce, they do not modify the accuracy of the efficient GMM estimator of the parameters of interest. As already announced in the introduction, what really matters for the efficiency of the II estimators is a choice of the auxiliary model well focused on the structural model. Hence the definition of the optimal selection matrix K^*.

To better understand this issue, imagine a favorable case where the overidentifying information pertains to the complete path of the binding function, meaning that, for all $\theta \in \Theta$,

$$E_\theta \left[g(y, b(\theta)) \right] = 0. \tag{17}$$

In other words, the binding function $\theta \mapsto b(\theta)$ does not depend on a specific selection matrix K, and thus there is no conflict between efficient estimation of β^0 and efficient estimation of θ^0. We can actually check this directly, since differentiating the identity (17) yields

$$\Gamma(\theta^0) + J \frac{\partial b(\theta^0)}{\partial \theta'} = 0. \tag{18}$$

Therefore,

$$\Omega^{-1}\Gamma(\theta^0) = -\Omega^{-1} J \frac{\partial b(\theta^0)}{\partial \theta'} = -K' \frac{\partial b(\theta^0)}{\partial \theta'} \tag{19}$$

when K has been chosen for efficient GMM estimation of β^0 (i.e., $K' = \Omega^{-1}J$). In this case, the selection matrix K corresponding to efficient GMM estimation of β^0 does fulfill the optimality condition (11) that produces an efficient II estimator of θ^0. The identity (19) indeed demonstrates that, for this choice of K, the columns of $\Omega^{-1}\Gamma(\theta^0)$ are linear combinations of the columns of K'. However, it may be argued that an identity like (17) should be the exception rather than the rule. There is a striking difference between the identity (16) that we obtained in a general setting (but for a given selection matrix K) and the much stronger condition (18) that does not depend on K and allows us to get altogether efficient II estimation of θ^0 and efficient GMM estimation of β^0. In contrast, it is worth stressing that any choice of K based on a GMM estimator of β^0 is generically inconsistent with efficient II estimation of θ^0 since, in general, for any weighting matrix W and any matrix Λ

$$K = J'W \Longrightarrow K'\Lambda = WJ\Lambda \neq \Omega^{-1}\Gamma.$$

It is worth realizing that situations of tension between efficient estimation of β^0 on the one hand and efficient estimation of θ^0 on the other hand (typically when (17) is violated) go beyond the simple framework considered in this paper. For instance, Sargan (1983) and Dovonon and Renault (2013) have stressed that for non-linear GMM, β may be globally identified by (1) while first-order identification may fail at some particular value β^0 because the matrix Γ is not full column rank. It turns out that in many circumstances (see, e.g., (Dovonon and Renault 2013)) the particular value at which rank deficiency occurs is precisely the case of interest. Dovonon and Hall (2018) have documented the

implication of such a lack of first-order identification for II when using the naive selection matrix $K = J'\Omega^{-1}$.

Recall that the messages of this paper are two-fold: one, the naive selection matrix $K = J'\Omega^{-1}$ may not be an efficient choice; two, more importantly, the efficient choice is based on a matrix Γ, the rank of which has no reason to be deficient when there is a rank deficiency in the matrix J. Therefore, it may well be the case that standard asymptotic theory for II is still valid, in contrast with the case of Dovonon and Hall (2018), when II is performed efficiently.

A similar argument applies in the case of weak identification (see, e.g., (Stock and Wright 2000) and (Kleibergen 2005)), that is, when the matrix Γ is only asymptotically rank deficient. A general theory of II in the case of first-order under-identification or weak identification of the auxiliary parameters is left for future research.

4. Conclusions

The overall message of this paper can be summarized as follows: application of the II methodology may require, for the sake of finite-sample performance, the imposition of certain constraints on the auxiliary model, leading to auxiliary parameters that are defined by an overidentified system of moment conditions. Typically, when II is based on a score generator, a la GT, possibly due to some interpretation of the auxiliary parameters, one would imagine that there exist more restrictions than are needed to identify the auxiliary parameters. In this context, we demonstrate that efficient indirect estimators of the structural parameters should take advantage of the overidentifying restrictions, but for the purpose of optimizing the accuracy of the II estimator of the structural parameters θ, not for the GMM estimator of the auxiliary parameters β, which is generally not equivalent.

The general characterization of a two-step efficient II estimator proposed in this paper may be used in various contexts. As a first example, we have in mind procedures that select valid and relevant moments to match, as recently devised by Cheng and Liao (2015), through an "information-based LASSO". If we contemplate applying this procedure to the choice of the auxiliary moment model, the information criterion of Cheng and Liao (2015) would be based on the asymptotic variance $\left[J'\Omega^{-1}J\right]^{-1}$ of the efficient GMM estimator of β. In contrast, for the purpose of efficient II estimation of θ^0, it would be more relevant to use as an information criterion the efficient asymptotic variance $\left[\Gamma'\Omega^{-1}\Gamma\right]^{-1}$ of an II estimator of θ (up to a correction factor for the number of simulations). Theoretically speaking, nothing prevents us to revisit the theory of Cheng and Liao (2015) in this new context of indirect moment-based estimation.

A second example may be inspired by the recent work of Hansen (2016) on "Stein Combination Shrinkage" for Vector Auto-Regressions (VAR). Since the work of Smith (1993), VAR models have been a popular class of auxiliary models for II estimation of structural models stemming from macroeconomic theory, however, there exists little guidance in the extant literature about the trade-off between efficiency and parsimony in the specification of these VAR models for II estimation. Moreover, the poor finite-sample performance of estimators for VAR models of dimension larger than two, and with more than one lag, has been widely documented, but the consequences for employing such a class of auxiliary models within II estimation have not been meaningfully studied. Hansen (2016) uses the tool of model averaging for the aggregation of estimators of VAR parameters provided by different possible shrinkage strategies. In our framework, this can be understood as averaging over estimators of auxiliary parameters β provided by different selection matrices K. Then, in our context, the procedure of model averaging should be elicited with regards to efficient II estimation of the structural parameters θ, which differs from the issue of efficient estimation of the VAR parameters β.

More generally this paper contributes to the search for efficiency in the context of II estimation. We emphasize the fact that the moments to match, or equivalently, the score generator provided by the auxiliary model, should not be treated as a statistical object whose inference must be efficient within the logic of the auxiliary world. Instead, auxiliary models should only be used as lenses focused on minimizing the asymptotic variance of the indirect estimator of θ obtained by calibrating the estimating

equations on β, without overlooking some finite sample issues related to the parsimony in the choice of these equations. Future research includes extensions to the cases of weak identification, first-order under-identification, and model misspecification.

Author Contributions: The creation and publication of this manuscript has been a collaborative effort, with both authors contributing to every facet of the manuscript. This includes conceptualization and investigation of the main ideas in the manuscript, methodology proposals, and formal analysis, as well as all aspects of the writing process.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proofs

Appendix A.1. Proof of Equation (6)

The first-order conditions that define $\hat{\theta}_{T,H}(K,W)$ can be written

$$\Gamma'(\theta^0)W\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\{\frac{1}{H}\sum_{h=1}^{H}g\left(y_t^{(h)}\left(\hat{\theta}_{T,H}(W)\right),\hat{\beta}_T(K)\right)\right\}=o_P(1),$$

which, after a first order Taylor expansion gives

$$\begin{aligned}o_P(1) &= \Gamma'(\theta^0)W\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\{\frac{1}{H}\sum_{h=1}^{H}g\left(y_t^{(h)}\left(\theta^0\right),\beta^0\right)\right\}+\Gamma'(\theta^0)W\Gamma(\theta^0)\sqrt{T}\left[\hat{\theta}_{T,H}(W)-\theta^0\right]\\ &\quad+\Gamma'(\theta^0)WJ\sqrt{T}\left[\hat{\beta}_T(K)-\beta^0\right].\end{aligned}$$

Plugging in the linear expansion (4) of the estimator $\hat{\beta}_T(K)$, i.e.,

$$\sqrt{T}\left[\hat{\beta}_T(K)-\beta^0\right]=-(KJ)^{-1}K\sqrt{T}\bar{g}_T(\beta^0)+o_P(1).$$

we obtain

$$\begin{aligned}\sqrt{T}\left[\hat{\theta}_{T,H}(K,W)-\theta^0\right] &= \left[\Gamma'(\theta^0)W\Gamma(\theta^0)\right]^{-1}\Gamma'(\theta^0)W\\ &\quad\left\{J(KJ)^{-1}K\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\{g\left(y_t,\beta^0\right)\right\}-\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\{\frac{1}{H}\sum_{h=1}^{H}g\left(y_t^{(h)}(\theta^0),\beta^0\right)\right\}\right\}+o_P(1)\end{aligned}\tag{A1}$$

In other words, $\sqrt{T}\left[\hat{\theta}_{T,H}(K,W)-\theta^0\right]$ is asymptotically a linear function of the asymptotically Gaussian vector

$$\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\{g\left(y_t,\beta^0\right)\right\},\frac{1}{\sqrt{T}}\sum_{t=1}^{T}g\left(y_t^{(h)}(\theta^0),\beta^0\right)\right]_{1\le h\le H}$$

Appendix A.2. Proof of Equation (15)

If $\hat{\beta}_T(K)$ has been computed from the sample counterpart of estimating equations

$$K\bar{g}_T\left(\hat{\beta}_T(K)\right)=0$$

for some matrix K of dimension $d_\beta\times k$ and of rank d_β, we have, by assumption,

$$\sqrt{T}\left[\hat{\beta}_T(K)-\beta^0\right]=-(KJ)^{-1}K\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\{g\left(y_t,\beta^0\right)\right\}+o_P(1)$$

and

$$\sqrt{T}\left[\tilde{\beta}_{T,H}(\theta^0,K)-\beta^0\right]=-(KJ)^{-1}K\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{H}\sum_{h=1}^{H}g\left(y_t^{(h)}(\theta^0),\beta^0\right)+o_P(1).$$

Thus, taking the difference

$$\begin{aligned}&\sqrt{T}\left[\hat{\beta}_T(K)-\tilde{\beta}_{T,H}(\theta^0,K)\right]\\ =\;&-(KJ)^{-1}K\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\{g\left(y_t,\beta^0\right)-\frac{1}{H}\sum_{h=1}^{H}g\left(y_t^{(h)}(\theta^0),\beta^0\right)\right\}+o_P(1)\end{aligned}$$

Therefore, $\sqrt{T}\left[\hat{\beta}_T(K)-\tilde{\beta}_{T,H}(\theta^0,K)\right]$ is asymptotically normal with asymptotic variance

$$\left(1+\frac{1}{H}\right)(KJ)^{-1}K\Omega K'[(KJ)']^{-1}$$

Hence, by using the result of GMR, the asymptotic variance of the indirect inference estimator is then

$$\left\{\text{plim}_{T\to\infty}\text{Var}\left[\sqrt{T}\left[\hat{\theta}_{T,H}^{P}[K,W(K)]-\theta^0\right]\right]\right\}=\left(1+\tfrac{1}{H}\right)\left\{\tfrac{\partial b_K'(\theta^0)}{\partial\theta}(KJ)'(K\Omega K')^{-1}(KJ)\tfrac{\partial b_K(\theta^0)}{\partial\theta'}\right\}^{-1}\tag{A2}$$

Moreover, by differentiating the identity

$$KE\left[g\left(y_t^{(h)}(\theta),b_K(\theta)\right)\right]=0$$

we obtain

$$K\Gamma(\theta^0)+KJ\frac{\partial b_K(\theta^0)}{\partial\theta'}=0$$

so that (A2) can be rewritten:

$$\left\{\underset{T\to\infty}{\text{plim}}\,\text{Var}\left[\sqrt{T}\left[\hat{\theta}_{T,H}^{P}P[K,W(K)]-\theta^0\right]\right]\right\}=\left(1+\frac{1}{H}\right)\left\{\Gamma'(\theta^0)K'(K\Omega K')^{-1}K\Gamma(\theta^0)\right\}^{-1}.$$

References

Andersen, Torben G., and Lund Jesper. 1997. Estimating continuous-time stochastic volatility models of the short-term interest rate. *Journal of Econometrics* 77: 343–77. [CrossRef]

Bates, Charles E., and White Halbert. 1993. Determination of estimators with minimum asymptotic covariance matrices. *Econometric Theory* 9: 633–48. [CrossRef]

Breusch, Trevor, Hailong Qian, Peter Schmidt, and Wyhowski Donald. 1999. Redundancy of moment conditions. *Journal of Econometrics* 91: 89–111. [CrossRef]

Cheng, Xu, and Zhipeng Liao. 2015. Select the valid and relevant moments: An information-based lasso for GMM with many moments. *Journal of Econometrics* 186: 443–64. [CrossRef]

Dovonon, Prosper, and Alastair R. Hall. 2018. The asymptotic properties of GMM and indirect inference under second-order identification. *Journal of Econometrics* 205: 76–111.

Dovonon, Prosper, and Renault Eric. 2013. Testing for common conditionally heteroskedastic factors. *Econometrica* 81: 2561–86. [CrossRef]

Eastwood, Brian J. 1991. Asymptotic normality and consistency of semi-nonparametric regression estimators using an upwards F test truncation rule. *Journal of Econometrics* 48: 151–81. [CrossRef]

Gallant, A. Ronald, and Jonathan R. Long. 1997. Estimating stochastic differential equations efficiently by minimum chi-squared. *Biometrika* 84: 125–41. [CrossRef]

Gallant, A. Ronald, and Douglas W. Nychka. 1987. Semi-nonparametric maximum likelihood estimation. *Econometrica: Journal of the Econometric Society* 55: 363–90. [CrossRef]

Gallant, A. Ronald, and George Tauchen. 1996. Which moments to match? *Econometric Theory* 12: 657–81. [CrossRef]

Gallant, A. Ronald, Hsieh David, and George Tauchen. 1997. Estimation of stochastic volatility models with diagnostics. *Journal of Econometrics* 81: 159–92. [CrossRef]

Gourieroux, Christian, Monfort Alain, and Renault Eric. 1993. Indirect inference. *Journal of Applied Econometrics* 8: S85–S118. [CrossRef]

Hansen, Bruce E. 2016. Stein Combination Shrinkage for Vector Autoregressions. Working Paper A39, Sir Clive Granger Building, University Park, PA, USA.

Kleibergen, Frank. 2005. Testing parameters in GMM without assuming that they are identified. *Econometrica* 73: 1103–23. [CrossRef]

Sargan, J. D. 1983. Identification and lack of identification. *Econometrica: Journal of the Econometric Society* 51: 1605–33. [CrossRef]

Smith, A. A., Jr. 1993. Estimating nonlinear time-series models using simulated vector autoregressions. *Journal of Applied Econometrics* 8: S63–S84. [CrossRef]

Stock, James H., and Jonathan H. Wright. 2000. GMM with weak identification. *Econometrica* 68: 1055–96. [CrossRef]

Tauchen, George. 1997. New minimum chi-square methods in empirical finance. *Econometric Society Monographs* 28: 279–317.