Noname manuscript No.
(will be inserted by the editor)

Analytical Finite Sample Econometrics – from A. L. Nagar to Now

Yong Bao · **Aman Ullah**

Received: date / Accepted: date

Abstract Professor A.L. Nagar was a world-renowned econometrician and an international authority on finite sample econometrics with many path-breaking papers on the statistical properties of econometric estimators and test statistics. His contributions to applied econometrics have been also widely recognized. Nagar's 1959 *Econometrica* paper on the so-called k-class estimators, together with a later one in 1962 on the double-k-class estimators, provided a very general framework of bias and mean squared error approximations for a large class of estimators and had motivated researchers to study a wide variety of issues such as many and weak instruments for many decades to follow. This paper reviews Nagar's seminal contributions to analytical finite sample econometrics by providing historical backgrounds, discussing extensions and generalization of Nagar's approach, and suggesting future directions of this literature.

Keywords Nagar · finite sample econometrics · k-class estimators

JEL Classification C10 C13 C18

1 Introduction

In the early 20th century, the legendary statisticians Sir R.A. Fisher, Neyman, Pearson and others set in motion what is known today as the classical parametric approach to statistics, estimation and testing of a finite number of population

Y. Bao
Department of Economics, Purdue University, West Lafayette, IN 47907, USA
Tel.: +1-765-496-2313
Fax: +1-765-496-1778
E-mail: ybao@purdue.edu

A. Ullah
Department of Economics, University of California, Riverside, CA 92521, USA
Tel.: +1-951-827-1591
Fax: +1-951-827-5685
E-mail: aman.ullah@ucr.edu

parameters using sample data. Thus began the practice of statistical inference (estimation, testing and confidence intervals, and their sampling properties) and this laid the foundation of inference procedures in econometrics. Studies on the properties of finite (small) and large sample properties of statistics based on any sample data became a hot topic from that time onward. While this statistical revolution was taking place, in 1930 the international Econometric Society was established in the USA, with the purpose of advancing the study of econometrics, formally described as the data analysis (measurement) of mathematical economic models using statistical-inference methods. Further, under this society the first econometric journal, *Econometrica*, came into existence in 1933 with a supporting fund by Alfred Cowles and Nobel Laureate Ragnar Frisch being its first editor. Earlier, the Statistical Society of London was established in 1834 at England, which began the publication of the *Journal of the Royal Statistical Society* (JRSS) in 1838, but an unrelated London Statistical Society was in existence at least as early as 1824. During this period, International Statistical Institute was established in 1885 in the Netherlands, and the American Statistical Association (ASA) was established in the USA in 1839 with a journal published under it in 1888 (*Quarterly Publications of the American Statistical Association*, and later as the *Journal of the American Statistical Association* (JASA) in 1922). Later on, Institute of Mathematical Statistics came in 1935 in the USA, and the *Annals of Mathematical Statistics* started publishing from 1930 to 1972 and then split into the *Annals of Statistics* and the *Annals of Probability* since 1973. With all these new excitements and developments in the area of econometrics (statistical) inference there was a keen interest in studying the sampling properties of econometric estimators such as unbiasedness, efficiency (lower mean squared error (MSE)), and distribution.

While the properties of the ordinary least squares (OLS) estimator in a classical linear regression model were well developed as the best linear unbiased estimator (the Gauss–Markov Theorem), the small sample bias of the OLS estimator of the first-order autoregressive (AR(1)) coefficient was developed by Hurwitz (1950). However, it was quite challenging for econometricians to develop properties of econometric estimators that are nonlinear functions of random variables. These include estimators in dynamic time-series models, operational generalized least squares estimators, estimators in no-linear regression models, and the two-stage least squares (2SLS), limited information maximum likelihood (LIML), instrumental variables (IV), and k-class estimators in the structural models. Essentially, the challenges appeared because almost all the econometric estimators and test statistics were either nonlinear functions of stochastic random variables or came as iterative solutions to some nonlinear optimizations of random functions. Therefore, deriving the finite sample distributions, bias, and MSE were statistically challenging. The results for many cases are still unknown in the literature, even at this stage.

Exact sampling distribution theory is an implicit promise of inference in statistical sciences. Given its near impossibility in general, even under normally distributed data, for many relevant estimators and test functions of the data, higher-order approximations than the first order asymptotics is the next best thing. Given a common inadequacy of the first order asymptotics, currently called asymptotic inference, the need and promise of higher order asymptotic approximations naturally arise.[1] This is what he anticipated with his higher order approximations to moments, when Nagar (1959) published his first seminal paper to derive the bias, up to order $O(n^{-1})$, where n is the sample size, and MSE, up to $O(n^{-2})$, of the k-class estimator of the vector coefficients, say $\hat{\boldsymbol{\beta}}$, of a structural equation in the simultaneous equations model. His approach was to simply first write the Taylor series stochastic expansion of the estimator in a decreasing order in probability like:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \boldsymbol{a}_{-1/2} + \boldsymbol{a}_{-1} + \boldsymbol{a}_{-3/2}, \tag{1}$$

where the $p \times 1$ vector $\boldsymbol{\beta}$ has true value $\boldsymbol{\beta}_0$ and $\boldsymbol{a}_{-j/2}$, for $j = 1, 2, 3$, is a term of $O_p(n^{-j/2})$. Then the bias, up to $O(n^{-1})$, and MSE, up to $O(n^{-2})$, denoted by $\boldsymbol{b}$ and $\boldsymbol{M}$, respectively, can be obtained by taking the following term by term expectations, namely,

$$\boldsymbol{b} = \mathrm{E}(\boldsymbol{a}_{-1/2} + \boldsymbol{a}_{-1}) \tag{2}$$

and

$$\begin{aligned}\boldsymbol{M} = \mathrm{E}(&\boldsymbol{a}_{-1/2}\boldsymbol{a}'_{-1/2} + \boldsymbol{a}_{-1/2}\boldsymbol{a}'_{-1} + \boldsymbol{a}_{-1}\boldsymbol{a}'_{-1/2} \\ &+\boldsymbol{a}_{-1/2}\boldsymbol{a}'_{-3/2} + \boldsymbol{a}_{-3/2}\boldsymbol{a}'_{-1/2} + \boldsymbol{a}_{-1}\boldsymbol{a}'_{-1}).\end{aligned} \tag{3}$$

Usually $\mathrm{E}(\boldsymbol{a}_{-1/2}) = \mathbf{0}$ and the $O(n^{-1})$ bias of $\hat{\boldsymbol{\beta}}$ will be $\mathrm{E}(\boldsymbol{a}_{-1})$. Further, the asymptotic MSE (of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$) is given by $\lim_{n\to\infty} n\mathrm{E}(\boldsymbol{a}_{-1/2}\boldsymbol{a}'_{-1/2})$. Thus, one may interpret $\mathrm{E}(\boldsymbol{a}_{-1})$ as the small sample correction in the asymptotic mean of $\hat{\boldsymbol{\beta}}$, which is zero, and $\mathrm{E}(\boldsymbol{a}_{-1/2}\boldsymbol{a}'_{-1} + \boldsymbol{a}_{-1}\boldsymbol{a}'_{-1/2} + \boldsymbol{a}_{-1/2}\boldsymbol{a}'_{-3/2} + \boldsymbol{a}_{-3/2}\boldsymbol{a}'_{-1/2} + \boldsymbol{a}_{-1}\boldsymbol{a}'_{-1})$ is the small sample correction in the MSE. A (feasible) bias adjusted estimator can be given as:

$$\hat{\boldsymbol{\beta}}_{bc} = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{b}} \tag{4}$$

where $\hat{\boldsymbol{b}}$ consistently estimates $\boldsymbol{b}$, usually available by substituting some consistent estimator of $\boldsymbol{\beta}_0$ into the expression of $\boldsymbol{b}$. This bias-corrected estimator $\hat{\boldsymbol{\beta}}_{bc}$ is almost unbiased up to $O(n^{-1})$. Also, see Mariano and Sawa (1972) for an almost unbiased estimator by combining the OLS and 2SLS estimators.

On many occasions, the $\boldsymbol{a}_{-j/2}$'s are in terms of linear and quadratic forms in a random vector. Bao and Ullah (2010) derived expectations of nonnormal quadratic forms, up to order 4, and for normal quadratic forms, a recursion algorithm was proposed. Explicit results for normal quadratic forms, up to order 4, are collected

[1] There is a subtle connection between resampling techniques and higher order expansions, see, for example, the classical work of Hall (1992) and the review by Horowitz (2001). We refrain from discussing the data-driven resampling methods and focus on analytical results in this paper.

as follows:

$$
\begin{aligned}
&\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon}) = \boldsymbol{\mu}'\boldsymbol{A}_1\boldsymbol{\mu} + \mathrm{tr}(\boldsymbol{A}_1),\\
&\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{A}_2\boldsymbol{\varepsilon}) = \mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon})\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{A}_2\boldsymbol{\varepsilon}) + 4\boldsymbol{\mu}'\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{\mu} + 2\mathrm{tr}(\boldsymbol{A}_1\boldsymbol{A}_2),\\
&\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{A}_2\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{A}_3\boldsymbol{\varepsilon}) = \mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon})\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{A}_2\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{A}_3\boldsymbol{\varepsilon})\\
&\quad + [4\boldsymbol{\mu}'\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{\mu} + 2\mathrm{tr}(\boldsymbol{A}_1\boldsymbol{A}_2)]\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{A}_3\boldsymbol{\varepsilon}) + [4\boldsymbol{\mu}'\boldsymbol{A}_1\boldsymbol{A}_3\boldsymbol{\mu} + 2\mathrm{tr}(\boldsymbol{A}_1\boldsymbol{A}_3)]\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{A}_2\boldsymbol{\varepsilon})\\
&\quad + 8[\boldsymbol{\mu}'\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{A}_3\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{A}_1\boldsymbol{A}_3\boldsymbol{A}_2\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{A}_2\boldsymbol{A}_1\boldsymbol{A}_3\boldsymbol{\mu} + \mathrm{tr}(\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{A}_3)],\\
&\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{A}_2\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{A}_3\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{A}_4\boldsymbol{\varepsilon}) = \mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon})\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{A}_2\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{A}_3\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{A}_4\boldsymbol{\varepsilon})\\
&\quad + [4\boldsymbol{\mu}'\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{\mu} + 2\mathrm{tr}(\boldsymbol{A}_1\boldsymbol{A}_2)]\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{A}_3\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{A}_4\boldsymbol{\varepsilon})\\
&\quad + [4\boldsymbol{\mu}'\boldsymbol{A}_1\boldsymbol{A}_3\boldsymbol{\mu} + 2\mathrm{tr}(\boldsymbol{A}_1\boldsymbol{A}_3)]\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{A}_2\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{A}_4\boldsymbol{\varepsilon})\\
&\quad + [4\boldsymbol{\mu}'\boldsymbol{A}_1\boldsymbol{A}_4\boldsymbol{\mu} + 2\mathrm{tr}(\boldsymbol{A}_1\boldsymbol{A}_4)]\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{A}_2\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{A}_3\boldsymbol{\varepsilon})\\
&\quad + 8[\boldsymbol{\mu}'\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{A}_3\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{A}_1\boldsymbol{A}_3\boldsymbol{A}_2\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{A}_2\boldsymbol{A}_1\boldsymbol{A}_3\boldsymbol{\mu} + \mathrm{tr}(\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{A}_3)]\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{A}_4\boldsymbol{\varepsilon})\\
&\quad + 8[\boldsymbol{\mu}'\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{A}_4\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{A}_1\boldsymbol{A}_4\boldsymbol{A}_2\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{A}_2\boldsymbol{A}_1\boldsymbol{A}_4\boldsymbol{\mu} + \mathrm{tr}(\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{A}_4)]\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{A}_3\boldsymbol{\varepsilon})\\
&\quad + 8[\boldsymbol{\mu}'\boldsymbol{A}_1\boldsymbol{A}_3\boldsymbol{A}_4\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{A}_1\boldsymbol{A}_4\boldsymbol{A}_3\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{A}_3\boldsymbol{A}_1\boldsymbol{A}_4\boldsymbol{\mu} + \mathrm{tr}(\boldsymbol{A}_1\boldsymbol{A}_3\boldsymbol{A}_4)]\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{A}_2\boldsymbol{\varepsilon})\\
&\quad + 16[\boldsymbol{\mu}'\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{A}_3\boldsymbol{A}_4\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{A}_4\boldsymbol{A}_3\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{A}_1\boldsymbol{A}_3\boldsymbol{A}_2\boldsymbol{A}_4\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{A}_1\boldsymbol{A}_3\boldsymbol{A}_4\boldsymbol{A}_2\boldsymbol{\mu}\\
&\qquad + \boldsymbol{\mu}'\boldsymbol{A}_1\boldsymbol{A}_4\boldsymbol{A}_2\boldsymbol{A}_3\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{A}_1\boldsymbol{A}_4\boldsymbol{A}_3\boldsymbol{A}_2\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{A}_2\boldsymbol{A}_1\boldsymbol{A}_3\boldsymbol{A}_4\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{A}_3\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{A}_4\boldsymbol{\mu}\\
&\qquad + \boldsymbol{\mu}'\boldsymbol{A}_2\boldsymbol{A}_1\boldsymbol{A}_4\boldsymbol{A}_3\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{A}_4\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{A}_3\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{A}_3\boldsymbol{A}_1\boldsymbol{A}_4\boldsymbol{A}_2\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{A}_4\boldsymbol{A}_1\boldsymbol{A}_3\boldsymbol{A}_2\boldsymbol{\mu}\\
&\qquad + \mathrm{tr}(\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{A}_3\boldsymbol{A}_4) + \mathrm{tr}(\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{A}_4\boldsymbol{A}_3) + \mathrm{tr}(\boldsymbol{A}_1\boldsymbol{A}_3\boldsymbol{A}_2\boldsymbol{A}_4)],
\end{aligned}
\tag{5}
$$

where $\boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{I})$ and $\boldsymbol{A}_i$ are nonstochastic symmetric matrices.[2]

2 The k-Class Estimators

For the explicit expressions of bias and MSE of the k-class estimators, consider a simple triangular structural model

$$\boldsymbol{y}_1 = \boldsymbol{y}_2\beta + \boldsymbol{u}, \quad \boldsymbol{y}_2 = \boldsymbol{Z}_2\boldsymbol{\pi} + \boldsymbol{v}, \tag{6}$$

where $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are $n \times 1$ vectors of observations on the scalar dependent and endogenous variables, respectively, the $n \times K$ matrix $\boldsymbol{Z}_2$ collects observations on $K \geq 1$ instrumental variables, $\boldsymbol{\pi}$ is a $K \times 1$ vector of parameters, and $\boldsymbol{u}$ and $\boldsymbol{v}$ are error vectors. It is assumed that the rows of $[\boldsymbol{u}, \boldsymbol{v}]$ are i.i.d normal with covariance matrix $\boldsymbol{\Sigma} = ((\sigma_u^2, \sigma_{uv})', (\sigma_{uv}, \sigma_v^2)')'$, where $\sigma_{uv} = \rho\sigma_u\sigma_v$. The k-class estimator is given by

$$\hat{\beta}(k) = (\boldsymbol{y}_2'\boldsymbol{N}_k\boldsymbol{y}_2)^{-1}\boldsymbol{y}_2'\boldsymbol{N}_k\boldsymbol{y}_1 = \beta_0 + (\boldsymbol{y}_2'\boldsymbol{N}_k\boldsymbol{y}_2)^{-1}\boldsymbol{y}_2'\boldsymbol{N}_k\boldsymbol{u}, \tag{7}$$

where $\boldsymbol{N}_k = \boldsymbol{P} - k\,\boldsymbol{Q} = \boldsymbol{I} - \kappa\boldsymbol{Q}$, in which $\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{P}$ and $\boldsymbol{P} = \boldsymbol{Z}_2(Z_2'\boldsymbol{Z}_2)^{-1}\boldsymbol{Z}_2'$. (This notation of κ corresponds to the "k" in the original papers of Nagar (1959) and Sawa (1972). Obviously, $\kappa = k + 1$.)

[2] Some terms were missing in the formulae given by Bao and Ullah (2010) and Ullah (2004, page 186). The authors thank Raymond Kan for pointing this out.

Denote $\overline{\boldsymbol{y}}_2 = \boldsymbol{Z}_2\boldsymbol{\pi}$ and $\theta = \overline{\boldsymbol{y}}_2'\overline{\boldsymbol{y}}_2$. Note that $\boldsymbol{y}_2'\boldsymbol{N}_k\boldsymbol{y}_2 = \theta + 2\overline{\boldsymbol{y}}_2'\boldsymbol{v} + \boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{v}$, where $\theta = O(n)$ and $\overline{\boldsymbol{y}}_2'\boldsymbol{v} = O_p(n^{1/2})$. Further, $\boldsymbol{y}_2'\boldsymbol{N}_k\boldsymbol{u} = \overline{\boldsymbol{y}}_2'\boldsymbol{u} + \boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{u}$, where $\overline{\boldsymbol{y}}_2'\boldsymbol{u} = O_p(n^{1/2})$. In light of (5),

$$\begin{aligned}\mathrm{E}(\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{v}) &= \sigma_v^2[K - k(n-K)],\\ \mathrm{Var}(\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{v}) &= 2\sigma_v^4\mathrm{tr}(\boldsymbol{N}_k^2) = 2\sigma_v^4[K + k^2(n-K)].\end{aligned} \tag{8}$$

Let $\boldsymbol{\epsilon} = (\boldsymbol{u}', \boldsymbol{v}')'$. It is obvious that $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma} \otimes \boldsymbol{I})$ and $\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{u} = \boldsymbol{\epsilon}'(\boldsymbol{J} \otimes \boldsymbol{N}_k)\boldsymbol{\epsilon}$, where $\boldsymbol{J} = ((0, 1/2)', (1/2, 0)')'$. Then

$$\begin{aligned}\mathrm{E}(\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{u}) &= \mathrm{tr}[(\boldsymbol{J} \otimes \boldsymbol{N}_k)(\boldsymbol{\Sigma} \otimes \boldsymbol{I})] = \rho\sigma_u\sigma_v[K - k(n-K)],\\ \mathrm{Var}(\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{u}) &= 2\mathrm{tr}[(\boldsymbol{J} \otimes \boldsymbol{N}_k)(\boldsymbol{\Sigma} \otimes \boldsymbol{I})(\boldsymbol{J} \otimes \boldsymbol{N}_k)(\boldsymbol{\Sigma} \otimes \boldsymbol{I})]\\ &= (1+\rho^2)\sigma_u^2\sigma_v^2[K + k^2(n-K)].\end{aligned} \tag{9}$$

It follows that for any finite $k \neq 0$, both $\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{v}$ and $\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{u}$ are $O_p(\max\{n, K\})$. In light of (7), one needs to have the order of $K - k(n-K)$ to be $o(n)$ for a possible estimator of β_0 to be consistent. When $k = -1$, one gets the inconsistent OLS estimator; when $k = 0$, one gets the 2SLS estimator, which is consistent as long as $K = o(n)$. Further, when $k = e$, where e is the smallest eigenvalue of the matrix $\boldsymbol{Y}'\boldsymbol{P}\boldsymbol{Y}(\boldsymbol{Y}'\boldsymbol{Q}\boldsymbol{Y})^{-1}$, in which $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2]$, one gets the LIML estimator; when $k = e - 1/(n-K)$, one gets Fuller's (1971) estimator; when $k = (K-2)/(n-K)$, one gets Nagar's (1959) second-order unbiased estimator; when $k = (K-2)/[(n-(K-2)]$, one gets the modified bias-adjusted 2SLS (B2SLS) estimator by Donald and Newey (2001); when $k = B(\theta/\sigma_v^2, K)(\theta/\sigma_v^2)/(n-K)$, where $B(c, K) = \exp(c/2)\,{}_1F_1(K/2 - 1; K/2; c/2)$ with ${}_1F_1$ denoting the confluent hypergeometric function, one gets the exact finite sample unbiased estimator of Harding et al. (2015).

2.1 Nagar's Bias and MSE Results

When K is finite and $k = O(n^{-1})$, it follows that $\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{u} = O_p(1)$ and $\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{v} = O_p(1)$. The Nagar (1959)-type expansion follows from

$$\begin{aligned}\hat{\beta}(k) - \beta_0 &= \frac{\overline{\boldsymbol{y}}_2'\boldsymbol{u} + \boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{u}}{\theta + 2\overline{\boldsymbol{y}}_2'\boldsymbol{v} + \boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{v}}\\ &= \frac{\overline{\boldsymbol{y}}_2'\boldsymbol{u} + \boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{u}}{\theta}\left(1 + \frac{2\overline{\boldsymbol{y}}_2'\boldsymbol{v} + \boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{v}}{\theta}\right)^{-1}\\ &= \frac{\overline{\boldsymbol{y}}_2'\boldsymbol{u} + \boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{u}}{\theta}\left[1 + \sum_{j=1}^{\infty}(-1)^j\left(\frac{2\overline{\boldsymbol{y}}_2'\boldsymbol{v} + \boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{v}}{\theta}\right)^j\right].\end{aligned} \tag{10}$$

In light of (8) and (9), for finite K and $k = O(n^{-1})$, in the notation of (1),

$$\begin{aligned}a_{-1/2} &= \frac{\overline{\boldsymbol{y}}_2'\boldsymbol{u}}{\theta},\\ a_{-1} &= \frac{\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{u}}{\theta} - \frac{2\boldsymbol{u}'\overline{\boldsymbol{y}}_2\overline{\boldsymbol{y}}_2'\boldsymbol{v}}{\theta^2},\\ a_{-3/2} &= -\frac{\overline{\boldsymbol{y}}_2'\boldsymbol{u}\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{v} + 2\overline{\boldsymbol{y}}_2'\boldsymbol{v}\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{u}}{\theta^2} + \frac{4\overline{\boldsymbol{y}}_2'\boldsymbol{v}\boldsymbol{u}'\overline{\boldsymbol{y}}_2\overline{\boldsymbol{y}}_2'\boldsymbol{v}}{\theta^3}.\end{aligned} \tag{11}$$

One can verify the following results by using (5):

$$\begin{aligned}
&\mathrm{E}(\boldsymbol{u}'\overline{\boldsymbol{y}}_2\overline{\boldsymbol{y}}_2'\boldsymbol{v}) = \rho\sigma_u\sigma_v\theta, \\
&\mathrm{E}[(\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{u})^2] = \rho^2\sigma_u^2\sigma_v^2[K-k(n-K)]^2 + (1+\rho^2)\sigma_u^2\sigma_v^2[K+k^2(n-K)], \\
&\mathrm{E}[(\boldsymbol{u}'\overline{\boldsymbol{y}}_2\overline{\boldsymbol{y}}_2'\boldsymbol{v})^2] = \rho^2\sigma_u^2\sigma_v^2\theta^2 + (1+\rho^2)\sigma_u^2\sigma_v^2\theta^2, \\
&\mathrm{E}(\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{u}\cdot\boldsymbol{u}'\overline{\boldsymbol{y}}_2\overline{\boldsymbol{y}}_2'\boldsymbol{v}) = \rho^2\sigma_u^2\sigma_v^2\theta[K-k(n-K)] + (1+\rho^2)\sigma_u^2\sigma_v^2\theta, \\
&\mathrm{E}(\boldsymbol{u}'\overline{\boldsymbol{y}}_2\overline{\boldsymbol{y}}_2'\boldsymbol{u}\cdot\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{v}) = \sigma_u^2\sigma_v^2\theta[K-k(n-K)] + 2\rho^2\sigma_u^2\sigma_v^2\theta. \qquad (12)
\end{aligned}$$

Therefore, the bias, up to $O(n^{-1})$, in view of (12), is

$$b = \mathrm{E}(a_{-1}) = \frac{\rho\sigma_u\sigma_v[K-2-k(n-K)]}{\theta}. \qquad (13)$$

Setting $k=(K-2)/(n-K)$ gives rise to Nagar's (1959) second-order unbiased estimator. Given that $\mathrm{E}(a_{-1/2}a_{-1})=0$, the MSE result, up to $O(n^{-2})$, is

$$\begin{aligned}
M &= \mathrm{E}(a_{-1/2}^2) + \mathrm{E}(a_{-1}^2) + 2\mathrm{E}(a_{-1/2}a_{-3/2}) \\
&= \frac{\sigma_u^2}{\theta} + \frac{\mathrm{E}[(\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{u})^2]}{\theta^2} + \frac{4\mathrm{E}[(\boldsymbol{u}'\overline{\boldsymbol{y}}_2\overline{\boldsymbol{y}}_2'\boldsymbol{v})^2]}{\theta^4} - \frac{4\mathrm{E}(\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{u}\cdot\boldsymbol{u}'\overline{\boldsymbol{y}}_2\overline{\boldsymbol{y}}_2'\boldsymbol{v})}{\theta^3} \\
&\quad - \frac{2\mathrm{E}(\boldsymbol{u}'\overline{\boldsymbol{y}}_2\overline{\boldsymbol{y}}_2'\boldsymbol{u}\cdot\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{v}) + 4\mathrm{E}(\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{u}\cdot\boldsymbol{u}'\overline{\boldsymbol{y}}_2\overline{\boldsymbol{y}}_2'\boldsymbol{v})}{\theta^3} + \frac{8\mathrm{E}[(\boldsymbol{u}'\overline{\boldsymbol{y}}_2\overline{\boldsymbol{y}}_2'\boldsymbol{v})^2]}{\theta^4} \\
&= \frac{\sigma_u^2}{\theta} + \frac{K^2(k+1)^2\rho^2\sigma_u^2\sigma_v^2}{\theta^2} - \frac{K(k+1)\sigma_u^2\sigma_v^2[k(2n+1)\rho^2+k+7\rho^2+1]}{\theta^2} \\
&\quad + \frac{\sigma_u^2\sigma_v^2\{k^2n[(n+1)\rho^2+1]+2kn(4\rho^2+1)+12\rho^2+4\}}{\theta^2}. \qquad (14)
\end{aligned}$$

2.2 The Double k-Class Estimators

Sawa (1972) showed that, under normality, for $k \le 0$, the k-class estimator has moments while for $k > 0$ the first moment may not even exist.[3] Thus, the exact moments of both 2SLS and OLS estimators exist under certain conditions but the LIML, Nagar, and the exact finite sample unbiased estimators all have k greater than 0 and their moments may not exist. This explains, due to lack of existence of moments, these biased adjusted estimators tend to fail in practice, especially under weak instruments and/or many instruments. However, Fuller (1977) indicated that for $k > 0$ but less than the k assigned by LIML or Nagar, the moments of $\hat{\beta}(k)$ do exist, resulting in the Fuller-type estimator being a useful solution to the problem of weak and/or large instruments. In view of these, Harding et al. (2015) looked into detail of the question as to how useful the k-class estimators are given the trade-off between bias reduction and existence of moments especially for many or weak instrument cases. Their findings suggest a failure of good behavior of the k-class estimators, leading them to consider the double-k-class estimator of Nagar(1962):

$$\hat{\beta}(k_1,k_2) = (\boldsymbol{y}_2'\boldsymbol{N}_{k_1}\boldsymbol{y}_2)^{-1}\boldsymbol{y}_2'\boldsymbol{N}_{k_2}\boldsymbol{y}_1. \qquad (15)$$

[3] When $K=1$, i.e., the just identified case, the 2SLS estimator (corresponding to $k=0$), does not have even the first moment, see Sawa (1969). For a definitive treatment of the issue of existence of moments, see Kinal (1980).

Dwivedi and Srivastava (1984) derived the exact bias and MSE results of this estimator. Based on their results, Harding et al. (2015) provided a choice of k_2 under which the double-k-class estimator possesses much smaller MSE than the 2SLS and Fuller estimators when there are many instruments.

In fact, from

$$\hat{\beta}(k_1,k_2)-\beta_0=\frac{\boldsymbol{y}_2'\boldsymbol{N}_{k_2}\boldsymbol{u}}{\boldsymbol{y}_2'\boldsymbol{N}_{k_1}\boldsymbol{y}_2}=\frac{\begin{pmatrix}\boldsymbol{y}_2\\ \boldsymbol{u}\end{pmatrix}'(\boldsymbol{J}\otimes\boldsymbol{N}_{k_1})\begin{pmatrix}\boldsymbol{y}_2\\ \boldsymbol{u}\end{pmatrix}}{\begin{pmatrix}\boldsymbol{y}_2\\ \boldsymbol{u}\end{pmatrix}'(\boldsymbol{J}_1\otimes\boldsymbol{N}_{k_2})\begin{pmatrix}\boldsymbol{y}_2\\ \boldsymbol{u}\end{pmatrix}}=\frac{\boldsymbol{\alpha}'\boldsymbol{A}\boldsymbol{\alpha}}{\boldsymbol{\alpha}'\boldsymbol{B}\boldsymbol{\alpha}}, \tag{16}$$

where $\boldsymbol{J}_1=((1,0)',(0,0)')'$, $\boldsymbol{\alpha}=(\boldsymbol{\Sigma}\otimes\boldsymbol{I})^{-1/2}(\boldsymbol{y}_2',\boldsymbol{u}')'\sim\mathrm{N}((\overline{\boldsymbol{y}}_2',\boldsymbol{0}')',\boldsymbol{I})$, $\boldsymbol{A}=(\boldsymbol{\Sigma}\otimes\boldsymbol{I})^{1/2}(\boldsymbol{J}\otimes\boldsymbol{N}_{k_1})(\boldsymbol{\Sigma}\otimes\boldsymbol{I})^{1/2}=\boldsymbol{\Sigma}\boldsymbol{J}\otimes\boldsymbol{N}_{k_1}$, $\boldsymbol{B}=(\boldsymbol{\Sigma}\otimes\boldsymbol{I})^{1/2}(\boldsymbol{J}_1\otimes\boldsymbol{N}_{k_2})(\boldsymbol{\Sigma}\otimes\boldsymbol{I})^{1/2}=\boldsymbol{\Sigma}\boldsymbol{J}_1\otimes\boldsymbol{N}_{k_2}$, one can evaluate the exact moments of $\hat{\beta}(k_1,k_2)$ (as well as those of the k-class estimators), when existing, using the fast algorithm in Bao and Kan (2013).[4]

2.3 Many Instruments and Weak Instruments

When K is allowed to increase with the sample size (but at a slower rate), namely, the so-called many instruments case, the $a_{-j/2}$'s in the Nagar-type expansion need to be modified, depending on how fast K diverges, relative to n. Donald and Newey (2001) showed that the higher-order MSE results are

$$\begin{aligned}M_{2SLS}&=\frac{\sigma_u^2}{\theta}+\frac{K^2\rho^2\sigma_u^2\sigma_v^2}{\theta^2},\quad K\to\infty\text{ and }K^2/n\to 0,\\ M_{LIML}&=\frac{\sigma_u^2}{\theta}+\frac{K\sigma_u^2\sigma_v^2(1-\rho^2)}{\theta^2},\quad K\to\infty\text{ and }K/n\to 0,\\ M_{B2SL}&=\frac{\sigma_u^2}{\theta}+\frac{K\sigma_u^2\sigma_v^2(1+\rho^2)}{\theta^2}\quad K\to\infty\text{ and }K/n\to 0.\end{aligned} \tag{17}$$

Recall that $\theta=\overline{\boldsymbol{y}}_2'\overline{\boldsymbol{y}}_2=\boldsymbol{\pi}'\boldsymbol{Z}_2'\boldsymbol{Z}_2\boldsymbol{\pi}=O(n)$. When K is allowed to increase with the sample size, the above MSEs here are not up to the typical second order, namely, $O(n^{-2})$. They are of higher order in the sense that the remainder terms in the relevant expansions are of smaller order relative to the second terms in the above expressions. Notably, the first term, when scaled by n, in each case is the same and it represents the asymptotic variance of $\sqrt{n}[\hat{\beta}(k)-\beta_0]$. The finite sample corrections are given by the second terms. The leading term in the higher-order bias of the 2SLS estimator, when $K\to\infty$, is $K\sigma_{uv}/\theta=K\rho\sigma_u\sigma_v/\theta$, thus the second term of M_{2SLS} is the squared bias, arising because of the higher order bias of the 2SLS estimator. The second terms for the LIML and B2SLS estimators represent the higher-order variances. Further their MSE expressions clearly indicate that they should be preferred, for any $\rho\neq 0$, to the 2SLS estimators in the lower MSE sense. However, it is well known that the LIML, B2SLS, and Nagar estimators may not possess moments, thus have thicker tails, but that is not the case for

[4] Notably, Dwivedi and Srivastava (1984) presented the first and second moments in terms of double infinite series, whereas Bao and Kan (2013) presented, for any finite integer j, the j-th moment using a single infinite series.

the 2SLS estimator, see, for example, Mariano and Sawa (1972) and Sawa (1972). Disappointing behaviors, in terms of the dispersion, of the LIML and B2SLS estimators have been documented in the literature, especially when identification of the model is weak. However, the 2SLS estimator may have higher MSE due to its bad bias behavior. In view of this, Hahn et al. (2004) introduced the jackknife IV estimator (JIVE), aiming to improve the 2SLS estimator by reducing its finite-sample bias through the jackknife, and they showed that its higher order MSE is the same as that of the B2SLS estimator.

Noticeably,

$$\sqrt{\theta/\sigma_v^2}[\hat{\beta}(k)-\beta_0]=\frac{\sigma_u}{\sigma_v}\frac{\frac{\overline{\boldsymbol{y}}_2'\boldsymbol{u}}{\sigma_u\sqrt{\theta}}+\frac{\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{u}}{\sigma_v\sigma_u}\frac{\sigma_v}{\sqrt{\theta}}}{1+\frac{2\overline{\boldsymbol{y}}_2'\boldsymbol{v}}{\sigma_v\sqrt{\theta}}\frac{\sigma_v}{\sqrt{\theta}}+\frac{\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{v}}{\sigma_v^2}\frac{\sigma_v^2}{\theta}}, \tag{18}$$

where, under normality, $\overline{\boldsymbol{y}}_2'\boldsymbol{u}/(\sigma_u\sqrt{\theta})$ and $\overline{\boldsymbol{y}}_2'\boldsymbol{v}/(\sigma_v\sqrt{\theta})$ are two standard normal random variables, $\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{v}/\sigma_v^2=\boldsymbol{v}'\boldsymbol{v}/\sigma_v^2-k\boldsymbol{v}'\boldsymbol{Q}\boldsymbol{v}/\sigma_v^2$ is a linear combination of two chi-squared random variables, and $\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{u}/(\sigma_v\sigma_u)=\boldsymbol{\epsilon}'(\boldsymbol{J}\otimes\boldsymbol{N}_k)\boldsymbol{\epsilon}/(\sigma_v\sigma_u)$ follows a linear combination (with weights related to the eigenvalues of $\boldsymbol{\Sigma}^{1/2}\boldsymbol{J}\boldsymbol{\Sigma}^{1/2}\otimes\boldsymbol{N}_k$) of chi-squared random variables. These well-defined distributions are independent of the sample size n. So θ/σ_v^2, also termed as the concentration parameter, plays the role of the sample size in determining the sampling distribution of $\hat{\beta}(k)$. From $\theta/\sigma_v^2\approx(n-K)R^2/(1-R^2)\approx KF$, where $R^2$and F are the first-stage (regression of $\boldsymbol{y}_2$ on $\boldsymbol{Z}_2$) R-squared and F-statistic, one sees that the sampling distribution of $\hat{\beta}(k)$ crucially depends on the number and strength of instruments used.

Bekker (1994) considered a different type of asymptotics such that as $n\to\infty$, the ratio of the number of instruments to n is of the order no greater than the observed sample ratio. Specifically, as $n\to\infty$, $K/n=\alpha+o(n^{-1/2})$, $0\le\alpha=(K_0-1)/(n_0-1)\le K_0/n_0$, where K_0 and n_0 denote the sample values of K and n, respectively. (Additionally, it is assumed that $n^{-1}\theta$ is fixed at the sample value $n_0^{-1}\theta$.) Under this asymptotic regime,

$$\begin{aligned}\text{plim}(\hat{\beta}_{OLS})&=\beta_0+\frac{\rho\sigma_u\sigma_v}{n_0^{-1}\theta},\\ \text{plim}(\hat{\beta}_{2SLS})&=\beta_0+\frac{\alpha\rho\sigma_u\sigma_v}{n_0^{-1}\theta+\alpha\sigma_v^2}.\end{aligned} \tag{19}$$

Under the traditional large-sample asymptotic regime, namely, $\alpha=0$ (K is fixed at K_0 and $n\to\infty$), the 2SLS estimator is consistent and the OLS estimator is inconsistent. Yet, under Bekker's regime ($\alpha\neq0$), $\hat{\beta}_{2SLS}$ is inconsistent (recall that for a possible consistent k-class estimator of β_0, one needs to have the order of $K-k(n-K)$ to be $o(n)$) and its asymptotic bias, relative to that of the OLS estimator, is related to the quality of the instruments:

$$\left[\frac{\text{plim}(\hat{\beta}_{2SLS})-\beta_0}{\text{plim}(\hat{\beta}_{OLS})-\beta_0}\right]^2\le\frac{\alpha^2}{q^2}, \tag{20}$$

where $q=\text{plim}\hat{q}$, $\hat{q}=\min_{\boldsymbol{l}\in\mathbb{R}^n}(\boldsymbol{l}'\boldsymbol{x}'\boldsymbol{P}\boldsymbol{x}\boldsymbol{l}/\boldsymbol{l}'\boldsymbol{x}'\boldsymbol{x}\boldsymbol{l})$, and $\hat{q}$ is considered a measure of the quality of the instruments. For $\alpha\neq0$ (such that $K=O(\alpha n)+o(n^{1/2})$) and $k=O(n^{-1})$, one can see that $\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{v}=O_p(\alpha n)$ and $\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{u}=O_p(\alpha n)$. Then it follows from $\hat{\beta}(k)-\beta_0=(\overline{\boldsymbol{y}}_2'\boldsymbol{u}+\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{u})/(\theta+2\overline{\boldsymbol{y}}_2'\boldsymbol{v}+\boldsymbol{v}'\boldsymbol{N}_k\boldsymbol{v})$ that no k-class estimator can be consistent under Bekker's (1994) asymptotics. Recognizing this, Bekker (1994) proposed his method of moments (MM) estimator that is consistent under both the traditional and his asymptotic regimes. It would be interesting to investigate the higher-order properties of his MM estimator.

2.4 Results Evolved over the Years

Fisher (1921, 1922, 1928, 1935), about a century ago, laid the foundation of statistical finite sample theory exploring distributions of various sample based statistics. It was brought into econometrics by the seminal works of Haavelmo (1947), Anderson and Rubin (1949), and Hurwicz(1950), among others, who analyzed the finite sample properties of various econometric statistics. Most of these works were related to the exact bias or distribution by using different techniques. However, the pioneering work of Nagar (1959) established the foundation for developing analytical approximate finite sample econometrics and its future developments. Although his results were on the approximate bias and MSE of the k-class estimators, the technique he proposed was easier to implement for a wide variety of econometric statistics which are nonlinear in stochastic variables. Essentially, it involved a Taylor series expansion in a decreasing order of random terms, and then taking their expectations one by one, see (1). Soon, Nagar and his students applied this technique to derive small sample approximate bias and MSE of a large number of estimators in different econometric models. For example, see Kakwani (1967), Nagar and Kakwani (1965), Nagar and Gupta (1968), Gupta and Ullah (1970), Kakwani (1971), and Nagar and Ullah (1973), among others. Since then, a large number of papers have appeared on the approximate results, see Kadane (1971), Maasoumi and Phillips (1982), Rothenberg (1984a, 1984b), Srivastava and Giles (1987), Kiviet and Phillips (1993), Srivastava and Maekawa (1995), Donald and Newey (2001), Hahn and Newey (2004), and others. However, Srinivasan (1970) raised some concerns and cautions on using these approximate moments for the unknown exact moments, which were often difficult to derive. This is because, as he emphasized, for some estimators the exact moments may not exist or would be infinite, but the Nagar's approximate moment results would be finite. Thus, he raised the question of the meaning of Nagar's approximate results without first checking the existence of exact moments. This has lead the finite sample econometrics into two different directions. One is to look into the validity of Nagar's expansion, see important contributions on this by Sargan (1975, 1976, 1980) and Phillips (1977, 1978, 1980). In these studies they discussed the conditions of validity for Nagar's expansion and its link to the theory and applications of Edgeworth's expansion (1896, 1905) to derive the approximate moments and distributions of econometric statistics. The second direction took place in terms of deriving exact moments and distributions of econometric statistics and important contributions in this direction include Basmann (1961), Kabe (1963, 1964), Richardson (1968), Anderson and Sawa (1973), Ullah and Nagar (1974), Maasoumi (1978), Kinal (1980), Mariano (1982), Phillips (1980, 1983). Rothenberg (1984a, 1984b), Taylor (1983), and Hillier et al. (1984). A summary of these extensive exact and approximate results above can be found in the book by Ullah (2004), where it is also shown that some of these results can be developed by studying the distribution and moments of quadratic forms as developed in Imhof (1961), Forchini (2002), Bao and Ullah

(2010), and Bao and Kan (2013). The applications of these theoretical results are explored in many practical applications such as Bao and Ullah (2007), Bao et al. (2019), Bao et al. (2021), and Chu et al. (2021).

The major breakthrough in the applications of Nagar's approximate results for the 2SLS and OLS estimators came in the context of growing literature on weak instruments and large number of instruments, see Section 2.3 and Angrist and Pischke (2009) for references. While many of the Nagar-type approximate results are derived for the estimators with multivariate stochastic variables, the exact results have been restricted to estimators in bivariate cases. However, this gap is likely to be filled in future with the development of the general recent exact results and tools by Hillier and Kan (this issue, 2021), Hillier et al. (2009, 2014). These papers will be fundamental for the future growth of exact results for econometric statistics in multivariate stochastic cases.

3 Generalization of Nagar's Expansion

The expansion (1) and various results based on it originated in linear simultaneous models with i.i.d. normal data. For example, the expansion in (10), the corresponding stochastic terms (11), and the bias and MSE results (13) and (14). Rilstone et al. (1996) generalized it to a class of extremum estimators in linear and nonlinear models with possibly nonnormally distributed data. The estimator in question is defined as

$$\hat{\boldsymbol{\beta}} = \arg_{\boldsymbol{\beta}} \psi_n(\boldsymbol{\beta}) = \arg_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n} q_i(\boldsymbol{\beta}) = 0, \tag{21}$$

where $q_i(\boldsymbol{\beta}) = q(\boldsymbol{\beta}, \boldsymbol{w}_i)$ is a known $p \times 1$ vector-valued function of the i.i.d. random vectors $\boldsymbol{w}_i$, consisting of observations on variables in the system, and the parameter vector $\boldsymbol{\beta}$ such that $\mathrm{E}(\psi_n(\boldsymbol{\beta})) = \mathbf{0}$ (conditional on the stochastic regressors, if any) only at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. One can think of $\psi_n(\boldsymbol{\beta})$ as the orthogonality condition between the regressors and the error terms, or the first-order condition of some optimization criterion. The class of estimators identified by (21) include many estimators in linear and nonlinear models, such as the maximum likelihood (ML), least squares (LS), and generalized method of moments (GMM) estimators. Under some smoothness conditions on $\psi_n(\boldsymbol{\beta})$ and its higher-order derivatives, one can implement a Taylor-series expansion such that

$$\psi_n(\hat{\boldsymbol{\beta}}) = \psi_n(\boldsymbol{\beta}_0) + \sum_{i=1}^{3} \frac{1}{i!} \nabla^i \psi_n(\boldsymbol{\beta}_0)[\otimes^i(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)] + o(n^{-3/2}), \tag{22}$$

where $\nabla^i \psi_n(\boldsymbol{\beta}_0)$ denotes the i-th order derivative of $\psi_n(\boldsymbol{\beta})$ with respect to the parameter vector $\boldsymbol{\beta}$, evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and defined recursively as in Rilstone et al. (1996), and $\otimes^i(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \otimes (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \cdots \otimes (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is the Kronecker product of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ and itself i times. Given the expansion (22), by solving for $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ recursively and expanding $[\nabla \psi_n(\boldsymbol{\beta}_0)]^{-1} = (\mathrm{E}[\nabla \psi_n(\boldsymbol{\beta}_0)])^{-1}\{\boldsymbol{I} + (\mathrm{E}[\nabla \psi_n(\boldsymbol{\beta}_0)])^{-1}(\nabla \psi_n(\boldsymbol{\beta}_0) - \mathrm{E}[\nabla \psi_n(\boldsymbol{\beta}_0)])\}^{-1}$ in terms of powers of $\nabla \psi_n(\boldsymbol{\beta}_0) - \mathrm{E}[\nabla \psi_n(\boldsymbol{\beta}_0)]$, one can write down $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ as in (1). One can see that the expansion (1) used for the k-class estimator is a special case of the expansion of $[\nabla \psi_n(\boldsymbol{\beta}_0)]^{-1}$. (It is not necessarily $\mathrm{E}(\psi_n(\boldsymbol{\beta}_0)) = \mathbf{0}$ though for the k-class estimator, except for the special case of 2SLS.)

Bao and Ullah (2017a, 2017b) recognized that Rilstone et al. (1996) can be further generalized to models for non-i.i.d. data, including time-series and spatial data, where the identification condition $\psi_n(\boldsymbol{\beta}) = \psi_n(\boldsymbol{\beta}, \boldsymbol{W})$, $\boldsymbol{W} = (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)'$. Though not obvious, on many occasions, one may write $\psi_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} q(\boldsymbol{\beta}, \boldsymbol{w}_i)$ such that $\mathrm{E}(q(\boldsymbol{\beta}, \boldsymbol{w}_i)|\sigma(\boldsymbol{w}_{i-1}, \cdots, \boldsymbol{w}_1)) = \boldsymbol{0}$, where $\sigma(\boldsymbol{w}_{i-1}, \cdots, \boldsymbol{w}_1)$ is the sigma-field generated by $\boldsymbol{w}_{i-1}, \cdots, \boldsymbol{w}_1$. In other words, $\{q(\boldsymbol{\beta}, \boldsymbol{w}_i)\}_{i=1}^{n}$ forms a martingale difference sequence. Based on this, Bao and Ullah (2017a) derived the second-order bias and MSE of the ML estimator in pure first-order spatial autoregressions under normality and Bao and Ullah (2017b) presented results for some commonly-used time-series models under nonnormality. Results for the first-order spatial autoregressions with exogenous regressors under nonnormality were given by Bao (2013). In fact, Bao and Ullah (2009) used the expansion (10) based on the general identification condition $\psi_n(\boldsymbol{\beta}, \boldsymbol{W})$ to derive the the approximate standardized measures of deviation from normality, namely, the skewness and excess kurtosis coefficients, for a class of econometric estimators. Given the skewness and kurtosis results above, one may follow the lines of Rothenberg (1984a) to use the two standardized measures to construct an Edgeworth-type approximation to the distribution of a non-linear estimator, though it is an open question as to whether the Edgeworth distribution is a valid approximation to the true distribution.

When one tries to evaluate the various terms involved in (2) and (3), the results on moments of quadratic forms in (5) as well as those under nonnormality in Bao and Ullah (2010) are useful for one to simplify and derive analytical results.

4 The Exact and Approximate Distribution Results

In addition to the approximate moments, one may be also interested in the exact and (higher-order) approximate distributions of the estimator in question. This is relevant since it has long been observed that the celebrated asymptotic normal distribution, with the help of various version of central limit theorems, depending on the degree of memory and heterogeneity in the data, may provide poor approximation to the estimator's finite-sample distribution. It may also happen that the asymptotic distribution depends on the asymptotic regime used. For example, for the continuous time Ornstein-Uhlenbeck process, the asymptotic distribution of the MLE of the mean reversion parameter depends on how the data are sampled, see Zhou and Yu (2015) for the different asymptotic distribution results under infill, expanding, and double asymptotic regimes. In practice, the realized sample data does not manifest itself which asymptotic regime should be used. This calls for the exact and/or higher-order approximate distributions.

4.1 The Exact Distribution

When the (scalar) estimator can be written as a ratio of quadratic forms, then under normality, the exact distribution of the estimator can be evaluated using the technique of Imhof (1960). Consider the following ratio

$$r = \frac{\boldsymbol{y}'\boldsymbol{A}_1\boldsymbol{y}}{\boldsymbol{y}'\boldsymbol{A}_2\boldsymbol{y}}, \tag{23}$$

where $\boldsymbol{y}$ is an $n \times 1$ normal random vector with its mean vector $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{\mu}$ and variance matrix $\mathrm{Var}(\boldsymbol{y}) = \boldsymbol{\Omega}$ being positive definite, $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are $n \times n$ non-stochastic symmetric matrices, and $\boldsymbol{A}_2$ is a positive semi-definite. The cumulative distribution function (CDF) of this ratio is

$$F(r_0) = \Pr(r \leq r_0) = \Pr(\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y} \leq 0), \tag{24}$$

where $\boldsymbol{A} = \boldsymbol{A}_1 - r_0\boldsymbol{A}_2$. Rewriting $\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{\Omega}^{-1/2}\boldsymbol{B}\boldsymbol{B}'\boldsymbol{\Omega}^{1/2}\boldsymbol{A}\boldsymbol{\Omega}^{1/2}\boldsymbol{B}\boldsymbol{B}'\boldsymbol{\Omega}^{-1/2}\boldsymbol{y} \equiv \boldsymbol{y}^{\star\prime}\boldsymbol{\Lambda}\boldsymbol{y}^{\star} = \sum_{i=1}^{n}\lambda_i y_i^{\star 2}$, where $\boldsymbol{y}^{\star} = \boldsymbol{B}'\boldsymbol{\Omega}^{-1/2}\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{\mu}^{\star}, \boldsymbol{I})$, $\boldsymbol{\mu}^{\star} = \boldsymbol{B}'\boldsymbol{\Omega}^{-1/2}\boldsymbol{\mu}$, $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues of $\boldsymbol{\Omega}^{1/2}\boldsymbol{A}\boldsymbol{\Omega}^{1/2}$, and $\boldsymbol{B}$ is an orthogonal matrix of eigenvectors of $\boldsymbol{\Omega}^{1/2}\boldsymbol{A}\boldsymbol{\Omega}^{1/2}$ such that $\boldsymbol{B}'\boldsymbol{\Omega}^{1/2}\boldsymbol{A}\boldsymbol{\Omega}^{1/2}\boldsymbol{B} = \boldsymbol{\Lambda}$, one can see that the distribution of the ratio of quadratic forms translates to that of a linear combination of independent non-central chi-squared random variables. Let λ_j, $j = 1, \cdots, s \leq n$, denote non-zero distinct elements of $\boldsymbol{\Lambda}$, n_j be the corresponding multiplicities, $\delta_j = \sum_{i \to j}\mu_i^{\star 2}$, where $\sum_{i \to j}$ denotes summing over i such that the i-th element of $\boldsymbol{\Lambda}$ equals λ_j. Then $\boldsymbol{y}^{\star\prime}\boldsymbol{\Lambda}\boldsymbol{y}^{\star} = \sum_{j=1}^{s}\lambda_j\zeta_j^2$, where $\zeta_j^2 \sim \chi^2_{n_j}(\delta_j)$ and they are independent of each other. For a linear combination (with weights λ_j) of independent noncentral chi-squared variables ζ_j (with noncentrality parameter δ_j and degrees of freedom n_j), Imhof (1961) showed that

$$\Pr(\sum_{j=1}^{s}\lambda_j\zeta_j^2 \leq c) = \frac{1}{2} - \frac{1}{\pi}\int_0^{\infty}\frac{\sin\rho_1(t)}{t\rho_2(t)}\mathrm{d}t, \tag{25}$$

where

$$\rho_1(t) = -\frac{ct}{2} + \sum_{j=1}^{s}\left[\frac{n_j}{2}\tan^{-1}(\lambda_j t) + \frac{\lambda_j\delta_j t}{2(1+\lambda_j^2 t^2)}\right],$$
$$\rho_2(t) = \prod_{j=1}^{s}(1+\lambda_j^2t^2)^{n_j/4}\exp\left[\frac{\lambda_j^2 t^2\delta_j}{2(1+\lambda_j^2t^2)}\right]. \tag{26}$$

Setting $c = 0$ in (25), one has $\Pr(\sum_{j=1}^{r}\lambda_j\zeta_j^2 \leq 0) = F(r_0) = \Pr(\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y} \leq 0) = \Pr(\boldsymbol{y}'\boldsymbol{A}_1\boldsymbol{y}/\boldsymbol{y}'\boldsymbol{A}_2\boldsymbol{y} \leq r_0)$.

For one to be able to use (25), the essential task is to compute eigenvalues of an $n \times n$ matrix, which may become very cumbersome if n is moderately large. A somewhat different approach is given by Gurland (1948) and Gil-Pelaez (1951). Suppose one can work out the joint characteristic function of Y_1 and Y_2, where $Y_1 = \boldsymbol{y}'\boldsymbol{A}_1\boldsymbol{y}$ and $Y_2 = \boldsymbol{y}'\boldsymbol{A}_2\boldsymbol{y}$, denoted by $\varphi(t_1, t_2) = \mathrm{E}(\exp(\mathrm{i}t_1Y_1 + \mathrm{i}t_2Y_2))$, then

$$F(r_0) = \frac{1}{2} - \frac{1}{\pi}\int_0^{\infty}\mathrm{Im}\left(\frac{\varphi(t_1, -t_1r_0)}{t_1}\right)\mathrm{d}t_1. \tag{27}$$

Bao et al. (2019) used (27) to evaluate the exact distribution of the MLE of the mean reversion parameter in the Ornstein-Uhlenbeck process with the help of analytical evaluation of the joint characteristic function involved.

When the estimator is multivariate, the issue of deriving the exact distribution is more complicated. A classical example is Phillips (1980) when he derived the exact probability distribution function (PDF) of the IV estimator of the coefficient vector of the endogenous variables in a structural equation under over-identification. Let the structural equation be

$$\boldsymbol{y}_1 = \boldsymbol{Y}_2\boldsymbol{\beta} + \boldsymbol{Z}_1\gamma + \boldsymbol{u} \tag{28}$$

and the reduced form be

$$[\boldsymbol{y}_1, \boldsymbol{Y}_2] = [\boldsymbol{Z}_1, \boldsymbol{Z}_2]\begin{bmatrix}\boldsymbol{\pi}_{11} & \boldsymbol{\Pi}_{12} \\ \boldsymbol{\pi}_{21} & \boldsymbol{\Pi}_{22}\end{bmatrix} + [\boldsymbol{v}_1, \boldsymbol{V}_2] = \boldsymbol{Z\Pi} + \boldsymbol{V}, \tag{29}$$

where $\boldsymbol{y}_1$ is an $n \times 1$ observational vector on the outcome variable, $\boldsymbol{Y}_2$ is an $n \times g$ observational matrix on g endogenous variables, $\boldsymbol{Z}_1$ is an $n \times K_1$ observational matrix of included exogenous regressors, and $\boldsymbol{Z}_2$ is an $n \times K_2$ observational matrix of excluded exogenous regressors. Suppose $[\boldsymbol{Z}_1, \boldsymbol{Z}_3]$ be an observational matrix of instruments, where the $n \times K_3$ matrix $\boldsymbol{Z}_3$ is a submatrix of $\boldsymbol{Z}_2$. The IV estimator is

$$\hat{\boldsymbol{\beta}}_{IV} = (\boldsymbol{Y}_2'\boldsymbol{Z}_3\boldsymbol{Z}_3'\boldsymbol{Y}_2)^{-1}\boldsymbol{Y}_2'\boldsymbol{Z}_3\boldsymbol{Z}_3'\boldsymbol{y}_1 = \boldsymbol{H}_{22}^{-1}\boldsymbol{h}_{21}, \tag{30}$$

where

$$\boldsymbol{H} = \begin{bmatrix}\boldsymbol{h}_{11} & \boldsymbol{h}_{21}' \\ \boldsymbol{h}_{21} & \boldsymbol{H}_{22}\end{bmatrix} = n^{-1}\begin{bmatrix}\boldsymbol{y}_1'\boldsymbol{Z}_3\boldsymbol{Z}_3'\boldsymbol{y}_1 & \boldsymbol{y}_1'\boldsymbol{Z}_3\boldsymbol{Z}_3'\boldsymbol{Y}_2 \\ \boldsymbol{Y}_2'\boldsymbol{Z}_3\boldsymbol{Z}_3'\boldsymbol{y}_1 & \boldsymbol{Y}_2'\boldsymbol{Z}_3\boldsymbol{Z}_3'\boldsymbol{Y}_2\end{bmatrix}. \tag{31}$$

With the normalization $n^{-1}\boldsymbol{Z}'\boldsymbol{Z} = \boldsymbol{I}$ and the covariance matrix of rows of $\boldsymbol{V}$ being the identity matrix, the matrix $\boldsymbol{H}$ is noncentral Wishart of order $g+1$. Phillips (1980) derived the joint PDF of $\hat{\boldsymbol{\beta}}_{IV}$ by the following steps. First, write the PDF of $\boldsymbol{A}$ in terms of a hypergeometric function with a matrix argument. In the next step, a variable transformation is carried out so that the PDF is expressed in terms of $\boldsymbol{H}_{22}$ and $\boldsymbol{h}_{21}$. In the final step, the PDF of $\hat{\boldsymbol{\beta}}_{IV}$ is derived by integrating over the space of $(\boldsymbol{h}_{21}, \boldsymbol{H}_{22})$. The final expression of the PDF of $\hat{\boldsymbol{\beta}}_{IV}$ involves an infinite sum of hypergeometric functions with a matrix argument. For the 2SLS estimator in the simple case (6) (with $g = 1$, $K_3 = K_2 = K$, and $K_1 = 0$), the PDF of $\hat{\beta}_{2SLS}$ can be written as

$$f(r_0) = \frac{\exp\left[-\frac{\theta}{2}(1+\beta_0^2)\right]\Gamma\left(\frac{K+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{K}{2}\right)(1+r_0^2)^{(K+1)/2}} \times \sum_{j=0}^{\infty}\frac{\left(\frac{K-1}{2}\right)_j}{j!\left(\frac{K}{2}\right)_j}\left(\frac{\theta\beta_0^2}{2}\right)^j {}_1F_1\left(\frac{K+1}{2}; \frac{K}{2}+j; \frac{\theta}{2}\frac{(1+\beta_0 r_0)^2}{1+r_0^2}\right). \tag{32}$$

4.2 The Approximate Distribution

When $g = 1$, namely, when there is only one right-hand-side endogenous variable, the matrix argument hypergeometric function has an explicit representation (in terms of the univariate ${}_1F_1$ function and the elementary symmetric functions of the matrix argument). For the general case of $g > 1$, its calculation is a challenge and hence the exact PDF of $\hat{\boldsymbol{\beta}}_{IV}$ is not feasible. Phillips (1980) recognized this and proposed approximating the matrix argument hypergeometric function such that approximation error is $O(n^{-1})$. The approximation error of the resulting PDF, however, is not clear.

Instead, one may consider developing approximate distribution results using an Edgeworth-type approximation. A multivariate Edgeworth approximation is typically complicated, if not impossible. Classical references and recent contributions include Bhattacharya and Ghosh (1978), Götze (1987), Kollo and von Rosen (1998), and Kundhi and Rilstone (2020), among others. Usually, for tractable results, one needs to employ complicated matrix and tensor notation. Skovgaard (1986) proposed a directional approach such that the expansion is implemented in one direction at a time. Kundhi and Rilstone (2013) considered a scalar statistic of $\hat{\boldsymbol{\beta}}$ that is based on (21), also see Kundhi and Rilstone (2012) for the special case when $\hat{\boldsymbol{\beta}}$ is the generalized empirical likelihood (GEL) estimator.

Suppose one is interested in the distribution of the scalar statistic

$$\hat{t} = \frac{\sqrt{n}\boldsymbol{l}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{\sqrt{\boldsymbol{l}'\hat{\boldsymbol{\mathcal{V}}}\boldsymbol{l}}}, \tag{33}$$

where $\boldsymbol{l} \in \mathbb{R}^p$ is a $p \times 1$ vector of constants and $\hat{\boldsymbol{\mathcal{V}}}$ is a consistent estimator of $\boldsymbol{\mathcal{V}} = n\mathrm{E}(\boldsymbol{a}_{-1/2}\boldsymbol{a}'_{-1/2})$. In addition to the smoothness assumptions in Rilstone et al. (1996), if one also assumes that an expansion of $\boldsymbol{l}'\hat{\boldsymbol{\mathcal{V}}}\boldsymbol{l}$ holds, namely,

$$\boldsymbol{l}'\hat{\boldsymbol{\mathcal{V}}}\boldsymbol{l} = \boldsymbol{l}'\boldsymbol{\mathcal{V}}\boldsymbol{l} + b_{-1/2} + b_{-1}, \tag{34}$$

where $b_{-j/2} = O_p(n^{-j/2})$, $j = 1, 2$, and $b_{-1/2} = n^{-1}\sum_{i=1}^n b_i$, Kundhi and Rilstone (2013) showed that the CDFs of $\hat{\xi}$ and $\hat{t}$ coincide to order $o(n^{-1/2})$, where

$$\hat{\xi} = \frac{\sqrt{n}\boldsymbol{l}'\boldsymbol{a}_{-1/2}}{\sqrt{\boldsymbol{l}'\boldsymbol{\mathcal{V}}\boldsymbol{l}}} + \frac{\sqrt{n}\boldsymbol{l}'\boldsymbol{a}_{-1}}{\sqrt{\boldsymbol{l}'\boldsymbol{\mathcal{V}}\boldsymbol{l}}} - \frac{\sqrt{n}\boldsymbol{l}'\boldsymbol{a}_{-1/2}}{\sqrt{\boldsymbol{l}'\boldsymbol{\mathcal{V}}\boldsymbol{l}}}\frac{b_{-1/2}}{2\boldsymbol{l}'\boldsymbol{\mathcal{V}}\boldsymbol{l}}. \tag{35}$$

Note that the third term in the above expression is due to the correlation of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\mathcal{V}}}$. (If $\boldsymbol{\mathcal{V}}$ is known, then the assumption (34) is not relevant and the third term in (35) is not needed.) Kundhi and Rilstone (2013) derived the first three (approximate) cumulants of $\hat{\xi}$, up to $o(n^{-1/2})$, in terms of cumulants of ξ, defined as the sum of the first two terms in (35). Suppose one can write $\boldsymbol{a}_{-1/2} = -n^{-1}\sum_{i=1}^n \boldsymbol{d}_i$. (For the case of i.i.d. data, $\boldsymbol{d}_i = [\mathrm{E}(\nabla q_1(\boldsymbol{\beta}_0)]^{-1} q_1(\boldsymbol{\beta}_0)$.) With $\mathrm{E}(\xi) = n^{-1/2}\kappa_1$, $\mathrm{Var}(\xi) = 1 + O(n^{-1})$, and $\mathrm{E}((\xi - \mathrm{E}(\xi))^3) = n^{-1/2}\kappa_3 + O(n^{-3/2})$, Kundhi and Rilstone (2013) defined the approximate cumulants of $\hat{\xi}$ as

$$\hat{\kappa}_1 = \kappa_1 + \frac{\mathrm{E}(\boldsymbol{l}'\boldsymbol{d}_1 b_1)}{2(\boldsymbol{l}'\boldsymbol{\mathcal{V}}\boldsymbol{l})^{3/2}}, \quad \hat{\kappa}_3 = \kappa_3 + \frac{3\mathrm{E}(\boldsymbol{l}'\boldsymbol{d}_1 b_1)}{(\boldsymbol{l}'\boldsymbol{\mathcal{V}}\boldsymbol{l})^{3/2}}. \tag{36}$$

Then the Edgeworth approximation to the distribution of $\hat{t}$, defined as

$$P(t) = \Phi(t) - \frac{\hat{\kappa}_1}{\sqrt{n}}\phi(t) - \frac{1}{6}\frac{\hat{\kappa}_3}{\sqrt{n}}(t^3 - 3t)\phi(t), \tag{37}$$

where $\Phi(t)$ and $\phi(t)$ are, respectively, the CDF and PDF of a standard normal distribution, is shown to approximate the true distribution of $\hat{t}$ to order $o(n^{-1/2})$ in the sense that $\sup_{t\in\mathbb{R}} |P(t) - \Pr(\hat{t} \le t)| = o(n^{-1/2})$. One can use (35) to handle straightforwardly the scalar case ($p = 1$) or non-studentized statistic.

Kundhi and Rilstone (2013) showed that the approximation (35) is valid by verifying the conditions in Bhattacharya and Ghosh (1978). For the more general case of nonlinear estimators with non-i.i.d. data, Bao and Ullah (2009) derived the approximate cumulants of $\boldsymbol{l}_j'\hat{\boldsymbol{\beta}}$, where $\boldsymbol{l}_j$ is the j-th unit vector in $\mathbb{R}^p$. It is expected that a similar adjustment, taking into account the correlation of $\hat{\boldsymbol{\beta}}$ and the estimated asymptotic variance matrix, is needed when a studentized statistic is used. However, it is still an open question regarding whether the resulting Edgeworth approximation using the approximate cumulants is a valid distribution function.

For the univariate first-order dynamic model, Phillips (1977) derived the exact cumulants of the LS estimator of the AR(1) coefficient, assuming normality, and used them to construct the Edgeworth approximation. He also used the approximate cumulants of the associated t statistic to construct the corresponding Edgeworth approximate distribution. Phillips and Park (1988) developed the Edgeworth-type expansion of the Wald statistic by first employing Taylor-series expansions of both the restrictions (evaluated at the sample estimate) and its estimated asymptotic covariance matrix. Then they derived the cumulants of the resulting approximate Wald statistic and used them to construct the Edgeworth approximate distribution of the Wald statistic, up to $o(n^{-1})$. Note that Phillips and Park (1988) assumed either the distribution of the estimator is exactly normal or has a valid Edgeworth distribution to begin with.

5 Future Directions

Recent advancements in computation, symbolic calculation, data science, and machine learning, with the presence of big data, may make one wonder about the future of finite sample econometrics. Even with big data, inference remains an important and fundamental issue, especially when some covariates are rare-event type. While big data may make frequent inference (and forecasting) updating possible, it never dismisses the relevance of inferences, whether they are based on the asymptotic or finite sample theory. There are occasions where data frequency is key factor in determining the properties of estimators and statistics of interest and it can happen that even in the presence of big data, the asymptotic and finite sample distributions give rise to quite different descriptions of their sampling distributions. Of course, there are many areas in the social sciences where the data are limited due to the nature of the variables involved and developing the finite sample theory is even more needed. Improvements in computational methods have made it promising for a return to the topics of exact theory and higher-order approximations. The following are fruitful directions to explore and improvements in computational technology may yield some useful results in the near future.

1. The expansion as in (1) is in the order of power series of $n^{-1/2}$, which is a natural choice given a typical estimator is $\sqrt{n}$-consistent. However, under infill asymptotics, the convergence rate is different. Another example is the role of the concentration parameter as demonstrated by (18). Related literature includes the fixed-b asymptotics for nonparametric estimation; large-N/large-T asymptotics in panel; sparse/dense weight matrices in spatial econometrics. Suppose the convergence rate is h_n, possibly a function of n, then a natural extension is to expand by power series of h_n.
2. The exact moments/distribution for the multivariate case seems to be a largely unexplored area. Recent contributions by Hashiguchi et al. (2018) and Hillier and Kan (this issue) provide some insights into the properties of the Wishart

matrix, which arise naturally in multivariate analysis, but still there appears to be no obvious way of deriving the exact moments.

3. The approximate skewness and kurtosis derived in Bao and Ullah (2009) may be used to construct the Edgeworth distribution of a general class of estimators and the associated t statistics and maybe Wald statistics. The challenge is to verify the resulting approximation constitutes a valid approximate distribution.
4. One major area where the Nagar-type approximate moments have not been developed is the class of estimators derived by solving discrete moment function, such as the quantile estimators and estimators under asymmetric loss functions. Only asymptotic theory results have been developed for such cases but not the analytical finite sample results. This is an open area for future research, but see recent attempts by Lee et al. (2018) and Franguridi et al. (2021).
5. Many financial and macro-economic models involve non-stationary/long-memory variables. The analytical finite sample results in this area are not yet explored extensively, although see the works by Abadir (1993), Kiviet and Phillips (2005), Phillips and Lee (2013), and Bao et al. (2014).
6. A number of non-linear applied micro models such as logit/probit models, censored models, truncated models have not been explored for their analytical finite sample properties. Rilstone and Ullah (2002) derived the second-order bias of Heckman's two-step estimator for sample selection models and one has yet to extend to other frequently used micro models.

Acknowledgements This paper is dedicated to Professor A.L.Nagar for this special issue. We thank Essie Maasoumi for his helpful comments.

References

1. Abadir, K.M. 1993. OLS bias in a nonstationary regression. *Econometric Theory* 9(1): 81–93.
2. Anderson, T.W., Rubin, H. 1949. Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* 20(1): 46–63.
3. Anderson, T.W., Sawa, T. 1973. Distributions of estimates of coefficients of a single equation in a simultaneous system and their asymptotic expansions. *Econometrica* 41(4): 683–714.
4. Angrist, J.D., Pischke, J. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
5. Bekker, P.A. 1994. Alternative approximations to the distributions of instrumental variable estimators. *Econometrica* 62(3): 657–681.
6. Bao, Y. 2013. Finite-sample bias of the QMLE in spatial autoregressive models. *Econometric Theory* 29(1): 68–88.
7. Bao, Y., Kan, R. 2013. On the moments of ratios of quadratic forms in normal random variables. *Journal of Multivariate Analysis* 117: 229–245.
8. Bao, Y., Ullah, A., Zhang, R. 2014. Moment approximation for least-squares estimator in first-order regression models with unit root and nonnormal errors. *Advances in Econometrics* 33: 65–92.
9. Bao, Y., Ullah, A., Wang, Y. 2019. Distribution of the mean reversion estimator in the Ornstein–Uhlenbeck process. *Econometric Reviews* 36 (6-9): 1039–1056.
10. Bao, Y., Ullah, A. 2007a. Finite sample properties of maximum likelihood estimator in spatial models. *Journal of Econometrics* 137(2): 396–413.
11. Bao, Y., Ullah, A. 2007b. The second-order bias and mean squared error of estimators in time-series models. *Journal of Econometrics* 140(2): 650–669.

12. Bao, Y., Ullah, A. 2009. On skewness and kurtosis of econometric estimators. *The Econometrics Journal* 12(2): 232–247.
13. Bao, Y., Ullah, A. 2010. Expectation of quadratic forms in normal and nonnormal variables with applications. *Journal of Statistical Planning and Inference* 140(5): 1193–1205.
14. Bao, Y., Liu, X., Ullah, A. 2021. On the exact statistical distribution of econometric estimators and test statistics. In *Advances in Statistics–Theory and Applications*, ed. I. Ghosh, N. Balakrishnan, and T. Ng, 119–131. Springer.
15. Basmann, R.L. 1961. A note on the exact finite sample frequency functions of generalized classical linear estimators in two leading over-identified cases. *Journal of the American Statistical Association* 56(295): 619–636.
16. Bhattacharya, R.N., Ghosh., J.K. 1978. On the validity of the formal Edgeworth expansion. *Annals of Statistics* 6(2): 434–451.
17. Chu, J., Lee, T.-H., Ullah, A., Xu, H. 2021. Exact distribution of the f-statistic under heteroskedasticity of unknown form for improved inference. *Journal of Statistical Computation and Simulation* 91(9): 1782–1801.
18. Dwivedi, T. D., Srivastava, V. K. 1984. Exact finite sample properties of double k-class estimators in simultaneous equations. *Journal of Econometrics* 25(3): 263–283.
19. Donald, S.G., Newey, W.K. 2001. Choosing the number of instruments. *Econometrica* 69(5): 1161–1191.
20. Edgeworth, F.Y. 1896. The asymmetrical probability curve. *Philosophical Magazine Series* 5 41(249): 90–99.
21. Edgeworth, F.Y. 1905. The law of error. *Transactions of the Cambridge Philosophical Society* 20(36-65): 113–141.
22. Fisher, R.A. 1921. On the "probable error" of a coefficient of correlation deduced from a small sample *Metron* 1: 3–32.
23. Fisher, R.A. 1922. The goodness of fit of regression formulae and the distribution of regression coefficients. *Journal of the Royal Statistical Society* 85 (4): 597–612.
24. Fisher, R.A. 1928. The general sampling distribution of the multiple correlation coefficient. *Proceedings of the Royal Society A* 121 (788): 654–673.
25. Fisher, R.A. 1935. The mathematical distributions used in the common tests of significance. *Econometrica* 3 (4): 353–365.
26. Forchini, G. 2002. The exact cumulative distribution function of a ratio of quadratic forms in normal variables, with application to the AR (1) model. *Econometric Theory* 18 (4): 823–852.
27. Franguridi, G., Gafarov, B., Wüthrich, K. 2021. Conditional quantile estimators: a small sample theory. CESifo Working Paper Series 9046, CESifo.
28. Fuller, W.A. 1977. Some properties of a modification of the limited information estimator. *Econometrica* 45(4): 939–953.
29. Gil-Pelaez, J. 1951. Note on the inversion theorem. *Biometrika* 38(3-4): 481–482.
30. Götze, F. 1987. Approximations for multivariate U-statistics. *Journal of Multivariate Analysis* 22(2): 212–229.
31. Gupta,Y. P., Ullah, A. 1970. A note on the moments of the Wald's estimator. *Statistica Neerlandica* 24(3): 109–123.
32. Gurland, J. 1948. Inversion formula for the distribution of ratios. *Annals of Mathematical Statistics* 19(2): 228–237.
33. Haavelmo, T. 1947. Methods of measuring the marginal propensity to consume. *Journal of the American Statistical Association* 42(237): 105–122.
34. Hahn, J., Hausman, J., Kuersteiner, G. 2004. Estimation with weak instruments: Accuracy of higher-order bias and MSE approximations. *The Econometrics Journal* 7(1): 272–306.
35. Hahn, J., Newey, W. 2004. Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica* 72(4): 1295–1319.
36. Hall, P. 1992. *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.
37. Harding, M., Hausman, J., Palmer, C.J. 2016. Finite sample bias corrected IV estimation for weak and many Instruments. *Advances in Econometrics* 36: 245–273.
38. Hashiguchi, H., Takayama, N., Takemura, A. 2018. Distribution of the ratio of two Wishart matrices and cumulative probability evaluation by the holonomic gradient method. *Journal of Multivariate Analysis* 165: 270–278.
39. Hillier, G., Kan, R. (this issue). Moments of a Wishart matrix. *Journal of Quantitative Economics.*
40. Hillier, G., Kan, R. 2021. Properties of the inverse of a noncentral Wishart matrix. *Econometric Theory*, forthcoming.

41. Hillier, G., Kan, R., Wang, X. 2009. Computationally efficient recursions for top-order invariant polynomials with applications. *Econometric Theory* 25(1): 211–242.
42. Hillier, G., Kan, R., Wang, X. 2014. Generating functions and short recursions, with applications to the moments of quadratic forms in noncentral normal vectors. *Econometric Theory* 30(2): 436–473.
43. Hillier, G., Kinal, T.W., Srivastava, V.K. 1984. On the moments of ordinary least squares and instrumental variables estimators in a general structural equation. *Econometrica* 52(1): 185–202.
44. Horowitz, J.L. 2001. The bootstrap. In *Handbook of Econometrics*, vol. 5, ed. J. Heckman & E. Leamer, 3159–3228. Amsterdam: North Holland.
45. Hurwicz, L. 1950. Least squares bias in time series. In *Statistical Inference in Dynamic Economic Models*, ed. T.C. Koopmans, 365–383. New York: Wiley.
46. Imhof, J.P. 1961. Computing the distribution of quadratic forms in normal variables. *Biometrika* 48(3-4): 419–426.
47. Kabe, D.G. 1963. A note on the exact distributions of the gcl estimators in two leading over identified cases. *Journal of the American Statistical Association* 58(302): 535–537.
48. Kabe, D.G. 1964. On the exact distributions of the GCL estimators in a leading three-equation case. *Journal of the American Statistical Association* 59(307): 881–894.
49. Kadane, J.B. 1971. Comparison of k-class estimators when the disturbances are small. *Econometrica* 39(5): 723–737.
50. Kakwani, N.C. 1967. The unbiasedness of Zellner's seemingly unrelated regression equations estimators. *Journal of the American Statistical Association* 62(317): 141–142,
51. Kakwani, N.C. 1971. The bias of the three-stage least-squares estimators in simultaneous equations. *Australian Journal of Statistics* 13: 19–26.
52. Kinal, T.W. 1980. The existence of moments of k-class estimators. *Econometrica* 48(1): 241–249.
53. Kiviet, J. F., Phillips, G.D.A. 1993. Alternative bias approximations in regressions with a lagged-dependent variable. *Econometric Theory* 9(1): 62–80.
54. Kiviet, J. F., Phillips, G.D.A. 2005. Moment approximation for least-squares estimators in dynamic regression models with a unit root. *Econometrics Journal* 8(2): 115–142.
55. Kollo, T., von Rosen, D. 1998. A unified approach to the approximation of multivariate densities. *Scandinavian Journal of Statistics* 25(1): 93–109.
56. Kundhi, G., Rilstone, P. 2012. Edgeworth expansions for GEL estimators. *Journal of Multivariate Analysis* 106: 118–146.
57. Kundhi, G., Rilstone, P. 2013. Edgeworth and saddlepoint expansions for nonlinear estimators. *Econometric Theory* 29(5): 1–22.
58. Kundhi, G., Rilstone, P. 2020. Simplified matrix methods for multivariate Edgeworth expansions. *Journal of Quantitative Economics* 18: 293–326.
59. Lee, T.-H., Ullah, A., Wang, H. 2018. The second-order bias of quantile estimators. *Economics Letters* 173, 143–147.
60. Maasoumi, E., Phillips, P.C.B. 1982. On the behavior of inconsistent instrumental variable estimators. *Journal of Econometrics* 19(2–3): 183–201.
61. Maasoumi, E. 1978. A modified Stein-like estimator for the reduced form coefficients of simultaneous equations, *Econometrica* 46(3): 695-703
62. Mariano, R.S. 1982. Analytical small-sample distribution theory in econometrics: the simultaneous-equations case. *International Economic Review* 23(3): 503–533.
63. Mariano, R., Sawa, T. 1972. The exact finite-sample distribution of the limited-information maximum likelihood estimator in the case of two included endogenous variables. *Journal of the American Statistical Association* 67(337): 159–163,
64. Nagar, A.L. 1959. The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica* 27(4): 575–595.
65. Nagar, A.L. 1962. Double k-class estimators of parameters in simultaneous equations and their small sample properties. *International Economic Review* 3(2): 168-188.
66. Nagar, A.L., Gupta, Y.P. 1968. The bias of Liviatan's consistent estimator in a distributed lag model. *Econometrica* 36(2): 337–342.
67. Nagar, A.L, Kakwani, N.C. 1965. Note on the use of prior information in statistical estimation of economic relations. *Sankhyā: The Indian Journal of Statistics, Series A* 27(1): 105–112.
68. Nagar, A.L., Ullah, A. 1973. Note on approximate skewness, and kurtosis of the two stage least squares estimator. *Indian Economic Review* 7:70–80.

69. Phillips, P.C.B. 1977. Approximations to some finite sample distributions associated with a first-order stochastic difference equation. *Econometrica* 45(2): 463–485.
70. Phillips, P.C.B. 1978. Edgeworth and saddlepoint approximations in the first-order noncircular autoregression. *Biometrika* 65(1): 91–98.
71. Phillips, P.C.B. 1980. The exact distribution of instrumental variable estimators in an equation containing $n+1$ endogenous variables. *Econometrica* 48(4): 861–878.
72. Phillips, P.C.B. 1983. Exact small sample theory in the simultaneous equations model. In *Handbook of Econometrics*, vol. 1, ed. Z. Griliches & M. Intriligator, 449–516. Amsterdam: North Holland.
73. Phillips, P.C.B., Park, J.Y. 1988. On the formulation of Wald tests of nonlinear restrictions. *Econometrica* 56(5): 1065–1083.
74. Phillips, P.C.B., Lee, J. H. 2013. Predictive regression under various degrees of persistence and robust long-horizon regression. *Journal of Econometrics* 177 (2): 250–264.
75. Richardson, D.H. 1968. The exact distribution of a structural coefficient estimator. *Journal of the American Statistical Association* 63(324): 1214–1226.
76. Rilstone, P., Srivastava, V.K., Ullah, A. 1996. The second-order bias and mean squared error of nonlinear estimators. *Journal of Econometrics* 75 (2): 369–395.
77. Rilstone, P., Ullah, A. 2002. Sampling bias in Heckman's sample selection estimator. In *Recent Advances in Statistical Methods*, ed. Y.P. Chaubey, 263–273, World Scientific Publishing Company.
78. Rothenberg, T.J. 1984a. Approximating the distributions of econometric estimators and test statistics. In *Handbook of econometrics*, vol. 2, ed. Z. Griliches & M. Intriligator, 881–893. Amsterdam: North Holland.
79. Rothenberg, T.J. 1984b. Approximate normality of generalized least squares estimates. *Econometrica* 52(4): 811–825.
80. Sargan, J.D. 1975. Gram-Charlier approximations applied to t ratios of k-class estimators. *Econometrica* 43(2): 327–346.
81. Sargan, J.D. 1976. Econometric estimators and the Edgeworth approximation. *Econometrica* 44(3): 421–448.
82. Sargan, J.D. 1980. Some tests for dynamic specification for a single equation. *Econometrica* 48(4): 879–897.
83. Sawa, T. 1969. The exact sampling distribution of ordinary least squares and two-stage least squares estimators. *Journal of the American Statistical Association* 64(327): 923–937.
84. Sawa, T. 1972. Finite sample properties of the k-class estimators. *Econometrica* 40(4): 653–680.
85. Skovgaard, Ib M. 1986. On multivariate Edgeworth expansions. *International Statistical Review* 54(2): 169–186.
86. Srinivasan, T. 1970. Approximations to finite sample moments of estimators whose exact sampling distributions are unknown. *Econometrica* 38(3): 533–541.
87. Srivastava, V., Giles, D.E.A. 1987. *Seemingly Unrelated Regression Equations Models: Estimation and Inference*. Marcel Dekker.
88. Srivastava, V.K., Maekawa, K. 1995. Efficiency properties of feasible generalized least squares estimators in SURE models under non-normal disturbances. *Journal of Econometrics* 66(1–2): 99–121.
89. Taylor, W.E. 1983. On the relevance of finite sample distribution theory. *Econometric Reviews* 2(1): 1–39.
90. Zhou, Q., Yu, J. 2015. Asymptotic theory for linear diffusions under alternative sampling schemes. *Economics Letters* 128, 1–5.
91. Ullah, A. 2004. *Finite sample econometrics*. New York: Oxford University Press.
92. Ullah, A., Nagar, A.L. 1974. The exact mean of the two stage least squares estimator of the structural parameters in the equation having three endogenous variables. *Econometrica* 42(4): 749–758.