JOURNAL OF FINANCIAL AND QUANTITATIVE ANALYSIS Vol. 48, No. 1, Feb. 2013, pp. 277–307
COPYRIGHT 2013, MICHAEL G. FOSTER SCHOOL OF BUSINESS, UNIVERSITY OF WASHINGTON, SEATTLE, WA 98195
doi:10.1017/S0022109013000070

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

Using Samples of Unequal Length in Generalized Method of Moments Estimation

Anthony W. Lynch and Jessica A. Wachter*

Abstract

This paper describes estimation methods, based on the generalized method of moments (GMM), applicable in settings where time series have different starting or ending dates. We introduce two estimators that are more efficient asymptotically than standard GMM. We apply these to estimating predictive regressions in international data and show that the use of the full sample affects inference for assets with data available over the full period as well as for assets with data available for a subset of the period. Monte Carlo experiments demonstrate that reductions hold for small-sample standard errors as well as asymptotic ones.

I. Introduction

Many applications in financial economics involve data series that have different starting dates or, more rarely, different ending dates. Settings where some data series are available over a much shorter time frame than others include estimation and testing using international data and performance evaluation of mutual funds. These problems represent only the most extreme examples of differences in data length. More broadly, aggregate stock return data may be available over a longer time frame than macroeconomic data, cash flow and earnings data, term structure data, or options data.

Building on work of Anderson (1957), Stambaugh (1997) derives estimates of the mean and variance of financial time series, as well as the posterior distribution of returns, in a setting in which some return series start at a later date than others. Pastor and Stambaugh (2002a), (2002b) apply these methods to the derivation of Bayesian posteriors for means and variances of mutual fund returns

*Lynch, alynch@stern.nyu.edu, Stern School of Business, New York University, 44 W 4th St, New York, NY 10012; Wachter, jwachter@wharton.upenn.edu, Wharton School, University of Pennsylvania, 3620 Locust Walk, Philadelphia, PA 19104. We thank Yacine Ait-Sahalia, David Chapman, Robert Engle, Martin Lettau, Andrew Lo, Paul Malatesta (the editor), Kenneth Singleton, Robert Stambaugh, Jim Stock, Amir Yaron, Motohiro Yogo, and Guofu Zhou (associate editor and referee), as well as seminar participants at the 2005 American Finance Association meetings, New York University, and University of Pennsylvania for their comments and suggestions.

when a longer data set for factors is observed. These studies assume that returns are normal and that they are independent and identically distributed (IID).[1]

While the assumption of IID normality allows much progress to be made, financial return data often exhibit significant nonnormalities (MacKinlay and Richardson (1991), Zhou (1993)). Moreover, it is of interest to have a general method for estimating models that goes beyond means and variances. We develop a method of incorporating samples of unequal lengths, based on the generalized method of moments (GMM). We show that our method is more efficient than standard GMM and more efficient than introducing the data from the longer series in a "naive" way. Because it is based on GMM, our method can be used for nonlinear estimation and for processes that are serially correlated and feature conditional heteroskedasticity. We focus on GMM because, as shown as Cochrane (2001), many common estimation techniques used in finance can be seen as special cases of GMM. Further, many empirical studies in finance explicitly use GMM (e.g., Hansen and Singleton (1982), Brandt (1999), Harvey (1989), and Zhou (1994)). Assumptions required for the consistency and asymptotic normality of the standard GMM estimator are also required here. We adopt the mixing assumption of White and Domowitz (1984) as a means of limiting the temporal dependence of the underlying stochastic process. Intuitively, mixing requires that autocovariances vanish as the lag length increases. This assumption allows for many processes of interest in financial economics, such as finite autoregressive moving average processes with general conditions on the underlying errors (see Phillips (1987)).[2]

Because our method is based on GMM, many of our results are asymptotic.[3] So that the asymptotic approximation is reasonable, care must be taken to ensure that the missing data problem does not become trivial as the sample size becomes large. We thus develop an asymptotic theory that keeps the fraction of missing data fixed as the sample size approaches infinity. To be precise, if T denotes the length of the longer sample, we say that λT is the length of the shorter sample, for $0 < \lambda \leq 1$. We hold λ constant as T approaches infinity. This approach has a parallel in the simulated method of moments estimation technique (see Duffie and Singleton (1993)), where the length of the simulated series divided by the length of the observed series is assumed to be constant as both series lengths approach

[1]Other studies in economics that make use of samples of unequal length include Harvey, Koopman, and Penzer (1998), Schmidt (1977), Swamy and Mehta (1975), Conniffe (1985), Storesletten, Telmer, and Yaron (2004), and Patton (2006). These studies all take a likelihood-based approach and make specific assumptions on the form of the distribution as well as on the function of the data to be estimated. Another strand of literature considers the problem of n independent individuals observed at up to T time periods, where some individuals drop out of the study (see, e.g., Robins and Rotnitsky (1995)). The independence across individuals and the fact that asymptotics are derived as n, rather than T, approaches infinity differentiates this problem from the one considered here. Little and Rubin (2002) survey the statistical literature confronting missing data problems.

[2]Like many of the studies mentioned here, we do assume that the data are missing at random, in the sense defined by Little and Rubin (2002). Stambaugh (1997) discusses cases where this assumption holds in financial time series, such as when the start date depends only on the long-history asset returns, and cases where it does not, such as when the decision to add a country to a list of emerging markets depends on past unobserved returns on that country (see Goetzmann and Jorion (1999)).

[3]We also verify, in Monte Carlo experiments, that our methods deliver efficiency gains in small samples.

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

infinity, and also in the literature on structural breaks (Andrews and Fair (1988), Ghysels and Hall (1990), Andrews and Ploberger (1994), Stock (1994), Sowell (1996), and Ghysels, Guay, and Hall (1997)).

We focus on the case in which some moment conditions are observed over the full range of dates, while others are observed over a time span that has the same ending date but a later starting date because this is the most common pattern in finance applications (we later generalize this to other patterns of missing data).[4] While general, these estimators are straightforward to implement, as we show in an application involving international data (Section III), and have natural and intuitive interpretations.

The 1st estimator (which we call the adjusted-moment (AM) estimator) uses full-sample averages to estimate the moments for which full-sample data are available, and short-sample averages to estimate moments for which only short-sample data are available. Then the moments for which only the short sample is available are "adjusted" using coefficients from a regression of the short-sample moments on the full-sample moments. This is reminiscent of an adjustment that appears in Stambaugh (1997) and Little and Rubin (2002) but here operates in a more general context. The 2nd estimator (which we call the overidentified (OI) estimator) uses the extra data available from the full sample as a new set of moment conditions. This estimator was suggested by Stambaugh (1997) and, in the linear context of that paper, turns out to be identical to our AM estimator (and the maximum-likelihood estimator proposed in that paper). In the more general context of our paper, the 2 estimators are equivalent asymptotically but typically differ in finite samples.

In that it is based on GMM, our study is closely related to that of Singleton ((2006), chap. 4.5). Besides placing the missing data problem within the context of GMM, Singleton also takes the same approach to asymptotics: Namely, the ratio of the length of the shorter sample to that of the longer sample remains constant as the total length goes to infinity in both his study and ours. Singleton proposes moment conditions that are the same as those for our OI estimator. However, he derives a different weighting matrix. The weighting matrix that we derive allows us to show that our estimators are more efficient than standard GMM and more efficient than a naive approach to using the full sample. We also depart from Singleton's study in that we define the asymptotically equivalent AM estimator, study the finite-sample performance of the estimators, and apply them to predictive regressions.

The organization of the paper is as follows: Section II defines our estimators and discusses their efficiency properties. Section III provides intuition for the efficiency gains from using the full sample. Section IV illustrates our methods through an application to international data. Section V presents a Monte Carlo analysis showing that the efficiency gains are present in small samples. Most of our paper focuses on the case where one data series begins earlier than another. However, our approach can easily be generalized to other patterns of missing

[4]In its focus on the efficiency results of carefully including additional data, this study has parallels in studies that focus on including high-frequency data in estimation while accounting for market microstructure effects (see Ait-Sahalia, Mykland, and Zhang (2005), Bandi and Russell (2006)).

data, provided that the data are missing in large blocks.[5] Section VI provides this generalization. Section VII concludes.

II. GMM Estimators for Samples of Unequal Length

A. Definitions

Following Cochrane (2001), assume that the model to be estimated can be expressed as

$$\mathrm{E}\left[f(x_t, \theta_0)\right] \quad = \quad 0.$$

Here, f is a vector of l restrictions. The true parameters of the model are represented by the q-vector θ_0. Finally, x_t is a vector-valued stochastic process. In what follows, we derive results based on assumptions that are standard in a GMM setting (see Appendix A for more details).

In many applications, it happens that data are missing for the early part of the sample period for some moment conditions (see Section III for an application to international data). We partition the elements of x_t so that $x_t = [x_{1t}^\top,\ x_{2t}^\top]^\top$, where data on x_{1t} are assumed to be available for the full period, and data on x_{2t} are assumed to be available for only the later part of the sample period. Similarly, we can partition the elements of f into those that depend only on x_{1t} and those that depend on both x_{1t} and x_{2t}: $f(x_t, \theta) = [f_1(x_{1t}, \theta)^\top,\ f_2(x_t, \theta)^\top]^\top$. Let f_1 be $l_1 \times 1$ and f_2 be $l_2 \times 1$, where $l_1 + l_2 = l$.

Let λ denote the fraction of the period for which all data are available. Then, x_{1t} is observed from $t = 1, \ldots, T$, while x_{2t} is observed from $t = (1-\lambda)T + 1, \ldots, T$. Define the following partial sums:[6]

$$\begin{aligned}
g_{1,T}(\theta) &= \frac{1}{T}\sum_{t=1}^{T} f_1(x_{1t}, \theta), \\
g_{1,(1-\lambda)T}(\theta) &= \frac{1}{(1-\lambda)T}\sum_{t=1}^{(1-\lambda)T} f_1(x_{1t}, \theta), \\
g_{1,\lambda T}(\theta) &= \frac{1}{\lambda T}\sum_{t=(1-\lambda)T+1}^{T} f_1(x_{1t}, \theta), \qquad \text{and} \\
g_{2,\lambda T}(\theta) &= \frac{1}{\lambda T}\sum_{t=(1-\lambda)T+1}^{T} f_2(x_t, \theta).
\end{aligned}$$

[5] Under general assumptions on the dependence of the underlying stochastic process, it is necessary that the number of "blocks" remains fixed asymptotically. This distinguishes the problem we tackle from the problem posed by data sampled at different frequencies (see Ghysels, Santa-Clara, and Valkanov (2005)).

[6] Formally, λ is a rational number strictly between 0 and 1. Define n_0 to be the smallest positive integer n such that $n\lambda$ is an integer. We consider partial sums of f of length λT and $(1-\lambda)T$ for T a multiple of n_0. For the remainder of the paper, we let T approach infinity along the subsequence of integer multiples of n_0. Alternatively, we could define partial sums of length $\lambda n_0 T'$ and $(1-\lambda)n_0 T'$ for any integer T'. The results would be identical, but the notation would be more cumbersome.

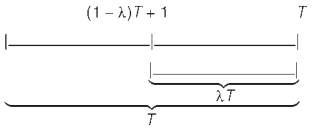
https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

Sums of f are indexed by the length of the sample. This is a slight abuse of notation because the subscript λT does not refer to the sum taken over observations $1, \ldots, \lambda T$. The subscripts λT, $(1 - \lambda)T$, and T can be understood as referring to intervals of the data rather than the ending point of the sample. Figure 1 illustrates the notation.

FIGURE 1

Notation for Data Missing at the Start of the Sample

In Figure 1, T denotes the end date of both series, 1 denotes the start date of the 1st series, and $(1 - \lambda)T + 1$ denotes the start date of the 2nd series. Note that λT is the length of the 2nd series.



Let $w_{1t} = f_1(x_{1t}, \theta_0)$ and $w_{2t} = f_2(x_t, \theta_0)$. Following Hansen (1982), define matrices

$$R_{ij}(\tau) \quad = \quad \mathrm{E}\left[w_{i0} w_{j,-\tau}^{\top}\right], \qquad i, j \; = \; 1, 2.$$

Under standard assumptions, these sums converge (see White (1994), prop. 3.44). Let

$$S_{ij} \; = \; \sum_{\tau=-\infty}^{\infty} R_{ij}(\tau) \qquad \text{and} \qquad S \; = \; \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}.$$

It is also useful to define the matrix of coefficients from a regression of the 2nd series on the 1st:

$$B_{21} \quad = \quad S_{21} S_{11}^{-1}.$$

The residual variance from this regression will be denoted Σ, where

$$\Sigma \quad = \quad S_{22} - S_{21} S_{11}^{-1} S_{12}. \tag{1}$$

Note that S is known as the spectral density matrix.

In this setting, standard GMM corresponds to using moment conditions measured over the subperiod for which all the data are available. That is, the standard GMM estimator solves

$$\min_{\theta} \left[g_{1,\lambda T}(\theta)^{\top}, \; g_{2,\lambda T}(\theta)^{\top}\right]^{\top} W_T \begin{bmatrix} g_{1,\lambda T}(\theta) \\ g_{2,\lambda T}(\theta) \end{bmatrix}$$

for a positive-definite and symmetric weighting matrix W_T. In what follows, we focus on the case where the weighting matrix is asymptotically efficient. In the case of standard GMM, this implies that the weighting matrix asymptotically approaches S^{-1}. We let $\hat{S}_T$ denote an estimator of S.[7] Let

$$\hat{\theta}_T^{S} \quad = \quad \underset{\theta}{\operatorname{argmin}} \left[g_{1,\lambda T}(\theta)^{\top}, \; g_{2,\lambda T}(\theta)^{\top}\right]^{\top} \hat{S}_T^{-1} \begin{bmatrix} g_{1,\lambda T}(\theta) \\ g_{2,\lambda T}(\theta) \end{bmatrix}. \tag{2}$$

[7] Stated more precisely, we choose $\hat{S}_T$ to converge to S almost surely. Convergence for estimates of variance-covariance matrices that follow should be interpreted similarly.

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

We call this the short estimator.

Standard arguments show that the short estimator is consistent and asymptotically normal. However, the short estimator does not use all of the data available. A natural estimator to consider takes the same form as the short estimator, except with $g_{1,\lambda T}(\theta)$ replaced by its full-sample counterpart, $g_{1,T}(\theta)$. Because this is the simplest estimator that makes use of all of the data, we call this the long estimator and let

$$\hat{\theta}_T^{\mathcal{L}} = \underset{\theta}{\operatorname{argmin}} \left[g_{1,T}(\theta)^\top,\ g_{2,\lambda T}(\theta)^\top\right]^\top \left(\hat{S}_T^{\mathcal{L}}\right)^{-1} \begin{bmatrix} g_{1,T}(\theta) \\ g_{2,\lambda T}(\theta) \end{bmatrix}, \tag{3}$$

where $\hat{S}_T^{\mathcal{L}}$ is an estimate of $S^{\mathcal{L}}$, the asymptotic variance of:

$$\sqrt{\lambda T}[g_{1,T}(\theta)^\top,\ g_{2,\lambda T}(\theta)^\top]^\top.$$

We argue, however, that the long estimator introduces new data in a suboptimal way. We define 2 alternative estimators. The 1st set takes $\hat{\theta}_T^{\mathcal{L}}$ as a starting point and adjusts the 2nd set of moment conditions based on sample properties of the 1st set of moment conditions. To define this estimator, let $\hat{B}_{21,\lambda T}$ be a matrix converging to B_{21}. The AM estimator, $\hat{\theta}_T^{\mathcal{A}}$, solves

$$\hat{\theta}_T^{\mathcal{A}} = \underset{\theta}{\operatorname{argmin}} \left[g_{1,T}(\theta)^\top,\ g_{2,T}^{\mathcal{A}}(\theta)^\top\right]^\top \left(\hat{S}_T^{\mathcal{A}}\right)^{-1} \begin{bmatrix} g_{1,T}(\theta) \\ g_{2,T}^{\mathcal{A}}(\theta) \end{bmatrix}, \tag{4}$$

where

$$g_{2,T}^{\mathcal{A}}(\theta) = g_{2,\lambda T}(\theta) + \hat{B}_{21,\lambda T}(1-\lambda)(g_{1,(1-\lambda)T}(\theta) - g_{1,\lambda T}(\theta))$$

and $\hat{S}_T^{\mathcal{A}}$ is an estimate of $S^{\mathcal{A}}$, the asymptotic variance of:

$$\sqrt{\lambda T}[g_{1,T}(\theta)^\top,\ g_{2,T}^{\mathcal{A}}(\theta)^\top]^\top.$$

The difference between equations (3) and (4) lies in the 2nd set of moment conditions, for which only the short sample is available. Because

$$g_{1,T} = (1-\lambda)g_{1,(1-\lambda)T} + \lambda g_{1,\lambda T},$$

the 2nd set of moment conditions for the AM estimator, $g_{2,T}^{\mathcal{A}}$, can be written as

$$g_{2,T}^{\mathcal{A}} = g_{2,\lambda T} + \hat{B}_{21,\lambda T}(g_{1,T} - g_{1,\lambda T}).$$

The previous expression illustrates the role of the longer sample in helping to estimate the 2nd set of moment conditions. Consider, for example, the case where g_1 and g_2 are univariate. If g_1 is below average in the 2nd part of the sample, and if g_1 and g_2 are positively correlated, g_2 is also likely to be below average. Thus, the estimate of $\mathrm{E}[f_2(x_0,\theta)]$ should be adjusted upward relative to g_2.

Finally, we define an estimator that makes use of a longer data sample to add overidentifying restrictions. The OI estimator solves

$$\hat{\theta}_T^{\mathcal{I}} = \underset{\theta}{\operatorname{argmin}} \left[g_{1,(1-\lambda)T}(\theta)^\top,\ g_{1,\lambda T}(\theta)^\top,\ g_{2,\lambda T}(\theta)^\top\right]^\top \left(\hat{S}_T^{\mathcal{I}}\right)^{-1} \begin{bmatrix} g_{1,(1-\lambda)T}(\theta) \\ g_{1,\lambda T}(\theta) \\ g_{2,\lambda T}(\theta) \end{bmatrix}, \tag{5}$$

where $\hat{S}_T^{\mathcal{I}}$ is an estimate of $S^{\mathcal{I}}$, the asymptotic variance of:

$$\sqrt{\lambda T}[g_{1,(1-\lambda)T}(\theta)^\top,\ g_{1,\lambda T}(\theta)^\top,\ g_{2,\lambda T}(\theta)^\top].$$

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

B. Asymptotic Distribution

Theorems A.2 and A.3 in Appendix A show that each estimator is consistent for θ_0 and is asymptotically normal. Standard errors can be obtained using the same results as in previous work on GMM. Standard errors depend on the derivative of the moment conditions. Define

$$\begin{aligned} D_{0,i} &= \mathrm{E}\left[\left(\partial f_i/\partial\theta\right)|_{\theta_0}\right], \qquad \text{for } i = 1, 2, \quad \text{and} \\ D_0 &= \begin{bmatrix} D_{0,1} \\ D_{0,2} \end{bmatrix}. \end{aligned}$$

For the short, long, and AM estimators, D_0 is the derivative of the moment condition evaluated at θ_0. As shown in Theorem A.3, the asymptotic distributions of the estimators are normal and centered around θ_0. For example, for the AM estimator,

$$\sqrt{\lambda T}\left(\hat{\theta}_T^{\mathcal{A}} - \theta_0\right) \underset{\mathrm{d}}{\rightarrow} N\left(0, \left(D_0^\top \left(S^{\mathcal{A}}\right)^{-1} D_0\right)^{-1}\right). \tag{6}$$

Analogous equations hold for the short and long estimator, where $S^{\mathcal{A}}$ is replaced by S and $S^{\mathcal{L}}$, respectively (recall that the short estimator is standard GMM). Similarly, for the OI estimator, the derivative D_0 is replaced by $[D_{0,1}^\top,\ D_0^\top]$, and $S^{\mathcal{A}}$ is replaced by $S^{\mathcal{I}}$. In each case, the difference between the estimator and θ_0 is scaled by $\sqrt{\lambda T}$. This an arbitrary choice: We could have equally well have chosen $\sqrt{T}$ (or indeed any constant multiplied by $\sqrt{T}$) and adjusted the variance-covariance matrix in equation (6) appropriately. Regardless of this choice, it is convenient to keep it the same for all 4 estimators.

An important practical step in implementing these estimators is obtaining estimates of the spectral density matrices $S^{\mathcal{L}}$, $S^{\mathcal{A}}$, and $S^{\mathcal{I}}$ to substitute into the equations above. Conveniently, these estimates can be obtained with no more difficulty than estimating the matrix S because these matrices can be completely characterized in terms of the submatrices S_{ij} of S. As shown in Theorem A.1,[8]

$$S^{\mathcal{L}} = \begin{bmatrix} \lambda S_{11} & \lambda S_{12} \\ \lambda S_{21} & S_{22} \end{bmatrix}, \tag{7}$$

$$S^{\mathcal{A}} = \begin{bmatrix} \lambda S_{11} & \lambda S_{12} \\ \lambda S_{21} & S_{22} - (1-\lambda) S_{21} S_{11}^{-1} S_{12} \end{bmatrix}, \qquad \text{and} \tag{8}$$

$$S^{\mathcal{I}} = \begin{bmatrix} \frac{\lambda}{1-\lambda} S_{11} & 0 & 0 \\ 0 & S_{11} & S_{12} \\ 0 & S_{21} & S_{22} \end{bmatrix}. \tag{9}$$

These formulas show that it suffices to have an estimate of the original spectral density matrix S (see Cochrane (2001) for a discussion). Given such an estimate, $S^{\mathcal{A}}$ and $S^{\mathcal{I}}$ are easily constructed by extracting submatrices: S_{11} is the upper left

[8] Our proposed weighting matrix for the OI estimator can be contrasted with that proposed by Singleton (2006). The weighting matrix he proposes is equivalent to the inverse of the matrix given in equation (9), without the $\lambda/(1-\lambda)$ term in the upper left block.

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

$l_1 \times l_1$ submatrix, S_{22} is the $l_2 \times l_2$ lower right submatrix, and so on. Such an estimate will generally make use of the last λT observations, because it is necessary to have all series available. In Section III, we discuss a means of constructing an estimate of S that uses all of the data.

Underlying these results is the asymptotic independence between nonoverlapping samples. That is, $\sqrt{\lambda T} g_{1,(1-\lambda)T}$ and $\sqrt{\lambda T} g_{i,\lambda T}$ are jointly normally distributed and have 0 covariance in the limit as the sample size approaches infinity. Please see Appendix A for a formal statement and proof. Asymptotic independence is intuitive: As more and more data become available, the parts of the nonoverlapping samples that are close to one another become an ever smaller part of the whole. The samples come to be dominated by terms that are far away and thus nearly independent.

C. Efficiency Properties

We now compare the asymptotic efficiency of the 4 estimators. The proof of the following theorem can be found in Appendix B.

Theorem 1. Assume the short, long, AM, and OI estimators are defined as in equations (2)–(5), respectively. Then:

1. The asymptotic distribution of the AM estimator is identical to that of the OI estimator.

2. The AM and OI estimators are more efficient than the short estimator.

3. The AM and OI estimators are more efficient than the long estimator.

Theorem 1 shows that asymptotically, the AM and OI estimators are the same despite the fact that they take very different forms. The 2nd statement shows that there is indeed an efficiency gain from using the longer sample. Moreover, it is more efficient to use the AM or OI estimators than to use the longer sample in a "naive" way, as the 3rd statement shows.

In contrast, the long estimator, despite its use of all the data, may not be more efficient than the short estimator. Statement 2 relies on the fact that

$$S - S^{\mathcal{A}} = (1-\lambda)\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{21}S_{11}^{-1}S_{12} \end{bmatrix}$$

is positive semidefinite (i.e., S is at least as large, in a matrix sense, as $S^{\mathcal{A}}$). However, the analogous quantity for the long estimator,

$$S - S^{\mathcal{L}} = (1-\lambda)\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & 0 \end{bmatrix},$$

will generally not be positive semidefinite. Thus, it is not sufficient to simply include the full data in the estimation. The nonoverlapping part of the sample must be introduced in precisely the right way to produce a gain in efficiency. The difference between the efficient estimators (the AM and OI estimators) and the long estimator is especially surprising given that, when attention is restricted to estimating $f_1(x, \theta)$, the 3 estimators are asymptotically identical. In fact, the

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

gains in efficiency occur because the method uncovers the deviation of $g_{1,\lambda T}$ from 0. Because of the correlation between $g_{1,\lambda T}$ and $g_{2,\lambda T}$, the deviation of $g_{1,\lambda T}$ from 0 implies that $g_{2,\lambda T}$ is also likely to deviate from 0. The efficient estimators make use of this information to construct an estimator of the mean of $f_2(x,\theta)$ that improves on $g_{2,\lambda T}$.

The previous results address the case when the efficient weighting matrix for each estimator is used. Sometimes, it is of interest to use a weighting matrix that is asymptotically inefficient because of small-sample considerations. As the next theorem shows, there is an efficiency gain for using the full sample in this setting as well. The proof is in Appendix B.

Theorem 2. Assume that the weighting matrices approach a positive-definite matrix W. The AM estimator is more efficient than the short and long estimators.

D. Comparing the Efficient Estimators

Because the AM and OI estimators are asymptotically identical, we refer to them as the *efficient estimators*. The previous results raise the question of whether these estimators are identical in finite samples and, if not, what the differences are. We answer these questions by deriving the first-order conditions (FOCs) that determine the estimators. For the purpose of this discussion, we assume that $\hat{S}_T^{\mathcal{I}} = S^{\mathcal{I}}$, $\hat{S}_T^{\mathcal{A}} = S^{\mathcal{A}}$, and $\hat{B}_{21,\lambda T} = B_{21}$. However, the results apply as long as these matrices are constructed using the same estimated submatrices of S.

As shown in Appendix B, the FOC determining the OI estimator is equal to

$$\begin{aligned}(10)\qquad &\frac{1-\lambda}{\lambda} g_{1,(1-\lambda)T}^{\top} S_{11}^{-1} \frac{\partial g_{1,(1-\lambda)T}}{\partial \theta} + g_{1,\lambda T}^{\top} S_{11}^{-1} \frac{\partial g_{1,\lambda T}}{\partial \theta} \\ &\quad + \left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right)^{\top} \Sigma^{-1} \frac{\partial}{\partial \theta} \left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right) \quad = \quad 0.\end{aligned}$$

The FOC determining the AM estimator is

$$\begin{aligned}(11)\qquad &\frac{1}{\lambda} g_{1,T}^{\top} S_{11}^{-1} \frac{\partial g_{1,T}}{\partial \theta} \\ &\quad + \left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right)^{\top} \Sigma^{-1} \frac{\partial}{\partial \theta} \left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right) \quad = \quad 0.\end{aligned}$$

According to Theorem 1, these 2 FOCs must be equivalent as $T \to \infty$. Indeed, they are, because

$$\begin{aligned}\lim_{T\to\infty} \left.\frac{\partial g_{1,(1-\lambda)T}}{\partial \theta}\right|_{\hat{\theta}_T^{\mathcal{I}}} &= \lim_{T\to\infty} \left.\frac{\partial g_{1,\lambda T}}{\partial \theta}\right|_{\hat{\theta}_T^{\mathcal{I}}} \\ &= \lim_{T\to\infty} \left.\frac{\partial g_{1,T}}{\partial \theta}\right|_{\hat{\theta}_T^{\mathcal{A}}} \quad = \quad D_{0,1},\end{aligned}$$

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

and

$$\begin{aligned}\frac{1-\lambda}{\lambda} g_{1,(1-\lambda)T}^{\top} S_{11}^{-1} D_{0,1} + g_{1,\lambda T}^{\top} S_{11}^{-1} D_{0,1} \\ = \frac{1}{\lambda}\left((1-\lambda) g_{1,(1-\lambda)T}^{\top} + \lambda g_{1,\lambda T}^{\top}\right) S_{11}^{-1} D_{0,1} \\ = \frac{1}{\lambda} g_{1,T}^{\top} S_{11}^{-1} D_{0,1}.\end{aligned}$$

For finite T, however, equations (10) and (11) are generally not equivalent. Therefore, the values of the AM and OI estimators differ as well.

There is a special case when the 2 estimators are the same, even in finite samples. The estimators are identical when

$$\frac{\partial g_{1,(1-\lambda)T}}{\partial \theta} = \frac{\partial g_{1,\lambda T}}{\partial \theta},$$

which occurs, for example, when the parameter to be estimated is the mean of x.

E. The Effect of the Full Sample

How does including the full sample influence the parameter estimates? For convenience, we consider an often-encountered special case. We assume that the system is exactly identified and, moreover, that the 1st set of moment conditions (of length T) is sufficient to identify a subset θ_1 of the parameters. That is, θ_1 is exactly identified by those moment conditions available over the full sample. We will call the remaining parameters θ_2.

We first discuss the effect of using the full sample on estimation of θ_1. A natural conjecture is that the long, AM, and OI estimators all produce the same estimates of θ_1, namely, those found by setting g_{1T} equal to 0. This is clearly the case for the AM and long estimators. For the OI estimator, we use the argument in the previous section to decompose the FOCs as follows:

$$\begin{aligned}(12)\qquad \frac{1-\lambda}{\lambda} g_{1,(1-\lambda)T}^{\top} S_{11}^{-1} \frac{\partial g_{1,(1-\lambda)T}}{\partial \theta_1} + g_{1,\lambda T}^{\top} S_{11}^{-1} \frac{\partial g_{1,\lambda T}}{\partial \theta_1} \\ + \left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right)^{\top} \Sigma^{-1} \frac{\partial}{\partial \theta_1}\left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right) = 0,\end{aligned}$$

and

$$\begin{aligned}(13)\qquad \frac{1-\lambda}{\lambda} g_{1,(1-\lambda)T}^{\top} S_{11}^{-1} \frac{\partial g_{1,(1-\lambda)T}}{\partial \theta_2} + g_{1,\lambda T}^{\top} S_{11}^{-1} \frac{\partial g_{1,\lambda T}}{\partial \theta_2} \\ + \left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right)^{\top} \Sigma^{-1} \frac{\partial}{\partial \theta_2}\left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right) = 0.\end{aligned}$$

Under our stated assumptions, f_1 is only a function of θ_1, not of θ_2. Therefore, equation (13) reduces to

$$\left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right)^{\top} \Sigma^{-1} \frac{\partial}{\partial \theta_2} g_{2,\lambda T} = 0.$$

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

Furthermore, because the system is exactly identified, and because f_1 can identify θ_1, it follows that $(\partial/\partial\theta_2)g_{2,\lambda T}$ is invertible and that

$$g_{2,\lambda T} - B_{21}g_{1,\lambda T} = 0.$$

Therefore,

$$\frac{1-\lambda}{\lambda}g_{1,(1-\lambda)T}^{\top}S_{11}^{-1}\frac{\partial g_{1,(1-\lambda)T}}{\partial\theta_1} + g_{1,\lambda T}^{\top}S_{11}^{-1}\frac{\partial g_{1,\lambda T}}{\partial\theta_1} = 0 \tag{14}$$

is the FOC that identifies θ_1 for the OI estimator. As discussed in the previous section, this set of equations will in general not be satisfied by the value of θ_1 that sets $g_{1T} = 0$. To summarize, the AM estimator gives the same estimate for θ_1 as simply using the full sample. The OI estimator gives a possibly different estimate, one that depends on the point in time in which the 2nd series begins. Equation (14) is a weighted average of the moment conditions from the earlier and later parts of the sample, where the weights are proportional to the derivatives, and thus to the amount of information contained in each part of the sample.

We now ask how including the full sample might affect the standard errors of θ_1 and θ_2. We focus on asymptotic results, so the results for the OI and AM estimators will be the same. Now,

$$D_0 = \begin{bmatrix} D_{0,1} \\ D_{0,2} \end{bmatrix} = \begin{bmatrix} d_{11} & 0 \\ d_{21} & d_{22} \end{bmatrix},$$

where

$$d_{ij} = \frac{\partial f_i}{\partial\theta_j}, \qquad i = 1, 2,$$

and where d_{11} and d_{22} are invertible.

The inverse of D_0 takes the form

$$D_0^{-1} = \begin{bmatrix} d_{11}^{-1} & 0 \\ -d_{22}^{-1}d_{21}d_{11}^{-1} & d_{22}^{-1} \end{bmatrix}.$$

Therefore, the asymptotic variance of the short estimator of θ_1 equals $d_{11}^{-1}S_{11}(d_{11}^{-1})^{\top}$. Similarly, the asymptotic variance of the efficient estimators of θ_1 (the 1st block of $(D_0^{\top}\left(S^{\mathcal{A}}\right)^{-1}D_0)^{-1}$) can be written as

$$d_{11}^{-1}S_{11}^{\mathcal{A}}(d_{11}^{-1})^{\top} = \lambda d_{11}^{-1}S_{11}(d_{11}^{-1})^{\top}.$$

This shows that asymptotic standard errors for the estimates of θ_1 shrink by a factor of $1 - \sqrt{\lambda}$ when the efficient estimators are used rather than the short estimator. As shown above, the 2nd set of moment conditions f_2 has no effect on the estimation of θ_1 (see also Ahn and Schmidt (1995)).

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

The standard errors of the efficient estimators for θ_2 are determined by the 2nd diagonal block of $(D_0^\top \left(S^{\mathcal{A}}\right)^{-1} D_0)^{-1}$, which reduces to

$$\begin{aligned}(15)\qquad &\left(D_0^\top \left(S^{\mathcal{A}}\right)^{-1} D_0\right)_{22}^{-1} \\ &= d_{22}^{-1}\left[d_{21}d_{11}^{-1}S_{11}^{\mathcal{A}} - S_{21}^{\mathcal{A}}\right]\left(S_{11}^{\mathcal{A}}\right)^{-1}\left[d_{21}d_{11}^{-1}S_{11}^{\mathcal{A}} - S_{21}^{\mathcal{A}}\right]^\top \left(d_{22}^{-1}\right)^\top \\ &\quad + d_{22}^{-1}\left[S_{22}^{\mathcal{A}} - S_{21}^{\mathcal{A}}\left(S_{11}^{\mathcal{A}}\right)^{-1} S_{12}^{\mathcal{A}}\right]\left(d_{22}^{-1}\right)^\top.\end{aligned}$$

Thus, the variance for the 2nd set of parameters can be decomposed into 2 parts. The 1st part represents the effect of the 1st set of moment conditions on the 2nd set of parameters. The 2nd part represents the variance due only to the residual variance of the 2nd set of moment conditions: $S_{22}^{\mathcal{A}} - S_{21}^{\mathcal{A}}\left(S_{11}^{\mathcal{A}}\right)^{-1} S_{12}^{\mathcal{A}}$ is the variance-covariance matrix of the 2nd set of moment conditions conditional on the 1st set. The 2nd diagonal block of $\left(D_0^\top S^{-1} D_0\right)_{22}^{-1}$, which gives the standard errors for θ_2 under standard GMM, has an analogous decomposition:

$$\begin{aligned}(16)\qquad &\left(D_0^\top (S)^{-1} D_0\right)_{22}^{-1} \\ &= d_{22}^{-1}\left[d_{21}d_{11}^{-1}S_{11} - S_{21}\right](S_{11})^{-1}\left[d_{21}d_{11}^{-1}S_{11} - S_{21}\right]^\top \left(d_{22}^{-1}\right)^\top \\ &\quad + d_{22}^{-1}\left[S_{22} - S_{21}(S_{11})^{-1} S_{12}\right]\left(d_{22}^{-1}\right)^\top.\end{aligned}$$

Comparing equations (15) and (16) reveals the source of the efficiency gain. The 1st term in equation (15) is equal to λ multiplied by the 1st term in equation (16):

$$\begin{aligned}&\left[d_{21}d_{11}^{-1}S_{11}^{\mathcal{A}} - S_{21}^{\mathcal{A}}\right]\left(S_{11}^{\mathcal{A}}\right)^{-1}\left[d_{21}d_{11}^{-1}S_{11}^{\mathcal{A}} - S_{21}^{\mathcal{A}}\right]^\top \\ &\quad = \lambda\left[d_{21}d_{11}^{-1}S_{11} - S_{21}\right]S_{11}^{-1}\left[d_{21}d_{11}^{-1}S_{11} - S_{21}\right]^\top.\end{aligned}$$

However, the 2nd terms are the same, which is not surprising because they represent the variance of the 2nd set of moment conditions conditional on the value of the 1st set of moment conditions:

$$S_{22}^{\mathcal{A}} - S_{21}^{\mathcal{A}}\left(S_{11}^{\mathcal{A}}\right)^{-1} S_{12}^{\mathcal{A}} \quad = \quad S_{22} - S_{21}S_{11}^{-1}S_{12}.$$

The percentage decline in standard errors depends on the 1st term relative to the whole. For example, when the 2nd set of moments is perfectly correlated with the 1st set, the residual variance is 0,

$$(17)\qquad S_{22} - S_{21}S_{11}^{-1}S_{12} \quad = \quad 0,$$

and the standard errors for θ_2 also shrink by a factor of $1 - \sqrt{\lambda}$. At the other extreme, suppose that f_2 tells you nothing about θ_1; that is, $d_{21} = 0$ (θ_1 does not enter into f_2) and $S_{21} = S_{12}^\top = 0$ (the moment conditions are independent). Then, the inclusion of the longer series leads to no shrinkage in the asymptotic variance of θ_2.

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

Of course, even if the 2 sets of moment conditions are independent ($S_{21} = S_{12}^{\top} = 0$), the sampling variance of θ_2 may still fall because the sampling variance of θ_1 is reduced. As long as $d_{21} \neq 0$, the 1st term in equation (15) is nonzero, and there is an effect on the standard errors of θ_2. Similarly, even if there is no impact of θ_1 on the 2nd set of moment conditions ($d_{21} = 0$), the 1st set of moment conditions helps to estimate θ_2 if the covariance between the 2 moment conditions is nonzero.

Imposing the restriction $d_{21} = 0$ allows us to extend the previous discussion to the long estimator. In this exactly identified case, the long estimator $\hat{\theta}_T^{\mathcal{L}}$ solves

$$g_{1,T}(x, \hat{\theta}_T^{\mathcal{L}}) = 0 \quad \text{and} \tag{18}$$

$$g_{2,\lambda T}(x, \hat{\theta}_T^{\mathcal{L}}) = 0. \tag{19}$$

It follows that long estimates for θ_1 are asymptotically identical to the efficient estimates for these parameters (they are numerically identical to the AM estimates and asymptotically identical to the OI estimates). However, the long estimates of θ_2 are numerically identical to the short estimates, not to the efficient estimates.[9] This follows because the efficient estimates for θ_2 solve

$$g_{2,\lambda T}\left(x, \hat{\theta}_T^{\mathcal{A}}\right) + \hat{B}_{21,\lambda T}\left(g_{1,T}\left(x, \hat{\theta}_T^{\mathcal{A}}\right) - g_{1,\lambda T}\left(x, \hat{\theta}_T^{\mathcal{A}}\right)\right) = 0$$

rather than the expression in equation (19). This starkly illustrates the surprising role of the long sample in helping to estimate θ_2.[10] As we illustrate in the section that follows, this surprising result occurs because the separation uncovers the deviation of $g_{1,\lambda T}$ from 0. Because of the correlation between $g_{1,\lambda T}$ and $g_{2,\lambda T}$, the deviation of $g_{1,\lambda T}$ from 0 implies that $g_{2,\lambda T}$ is also likely to deviate from 0. The efficient estimators make use of this information to construct an estimator of the mean of $f_2(x, \theta)$ that improves on $g_{2,\lambda T}$.

III. Application to Predictive Regressions in International Data

This section applies our method to estimating predictive regressions for returns in international data. Reliable international data typically begin substantially later than U.S. data. At the same time, predictive regressions are often measured with noise, making it desirable to use as long a data series as possible. Our methods allow international data to be used at the same time as longer U.S. data.

A. Data

For the United States, we use the annual data of Shiller ((1989), chap. 26), which begin in 1871 and are updated through 2005. Stock returns, prices, and

[9] It is tempting to conclude that the lower variance for $\hat{\theta}_{1,T}^{\mathcal{L}}$ and the same variance for $\hat{\theta}_{2,T}^{\mathcal{L}}$ imply that the long estimator is more efficient than the short estimator. This is not the case, however. Efficiency requires that any linear combination of $\hat{\theta}_{1,T}^{\mathcal{L}}$ and $\hat{\theta}_{2,T}^{\mathcal{L}}$ have lower variance than the same linear combination of short estimates.

[10] The result is even more surprising given that the presence of the 2nd set of moment conditions does not affect estimation of the 1st set in this exactly identified case, as shown by Ahn and Schmidt (1995).

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

earnings are for the Standard & Poor's (S&P) 500 index. The predictor variable we use is the ratio of previous 10-year earnings to current stock price. We refer to this as the smoothed earnings-price (EP) ratio. This ratio is motivated by the present-value formula linking the EP ratio to returns; normalizing by smoothed earnings rather than earnings has the advantage that it eliminates short-term cyclical noise (see Campbell and Shiller (1988), Campbell and Thompson (2008)). The risk-free rate is the return on 6-month commercial paper purchased in January and rolled over in July. Because the first 10 years of the sample are used to construct the predictor variable, the data series of the predictor and returns begins in 1881 and ends in 2005. All variables are deflated using the consumer price index (CPI).

Data on international indices come from Ken French's Web site (http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html). The raw data on international indices come from Morgan Stanley's Capital International (MSCI) Perspectives. Fama and French (1989) discuss details of the construction of these data. The EAFE is a value-weighted index for Europe, Australia, and the Far East; within the EAFE, countries are added when data become available. For each country, returns are value-weighted, and countries are then weighted in proportion to their market values in the index. We also examine results for subindices. These are Asia-Pacific (Australia, Hong Kong, Japan, Malaysian, New Zealand, Singapore), Europe without the United Kingdom (Austria, Belgium, Switzerland, Germany, Spain, France, Italy, Netherlands), Europe with the United Kingdom (same as previous with Great Britain and Ireland), and Scandinavia (Denmark, Finland, Norway, Sweden). Data are monthly from 1975 to 2005. We compound the monthly dollar returns on these indices to create annual returns. We then subtract changes in the CPI from the Shiller (http://www.econ.yale.edu/∼shiller/data.htm) data set described above from the log of these returns to create real, continuously compounded returns.

We test the returns for nonnormality using the statistics proposed by Zhou (1993). We find that the skewness test is significant at the 5% level for the excess return on the U.S. index when using the long sample. While we are unable to reject normality for the series available over the short samples, this seems likely to be due to insufficient data and therefore a lack of test power. An advantage of basing our methods on GMM is that they are robust to nonnormalities in the data.

B. Applying the Estimators

Let $r_{1,t}$ denote the excess return on the long-history asset (the S&P 500 index) and $r_{2,t}$ the excess return on the short-history asset (the EAFE or one of the subindices). We estimate the predictive regressions

$$r_{1,t+1} = \alpha_1 + \beta_1 z_t + \epsilon_{1,t+1} \quad \text{and} \tag{20}$$

$$r_{2,t+1} = \alpha_2 + \beta_2 z_t + \epsilon_{2,t+1} \tag{21}$$

jointly for the S&P 500 index and international index excess returns, where z_t is the smoothed EP ratio on the S&P. It follows from the discussion in Section II.E that the source of the gain in estimating α_2 and β_2 will be the correlation between shocks to $r_{1,t}$ and $r_{2,t}$.

Define matrices

$$Z_T = \begin{bmatrix} 1 & z_0 \\ \vdots & \vdots \\ 1 & z_{T-1} \end{bmatrix}, \quad Z_{\lambda T} = \begin{bmatrix} 1 & z_{(1-\lambda)T} \\ \vdots & \vdots \\ 1 & z_{T-1} \end{bmatrix},$$
$$Z_{(1-\lambda)T} = \begin{bmatrix} 1 & z_0 \\ \vdots & \vdots \\ 1 & z_{(1-\lambda)T-1} \end{bmatrix},$$

and similarly,

$$R_{1,T} = \begin{bmatrix} 1 & r_{1,1} \\ \vdots & \vdots \\ 1 & r_{1,T} \end{bmatrix}, \quad R_{1,\lambda T} = \begin{bmatrix} 1 & r_{1,(1-\lambda)T+1} \\ \vdots & \vdots \\ 1 & r_{1,T} \end{bmatrix},$$
$$R_{1,(1-\lambda)T} = \begin{bmatrix} 1 & r_{1,1} \\ \vdots & \vdots \\ 1 & r_{1,(1-\lambda)T} \end{bmatrix},$$

and

$$R_{2,\lambda T} = \begin{bmatrix} 1 & r_{2,(1-\lambda)T+1} \\ \vdots & \vdots \\ 1 & r_{2,T} \end{bmatrix}.$$

Let $x_t = (r_{1,t}, r_{2,t}, z_{t-1})$ and $\theta_i = [\alpha_i, \beta_i]^\top$ for $i = 1, 2$. The moment conditions are estimated using

$$g_{1,T}(x,\theta) = \frac{1}{T} Z_T^\top \left(R_{1,T} - Z_T \theta_1\right), \tag{22}$$

$$g_{1,(1-\lambda)T}(x,\theta) = \frac{1}{(1-\lambda)T} Z_{(1-\lambda)T}^\top \left(R_{1,(1-\lambda)T} - Z_{(1-\lambda)T} \theta_1\right), \tag{23}$$

$$g_{1,\lambda T}(x,\theta) = \frac{1}{\lambda T} Z_{\lambda T}^\top \left(R_{1,\lambda T} - Z_{\lambda T} \theta_1\right), \quad \text{and} \tag{24}$$

$$g_{2,\lambda T}(x,\theta) = \frac{1}{\lambda T} Z_{\lambda T}^\top \left(R_{2,\lambda T} - Z_{\lambda T} \theta_2\right). \tag{25}$$

Setting equations (24) and (25) to 0 gives the short estimator (which, in this context, is the same as ordinary least squares (OLS) regression over the 1975–2005 period). The AM estimator requires an estimate of B_{21}. In this context, this is a 2×2 matrix of coefficients of a multivariate regression of errors from the short-history moment conditions on errors from the long-history moment conditions. To calculate this regression, we first estimate the system using the short method and evaluate f_1 and f_2 at the short estimates. We then have a sequence of observations on the errors for the moment conditions from 1975–2005. Regressing the errors that correspond to f_2 on the errors that correspond to f_1 yields the 4 entries of the

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

matrix $\hat{B}_{21,\lambda T}$. Given $\hat{B}_{21,\lambda T}$, the AM estimator is the solution to equations defined by setting equations (22) and

$$g_{2,\lambda T}(x,\theta) + \hat{B}_{21,\lambda T}\left(g_{1,T}(x,\theta) - g_{1,\lambda T}(x,\theta)\right)$$

to 0. For the long-history asset, this corresponds to OLS regression over the 1881–2005 period. For the short-history asset, this corresponds to a regression over the later part of the sample period, plus an adjustment that, as we show later, can be quite substantial.

While the AM and short estimators are exactly identified, the OI estimator is not, as its name suggests. Moment conditions for the OI estimator are given by equations (23), (24), and (25). The weighting matrix is the inverse of an estimate of $S^{\mathcal{I}}$, which can be calculated based on submatrices of an estimate of S as in equation (9). Below, we explain how we estimate S.

Obtaining standard errors requires an estimate of the derivative matrix D_0 and an estimate for the variance matrix S. The results of Section II require only that we choose estimators that are consistent. However, it is most in the spirit of our approach to use the full data in constructing $\hat{D}_T$ and $\hat{S}_T$. A consistent estimator of the derivative matrix D_0 that makes use of the full sample is

$$\hat{D}_T \quad = \quad I_2 \otimes \frac{1}{T} Z_T^{\top} Z_T,$$

where I_2 is the 2×2 identity matrix.

To construct an estimate for $\hat{S}_T$ that makes use of the full sample, we apply the procedure outlined in Stambaugh (1997) for constructing a positive-definite variance-covariance matrix for data of unequal lengths. Let $\hat{S}_{11,T}$ denote the White (1980) estimate of S_{11} using the full sample. Let $\hat{B}_{21,\lambda T}$ and $\hat{\Sigma}$ be estimates of B_{21} and Σ, respectively, over the sample period that is available for all moment conditions (these matrices can be constructed by regressing the short-sample moment conditions on the long-sample moment conditions over the period in which the samples overlap). We take the following as the estimator of S:

$$\hat{S}_T \quad = \quad \begin{bmatrix} \hat{S}_{11,T} & \hat{S}_{11,T}\hat{B}_{21,\lambda T}^{\top} \\ \hat{B}_{21,\lambda T}\hat{S}_{11,T} & \hat{\Sigma} + \hat{B}_{21,\lambda T}\hat{S}_{11,T}\hat{B}_{21,\lambda T}^{\top} \end{bmatrix}.$$

C. Results

Prior to reporting the results for the predictive regressions, we briefly discuss the estimates of the mean returns implied by our methods. The implementation for this estimation is very similar to, and less complicated than, the implementation of the predictability estimation described previously. Note that our estimators take the same form as those of Stambaugh (1997) in the setting of estimating sample means.[11]

[11] In the IID normal setting of Stambaugh (1997), the estimate of the matrix B_{21} is comprised of regression coefficients of the short-history series on the long-history series. This estimate will also be consistent under more general distributional assumptions, including conditional heteroskedasticity. However, allowing for serial correlation would require a different estimate of B_{21} (which could be derived from submatrices of S). For the current application (which uses annual nonoverlapping observations), allowing for serial correlation is unlikely to have a large effect.

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

The first 2 columns of Table 1 report results and standard errors for the short estimator; the second 2 columns report results and standard errors for the AM and OI estimators. Because these are numerically equivalent in the setting of estimating means, we refer to them jointly as the efficient estimator. As the columns for short show, the sample mean for excess returns on the S&P 500 index in the 1975–2005 period was 5.64% with a standard error of 3.16%. The sample mean over the full period is 3.96%, as the efficient column shows. It is also estimated much more precisely: The standard error falls from 3.16% to 1.55%.

TABLE 1

Mean Excess Returns on International Indices

Means are estimated for excess returns on international indices. Returns are annual, continuously compounded, and in excess of the risk-free rate. United States refers to returns on the S&P 500 index; EAFE refers to returns on an index for Europe, Asia, and the Far East; Asia-Pacific, Europe, Europe without the United Kingdom, and Scandinavia are subindices of EAFE. "Short" denotes estimates obtained using standard GMM; "Efficient" denotes estimates obtained using the adjusted-moment method or overidentified method, which are numerically identical in this application. Results for the "Long" estimator (not shown) are equal to the results for Efficient for the United States and the results for Short for all other assets. Standard errors are computed using efficient estimates and are robust to conditional heteroskedasticity. Data for the United States span the 1881–2005 period; data for the other indices span the 1975–2005 period. Means and standard errors are reported in percentage terms.

	Short		Efficient	
Series	Mean	SE	Mean	SE
United States	5.64	3.16	3.96	1.55
EAFE	5.29	3.79	3.95	3.09
Asia-Pacific	3.52	4.80	2.43	4.46
Europe	6.46	3.68	4.92	2.68
Europe without the United Kingdom	5.40	4.17	3.69	3.09
Scandinavia	7.00	4.58	5.27	3.60

Introducing data from 1881–1975 also results in more precise estimates of the excess return on short-history assets. For the EAFE index, the standard errors falls from 3.79 for the short method to 3.09 for the efficient methods (the correlation between the S&P 500 index and the EAFE portfolios is 0.67). It is this correlation that leads to the reduced standard errors. In particular, the fact that the mean return for the S&P 500 index was somewhat higher in the later part of the sample than the earlier part implies that shocks during the 1975–2005 period had a positive mean, on average. The efficient estimators therefore adjust the mean excess return on the EAFE downward.

Estimation for the subindices also improves, more dramatically for the European indices and less so for the Asia-Pacific index. While the correlation between the Asia-Pacific index and the S&P 500 index is 0.43, the correlation between the European indices and the S&P 500 index exceeds 0.70. As shown in Section II.E, higher correlations between the moment conditions lead to greater improvement for the short-history asset.

Table 2 reports results of estimating the predictive regressions. We present the coefficients on the predictive variable, the standard error on this variable, and the R^2, computed as the sample variance of the predicted return divided by the sample variance of the total return.[12] In the short sample, the point estimates for

[12]For each series, we use the data that are available in computing the R^2.

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

all the portfolios are positive but insignificant: The *t*-statistics are below 1 for all portfolios. The R^2 values are also small (e.g., 0.6% for the S&P 500 index). There is more evidence for predictability in the longer sample. For the S&P 500 index, the AM method leads to an estimated coefficient of 0.093 with a standard error of 0.038 and an R^2 of 4.2%. The OI method leads to an estimated coefficient of 0.065 and an R^2 of 2.1%.

TABLE 2

Predictive Regressions for Excess Returns on International Indices

Predictive regressions are estimated in annual data for excess returns on international indices. Table 2 reports the estimate of the coefficient on the predictor variable (Coef.), the standard error (SE) on this coefficient, and the R^2 from the regression. The predictive variable is the log of the smoothed earnings-price (EP) ratio. Returns are annual, continuously compounded, and in excess of the risk-free rate. United States refers to returns on the S&P 500 index; EAFE refers to returns on an index for Europe, Asia, and the Far East; Asia-Pacific, Europe, Europe without the United Kingdom, and Scandinavia are subindices of EAFE. "Short" denotes standard GMM; "AM" denotes the adjusted-moment method; "OI" denotes the overidentified method. Results for the "Long" estimator (not shown) are equal to the results for AM for the United States and the results for Short for all other assets. Standard errors are computed using AM estimates and are robust to conditional heteroskedasticity. Data for the United States span the 1881–2005 period; data for the other indices span the 1975–2005 period.

	Short			AM			OI		
Series	Coef.	SE	R^2	Coef.	SE	R^2	Coef.	SE	R^2
United States	0.036	0.077	0.006	0.093	0.038	0.042	0.065	0.038	0.021
EAFE	0.073	0.118	0.038	0.128	0.097	0.117	0.101	0.097	0.073
Asia-Pacific	0.121	0.185	0.059	0.170	0.175	0.117	0.147	0.175	0.086
Europe	0.038	0.103	0.012	0.097	0.076	0.077	0.068	0.076	0.037
Europe without the United Kingdom	0.018	0.114	0.002	0.080	0.088	0.040	0.050	0.088	0.016
Scandinavia	0.015	0.164	0.001	0.093	0.139	0.044	0.058	0.139	0.017

Well-known theoretical results demonstrate that the OLS regression produces the best-fitting regression estimates, given a single series of data. In this predictive regression setting, OLS is equivalent to our short method. Our results show that one can improve on OLS if one has data on a series for which a longer sample is available. Indeed, this application shows that including the earlier period of the sample has a substantial impact on the estimation for the EAFE and other short-history assets. The AM method leads to an estimated coefficient of 0.128, as opposed to 0.073 using the short method. Moreover, the standard error on this estimate falls from 0.118 for the short method to 0.097 for the AM method. The implied R^2 is 12%, up from 3.8% when the short method is used. Results for the OI estimator are similar: The coefficient is 0.101 with an R^2 of 7.3%. Similar effects are present for subindices of the EAFE.

Essentially, our methods exploit the information that the evidence for predictability in the United States is stronger in the full sample than over the latter half. Under the assumption of stationarity, shocks to returns and to the predictive variable over the latter half of the sample must be such that the predictive coefficient estimated over this data range is too small. Because of the correlation between international returns and U.S. returns, OLS regression for international returns over the same data range would also be likely to understate the extent of predictability. Thus, the efficient estimators adjust the OLS (short) estimate upward. The resulting estimates have less noise, as represented by the smaller standard errors in Table 2. While the annual data sample is still too short to comment on statistical significance (except for the United States), it is clear that the

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

economic significance of predictability is a great deal higher when the early part of the sample is included.

D. Efficient versus Inefficient Use of the Full Data

Finally, we use this application to contrast the efficient estimators with the long estimator, which uses the full sample but in an inefficient way. In so doing, we illustrate the theoretical results presented at the end of Section II.E.

The results in Section II.E imply that the long estimate for the predictability coefficient β_1 is numerically identical to the AM estimate and asymptotically equal to the OI estimate of this coefficient. Both the long and the AM estimates are equal to the value obtained from an OLS regression of S&P 500 index returns on the predictor variable over the full sample of data. In contrast, the long estimate for β_2 is not equal to the AM or OI estimate. Rather it is equal to the short estimate of 0.073 (in the case of the EAFE), which is the value obtained from an OLS regression of EAFE returns on the predictor variable over the 1975–2005 period. The AM and OI estimates are substantially higher, at 0.128 and 0.101, respectively.

The efficient estimators differ from the long estimator in that they divide the data on the S&P 500 index into 2 moment conditions, one defined over 1881–1975 and one defined over 1975–2005. Dividing the data in this way does not alter the estimate (asymptotically) of the predictive coefficient for the S&P 500 index. However, this division does create more information: It uncovers the fact that there is less predictability in the S&P 500 index over the 1975–2005 period than over the full period. As discussed in the previous section, our efficient estimators correctly incorporate this information to improve estimation of β_2.

IV. Monte Carlo Analysis

In the previous sections, we introduced 2 methods of implementing GMM with unequal sample lengths and showed that these methods lead to improvements in asymptotic efficiency. More precise estimates can be obtained both for assets with data available for the full period and, more surprisingly, for assets with data available for the later part of the period. A natural question is whether these gains are present for the small-sample distribution of the estimates.

In this section, we answer this question using a Monte Carlo experiment modeled after the estimation of predictability. It is particularly useful to investigate this case in a small-sample setting, as it is well known that asymptotic properties can fail noticeably for predictive regressions when the regressors are persistent (e.g., Cavanagh, Elliott, and Stock (1995), Nelson and Kim (1993), and Stambaugh (1999)).[13]

[13] In contrast, the small-sample standard errors for the means are nearly identical to the asymptotic ones.

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

We simulate from the system of equations (20)–(21) using the AM regression coefficients to determine the data-generating process. We augment this system with an autoregression for the log of the smoothed EP ratio z_t:

$$z_{t+1} = \rho_0 + \rho_1 z_t + \epsilon_{z,t+1}. \tag{26}$$

We assume that the shocks are IID and normally distributed. Estimates for ρ_0 and ρ_1 are obtained using the full data set and are equal to -0.294 and 0.892, respectively. For each index, we estimate the variance-covariance matrix of errors from equations (20), (21), and (26) using the method described above. Table 3 reports the variances and correlations. The contemporaneous correlation between innovations to z_t and to S&P 500 index returns $r_{1,t}$ is -0.91: This large negative value is due to the fact that price is in the denominator of the smoothed EP ratio. Innovations to z_t are also negatively correlated with innovations to returns on the short-history assets. For example, the correlation with innovations to returns on the EAFE is -0.515. Innovations to returns on the S&P 500 index are also highly correlated with innovations to international returns: This correlation is 0.65 for the EAFE and over 0.70 for the European subindices. Therefore, it is reasonable to expect that incorporating the earlier data period will affect the precision of the estimates for the short-history assets.

TABLE 3

Monte Carlo Parameters for Predictive Regressions

Standard deviations and correlations are estimated in annual data for errors from predictive regressions for use in constructing simulated data. Right-hand-side variables are the U.S. return, an international index return (EAFE or a subindex of the EAFE), and the predictor variable, the log of the smoothed earnings-price (EP) ratio. Returns are annual, continuously compounded, and in excess of the risk-free rate. United States refers to returns on the S&P 500 index; EAFE refers to returns on an index for Europe, Asia, and the Far East; Asia-Pacific, Europe, Europe without the United Kingdom, and Scandinavia are subindices of EAFE. Data for the United States span the 1881–2005 period; data for the other indices span the 1975–2005 period. Predictive coefficients for returns are reported in Table 2 under the heading "AM." The coefficient for the predictor variable is 0.89. Data on international index returns are annual and span the 1975–2005 period. Data on U.S. returns are annual and span the 1881–2005 period.

Series	Standard Deviation	Correlation with log(EP)	Correlation with the United States
log(EP)	0.179		
United States	0.170	−0.912	
EAFE	0.207	−0.515	0.653
Asia-Pacific	0.259	−0.309	0.409
Europe	0.205	−0.616	0.775
Europe without the United Kingdom	0.229	−0.666	0.769
Scandinavia	0.255	−0.578	0.710

For each international index, we simulate 50,000 samples of returns on the S&P 500 index, values for the predictor variable, and returns on that index. The sample length for the S&P 500 index (the long-history asset) and the predictor variable is 124 years; the sample length for the short-history asset is 30 years. We repeat the short, AM, and OI estimations in each. We report both the standard deviations of the estimates (Table 4) and the bias (Table 5, measured as the difference between the mean estimated and the true coefficient).

Table 4 shows that the asymptotic efficiency gains discussed in Section II.C also appear in finite samples. For the long-history asset, the standard deviation of the predictive coefficient falls from 0.133 to 0.48 for both the AM and OI methods.

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

TABLE 4

Predictive Regressions in Repeated Samples

In Table 4, 50,000 samples of returns are simulated assuming joint normality of excess returns and the predictor variable. This table reports standard deviations of estimates of the predictive coefficient. In each set of samples, there is a long-history asset calibrated to the S&P 500 index and a short-history asset calibrated to the EAFE or subindex. The long-history asset has 124 years of data; the short-history asset has 30 years of data. Predictive coefficients are reported in Table 2 under the heading "AM" and standard deviations and correlations of errors in Table 3. The predictor variable has an autocorrelation coefficient of 0.89. "Short" denotes standard GMM; "AM" denotes the adjusted-moment method; and "OI" denotes the overidentified method.

Series	Short	AM	OI
United States	0.133	0.048	0.048
EAFE	0.156	0.134	0.135
Asia-Pacific	0.193	0.196	0.197
Europe	0.156	0.116	0.116
Europe without the United Kingdom	0.175	0.130	0.131
Scandinavia	0.194	0.156	0.157

TABLE 5

Bias in Predictive Coefficients

In Table 5, 50,000 samples of returns are simulated assuming joint normality of excess returns and the predictor variable. This table reports the difference between the estimated mean of the predictive coefficient and the true mean. In each set of samples, there is a long-history asset calibrated to the S&P 500 index and a short-history asset calibrated to the EAFE or subindex. The long-history asset has 124 years of data; the short-history asset has 30 years of data. Predictive coefficients are reported in Table 2 and standard deviations and correlations of errors in Table 3. The predictor variable has an autocorrelation coefficient of 0.89. "Short" denotes standard GMM; "AM" denotes the adjusted-moment method; and "OI" denotes the overidentified method.

Series	Short	AM	OI
United States	0.120	0.028	0.015
EAFE	0.083	0.008	−0.003
Asia-Pacific	0.063	0.004	−0.005
Europe	0.098	0.011	−0.002
Europe without the United Kingdom	0.119	0.023	0.008
Scandinavia	0.115	0.015	0.001

For the short-history assets, there is improvement in all but one case (when this asset is calibrated to the Asia-Pacific index). When the asset is calibrated to the EAFE, for example, the short method delivers a standard deviation of 0.156. The AM method delivers a standard deviation of 0.134, and the OI method delivers a standard deviation of 0.135. When the asset is calibrated to the European index, the standard deviation of the estimate falls from 0.156 to 0.116 for both the AM and OI methods. In each case, the improvement in the small-sample standard errors is of the same magnitude as the asymptotic standard errors.

The theory presented in Section II is silent on the subject of bias. However, it is of interest to compare the performance of the efficient estimators to the standard estimator in this regard. It is not surprising that the bias is reduced under the AM and OI estimators for the long-history asset. Because these estimators are consistent, introducing the longer data should result in a lower bias. Indeed, while the bias for a long-history asset is 0.120 under the short estimator, it is 0.028 under the AM estimator and 0.015 under the OI estimator.

More surprising is the reduction in the bias for the short-history assets. When the short-history asset is calibrated to the EAFE, the bias is 0.083 under the short estimator. Under both the AM and OI estimators, the bias is about equal to 0

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

(it is, in fact, very slightly negative for the OI estimator). Similar results are apparent when the short-history asset is calibrated to the other indices.

While a full investigation is outside the scope of this study, the form of the estimators gives some insight into the source of the bias reduction. Both the AM and OI estimators use the fact that the standard GMM estimates for the long-history asset differ between the full sample and the later part of the sample. Because standard GMM is consistent, some of this difference arises from the bias in the coefficient (because the bias, on average, will be worse in the later part of the sample than in the full sample). Given that the moment conditions are correlated, the bias in estimates for the long-history asset (measured over the later part of the sample) is also likely to appear for the short-history asset (measured over the same period). The estimators can then use the information on the bias for the long-history asset to correct the bias in the short-history asset.

V. Extensions

In this section, we briefly outline how our estimators can be extended to more general patterns of missing data. We focus on the OI estimator that has a direct extension.[14]

Consider intervals of the data defined by points in time where at least one sample moment starts or ends. Say these points in time divide the sample into disjoint intervals $1, \ldots, n$. Let λ_1 denote the ratio of the length of the 1st region to the length of the entire sample, λ_2 denote the ratio of the length of the 2nd region to the length of the entire sample, etc. Note that $\sum_{i=1}^{n} \lambda_i = 1$. Define points $t_1, \ldots, t_n$ so that the 1st data segment begins at $t_1 + 1$, the 2nd data segment at $t_2 + 1$, etc. Then,

$$g_{\lambda_j T}(\theta) \quad = \quad \frac{1}{\lambda_j T} \sum_{t=t_j+1}^{t_j+\lambda_j T} f(x_t, \theta), \qquad j = 1, \ldots, n.$$

For the case described in Section II, the 1st segment consist of points 1 to $(1-\lambda)T$, while the 2nd segment consists of points $(1 - \lambda)T + 1$ to T. We adopt the same notational convention as in Section II: Here, $\lambda_j t$ will refer to the length of the segment between $t_j + 1$ and $t_j + \lambda_j T$, and the segment itself.

Let ϕ_i denote the set of moment conditions that are observed in data segment λ_i, and let π_i denote the number of such moment conditions. Define

$$f_{\phi_j}(x_t, \theta) \quad = \quad \left[f_{i_1}(x_t, \theta), \ldots, f_{i_{\pi_j}}(x_t, \theta) \right]^{\top},$$

where $\{i_1, \ldots, i_{\pi_j}\} \in \phi_j$ and $i_1 < \cdots < i_{\pi_j}$. Then, f_{ϕ_j} are the components of f observed over the segment $\lambda_j T$. Define the $\pi_j \times 1$ vector

$$g_{\phi_j, \lambda_j T}(\theta) \quad = \quad \frac{1}{\lambda_j T} \sum_{t=t_j+1}^{t_j+\lambda_j T} f_{\phi_j}(x_t, \theta)$$

[14]The extension for the AM estimator, as well as examples for various patterns of missing data, can be found in Lynch and Wachter (2004).

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

and the $\pi_j \times \pi_j$ matrices

$$\begin{aligned} R_{\phi_j}(\tau) &= \mathrm{E}\left[f_{\phi_j}(x_0, \theta_0) f_{\phi_j}(x_{-\tau}, \theta_0)^\top\right] \qquad \text{and} \\ S_{\phi_j} &= \sum_{\tau=-\infty}^{\infty} R_{\phi_j}(\tau). \end{aligned}$$

Define

$$h_T^{\mathcal{I}_n}(\theta) = \left[g_{\phi_1,\lambda_1 T}(\theta)^\top,\ g_{\phi_2,\lambda_2 T}(\theta)^\top, \ldots,\ g_{\phi_n,\lambda_n T}(\theta)^\top\right]^\top. \tag{27}$$

The $\mathcal{I}_n$ superscript refers to the fact that these are moment conditions for the OI estimator, and that there are n nonoverlapping intervals. The T subscript refers to the fact that the data length is T.[15] Let $S^{\mathcal{I}_n}$ be the variance-covariance matrix and $\hat{S}_T^{\mathcal{I}_n}$ be an estimate of $S^{\mathcal{I}_n}$. We can then define the extended OI estimator as

$$\hat{\theta}_T^{\mathcal{I}_n} = \underset{\theta}{\operatorname{argmin}}\ h_T^{\mathcal{I}_n}(\theta)^\top \left(\hat{S}_T^{\mathcal{I}_n}\right)^{-1} h_T^{\mathcal{I}_n}(\theta). \tag{28}$$

The same consistency and asymptotic normality results go through for the extended OI estimator as for the original OI estimator. Moreover,

$$S^{\mathcal{I}_n} = \begin{bmatrix} \frac{1}{\lambda_1} S_{\phi_1} & 0 & \ldots & 0 \\ 0 & \frac{1}{\lambda_2} S_{\phi_2} & \ldots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \ldots & \frac{1}{\lambda_n} S_{\phi_n} \end{bmatrix}.$$

We now state a result analogous to Theorem 1. That theorem showed that including the data segment for which some data were missing improved efficiency relative to standard GMM. Here, we show that including a new data segment improves efficiency relative to the estimator that includes all data but this segment. Without loss of generality, we consider the full OI estimator relative to the OI estimator defined over the first $n-1$ blocks of data. The proof (available from the authors) is similar to that of Theorem 1.

Theorem 3. Assume the OI estimator $\hat{\theta}_T^{\mathcal{I}_n}$ is defined by equation (28). Then, this estimator is asymptotically more efficient than $\hat{\theta}_{(1-\lambda_n)T}^{\mathcal{I}_{n-1}}$, the analogous estimator that is defined over the first $n-1$ blocks of data.

VI. Conclusion

This paper introduces 2 estimators that extend the GMM of Hansen (1982) to cases where moment conditions are observed over different sample periods. Most estimation procedures, when confronted with data series that are of

[15] This notation does not, of course, completely define the OI estimator. For that, one would need the points at which the data intervals begin, $t_1, \ldots, t_n$. These points, in turn, depend on $\lambda_1, \ldots, \lambda_n$ and T.

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

unequal length, require the researcher to truncate the data so that all series are observed over the same interval. This paper provides an alternative that allows the researcher to use all the data available for each moment condition.

Under assumptions of mixing and stationarity, we demonstrate consistency, asymptotic normality, and efficiency over both standard GMM and an extension of GMM that uses the full data in a naive way. Our base case assumes that the 2 series had the same end date but different start dates. We then generalize our results to cases where the start date and the end date may differ over multiple series. In all cases, using all the data produces more efficient estimates. Moreover, the impact of including the nonoverlapping portion of the data is not limited to estimating moment conditions that are available for the full period. As long as there is some interaction between the moment conditions observed over the full period and those observed over the shorter period, there will be an impact on all the parameters. This interaction can be through covariances between the moment conditions, or through the fact that some parameters appear in both the moment conditions available over the full sample and those available over the shorter sample. In an application of our methods to estimation of conditional and unconditional means in international data, we show that this impact can be large.

Our estimators are as straightforward to implement as standard GMM and have intuitive interpretations. The AM estimator calculates moments using all the data available for each series and then adjusts the moments available over the shorter series using coefficients from a regression of the short-sample moment conditions on the full-sample moment conditions. The OI estimator uses the nonoverlapping data to form additional moment conditions. These estimators are equivalent asymptotically and superior to standard GMM, but they differ in finite samples. We leave the question of which estimator has superior finite-sample properties to future work.

Appendix A. Deriving the Asymptotic Distribution

Let $\{x_t\}_{t=-\infty}^{\infty}$ denote a p-component stochastic process defined over a probability space $(\Omega, \mathcal{F}, P)$. Let $\mathcal{F}_a^b \equiv \sigma(x_t\,;\, a \leq t \leq b)$, the Borel σ-algebra of events generated by $x_a, \ldots, x_b$. Consider a function $f : \mathbf{R}^p \times \Theta \to \mathbf{R}^l$ for Θ, a compact subset of $\mathbf{R}^q$. For a precise statement of the required conditions on $\{x_t\}$ and f, see White and Domowitz (1984).

Define

$$w_t \quad = \quad f(x_t, \theta_0).$$

It is useful to slightly generalize the notation of Section II. Let

$$\begin{aligned} g_T(\theta) &= \frac{1}{T}\sum_{t=1}^{T} f(x_t, \theta), \\ g_{(1-\lambda)T}(\theta) &= \frac{1}{(1-\lambda)T}\sum_{t=1}^{(1-\lambda)T} f(x_t, \theta), \qquad \text{and} \\ g_{\lambda T}(\theta) &= \frac{1}{\lambda T}\sum_{t=(1-\lambda)T+1}^{T} f(x_t, \theta). \end{aligned}$$

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

The following lemma states that partial sums taken over disjoint intervals are asymptotically independent.

Lemma A.1. Let $F \in \mathcal{F}_{-\infty}^{0}$. Let μ be a $1 \times l$ vector, and let c be a scalar. Let

$$P_g \quad = \quad \lim_{T\to\infty} P\left(\sqrt{T}\mu g_T(\theta_0) < c\right).$$

Then,

$$\lim_{T\to\infty} P\left(\left(\sqrt{T}\mu g_T(\theta_0) < c\right) F\right) \quad = \quad P_g P(F).$$

Proof. For any integer T,

$$\sqrt{T} g_T(\theta_0) \quad = \quad \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor\sqrt{T}\rfloor} w_t + \frac{1}{\sqrt{T}} \sum_{t=\lfloor\sqrt{T}\rfloor+1}^{T} w_t,$$

where $\lfloor\sqrt{T}\rfloor$ is the largest integer less than the square root of T. Then,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor\sqrt{T}\rfloor} w_t \quad = \quad \frac{\lfloor\sqrt{T}\rfloor}{\sqrt{T}} \frac{1}{\lfloor\sqrt{T}\rfloor} \sum_{t=1}^{\lfloor\sqrt{T}\rfloor} w_t \quad \underset{\text{a.s.}}{\to} \quad 0$$

as $T \to \infty$, by Theorem 2.3 of White and Domowitz (1984). Because

$$\frac{1}{\sqrt{T}} \sum_{t=\lfloor\sqrt{T}\rfloor+1}^{T} w_t \ \in \ \mathcal{F}_{\sqrt{T}}^{\infty},$$

$$\left| P\left(\left[\frac{1}{\sqrt{T}} \sum_{t=\lfloor\sqrt{T}\rfloor+1}^{T} \mu w_t < c\right] F\right) - P\left(\frac{1}{\sqrt{T}} \sum_{t=\lfloor\sqrt{T}\rfloor+1}^{T} \mu w_t < c\right) P(F) \right| \quad < \quad \alpha(\sqrt{T}).$$

White and Domowitz show that w_t is α-mixing. Therefore, $\alpha(\sqrt{T})$ goes to 0 as $T \to \infty$. By the Slutsky theorem,

$$\begin{aligned}
\lim_{T\to\infty} P\left(\left(\sqrt{T}\mu g_T(\theta_0) < c\right) F\right) \quad &= \quad \lim_{T\to\infty} P\left(\left[\frac{1}{\sqrt{T}} \sum_{t=\lfloor\sqrt{T}\rfloor+1}^{T} \mu w_t < c\right] F\right) \\
&= \quad \lim_{T\to\infty} P\left(\frac{1}{\sqrt{T}} \sum_{t=\lfloor\sqrt{T}\rfloor+1}^{T} \mu w_t < c\right) P(F) \\
&= \quad P_g P(F),
\end{aligned}$$

where the 2nd line follows because w_t is α-mixing, and the last line follows from a 2nd application of the Slutsky theorem. □

Lemma A.2. As $T \to \infty$,

$$\sqrt{T}\begin{bmatrix} \sqrt{(1-\lambda)} g_{(1-\lambda)T}(\theta_0) \\ \sqrt{\lambda} g_{\lambda T}(\theta_0) \end{bmatrix} \quad \underset{\text{d}}{\to} \quad N\left(0, \begin{bmatrix} S & 0 \\ 0 & S \end{bmatrix}\right).$$

Proof. White and Domowitz ((1984), thm. 2.4) show

$$\text{(A-1)} \qquad \sqrt{(1-\lambda)T} g_{(1-\lambda)T}(\theta_0) \quad \underset{\text{d}}{\to} \quad N(0, S) \qquad \text{and}$$

$$\text{(A-2)} \qquad \sqrt{\lambda T} g_{\lambda T}(\theta_0) \quad \underset{\text{d}}{\to} \quad N(0, S).$$

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

Stationarity of x_t implies that random variables $f(x_{-(1-\lambda)T+1},\theta),\ldots,f(x_{\lambda T},\theta)$ have the same joint distribution as random variables $f(x_1,\theta),\ldots,f(x_T,\theta)$. Thus, partial sums taken over $f(x_{-(1-\lambda)T+1},\theta),\ldots,f(x_{\lambda T},\theta)$ have the same distribution as the corresponding partial sums taken over $f(x_1,\theta),\ldots,f(x_T,\theta)$. Define

$$\begin{aligned}\tilde{g}_{\lambda T}(\theta) &= \frac{1}{\lambda T}\sum_{t=1}^{\lambda T} f(x_t,\theta),\\ \tilde{g}_{(1-\lambda)T}(\theta) &= \frac{1}{(1-\lambda)T}\sum_{t=0}^{(1-\lambda)T-1} f(x_{-t},\theta).\end{aligned}$$

It suffices to prove the results for $\tilde{g}_{\lambda T}$ and $\tilde{g}_{(1-\lambda)T}$.

Let $\mathcal{N}(c)$ denote the cumulative distribution function of the standard normal distribution evaluated at c. Let μ_1 and μ_2 be $1\times l$ vectors such that $\mu_1\mu_1^\top=\mu_2\mu_2^\top=1$. By Lemma A.1,

$$\begin{aligned}&\lim_{T\to\infty} P\left(\mu_1\sqrt{(1-\lambda)T}S^{-1}\tilde{g}_{(1-\lambda)T}(\theta_0)<c_1,\mu_2\sqrt{\lambda T}S^{-1}\tilde{g}_{\lambda T}(\theta_0)<c_2\right)\\ &\quad= \lim_{T\to\infty} P\left(\mu_1\sqrt{(1-\lambda)T}S^{-1}g_{(1-\lambda)T}(\theta_0)<c_1\right)\lim_{T\to\infty}\left(\mu_2\sqrt{\lambda T}S^{-1}\tilde{g}_{\lambda T}(\theta_0)<c_2\right)\\ &\quad= \mathcal{N}(c_1)\mathcal{N}(c_2),\end{aligned}$$

for scalars a and b. This shows $\tilde{g}_{\lambda T}(\theta_0)$ and $\tilde{g}_{(1-\lambda)T}(\theta_0)$ are asymptotically independent, and therefore that $g_{\lambda T}(\theta_0)$ and $g_{(1-\lambda)T}(\theta_0)$ are asymptotically independent. The result follows from equations (A-1) and (A-2). □

For notational convenience, it is useful to define functions h^k that give the moment conditions for each estimator. Assume that the weighting matrices for each estimator converge to a positive-definite weighting matrix. Now,

$$\begin{aligned}h_T^{\mathcal{S}}(\theta) &= \left[g_{1,\lambda T}(\theta)^\top,\ g_{2,\lambda T}(\theta)^\top\right]^\top,\\ h_T^{\mathcal{L}}(\theta) &= \left[g_{1,T}(\theta)^\top,\ g_{2,\lambda T}(\theta)^\top\right]^\top,\\ h_T^{\mathcal{A}}(\theta) &= \left[g_{1,T}(\theta)^\top,\ \left(g_{2,\lambda T}(\theta)+\hat{B}_{21,\lambda T}(1-\lambda)(g_{1,(1-\lambda)T}(\theta)-g_{1,\lambda T}(\theta))\right)^\top\right]^\top, \quad \text{and}\\ h_T^{\mathcal{I}}(\theta) &= \left[g_{1,(1-\lambda)T}(\theta)^\top,\ g_{1,\lambda T}(\theta)^\top,\ g_{2,\lambda T}(\theta)^\top\right]^\top,\end{aligned}$$

where for $k\in\mathcal{S},\mathcal{L},\mathcal{A},\mathcal{I}$, given a weighting matrix W_T^k, let

$$\theta^k = \underset{\theta}{\operatorname{argmin}}\ h_T^k(\theta)^\top W_T^k h_T^k(\theta).$$

Theorem A.1. As $T\to\infty$,

$$\sqrt{\lambda T}h_T^k(\theta_0) \underset{\mathrm{d}}{\to} N(0,S^k),$$

where $S^{\mathcal{S}}=S$, and $S^{\mathcal{L}}$, $S^{\mathcal{A}}$, and $S^{\mathcal{I}}$ are defined in equations (7)–(9).

Proof. The result for the short estimator follows directly from Lemma A.2. To illustrate the proof for the remaining matrices, we derive equation (8); the proofs of equations (7) and (9) are similar. In what follows, the argument θ_0 is suppressed, and convergence is in the sense of almost surely.

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

Stationarity implies that $S_{11}^{\mathcal{A}} = \lambda S_{11}$. By Lemma A.2,

$$\begin{aligned}(\text{A-3}) \qquad \lim_{T\to\infty} \mathrm{E}\left[\sqrt{\lambda T}\left(\lambda g_{i,\lambda T} + (1-\lambda) g_{i,(1-\lambda)T}\right)\sqrt{\lambda T}\left(g_{j,(1-\lambda)T} - g_{j,\lambda T}\right)^{\top}\right] \\ = \lim_{T\to\infty}\left(-\mathrm{E}\left[\sqrt{\lambda T}\lambda g_{i,\lambda T}\sqrt{\lambda T} g_{j,\lambda T}^{\top}\right]\right. \\ \left. + \mathrm{E}\left[\sqrt{\lambda T}(1-\lambda) g_{i,(1-\lambda T)}\sqrt{\lambda T} g_{j,(1-\lambda)T}^{\top}\right]\right) \\ = \lambda S_{ij} - \lambda S_{ij} = 0\end{aligned}$$

for $i, j = 1, 2$. Therefore,

$$\begin{aligned} S_{12}^{\mathcal{A}} &= \lim_{T\to\infty} \mathrm{E}\left[\sqrt{\lambda T}\left(\lambda g_{1,\lambda T} + (1-\lambda) g_{1,(1-\lambda)T}\right)\right. \\ &\qquad \left.\times\sqrt{\lambda T}\left(g_{2,\lambda T} + B_{21}(1-\lambda)(g_{1,(1-\lambda)T} - g_{1,\lambda T})\right)^{\top}\right] \\ &= \lim_{T\to\infty} \mathrm{E}\left[\sqrt{\lambda T}\left(\lambda g_{1,\lambda T} + (1-\lambda) g_{1,(1-\lambda)T}\right)\sqrt{\lambda T} g_{2,\lambda T}^{\top}\right] \\ &= \lim_{T\to\infty} \mathrm{E}\left[\sqrt{\lambda T}\lambda g_{1,\lambda T}\sqrt{\lambda T} g_{2,\lambda T}^{\top}\right] \\ &= \lambda S_{12}. \end{aligned}$$

The 2nd line follows from equation (A-3), and the 3rd and 4th lines follow from Theorem A.2. Using similar reasoning,

$$\begin{aligned} S_{22}^{\mathcal{A}} &= \lim_{T\to\infty} \mathrm{E}\left[\sqrt{\lambda T} g_{2,\lambda T}\sqrt{\lambda T} g_{2,\lambda T}^{\top}\right] - 2\lim_{T\to\infty}(1-\lambda)\mathrm{E}\left[\sqrt{\lambda T} g_{2,\lambda T}\sqrt{\lambda T} g_{1,\lambda T}^{\top}\right] B_{21}^{\top} \\ &\quad + \lim_{T\to\infty} B_{21}(1-\lambda)^2\mathrm{E}\left[\sqrt{\lambda T}(g_{1,(1-\lambda)T} - g_{1,\lambda T})\sqrt{\lambda T}(g_{1,(1-\lambda)T} - g_{1,\lambda T})^{\top}\right] B_{21}^{\top} \\ &= S_{22} - 2(1-\lambda)S_{21}S_{11}^{-1}S_{12} + (1-\lambda)^2\left(\frac{\lambda}{1-\lambda} + 1\right) S_{21}S_{11}^{-1}S_{12} \\ &= S_{22} - (1-\lambda)S_{21}S_{11}^{-1}S_{12}, \end{aligned}$$

which completes the derivation of equation (8). □

Theorem A.2 establishes consistency of the estimators.

Theorem A.2. As $T \to \infty$, $\hat{\theta}_T^k \underset{\text{a.s.}}{\to} \theta_0$ for $k \in \{\mathcal{S}, \mathcal{L}, \mathcal{A}, \mathcal{I}\}$.

Proof. White and Domowitz (1984) show that under these assumptions,

$$\begin{aligned} |g_{\lambda T}(\theta) - \mathrm{E} f(x_t, \theta)| &\underset{\text{a.s.}}{\to} 0 \qquad \text{and} \\ |g_{(1-\lambda)T}(\theta) - \mathrm{E} f(x_t, \theta)| &\underset{\text{a.s.}}{\to} 0, \end{aligned}$$

as $T \to \infty$ uniformly in $\theta \in \Theta$. By the continuous mapping theorem,

$$h_T^k(\theta)^{\top} W_T^k h_T^k(\theta) \underset{\text{a.s.}}{\to} \mathrm{E}[f(x_t, \theta)]^{\top} W^k \mathrm{E}[f(x_t, \theta)]$$

for $k \in \{\mathcal{S}, \mathcal{L}, \mathcal{A}\}$, and

$$h_T^{\mathcal{I}}(\theta)^{\top} W_T^{\mathcal{I}} h_T^{\mathcal{I}}(\theta) \underset{\text{a.s.}}{\to} \mathrm{E}\left[f_1(x_{1t}, \theta)^{\top}, f(x_t, \theta)^{\top}\right]^{\top} W^{\mathcal{I}} \mathrm{E}\begin{bmatrix} f_1(x_{1t}, \theta) \\ f(x_t, \theta) \end{bmatrix}$$

uniformly in θ. The result then follows from Amemiya ((1985), thm. 4.1.1). □

For convenience, define the notation

$$\begin{aligned} D_0^k &= D_0, \quad k \in \{\mathcal{S}, \mathcal{L}, \mathcal{A}\}, \\ D_0^{\mathcal{I}} &= \left[D_{0,1}^{\top}, D_{0,1}^{\top}, D_{0,2}^{\top}\right]^{\top}. \end{aligned}$$

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

The following theorem establishes asymptotic normality.

Theorem A.3.

$$\sqrt{\lambda T}(\hat{\theta}_T^k - \theta_0) \quad \underset{\text{d}}{\to} \quad N\left(0, \left((D_0^k)^\top W^k D_0^k\right)^{-1} \left((D_0^k)^\top W^k S^k W^k D_0^k\right) \left((D_0^k)^\top W^k D_0^k\right)^{-1}\right).$$

Proof. Define

$$D_T^k(\theta) \quad = \quad \frac{\partial h_T^k}{\partial \theta}(\theta)$$

for θ in the interior of Θ. For T sufficiently large, $\hat{\theta}_T^k$ lies in the interior of Θ. By the mean value theorem, there exists a $\tilde{\theta}^k$ in the segment between θ_0 and $\hat{\theta}_T^k$ such that

$$h_T^k(\hat{\theta}_T^k) - h_T^k(\theta_0) \quad = \quad D_T^k(\tilde{\theta}^k)(\hat{\theta}_T^k - \theta_0).$$

Premultiplying by $D_T^k(\hat{\theta}_T^k)^\top W_T^k$,

$$D_T^k(\hat{\theta}_T^k)^\top W_T^k \left(h_T^k(\hat{\theta}_T^k) - h_T^k(\theta_0)\right) \quad = \quad D_T^k(\hat{\theta}_T^k)^\top W_T^k D_T^k(\tilde{\theta}^k)(\hat{\theta}_T^k - \theta_0).$$

By the FOC of the optimization problem,

$$D_T^k(\hat{\theta}_T^k)^\top W_T^k D_T^k(\tilde{\theta}^k)(\hat{\theta}_T^k - \theta_0) \quad = \quad -D_T^k(\hat{\theta}_T^k)^\top W_T^k h_T^k(\theta_0).$$

Theorem 2.3 of White and Domowitz (1984) implies that

$$D_T^k(\theta) \quad \underset{\text{a.s.}}{\to} \quad \text{E}\left[\frac{\partial f}{\partial \theta}(x_t, \theta)\right], \quad \text{for } k \in \{\mathcal{S}, \mathcal{L}, \mathcal{A}\}, \text{ and}$$

$$D_T^{\mathcal{I}}(\theta) \quad \underset{\text{a.s.}}{\to} \quad \text{E}\begin{bmatrix} \frac{\partial f_1}{\partial \theta}(x_{1t}, \theta) \\ \frac{\partial f}{\partial \theta}(x_t, \theta) \end{bmatrix}$$

uniformly in θ. Therefore by Theorem A.2 and Amemiya ((1985), thm. 4.1.5),

$$D_T^k(\hat{\theta}_T^k) \underset{\text{a.s.}}{\to} D_0^k \quad \text{and} \quad D_T^k(\tilde{\theta}^k) \underset{\text{a.s.}}{\to} D_0^k.$$

The result follows from the Slutsky theorem. □

As in Hansen (1982), choosing the weighting matrix that is a consistent estimator of the inverse variance-covariance matrix is efficient for a given set of moment conditions.

Theorem A.4. Suppose $W_{\lambda T}^k \underset{\text{a.s.}}{\to} W_k = (S^k)^{-1}$. Then,

$$\sqrt{\lambda T}(\hat{\theta}_T^k - \theta_0) \quad \underset{\text{d}}{\to} \quad N\left(0, \left((D_0^k)^\top \left(S^k\right)^{-1} (D_0^k)\right)^{-1}\right).$$

Moreover, this choice of W^k is efficient for each estimator.

Appendix B. Proofs of the Theorems in the Text

1. Proof of Theorem 1

We first show that the AM and OI estimators are asymptotically equivalent. Because the means are the same, it suffices to show that

$$\text{(B-1)} \qquad \left(\begin{bmatrix} D_{0,1}^\top & D_0^\top \end{bmatrix}^\top \left(S^{\mathcal{I}}\right)^{-1} \begin{bmatrix} D_{0,1} \\ D_0 \end{bmatrix}\right)^{-1} \quad = \quad \left(D_0^\top \left(S^{\mathcal{A}}\right)^{-1} D_0\right)^{-1}.$$

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

From equation (9) and from the expression for the inverse of an invertible matrix, it follows that

$$\left(S^{\mathcal{I}}\right)^{-1} = \begin{bmatrix} \frac{1-\lambda}{\lambda}S_{11}^{-1} & 0 \\ 0 & S^{-1} \end{bmatrix} = \begin{bmatrix} \frac{1-\lambda}{\lambda}S_{11}^{-1} & 0 & 0 \\ 0 & S_{11}^{-1}+B_{21}^{\top}\Sigma^{-1}B_{21} & -B_{21}^{\top}\Sigma^{-1} \\ 0 & -\Sigma^{-1}B_{21} & \Sigma^{-1} \end{bmatrix}.$$

Moreover, it follows from the formula for the matrix inverse (see Green (1997), chap. 2) that

$$\left(S^{\mathcal{A}}\right)^{-1} = \begin{bmatrix} \frac{1}{\lambda}S_{11}^{-1}+B_{21}^{\top}\Sigma^{-1}B_{21} & -B_{21}^{\top}\Sigma^{-1} \\ -\Sigma^{-1}B_{21} & \Sigma^{-1} \end{bmatrix}.$$

Equation (B-1) follows.

We now show that the AM estimator is more efficient than the short estimator; namely, that

$$\left(D_0^{\top}\left(S^{\mathcal{A}}\right)^{-1}D_0\right)^{-1} \leq \left(D_0^{\top}S^{-1}D_0\right)^{-1},$$

where $A \leq B$ should be interpreted as stating that $B-A$ is positive semidefinite. It suffices to show that $S^{\mathcal{A}} \leq S$.[16] Note that

$$S - S^{\mathcal{A}} = (1-\lambda)\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{21}S_{11}^{-1}S_{12} \end{bmatrix}.$$

For any $l \times 1$ vector $v = [v_1^{\top},\ v_2^{\top}]^{\top}$,

$$\begin{aligned} v^{\top}(S-S^{\mathcal{A}})v &= (1-\lambda)\left(v_1^{\top}S_{11}v_1 + v_1^{\top}S_{12}v_2 + v_2^{\top}S_{21}v_1 + v_2^{\top}S_{21}S_{11}^{-1}S_{12}v_2\right) \\ &= (1-\lambda)(S_{11}v_1 + S_{12}v_2)^{\top}S_{11}^{-1}(S_{11}v_1 + S_{12}v_2) \geq 0, \end{aligned}$$

because S_{11}^{-1} is positive semidefinite and $\lambda < 1$. An analogous proof shows that the AM estimator is more efficient than the long estimator.

2. Proof of Theorem 2

Define

$$U = WD_0\left(D_0^{\top}WD_0\right)^{-1}.$$

By Theorem A.3, it suffices to show that $U^{\top}SU - U^{\top}S^{\mathcal{A}}U$ and that $U^{\top}S^{\mathcal{L}}U - U^{\top}S^{\mathcal{A}}U$ are positive semidefinite. For any vector v,

$$v^{\top}\left(U^{\top}SU - U^{\top}S^{\mathcal{A}}U\right)v = (Uv)^{\top}(S-S^{\mathcal{A}})Uv > 0,$$

because $S-S^{\mathcal{A}}$ is positive semidefinite. A similar argument shows that $U^{\top}S^{\mathcal{L}}U - U^{\top}S^{\mathcal{A}}U$ is positive semidefinite.

[16] For invertible matrices U_1 and U_2, if $U_1 - U_2$ is positive semidefinite, then $U_2^{-1} - U_1^{-1}$ is positive semidefinite (Goldberger (1964), chap. 2.7). It follows that for a conforming matrix M, $(M^{\top}U_1^{-1}M)^{-1} - (M^{\top}U_2^{-1}M)^{-1}$ is positive semidefinite.

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

3. Derivation of the FOCs for the Efficient Estimators

Differentiating the objective function for the OI estimator with respect to θ yields

$$\begin{aligned}\frac{1-\lambda}{\lambda} g_{1,(1-\lambda)T}^{\top} S_{11}^{-1} \frac{\partial g_{1,(1-\lambda)T}}{\partial \theta} + g_{1,\lambda T}^{\top} S_{11}^{-1} \frac{\partial g_{1,\lambda T}}{\partial \theta} \\ + \left[g_{1,\lambda T}^{\top}, \; g_{2,\lambda T}^{\top} \right] \begin{bmatrix} B_{21}^{\top} \Sigma^{-1} B_{21} & -B_{21}^{\top} \Sigma^{-1} \\ -\Sigma^{-1} B_{21} & \Sigma^{-1} \end{bmatrix} \begin{bmatrix} \frac{\partial g_{1,\lambda T}}{\partial \theta} \\ \frac{\partial g_{2,\lambda T}}{\partial \theta} \end{bmatrix} &= 0,\end{aligned}$$

which reduces to

$$\begin{aligned}\frac{1-\lambda}{\lambda} g_{1,(1-\lambda)T}^{\top} S_{11}^{-1} \frac{\partial g_{1,(1-\lambda)T}}{\partial \theta} + g_{1,\lambda T}^{\top} S_{11}^{-1} \frac{\partial g_{1,\lambda T}}{\partial \theta} \\ + \left(g_{2,\lambda T} - B_{21} g_{1,\lambda T} \right)^{\top} \Sigma^{-1} \frac{\partial}{\partial \theta} \left(g_{2,\lambda T} - B_{21} g_{1,\lambda T} \right) &= 0.\end{aligned}$$

Differentiating the objective function for the AM estimator with respect to θ yields

$$\frac{1}{\lambda} g_{1,T}^{\top} S_{11}^{-1} \frac{\partial g_{1,T}}{\partial \theta} + \left[g_{1,T}^{\top}, \; \left(g_{2,T}^{\mathcal{A}} \right)^{\top} \right] \begin{bmatrix} B_{21}^{\top} \Sigma^{-1} B_{21} & -B_{21}^{\top} \Sigma^{-1} \\ -\Sigma^{-1} B_{21} & \Sigma^{-1} \end{bmatrix} \begin{bmatrix} \frac{\partial g_{1,T}}{\partial \theta} \\ \frac{\partial g_{2,T}^{\mathcal{A}}}{\partial \theta} \end{bmatrix} = 0,$$

which reduces to

$$\frac{1}{\lambda} g_{1,T}^{\top} S_{11}^{-1} \frac{\partial g_{1,T}}{\partial \theta} + \left(B_{21} g_{1,\lambda T} - g_{2,\lambda T} \right)^{\top} \Sigma^{-1} \frac{\partial}{\partial \theta} \left(B_{21} g_{1,\lambda T} - g_{2,\lambda T} \right) = 0.$$

References

Ahn, S. C., and P. Schmidt. "A Separability Result for GMM Estimation with Applications to GLS Prediction and Conditional Moment Tests." *Econometric Reviews*, 14 (1995), 19–34.

Ait-Sahalia, Y.; P. A. Mykland; and L. Zhang. "How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise." *Review of Financial Studies*, 18 (2005), 351–416.

Amemiya, T. *Advanced Econometrics*. Cambridge, MA: Harvard University Press (1985).

Anderson, T. W. "Maximum Likelihood Estimates for a Multivariate Normal Distribution When Some Observations Are Missing." *Journal of the American Statistical Association*, 52 (1957), 200–203.

Andrews, D. W. K., and R. C. Fair. "Inference in Nonlinear Econometric Models with Structural Change." *Review of Economic Studies*, 55 (1988), 615–640.

Andrews, D. W. K., and W. Ploberger. "Optimal Tests When a Nuisance Parameter Is Present Only under the Alternative." *Econometrica*, 62 (1994), 1383–1414.

Bandi, F. M., and J. R. Russell. "Separating Microstructure Noise from Volatility." *Journal of Financial Economics*, 79 (2006), 655–692.

Brandt, M. W. "Estimating Portfolio and Consumption Choice: A Conditional Euler Equations Approach." *Journal of Finance*, 54 (1999), 1609–1645.

Campbell, J. Y., and R. J. Shiller. "Stock Prices, Earnings, and Expected Dividends." *Journal of Finance*, 43 (1988), 661–676.

Campbell, J. Y., and S. B. Thompson. "Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?" *Review of Financial Studies*, 21 (2008), 1509–1531.

Cavanagh, C. L.; G. Elliott; and J. H. Stock. "Inference in Models with Nearly Integrated Regressors." *Econometric Theory*, 11 (1995), 1131–1147.

Cochrane, J. H. *Asset Pricing*. Princeton, NJ: Princeton University Press (2001).

Conniffe, D. "Estimating Regression Equations with Common Explanatory Variables but Unequal Numbers of Observations." *Journal of Econometrics*, 27 (1985), 179–196.

Duffie, D., and K. J. Singleton. "Simulated Moments Estimation of Markov Models of Asset Prices." *Econometrica*, 61 (1993), 929–952.

Fama, E. F., and K. R. French. "Business Conditions and Expected Returns on Stocks and Bonds." *Journal of Financial Economics*, 25 (1989), 23–49.

Ghysels, E.; A. Guay; and A. Hall. "Predictive Tests for Structural Change with Unknown Breakpoint." *Journal of Econometrics*, 82 (1997), 209–233.

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press

Ghysels, E., and A. Hall. "A Test for Structural Stability of Euler Conditions Parameters Estimated via the Generalized Method of Moments Estimator." *International Economic Review*, 31 (1990), 335–364.
Ghysels, E.; P. Santa-Clara; and R. Valkanov. "There Is a Risk-Return Trade-Off After All." *Journal of Financial Economics*, 76 (2005), 509–548.
Goetzmann, W. N., and P. Jorion. "Re-Emerging Markets." *Journal of Financial and Quantitative Analysis* (1999), 1–32.
Goldberger, A. S. *Econometric Theory*. New York, NY: John Wiley and Sons (1964).
Green, W. H. *Econometric Analysis*. Upper Saddle River, NJ: Prentice-Hall, Inc. (1997).
Hansen, L. P. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica*, 50 (1982), 1029–1054.
Hansen, L. P., and K. Singleton. "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models." *Econometrica*, 50 (1982), 1269–1286.
Harvey, A.; S. J. Koopman; and J. Penzer. "Messy Time Series: A Unified Approach." In *Advances in Econometrics*, Vol. 13. Greenwich, CT: JAI Press Inc. (1998), 103–143.
Harvey, C. "Time-Varying Conditional Covariances in Tests of Asset Pricing Models." *Journal of Financial Economics*, 24 (1989), 289–317.
Little, R. J. A., and D. B. Rubin. *Statistical Analysis with Missing Data, 2nd ed.* Hoboken, NJ: John Wiley & Sons (2002).
Lynch, A. W., and J. A. Wachter. "Using Samples of Unequal Length in Generalized Method of Moments Estimation." New York University Working Paper FIN-05-021 (2004).
MacKinlay, A. C., and M. P. Richardson. "Using Generalized Method of Moments to Test Mean-Variance Efficiency." *Journal of Finance*, 46 (1991), 511–527.
Nelson, C. R., and M. J. Kim. "Predictable Stock Returns: The Role of Small Sample Bias." *Journal of Finance*, 48 (1993), 641–661.
Pastor, L., and R. F. Stambaugh. "Investing in Equity Mutual Funds." *Journal of Financial Economics*, 63 (2002a), 351–380.
Pastor, L., and R. F. Stambaugh. "Mutual Fund Performance and Seemingly Unrelated Assets." *Journal of Financial Economics*, 63 (2002b), 315–349.
Patton, A. J. "Estimation of Multivariate Models for Time Series of Possibly Different Lengths." *Journal of Applied Econometrics*, 21 (2006), 147–173.
Phillips, P. C. "Time Series Regressions with a Unit Root." *Econometrica*, 55 (1987), 277–301.
Robins, J. M., and A. Rotnitsky. "Semiparametric Efficiency in Multivariate Regression Models with Missing Data." *Journal of the American Statistical Association*, 90 (1995), 122–129.
Schmidt, P. "Estimation of Seemingly Unrelated Regressions with Unequal Numbers of Observations." *Journal of Econometrics*, 5 (1977), 365–377.
Shiller, R. J. *Market Volatility*. Cambridge, MA: MIT Press (1989).
Singleton, K. *Empirical Dynamic Asset Pricing: Model Specification and Econometric Assessment*. Princeton, NJ: Princeton University Press (2006).
Sowell, F. "Optimal Tests for Parameter Instability in the Generalized Method of Moments Framework." *Econometrica*, 64 (1996), 1085–1107.
Stambaugh, R. F. "Analyzing Investments Whose Histories Differ in Length." *Journal of Financial Economics*, 45 (1997), 285–331.
Stambaugh, R. F. "Predictive Regressions." *Journal of Financial Economics*, 54 (1999), 375–421.
Stock, J. H. "Unit Roots, Structural Breaks and Trends." In *Handbook of Econometrics*, Vol. IV, R. Engle and D. L. McFadden, eds. Amsterdam, The Netherlands: North-Holland (1994), 2740–2841.
Storesletten, K.; C. I. Telmer; and A. Yaron. "Cyclical Dynamics in Idiosyncratic Labor Market Risk." *Journal of Political Economy*, 112 (2004), 695–717.
Swamy, P. A. V. B., and J. S. Mehta. "On Bayesian Estimation of Seemingly Unrelated Regressions When Some Observations Are Missing." *Journal of Econometrics*, 3 (1975), 157–169.
White, H. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, 24 (1980), 817–838.
White, H. *Asymptotic Theory for Econometricians*. San Diego, CA: Academic Press, Inc. (1994).
White, H., and I. Domowitz. "Nonlinear Regression with Dependent Observations." *Econometrica*, 52 (1984), 143–162.
Zhou, G. "Asset-Pricing Tests under Alternative Distributions." *Journal of Finance*, 48 (1993), 1927–1942.
Zhou, G. "Analytical GMM Tests: Asset Pricing with Time-Varying Risk Premiums." *Review of Financial Studies*, 7 (1994), 687–709.

https://doi.org/10.1017/S0022109013000070 Published online by Cambridge University Press