Using Samples of Unequal Length in Generalized Method of Moments Estimation *

Anthony W. Lynch
New York University and NBER

Jessica A. Wachter
University of Pennsylvania and NBER

January 18, 2008

Abstract

Many applications in financial economics use data series with different starting or ending dates. This paper describes estimation methods, based on the generalized method of moments (GMM), which make use of all available data for each moment condition. We introduce two asymptotically equivalent estimators that are consistent, asymptotically normal, and more efficient asymptotically than standard GMM. We apply these methods to estimating predictive regressions in international data and show that the use of the full sample affects point estimates and standard errors for both assets with data available for the full period and 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. These methods are extended to more general patterns of missing data, and are shown to be more efficient than estimators that ignore intervals of the data, and thus more efficient than standard GMM.

*Lynch: Department of Finance, Stern School of Business, New York University, 44 West Fourth Street, New York, NY 10012-1126; Email: alynch@stern.nyu.edu; Tel: (212) 998-0350. Wachter: Department of Finance, The Wharton School, University of Pennsylvania, 3620 Locust Walk, Philadelphia, PA 19104; Tel: (215) 898-7634; Email: jwachter@wharton.upenn.edu. We thank Yacine Ait-Sahalia, David Chapman, Robert Engle, Martin Lettau, Andrew Lo, Kenneth Singleton, Robert Stambaugh, Jim Stock, Amir Yaron, Motohiro Yogo, as well as seminar participants at the 2005 AFA meetings, at New York University, at the Wharton School and at the University of Pennsylvania Department of Economics for their comments and suggestions.

Using Samples of Unequal Length in Generalized Method of Moments Estimation

Abstract

Many applications in financial economics use data series with different starting or ending dates. This paper describes estimation methods, based on the generalized method of moments (GMM), which make use of all available data for each moment condition. We introduce two asymptotically equivalent estimators that are consistent, asymptotically normal, and more efficient asymptotically than standard GMM. We apply these methods to estimating predictive regressions in international data and show that the use of the full sample affects point estimates and standard errors for both assets with data available for the full period and 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. These methods are extended to more general patterns of missing data, and are shown to be more efficient than estimators that ignore intervals of the data, and thus more efficient than standard GMM.

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.

When data are missing as described above, common practice is to take the intersection of the sample periods over which the data are observed. The intersection then becomes the sample period for the study and the rest of the data are ignored. This paper introduces an alternative, based on the generalized method of moments (GMM), that allows the researcher to make use of all of the data available for each moment condition.[1] We show, moreover, that our method is more efficient than standard GMM, and more efficient than introducing the data from the longer series in a "naive" way. We then apply our methods to estimating predictive regressions in international data.

The econometrics literature on unequal sample lengths goes back at least as far as Anderson (1957), who derives a maximum likelihood estimator for a bivariate normal distribution in which one variable has more observations than another. More recently, Harvey, Koopman, and Penzer (1998) develop a Kalman-filter approach to missing data, while Schmidt (1977), Swamy and Mehta (1975), and Conniffe (1985) focus on extending the seemingly unrelated regression approach to cases in which more data is available for one equation than the other. Little and Rubin (2002) survey the statistical literature on missing-data problems. Stambaugh (1997) builds on these methods to estimate the mean and variance of financial time series assuming returns are normally and independently distributed, in a setting where some return series start at a later date than others.[2]

[1] Burguete, Gallant, and Souza (1982) and Hansen (1982) describe the GMM estimator and derive its asymptotic properties. Hansen and Singleton (1982) derive implications for estimation and testing of financial models; Brandt (1999) derives implications for the estimation of optimal portfolio and consumption choice. See Newey and McFadden (1994) for a survey of recent work on GMM and related estimators.

[2] Pastor and Stambaugh (2002a, 2002b) derive Bayesian posteriors for means and variances of mutual fund returns using samples of unequal length, under the assumption of normality and identically and independently distributed returns. Storesletten, Telmer, and Yaron (2004) combine a time series of macro-economic variables dating back to 1930 with the shorter Panel Study of Income Dynamics to estimate the relationship between cross-sectional variance and recessions.

Following Anderson (1957), these previous studies take a likelihood-based approach. In contrast, our approach, because it is based on GMM, does not require the data generating process to be normal. It can be used for dependent, stationary processes, and it permits estimation of parameters that are related to the observed functions in non-linear ways.[3] As shown in Cochrane (2001), many common estimation techniques used in finance can be seen as special cases of GMM. 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 ARMA processes with general conditions on the underlying errors (see Phillips (1987)).[4]

Because our method is based on GMM, many of our results are asymptotic.[5] So that the asymptotic approximation is reasonable, care must be taken to insure 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 infinity, and also in the literature on structural breaks (Andrews and Fair (1988), Ghysels and Hall (1990), Stock (1994), Sowell (1996), 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

[3]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.

[4]Like many of the studies mentioned above, we do assume that the data is 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)).

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

generalize this to other patterns of missing data). The two sets of moment conditions may depend on the same or different underlying parameters. We develop two asymptotically equivalent estimators that make use of all of the data.[6] While general, these estimators are straightforward to implement, as we show in an application involving international data (Section 3), and have natural and intuitive interpretations.

The first estimator (which we call the *adjusted-moment* 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 second estimator, (which we call the *over-identified* 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 adjusted-moment estimator (and the maximum-likelihood estimator proposed in that paper). In the more general context of our paper, the two 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, Chapter 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 over-identified 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.

Our approach can be extended to many other patterns of missing data. One pattern of interest is the case where there are more than two starting dates but all series end at the same date (this case satisfies a condition that Little and Rubin (2002) call monotonicity). Both of our estimators can be extended not only to this case, but further, to cases where

[6] 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 (2004)).

the series do not satisfy monotonicity. The extension works for an arbitrary number of ending dates and starting dates. It is also possible to have data missing in the middle of the sample.[7] We show that it is always more efficient to "add" an interval of data, even if some series are not observed over the interval. By implication, these generalized estimators are also more efficient than standard GMM.

The organization of the paper is as follows. Section 1 defines our estimators and discusses their efficiency properties. Section 2 provides intuition for the efficiency gains from using the full sample. Section 3 illustrates our methods through an application to international data. Section 4 presents a Monte Carlo analysis showing that the efficiency gains are present in small samples. Section 5 outlines extensions to more general patterns of missing data, and Section 6 concludes.

1 GMM estimators for samples of unequal length

1.1 Definitions

Let x_t denote observations on a vector-valued stochastic process. Let θ be a vector of parameters. The underlying economic model is captured by a function $f(x_t, \theta)$ in the sense that

$$E\left[f(x_t, \theta_0)\right] = 0$$

for a unique value θ_0 of θ. In Appendix A we formally define the stochastic setting and state assumptions on x_t and f. These assumptions are standard (see, e.g. White and Domowitz (1984)) and do not include assumptions on the distribution of x. We do require that x_t be stationary, and that observations satisfy a notion of dependence known as mixing. Mixing guarantees that autocovariances vanish at sufficiently long lags and is convenient because it allows a tradeoff between the amount of dependence that x_t can exhibit and regularity conditions on f. Mixing is a relatively weak assumption: for example, it allows all finite-order ARMA processes.

In many applications, it happens that data are missing for the early part of the sample period for some moment conditions (see Section 3 for an application to international data). We partition the elements of x_t so that $x_t = \left[x_{1t}^\top \; x_{2t}^\top\right]^\top$, where data on x_{1t} are assumed

[7] 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)).

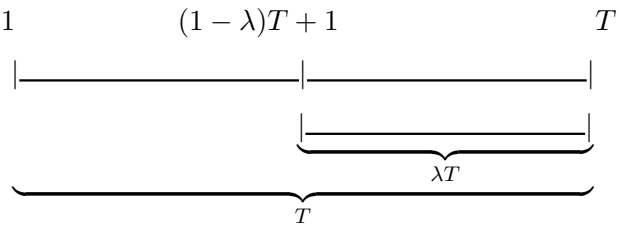

Figure 1: Notation for data missing at the start of the sample

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 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:[8]

$$\begin{aligned} g_{1,T}(\theta) &= \frac{1}{T}\sum_{t=1}^{T} f_1(x_{1t}, \theta), \\ g_{1,(1-\lambda)T}(\theta) &= \frac{1}{(1-\lambda)T}\sum_{t=1}^{(1-\lambda)T} f_1(x_{1t}, \theta), \\ g_{1,\lambda T}(\theta) &= \frac{1}{\lambda T}\sum_{t=(1-\lambda)T+1}^{T} f_1(x_{1t}, \theta), \end{aligned}$$

and

$$g_{2,\lambda T}(\theta) = \frac{1}{\lambda T}\sum_{t=(1-\lambda)T+1}^{T} f_2(x_t, \theta).$$

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.

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

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

[8]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.

Under the assumptions of Appendix A, these sums converge (see White (1994, Proposition 3.44)). Let

$$S_{ij} = \sum_{\tau=-\infty}^{\infty} R_{ij}(\tau).$$

and

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}.$$

It is also useful to define the matrix of coefficients from a regression of the second series on the first:

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

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

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

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

$$\min_{\theta} \begin{bmatrix} g_{1,\lambda T}(\theta)^\top & g_{2,\lambda T}(\theta)^\top \end{bmatrix}^\top W_T \begin{bmatrix} g_{1,\lambda T}(\theta) \\ g_{2,\lambda T}(\theta) \end{bmatrix}.$$

for a positive definite and symmetric weighting matrix W_T. In what follows, we will 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.[9] Let

$$\hat{\theta}_T^{\mathcal{S}} = \text{argmin}_\theta \begin{bmatrix} g_{1,\lambda T}(\theta)^\top & g_{2,\lambda T}(\theta)^\top \end{bmatrix}^\top \hat{S}_T^{-1} \begin{bmatrix} g_{1,\lambda T}(\theta) \\ g_{2,\lambda T}(\theta) \end{bmatrix}. \tag{2}$$

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 (2), except $g_{1,\lambda T}(\theta)$ is 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

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

[9] 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.

where $\hat{S}_T^{\mathcal{L}}$ is an estimate of $S^{\mathcal{L}}$, the asymptotic variance of $\sqrt{\lambda T}\left[g_{1,T}(\theta)^\top \; g_{2,\lambda T}(\theta)^\top\right]^\top$.

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

$$\hat{\theta}_T^{\mathcal{A}} = \text{argmin}_\theta \left[g_{1,T}(\theta)^\top \; g_{2,T}^{\mathcal{A}}(\theta)^\top\right]^\top \left(\hat{S}_T^{\mathcal{A}}\right)^{-1} \left[\begin{array}{c} g_{1,T}(\theta) \\ g_{2,T}^{\mathcal{A}}(\theta) \end{array}\right], \tag{4}$$

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

The difference between (3) and (4) lies in the second 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 second set of moment conditions for the adjusted-moment estimator, $g_{2,T}^{\mathcal{A}}(\theta)$, can be written as

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

The expression above illustrates the role of the longer sample in helping to estimate the second 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 second 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 $E[f_2(x_0,\theta)]$ should be adjusted upward relative to g_2.

Finally, we define an estimator that makes use of longer data sample to add over-identifying restrictions. The *over-identified* estimator solves

$$\hat{\theta}_T^{\mathcal{I}} = \text{argmin}_\theta \left[g_{1,(1-\lambda)T}(\theta)^\top \; g_{1,\lambda T}(\theta)^\top g_{2,\lambda T}(\theta)^\top\right]^\top \left(\hat{S}_T^{\mathcal{I}}\right)^{-1} \left[\begin{array}{c} g_{1,(1-\lambda)T}(\theta) \\ g_{1,\lambda T}(\theta) \\ g_{2,\lambda T}(\theta) \end{array}\right], \tag{5}$$

where $\hat{S}_T^{\mathcal{I}}$ is an estimate of $S^{\mathcal{I}}$, the asymptotic variance of $\sqrt{\lambda T}\left[g_{1,(1-\lambda)T}(\theta)^\top \; g_{1,\lambda T}(\theta)^\top g_{2,\lambda T}(\theta)^\top\right]$.

1.2 Asymptotic distribution

Theorems C.2 and C.3 in Appendix C 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

$$D_{0,i} = E\left[\left(\partial f_i/\partial\theta\right)|_{\theta_0}\right],$$

for $i = 1, 2$, and

$$D_0 = \left[\begin{array}{c} D_{0,1} \\ D_{0,2} \end{array}\right].$$

For the short, long, and adjusted-moment estimators, D_0 is the derivative of the moment condition evaluated at θ_0. Therefore, as shown in Theorem C.3, the asymptotic distribution of the estimators is given by

$$\sqrt{\lambda T}(\hat{\theta}_T^{\mathcal{S}} - \theta_0) \to_{\mathrm{d}} N\left(0, \left(D_0^\top S^{-1} D_0\right)^{-1}\right)$$

for the short estimator,

$$\sqrt{\lambda T}(\hat{\theta}_T^{\mathcal{L}} - \theta_0) \to_{\mathrm{d}} N\left(0, \left(D_0^\top \left(S^{\mathcal{L}}\right)^{-1} D_0\right)^{-1}\right)$$

for the long estimator, and

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

for the adjusted-moment estimator. The over-identified estimator has derivatives $[D_{0,1}^\top \; D_0^\top]$, and therefore its asymptotic distribution is

$$\sqrt{\lambda T}(\hat{\theta}_T^{\mathcal{I}} - \theta_0) \to_{\mathrm{d}} N\left(0, \left([D_{0,1}^\top \; D_0^\top]^\top \left(S^{\mathcal{I}}\right)^{-1} \left[\begin{array}{c} D_{0,1} \\ D_0 \end{array}\right]\right)^{-1}\right).$$

An important practical step in implementing these estimators is obtaining estimates of the variance 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 C.1:[10]

$$S^{\mathcal{L}} = \begin{bmatrix} \lambda S_{11} & \lambda S_{12} \\ \lambda S_{21} & S_{22} \end{bmatrix} \tag{6}$$

$$S^{\mathcal{A}} = \begin{bmatrix} \lambda S_{11} & \lambda S_{12} \\ \lambda S_{21} & S_{22} - (1-\lambda) S_{21} S_{11}^{-1} S_{12} \end{bmatrix} \tag{7}$$

$$S^{\mathcal{I}} = \begin{bmatrix} \frac{\lambda}{1-\lambda} S_{11} & 0 & 0 \\ 0 & S_{11} & S_{12} \\ 0 & S_{21} & S_{22} \end{bmatrix}. \tag{8}$$

Therefore it suffices to have estimates of submatrices of S. Underlying (6)–(8) is asymptotic independence between non-overlapping samples. That is, $\sqrt{\lambda T} g_{1,(1-\lambda)T}$ and $\sqrt{\lambda T} g_{i,\lambda T}$ are jointly normally distributed and

$$\lim_{T \to \infty} E\left[\sqrt{\lambda T} g_{1,(1-\lambda)T} \sqrt{\lambda T} g_{i,\lambda T}\right] = 0 \quad i = 1, 2.$$

This result is shown formally in Appendix B. This statement is intuitive: as more and more data become available, the part of the series the non-overlapping samples that is close becomes an ever smaller percent of the whole and the series is dominated by terms that are far away and therefore independent.

1.3 Efficiency properties

We now compare the asymptotic efficiency of the four estimators.

Theorem 1.1 *Assume the short, long, adjusted-moment and over-identified estimators are defined as (2)–(5). The assumptions in Appendix A imply*

1. *The asymptotic distribution of the adjusted-moment estimator is identical to that of the over-identified estimator.*

2. *The adjusted-moment estimator and over-identified estimator are more efficient than the short estimator.*

3. *The adjusted-moment estimator and over-identified estimator are more efficient than the long estimator.*

[10] Our proposed weighting matrix for the over-identified 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 (8), without the $\lambda/(1-\lambda)$ term in the upper left block.

PROOF: It suffices to compare the asymptotic variances of each estimator because the mean is the same for all of them. That is, it suffices to show that the variance in these expressions is equal for the adjusted-moment and over-identified estimators, and is smaller (in a matrix sense) for these estimators than for the long and short estimator. In the notation of the previous section, this is equivalent to showing

$$\left([D_{0,1}^\top \; D_0^\top]^\top \left(S^{\mathcal{I}}\right)^{-1} \left[\begin{array}{c} D_{0,1} \\ D_0 \end{array} \right] \right)^{-1} = \left(D_0^\top \left(S^{\mathcal{A}}\right)^{-1} D_0 \right)^{-1} \leq \left(D_0^\top S^{-1} D_0 \right)^{-1} \tag{9}$$

and that

$$\left(D_0^\top \left(S^{\mathcal{A}}\right)^{-1} D_0 \right)^{-1} \leq \left(D_0^\top \left(S^{\mathcal{L}}\right)^{-1} D_0 \right)^{-1}. \tag{10}$$

where $A \leq B$ should be interpreted as stating that $B - A$ is positive semi-definite.

We begin by showing the equivalence of the adjusted-moment and over-identified estimators. From (8) and from the expression for the inverse of an invertible matrix it follows that

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

Moreover, it follows from Lemma (D.3) that

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

The equality in (9) follows.

To show the remaining statements, we use Lemma D.2, which states that it suffices to show $S^{\mathcal{A}} \leq S$, and $S^{\mathcal{A}} \leq S^{\mathcal{L}}$. To show $S^{\mathcal{A}} \leq S$, note that

$$S - S^{\mathcal{A}} = (1-\lambda) \left[\begin{array}{cc} S_{11} & S_{12} \\ S_{21} & S_{21} S_{11}^{-1} S_{12} \end{array} \right].$$

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-semi-definite and $\lambda < 1$. To show that $S^{\mathcal{A}} \leq S^{\mathcal{L}}$

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

which is positive semi-definite by the same reasoning. The first statement of the theorem then implies that $\hat{\theta}_T^{\mathcal{I}}$ is also more efficient than $\hat{\theta}_T^{\mathcal{S}}$ and $\hat{\theta}_T^{\mathcal{L}}$. □

Theorem 1.1 shows that asymptotically, the adjusted-moment and over-identified estimators are the same despite the fact that they take very different forms. The second statement shows that there is indeed an efficiency gain from using the longer sample. Moreover, it is more efficient to use the adjusted-moment or over-identified estimators than to use the longer sample in a "naive" way, as the third statement shows.

In contrast, the long estimator, despite its use of all the data, may not be more efficient than the short estimator. Using the same reasoning as above, it suffices to compare $S^{\mathcal{L}}$ with S. From (6), it follows that

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

If the covariances between the first and second set of moment conditions are nonzero, this matrix may not be positive semi-definite. Thus it is not sufficient to simply include the full data in the estimation. The non-overlapping 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 adjusted-moment and over-identified estimators) and the long estimator is especially surprising given that, when attention is restricted to estimating $f_1(x, \theta)$, the three estimators are asymptotically identical.

Theorem 1.1 addresses 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.

Theorem 1.2 *Assume that the weighting matrices approach a positive-definite matrix W. The adjusted-moment estimator is more efficient than the short estimator and the long estimator.*

PROOF: Define

$$U = W D_0 \left(D_0^\top W D_0 \right)^{-1}. \tag{11}$$

By Theorem C.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 semi-definite. For any vector v,

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

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

We further examine the relation between the adjusted-moment and over-identified estimators by examining the first order conditions. 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. Differentiating the over-identified estimator with respect to θ yields

$$\frac{1-\lambda}{\lambda} g_{1,(1-\lambda)T}^\top S_{11}^{-1} \frac{\partial g_{1,(1-\lambda)T}}{\partial \theta} + g_{1,\lambda T}^\top S_{11}^{-1} \frac{\partial g_{1,\lambda T}}{\partial \theta} + \left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right)^\top \Sigma^{-1} \frac{\partial}{\partial \theta}\left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right) = 0. \quad (12)$$

Equation (12) is the first-order condition that determines the over-identified estimator $\hat{\theta}_T^{\mathcal{I}}$. By contrast, the first order condition associated with the adjusted-moment estimator is

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

which reduces to

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

Equation (13) is the first-order condition that determines the adjusted-moment estimator $\hat{\theta}_T^{\mathcal{A}}$. According to Theorem 1.1, these two first order conditions must be equivalent as $T \to \infty$. Indeed they are, because

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

and

$$\begin{aligned} \frac{1-\lambda}{\lambda} g_{1,(1-\lambda)T}^\top S_{11}^{-1} D_{0,1} + g_{1,\lambda T}^\top S_{11}^{-1} D_{0,1} &= \frac{1}{\lambda}\left((1-\lambda) g_{1,(1-\lambda)T}^\top + \lambda g_{1,\lambda T}^\top\right) S_{11}^{-1} D_{0,1} \\ &= \frac{1}{\lambda} g_{1,T}^\top S_{11}^{-1} D_0. \end{aligned}$$

In finite samples however, they will generally be equivalent only when

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

which occurs when the moment conditions are linear with coefficients that do not depend on the data. This corresponds to the case examined by Stambaugh (1997) in a maximum likelihood context.

Of interest is the special case in which the model is exactly identified, and when the moment conditions g_1 depend only on a subset of the parameters θ_1. Denote the remaining parameters by θ_2. In this case, the first order conditions for the over-identified estimator $\hat{\theta}_T^{\mathcal{T}}$ reduce to

$$g_{2,\lambda T} - B_{21} g_{1,\lambda T} = 0 \tag{14}$$

for θ_2 (because the derivative of g_1 with respect to θ_2 is zero) and

$$\frac{1-\lambda}{\lambda} g_{1,(1-\lambda)T}^{\top} S_{11}^{-1} \frac{\partial g_{1,(1-\lambda)T}}{\partial \theta_1} + g_{1,\lambda T}^{\top} S_{11}^{-1} \frac{\partial g_{1,\lambda T}}{\partial \theta_1} = 0 \tag{15}$$

for θ_1 In comparison, the first order conditions for the adjusted-moment estimator $\hat{\theta}_T^{\mathcal{A}}$ reduce to (14) for θ_2 and $g_{1,T} = 0$ for θ_1.[11] Thus the adjusted-moment estimator gives the same estimate for θ_1 as simply using the full sample. The over-identified estimator gives a possibly different estimate, one that depends on the point in time in which the second series begins. While this dependence is possibly unattractive, (15) nonetheless has an interpretation: it 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.

2 The effect of the full sample

This section provides additional intuition on the source of gains in efficiency. For simplicity, we maintain the assumptions of the previous paragraph, namely that the system is exactly identified, and that the variables can be decomposed into $\theta = [\theta_1^{\top} \ \theta_2^{\top}]^{\top}$, where f_1 is a function of θ_1 alone. This section focuses on results that hold asymptotically. For this

[11] Note that, even though the first-order conditions for θ_2 have the same form, the finite-sample adjusted-moment estimate will generally be different from the finite-sample over-identified estimate. The reason is that the term $g_{1,\lambda T}$ depends on θ_1 (it is also possible that $g_{2,\lambda T}$ depends on θ_1 as well). Therefore the second term in (14) will differ across the two estimators because the estimate of θ_1 will differ, and as a result the equations will yield different solutions for θ_2.

reason, the adjusted-moment and the over-identified estimators will be the same, hence we refer to them as the *efficient estimators*.

Under these assumptions, the derivative matrix is invertible and takes the form

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

where

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

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 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 (first block of $\left(D_0^\top \left(S^{\mathcal{A}} \right)^{-1} D_0 \right)^{-1}$) 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. Note that the second set of moment conditions f_2 has no effect on the estimation of θ_1 (see Ahn and Schmidt (1995)).

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

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

Thus the variance for the second set of parameters can be decomposed into two parts. The first part represents the effect of the first moment conditions on the second variables. The second part represents the variance due only to the residual variance of the second 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 second set of moment conditions conditional on the first. The second diagonal block of $\left(D_0^\top S^{-1} D_0 \right)^{-1}_{22}$,

which gives the standard errors for θ_2 under standard GMM, has an analogous decomposition:

$$\left(D_0^\top (S)^{-1} D_0\right)^{-1}_{22} = d_{22}^{-1}\left[d_{21}d_{11}^{-1}S_{11} - S_{21}\right](S_{11})^{-1}\left[d_{21}d_{11}^{-1}S_{11} - S_{21}\right]^\top (d_{22}^{-1})^\top + d_{22}^{-1}\left[S_{22} - S_{21}(S_{11})^{-1}S_{12}\right](d_{22}^{-1})^\top. \quad (17)$$

Comparing (16) with (17) reveals the source of the efficiency gain. The first term in (16) is equal to λ multiplied by the first term in (17):

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

However the second terms are the same, not surprisingly because they represent the variance of the second moment conditions conditional on the value of the first:

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

Thus the decrease in the standard errors depends on the extent to which the first term dominates the second term. For example, when the second set of moments are perfectly correlated with the first set, the residual variance is zero,

$$S_{22} - S_{21}S_{11}^{-1}S_{12} = 0, \quad (18)$$

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, i.e. $d_{21} = 0$ (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.

Of course, even if the two 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 first term in (16) 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 second set of moment conditions ($d_{21} = 0$) the first set of moment conditions help to estimate θ_2 if the covariance between the two moment conditions is nonzero.

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

$$g_{1,T}(x, \hat{\theta}_T^{\mathcal{L}}) = 0 \quad (19)$$

$$g_{2,\lambda T}(x, \hat{\theta}_T^{\mathcal{L}}) = 0. \quad (20)$$

It follows that long estimates for θ_1 are asymptotically identical to the efficient estimates for these parameters (they are numerically identical to the adjusted-moment estimates and asymptotically identical to the over-identified estimates). However, the long estimates of θ_2 are numerically identical to the short estimates, not to the efficient estimates.[12] This follows because the efficient estimates for θ_2 solve

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

rather than (20).

This result at first seems paradoxical. It implies that separately considering observations on $f_1(x, \theta)$ from $1, \ldots, (1-\lambda)T$ and observations from $(1-\lambda)T+1, \ldots, T$ does not (asymptotically) affect the estimate of θ_1, but does affect the estimate of θ_2.[13] As we illustrate in the section that follows, this surprising result occurs because the separation uncovers the deviation of $g_{1,\lambda T}$ from zero. Because of the correlation between $g_{1,\lambda T}$ and $g_{2,\lambda T}$, the deviation of $g_{1,\lambda T}$ from zero implies that $g_{2,\lambda T}$ is also likely to deviate from zero. The efficient estimators make use of this information to construct an estimator of the mean of $f_2(x, \theta)$ that improves on $g_{2,\lambda T}$.

3 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 US data.

3.1 Data

For the U.S., we use the annual data of Shiller (1989, Chap. 26), which begin in 1871 and are updated through 2005. Stock returns, prices and earnings are for the S&P 500 index. The predictor variable we use is the ratio of previous ten-year earnings to current stock

[12] It is tempting to conclude that the lower variance for $\hat{\theta}_{1,T}^{\mathcal{L}}$ and the same variance for $\hat{\theta}_{2,T}^{\mathcal{L}}$ implies 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.

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

price. We refer to this as the smoothed earnings-price ratio. This ratio is motivated by the present-value formula linking the earnings-price 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 (2007)). The riskfree rate is the return on six-month commercial paper purchased in January and rolled over in July. Because the first ten 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 website. The raw data on international indices come from Morgan Stanley's Capital International Perspectives (MSCI). 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 sub-indices. These are Asia-Pacific (Australia, Hong Kong, Japan, Malaysian, New Zealand, Singapore), Europe without the UK (Austria, Belgium, Switzerland, Germany, Spain, France, Italy, Netherlands), Europe with the UK (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 data set described above from the log of these returns to create real continuously compounded returns.

3.2 Applying the estimators

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

$$r_{1,t+1} = \alpha_1 + \beta_1 z_t + \epsilon_{1,t+1} \tag{21}$$

$$r_{2,t+1} = \alpha_2 + \beta_2 z_t + \epsilon_{2,t+1} \tag{22}$$

jointly for S&P 500 and international index excess returns, where z_t is the smoothed earnings-price ratio on the S&P. Moment conditions are determined by

$$f_1(x_t, \theta) = \begin{bmatrix} 1 \\ z_t \end{bmatrix} (r_{1,t+1} - \alpha_1 - \beta_1 z_t) \tag{23}$$

$$f_2(x_t, \theta) = \begin{bmatrix} 1 \\ z_t \end{bmatrix} (r_{2,t+1} - \alpha_2 - \beta_2 z_t), \tag{24}$$

where $x_t = (r_{1,t}, r_{2,t}, z_{t-1})$,

$$\theta_i = [\alpha_i, \beta_i]^\top, \quad i = 1, 2,$$

and $\theta = [\theta_1^\top \ \theta_2^\top]^\top$. The regression coefficients are identified by the conditions

$$E\left[f_1(x_t, \theta_0)\right] = E\left[f_2(x_t, \theta_0)\right] = 0.$$

The system is exactly identified and f_1 is sufficient to identify α_1 and β_1. Therefore we are in the setting of Section 2. Moreover, α_1 and β_1 do not appear as arguments in f_2. The source of the gain in estimating α_2 and β_2 will therefore be the correlation in the moment conditions, which arises from the correlation between shocks to $r_{1,t}$ and $r_{2,t}$, as shown in Section 2. We refer to the two moment conditions implied by f_1 as the long-history moment conditions and the moment conditions implied by f_2 as the short-history moment conditions.

Define matrices

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

and similarly,

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

and

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

The partial sums in Section 1.1 can then be written as

$$g_{1,T}(x,\theta) = \frac{1}{T}Z_T^\top\left(R_{1,T} - Z_T\theta_1\right) \tag{25}$$

$$g_{1,(1-\lambda)T}(x,\theta) = \frac{1}{(1-\lambda)T}Z_{(1-\lambda)T}^\top\left(R_{1,(1-\lambda)T} - Z_{(1-\lambda)T}\theta_1\right) \tag{26}$$

$$g_{1,\lambda T}(x,\theta) = \frac{1}{\lambda T}Z_{\lambda T}^\top\left(R_{1,\lambda T} - Z_{\lambda T}\theta_1\right) \tag{27}$$

$$g_{2,\lambda T}(x,\theta) = \frac{1}{\lambda T}Z_{\lambda T}^\top\left(R_{2,\lambda T} - Z_{\lambda T}\theta_2\right) \tag{28}$$

The short estimator is the solution to equations defined by setting (27) and (28) to zero. This is the same as ordinary least squares (OLS) regression over the 1975–2005 period. The adjusted-moment 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 matrix $\hat{B}_{21,\lambda T}$. Given $\hat{B}_{21,\lambda T}$, the adjusted-moment estimator is the solution to equations defined by setting (25) 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 zero. 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 which, as we show below, can be quite substantial.

While the adjusted-moment and short estimators are exactly identified, the over-identified estimator is not, as its name suggests. Moment conditions for the over-identified estimator are (26), (27) and (28). 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 (8). 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 1 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

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

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. Specifically, to obtain an estimate for the 4×4 matrix S, we take the moment conditions (23) and (24) evaluated at the adjusted-moment estimates and apply Stambaugh's method to the resulting time series of errors.[14] While Stambaugh assumes that the errors are normal, independent, and identically distributed, the resulting estimate will be consistent for S under a broader set of assumptions, including heteroskedasticity, as can be seem by applying arguments of White (1980). We then use submatrices of $\hat{S}_T$ to construct $\hat{S}_T^{\mathcal{A}}$ as in (7) and $\hat{S}_T^{\mathcal{I}}$ as in (8). The inverse of $\hat{S}_T^{\mathcal{I}}$ forms the weighting matrix for the over-identified estimator.

3.3 Results

Prior to reporting the results from 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 is less complicated than the implementation of the predictability estimation described above. The first two columns of Table 1 report results and standard errors for the short estimator; the second two columns report results and standard errors for the adjusted-moment and over-identified estimators. Because these are numerically equivalent in the setting of estimating means, we jointly to them as efficient estimators. As the columns for short show, the sample mean for excess returns on the S&P 500 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%.

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 and the EAFE portfolios is 0.67). It is this correlation that leads to the reduced standard

[14] This method can be outlined as follows. 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 over the sample period that is available for all moment conditions. We take the following as the estimator of S:

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

errors. In particular, the fact that the mean return for the S&P 500 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 efficients estimators therefore adjust the mean excess return on the EAFE downward.

Estimation for the sub-indices 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 exceed 0.70. As shown in Section 2, higher correlations between the moment conditions lead to greater improvement for the short-history asset.

Table 2 reports results of estimating the predictive regressions. In the short sample, the point estimates for all the portfolios are positive but insignificant: t-statistics are below 1 for all portfolios. The R^2 values are also small, e.g. 0.6% for the S&P 500. There is more evidence for predictability in the longer sample. For the S&P 500, the adjusted-moment method leads to an estimated coefficient of 0.093 with a standard error of 0.038 and an R^2 of 4.2%. The over-identified method leads to an estimated coefficient of 0.065 and an R^2 of 2.1%.

Including the earlier period of the sample has a substantial impact on the estimation for the EAFE and other short-history assets. The adjusted-moment 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 to 0.097. The implied R^2 is 12%, up from 3.8% when the short method is used. Results for the over-identified estimator are similar: the coefficient is 0.101 with an R^2 of 7.3%. The reason for this change lies in the covariance between the errors of the moment coefficients, as captured by the 2×2 matrix $\hat{B}_{21}$. This covariance, in turn, originates in the correlation between the U.S. and EAFE returns (0.65 as show in Table 3). The intuition for the effect is as follows: because estimated predictability is lower in the later part of the sample period than in the full period, it must be that positive shocks to the predictor variable are followed by, on average, negative shocks to S&P 500 returns over the later period. These will also tend to be followed by negative shocks to EAFE returns, because of the positive correlation between the S&P 500 and the EAFE. Therefore, the short estimator will understate the amount of predictability for the EAFE, and the efficient estimators adjust this upward. The resulting estimates have less noise, as represented by the smaller standard errors in Table 2. Similar effects are present for

sub-indices of the EAFE.

3.4 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 2.

The results in Section 2 imply that the long estimate for the predictability coefficient β_1 is numerically identical to the adjusted-moment estimate and asymptotically equal to the over-identified estimate of this coefficient. Both the long and the adjusted-moment estimate are equal to the value obtained from an OLS regression of S&P 500 returns on the predictor variable over the full sample of data. In contrast, the long estimate for β_2 is not equal to the adjusted-moment or over-identified 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 adjusted-moment estimate and the over-identified estimate 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 into two 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. However, this division does create more information: it uncovers the fact that there is less predictability in the S&P 500 over the 1975–2005 period than over the full period. Because the regression estimate is consistent, this could only occur if, on average, shocks to the predictor variable were followed by shocks to the S&P 500 return of the opposite sign over this period. Because of the positive correlation between returns on the S&P 500 and the EAFE, shocks to the predictor variable would also tend to be followed by shocks to the EAFE of the opposite sign. The efficient estimators use this information, in this case, to shift the estimate of β_2 upward.

4 Monte Carlo Analysis

In the previous sections we introduced two 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 apparent 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), Stambaugh (1999)).[15]

We simulate from the system (21)–(22) using the adjusted-moment regression coefficients to determine the data-generating process. We augment this system with an autoregression for the log of the smoothed earnings-price ratio z_t:

$$z_{t+1} = \rho_0 + \rho_1 z_t + \epsilon_{z,t+1}. \tag{29}$$

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 (21), (22) and (29) using the method of Stambaugh (1997) as described above. Table 3 reports the variances and correlations. The contemporaneous correlation between innovations to z_t and to S&P 500 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 earnings-price 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 are also highly correlated with innovations to international returns: this correlation is 0.65 for the EAFE and over 0.70 for the European sub-indices. Therefore it is reasonable to expect that incorporating the earlier data period will affect the precision of the estimates for the short-history assets.

For each international index, we simulate 50,000 samples of returns on the S&P 500, values for the predictor variable, and returns on that index. The sample length for the S&P 500 (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, adjusted-moment and over-identified 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).

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

Table 4 shows that the asymptotic efficiency gains discussed in Section 3.3 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 adjusted-moment and over-identified methods. 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 adjusted-moment method delivers a standard deviation of 0.134, the over-identified method 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 adjusted-moment and over-identified 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 1 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 adjusted-moment and over-identified 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 long-history asset is 0.120 under the short estimator, it is 0.028 under the adjusted-moment estimator and 0.015 under the over-identified 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 adjusted-moment and over-identified estimators, the bias is about equal to zero (it is in fact very slightly negative for the over-identified 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 adjusted-moment and the over-identified estimator 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.

5 Extensions

In this section we briefly outline how our estimators can be extended to more general patterns of missing data. We focus on the over-identified estimator which has a direct extension.[16]

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 up into disjoint intervals $1, \ldots, n$. Let λ_1 denote the ratio of the length of the first region to the length of the entire sample, λ_2 the ratio of the length of the second 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 first data segment begins at $t_1 + 1$, the second data segment at $t_2 + 1$, etc. Then

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

For the case described in Section 1, the first segment consist of points 1 to $(1-\lambda)T$, while the second segment consists of points $(1-\lambda)T + 1$ to T. We adopt the same notational convention as in Section 1: $\lambda_j t$ 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

$$f_{\phi_j}(x_t, \theta) \quad = \quad \Big(f_{i_1}(x_t, \theta), \ \ldots, \ f_{i_{\pi_j}}(x_t, \theta) \Big)^\top,$$

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

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

and the $\pi_j \times \pi_j$ matrices

$$R_{\phi_j}(\tau) = E\left[f_{\phi_j}(x_0, \theta_0) f_{\phi_j}(x_{-\tau}, \theta_0)^\top \right]$$

[16] The extension for the adjusted-moment estimator as well as examples for various patterns of missing data can be found in the working paper Lynch and Wachter (2004).

and

$$S_{\phi_j} = \sum_{\tau=-\infty}^{\infty} R_{\phi_j}(\tau).$$

Define

$$h_T^{\mathcal{I}_n}(\theta) = \left[g_{\phi_1,\lambda_1 T}(\theta)^\top,\ g_{\phi_2,\lambda_2 T}(\theta)^\top, \ldots,\ g_{\phi_n,\lambda_n T}(\theta)^\top \right]^\top. \tag{30}$$

The $\mathcal{I}_n$ superscript refers to the fact that these are moment conditions for the over-identified estimator, and that there are n non-overlapping intervals. The T subscript refers to the fact that the data length is T.[17] 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 over-identified estimator as

$$\hat{\theta}_T^{\mathcal{I}_n} = \operatorname{argmin} h_T^{\mathcal{I}_n}(\theta)^\top \left(\hat{S}_T^{\mathcal{I}_n} \right)^{-1} h_T^{\mathcal{I}_n}(\theta). \tag{31}$$

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

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

We now state a result analogous to Theorem 1.1. 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 over-identified estimator relative to the over-identified estimator defined over the first $n-1$ blocks of data. The proof (available from the authors) is similar to that of Theorem 1.1.

Theorem 5.1 *Assume the over-identified estimator $\hat{\theta}_T^{\mathcal{I}_n}$ is defined as (31). The Assumptions in Appendix A imply that 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.*

[17] This notation does not, of course, completely define the over-identified 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.

6 Conclusion

This paper has introduced two estimators that extend the generalized method of moments 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 unequal length, require the researcher to truncate the data so that all series are observed over the same interval. This paper has provided an alternative that allows the researcher to use all the data available for each moment condition.

Under assumptions of mixing and stationarity, we demonstrated 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 assumed that the two series had the same end date but different start dates. We then generalized 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 non-overlapping portion of the data is not limited to estimating moment conditions which 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 two estimators are as straightforward to implement as standard GMM and have intuitive interpretations. The adjusted-moment 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 over-identified estimator uses the non-overlapping data to form additional moment conditions. These two estimators are equivalent asymptotically, and superior to standard GMM, but differ in finite samples. We leave the question of which estimator has superior finite-sample properties to future work.

Appendix

A Stochastic setting

Let $\{x_t\}_{t=-\infty}^{\infty}$ denote a p-component stochastic process defined over an underlying 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 $\mathbf{R}^q$. The function f provides the restrictions that determine θ based on the observations of x_t. Following White and Domowitz (1984), define

$$\alpha(\mathcal{F}, \mathcal{G}) \equiv \sup_{\{F \in \mathcal{F}, G \in \mathcal{G}\}} |P(FG) - P(F)P(G)|$$

for σ-algebras $\mathcal{F}$ and $\mathcal{G}$, and

$$\alpha(s) \equiv \sup_t \alpha\left(\mathcal{F}_{-\infty}^t, \mathcal{F}_{t+s}^{\infty}\right).$$

The process $\{x_t\}$ is said to be α-mixing if $\alpha(s) \to 0$ as $s \to \infty$. As White and Domowitz (1984) discuss, α-mixing guarantees that autocovariances vanish at at arbitrarily long lags. Mixing is a convenient assumption because it allows a trade-off between the speed at which $\alpha(s)$ approaches zero and the conditions required on the function f. An ARMA process, for example, entails relatively fast convergence of $\alpha(s)$, and thus requires only weak conditions on f.

A process is said to be α-mixing of size $r/(r-1)$ for $r > 1$ if for some $\kappa > r/(r-1)$, $\alpha(s)$ is $O(s^{-\kappa})$. We assume that $\{x_t\}$ is mixing:

Assumption 1 *$\{x_t\}_{t=-\infty}^{\infty}$ is α-mixing of size $\frac{r}{r-1}$ for $r > 1$, and stationary.*

Assumption 2 guarantees that $f(x_t, \theta)$ is also mixing.

Assumption 2 *$f(\cdot, \theta)$ is measurable for all $\theta \in \Theta$.*

The following assumption specifies the sense in which $f(x_t, \theta)$ determines θ given observations on x_t.

Assumption 3 *There exists a unique $\theta_0 \in \Theta$ such that $E[f(x_t, \theta_0)] = 0$.*

The next assumptions form the basis for the consistency and asymptotic normality results of estimators based on partial sums of $f(x_t, \theta)$.

Assumption 4 *There exists* $\Delta \in \mathbf{R}$ *such that* $E\left(\left|f_i(x_t,\theta_0)^{2r}\right|\right) < \Delta$, $i = 1,\ldots,l$.

Assumption 5 $f(x_t,\theta)$ *is continuous in* θ. *There exists a measurable function* $H(x_t) \in \mathbf{R}^l$ *such that* $|f_i(x_t,\theta)| \leq H_i(x_t)$ *for all* $\theta \in \Theta$ *and such that* $E|H_i(x_t)|^{r+\delta} \leq \Delta < \infty$, *for some* $\delta > 0$ *and all* $i = 1,\ldots,l$.

Assumptions 4 and 5 illustrate the usefulness of the definition of mixing. As White and Domowitz (1984) explain, the greater is r, the more dependence is allowed for the process x_t, but the stronger are the required conditions on the function f. For example, if x_t is independent then $\alpha(s) = 0$ for all s, and hence we can set $r = 1$. If x_t follows an ARMA process, r can be taken to be arbitrarily close to 1.

The following assumptions allow us to establish consistency and asymptotic normality.

Assumption 6 *For* $k \in \{\mathcal{S},\mathcal{L},\mathcal{A},\mathcal{I}\}$, *the weighting matrix* W_T^k *converges almost surely to a positive-definite matrix* W^k.

Assumption 7 θ_0 *lies in the interior of* Θ.

Assumption 8 $f(x,\theta)$ *is continuously differentiable in* θ.

Assumption 9 *There exists a measurable matrix-valued function* $\hat{H}(x_t) \in \mathbf{R}^{l\times q}$ *such that* $|\frac{\partial f_i}{\partial \theta_j}(x_t,\theta)| < \hat{H}(x_t)_{(i,j)}$ *for all* θ *in the interior of* Θ *and such that for some* $\delta > 0$, $E|\hat{H}(x_t)_{(i,j)}|^{r+\delta} \leq \Delta < \infty$ *for all* $i = 1,\ldots,l$, $j = 1,\ldots q$.

B Independence results

Define

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

It is useful to slightly generalize the notation of Section 1. Let

$$g_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} f(x_t,\theta) \tag{32}$$

$$g_{(1-\lambda)T}(\theta) = \frac{1}{(1-\lambda)T}\sum_{t=1}^{(1-\lambda)T} f(x_t,\theta) \tag{33}$$

$$g_{\lambda T}(\theta) = \frac{1}{\lambda T}\sum_{t=(1-\lambda)T+1}^{T} f(x_t,\theta). \tag{34}$$

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

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

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

Then Assumptions 1–3 and 5 imply that

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

PROOF: For any integer T,

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

where $\lfloor\sqrt{T}\rfloor$ is the largest integer less than the square root of T. Assumptions 1–3, and 5 imply that

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

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

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

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

White and Domowitz (1984) show that w_t is α-mixing. Therefore $\alpha(\sqrt{T})$ goes to 0 as $T \to \infty$. By the Slutsky theorem,

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

where the second line follows because w_t is α-mixing, and the last line follows from a repeated application of the Slutsky Theorem. □

The following theorem states that partial sums taken over disjoint intervals are asymptotically independent and normally distributed. Related results appear in the literature on GMM and structural breaks; see Andrews and Fair (1988), Ghysels and Hall (1990), Andrews (1993), Andrews and Ploberger (1994), Sowell (1996), and Ghysels, Guay, and Hall (1997).

Theorem B.1 *Assumptions 1–5 imply that as* $T \to \infty$,

$$\sqrt{T}\begin{bmatrix} \sqrt{(1-\lambda)}g_{(1-\lambda)T}(\theta_0) \\ \sqrt{\lambda}g_{\lambda T}(\theta_0) \end{bmatrix} \to_{\mathrm{d}} N\left(0, \begin{bmatrix} S & 0 \\ 0 & S \end{bmatrix}\right). \tag{35}$$

Proof: Assumptions 1–4 imply that

$$\sqrt{(1-\lambda)T}g_{(1-\lambda)T}(\theta_0) \to_{\mathrm{d}} N(0,S) \tag{36}$$

and

$$\sqrt{\lambda T}g_{\lambda T}(\theta_0) \to_{\mathrm{d}} N(0,S) \tag{37}$$

by Theorem 2.4 of White and Domowitz (1984). Stationarity of x_t (Assumption 1) 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 B.1,

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

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 (36) and (37). □

C Deriving the Asymptotic Distribution

This Appendix derives the asymptotic distribution for the four estimators we consider. For notational convenience, it is useful to define functions h^k that give the moment conditions for each estimator.

$$h_T^{\mathcal{S}}(\theta) = \left[g_{1,\lambda T}(\theta)^\top \; g_{2,\lambda T}(\theta)^\top\right]^\top \tag{38}$$

$$h_T^{\mathcal{L}}(\theta) = \left[g_{1,T}(\theta)^\top \; g_{2,\lambda T}(\theta)^\top\right]^\top \tag{39}$$

$$h_T^{\mathcal{A}}(\theta) = \left[g_{1,T}(\theta)^\top \; \left(g_{2,\lambda T}(\theta) + \hat{B}_{21,\lambda T}(1-\lambda)(g_{1,(1-\lambda)T}(\theta) - g_{1,\lambda T}(\theta))\right)^\top\right]^\top \tag{40}$$

$$h_T^{\mathcal{I}}(\theta) = \left[g_{1,(1-\lambda)T}(\theta)^\top \; g_{1,\lambda T}(\theta)^\top \; g_{2,\lambda T}(\theta)^\top\right]^\top, \tag{41}$$

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

$$\theta^k = \text{argmin}_\theta \, h_T^k(\theta)^\top W_T^k h_T^k(\theta), \tag{42}$$

Theorem C.1 *Assumptions 1–5 imply that as $T \to \infty$,*

$$\sqrt{\lambda T} h_T^k(\theta_0) \to_{\text{d}} N(0, S^k),$$

where $S^{\mathcal{S}} = S$, and $S^{\mathcal{L}}$, $S^{\mathcal{A}}$ and $S^{\mathcal{I}}$ are defined in (6)–(8).

Proof: The result for the short estimator follows directly from Theorem B.1. To illustrate the proof for the remaining matrices, we derive (7); the proofs of (6) and (8) are similar. In what follows, the argument θ_0 is suppressed and convergence is in the sense of almost surely.

Stationarity implies that $S_{11}^{\mathcal{A}} = \lambda S_{11}$. By Theorem B.1,

$$\begin{aligned}\lim_{T\to\infty} E\left[\sqrt{\lambda T}\left(\lambda g_{i,\lambda T} + (1-\lambda) g_{i,(1-\lambda)T}\right)\sqrt{\lambda T}\left(g_{j,(1-\lambda)T} - g_{j,\lambda T}\right)^\top\right] \\ = \lim_{T\to\infty}\left(-E\left[\sqrt{\lambda T}\lambda g_{i,\lambda T}\sqrt{\lambda T} g_{j,\lambda T}^\top\right] + E\left[\sqrt{\lambda T}(1-\lambda) g_{i,(1-\lambda T)}\sqrt{\lambda T} g_{j,(1-\lambda)T}^\top\right]\right) \\ = \lambda S_{ij} - \lambda S_{ij} = 0\end{aligned} \tag{43}$$

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

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

The second line follows from (43) and the third and fourth lines follow from Theorem B.1. Using similar reasoning,

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

which completes the derivation of (7). □

Theorem C.2 establishes consistency of the estimators.

Theorem C.2 *Assumptions 1–6 imply that as $T \to \infty$, $\hat{\theta}_T^k \to_{\text{a.s.}} \theta_0$ for $k \in \{\mathcal{S}, \mathcal{L}, \mathcal{A}, \mathcal{I}\}$.*

PROOF: White and Domowitz (1984) show that under these assumptions

$$|g_{\lambda T}(\theta) - E f(x_t, \theta)| \to_{\text{a.s.}} 0$$
$$|g_{(1-\lambda)T}(\theta) - E f(x_t, \theta)| \to_{\text{a.s.}} 0$$

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

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

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

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

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

For convenience, define the notation

$$D_0^k = D_0 \quad k \in \{\mathcal{S}, \mathcal{L}, \mathcal{A}\} \tag{44}$$

$$D_0^{\mathcal{I}} = \left[D_{0,1}^\top \; D_{0,1}^\top \; D_{0,2}^\top\right]^\top. \tag{45}$$

The following theorem establishes asymptotic normality.

Theorem C.3 *Assumptions 1–9 imply*

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

Proof: Define

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

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

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

Pre-multiplying by $D_T^k(\hat{\theta}_T^k)^\top W_T^k$:

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

By the first-order condition of the optimization problem,

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

The assumptions and Theorem 2.3 of White and Domowitz (1984) imply that

$$D_T^k(\theta) \to_{\text{a.s.}} E\left[\frac{\partial f}{\partial \theta}(x_t, \theta)\right]$$

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

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

uniformly in θ. Therefore by Theorem C.2 and Assumptions 7 and 8, Amemiya (1985, Theorem 4.1.5) implies

$$D_T^k(\hat{\theta}_T^k) \to_{\text{a.s.}} D_0^k \tag{46}$$

$$D_T^k(\tilde{\theta}^k) \to_{\text{a.s.}} D_0^k \tag{47}$$

$$W_T^k \to_{\text{a.s.}} W^k. \tag{48}$$

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 C.4 *Suppose* $W_{\lambda T}^k \to_{\text{a.s.}} W_k = (S^k)^{-1}$. *Then Assumptions 1–5 and 7-9*

$$\sqrt{\lambda T}(\hat{\theta}_T^k - \theta_0) \to_{\text{d}} N\left(0, \left((D_0^k)^\top \left(S^k\right)^{-1} (D_0^k)\right)^{-1}\right). \tag{49}$$

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

D Matrix Algebra Results

Lemma D.1 *Assume* $m \times m$ *matrices* U_1 *and* U_2 *are invertible. If* $U_1 - U_2$ *is positive semi-definite, then* $U_2^{-1} - U_1^{-1}$ *is also positive semi-definite.*

PROOF: See Goldberger (1964, Chapter 2.7). □

Lemma D.2 *Assume* $m \times m$ *matrices* U_1 *and* U_2 *are invertible. If* $U_1 - U_2$ *is positive semi-definite, then for any conforming matrix* M, $(M^\top U_1^{-1} M)^{-1} - (M^\top U_2^{-1} M)^{-1}$ *is also positive semi-definite.*

PROOF: Assume $U_1 - U_2$ is positive semi-definite. By Lemma D.1, $U_2^{-1} - U_1^{-1}$ is positive semi-definite. For any vector v and matrix M,

$$(Mv)^\top (U_2^{-1} - U_1^{-1})(Mv) \geq 0.$$

Therefore

$$v^\top M^\top (U_2^{-1} - U_1^{-1}) M v \geq 0$$

which shows $M^\top (U_2^{-1} - U_1^{-1}) M$ is positive semi-definite. Applying Lemma D.1 a second time shows that $(M^\top U_1^{-1} M)^{-1} - (M^\top U_2^{-1} M)^{-1}$ is positive semi-definite as required. □

Lemma D.3 *Let*

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$$

be a symmetric invertible matrix. Then

$$S^{-1} = \begin{bmatrix} S_{11}^{-1} + B_{21}^{\top}\Sigma^{-1}B_{21} & -B_{21}^{\top}\Sigma^{-1} \\ -\Sigma^{-1}B_{21} & \Sigma^{-1} \end{bmatrix}, \tag{50}$$

where Σ is defined by (1). Moreover, if $\bar{S}$ is defined as

$$\bar{S} = \begin{bmatrix} \lambda S_{11} & \lambda S_{12} \\ \lambda S_{21} & S_{22} - (1-\lambda)S_{21}S_{11}^{-1}S_{12} \end{bmatrix},$$

with $\lambda \neq 0$, then

$$\bar{S}^{-1} = \begin{bmatrix} \frac{1}{\lambda}S_{11}^{-1} + B_{21}^{\top}\Sigma^{-1}B_{12} & -B_{21}^{\top}\Sigma^{-1} \\ -\Sigma^{-1}B_{12} & \Sigma^{-1} \end{bmatrix}. \tag{51}$$

PROOF: The first statement follows from the expression for the matrix inverse (see e.g. Green (1997, Chapter 2)). Applying the same formula to $\bar{S}$ results in

$$\bar{S}^{-1} = \begin{bmatrix} \bar{S}_{11}^{-1} + \bar{B}_{21}^{\top}\bar{\Sigma}^{-1}\bar{B}_{21} & -\bar{B}_{21}^{\top}\bar{\Sigma}^{-1} \\ -\bar{\Sigma}^{-1}\bar{B}_{12} & \bar{\Sigma}^{-1} \end{bmatrix},$$

where

$$\bar{B}_{21} = \bar{S}_{21}\left(\bar{S}_{11}\right)^{-1} = S_{21}S_{11}^{-1} = B_{21},$$

and

$$\begin{aligned} \bar{\Sigma} &= \bar{S}_{22} - \bar{S}_{21}\bar{S}_{11}^{-1}\bar{S}_{12} \\ &= S_{22} - (1-\lambda)S_{21}S_{11}^{-1}S_{12} - \lambda S_{21}S_{11}^{-1}S_{12} = \Sigma. \end{aligned}$$

Therefore (51) holds. □

References

Ahn, Seung C., and Peter Schmidt, 1995, A separability result for GMM estimation with applications to GLS prediction and conditional moment tests, *Econometric Reviews* 14, 19–34.

Ait-Sahalia, Yacine, Per A. Mykland, and Lan Zhang, 2005, How often to sample a continuous-time process in the presence of market microstructure noise, *Review of Financial Studies* 18, 351–416.

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

Anderson, T. W., 1957, Maximum likelihood estimates for a multivariate normal distribution when some observations are missing, *Journal of the American Statistical Association* 52, 200–203.

Andrews, Donald W. K., 1993, Tests for parameter instability and structural change with unknown change point, *Econometrica* 61, 821–856.

Andrews, Donald W. K., and Ray C. Fair, 1988, Inference in Nonlinear Econometric Models with Structural Change, *Review of Economic Studies* 55, 615–639.

Andrews, Donald W. K., and Werner Ploberger, 1994, Optimal tests when a nuisance parameter is present only under the alternative, *Econometrica* 62, 1383–1414.

Bandi, Frederico M., and Jeffrey R. Russell, 2004, Separating microstructure noise from volatility, forthcoming, *Journal of Financial Economics*.

Brandt, Michael W., 1999, Estimating portfolio and consumption choice: A conditional Euler equations approach, *Journal of Finance* 54, 1609–1645.

Burguete, Jose F., A. Ronald Gallant, and Geraldo Souza, 1982, On unification of the asymptotic theory of nonlinear econometric models, *Econometric Reviews* 1, 151–190.

Campbell, John Y., and Robert J. Shiller, 1988, Stock prices, earnings, and expected dividends, *Journal of Finance* 43, 661–676.

Campbell, John Y., and Samuel Thompson, 2007, Predicting excess stock returns out of sample: Can anything beat the historical average?, forthcoming, *Review of Financial Studies*.

Cavanagh, Christopher L., Graham Elliott, and James H. Stock, 1995, Inference in models with nearly integrated regressors, *Econometric Theory* 11, 1131–1147.

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

Conniffe, Denis, 1985, Estimating regression equations with common explanatory variables but unequal numbers of observations, *Journal of Econometrics* 27, 179–196.

Duffie, Darrell, and Kenneth J. Singleton, 1993, Simulated moments estimation of Markov models of asset prices, *Econometrica* 61, 929–952.

Fama, Eugene F., and Kenneth R. French, 1989, Business conditions and expected returns on stocks and bonds, *Journal of Financial Economics* 25, 23–49.

Ghysels, Eric, Alain Guay, and Alastair Hall, 1997, Predictive tests for structural change with unknown breakpoint, *Journal of Econometrics* 82, 209–233.

Ghysels, Eric, and Alastair Hall, 1990, A test for structural stability of Euler conditions parameters estimated via the generalized method of moments estimator, *International Economic Review* 31, 335–364.

Ghysels, Eric, Pedro Santa-Clara, and Rossen Valkanov, 2005, There is a risk-return trade-off after all, *Journal of Financial Economics* 76, 509–548.

Goetzmann, William N., and Philippe Jorion, 1999, Re-emerging markets, *Journal of Financial and Quantitative Analysis* pp. 1–32.

Goldberger, Arthur S., 1964, *Econometric Theory*. (John Wiley and Sons New York).

Green, William H., 1997, *Econometric Analysis*. (Prentice-Hall, Inc. Upper Saddle River, NJ).

Hansen, Lars Peter, 1982, Large sample properties of generalized method of moments estimators, *Econometrica* 50, 1029–1054.

Hansen, Lars Peter, and Ken Singleton, 1982, Generalized instrumental variables estimation of nonlinear rational expectations models, *Econometrica* 50, 1269–1286.

Harvey, Andrew, Siem Jan Koopman, and Jeremy Penzer, 1998, Messy time series: A unified approach, in *Advances in Econometrics, Volume 13* (JAI Press Inc.,).

Little, Roderick J. A., and Donald B. Rubin, 2002, *Statistical analysis with missing data*. (John Wiley & Sons Hoboken, NJ) 2 edn.

Lynch, Anthony W., and Jessica A. Wachter, 2004, Using samples of unequal length in generalized method of moments estimation, New York University Working Paper FIN-05-021.

Nelson, C. R., and M. J. Kim, 1993, Predictable stock returns: The role of small sample bias, *Journal of Finance* 48, 641–661.

Newey, Whitney K., and Daniel McFadden, 1994, Large sample estimation and hypothesis testing, in R.F. Engle, and D. L. McFadden, eds.: *Handbook of Econometrics, Volume IV* (North-Holland, Amsterdam).

Pastor, Lubos, and Robert F. Stambaugh, 2002a, Investing in equity mutual funds, *Journal of Financial Economics* 63, 351–380.

Pastor, Lubos, and Robert F. Stambaugh, 2002b, Mutual fund performance and seemingly unrelated assets, *Journal of Financial Economics* 63, 315–349.

Phillips, Peter C.B., 1987, Time series regressions with a unit root, *Econometrica* 55, 277–301.

Robins, James M., and Andrea Rotnitsky, 1995, Semiparametric efficiency in multivariate regression models with missing data, *Journal of the American Statistical Association* 90, 122–129.

Schmidt, Peter, 1977, Estimation of seemingly unrelated regressions with unequal numbers of observations, *Journal of Econometrics* 5, 365–377.

Shiller, Robert J., 1989, *Market volatility*. (MIT Press Cambridge, MA).

Singleton, Kenneth, 2006, *Empirical dynamic asset pricing: Model specification and econometric assessment*. (Princeton University Press Princeton, NJ).

Sowell, Fallaw, 1996, Optimal tests for parameter instability in the generalized method of moments framework, *Econometrica* 64, 1085–1107.

Stambaugh, Robert, 1997, Analyzing investments whose histories differ in length, *Journal of Financial Economics* 45, 285–331.

Stambaugh, Robert, 1999, Predictive regressions, *Journal of Financial Economics* 54, 375–421.

Stock, James H., 1994, Unit roots, structural breaks and trends, in R.F. Engle, and D. L. McFadden, eds.: *Handbook of Econometrics, Volume IV* (North-Holland, Amsterdam).

Storesletten, Kjetil, Chris I. Telmer, and Amir Yaron, 2004, Cyclical dynamics in idiosyncratic labor market risk, *Journal of Political Economy* 112, 695–717.

Swamy, P. A. V. B., and J. S. Mehta, 1975, On Bayesian estimation of seemingly unrelated regressions when some observations are missing, *Journal of Econometrics* 3, 157–169.

White, Halbert, 1980, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica* 24, 817–838.

White, Halbert, 1994, *Asymptotic Theory for Econometricians*. (Academic Press, Inc.).

White, Halbert, and Ian Domowitz, 1984, Nonlinear regression with dependent observations, *Econometrica* 52, 143–162.

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 riskfree rate. US refers to returns on the S&P 500; EAFE refers to returns on an index for Europe, Asia and the Far East; Asia-Pacific, Europe, Europe without UK and Scandinavia are sub-indices of the EAFE. "Short" denotes estimates obtained using standard GMM; "Efficient" denotes estimates obtained using the adjusted-moment method or over-identified method, which are numerically identical in this application. Standard errors are computed using efficient estimates and are robust to heteroskedasticity. Data for the US 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	
	Mean	SE	Mean	SE
US	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 UK	5.40	4.17	3.69	3.09
Scandinavia	7.00	4.58	5.27	3.60

Table 2: Predictive Regressions for Excess Returns on International Indices

Predictive regressions are estimated in annual data for excess returns on international indices. The table 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 ratio. Returns are annual, continuously compounded, and in excess of the riskfree rate. US refers to returns on the S&P 500; EAFE refers to returns on an index for Europe, Asia and the Far East; Asia-Pacific, Europe, Europe without UK and Scandinavia are sub-indices of the EAFE. "Short" denotes standard GMM; "AM" denotes the adjusted-moment method; "OI" denotes the over-identified method. Standard errors are computed using AM estimates and are robust to heteroskedasticity. Data for the US span the 1881-2005 period; data for the other indices span the 1975–2005 period.

	Short			AM			OI		
	Coef.	SE	R^2	Coef.	SE	R^2	Coef.	SE	R^2
US	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 UK	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

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 US return, an international index return (EAFE or sub-index of the EAFE) and the predictor variable, the log of the smoothed earnings-price ratio. Returns are annual, continuously compounded, and in excess of the riskfree rate. US refers to returns on the S&P 500; EAFE refers to returns on an index for Europe, Asia and the Far East; Asia-Pacific, Europe, Europe without UK and Scandinavia are sub-indices of the EAFE. Data for the US 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 US returns are annual and span the 1881–2005 period.

	Standard deviation	Correlation with $\log(E/P)$	Correlation with U.S.
$\log(E/P)$	0.179		
US	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 UK	0.229	-0.666	0.769
Scandinavia	0.255	-0.578	0.710

Table 4: Predictive Regressions in Repeated Samples

50,000 samples of returns are simulated assuming joint normality of excess returns and the predictor variable. The 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 and a short-history asset calibrated to the EAFE or sub-index. 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; "OI" denotes the over-identified method.

	Short	AM	OI
US	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 UK	0.175	0.130	0.131
Scandinavia	0.194	0.156	0.157

Table 5: Bias in Predictive Coefficients

50,000 samples of returns are simulated assuming joint normality of excess returns and the predictor variable. The 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 and a short-history asset calibrated to the EAFE or sub-index. 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; "OI" denotes the over-identified method.

	Short	AM	OI
US	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 UK	0.119	0.023	0.008
Scandinavia	0.115	0.015	0.001