Generalized Method of Moments for Samples of Unequal Length [*]

Anthony W. Lynch[†]
New York University and NBER

Jessica A. Wachter[‡]
University of Pennsylvania and NBER

July 14, 2004

Abstract

This paper extends the generalized method of moments technique of Hansen (1982) to cases where moment conditions are observed over different sample periods. Many applications in financial economics use data series that have different starting dates or different ending dates. 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 describes an alternative that allows the researcher to make use of all of the data available for each moment condition. We describe two asymptotically equivalent estimators and show that these estimators are consistent, asymptotically normal, and more efficient asymptotically than standard GMM. We then extend both of these estimators to settings with more general patterns of missing data and show that the extended estimators are asymptotically more efficient than estimators that ignore intervals of the sample, even if some series are not observed over all these intervals. By implication, the extended estimators are more efficient than standard GMM.

[*] The authors would like to thank Martin Lettau, Andrew Lo, Robert Stambaugh, and Amir Yaron as well as seminar participants at New York University, and at the Wharton School for their comments and suggestions.

[†] Department of Finance, Stern School of Business, New York University, 44 West Fourth Street, New York, NY 10012-1126, alynch@stern.nyu.edu, (212) 998-0350.

[‡] Department of Finance, The Wharton School, University of Pennsylvania, 2300 SH-DH, Philadelphia, PA, 19104. jwachter@wharton.upenn.edu, (215) 898-7634.

Introduction

This paper extends the generalized method of moments technique of Hansen (1982) to cases where moment conditions are observed over different sample periods. 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 other series 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, or term structure data. 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 describes an alternative that allows the researcher to make use of all of the data available for each moment condition. A key question is whether using the full data set provides more reliable estimates of parameters. We will show that our estimator is indeed more efficient than standard GMM. However, not all ways of introducing the additional data are the same. It must be done carefully or efficiency may deteriorate rather than improve.

The problem of unequal sample lengths in financial time series was first addressed by Stambaugh (1997).[1] Stambaugh derives a maximum likelihood and a Bayesian estimator for the mean and the variance of a joint normal distribution, assuming returns are homoscedastic and independently distributed, in a setting where some return series start at a later date than others. Little and Rubin (2002) also derive maximum likelihood estimators when a portion of the data is missing (generally in non-economic applications), but their approach is similarly dependent on the specifics of the data generating process. 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. As shown in Cochrane (2001), most common estimation techniques 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. In particular, we adopt the mixing assumption

[1] 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.

of White and Domowitz (1984) as a means of limiting the temporal dependence of the underlying stochastic process.[2]

Because our method is based on GMM, the results we derive are asymptotic. Developing an asymptotic approach to a problem of missing data may at first seem strange. After all, asymptotics involve taking the sample size to infinity, which seems in opposition to the notion of missing data. We argue, however, that it is no more strange than applying asymptotics in the usual setting, where samples have the same length. In such cases, the number of data points is, of course, finite, so the asymptotic distribution must be treated as an approximation to the true sampling distribution of the parameters. Our asymptotic distribution can be thought of as an approximation in exactly the same way.

So that this approximation is not unreasonable, care must be taken to insure that the missing data problem does not become trivial as the sample size becomes large. For this reason, we 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 the both series lengths approach infinity.

Our initial setting supposes that some moment conditions are observed over the full data set, and some moment conditions are observed over a data set that has the same ending date but a later starting date (we later generalize this to other patterns of missing data). The two sets of moment conditions may depend on the same underlying parameters, or on different underlying parameters. We develop two asymptotically equivalent estimators that make use of all of the data. While general, these estimators are straightforward to implement (as shown in Lynch, Wachter, and Boudry (2004) for the case of performance evaluation of mutual funds), and have natural and intuitive interpretations.

The first estimator (which we call the *adjusted-moment* estimator) uses sample averages over the full sample to estimate the moments for which full-sample data are available, and sample averages over the short sample to estimate moments for which only the short-sample data are available. Then the moments for which only the short sample is available are "adjusted" using coefficients

[2] Under stationarity, mixing is a slightly stronger condition than ergodicity. Intuitively, mixing requires that autocovariances vanish as the lag length increases, but sufficiently slowly to allow processes with memories much longer than finite ARMA processes.

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 in 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. As we show, both estimators are consistent and asymptotically normal. Both estimators we derive are asymptotically more efficient than truncating the sample at the beginning of the shorter data set.

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). This pattern is analyzed in detail in a maximum likelihood setting for independent and identically distributed normal observations by Little and Rubin, and by Stambaugh (1997). Both of our estimators can be extended not only to this case, but further, to cases where 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. Despite the general nature of this problem, it is still possible to prove that the adjusted-moment and the over-identified estimator are asymptotically equivalent, though different in finite samples. Each preserves key properties of its counterpart when there are only two starting dates. Moreover, 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. The first section develops the asymptotic theory that will be the basis for the consistency and asymptotic normality proofs. The key result in this section is that sample averages (multiplied by the square root of the length of the sample) taken over disjoint intervals of data are independent as the number of data points in each interval approaches infinity. This result clearly holds when observations are independent. Even when there is dependence the result holds, provided that the process satisfies a mixing condition in the sense of White and Domowitz (1984). Intuitively, mixing insures that autocovariances are small at arbitrarily long lags. As the number of data points approaches infinity, the data from one partial sum that is "near" the data from the other partial sum becomes negligible in the overall average, implying that the partial sums are independent.

With this result as background, the second section defines four estimators in a setting where data are missing at the beginning of the sample for some of the moment conditions. The first of these estimators is the standard GMM estimator (we call it "short"). The second of these estimators combines the long and short data in a naive way (we call it "long"). The third and the fourth are the adjusted-moment and the over-identified estimator mentioned above. All four estimators are shown to be consistent and asymptotically normal under standard assumptions. Moreover, for each estimator, the efficient weighting matrix is the inverse of the variance-covariance matrix of the moments, just as in standard GMM.

The third section shows that the adjusted-moment estimator and the over-identified estimator are asymptotically equivalent. They are both more efficient than the short estimator (standard GMM), and more efficient than the long estimator (which takes into account the additional data in a naive way). The long estimator is not necessarily more efficient than the short estimator; thus including the additional data in a naive way could cause the efficiency of the estimators to deteriorate rather than improve. Fortunately, the adjusted-moment and the over-identified estimator are just as easy to compute as the short and long estimator. Finally, this section shows that in finite samples, the adjusted-moment and the over-identified estimators generally differ.

The fourth section investigates a special case in which the original system is exactly identified, and some variables can be identified by the long-sample data alone. In this case, it is possible to gain additional intuition about the forms of the adjusted-moment and over-identified estimators, and to estimate the size of the efficiency gain from using the adjusted-moment or the over-identified estimator. For simplicity, moments for which the full sample is available are assumed to depend on a subset of the parameters θ_1, while moments for which only the short sample is available depend on θ_2 as well as, possibly, θ_1. Asymptotic standard errors for θ_1 are a fraction $\sqrt{\lambda}$ of their values under standard GMM; thus the percent decrease is $1 - \sqrt{\lambda}$. For θ_2, $1 - \sqrt{\lambda}$ represents an upper bound on the percent decrease. The actual decrease depends on the correlation between the moment conditions, and the extent to which θ_1 influences the moments for which only the short sample is available.

The fifth section extends the analysis to the general case, where there can be an arbitrary number of starting dates, ending dates, and data can be missing in the middle of the sample. Data interval endpoints are identified by points in time at which data for at least one sample moment starts or ends. Asymptotic theory is developed assuming that the ratios of these various data intervals stay the stay same as the sample size grows large. Both the adjusted-moment and over-identified estimators are extended in this more general setting in natural ways. The over-identified estimator is obtained by treating the sample moments for each data interval as separate sample moments in the GMM estimation. The adjusted-moment estimator is defined inductively: the moments used when a data interval is added are obtained by taking the moments used without that interval and adding an adjustment term that uses the data in the added interval. It is shown that these extensions to the two estimators, while different in finite samples, are asymptotically equivalent, and moreover, that adding an additional data interval, even though not observed for all moments, improves efficiency. The sixth section concludes.

1 Large sample theory for sums covering different sample periods

Let $\{x_t\}_{t=-\infty}^{\infty}$ denote a p-component stochastic process defined over an underlying probability space $(\Omega, \mathcal{F}, P)$. Let $\mathcal{F}_a^b \equiv \sigma(x_t; a \le t \le b)$, 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. In what follows we make standard assumptions on $\{x_t\}$ and f in order to guarantee consistency and asymptotic normality of the estimates. Particularly useful is a notion of dependence known as mixing.

Following White and Domowitz (1984), define

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

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

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

The process $\{x_t\}$ is said to be α- mixing if $\alpha(m) \to 0$ as $m \to \infty$. 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 $\alpha(m)$ approaches zero and the conditions required on $\{x_t\}$. In particular, a process is said to be α-mixing of size $r/(r-1)$ for $r > 1$ if for some $\eta > r/(r-1)$, $\alpha(m)$ is $O(m^{-\eta})$. 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 $E\left[f(x_t, \theta_0)\right] = 0$.*

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 $M(x_t) \in \mathbf{R}^l$ such that $|f_i(x_t, \theta)| \leq M_i(x_t)$ for all $\theta \in \Theta$ and such that $E|M_i(x_t)|^{r+\delta} \leq \Delta < \infty$, 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 $\alpha(m) = 0$ for all m, and hence we can set $r = 1$. If x_t follows an ARMA process, r can be taken to be arbitrarily close to 1.

White and Domowitz (1984) prove the following:

Lemma 1.1 *Assumptions 1 and 2 imply that $\{f(x_t, \theta)\}_{t=-\infty}^{\infty}$ is $\alpha(m)$ mixing of size $r/(r-1)$ and stationary.*

Following Hansen (1982), define the $l \times l$ matrix $R(\tau) = E\left[f(x_0, \theta_0) f(x_{-\tau}, \theta_0)^\top\right]$ and let

$$S = \sum_{\tau=-\infty}^{\infty} R(\tau) = R(0) + \sum_{\tau=1}^{\infty} (R(\tau) + R(\tau)^\top). \tag{1}$$

Lemma 1.1 implies that this sum converges, because $\alpha(m)$ mixing combined with stationarity implies that the series is ergodic (see White (1994, Proposition 3.44)). Define

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

for $\theta \in \Theta$, and

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

Lemma 1.2 *Let $F \in \mathcal{F}_{-\infty}^{0}$. Let μ be $1 \times l$, and let a be a scalar. Then Assumptions 1–3 and 5 imply that*

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

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

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

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

By Lemma 1.1, w_t is α-mixing. Therefore $\alpha(\sqrt{T})$ goes to 0 as $T \to \infty$. By the Slutsky theorem,

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

where the second line follows from Lemma 1.1, and the last line follows from a repeated application of the Slutsky Theorem. ▮

Let λ be a rational number between 0 and 1, and 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

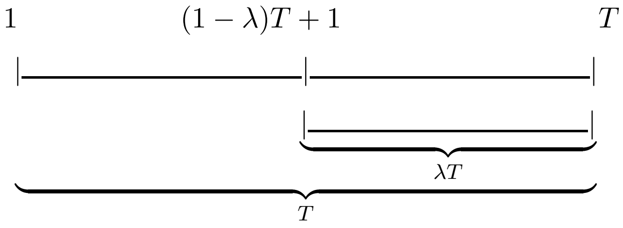

Figure 1: Notation for data missing at the start of the sample

multiple of n_0. For T a multiple of n_0, define

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

$$g_{\lambda T}(\theta) = \frac{1}{\lambda T}\sum_{t=(1-\lambda)T+1}^{T} f(x_t,\theta). \tag{3}$$

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 $1,\ldots,\lambda T$. Figure 1 illustrates the notation. Thus 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.

For the next theorem and in the remainder of the paper, we let T approach infinity along the subsequence of integer multiples of n_0.[3]

Theorem 1.1 *Assumptions 1–5 imply that as $T\to\infty$,*

$$\sqrt{T}\begin{pmatrix}\sqrt{(1-\lambda)}g_{(1-\lambda)T}(\theta_0)\\ \sqrt{\lambda}g_{\lambda T}(\theta_0)\end{pmatrix} \to_{\mathrm{d}} N\left(0,\begin{bmatrix} S & 0\\ 0 & S\end{bmatrix}\right). \tag{4}$$

Proof Assumptions 1–4 imply that

$$\sqrt{(1-\lambda)T}g_{(1-\lambda)T}(\theta_0) \to_{\mathrm{d}} N(0,S) \tag{5}$$

and

$$\sqrt{\lambda T}g_{\lambda T}(\theta_0) \to_{\mathrm{d}} N(0,S) \tag{6}$$

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

[3] Alternatively, we could define partial sums of length $\lambda n_0 T'$ and $(1-\lambda)n_0T'$ for any integer T'. The results would be identical, but the notation would be more cumbersome.

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 $\mathcal{N}(a)$ denote the cumulative distribution function of the standard normal distribution evaluated at a. 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 1.2,

$$\lim_{T\to\infty} P\left(\mu_1\sqrt{(1-\lambda)T}S^{-1}\tilde{g}_{(1-\lambda)T}(\theta_0) < a, \mu_2\sqrt{\lambda T}S^{-1}\tilde{g}_{\lambda T}(\theta_0) < b\right) =$$
$$\lim_{T\to\infty} P\left(\mu_1\sqrt{(1-\lambda)T}S^{-1}g_{(1-\lambda)T}(\theta_0) < a\right)\lim_{T\to\infty}\left(\mu_2\sqrt{\lambda T}S^{-1}\tilde{g}_{\lambda T}(\theta_0) < b\right) = \mathcal{N}(a)\mathcal{N}(b)$$

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 (5) and (6). ∎

2 Consistency and asymptotic normality of estimators

In many applications, it happens that data is missing for the early part of the sample period for some moment conditions (see Stambaugh (1997) for an application to international data and Lynch, Wachter, and Boudry (2004) for an application to mutual funds). In the notation of Section 1, some elements of the vector x_t are observed for dates $1, \ldots, T$, while others are observed only for the last fraction λ of the sample, namely dates $(1-\lambda)T + 1, \ldots T$.

Without loss of generality, order the elements in x_t so that

$$x_t = \begin{pmatrix} x_{1t} \\ x_{2t} \end{pmatrix} \tag{7}$$

and $f(x_t, \theta)$ so that

$$f(x_t, \theta) = \begin{pmatrix} f_1(x_{1t}, \theta) \\ f_2(x_t, \theta) \end{pmatrix},$$

where $x_{1t} \in \mathbf{R}^{p_1}$, $x_{2t} \in \mathbf{R}^{p_2}$, $f_1 : \mathbf{R}^{p_1} \times \Theta \to \mathbf{R}^{l_1}$, and $f_2 : \mathbf{R}^{p} \times \Theta \to \mathbf{R}^{l_2}$. In an application, this corresponds to the case where data on x_{2t} is available for only the last λT dates of the sample.

Analogously, define

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

and

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

It is useful to define partitions of the matrix S. Let $R_{ij}(\tau)$ be the $l_i \times l_j$ matrix

$$R_{ij}(\tau) = E\left[f_i(x_0,\theta_0)f_j(x_{-\tau},\theta_0)^\top\right], \quad i,j = 1,2,$$

and define

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

Then

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

It is also useful to define the matrix of coefficients from a regression of the second series on the first. This is the $l_2 \times l_1$ matrix

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

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

$$\Sigma = S_{22} - S_{21}S_{11}^{-1}S_{12}. \tag{8}$$

We consider four estimators, distinguished by their moment conditions. In what follows, we will emphasize the case where the weighting matrix converges almost surely to the inverse of the variance-covariance matrix of the moments. Define

$$\begin{aligned} h_T^{\mathcal{S}}(\theta) &= \left[g_{1,\lambda T}(\theta)^\top \; g_{2,\lambda T}(\theta)^\top\right]^\top & (9) \\ h_T^{\mathcal{L}}(\theta) &= \left[g_{1,T}(\theta)^\top \; g_{2,\lambda T}(\theta)^\top\right]^\top & (10) \\ h_T^{\mathcal{A}}(\theta) &= \left[g_{1,T}(\theta)^\top \; \left(g_{2,\lambda T}(\theta) + \hat{B}_{21,\lambda T}(1-\lambda)(g_{1,(1-\lambda)T}(\theta) - g_{1,\lambda T}(\theta))\right)^\top\right]^\top & (11) \\ h_T^{\mathcal{I}}(\theta) &= \left[g_{1,(1-\lambda)T}(\theta)^\top \; g_{1,\lambda T}(\theta)^\top \; g_{2,\lambda T}(\theta)^\top\right]^\top, & (12) \end{aligned}$$

where $\hat{B}_{21,\lambda T}$ is an $l_2 \times l_1$ matrix such that $\hat{B}_{21,\lambda T} \rightarrow_{\text{a.s.}} B_{21}$. Let

$$\hat{\theta}_T^{\mathcal{S}} = \text{argmin}_\theta \, h_T^{\mathcal{S}}(\theta)^\top W_T^{\mathcal{S}} h_T^{\mathcal{S}}(\theta). \tag{13}$$

The estimator $\hat{\theta}_T^{\mathcal{S}}$ corresponds to the standard GMM. Observations on x_{1t} for $t = 1, \dots, (1-\lambda)T$ are discarded. Let

$$\hat{\theta}_T^{\mathcal{L}} = \text{argmin}_\theta \, h_T^{\mathcal{L}}(\theta)^\top W_T^{\mathcal{L}} h_T^{\mathcal{L}}(\theta). \tag{14}$$

The estimator $\hat{\theta}_T^{\mathcal{L}}$ corresponds to incorporating all of the data in the most straightforward way. Let

$$\hat{\theta}_T^{\mathcal{A}} = \text{argmin}_\theta \, h_T^{\mathcal{A}}(\theta)^\top W_T^{\mathcal{A}} h_T^{\mathcal{A}}(\theta) \tag{15}$$

and

$$\hat{\theta}_T^{\mathcal{I}} = \text{argmin}_\theta \, h_T^{\mathcal{I}}(\theta)^\top W_T^{\mathcal{I}} h_T^{\mathcal{I}}(\theta) \tag{16}$$

Estimators $\hat{\theta}_T^{\mathcal{A}}$ and $\hat{\theta}_T^{\mathcal{I}}$ are less straightforward, but, as we will argue, superior ways of including the long data. Note that, because

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

the second component of (11) can be rewritten as

$$h_{2,T}^{\mathcal{A}} = g_{2,\lambda T} + \hat{B}_{21,\lambda T}(g_{1,T} - g_{1,\lambda T}). \tag{17}$$

Equation (17) 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. We call $\hat{\theta}_T^{\mathcal{A}}$ the "adjusted moment" estimator, because it involves adjusting the second set of moments. We refer to $\hat{\theta}_T^{\mathcal{I}}$ as the "over-identified" estimator, because it involves adding an additional moment condition. As we will show, $\hat{\theta}_T^{\mathcal{I}}$ has the same asymptotic properties as $\hat{\theta}_T^{\mathcal{A}}$. In order to present theorems that apply to all the estimators, we use the notation $\hat{\theta}_T^k$ to denote a member of the class of estimators defined above, and similarly for W^k.

Theorem 2.1 *Assumptions 1–5 imply that as $T \to \infty$,*

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

where

$$
\begin{aligned}
S^{\mathcal{S}} &= S &(18)\\
S^{\mathcal{L}} &= \begin{bmatrix} \lambda S_{11} & \lambda S_{12} \\ \lambda S_{21} & S_{22} \end{bmatrix} &(19)\\
S^{\mathcal{A}} &= \begin{bmatrix} S^{\mathcal{A}}_{11} & S^{\mathcal{A}}_{12} \\ S^{\mathcal{A}}_{21} & S^{\mathcal{A}}_{22} \end{bmatrix} = \begin{bmatrix} \lambda S_{11} & \lambda S_{12} \\ \lambda S_{21} & S_{22} - (1-\lambda) S_{21} S_{11}^{-1} S_{12} \end{bmatrix} &(20)\\
S^{\mathcal{I}} &= \begin{bmatrix} \frac{\lambda}{1-\lambda} S_{11} & 0 & 0 \\ 0 & S_{11} & S_{12} \\ 0 & S_{21} & S_{22} \end{bmatrix}. &(21)
\end{aligned}
$$

Proof Equation (18) follows from Theorem 1.1. We show (20); the proofs of (19) and (21) are similar. In what follows, the argument θ_0 is suppressed and convergence is in the sense of almost surely.

Stationarity implies that $S^{\mathcal{A}}_{11} = \lambda S_{11}$. By Theorem 1.1,

$$
\begin{aligned}
&\lim_{T\to\infty} E\left[\sqrt{\lambda T}\left(\lambda g_{i,\lambda T} + (1-\lambda) g_{i,(1-\lambda)T}\right)\sqrt{\lambda T}\left(g_{j,(1-\lambda)T} - g_{j,\lambda T}\right)\right)^{\top}\Big]\\
&\quad = \lim_{T\to\infty}\left(-E\left[\sqrt{\lambda T}\lambda g_{i,\lambda T}\sqrt{\lambda T} g^{\top}_{j,\lambda T}\right] + E\left[\sqrt{\lambda T}(1-\lambda) g_{i,(1-\lambda T)}\sqrt{\lambda T} g^{\top}_{j,(1-\lambda)T}\right]\right)\\
&\quad = \lambda S_{ij} - \lambda S_{ij} = 0 \quad (22)
\end{aligned}
$$

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

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

The second line follows from (22) and the third and fourth lines follow from Theorem 1.1. Using similar reasoning,

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

which completes the derivation of (20). ▮

To establish consistency, we require the following condition on the weighting matrices.

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

Theorem 2.2 establishes consistency of the estimators.

Theorem 2.2 *Assumptions 1–6 imply that as $T \to \infty$, $\hat{\theta}^k_T \to_{\text{a.s.}} \theta_0$ for $k \in \{\mathcal{S}, \mathcal{L}, \mathcal{A}, \mathcal{I}\}$.*

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

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

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

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

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

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

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

Three remaining assumptions allow us to establish asymptotic normality of the estimators:

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 $\hat{M}(x_t) \in \mathbf{R}^{l \times q}$ such that $|\frac{\partial f_i}{\partial \theta_j}(x_t, \theta)| < \hat{M}(x_t)_{(i,j)}$ for all θ in the interior of Θ and such that for some $\delta > 0$, $E|\hat{M}(x_t)_{(i,j)}|^{r+\delta} \leq \Delta < \infty$ for all $i = 1, \dots, l$, $j = 1, \dots q$.*

Define

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

and $D_0 = [D^\top_{0,1}, \;\; D^\top_{0,2}]^\top$. Let

$$D^k_0 = D_0 \quad k \in \{\mathcal{S}, \mathcal{L}, \mathcal{A}\} \tag{23}$$
$$D^{\mathcal{I}}_0 = \left[D^\top_{0,1} \; D^\top_{0,1} \; D^\top_{0,2} \right]^\top. \tag{24}$$

The following theorem establishes asymptotic normality.

Theorem 2.3 *Assumptions 1–9 imply*

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

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, and 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 2.2 and Assumptions 7 and 8, Amemiya (1985, Theorem 4.1.5) implies

$$D_T^k(\hat{\theta}_T^k) \to_{\text{a.s.}} D_0^k \tag{25}$$

$$D_T^k(\tilde{\theta}^k) \to_{\text{a.s.}} D_0^k \tag{26}$$

$$W_T^k \to_{\text{a.s.}} W^k. \tag{27}$$

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

$$\sqrt{\lambda T}(\hat{\theta}_T^k - \theta_0) \to_{\text{d}} N\left(0, \left((D_0^k)^\top \left(S^k\right)^{-1}(D_0^k)\right)^{-1}\right). \tag{28}$$

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

3 Comparison

Interestingly, the over-identified estimator and the adjusted-moment estimator have identical asymptotic properties when the optional weighting matrix is used.

Theorem 3.1 *Assume $W_T^{\mathcal{I}} \to_{\text{a.s.}} \left(S^{\mathcal{I}}\right)^{-1}$ and $W_T^{\mathcal{A}} \to_{\text{a.s.}} \left(S^{\mathcal{A}}\right)^{-1}$. Assumptions 1–5 and 7–9 imply that the asymptotic distribution of $\sqrt{\lambda T}\hat{\theta}_T^{\mathcal{I}}$ is identical to that of $\sqrt{\lambda T}\hat{\theta}_T^{\mathcal{A}}$.*

Proof It suffices to compare the asymptotic variances as the mean of both asymptotic distributions is $\sqrt{\lambda T}\theta_0$. In the case of the over-identified estimator, the inverse of the asymptotic variance of $\sqrt{\lambda T}\hat{\theta}_T^{\mathcal{I}}$ equals

$$\begin{aligned}
\left(D^{\mathcal{I}}\right)^\top \left(S^{\mathcal{I}}\right)^{-1} D^{\mathcal{I}} &= \frac{1-\lambda}{\lambda} D_{0,1}^\top S_{11}^{-1} D_{0,1} + D_{0,1}^\top S_{11}^{-1} D_{0,1} + D_0^\top \begin{bmatrix} B_{21}^\top \Sigma^{-1} B_{21} & -B_{21}^\top \Sigma^{-1} \\ -\Sigma^{-1} B_{21} & \Sigma^{-1} \end{bmatrix} D_0 \\
&= \frac{1}{\lambda} D_{0,1}^\top S_{11}^{-1} D_{0,1} + D_0^\top \begin{bmatrix} B_{21}^\top \Sigma^{-1} B_{21} & -B_{21}^\top \Sigma^{-1} \\ -\Sigma^{-1} B_{21} & \Sigma^{-1} \end{bmatrix} D_0, \qquad (29)
\end{aligned}$$

where Σ is defined by (8). This follows from Theorem 2.4 and Lemma A.3.

It follows from the distribution of the adjusted moment estimator (28) and Lemma A.3 that the inverse of the variance of $\sqrt{\lambda T}\hat{\theta}_T^{\mathcal{A}}$ equals

$$D_0^\top \left(S^{\mathcal{A}}\right)^{-1} D_0 = \frac{1}{\lambda} D_{0,1}^\top S_{11}^{-1} D_{0,1} + D_0^\top \begin{bmatrix} B_{21}^\top \Sigma^{-1} B_{21} & -B_{21}^\top \Sigma^{-1} \\ -\Sigma^{-1} B_{21} & \Sigma^{-1} \end{bmatrix} D_0,$$

which equals (29). Thus the estimators are asymptotically equivalent. ▮

Theorem 3.1 shows that asymptotically, the distributions of the two estimators are the same. However, the interpretation of the over-identified estimator is different from the adjusted-moment estimator. Rather than adjusting the second set of moments based on the covariance with the first, the over-identified estimator turns the early data into a new moment condition.

Now we ask whether there is indeed an efficiency gain from using the longer sample. Are the adjusted-moment estimator and the over-identified estimator indeed more efficient than the short estimator?

Theorem 3.2 *Assume 1–5 and 7–9 then*

(1) If $W_{\lambda T}^k \to_{\text{a.s.}} (S^k)^{-1}$, for $k \in \mathcal{S}, \mathcal{A}, \mathcal{I}$, the estimators $\hat{\theta}_T^{\mathcal{A}}$ and $\hat{\theta}_T^{\mathcal{I}}$ are asymptotically more efficient than $\hat{\theta}_T^{\mathcal{S}}$.

(2) If $W_{\lambda T}^k \to_{\text{a.s.}} W^k$ for W^k positive definite, and such that $W^{\mathcal{A}} = W^{\mathcal{S}}$ almost surely, $\hat{\theta}_T^{\mathcal{A}}$ is more efficient than $\hat{\theta}_T^{\mathcal{S}}$.

Proof We first prove statement (1) for $\hat{\theta}_T^{\mathcal{A}}$. By Theorem 2.4 and Lemma A.2, it suffices to show that $S - S^{\mathcal{A}}$ is positive semi-definite. Note

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

For any vector $n \times 1$ vector $c = [c_1^\top,\ c_2^\top]^\top$,

$$\begin{aligned} c^\top (S - S^{\mathcal{A}}) c &= (1-\lambda) \left(c_1^\top S_{11} c_1 + c_1^\top S_{12} c_2 + c_2^\top S_{21} c_1 + c_2^\top S_{21} S_{11}^{-1} S_{12} c_2 \right) \\ &= (1-\lambda) \left(c_1^\top S_{11}^\top S_{11}^{-1} S_{11} c_1 + c_1^\top S_{11} S_{11}^{-1} S_{12} c_2 + c_2^\top S_{12}^\top S_{11}^{-1} S_{11} c_1 + c_2 S_{12}^\top S_{11}^{-1} S_{12} c_2 \right) \\ &= (1-\lambda)(S_{11} c_1 + S_{12} c_2)^\top S_{11}^{-1} (S_{11} c_1 + S_{12} c_2) \geq 0 \end{aligned}$$

because S_{11}^{-1} is positive-semi-definite and $\lambda < 1$. Therefore $S - S^{\mathcal{A}}$ is positive semi-definite and, as a consequence, $\hat{\theta}_T^{\mathcal{A}}$ is more efficient than $\hat{\theta}_T^{\mathcal{S}}$. The statement for $\hat{\theta}_T^{\mathcal{I}}$ then follows from Theorem 3.1.

To prove statement (2), define

$$M = W^{\mathcal{A}} D_0^{\mathcal{A}} \left((D_0^{\mathcal{A}})^\top W^{\mathcal{A}} D_0^{\mathcal{A}} \right)^{-1}. \tag{30}$$

Because the weighting matrix is assumed to be the same for both estimators,

$$M = W^{\mathcal{S}} D_0^{\mathcal{S}} \left((D_0^{\mathcal{S}})^\top W^{\mathcal{S}} D_0^{\mathcal{S}} \right)^{-1}.$$

By Theorem 2.3, proving (2) is equivalent to showing $M^\top S M - M^\top S^{\mathcal{A}} M$ is positive semi-definite. But for any vector c,

$$c^\top (M^\top S M - M^\top S^{\mathcal{A}} M) c = (Mc)^\top (S - S^{\mathcal{A}}) Mc > 0$$

because $S - S^{\mathcal{A}}$ is positive semi-definite. Therefore $\hat{\theta}_T^{\mathcal{A}}$ is more efficient then $\hat{\theta}_T^{\mathcal{S}}$ when $W^{\mathcal{S}} = W^{\mathcal{A}}$. ∎

Note that statement (2) of Theorem 3.2 does not make sense for the over-identified estimator $\hat{\theta}_T^{\mathcal{I}}$ because $\hat{\theta}_T^{\mathcal{I}}$ has l_1 more moment conditions than $\hat{\theta}_T^{\mathcal{A}}$ and $\hat{\theta}_T^{\mathcal{S}}$. It is not possible to keep the weighting matrices the same.

Theorem 3.2 shows that introducing the extra data from the longer series reduces the variance of the estimates relative to using the shorter series alone. It is also interesting to ask whether the estimator is more efficient than the one that would result from using the longer sample in a more "naive" way, namely using the longer data series to estimate the first set of moment conditions, and the shorter series to estimate the second. In the terminology of Section 1 this involves comparing $\hat{\theta}_T^{\mathcal{A}}$ with the estimator $\hat{\theta}_T^{\mathcal{L}}$.

Theorem 3.3 *Assume 1–5 and 7–9 then*

(1) If $W^k_{\lambda T} \rightarrow_{\text{a.s.}} (S^k)^{-1}$, for $k \in \mathcal{L}, \mathcal{A}, \mathcal{I}$, the estimators $\hat{\theta}^{\mathcal{A}}_T$ and $\hat{\theta}^{\mathcal{I}}_T$ are asymptotically more efficient than $\hat{\theta}^{\mathcal{L}}_T$.

(2) If $W^k_{\lambda T} \rightarrow_{\text{a.s.}} W^k$ for W^k positive definite, and such that $W^{\mathcal{A}} = W^{\mathcal{L}}$ almost surely, $\hat{\theta}^{\mathcal{A}}_T$ is more efficient than $\hat{\theta}^{\mathcal{L}}_T$.

Proof As in Theorem 3.2, it suffices to show that

$$S^{\mathcal{L}} - S^{\mathcal{A}} = \left[\begin{array}{cc} 0 & 0 \\ 0 & (1-\lambda) S_{21} S_{11}^{-1} S_{12} \end{array} \right]$$

is positive semi-definite. Note that for any vector $c = [c_1^\top,\ c_2^\top]^\top$,

$$c^\top (S^{\mathcal{L}} - S^{\mathcal{A}}) c = (1-\lambda)(S_{12} c_2)^\top S_{11}^{-1} S_{12} c_2 \geq 0$$

because $\lambda < 1$ and S_{11} is positive semi-definite. Lemma A.2 then implies that $\hat{\theta}^{\mathcal{A}}_T$ is more efficient than $\hat{\theta}^{\mathcal{L}}_T$. By Theorem 3.1, $\hat{\theta}^{\mathcal{I}}_T$ is also more efficient then $\hat{\theta}^{\mathcal{L}}_T$. This proves (1).

To show the second statement, define M analogously to (30):

$$M = W^{\mathcal{L}} D_0^{\mathcal{L}} \left((D_0^{\mathcal{L}})^\top W^{\mathcal{L}} D_0^{\mathcal{L}} \right)^{-1},$$

and note that $W^{\mathcal{A}} = W^{\mathcal{L}}$. Because $S^{\mathcal{L}} - S^{\mathcal{A}}$ is positive semi-definite, for any vector c,

$$c^\top (M^\top S^{\mathcal{L}} M - M^\top S^{\mathcal{A}} M) c = (Mc)^\top (S^{\mathcal{L}} - S^{\mathcal{A}}) Mc > 0.$$

By Theorem 2.3, $\hat{\theta}^{\mathcal{A}}_T$ is more efficient then $\hat{\theta}^{\mathcal{L}}_T$ when $W^{\mathcal{A}} = W^{\mathcal{L}}$. This proves (2). ▮

Surprisingly, $\hat{\theta}^{\mathcal{L}}_T$ is not necessarily more efficient than $\hat{\theta}^{\mathcal{S}}_T$. Efficiency would require that

$$S^{\mathcal{L}} - S = (1-\lambda) \left[\begin{array}{cc} S_{11} & S_{12} \\ S_{21} & 0 \end{array} \right]$$

be positive semi-definite. However, if the covariances between the first and second set of moment conditions are nonzero, this may not be the case. Thus it is not sufficient to simply use the first part of the sample, it must be combined with the second part of the sample in precisely the right way to produce a gain in efficiency.

We further explore the relation between these estimators by looking at the first order conditions. For the purpose of this discussion, we assume $W^{\mathcal{I}}_T = \left(S^{\mathcal{I}}\right)^{-1}$, $W^{\mathcal{A}}_T = \left(S^{\mathcal{A}}\right)^{-1}$, and $\hat{B}_{21,\lambda T} = B_{21}$.

Differentiating (16) with respect to θ yields

$$\begin{aligned} 0 &= \frac{1-\lambda}{\lambda} g_{1,(1-\lambda)T}^{\top} S_{11}^{-1} \frac{\partial g_{1,(1-\lambda)T}}{\partial \theta} + g_{1,\lambda T}^{\top} S_{11}^{-1} \frac{\partial g_{1,\lambda T}}{\partial \theta} \\ &\qquad + \left[g_{1,\lambda T}^{\top} \quad g_{2,\lambda T}^{\top} \right] \left[\begin{array}{cc} B_{21}^{\top} \Sigma^{-1} B_{21} & -B_{21}^{\top} \Sigma^{-1} \\ -\Sigma^{-1} B_{21} & \Sigma^{-1} \end{array} \right] \left(\begin{array}{c} \frac{\partial g_{1,\lambda T}}{\partial \theta} \\ \frac{\partial g_{2,\lambda T}}{\partial \theta} \end{array} \right) \\ &= \frac{1-\lambda}{\lambda} g_{1,(1-\lambda)T}^{\top} S_{11}^{-1} \frac{\partial g_{1,(1-\lambda)T}}{\partial \theta} + g_{1,\lambda T}^{\top} S_{11}^{-1} \frac{\partial g_{1,\lambda T}}{\partial \theta} \\ &\qquad + \left(g_{2,\lambda T} - B_{21} g_{1,\lambda T} \right)^{\top} \Sigma^{-1} \frac{\partial}{\partial \theta} \left(g_{2,\lambda T} - B_{21} g_{1,\lambda T} \right). \end{aligned} \tag{31}$$

Equation (31) is the first-order condition that determines the over-identified estimator $\hat{\theta}_T^{\mathcal{I}}$. By contrast, the first order condition associated with (15) is

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

which reduces to

$$\begin{aligned} 0 &= \frac{1}{\lambda} g_{1,T}^{\top} S_{11}^{-1} \frac{\partial g_{1,T}}{\partial \theta} + \left(B_{21} g_{1,T} - h_{2,T} \right)^{\top} \Sigma^{-1} \frac{\partial}{\partial \theta} \left(B_{21} g_{1,T} - h_{2,T} \right) \\ &= \frac{1}{\lambda} g_{1,T}^{\top} S_{11}^{-1} \frac{\partial g_{1,T}}{\partial \theta} + \left(B_{21} g_{1,\lambda T} - g_{2,\lambda T} \right)^{\top} \Sigma^{-1} \frac{\partial}{\partial \theta} \left(B_{21} g_{1,\lambda T} - g_{2,\lambda T} \right). \end{aligned} \tag{32}$$

Equation (32) is the first-order condition that determines the adjusted-moment estimator $\hat{\theta}_T^{\mathcal{A}}$. According to Theorem 3.1, 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, for example, when the moment conditions are linear. This corresponds to the case examined by Stambaugh (1997) in a maximum likelihood context.

4 A special case

This section examines a special case of the set-up of Section 2. We assume 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. Let l_1 be the length of θ_1, and $l_2 = q - l_1$ the length of θ_2. In this setting, we can draw additional conclusions about the first-order conditions of the adjusted-moment and over-identified estimators, and we can quantify the gains from including the longer sample.

For convenience assume that $W_T^{\mathcal{I}} = \left(S^{\mathcal{I}}\right)^{-1}$ and $\hat{B}_{21,T} = B_{21}$. Because

$$\frac{\partial g_{1,(1-\lambda)T}}{\partial \theta_2} \equiv \frac{\partial g_{1,\lambda T}}{\partial \theta_2} \equiv \frac{\partial g_{1T}}{\partial \theta_2} \equiv 0,$$

and because $\Sigma^{-1}\frac{\partial}{\partial\theta}\left(B_{21}g_{1,\lambda T} - g_{2,\lambda T}\right)$ is invertible, the first order conditions for the over-identified estimator $\hat{\theta}_T^{\mathcal{I}}$ reduce to

$$g_{2,\lambda T} - B_{21}g_{1,\lambda T} = 0 \tag{33}$$

for θ_2, and thus

$$\frac{1-\lambda}{\lambda} g_{1,(1-\lambda)T}^\top S_{11}^{-1}\frac{\partial g_{1,(1-\lambda)T}}{\partial \theta_1} + g_{1,\lambda T}^\top S_{11}^{-1}\frac{\partial g_{1,\lambda T}}{\partial \theta_1} = 0 \tag{34}$$

for θ_1. The first order conditions for the adjusted-moment estimator $\hat{\theta}_T^{\mathcal{A}}$ reduce to (33) for θ_2 but

$$g_{1,T} = 0$$

for θ_1. This is no surprise. When the adjusted-moment estimator is exactly identified, the first-order conditions must be equivalent to setting $g_{1,T}$ equal to zero, and $h_{2,T}$ equal to zero. When $g_{1,T} = 0$, $h_{2,T}$ is equivalent to the left-hand side of (33).

We have shown that in the case considered here, the adjusted-moment estimator gives the same estimate for θ_1 as simply using the long 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, (34) nonetheless has an interpretation; it is a weighted average of the moment conditions from the first part and the second part of the sample, where the weights are proportional to the derivatives, and thus to the amount of information contained in each part of the sample.

We now quantify the effects of using the adjusted-moment estimator or the over-identified estimator on the standard errors for θ. In the special case where the system is exactly identified and f_1 depends on θ_1, the derivative matrix D_0 is invertible and takes the form

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

for an $l_1 \times l_1$ invertible matrix d_{11}, an $l_2 \times l_1$ matrix d_{21} and an $l_2 \times l_2$ invertible matrix d_{22}. The matrix d_{11} gives the derivatives of f_1 with respect to θ_1, d_{21} gives the derivatives of f_2 with respect to θ_1, and d_{22} gives the derivatives of f_2 with respect to θ_2.

The inverse of D_0 takes the form

$$D_0^{-1} = \left[\begin{array}{cc} d_{11}^{-1} & 0 \\ -d_{22}^{-1} d_{21} d_{11}^{-1} & d_{22}^{-1} \end{array} \right].$$

Therefore the first diagonal block of $\left(D_0^\top S^{-1} D_0\right)^{-1}$ equals $d_{11}^{-1} S_{11} (d_{11}^{-1})^\top$. Similarly, the first block of $\left(D_0^\top \left(S^{\mathcal{A}}\right)^{-1} D_0\right)^{-1}$ can be written as[4]

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

This shows that asymptotic standard errors for the estimates of θ_1 shrink by a factor of $1 - \sqrt{\lambda}$ when the adjusted-moment estimator is used rather than the short estimator. Because the over-identified estimator is asymptotically equivalent to the adjusted-moment estimator, the shrinkage is the same.

It is more interesting to look at the effect on the standard errors of the second set of parameters θ_2. The second diagonal block of $\left(D_0^\top \left(S^{\mathcal{A}}\right)^{-1} D_0\right)^{-1}$ equals

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

which reduces to:

$$\begin{aligned} \left(D_0^\top \left(S^{\mathcal{A}}\right)^{-1} D_0\right)^{-1}_{22} = {} & d_{22}^{-1} \left[d_{21} d_{11}^{-1} S_{11}^{\mathcal{A}} - S_{21}^{\mathcal{A}}\right] {S_{11}^{\mathcal{A}}}^{-1} \left[d_{21} d_{11}^{-1} S_{11}^{\mathcal{A}} - S_{21}^{\mathcal{A}}\right] (d_{22}^{-1})^\top \\ & + d_{22}^{-1} \left[S_{22}^{\mathcal{A}} - S_{21}^{\mathcal{A}} \left(S_{11}^{\mathcal{A}}\right)^{-1} S_{12}^{\mathcal{A}}\right] (d_{22}^{-1})^\top. \end{aligned} \tag{35}$$

The variance for the second set of variables 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: $S_{22} - S_{21} S_{11}^{-1} S_{12}$ is the variance-covariance matrix of the second set of moment conditions conditional on the first.

Because the new data reduces the asymptotic variance of the first set of moment conditions by a factor of λ, the data will also reduce the asymptotic variance of the second set of variables:

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

[4] Recall that $D_0^{\mathcal{A}} = D_0$.

However the second term in (35) does not change with the addition of new data, not surprisingly because it represents 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,

$$S_{22} - S_{21} S_{11}^{-1} S_{12} = 0, \tag{36}$$

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_2, 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 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 (35) is nonzero and there is an effect on the standard errors. 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.

5 Extensions

The previous sections considered cases where there were two relevant sample periods: a "short" sample period over which all data are observed, and a "long" sample period over which only some of the data are observed. This section extends the methods to cases where there are more than two different sample periods. In order to extend the estimators of Section 2, it is necessary to prove a theorem analogous to Theorem 1.1 for the case where the data of length T is divided into more than two blocks. Let $\eta_1, \eta_2, \ldots, \eta_n$ denote rational numbers such that $\sum_{k=1}^{n} \eta_k = 1$. Let n_0 be the smallest integer such that the product with η_j is an integer, for all j. As above, we will restrict attention to values T that are a multiple of n_0. Define the following partial sums of g:

$$g_{\eta_1 T}(\theta) = \frac{1}{\eta_1 T} \sum_{t=1}^{\eta_1 T} f(x_t, \theta) \tag{37}$$

$$g_{\eta_j T}(\theta) = \frac{1}{\eta_j T} \sum_{t=(\eta_1+\cdots+\eta_{j-1})T+1}^{(\eta_1+\cdots+\eta_j)T} f(x_t, \theta), \qquad j = 2, \ldots, n. \tag{38}$$

$-(\eta_1 + \cdots + \eta_{n-1})T + 1 \qquad - (\eta_2 + \cdots + \eta_{n-1})T + 1 \qquad 1 \qquad \eta_n T$

|———————|————————————————|———————|

Figure 2: Numbering scheme used in the proof of Theorem 5.1

.

Theorem 5.1 *Define $g_{\eta_1 T}$ as in (37) and $g_{\eta_j T}$ as in (38) for $j = 2, \ldots, n$. Assumptions 1–5 imply*

$$\sqrt{T} \begin{pmatrix} \sqrt{\eta_1} g_{\eta_1 T}(\theta_0) \\ \sqrt{\eta_2} g_{\eta_2 T}(\theta_0) \\ \vdots \\ \sqrt{\eta_n} g_{\eta_n T}(\theta_0) \end{pmatrix} \to_{\mathrm{d}} N \left(0, \begin{bmatrix} S & 0 & \ldots & 0 \\ 0 & S & \ldots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & S \end{bmatrix} \right) \tag{39}$$

as $T \to \infty$.

Proof We proceed by induction. The case for $n = 1$ follows from standard results (e.g. White and Domowitz (1984, Theorem 2.4)). Suppose (39) holds for $n - 1$. Because x_t is stationary, we can define a new set of partial sums $\tilde{g}$ with the same joint distribution as the partial sums g. Let

$$\tilde{g}_{\eta_n T}(\theta_0) = \frac{1}{\eta_n T} \sum_{t=1}^{\eta_n T} f(x_t, \theta),$$

while

$$\tilde{g}_{\eta_j T}(\theta_0) = \sum_{-(\eta_j + \ldots \eta_{n-1})T+1}^{-(\eta_{j+1} + \ldots \eta_{n-1})T} f(x_t, \theta), \qquad j = 1, \ldots, n-1.$$

As Figure 2 illustrates, the start data of new sample is $-(\eta_1 + \cdots + \eta_{n-1})T + 1$ while the end date is $\eta_n T$. Then $\tilde{g}_{\eta_1 T}(\theta_0), \ldots, \tilde{g}_{\eta_n T}(\theta_0)$ have the same joint distribution as $g_{\eta_1 T}(\theta_0), \ldots, g_{\eta_n T}(\theta_0)$. By Lemma 1.2, for any $1 \times l$ vectors $\mu_1, \ldots, \mu_n$ such that $\mu_j \mu_j^\top = 1$, and scalars $a_1, \ldots, a_n$,

$$\lim_{T \to \infty} P\left(\sqrt{\eta_n T} \mu_n S^{-1} \tilde{g}_{\eta_n T}(\theta_0) < a_n, \sqrt{\eta_{n-1} T} \mu_{n-1} S^{-1} \tilde{g}_{\eta_{n-1} T}(\theta_0) < a_{n-1}, \ldots, \sqrt{\eta_1 T} \mu_1 S^{-1} \tilde{g}_{\eta_1 T}(\theta_0) < a_1 \right) =$$
$$\lim_{T \to \infty} P\left(\sqrt{\eta_n T} \mu_n S^{-1} \tilde{g}_{\eta T}(\theta_0) < a_n \right) \times$$
$$\lim_{T \to \infty} P\left(\sqrt{\eta_{n-1} T} \mu_{n-1} S^{-1} \tilde{g}_{\eta_{n-1} T}(\theta_0) < a_{n-1}, \ldots, \sqrt{\eta_1 T} \mu_1 S^{-1} \tilde{g}_{\eta_1 T}(\theta_0) < a_1 \right).$$

The result then follows from the induction assumption and asymptotic normality of $\sqrt{\eta_n T} g_{\eta_n T}$. ▮

5.1 Extending the Over-Identified Estimator

An advantage of the over-identified estimator is that it is has a transparent extension to samples where there is a more general pattern of missing data. Theorem 5.1 gives the joint distribution of partial sums of g. We now use this result to extend the over-identified estimator.

As before, we consider the situation where not all moments are observed over the whole sample. Here, however, we allow for an arbitrary number of missing "blocks" of data, and they can occur anywhere in the sample, rather than simply at the beginning. Our asymptotic results will keep the size of these missing blocks of data proportional to the size of the overall sample, so that the missing data problem does not become trivial, just as in the case where there was data missing at the start of the sample.

Consider segments 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 regions $1, \ldots, n$. We can create a weak ordering on these segments by how many of the sample moments are observed over each segments. That is, π_1 sample moments are observed over the first segment, $\pi_2 \leq \pi_1$ over the second segment, etc. We let λ_1 denote the ratio of the length of the first region (the region over which the greatest number of moments are observed) 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. Then $\lambda_1, \ldots, \lambda_n$ can be thought of in the same way as $\eta_1, \ldots, \eta_n$ in the previous section, except that while the ηs are labeled according to their order in the sample, the λs are labeled according to how many data moments are observed over that segment. 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 2, the first segment consist of points $(1-\lambda)T + 1$ to T. All moments were observed over this segment. The second segment consists of points 1 to $(1-\lambda)T$. Only a subset of moments are observed over these points. In this example, $t_1 = (1-\lambda)T$, $t_2 = 0$, $\lambda_1 = \lambda$, and $\lambda_2 = (1-\lambda)$. We adopt the same notational convention as in Section 2: $\lambda_j T$ will refer to the length of the segment between $t_j + 1$ and $t_j + \lambda_j T$, and the segment itself.

Finally, let ϕ_i denote the set of data series that are observed in data segment λ_i. 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 $\lambda_j T$. Similarly, define the $\pi_j \times q$ matrix

$$D_{0,\phi_j} = E\left[\left. \frac{\partial f_{\phi_j}}{\partial \theta} \right|_{\theta_0} \right] = \Big(D_{0,i_1}^\top, \ldots, \; D_{0,i_{\pi_j}}^\top \Big)^\top,$$

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

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

and

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

The extended over-identified estimator, for the case where there are n blocks of data and the total data length is T, has moment conditions

$$h_T^{\mathcal{I}_n}(\theta) = \left[g_{\phi_1,\lambda_1 T}(\theta)^\top,\ g_{\phi_2,\lambda_2 T}(\theta)^\top, \ldots,\ g_{\phi_n,\lambda_n T}(\theta)^\top\right]^\top, \tag{40}$$

for $\theta \in \Theta$. 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 segments. The T subscript refers to the fact that the data length is T.[5] As in Section 2, $\sqrt{T} h_T^{\mathcal{I}_n}(\theta)$ is asymptotically normally distributed. The following is analogous to Theorem 2.1.

Theorem 5.2 *Assumptions 1–5 imply*

$$\sqrt{T} h_T^{\mathcal{I}_n}(\theta_0) \to_{\mathrm{d}} N\left(0, S^{\mathcal{I}_n}\right),$$

where

$$S^{\mathcal{I}_n} = \left[\begin{array}{cccc} \frac{1}{\lambda_1} S_{\phi_1} & 0 & \ldots & 0 \\ 0 & \frac{1}{\lambda_2} S_{\phi_2} & \ldots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \frac{1}{\lambda_n} S_{\phi_n} \end{array}\right] \tag{41}$$

as $T \to \infty$.

The extended over-identified estimator takes (40) as moment conditions. In principle, any weighting matrix can be used, but we will emphasize the case when the weighting matrix converges almost surely to (41). Define

$$\hat{\theta}_T^{\mathcal{I}_n} = \mathrm{argmin}_\theta\, h_T^{\mathcal{I}_n}(\theta)^\top W_T^{\mathcal{I}_n} h_T^{\mathcal{I}_n}(\theta). \tag{42}$$

[5] This notation does not, of course, completely define the over-identified estimator. For that, one would need the points at which the data segments begin, $t_1, \ldots, t_n$. These points in turn depend in a complicated way on $\lambda_1, \ldots, \lambda_n$ and T.

Not surprisingly, the same consistency and asymptotic efficiency results go through for the extended over-identified estimator as for the original over-identified estimator. Here, we repeat the results but omit the proofs, which follow along the same lines as the corresponding proofs in Section 2.

Assumption 10 *The weighting matrix $W_T^{\mathcal{I}_n}$ converges almost surely to a positive-definite matrix $W^{\mathcal{I}_n}$.*

Theorem 5.3 *Assumptions 1-5 and 10 imply that as $T \to \infty$, $\hat{\theta}_T^{\mathcal{I}_n} \to_{\text{a.s.}} \theta_0$.*

Define

$$D_0^{\mathcal{I}_n} = \left[D_{0,\phi_1}^\top \; D_{0,\phi_2}^\top \; \ldots \; D_{0,\phi_n}^\top \right]^\top.$$

Note that Assumptions 1–5 imply that

$$D_0^{\mathcal{I}_n} = \lim_{T\to\infty} \left. \frac{\partial h_T^{\mathcal{I}_n}}{\partial \theta} \right|_{\theta_0}.$$

Theorem 5.4 *Assumptions 1–5, and 7–10 imply that as $T \to \infty$,*

$$\sqrt{T}(\hat{\theta}_T^{\mathcal{I}_n} - \theta_0) \to_{\text{d}} N\left(0, \left((D_0^{\mathcal{I}_n})^\top W^{\mathcal{I}_n} D_0^{\mathcal{I}_n}\right)^{-1} \left((D_0^{\mathcal{I}_n})^\top W^{\mathcal{I}_n} S^{\mathcal{I}_n} W^{\mathcal{I}_n} D_0^{\mathcal{I}_n}\right) \left((D_0^{\mathcal{I}_n})^\top W^{\mathcal{I}_n} D_0^{\mathcal{I}_n}\right)^{-1}\right).$$

Theorem 5.5 *Suppose $W_T^{\mathcal{I}_n} \to_{\text{a.s.}} W^{\mathcal{I}_n} = (S^{\mathcal{I}_n})^{-1}$. Assumptions 1–5 and 7–10 imply*

$$\sqrt{T}(\hat{\theta}_T^{\mathcal{I}_n} - \theta_0) \to_{\text{d}} N\left(0, \left((D_0^{\mathcal{I}_n})^\top \left(S^{\mathcal{I}_n}\right)^{-1} (D_0^{\mathcal{I}_n})\right)^{-1}\right). \tag{43}$$

Moreover, this choice of $W^{\mathcal{I}_n}$ is efficient given the moment conditions (40).

The extended over-identified estimator reduces to the over-identified estimator considered in Section 2 when there is a single block of data. Section 5.3 gives examples where there are multiple blocks of data.

We now prove a result analogous to Theorem 3.2. 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.

Theorem 5.6 *Assume $W_T^{\mathcal{I}_n} \to \left(S^{\mathcal{I}_n}\right)^{-1}$ and $W_{(1-\lambda_n)T}^{\mathcal{I}_{n-1}} \to \left(S^{\mathcal{I}_{n-1}}\right)^{-1}$. Assumptions 1–5 and 7–9 imply $\hat{\theta}_T^{\mathcal{I}_n}$ is asymptotically more efficient than $\hat{\theta}_{(1-\lambda_n)T}^{\mathcal{I}_{n-1}}$.*

Proof It suffices to compare the asymptotic variance of $\sqrt{T}\hat{\theta}_T^{\mathcal{I}_n}$ with $\sqrt{T}\hat{\theta}_{(1-\lambda_n)T}^{\mathcal{I}_{n-1}}$. By Theorem 5.5,

$$E\left[(1-\lambda_n)T\left(\hat{\theta}_{(1-\lambda_n)T}^{\mathcal{I}_{n-1}}-\theta_0\right)\left(\hat{\theta}_{(1-\lambda_n)T}^{\mathcal{I}_{n-1}}-\theta_0\right)^\top\right]=\left((D_0^{\mathcal{I}_{n-1}})^\top\left(S^{\mathcal{I}_{n-1}}\right)^{-1}(D_0^{\mathcal{I}_{n-1}})\right)^{-1},$$

where

$$S^{\mathcal{I}_{n-1}}=\begin{bmatrix}\frac{1-\lambda_n}{\lambda_1}S_{\phi_1} & 0 & \dots & 0\\ 0 & \frac{1-\lambda_n}{\lambda_2}S_{\phi_2} & \dots & 0\\ 0 & 0 & \ddots & 0\\ 0 & 0 & 0 & \frac{1-\lambda_n}{\lambda_{n-1}}S_{\phi_{n-1}}\end{bmatrix},$$

because data segment λ_j occupies a fraction $\lambda_k/(1-\lambda_n)$ of the data segment $(1-\lambda_n)T$ (note that $\lambda_j < 1-\lambda_n$ because $\sum_{j=1}^{n-1}\lambda_j = 1-\lambda_n$). Therefore,

$$\begin{aligned}E\left[T\left(\hat{\theta}_{(1-\lambda_n)T}^{\mathcal{I}_{n-1}}-\theta_0\right)\left(\hat{\theta}_{(1-\lambda_n)T}^{\mathcal{I}_{n-1}}-\theta_0\right)^\top\right] &= \frac{1}{1-\lambda_n}\left((D_0^{\mathcal{I}_{n-1}})^\top\left(S^{\mathcal{I}_{n-1}}\right)^{-1}(D_0^{\mathcal{I}_{n-1}})\right)^{-1} \qquad (44)\\ &= \left((D_0^{\mathcal{I}_{n-1}})^\top\begin{bmatrix}\frac{1}{\lambda_1}S_{\phi_1} & 0 & \dots & 0\\ 0 & \frac{1}{\lambda_2}S_{\phi_2} & \dots & 0\\ 0 & 0 & \ddots & 0\\ 0 & 0 & 0 & \frac{1}{\lambda_{n-1}}S_{\phi_{n-1}}\end{bmatrix}^{-1}(D_0^{\mathcal{I}_{n-1}})\right)^{-1}.\end{aligned}$$

By (44) and Lemma A.1, it suffices to show that

$$(D_0^{\mathcal{I}_n})^\top\left(S^{\mathcal{I}_n}\right)^{-1}(D_0^{\mathcal{I}_n})-\frac{1}{1-\lambda_n}(D_0^{\mathcal{I}_{n-1}})^\top\left(S^{\mathcal{I}_{n-1}}\right)^{-1}(D_0^{\mathcal{I}_{n-1}})$$

is positive semi-definite. Applying (44), we have

$$S^{\mathcal{I}_n}=\begin{bmatrix}\frac{1}{1-\lambda_n}S^{\mathcal{I}_{n-1}} & 0\\ 0 & \frac{1}{\lambda_n}S_{\phi_n}\end{bmatrix}$$

and

$$D^{\mathcal{I}_n}=[(D^{\mathcal{I}_{n-1}})^\top\ D_{0,\phi_n}^\top]^\top.$$

Therefore

$$(D_0^{\mathcal{I}_n})^\top\left(S^{\mathcal{I}_n}\right)^{-1}(D_0^{\mathcal{I}_n})-(1-\lambda_n)(D_0^{\mathcal{I}_{n-1}})^\top\left(S^{\mathcal{I}_{n-1}}\right)^{-1}(D_0^{\mathcal{I}_{n-1}})=\lambda_n D_{0,\phi_n}^\top S_{\phi_n}^{-1}D_{0,\phi_n}, \qquad (45)$$

which is positive semi-definite. ∎

5.2 Extending the Adjusted-Moment Estimator

This section shows that the adjusted-moment estimator can also be extended to the case where there is are series of n lengths, where n is greater than 2. In fact, it is possible to define an adjusted-moment estimator that is asymptotically equivalent to the over-identified estimator, just as in the case where there were two blocks of data. Rather than a formulation (40), the extended adjusted-moment estimator is defined by induction.

An advantage of the adjusted moment estimator over the over-identified estimator was described in Section 3. When the model is exactly identified, and there is a set of series that have data throughout the sample period that depend on a subset of the parameters, the adjusted-moment estimator gives the same estimate for those parameters as simply using the long sample.

Consider the same set-up as in Section 5.1. To simplify notation, we consider a slightly less general problem than in Section 5.1. We require that all series have a segment in common, in other words, $\pi_1 = l$ and $\phi_1 = \{1, \ldots, l\}$. Of course, this segment could be a small portion of the total data available. To inductively define the adjusted-moment estimator, we first give a definition of the adjusted-moment estimator when all series are observed for all the data. This is standard GMM. Then we assume that the adjusted-moment estimator has been defined over the first $n-1$ segments of data, and extend the adjusted moment estimator to all n segments of data. This procedure can be used to construct the adjusted-moment estimator over the same patterns of missing data as for the over-identified estimator (assuming that all data has been observed over at least one segment), provided that one starts the construction with the segment over which all the data has been observed (of length $\lambda_1 T$), and then proceeds to the segment where $\pi_2 < \pi_1 = l$ moment conditions have been observed, and so forth. While the induction approach may appear somewhat cumbersome, the procedure is quite straightforward, as demonstrated by the examples in Section 5.3.

We begin by defining the adjusted-moment estimator when there is a single data segment and no missing data. This is the standard GMM estimator:

$$h^{\mathcal{A}_1}_{\lambda_1 T} = g_{\lambda_1 T}. \tag{46}$$

As in the previous section, the subscript on h denotes the data region over which h is measured. The superscript refers to the fact that it is the adjusted-moment estimator, while the subscript on $\mathcal{A}$ refers to the number of blocks of data. It follows from standard arguments that

$$E\left[h^{\mathcal{A}_1}_{\lambda_1 T}(\theta_0)\right] = E\left[f(x_t, \theta_0)\right] = 0,$$

and that

$$h^{\mathcal{A}_1}_{\lambda_1 T}(\theta) \rightarrow_{\text{a.s.}} E[f(x_t, \theta)]$$

as $T \to \infty$. Assume by induction that

$$h^{\mathcal{A}_{n-1}}_{(1-\lambda_n)T}(\theta) \to_{\text{a.s.}} E[f(x_t, \theta)], \tag{47}$$

and that

$$E\left[(1-\lambda_n)T\left(h^{\mathcal{A}_{n-1}}_{(1-\lambda_n)T}(\theta_0)\right)\left(h^{\mathcal{A}_{n-1}}_{(1-\lambda_n)T}(\theta_0)\right)^\top\right] \to_{\text{a.s.}} S^{\mathcal{A}_{n-1}}, \tag{48}$$

for some symmetric, positive-definite matrix $S^{\mathcal{A}_{n-1}}$. Finally, assume that $h^{\mathcal{A}_{n-1}}_{(1-\lambda_n)T}(\theta_0)$ is a linear combination of $g_{\phi_1}(\theta_0), \ldots, g_{\phi_{n-1}}(\theta_0)$ with non-random coefficients that do not depend on T. That is

$$h^{\mathcal{A}_{n-1}}_{(1-\lambda_n)T}(\theta_0) = M_{n-1}[g_{\phi_1}(\theta_0)^\top, \ldots, g_{\phi_{n-1}}(\theta_0)^\top]^\top. \tag{49}$$

This allows Theorem 5.1 to be applied. Note that (47) implies that there is a one-to-one correspondence between moment conditions in the adjusted-moment estimator, and moment conditions f_i.

Let $h^{\mathcal{A}_{n-1}}_{\phi_n,(1-\lambda_n)T}(\theta)$ denote the π_n components of $h^{\mathcal{A}_{n-1}}_{(1-\lambda_n)T}(\theta)$ that converge to $E\left[f_{\phi_n}(x_t, \theta)\right]$. These are the elements of $h^{\mathcal{A}_{n-1}}_{\phi_n,(1-\lambda_n)T}(\theta)$ corresponding to moments observed over the new data length. We define the adjusted-moment estimator for n segments as the residual from a regression of the previous adjusted-moment estimator on the difference between the components of the previous adjusted-moment estimator for which the new data is available, and the sample average over the new segment of data. Define

$$\begin{aligned} B^{\mathcal{A}_{n-1}} = \lim_{T\to\infty} E&\left[T h^{\mathcal{A}_{n-1}}_{(1-\lambda_n)T}(\theta_0)\left(h^{\mathcal{A}_{n-1}}_{\phi_n,(1-\lambda_n)T}(\theta_0) - g_{\phi_n,(1-\lambda_n)T}(\theta_0)\right)^\top\right] \times \\ &E\left[T\left(h^{\mathcal{A}_{n-1}}_{\phi_n,(1-\lambda_n)T}(\theta_0) - g_{\phi_n,(1-\lambda_n)T}(\theta_0)\right)\left(h^{\mathcal{A}_{n-1}}_{\phi_n,(1-\lambda_n)T}(\theta_0) - g_{\phi_n,(1-\lambda_n)T}(\theta_0)\right)^\top\right]^{-1}. \end{aligned} \tag{50}$$

$B^{\mathcal{A}_{n-1}}$ is the $l \times \pi_n$ matrix of asymptotic regression coefficients from a regression of the $(n-1)$st adjusted-moment estimator on $h^{\mathcal{A}_{n-1}}_{\phi_n,(1-\lambda_n)T}(\theta_0) - g_{\phi_n,\lambda_n T}(\theta_0)$, appropriately scaled by the square root of the sample length. In practice, $B^{\mathcal{A}_j}$ can be replaced by a sample estimate $\hat{B}^{\mathcal{A}_j}_T$ such that $\hat{B}^{\mathcal{A}_j}_T \to_{\text{a.s.}} B^{\mathcal{A}_j}$ as $T \to \infty$.[6] Finally define the nth adjusted-moment estimator as

$$h^{\mathcal{A}_n}_T(\theta) = h^{\mathcal{A}_{n-1}}_{(1-\lambda_n)T}(\theta) - B^{\mathcal{A}_{n-1}}\left(h^{\mathcal{A}_{n-1}}_{\phi_n,(1-\lambda_n)T}(\theta) - g_{\phi_n,\lambda_n T}(\theta)\right). \tag{51}$$

[6] In that case, (49) would be replaced by the requirement that

$$h^{\mathcal{A}_{n-1}}_{(1-\lambda_n)T}(\theta_0) = M_{n-1,T}[g_{\phi_1}(\theta_0)^\top, \ldots, g_{\phi_{n-1}}(\theta_0)^\top]^\top,$$

where $\lim_T M_{n-1,T} \to_{\text{a.s.}} M_{n-1}$. None of the arguments would change.

When evaluated at θ_0, $h_T^{\mathcal{A}_n}$ is a regression residual. This completes the definition of the adjusted-moment estimator.

We now verify that the induction hypotheses (47)–(49) are valid for n. These are necessary to define (51).[7] Because

$$h^{\mathcal{A}_{n-1}}_{\phi_n,(1-\lambda_n)T}(\theta) - g_{\phi_n,\lambda_n T}(\theta) \to_{\text{a.s.}} E\left[f_{\phi_n}(x_t,\theta)\right] - E\left[f_{\phi_n}(x_t,\theta)\right] = 0,$$

(47) is satisfied for n. To show (48) for n, note first that Theorem 5.1 implies

$$\lim_{T\to\infty} E\left[T g_{\phi_n,\lambda_n T}(\theta_0) h^{\mathcal{A}_{n-1}}_{\phi_n,(1-\lambda)T}(\theta_0)^\top\right] = 0.$$

Because $h_T^{\mathcal{A}_n}(\theta_0)$ is a regression residual, (51) implies

$$\lim_{T\to\infty} E\left[T h_T^{\mathcal{A}_n}(\theta_0) h_T^{\mathcal{A}_n}(\theta_0)^\top\right] = \frac{1}{(1-\lambda_n)} S^{\mathcal{A}_{n-1}} - B^{\mathcal{A}_{n-1}}\left[\frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{(1-\lambda_n)} S^{\mathcal{A}_{n-1}}_{\phi_n}\right]\left(B^{\mathcal{A}_{n-1}}\right)^\top, \tag{52}$$

where

$$S^{\mathcal{A}_{n-1}}_{\phi_n} = \lim_{T\to\infty} E\left[T h^{\mathcal{A}_{n-1}}_{\phi_n,(1-\lambda_n)T}(\theta_0) h^{\mathcal{A}_{n-1}}_{\phi_n,(1-\lambda_n)T}(\theta_0)^\top\right].$$

Clearly (52) is well-defined and symmetric. It is positive definite because ϕ_n is a strict subset of $\{1,\ldots,l\}$, so not all the variance in $h^{\mathcal{A}_{n-1}}_{(1-\lambda_n)T}$ can be explained by $h^{\mathcal{A}_{n-1}}_{\phi_n,(1-\lambda_n)T}$. Finally, (49) follows from the form of (51).

It may not be immediately clear that this estimator reduces to the one defined in Section 2 when there are only two blocks of data. In fact, it does reduce to the previously-defined adjusted-moment estimator. As stated in Section 5.1, $\lambda_1 = \lambda$ and $\lambda_2 = (1-\lambda)$. We also have $\phi_1 = \{1,2\}$, and $\phi_2 = \{1\}$. The moment conditions for the first adjusted-moment estimator are the same as in standard GMM:

$$h^{\mathcal{A}_1}_{\lambda_1 T} = g_{\lambda T}.$$

It follows from (51) that

$$\begin{aligned} h_T^{\mathcal{A}_2} &= h^{\mathcal{A}_1}_{\lambda_1 T} - B^{\mathcal{A}_1}\left(h^{\mathcal{A}_1}_{1,\lambda_1 T} - g_{\phi_2,\lambda_2 T}\right) \\ &= g_{\lambda T} - B^{\mathcal{A}_1}\left(g_{1,\lambda T} - g_{1,(1-\lambda)T}\right). \end{aligned}$$

[7] Equation (47) insures a one-to-one correspondence between moment conditions and components of $h^{\mathcal{A}_n}$. Equation (49) insures that $\sqrt{T} h^{\mathcal{A}_{n-1}}_{(1-\lambda_n)T}$ and $\sqrt{T} g_{\phi_n,(1-\lambda_n)}$ have an asymptotic distribution that is well-defined (by Theorem 5.1). This implies that $B^{\mathcal{A}_{n-1}}$ is well-defined. Equation (48) will be useful later in determining the asymptotic distribution of the adjusted-moment estimator.

By (50),

$$\begin{aligned}
B^{\mathcal{A}_1} &= \lim_{T\to\infty} E\left[Tg_{\lambda T}(g_{1,\lambda T}-g_{1,(1-\lambda)T})^\top\right]\left(E\left[T(g_{1,\lambda T}-g_{1,(1-\lambda)T})(g_{1,\lambda T}-g_{1,(1-\lambda)T})^\top\right]\right)^{-1} \\
&= \frac{1}{\lambda}\begin{pmatrix} S_{11} \\ S_{21} \end{pmatrix}\left[\left(\frac{1}{1-\lambda}+\frac{1}{\lambda}\right)S_{11}\right]^{-1} \\
&= \begin{pmatrix} (1-\lambda)I \\ (1-\lambda)B_{21} \end{pmatrix},
\end{aligned}$$

where we have suppressed the argument θ_0 in the first line. Therefore, the moment conditions for the adjusted-moment estimator equal

$$h_T^{\mathcal{A}_2} = \begin{pmatrix} g_{1,\lambda T} + (1-\lambda)\left(g_{1,(1-\lambda)T} - g_{1,\lambda T}\right) \\ g_{2,\lambda T} + (1-\lambda)B_{21}(g_{1,(1-\lambda)T} - g_{1,\lambda T}) \end{pmatrix}, \tag{53}$$

which are the same moment conditions as those given in Section 2.[8]

The usual asymptotic results hold for the extended adjusted-moment estimator. The following lemma is helpful:

Lemma 5.1 *Assumptions 1–5 imply*

$$\sqrt{T}\begin{pmatrix} h^{\mathcal{A}_{n-1}}_{(1-\lambda_n)T}(\theta_0) \\ g_{\phi_n,\lambda_n T}(\theta_0) \end{pmatrix} \to_{\mathrm{d}} N\left(0, \begin{bmatrix} \frac{1}{1-\lambda_n}S^{\mathcal{A}_{n-1}} & 0 \\ 0 & \frac{1}{\lambda_n}S_{\phi_n} \end{bmatrix}\right).$$

Proof It follows from (49) and Theorem 5.1 that $\sqrt{T}h^{\mathcal{A}_{n-1}}_{(1-\lambda_n)T}(\theta_0)$ and $\sqrt{T}g_{\phi_n,\lambda_n T}(\theta_0)$ are asymptotically independent, and that each are asymptotically normally distributed. The form of the asymptotic variance follows from (48) and Theorem 5.1. ▮

This lemma implies that the sample moment conditions for the nth adjusted-moment estimator, scaled by $\sqrt{T}$ are asymptotically normally distributed. Equation (51) implies an inductive equation for the variance.

Theorem 5.7 *Assumptions 1–5 imply*

$$\sqrt{T}h_T^{\mathcal{A}_n}(\theta_0) \to_{\mathrm{d}} N\left(0, S^{\mathcal{A}_n}\right),$$

where $S^{\mathcal{A}_n}$ is defined inductively as

$$S^{\mathcal{A}_n} = \frac{1}{1-\lambda_n}S^{\mathcal{A}_{n-1}} - B^{\mathcal{A}_{n-1}}\left[\frac{1}{\lambda_n}S_{\phi_n} + \frac{1}{1-\lambda_n}S^{\mathcal{A}_{n-1}}_{\phi_n}\right]\left(B^{\mathcal{A}_{n-1}}\right)^\top, \tag{54}$$

with

$$S^{\mathcal{A}_1} = S. \tag{55}$$

[8] Here we make use of the equation $g_{1T} = \lambda g_{1,\lambda_1 T} + (1-\lambda)g_{1,(1-\lambda)T}$.

As with the extended over-identified estimator, any positive definite weighting matrix can be used with moment conditions $h_T^{\mathcal{A}_n}$ to produce a consistent estimator. As usual, we will emphasize the case when the weighting matrix converges almost surely to $S^{\mathcal{A}_n}$. Define

$$\hat{\theta}_T^{\mathcal{A}_n} = \operatorname{argmin}_\theta h_T^{\mathcal{A}_n}(\theta)^\top W_T^{\mathcal{A}_n} h_T^{\mathcal{A}_n}(\theta), \tag{56}$$

where $W_T^{\mathcal{A}_n}$ satisfies Assumption 11:

Assumption 11 *The weighting matrix $W_T^{\mathcal{A}_n}$ converges almost surely to a positive-definite matrix $W^{\mathcal{A}_n}$.*

Consistency for the extended adjusted moment estimator follows from the fact that

$$h_T^{\mathcal{A}_n}(\theta) \to_{\text{a.s.}} E\left[f(x_t, \theta)\right]$$

(proved above by induction) and the arguments of Section 2.

Theorem 5.8 *Assumptions 1-5, and 11 imply that as $T \to \infty$, $\hat{\theta}^{\mathcal{A}_n} \to_{\text{a.s.}} \theta_0$.*

Similarly, it is possible to show that the estimator is asymptotically normally distributed:

Theorem 5.9 *Assumptions 1-5, 7–9 and 11 imply that as $T \to \infty$,*

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

Proof We show by induction on n that

$$\frac{\partial h_T^{\mathcal{A}_n}}{\partial \theta}(\theta_0) \to_{\text{a.s.}} E\left[\left.\frac{\partial f}{\partial \theta}(x_t, \theta)\right|_{\theta_0}\right] \equiv D_0. \tag{57}$$

By definition, and White and Domowitz (1984, Theorem 2.3) it follows that

$$\frac{\partial h_{\lambda_1 T}^{\mathcal{A}_1}}{\partial \theta}(\theta_0) = \frac{\partial g_{\lambda_1 T}}{\partial \theta}(\theta_0) \to_{\text{a.s.}} D_0.$$

Assume (57) holds for $n-1$. By (51),

$$h_T^{\mathcal{A}_n}(\theta) = h_{(1-\lambda_n)T}^{\mathcal{A}_{n-1}}(\theta) - B^{\mathcal{A}_{n-1}}\left(h_{\phi_n,(1-\lambda_n)T}^{\mathcal{A}_{n-1}}(\theta) - g_{\phi_n,\lambda_n T}(\theta)\right). \tag{58}$$

Applying White and Domowitz (1984, Theorem 2.3) again, it follows that

$$\frac{\partial g_{\phi_n,\lambda_n T}}{\partial \theta}(\theta) \to_{\text{a.s.}} D_0.$$

Taking limits on both sides of (58) and using the induction hypothesis produces the desired result. The rest of the proof follows along the same lines as that of Theorem 2.3 in Section 2. ▮

Lastly, given moments $h_T^{\mathcal{A}_n}$, the most efficient asymptotic weighting matrix is the inverse of the variance of these moments.

Theorem 5.10 *Suppose $W_T^{\mathcal{A}_n} \to_{\text{a.s.}} W^{\mathcal{A}_n} = (S^{\mathcal{A}_n})^{-1}$. Assumptions 1–5 and 7–9 imply that*

$$\sqrt{T}(\hat{\theta}_T^{\mathcal{A}_n} - \theta_0) \overset{a}{\sim} N\left(0, \left(D_0^\top \left(S^{\mathcal{A}_n}\right)^{-1} D_0\right)^{-1}\right). \tag{59}$$

Moreover, this choice of $W^{\mathcal{A}_n}$ is efficient given the moment conditions (40).

While the extended adjusted-moment estimator appears completely different from the extended over-identified estimator, they are asymptotically equivalent.

Theorem 5.11 *Assume that $W_T^{\mathcal{A}_n} \to \left(S^{\mathcal{A}_n}\right)^{-1}$ and $W_T^{\mathcal{I}_n} \to \left(S^{\mathcal{I}_n}\right)^{-1}$. For any integer n, assumptions 1–5 and 7–9 imply that the extended adjusted-moment estimator (56) is asymptotically equivalent to the extended over-identified estimator (42).*

A full proof is given in the Appendix. The structure of the proof is similar to that of Theorem 3.1. The preceding theorems show that it suffices to compare the asymptotic variances. Then matrix partition results are used to relate the inverse of the asymptotic variance for the over-identified estimator to the asymptotic variance of the adjusted-moment estimator.

Intuitively, the reason for the equivalence is that both estimators insure that each additional segment reduces the variance in the most efficient way. The variance reduction is easier to see in the case of the over-identified estimator, where each additional segment introduces a new moment condition. The efficient weighting matrix, along with a standard "diversification" argument insures that the new estimator will have a smaller variance than the old estimator. For the adjusted-moment estimator, each step of the further reduces the variance of the moment conditions, because the new moment conditions are defined as regression residuals from the previous step. As regression residuals, they must have smaller variance than the variable on the right-hand side of the regression – the previous moment conditions.

Theorem 5.11 shows that the extended adjusted moment estimator is asymptotically equivalent to the extended over-identified estimator. By Theorem 5.6, we can conclude that adding a block of data always increases efficiency for the adjusted-moment estimator.

Corollary 5.1 *Assume that $W_T^{\mathcal{A}_n} \to \left(S^{\mathcal{A}_n}\right)^{-1}$ and $W_{(1-\lambda_n)T}^{\mathcal{A}_{n-1}} \to \left(S^{\mathcal{A}_{n-1}}\right)^{-1}$. Assumptions 1–5 and 7–9 imply $\sqrt{T}\hat{\theta}_T^{\mathcal{A}_n}$ is asymptotically more efficient than $\sqrt{T}\hat{\theta}_{(1-\lambda_n)T}^{\mathcal{A}_{n-1}}$.*

Defining the adjusted-moment estimator as a regression residual has some appealing properties; it facilitates the proof of equivalence for the over-identified estimator, and it demonstrates clearly the reduction in variance. In other respects, it may appear counterintuitive. In the next section, we compute three examples of the adjusted-moment estimator and show that indeed, they have an interpretation that is equally appealing as in the case where only one data block is missing.

5.3 Examples

This section computes explicit estimators for three examples of missing data patterns. The first example is like that explored in Section 2, except here data is missing at both ends of the sample for one of the series, rather than just at the beginning. The second example is where there are three different starting dates, but all data have the same ending dates. This case is treated in Stambaugh (1997) in a maximum-likelihood setting, and applied to international data. This example shows that our methods can be easily applied to this setting as well. Given the form of the estimators for examples 1 and 2, one could easily combine the reasoning and put together an example where the data have both different starting and ending dates, but that the available data are "nested" (e.g. there are three series, the first of which is observed for the full sample, the second is observed for a subset of the dates, and the third is observed for a subset of the dates for which the second is observed). Little and Rubin (2002) refer to this condition as monotonicity, and derive a maximum likelihood estimator under normality and independent, identically distributed observations.

The last example is a case where the series have different starting dates and different ending dates, but that the series that start later also end later. This example illustrates the power of our generalization above, as its form for the adjusted-moment estimator is non-obvious.

In all of these cases, we derive both the over-identified estimator and the adjusted-moment estimator. For the over-identified estimator, we derive both the moment conditions, the optimal matrix, and the form of $D_0^{\mathcal{I}}$. For the adjusted-moment matrix, the derivative of the moments always equals D_0 asymptotically. The optimal weighting matrix is the inverse of the variance of the moments, which can be computed from (54). If the original problem is exactly identified, it will remain so with the adjusted-moment estimator. Also, the extended adjusted-moment estimator will be consistent, and efficient relative to the estimator that uses a shorter length of data, for any choice of positive-definite weighting matrix.

5.3.1 Data missing at both ends

The first example is similar to the case in Section 2, except that data from the second set of series is missing not only at the beginning of the sample, but also at the end.[9] Figure 3 illustrates this pattern of missing data. As in Section 2, group the moment conditions observed for the full data set into a vector $f_1(x_{1t}, \theta)$, and moment conditions only observed for the middle segment into a vector $f_2(x_t, \theta)$. This situation would occur if the series for which data is missing at the start of the sample also is updated less frequently. We use $g_{1,\cdot}(\theta)$ to denote partial sums of $f_1(x_{1t}, \theta)$ and $g_{2,\cdot}(\theta)$ to denote partial sums of $f_2(x_t, \theta)$, where $\cdot$ will represent the length of the segment over which the observation is taken. The notation for sub-matrices of S and D_0 follows the same conventions as in Section 2.

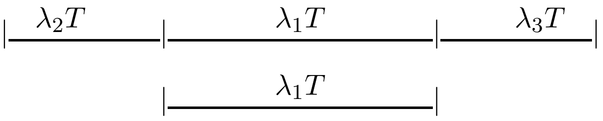

Figure 3: Illustration of Example 1. Example 1 explicitly calculates the extended estimators for data missing at both ends. The notation above the horizontal lines refers to the length of each segment as a function of the sample size T.

As shown in Figure 3, λ_1 is the length of the middle segment divided by the total data length. Without loss of generality, we assume λ_2 is the length of the first segment of missing data divided by the total length (we could also have set λ_2 equal to the length of the second segment of missing data divided by the total data length). The moment conditions for the over-identified estimator $h_T^{\mathcal{I}_3}$ are

$$h_T^{\mathcal{I}_3}(\theta) = \left[g_{1,\lambda_1 T}(\theta)^\top \; g_{2,\lambda_1 T}(\theta)^\top \; g_{1,\lambda_2 T}(\theta)^\top \; g_{1,\lambda_3 T}(\theta)^\top \right]^\top .$$

Then the results in Section 5.1 imply that $\sqrt{T} h_T^{\mathcal{I}_3}(\theta)$ has asymptotic variance[10]

$$S^{\mathcal{I}_3} = \begin{bmatrix} \frac{1}{\lambda_1} S_{11} & \frac{1}{\lambda_1} S_{12} & 0 & 0 \\ \frac{1}{\lambda_1} S_{21} & \frac{1}{\lambda_1} S_{22} & 0 & 0 \\ 0 & 0 & \frac{1}{\lambda_2} S_{11} & 0 \\ 0 & 0 & 0 & \frac{1}{\lambda_3} S_{11} \end{bmatrix} .$$

[9] If the observations were independent, then this case is clearly identical to that described in Section 2 because the data points could be rearranged without effecting the joint distribution. Under dependence, this does not follow immediately.

[10] The asymptotic variance does not take exactly the same form as $S^{\mathcal{I}}$ in Section 2. The reason for the discrepancy is that $S^{\mathcal{I}}$ was defined as the asymptotic variance of the moment conditions scaled by $\sqrt{\lambda T}$, while $S^{\mathcal{I}_3}$ is the asymptotic variance of the moment conditions scaled by $\sqrt{T}$.

The extended over-identified estimator with efficient weighting matrix is therefore

$$\hat{\theta}^{\mathcal{I}_3} = \operatorname{argmin}_\theta h_T^{\mathcal{I}_3}(\theta)^\top W_T^{\mathcal{I}_3} h_T^{\mathcal{I}_3}(\theta), \quad W_T^{\mathcal{I}_3} \to_{\text{a.s.}} \left(S^{\mathcal{I}_3}\right)^{-1}$$

for $W_T^{\mathcal{I}_3}$ positive definite. The asymptotic distribution is given by

$$\sqrt{T}(\hat{\theta}_T^{\mathcal{I}_3} - \theta_0) \to_{\text{d}} N\left(0, \left((D_0^{\mathcal{I}_3})^\top \left(S^{\mathcal{I}_3}\right)^{-1} (D_0^{\mathcal{I}_3})\right)^{-1}\right),$$

where

$$D^{\mathcal{I}_3} = \left[D_{0,1}^\top \; D_{0,2}^\top \; D_{0,1}^\top \; D_{0,1}^\top\right]^\top.$$

We now describe the extended adjusted-moment estimator. The moment conditions for the first adjusted-moment estimator are the same as in standard GMM:

$$h_{\lambda_1 T}^{\mathcal{A}_1} = g_{\lambda_1 T}. \tag{60}$$

Substituting (60) into (51) yields the adjusted-moment estimator that includes the λ_2 block:

$$h_{(\lambda_1+\lambda_2)T}^{\mathcal{A}_2} = g_{\lambda_1 T} - B^{\mathcal{A}_1}\left(g_{1,\lambda_1 T} - g_{1,\lambda_2 T}\right), \tag{61}$$

where

$$\begin{aligned} B^{\mathcal{A}_1} &= \lim_{T\to\infty} E\left[(\lambda_1+\lambda_2)T g_{\lambda_1 T}\left(g_{1,\lambda_1 T} - g_{1,\lambda_2 T}\right)^\top\right]\left(E\left[(\lambda_1+\lambda_2)T\left(g_{1,\lambda_1 T} - g_{1,\lambda_2 T}\right)\left(g_{1,\lambda_1 T} - g_{1,\lambda_2 T}\right)^\top\right]\right)^{-1} \\ &= \frac{\lambda_2}{\lambda_1+\lambda_2}\begin{pmatrix} I \\ B_{21} \end{pmatrix}, \end{aligned}$$

where we have suppressed the θ_0 argument in the first line. Substituting into (61) produces

$$h_{(\lambda_1+\lambda_2)T}^{\mathcal{A}_2} = \begin{pmatrix} g_{1,(\lambda_1+\lambda_2)T} \\ g_{2,\lambda_1 T} + \frac{\lambda_2}{\lambda_1+\lambda_2} B_{21}\left(g_{1,\lambda_2 T} - g_{1,\lambda_1 T}\right) \end{pmatrix}, \tag{62}$$

which is the same estimator described in Section 2, except that the length of the sample is taken to be $(\lambda_1+\lambda_2)T$ rather than T.[11]

To construct the full adjusted-moment estimator for this case, we apply (51) again:

$$h_T^{\mathcal{A}_3} = h_{(\lambda_1+\lambda_2)T}^{\mathcal{A}_2} - B^{\mathcal{A}_2}\left(h_{1,(\lambda_1+\lambda_2)T}^{\mathcal{A}_2} - g_{1,\lambda_3 T}\right), \tag{64}$$

[11] Here and in the following computations, we make use of the equation

$$g_{1,(\lambda_1+\lambda_2)T} = \frac{\lambda_1}{\lambda_1+\lambda_2} g_{1,\lambda_1 T} + \frac{\lambda_2}{\lambda_1+\lambda_2} g_{1,\lambda_2 T}. \tag{63}$$

where

$$\begin{aligned} B^{\mathcal{A}_2} =& \lim_{T\to\infty} E\left[Th^{\mathcal{A}_2}_{(\lambda_1+\lambda_2)T}\left(h^{\mathcal{A}_2}_{1,(\lambda_1+\lambda_2)T} - g_{1,\lambda_3 T}\right)^\top\right] \\ & \times\left(E\left[T\left(h^{\mathcal{A}_2}_{(\lambda_1+\lambda_2)T} - g_{1,\lambda_3 T}\right)\left(h^{\mathcal{A}_2}_{1,(\lambda_1+\lambda_2)T} - g_{1,\lambda_3 T}\right)^\top\right]\right)^{-1} \qquad (65) \\ =& \lim_{T\to\infty} E\left[Th^{\mathcal{A}_2}_{(\lambda_1+\lambda_2)T}\left(g_{1,(\lambda_1+\lambda_2)T} - g_{1,\lambda_3 T}\right)^\top\right]\left(E\left[T\left(g_{1,(\lambda_1+\lambda_2)T} - g_{1,\lambda_3 T}\right)\left(g_{1,(\lambda_1+\lambda_2)T} - g_{1,\lambda_3 T}\right)^\top\right]\right)^{-1}. \end{aligned}$$

It follows from Theorem 5.1 that

$$\lim_{T\to\infty} E\left[T\left(g_{1,(\lambda_1+\lambda_2)T} - g_{1,\lambda_3 T}\right)\left(g_{1,(\lambda_1+\lambda_2)T} - g_{1,\lambda_3 T}\right)^\top\right] = \left(\frac{1}{\lambda_3} + \frac{1}{\lambda_1+\lambda_2}\right)S_{11}. \qquad (66)$$

Using (62) and the same argument,

$$\lim_{T\to\infty} E\left[Th^{\mathcal{A}_2}_{1,(\lambda_1+\lambda_2)T}\left(g_{1,(\lambda_1+\lambda_2)T} - g_{1,\lambda_3 T}\right)^\top\right] = \frac{1}{\lambda_1+\lambda_2}S_{11}. \qquad (67)$$

Finally, Theorem 5.1 and the same reasoning used to show (22) that

$$\begin{aligned} \lim_{T\to\infty} E\left[Th^{\mathcal{A}_2}_{2,(\lambda_1+\lambda_2)T}\left(g_{1,(\lambda_1+\lambda_2)T} - g_{1,\lambda_3 T}\right)^\top\right] &= \lim_{T\to\infty} E\left[Th^{\mathcal{A}_2}_{2,(\lambda_1+\lambda_2)T}\; g^\top_{1,(\lambda_1+\lambda_2)T}\right] \\ &= \frac{1}{\lambda_1+\lambda_2}S_{21}, \qquad (68) \end{aligned}$$

where we have continued to suppressed the argument θ_0. Combining (66), (67), and (68), and rearranging,

$$B^{\mathcal{A}_2} = \lambda_3\begin{pmatrix} I \\ B_{21}\end{pmatrix}.$$

Substituting into (64) and rearranging produces[12]

$$h^{\mathcal{A}_3}_T = \begin{pmatrix} g_{1,T} \\ g_{2,\lambda_1 T} + (\lambda_2+\lambda_3)B_{21}\left(\frac{\lambda_2}{\lambda_2+\lambda_3}g_{1,\lambda_2 T} + \frac{\lambda_3}{\lambda_2+\lambda_3}g_{1,\lambda_3 T} - g_{1,\lambda_1 T}\right)\end{pmatrix}.$$

Several features of this extended adjusted-moment estimator are worth noting. First, the moment condition for the series observed for the full data set is the same as if these series were estimated independently of the second set of moments. The basic adjusted-moment estimator described in Section 2) also had this property, and, as we argued in Section 3 this may be a reason to prefer the adjusted-moment estimator over the over-identified estimator. Second, the adjustment to the moments of the second series is the same as if the segments λ_2 and λ_3 were contiguous rather than separated by λ_1. For our asymptotic results, it does not matter whether the blocks defined by starting and ending points are contiguous.

[12] Here and in the following example, we use the fact that $\lambda_1+\lambda_2+\lambda_3 = 1$, and that

$$\begin{aligned} g_{1,T} &= (\lambda_1+\lambda_2)g_{1,(\lambda_1+\lambda_2)T} + \lambda_3 g_{2,\lambda_3 T} \\ &= \lambda_1 g_{1,\lambda_1 T} + \lambda_2 g_{1,\lambda_2 T} + \lambda_3 g_{2,\lambda_3 T}. \end{aligned}$$

5.3.2 Data missing in a monotonic pattern

The second example represents a problem dealt with in detail in a maximum likelihood context by Little and Rubin (2002) and Stambaugh (1997). Here, the data series all end at the same point, but may start from more than two different points. This may occur, for example, if one is using international data as in the study by Stambaugh. Figure 4 illustrates the missing data pattern in this case. For ease of notation, we illustrate the extended over-identified and extended adjusted-moment estimators for the case where there are three starting dates. Extending the method further to more than three starting dates is straightforward.

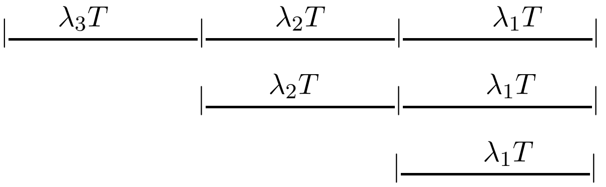

Figure 4: Illustration of Example 2. Example 2 explicitly calculates the extended estimators for data missing in a monotonic pattern. The notation above the horizontal lines refers to the length of each segment as a function of the sample size T.

As shown in Figure 4, λ_1 is the length of the final segment divided by the total data length. This is because all series are observed for the segment of length $\lambda_1 T$. A subset of these series are also observed for the middle segment: this has length $\lambda_2 T$. A smaller subset is also observed for the first segment, of length $\lambda_3 T = (1 - \lambda_1 - \lambda_2)T$. Following the notational convention of Section 2 and the previous example, we group the moment conditions observed for the full data set into a vector $f_1(x_{1t}, \theta)$, the moment conditions observed for the last two data segments into a vector $f_2(x_{1t}, x_{2t}, \theta)$, and the moment conditions observed only for the last data segment into a vector $f_3(x_{1t}, x_{2t}, x_{3t}, \theta)$. The notation for sub-vectors of g and submatrices of D_0 and S follows the same convention as in the previous example.

The moment conditions for the over-identified estimator $h_T^{\mathcal{I}_3}$ are

$$h_T^{\mathcal{I}_3}(\theta) = \left[g_{1,\lambda_1 T}(\theta)^\top \; g_{2,\lambda_1 T}(\theta)^\top \; g_{3,\lambda_1 T}(\theta)^\top \; g_{1,\lambda_2 T}(\theta)^\top \; g_{2,\lambda_2 T}(\theta)^\top g_{1,\lambda_3 T}(\theta)^\top \right].$$

The results of Section 5.1 imply that $\sqrt{T} h_T^{\mathcal{I}_3}$ has asymptotic variance

$$S^{\mathcal{I}_3} = \left[\begin{array}{cccccc} \frac{1}{\lambda_1} S_{11} & \frac{1}{\lambda_1} S_{12} & \frac{1}{\lambda_1} S_{13} & 0 & 0 & 0 \\ \frac{1}{\lambda_1} S_{21} & \frac{1}{\lambda_1} S_{22} & \frac{1}{\lambda_1} S_{23} & 0 & 0 & 0 \\ \frac{1}{\lambda_1} S_{31} & \frac{1}{\lambda_1} S_{32} & \frac{1}{\lambda_1} S_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\lambda_2} S_{11} & \frac{1}{\lambda_2} S_{12} & 0 \\ 0 & 0 & 0 & \frac{1}{\lambda_2} S_{21} & \frac{1}{\lambda_2} S_{22} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{\lambda_3} S_{11} \end{array} \right].$$

In this example the extended over-identified estimator is therefore

$$\hat{\theta}_T^{\mathcal{I}_3} = \text{argmin}_\theta \, h_T^{\mathcal{I}_3}(\theta)^\top W_T^{\mathcal{I}_3} h_T^{\mathcal{I}_3}(\theta),$$

for $W_T^{\mathcal{I}_3}$ positive definite and $W_T^{\mathcal{I}_3} \to_{\text{a.s.}} \left(S^{\mathcal{I}_3}\right)^{-1}$. The estimator has asymptotic distribution

$$\sqrt{T}(\hat{\theta}_T^{\mathcal{I}_3} - \theta_0) \to_{\text{d}} N\left(0, \left((D_0^{\mathcal{I}_3})^\top \left(S^{\mathcal{I}_3}\right)^{-1} (D_0^{\mathcal{I}_3})\right)^{-1}\right),$$

where

$$D^{\mathcal{I}_3} = \left[D_{0,1}^\top \; D_{0,2}^\top \; D_{0,3}^\top \;\; D_{0,1}^\top \; D_{0,2}^\top \; D_{0,1}^\top \right]^\top.$$

We now describe the extended adjusted-moment estimator. The first step is the same as standard GMM for the three series:

$$h_{\lambda_1 T}^{\mathcal{A}_1} = \left[\begin{array}{c} g_{1,\lambda_1 T} \\ g_{2,\lambda_1 T} \\ g_{3,\lambda_1 T} \end{array} \right].$$

The second step is the same as the second step in the example above. However, here two sets of series are observed for the longer sample, g_1 and g_2. Therefore

$$B^{\mathcal{A}_1} = \frac{\lambda_2}{\lambda_1 + \lambda_2} \left(\begin{array}{c} I \\ B_{3\cdot 12} \end{array} \right)$$

and

$$h_{(\lambda_1+\lambda_2)T}^{\mathcal{A}_2} = \left[\begin{array}{l} g_{1,(\lambda_1+\lambda_2)T} \\ g_{2,(\lambda_1+\lambda_2)T} \\ g_{3,\lambda_1 T} + \frac{\lambda_2}{\lambda_1+\lambda_2} B_{3\cdot 12} \left(\begin{array}{c} g_{1,\lambda_2 T} - g_{1,\lambda_1 T} \\ g_{2,\lambda_2 T} - g_{2,\lambda_1 T} \end{array} \right) \end{array} \right], \tag{69}$$

where $B_{3\cdot 12}$ are the coefficients from a multivariate regression on the third set of series on the first two:

$$B_{3\cdot 12} = [S_{31} \;\; S_{32}] \left[\begin{array}{cc} S_{11} & S_{12} \\ S_{21} & S_{22} \end{array} \right]^{-1}.$$

In the third step, we add the segment of length $\lambda_3 T$. Then

$$h_T^{\mathcal{A}_3} = h_{(\lambda_1+\lambda_2)T}^{\mathcal{A}_2} - B^{\mathcal{A}_2} \left(h_{1,(\lambda_1+\lambda_2)T}^{\mathcal{A}_2} - g_{1,\lambda_3 T} \right), \tag{70}$$

where the expression for $B^{\mathcal{A}_2}$ is given by (65). It follows from Theorem 5.1 that

$$\lim_{T\to\infty} E\left[T\left(h^{\mathcal{A}_2}_{1,(\lambda_1+\lambda_2)T} - g_{1,\lambda_3 T}\right)\left(h^{\mathcal{A}_2}_{1,(\lambda_1+\lambda_2)T} - g_{1,\lambda_3 T}\right)^{\top}\right]^{-1} = \left(\frac{1}{\lambda_3} + \frac{1}{\lambda_1+\lambda_2}\right) S_{11}.$$

Similar reasoning shows that

$$\lim_{T\to\infty} E\left[T h^{\mathcal{A}_2}_{j,(\lambda_1+\lambda_2)T}\left(h^{\mathcal{A}_2}_{1,(\lambda_1+\lambda_2)T} - g_{1,\lambda_3 T}\right)^{\top}\right] = \frac{1}{\lambda_1+\lambda_2} S_{j1}, \quad j = 1,2,3,$$

where we have made use of (22) for $j = 3$. Substituting into (70), and applying footnote 12 results in

$$h^{\mathcal{A}_3}_T = \left[\begin{array}{l} g_{1T} \\ g_{2,(\lambda_1+\lambda_2)T} + \lambda_3 B_{21}(g_{1,\lambda_3 T} - g_{1,(\lambda_1+\lambda_2)T}) \\ g_{3,\lambda_1 T} + \frac{\lambda_2}{\lambda_1+\lambda_2} B_{3\cdot 12}\left(\begin{array}{c} g_{1,\lambda_2 T} - g_{1,\lambda_1 T} \\ g_{2,\lambda_2 T} - g_{2,\lambda_1 T}\end{array}\right) + \lambda_3 B_{31}\left(g_{1,\lambda_3 T} - g_{1,(\lambda_1+\lambda_2)T}\right)\end{array}\right],$$

where $B_{31} = S_{31}S_{11}^{-1}$.

Note that the moment conditions for the data series observed for the full data set are the same as if these series were estimated independently of the second and third set of moments. Indeed, the moment conditions for the data series observed for both λ_1 and λ_2 are the same as if these series were estimated (using the adjusted-moment estimator) without the third set of moments. Thus the principle advantage of the adjusted-moment estimator for two starting dates is retained and extended in this example with multiple starting dates.

In constructing this estimator, we have assumed that all the missing data occurs at the beginning of the sample. However, the estimator would take the same form if the missing data were at the end. Indeed, as the previous section shows, it suffices to have the data observed for the third set of series be nested in the data observed for the second set, which is nested in the data observed for the first set. In other words, data could be missing at both ends of the sample. In this case, the adjusted-moment estimator would take the same form as above.

5.3.3 Data missing in a non-monotonic pattern

Our last example represents a case not handled in the maximum likelihood settings of Little and Rubin (2002) and Stambaugh (1997). In this example, there are two sets of moments. These moments have different starting dates and different ending dates, as in the first example. However, the series that ends earlier also starts earlier, so neither series is observed for the full length. Figure 5 illustrates the pattern of missing data in this example.

We refer to the length of the middle data segment as $\lambda_1 T$ because all data are observed over this segment. We could let $\lambda_2 T$ denote the length of the first or the last data segment. Without

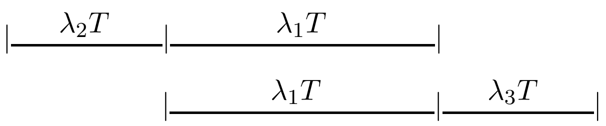

Figure 5: Illustration of Example 3. Example 3 explicitly calculates the extended estimators for data missing in a non-monotonic pattern. The notation above the horizontal lines refers to the length of each segment as a function of the sample size T.

loss of generality, we let it refer to the length of the first segment. We let $\lambda_3 T = (1 - \lambda_1 - \lambda_2)T$ denote the length of the final segment. Following the notation convention of Section 2 and the previous examples, we group the moment conditions observed for the first two segments into a vector $f_1(x_{1t}, \theta)$ and the moment conditions observed for the last two segments into a vector $f_2(x_{2t}, \theta)$. The notation for sub-vectors of g and submatrices of D_0 and S follows the same convention as in the previous example.

The moment conditions for the over-identified estimator $h_T^{\mathcal{I}_3}$ are

$$h_T^{\mathcal{I}_3}(\theta) = \left[g_{1,\lambda_1 T}(\theta)^\top \; g_{2,\lambda_1 T}(\theta)^\top \; g_{1,\lambda_2 T}(\theta)^\top \; g_{2,\lambda_3 T}(\theta)^\top \right]^\top .$$

Then the results in Section 5.1 imply that $\sqrt{T} h_T^{\mathcal{I}_3}(\theta)$ has asymptotic variance

$$S^{\mathcal{I}_3} = \begin{bmatrix} \frac{1}{\lambda_1} S_{11} & \frac{1}{\lambda_1} S_{12} & 0 & 0 \\ \frac{1}{\lambda_1} S_{21} & \frac{1}{\lambda_1} S_{22} & 0 & 0 \\ 0 & 0 & \frac{1}{\lambda_2} S_{11} & 0 \\ 0 & 0 & 0 & \frac{1}{\lambda_3} S_{22} \end{bmatrix} .$$

The extended over-identified estimator is therefore

$$\hat{\theta}_T^{\mathcal{I}_3} = \operatorname{argmin}_\theta h_T^{\mathcal{I}_3}(\theta)^\top W_T^{\mathcal{I}_3} h_T^{\mathcal{I}_3}(\theta)$$

for $W_T^{\mathcal{I}_3}$ positive definite and $W_T^{\mathcal{I}_3} \to_{\text{a.s.}} \left(S^{\mathcal{I}_3}\right)^{-1}$. The asymptotic distribution is

$$\sqrt{T}(\hat{\theta}_T^{\mathcal{I}_3} - \theta_0) \to_{\text{d}} N\left(0, \left((D_0^{\mathcal{I}_3})^\top \left(S^{\mathcal{I}_3}\right)^{-1} (D_0^{\mathcal{I}_3}) \right)^{-1} \right),$$

where

$$D^{\mathcal{I}_3} = \left[D_{0,1}^\top \; D_{0,2}^\top \; D_{0,1}^\top \; D_{0,2}^\top \right]^\top .$$

We now describe the adjusted-moment estimator. The first two steps in constructing the adjusted-moment estimator are identical to those in the first example. Therefore we can write

$$h_{(\lambda_1+\lambda_2)T}^{\mathcal{A}_2} = \begin{pmatrix} g_{1,(\lambda_1+\lambda_2)T} \\ g_{2,\lambda_1 T} + \frac{\lambda_2}{\lambda_1+\lambda_2} B_{21} \left(g_{1,\lambda_2 T} - g_{1,\lambda_1 T} \right) \end{pmatrix} . \tag{71}$$

We have

$$h_T^{\mathcal{A}_3} = h_{(\lambda_1+\lambda_2)T}^{\mathcal{A}_2} - B^{\mathcal{A}_2}\left(h_{2,(\lambda_1+\lambda_2)T}^{\mathcal{A}_2} - g_{2,\lambda_3 T}\right), \tag{72}$$

where

$$B^{\mathcal{A}_2} = \lim_{T\to\infty} E\left[T h_{(\lambda_1+\lambda_2)T}^{\mathcal{A}_2}\left(h_{2,(\lambda_1+\lambda_2)T}^{\mathcal{A}_2} - g_{2,\lambda_3 T}\right)^{\top}\right] \\ \times E\left(\left[T\left(h_{2,(\lambda_1+\lambda_2)T}^{\mathcal{A}_2} - g_{2,\lambda_3 T}\right)\left(h_{2,(\lambda_1+\lambda_2)T}^{\mathcal{A}_2} - g_{2,\lambda_3 T}\right)^{\top}\right]\right)^{-1},$$

where we have suppressed the argument θ_0. Define

$$\gamma = \frac{\lambda_2}{\lambda_1+\lambda_2}\frac{\lambda_3}{\lambda_1+\lambda_3}.$$

Standard arguments (given in the Appendix) show that

$$B^{\mathcal{A}_2} = \frac{\lambda_3}{\lambda_1+\lambda_3}\left(\begin{array}{l} \frac{\lambda_1}{\lambda_1+\lambda_2}S_{12} \\ S_{22} - \frac{\lambda_2}{\lambda_1+\lambda_2}S_{21}S_{11}^{-1}S_{12}\end{array}\right) S_{22}^{-1}\left(I - \gamma B_{21}B_{12}\right)^{-1}. \tag{73}$$

Given $B^{\mathcal{A}_2}$, (72) gives the moments for the adjusted-moment estimator. The first component is as follows:

$$h_{1T}^{\mathcal{A}_3} = g_{1,(\lambda_1+\lambda_2)T} + \frac{\lambda_1}{\lambda_1+\lambda_2}\frac{\lambda_3}{\lambda_1+\lambda_3}B_{12}\left(I-\gamma B_{21}B_{12}\right)^{-1}\left(g_{2,\lambda_3 T} - g_{2,\lambda_1 T} - \frac{\lambda_2}{\lambda_1+\lambda_2}B_{21}(g_{1,\lambda_2 T} - g_{1,\lambda_1 T})\right), \tag{74}$$

while more extensive matrix algebra results in the following expression for the second component:

$$h_{2T}^{\mathcal{A}_3} = g_{2,(\lambda_1+\lambda_3)T} + \frac{\lambda_1}{\lambda_1+\lambda_3}\frac{\lambda_2}{\lambda_1+\lambda_2}\left(I-\gamma B_{21}B_{12}\right)^{-1}B_{21}\left(g_{1,\lambda_2 T} - g_{1,\lambda_1 T} - \frac{\lambda_3}{\lambda_1+\lambda_3}B_{12}(g_{2,\lambda_3 T} - g_{2,\lambda_1 T})\right). \tag{75}$$

Because

$$B_{12}\left(I-\gamma B_{21}B_{12}\right)^{-1} = \left(I-\gamma B_{12}B_{21}\right)^{-1}B_{12}, \tag{76}$$

these expressions are symmetric.[13]

At first glance, the adjustments implicit in (74) and (75) do not seem as intuitive as their counterparts in Section 2, or, for that matter, in Sections 5.3.1 and 5.3.2. However, there is a

[13] Equation (76) can be shown by noting that

$$\left(I-\gamma B_{21}B_{12}\right)^{-1} = \sum_{m=0}^{\infty}\gamma^m (B_{21}B_{12})^m. \tag{77}$$

reason for the apparently strange form. It follows from (77) and (63) that (74) can be rewritten as

$$h_{1T}^{\mathcal{A}_3} = \frac{\lambda_2}{\lambda_1+\lambda_2} g_{1,\lambda_2 T} + \frac{\lambda_1}{\lambda_1+\lambda_2} g_{1,\lambda_1 T} + \\ \frac{\lambda_1}{\lambda_1+\lambda_2}\frac{\lambda_3}{\lambda_1+\lambda_3} B_{12}\left(\sum_{m=0}^{\infty}(\gamma B_{21}B_{12})^m\right)\left(g_{2,\lambda_3 T} - g_{2,\lambda_1 T} - \frac{\lambda_2}{\lambda_1+\lambda_2} B_{21}(g_{1,\lambda_2 T} - g_{1,\lambda_1 T})\right). \quad (78)$$

It is instructive to expand out the infinite sum explicitly:

$$\begin{aligned} h_{1T}^{\mathcal{A}_3} = \frac{\lambda_2}{\lambda_1+\lambda_2} g_{1,\lambda_2 T} + \frac{\lambda_1}{\lambda_1+\lambda_2} g_{1,\lambda_1 T} &+ \frac{\lambda_1}{\lambda_1+\lambda_2}\frac{\lambda_3}{\lambda_1+\lambda_3} B_{12}(g_{2,\lambda_3 T} - g_{2,\lambda_1 T}) \\ &- \frac{\lambda_1}{\lambda_1+\lambda_2}\frac{\lambda_3}{\lambda_1+\lambda_3} B_{12}\frac{\lambda_2}{\lambda_1+\lambda_2} B_{21}(g_{1,\lambda_2 T} - g_{1,\lambda_1 T}) \\ &+ \frac{\lambda_1}{\lambda_1+\lambda_2}\frac{\lambda_3}{\lambda_1+\lambda_3} B_{12}\frac{\lambda_2}{\lambda_1+\lambda_2} B_{21}\frac{\lambda_3}{\lambda_1+\lambda_3} B_{12}(g_{2,\lambda_3 T} - g_{2,\lambda_1 T}) - \ldots \end{aligned} \quad (79)$$

The first two terms are partial sums of g over the data segment of length $\lambda_2 T$ and the data segment of length $\lambda_1 T$, weighted appropriately. The third term is the adjustment to $g_{1,\lambda_1 T}$, given that g_2 is observed over the longer data segment (precisely, the segment of length $(\lambda_1 + \lambda_3)T$). This is the same adjustment as in Section 2, except here it is the first rather than the second series that is being adjusted. Because $g_{1,\lambda_1 T}$ is weighted by $\lambda_1/(\lambda_1 + \lambda_2)$, the adjustment also receives this weight. Note that there is no adjustment to $g_{1,\lambda_2 T}$ because the second data series is not observed over the period of length $\lambda_2 T$.

One possibility would be to stop with the third term. However, the resulting estimator would be inefficient relative to the generalized adjusted-moment estimator. Instead, the extended adjusted-moment estimator has additional terms. The reason is that the adjustment, $\frac{\lambda_3}{\lambda_1+\lambda_3} B_{12}(g_{2,\lambda_3 T} - g_{2,\lambda_1 T})$, must itself be adjusted to reflect the fact that the first set of series is observed over the data segment of length $\lambda_2 T$. More precisely, $-\frac{\lambda_3}{\lambda_1+\lambda_3} B_{12} g_{2,\lambda_1 T}$ must be adjusted. This is the reason for the fourth term. But then this must also be adjusted, and so forth. Repeating this argument results in the telescoping matrix series (79), which, by (77), converges to the extended adjusted-moment estimator (74). A symmetric explanation holds for (75). Thus even in this complicated problem, the adjusted-moment estimator produces moments that have intuitive appeal.

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 standard GMM. 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. Interestingly, this gain in efficiency is not only present for the parameters that enter into the moment conditions observed over the longer data. As long as there is some interaction between the moment conditions observed over the long data and the series observed over the short data there is an efficiency gain for all the parameters. This interaction can be through covariances between the moment conditions, or through the fact that some parameters appear in both the long-sample and short-sample moment conditions.

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 regression coefficients from a regression of the short-series moments on the long-series moments. 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

Lemma A.1 *Assume V_1 and V_2 are invertible. If $V_1 - V_2$ is positive semi-definite, then $V_2^{-1} - V_1^{-1}$ is also positive semi-definite.*

Lemma A.2 *Assume V_1 and V_2 are invertible. If $V_1 - V_2$ is positive semi-definite, then for any matrix Z, $(Z^\top V_1^{-1} Z)^{-1} - (Z^\top V_2^{-1} Z)^{-1}$ is also positive semi-definite.*

Proof Assume $V_1 - V_2$ is positive semi-definite. By Lemma A.1, $V_2^{-1} - V_1^{-1}$ is positive semi-definite. For any vector c and matrix Z,

$$(Zc)^\top (V_2^{-1} - V_1^{-1})(Zc) \geq 0.$$

Therefore

$$c^\top Z^\top (V_2^{-1} - V_1^{-1}) Zc \geq 0 \quad \forall c,$$

which shows $Z^\top (V_2^{-1} - V_1^{-1}) Z$ is positive semi-definite. Applying Lemma A.1 a second time shows that $(Z^\top V_1^{-1} Z)^{-1} - (Z^\top V_2^{-1} Z)^{-1}$ is positive semi-definite as required. ▮

Lemma A.3 *Let*

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

be a symmetric invertible matrix. Then

$$S^{-1} = \begin{bmatrix} S_{11}^{-1} + B_{21}^\top \Sigma^{-1} B_{21} & -B_{21}^\top \Sigma^{-1} \\ -\Sigma^{-1} B_{21} & \Sigma^{-1} \end{bmatrix}, \tag{80}$$

where Σ is defined by (8). 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

$$\bar{S}^{-1} = \begin{bmatrix} \frac{1}{\lambda} S_{11}^{-1} + B_{21}^\top \Sigma^{-1} B_{12} & -B_{21}^\top \Sigma^{-1} \\ -\Sigma^{-1} B_{12} & \Sigma^{-1} \end{bmatrix}. \tag{81}$$

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 (81) holds. ∎

Proof of Theorem 5.11:

By Theorem 5.10, it suffices to show that the asymptotic variance of $\sqrt{T}\hat{\theta}_T^{\mathcal{I}_n}$ is the same as the asymptotic variance of $\sqrt{T}\hat{\theta}_T^{\mathcal{A}_n}$. The proof is by induction on n. For $n = 1$,

$$\theta_{\lambda_1 T}^{\mathcal{I}_1} = \theta_{\lambda_1 T}^{\mathcal{A}_1}$$

because they both equal the standard GMM estimator over data of length $\lambda_1 T$. We assume by induction that

$$D_0^\top \left(S^{\mathcal{A}_{n-1}}\right)^{-1} D_0 = \left(D_0^{\mathcal{I}_{n-1}}\right)^\top \left(S^{\mathcal{I}_{n-1}}\right)^{-1} D_0^{\mathcal{I}_{n-1}}.$$

Without loss of generality, let $\phi_n = \{1, \ldots, \pi_n\}$. That is, the first π_n moment conditions are observed over data region λ_n. By (45) it suffices to show

$$D_0^\top \left(S^{\mathcal{A}_n}\right)^{-1} D_0 - (1-\lambda_n) D_0^\top \left(S^{\mathcal{A}_{n-1}}\right)^{-1} D_0 = \lambda_n D_{0,\phi_n}^\top S_{\phi_n}^{-1} D_{0,\phi_n}.$$

Equivalently, it suffices to show

$$\left(S^{\mathcal{A}_n}\right)^{-1} = (1-\lambda_n)\left(S^{\mathcal{A}_{n-1}}\right)^{-1} + \lambda_n \begin{bmatrix} S_{\phi_n}^{-1} & 0 \\ 0 & 0 \end{bmatrix}. \tag{82}$$

We use the formula for the inverse of a partitioned matrix (Lemma A.3). Let $-\phi_n$ denote the set of data series not observed over λ_n, i.e. the complement of ϕ_n. The assumption that $\phi_n = \{1, \ldots, \pi_n\}$ implies that $S^{\mathcal{A}_n}$ can be written as

$$S^{\mathcal{A}_n} = \begin{bmatrix} S_{\phi_n}^{\mathcal{A}_n} & S_{\phi_n,-\phi_n}^{\mathcal{A}_n} \\ S_{-\phi_n,\phi_n}^{\mathcal{A}_n} & S_{-\phi_n}^{\mathcal{A}_n} \end{bmatrix},$$

where

$$S_{-\phi_n}^{\mathcal{A}_n} = E\left[T\ h_{-\phi_n,T}^{\mathcal{A}_n}(\theta_0) h_{-\phi_n,T}^{\mathcal{A}_n}(\theta_0)^\top\right],$$

$$S_{-\phi_n,\phi_n}^{\mathcal{A}_n} = E\left[T h_{-\phi_n,T}^{\mathcal{A}_n}(\theta_0) h_{\phi_n,T}^{\mathcal{A}_n}(\theta_0)^\top\right],$$

and $S^{\mathcal{A}_n}_{\phi_n,-\phi_n} = \left(S^{\mathcal{A}_n}_{-\phi_n,\phi_n}\right)^\top$. Note that under this ordering,

$$B^{\mathcal{A}_{n-1}} = \frac{1}{1-\lambda_n}\begin{pmatrix} S^{\mathcal{A}_{n-1}}_{\phi_n} \\ S^{\mathcal{A}_{n-1}}_{-\phi_n,\phi_n} \end{pmatrix} \left[\frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1-\lambda_n} S^{\mathcal{A}_{n-1}}_{\phi_n}\right]^{-1}.$$

Analogously to B_{21} in Section 2, define

$$\begin{aligned} B^{\mathcal{A}_n}_{21} &= S^{\mathcal{A}_n}_{-\phi_n,\phi_n}\left(S^{\mathcal{A}_n}_{\phi_n}\right)^{-1} && (83) \\ B^{\mathcal{A}_{n-1}}_{21} &= S^{\mathcal{A}_{n-1}}_{-\phi_n,\phi_n}\left(S^{\mathcal{A}_{n-1}}_{\phi_n}\right)^{-1}. && (84) \\ & && (85) \end{aligned}$$

Analogously to Σ in Section 2, define

$$\begin{aligned} \Sigma^{\mathcal{A}_n} &= S^{\mathcal{A}_n}_{-\phi_n,-\phi_n} - S^{\mathcal{A}_n}_{-\phi_n,\phi_n}\left(S^{\mathcal{A}_n}_{\phi_n}\right)^{-1} S^{\mathcal{A}_n}_{\phi_n,-\phi_n} && (86) \\ \Sigma^{\mathcal{A}_{n-1}} &= S^{\mathcal{A}_{n-1}}_{-\phi_n,-\phi_n} - S^{\mathcal{A}_{n-1}}_{-\phi_n,\phi_n}\left(S^{\mathcal{A}_{n-1}}_{\phi_n}\right)^{-1} S^{\mathcal{A}_{n-1}}_{\phi_n,-\phi_n}. && (87) \end{aligned}$$

By Lemma A.3, (82) holds if and only if

$$\begin{aligned} \left(S^{\mathcal{A}_n}_{\phi_n}\right)^{-1} &= (1-\lambda_n)\left(S^{\mathcal{A}_{n-1}}_{\phi_n}\right)^{-1} + \lambda_n\left(S_{\phi_n}\right)^{-1} && (88) \\ B^{\mathcal{A}_n}_{21} &= B^{\mathcal{A}_{n-1}}_{21} && (89) \\ \Sigma^{\mathcal{A}_n} &= \frac{1}{1-\lambda_n}\Sigma^{\mathcal{A}_{n-1}}. && (90) \end{aligned}$$

We first show (88). Equation (54) implies

$$S^{\mathcal{A}_n}_{\phi_n} = \frac{1}{1-\lambda_n} S^{\mathcal{A}_{n-1}}_{\phi_n} - \frac{1}{1-\lambda_n} S^{\mathcal{A}_{n-1}}_{\phi_n}\left[\frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1-\lambda_n} S^{\mathcal{A}_{n-1}}_{\phi_n}\right]^{-1} \frac{1}{1-\lambda_n} S^{\mathcal{A}_{n-1}}_{\phi_n}.$$

Pre-multiplying by $(1-\lambda_n)\left(S^{\mathcal{A}_{n-1}}_{\phi_n}\right)^{-1}$ yields

$$\begin{aligned} (1-\lambda_n)\left(S^{\mathcal{A}_{n-1}}_{\phi_n}\right)^{-1} S^{\mathcal{A}_n}_{\phi_n} &= I - \left[\frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1-\lambda_n} S^{\mathcal{A}_{n-1}}_{\phi_n}\right]^{-1} \frac{1}{1-\lambda_n} S^{\mathcal{A}_{n-1}}_{\phi_n} \\ &= \left[\frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1-\lambda_n} S^{\mathcal{A}_{n-1}}_{\phi_n}\right]^{-1} \frac{1}{\lambda_n} S_{\phi_n}. \end{aligned}$$

Taking inverses yields

$$\begin{aligned} \frac{1}{1-\lambda_n}\left(S^{\mathcal{A}_n}_{\phi_n}\right)^{-1} S^{\mathcal{A}_{n-1}}_{\phi_n} &= \lambda_n\left(S_{\phi_n}\right)^{-1}\left[\frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1-\lambda_n} S^{\mathcal{A}_{n-1}}_{\phi_n}\right] \\ &= I + \frac{\lambda_n}{1-\lambda_n}\left(S_{\phi_n}\right)^{-1} S^{\mathcal{A}_{n-1}}_{\phi_n}. \end{aligned}$$

Post-multiplying by $(1-\lambda_n)\left(S_{\phi_n}^{\mathcal{A}_{n-1}}\right)^{-1}$ yields (88).

We now show (89). By (54),

$$S_{-\phi_n,\phi_n}^{\mathcal{A}_n} = \frac{1}{1-\lambda_n} S_{-\phi_n,\phi_n}^{\mathcal{A}_{n-1}} - \frac{1}{1-\lambda_n} S_{-\phi_n,\phi_n}^{\mathcal{A}_{n-1}} \left[\frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1-\lambda_n} S_{\phi_n}^{\mathcal{A}_{n-1}}\right]^{-1} \frac{1}{1-\lambda_n} S_{\phi_n}^{\mathcal{A}_{n-1}}.$$

Post-multiplying by $\left(S_{\phi_n}^{\mathcal{A}_n}\right)^{-1}$ and applying (88) produces

$$\begin{aligned} B_{21}^{\mathcal{A}_n} = S_{-\phi_n,\phi_n}^{\mathcal{A}_n}\left(S_{\phi_n}^{\mathcal{A}_n}\right)^{-1} &= \frac{1}{1-\lambda_n} S_{-\phi_n,\phi_n}^{\mathcal{A}_{n-1}} \left[(1-\lambda_n)\left(S_{\phi_n}^{\mathcal{A}_{n-1}}\right)^{-1} + \lambda_n S_{\phi_n}^{-1}\right] \\ &- \frac{1}{1-\lambda_n} S_{-\phi_n,\phi_n}^{\mathcal{A}_{n-1}} \left[\frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1-\lambda_n} S_{\phi_n}^{\mathcal{A}_{n-1}}\right]^{-1} \frac{1}{1-\lambda_n} S_{\phi_n}^{\mathcal{A}_{n-1}} \left[(1-\lambda_n)\left(S_{\phi_n}^{\mathcal{A}_{n-1}}\right)^{-1} + \lambda_n S_{\phi_n}^{-1}\right]. \end{aligned}$$

Expanding out the first term on the right hand side and multiplying through by $S_{\phi_n}^{\mathcal{A}_{n-1}}$ in the second term produces

$$\begin{aligned} B_{21}^{\mathcal{A}_n} = B_{21}^{\mathcal{A}_{n-1}} &+ \frac{\lambda_n}{1-\lambda_n} S_{-\phi_n,\phi_n}^{\mathcal{A}_{n-1}} S_{\phi_n}^{-1} \\ &- \frac{1}{1-\lambda_n} S_{-\phi_n,\phi_n}^{\mathcal{A}_{n-1}} \left[\frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1-\lambda_n} S_{\phi_n}^{\mathcal{A}_{n-1}}\right]^{-1} \left[I + \frac{\lambda_n}{1-\lambda_n} S_{\phi_n}^{\mathcal{A}_{n-1}} S_{\phi_n}^{-1}\right]. \end{aligned}$$

Factoring out $\lambda_n S_{\phi_n}^{-1}$ in the last term yields (89).

Lastly, we show (90). Equation (54) implies

$$S_{-\phi_n,-\phi_n}^{\mathcal{A}_n} = \frac{1}{1-\lambda_n} S_{-\phi_n,-\phi_n}^{\mathcal{A}_{n-1}} - \frac{1}{1-\lambda_n} S_{-\phi_n,\phi_n}^{\mathcal{A}_{n-1}} \left[\frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1-\lambda_n} S_{\phi_n}^{\mathcal{A}_{n-1}}\right]^{-1} \frac{1}{1-\lambda_n} S_{\phi_n,-\phi_n}^{\mathcal{A}_{n-1}}. \tag{91}$$

By (89),

$$\begin{aligned} S_{-\phi_n,\phi_n}^{\mathcal{A}_n}\left(S_{\phi_n}^{\mathcal{A}_n}\right)^{-1} S_{\phi_n,-\phi_n}^{\mathcal{A}_n} &= B_{21}^{\mathcal{A}_n} S_{\phi_n}^{\mathcal{A}_n} \left(B_{21}^{\mathcal{A}_n}\right)^\top \\ &= B_{21}^{\mathcal{A}_{n-1}} S_{\phi_n}^{\mathcal{A}_n} \left(B_{21}^{\mathcal{A}_{n-1}}\right)^\top. \end{aligned}$$

Therefore,

$$\begin{aligned} S_{-\phi_n,\phi_n}^{\mathcal{A}_n}\left(S_{\phi_n}^{\mathcal{A}_n}\right)^{-1} S_{\phi_n,-\phi_n}^{\mathcal{A}_n} &= S_{-\phi_n,\phi_n}^{\mathcal{A}_{n-1}}\left(S_{\phi_n}^{\mathcal{A}_{n-1}}\right)^{-1} \left[(1-\lambda_n)\left(S_{\phi_n}^{\mathcal{A}_{n-1}}\right)^{-1} + \lambda_n S_{\phi_n}^{-1}\right]^{-1} \left(S_{\phi_n}^{\mathcal{A}_{n-1}}\right)^{-1} S_{\phi_n,-\phi_n}^{\mathcal{A}_{n-1}} \\ &= S_{-\phi_n,\phi_n}^{\mathcal{A}_{n-1}} \left[(1-\lambda_n) S_{\phi_n} + \lambda_n S_{\phi_n}^{\mathcal{A}_{n-1}}\right]^{-1} S_{\phi_n} \left(S_{\phi_n}^{\mathcal{A}_{n-1}}\right)^{-1} S_{\phi_n,-\phi_n}^{\mathcal{A}_{n-1}}. \end{aligned} \tag{92}$$

Substituting in (91) and (92) into (86),

$$\begin{aligned} \Sigma^{\mathcal{A}_n} = \frac{1}{1-\lambda_n} S_{-\phi_n,-\phi_n}^{\mathcal{A}_{n-1}} - \\ \frac{1}{1-\lambda_n} S_{-\phi_n,\phi_n}^{\mathcal{A}_{n-1}} \left[\frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1-\lambda_n} S_{\phi_n}^{\mathcal{A}_{n-1}}\right]^{-1} \left[\frac{1}{1-\lambda_n} I + \frac{1}{\lambda_n} S_{\phi_n} \left(S_{\phi_n}^{\mathcal{A}_{n-1}}\right)^{-1}\right] S_{\phi_n,-\phi_n}^{\mathcal{A}_{n-1}}. \end{aligned}$$

Which implies

$$\begin{aligned}
\Sigma^{\mathcal{A}_n} &= \frac{1}{1-\lambda_n} S^{\mathcal{A}_{n-1}}_{-\phi_n,-\phi_n} - \\
&\quad \frac{1}{1-\lambda_n} S^{\mathcal{A}_{n-1}}_{-\phi_n,\phi_n} \left[\frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1-\lambda_n} S^{\mathcal{A}_{n-1}}_{\phi_n}\right]^{-1} \left[\frac{1}{1-\lambda_n} S^{\mathcal{A}_{n-1}}_{\phi_n} + \frac{1}{\lambda_n} S_{\phi_n}\right] \left(S^{\mathcal{A}_{n-1}}_{\phi_n}\right)^{-1} S^{\mathcal{A}_{n-1}}_{\phi_n,-\phi_n} \\
&= \frac{1}{1-\lambda_n} S^{\mathcal{A}_{n-1}}_{-\phi_n,-\phi_n} - \frac{1}{1-\lambda_n} S^{\mathcal{A}_{n-1}}_{-\phi_n,\phi_n} \left(S^{\mathcal{A}_{n-1}}_{\phi_n}\right)^{-1} S^{\mathcal{A}_{n-1}}_{\phi_n,-\phi_n} \\
&= \frac{1}{1-\lambda_n} \Sigma^{\mathcal{A}_{n-1}},
\end{aligned}$$

which shows (90). ▮

Proof that $B^{\mathcal{A}_2}$ in Section 5.3.3 is equal to (73):

From Theorem 5.1 and (71), it follows that

$$\begin{aligned}
&\lim_{T\to\infty} E\left[T\left(h^{\mathcal{A}_2}_{2,(\lambda_1+\lambda_2)T} - g_{2,\lambda_3 T}\right)\left(h^{\mathcal{A}_2}_{2,(\lambda_1+\lambda_2)T} - g_{2,\lambda_3 T}\right)^{\top}\right] \\
&\quad = \left(\frac{1}{\lambda_3} + \frac{1}{\lambda_1}\right) S_{22} + \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^2 \left(\frac{1}{\lambda_2} + \frac{1}{\lambda_1}\right) B_{21} S_{11} B_{21}^{\top} - \frac{1}{\lambda_1}\frac{\lambda_2}{\lambda_1+\lambda_2}\left(S_{21}B_{21}^{\top} + B_{21}S_{12}\right) \\
&\quad = \left(\frac{1}{\lambda_3} + \frac{1}{\lambda_1}\right) S_{22} - \frac{1}{\lambda_1}\frac{\lambda_2}{\lambda_1+\lambda_2} S_{21} S_{11}^{-1} S_{12}, \qquad (93)
\end{aligned}$$

and

$$\begin{aligned}
\lim_{T\to\infty} E\left[T h^{\mathcal{A}_2}_{1,(\lambda_1+\lambda_2)T}\left(h^{\mathcal{A}_2}_{2,(\lambda_1+\lambda_2)T} - g_{2,\lambda_3 T}\right)^{\top}\right] &= \lim_{T\to\infty} E\left[T g_{1,(\lambda_1+\lambda_2)T}\left(h^{\mathcal{A}_2}_{2,(\lambda_1+\lambda_2)T}\right)^{\top}\right] \\
&= \frac{1}{\lambda_1+\lambda_2} S_{12}. \qquad (94)
\end{aligned}$$

where we have applied the reasoning of (22). (94) multiplied by the inverse of (93) equals the first component of $B^{\mathcal{A}_2}$. The second component equals

$$\begin{aligned}
\lim_{T\to\infty} E\left[T h^{\mathcal{A}_2}_{2,(\lambda_1+\lambda_2)T}\left(h^{\mathcal{A}_2}_{2,(\lambda_1+\lambda_2)T} - g_{2,\lambda_3 T}\right)^{\top}\right] &= \lim_{T\to\infty} E\left[T h^{\mathcal{A}_2}_{2,(\lambda_1+\lambda_2)}\left(h^{\mathcal{A}_2}_{2,(\lambda_1+\lambda_2)T}\right)^{\top}\right] \\
&= \frac{1}{\lambda_1}\left(S_{22} - \frac{\lambda_2}{\lambda_1+\lambda_2} S_{21} S_{11}^{-1} S_{12}\right), \qquad (95)
\end{aligned}$$

multiplied by the inverse of (93). The resulting expression for $B^{\mathcal{A}_2}$ can be simplified considerably. The inverse of (93) equals

$$\begin{aligned}
\left(\left(\frac{1}{\lambda_3} + \frac{1}{\lambda_1}\right) S_{22} - \frac{1}{\lambda_1}\frac{\lambda_2}{\lambda_1+\lambda_2} S_{21} S_{11}^{-1} S_{12}\right)^{-1} &= S_{22}^{-1}\left(I - \gamma S_{21} S_{11}^{-1} S_{12} S_{22}^{-1}\right)^{-1} \frac{\lambda_1\lambda_3}{\lambda_1+\lambda_3} \\
&= S_{22}^{-1}\left(I - \gamma B_{21} B_{12}\right)^{-1} \frac{\lambda_1\lambda_3}{\lambda_1+\lambda_3}, \qquad (96)
\end{aligned}$$

where $B_{12} = S_{12}S_{22}^{-1}$. Therefore,

$$B^{\mathcal{A}_2} = \frac{\lambda_3}{\lambda_1 + \lambda_3} \begin{pmatrix} \frac{\lambda_1}{\lambda_1+\lambda_2} S_{12} \\ S_{22} - \frac{\lambda_2}{\lambda_1+\lambda_2} S_{21} S_{11}^{-1} S_{12} \end{pmatrix} S_{22}^{-1} \left(I - \gamma B_{21} B_{12}\right)^{-1}.$$

∎

References

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

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

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

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.

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

Lynch, Anthony W., Jessica A. Wachter, and Walter Boudry, 2004, Does mutual fund performance vary over the business cycle?, Working Paper, New York University and the University of Pennsylvania.

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.

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

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

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.