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

Anthony W. Lynch
New York University and NBER

Jessica A. Wachter
University of Pennsylvania and NBER

October 7, 2005

*Lynch: Department of Finance, Stern School of Business, New York University, 44 West Fourth Street, New York, NY 10012-1126; Email: alynch@stern.nyu.edu; Tel: (212) 998-0350. Wachter: Department of Finance, The Wharton School, University of Pennsylvania, 3620 Locust Walk, Philadelphia, PA 19104; Tel: (215) 898-7634; Email: jwachter@wharton.upenn.edu. We thank Yacine Ait-Sahalia, David Chapman, Robert Engle, Martin Lettau, Andrew Lo, Kenneth Singleton, Robert Stambaugh, Amir Yaron, and Motohiro Yogo as well as seminar participants at the 2005 AFA meetings, at New York University, at the Wharton School, and at the University of Pennsylvania Department of Economics for their comments and suggestions. The most recent version of this paper can be found at http://finance.wharton.upenn.edu/~jwachter/research/gmm.pdf.
Using Samples of Unequal Length in Generalized Method of Moments Estimation

Abstract

Many applications in financial economics use data series with different starting or ending dates. This paper describes estimation methods, based on the generalized method of moments (GMM), which make use of all available data for each moment condition. We introduce two asymptotically equivalent estimators that are consistent, asymptotically normal, and more efficient asymptotically than standard GMM. We apply these methods to estimating factor models and predictability regressions in international data and show that the use of the full sample impacts point estimates and standard errors for both the long and the short-history assets. A Monte Carlo experiment demonstrates that reductions hold for small-sample standard errors as well as asymptotic ones. These methods are extended to more general patterns of missing data, and shown to be more efficient than estimators that ignore intervals of the data, and thus more efficient than standard GMM.
Introduction

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

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

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

Following the work of Anderson, the approach in these previous studies is likelihood-based. In


\(^2\)Pastor and Stambaugh (2002a, 2002b) derive Bayesian posteriors for means and variances of mutual fund returns using samples of unequal length, under the assumption of normality and identically and independently distributed returns. Storesletten, Telmer, and Yaron (2004) combine a time series of macro-economic variables dating back to 1930 with the shorter Panel Study of Income Dynamics to estimate the relationship between cross-sectional variance and recessions.
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.\textsuperscript{3} As shown in Cochrane (2001), many common estimation techniques used in finance can be seen as special cases of GMM. Assumptions required for the consistency and asymptotic normality of the standard GMM estimator are also required here. We adopt the mixing assumption of White and Domowitz (1984) as a means of limiting the temporal dependence of the underlying stochastic process. Intuitively, mixing requires that autocovariances vanish as the lag length increases. This assumption allows for many processes of interest in financial economics, such as finite ARMA processes with general conditions on the underlying errors (see Phillips (1987)).\textsuperscript{4}

Because our method is based on GMM, the many of our results are asymptotic.\textsuperscript{5} So that the asymptotic approximation is reasonable, care must be taken to insure that the missing data problem does not become trivial as the sample size becomes large. We thus develop an asymptotic theory that keeps the fraction of missing data fixed as the sample size approaches infinity. To be precise, if $T$ denotes the length of the longer sample, we say that $\lambda T$ is the length of the shorter sample, for $0 < \lambda \leq 1$. We hold $\lambda$ constant, as $T$ approaches infinity. This approach has a parallel in the simulated method of moments estimation technique (see Duffie and Singleton (1993)), where the length of the simulated series divided by the length of the observed series is assumed to be constant as both series lengths approach infinity, as well as in the literature on structural breaks (see Stock (1994)).

Our initial setting supposes that some moment conditions are observed over the full data set while others 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 or 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 we show in an application involving international data (Section 5), and have natural

\textsuperscript{3}Another strand of literature considers the problem of $n$ independent individuals observed at up to $T$ time periods, where some individuals drop out of the study (see, e.g., Robins and Rotnitsky (1995)). The independence across individuals and the fact that asymptotics are derived as $n$, rather than $T$, approaches infinity differentiates this problem from the one considered here.

\textsuperscript{4}Like many of the studies mentioned above, we do assume that the data is missing at random, in the sense defined by Little and Rubin (2002). Stambaugh (1997) discusses cases where this assumption holds in financial time series, such as when the start date depends only on the long-history asset returns, and cases where it does not, such as when the decision to add a country to a list of emerging markets depends on past unobserved returns on that country (see Goetzmann and Jorion (1999)).

\textsuperscript{5}We also verify, in Monte Carlo experiments, that our methods deliver efficiency gains in small samples.
and intuitive interpretations.

The first estimator (which we call the *adjusted-moment* estimator) uses full sample averages to estimate the moments for which full-sample data are available, and short sample averages to estimate moments for which only short-sample data are available. Then the moments for which only the short sample is available are “adjusted” using coefficients from a regression of the short-sample moments on the full-sample moments. This is reminiscent of an adjustment that appears in Stambaugh (1997) and Little and Rubin (2002) but here operates in a more general context. The second estimator, (which we call the *over-identified* estimator) uses the extra data available from the full sample as a new set of moment conditions. This estimator was suggested by Stambaugh (1997) and, in the linear context of that paper, turns out to be identical to our adjusted-moment estimator (and the maximum-likelihood estimator proposed in that paper). In the more general context of our paper, the two estimators are equivalent asymptotically but typically differ in finite samples. Singleton (2004) proposes an extension of the estimator in Stambaugh (1997) that is similar to our over-identified estimator, but does not show asymptotic normality or efficiency over standard GMM. We build on this work by showing that both of our estimators are consistent and asymptotically normal, and that both are asymptotically more efficient than standard GMM.

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 estimators 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

---

6 Under general assumptions it is necessary that the “blocks” of missing data be large relative to the sample size. This distinguishes the problem we tackle from the problem posed by data sampled at different frequencies (see Ghysels, Santa-Clara, and Valkanov (2005)).

7 In its focus on the efficiency results of carefully including additional data, this study has parallels in studies that focus on including high-frequency data in estimation while accounting for market microstructure effects (see Ait-Sahalia, Mykland, and Zhang (2005), Bandi and Russell (2004))
generalized estimators are also more efficient than standard GMM.

The organization of the paper is as follows. Section 1 develops asymptotic theory that forms basis for the consistency and asymptotic normality proofs. The key result in this section is that sample averages (scaled appropriately) taken over disjoint intervals of data are independent as the number of data points in each interval approaches infinity. Section 2 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.

Section 3 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. Section 4 investigates a common 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.

Section 5 illustrates our methods in applications to value and growth portfolios in international data. While US data on portfolios sorted on book-to-market ratios go back to 1927, international data on these portfolios typically begin in the mid-1970s to early 1980s. For this reason, studies that examine properties of value and growth portfolios in international data typically restrict the sample to this period (e.g. Fama and French (1998)). Our methods allow the simultaneous use of the full US sample and the international sample. We apply these methods to estimating factor models and predictability regressions and show that the use of the full US sample leads to more precise estimates for both US and international returns. A Monte Carlo experiment demonstrates that reductions hold for small-sample standard errors as well as asymptotic ones. Finally, Section 6 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.
1 Stochastic setting

This section introduces the stochastic setting for the paper and states assumptions and results necessary for establishing properties of our estimators. The assumptions are standard (see, e.g. White and Domowitz (1984)). The main new result is that partial sums taken over disjoint intervals are asymptotically independent and normally distributed.

Let \( f_{xt} \) denote a \( p \)-component stochastic process defined over an underlying probability space \((\Omega, \mathcal{F}, P)\). Let \( \mathcal{F}_a^b \equiv \sigma(x_t; a \leq t \leq b) \), the Borel \( \sigma \)-algebra of events generated by \( x_a, \ldots, x_b \). Consider a function \( f : \mathbb{R}^p \times \Theta \rightarrow \mathbb{R}^l \) for \( \Theta \), a compact subset of \( \mathbb{R}^q \). The function \( f \) provides the restrictions that determine \( \mu \) based on the observations of \( x_t \). In what follows we make standard assumptions on \( f_{xt} \) 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

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

for \( \sigma \)-algebras \( \mathcal{F} \) and \( \mathcal{G} \), and

\[
\alpha(s) \equiv \sup_{t} \alpha(\mathcal{F}_t^t, \mathcal{F}_t^{t+s}).
\]

The process \( \{x_t\} \) is said to be \( \alpha \)-mixing if \( \alpha(s) \to 0 \) as \( s \to \infty \). As White and Domowitz (1984) discuss, \( \alpha \)-mixing guarantees that autocovariances vanish at arbitrarily long lags. Mixing is a convenient assumption because it allows a trade-off between the speed at which \( \alpha(s) \) approaches zero and the conditions required on \( \{x_t\} \). In particular, a process is said to be \( \alpha \)-mixing of size \( r/(r - 1) \) for \( r > 1 \) if for some \( \kappa > r/(r - 1) \), \( \alpha(s) \) is \( O(s^{-\kappa}) \). We assume that \( \{x_t\} \) is mixing:

**Assumption 1** \( \{x_t\}_{t=-\infty}^{\infty} \) is \( \alpha \)-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 \( \theta \) given observations on \( x_t \).

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

The next assumptions form the basis for the consistency and asymptotic normality results of estimators based on partial sums of \( f(x_t, \theta) \).
Assumption 4 There exists $\Delta \in \mathbb{R}$ such that $E \left( \left| f_i(x_t, \theta_0) \right|^{2r} \right) < \Delta$, $i = 1, \ldots, l$.

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

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

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

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

This sum converges because $\alpha$-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

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

The proof of the main result in this section relies on our first proving the following lemma, which establishes that partial sums taken over disjoint intervals are asymptotically independent. This new lemma is thus stronger than showing that the partial sums have zero covariance asymptotically. The proof is contained in Appendix A.

**Lemma 1.1** Let $F \in \mathcal{F}_{-\infty}^0$. Let $\mu$ be a $1 \times l$ vector, and let $c$ be a scalar. Let

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

Then Assumptions 1–3 and 5 imply that

$$\lim_{T \to \infty} P \left( \left( \sqrt{T} g_T(\theta_0) < c \right) F \right) = P_g P(F).$$
Let \( \lambda \) be a rational number strictly 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 \( \lambda T \) and \( (1 - \lambda)T \) for \( T \) a multiple of \( n_0 \). For \( T \) a multiple of \( n_0 \), define

\[
g_{(1-\lambda)T}(\theta) = \frac{1}{(1-\lambda)T} \sum_{t=1}^{(1-\lambda)T} f(x_t, \theta) \\
g_{\lambda T}(\theta) = \frac{1}{\lambda T} \sum_{t=(1-\lambda)T+1}^{T} f(x_t, \theta). \tag{2} \tag{3}
\]

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

The following theorem states the main result of this section: that partial sums taken over disjoint intervals are asymptotically independent and normally distributed. The proof uses the result in Lemma 1.1, and is contained in Appendix A. For the remainder of the paper, we let \( T \) approach infinity along the subsequence of integer multiples of \( n_0 \).

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

\[
\sqrt{T} \left[ \sqrt{(1-\lambda)}g_{(1-\lambda)T}(\theta_0) \right] \rightarrow_d N \left( 0, \begin{bmatrix} S & 0 \\ 0 & S \end{bmatrix} \right). \tag{4}
\]

# 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 Section 5 for an application to international data). 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 $\lambda$ of the sample, namely dates $(1 - \lambda)T + 1, \ldots, T$. In this section, we introduce estimators that are consistent and asymptotically normal in this setting.

Without loss of generality, partition the elements of $x_t$ so that $x_t = [x_{1t}^\top \ x_{2t}^\top]^\top$, where data on $x_{1t} \in \mathbb{R}^{p_1}$ is assumed to be available for the full sample, and data on $x_{2t} \in \mathbb{R}^{p_2}$ is assumed to be available for the last $\lambda T$ dates of the sample. Similarly, partition the elements of $f$ into those that depend only on $x_{1t}$ and those that depend on both $x_{1t}$ and $x_{2t}$: $f(x_t, \theta) = [f_1(x_{1t}, \theta)\ T \ f_2(x_t, \theta)]^\top$, where $f_1 : \mathbb{R}^{p_1} \times \Theta \to \mathbb{R}^{l_1}$, and $f_2 : \mathbb{R}^p \times \Theta \to \mathbb{R}^{l_2}$. Analogously, let

\[
g_{1T}(\theta) = \frac{1}{T} \sum_{t=1}^{T} f_1(x_t, \theta),
\]

\[
g_{1,(1-\lambda)T}(\theta) = \frac{1}{(1-\lambda)T} \sum_{t=1}^{(1-\lambda)T} f_1(x_t, \theta),
\]

\[
g_{1,\lambda T}(\theta) = \frac{1}{\lambda T} \sum_{t=(1-\lambda)T+1}^{T} f_1(x_t, \theta),
\]

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 $S$ that correspond to the partitions of $f$ and $g$. 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 $\Sigma$, where

\[
\Sigma = S_{22} - S_{21} S_{11}^{-1} S_{12}.
\]
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

\[
\begin{align*}
\hat{h}^S_T(\theta) &= \begin{bmatrix} g_{1,\lambda T}(\theta)^\top & g_{2,\lambda T}(\theta)^\top \end{bmatrix}^\top \\
\hat{h}^L_T(\theta) &= \begin{bmatrix} g_{1,\lambda T}(\theta)^\top & g_{2,\lambda T}(\theta)^\top \end{bmatrix}^\top \\
\hat{h}^A_T(\theta) &= \begin{bmatrix} g_{1,\lambda T}(\theta)^\top & \left( g_{1,\lambda T}(\theta) - \hat{g}_{21,\lambda T}(1 - \lambda)(g_{1,(1-\lambda)T}(\theta) - g_{1,\lambda T}(\theta)) \right)^\top \end{bmatrix}^\top \\
\hat{h}^I_T(\theta) &= \begin{bmatrix} g_{1,(1-\lambda)T}(\theta)^\top & g_{1,\lambda T}(\theta)^\top & g_{2,\lambda T}(\theta)^\top \end{bmatrix}^\top,
\end{align*}
\]

where \(\hat{B}_{21,\lambda T}\) is an \(l_2 \times l_1\) matrix such that \(\hat{B}_{21,\lambda T} \rightarrow_{a.s.} B_{21}\). For \(k \in S, L, A, I\), given a weighting matrix \(W^k_T\), let

\[
\hat{\theta}^k = \arg\min_{\theta} h^k_T(\theta)^\top W^k_T h^k_T(\theta),
\]

Equation (11) 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 \(\hat{\theta}^A_T\) the “adjusted moment” estimator, because it involves adjusting the second set of moments. We refer to \(\hat{\theta}^I_T\) as the “over-identified” estimator, because it involves adding an additional moment condition. As we will show, \(\hat{\theta}^I_T\) has the same asymptotic properties as \(\hat{\theta}^A_T\).

**Theorem 2.1** Assumptions 1–5 imply that as \(T \rightarrow \infty\),

\[
\sqrt{\lambda T} h^k_T(\theta_0) \rightarrow_d N(0, S^k),
\]
where

\[ S^S = S \]  

(12)

\[ S^C = \begin{bmatrix} \lambda S_{11} & \lambda S_{12} \\ \lambda S_{21} & S_{22} \end{bmatrix} \]  

(13)

\[ S^A = \begin{bmatrix} S_1^A & S_2^A \\ S_2^A & S_2^A \end{bmatrix} = \begin{bmatrix} \lambda S_{11} & \lambda S_{12} \\ \lambda S_{21} & S_{22} - (1 - \lambda)S_{21}S_{11}^{-1}S_{12} \end{bmatrix} \]  

(14)

\[ S^T = \begin{bmatrix} \frac{\lambda}{1 - \lambda} S_{11} & 0 & 0 \\ 0 & S_{11} & S_{12} \\ 0 & S_{21} & S_{22} \end{bmatrix}. \]  

(15)

**Proof:** Equation (12) follows from Theorem 1.1. We show (14); the proofs of (13) and (15) are similar. In what follows, the argument \( \theta_0 \) is suppressed and convergence is in the sense of almost surely.

Stationarity implies that \( S_{11}^A = \lambda S_{11} \). By Theorem 1.1,

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

(16)

for \( i, j = 1, 2 \). Therefore,

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

The second line follows from (16) and the third and fourth lines follow from Theorem 1.1. Using similar reasoning,

\[ S_{22}^A = \lim_{T \to \infty} E \left[ \sqrt{T} g_{2,\lambda T} \sqrt{T} g_{2,\lambda T}^\top \right] - 2 \lim_{T \to \infty} (1 - \lambda)E \left[ \sqrt{T} g_{2,\lambda T} \sqrt{T} g_{1,\lambda T}^\top \right] B_{21}^\top \\
+ \lim_{T \to \infty} B_{21}(1 - \lambda)^2 E \left[ \sqrt{T}(g_{1,(1-\lambda)T} - g_{1,\lambda T}) \sqrt{T}(g_{1,(1-\lambda)T} - g_{1,\lambda T})^\top \right] B_{21}^\top \\
= S_{22} - 2(1 - \lambda)S_{21}S_{11}^{-1}S_{12} + (1 - \lambda)^2 \left( \frac{\lambda}{1 - \lambda} + 1 \right) S_{21}S_{11}^{-1}S_{12} \\
= S_{22} - (1 - \lambda)S_{21}S_{11}^{-1}S_{12}, \]
which completes the derivation of (14). □

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

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

Theorem 2.2 establishes consistency of the estimators. The proof is standard and given in Appendix B.

**Theorem 2.2** Assumptions 1–6 imply that as \( T \to \infty \), \( \hat{\theta}_T^k \to_{a.s.} \theta_0 \) for \( k \in \{S, L, A, T\} \).

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

**Assumption 7** \( \theta_0 \) lies in the interior of \( \Theta \).

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

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

Define \( D_{0,i} = E \left[ \left. \left( \frac{\partial f_i}{\partial \theta} \right) \right|_{\theta_0} \right] \) and \( D_0 = [D_{0,1}^T, D_{0,2}^T] \). Let

\[
D_0^k = D_0 \quad k \in \{S, L, A\} \\
D_0^T = \begin{bmatrix} D_{0,1}^T & D_{0,1}^T & D_{0,2}^T \end{bmatrix}^T.
\]  

(17)  

(18)

The following theorem establishes asymptotic normality. The proof is given in Appendix B.

**Theorem 2.3** Assumptions 1–9 imply

\[
\sqrt{\lambda_T} (\hat{\theta}_T^k - \theta_0) \to_d N \left( 0, \left( \begin{bmatrix} D_0^k \end{bmatrix}^T W^k D_0^k \right)^{-1} \left( \begin{bmatrix} D_0^k \end{bmatrix}^T W^k S^k W^k D_0^k \right) \left( \begin{bmatrix} D_0^k \end{bmatrix}^T W^k D_0^k \right)^{-1} \right).
\]  

(19)

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

\[
\sqrt{\lambda_T} (\hat{\theta}_T^k - \theta_0) \to_d N \left( 0, \left( \begin{bmatrix} D_0^k \end{bmatrix}^T \left( S^k \right)^{-1} (D_0^k) \right)^{-1} \right).
\]  

(19)

Moreover, this choice of \( W^k \) is efficient for each estimator.
3 Comparison

The previous section defined four estimators: the “short” estimator, which corresponds to standard GMM; the “long” estimator, which makes use of the full sample in a naive way; and the adjusted-moment and over-identified estimators, which make use of the full sample in a slightly more complicated way than “long”. These four estimators were shown to be consistent and asymptotically normal. This section compares the asymptotic efficiency of these estimators.

The first result is that the over-identified estimator and the adjusted-moment estimator have identical asymptotic properties when the optional weighting matrix is used.

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

Proof: It suffices to compare the asymptotic variances as the mean of both asymptotic distributions is \( \sqrt{T} \hat{\theta}_T^0 \). In the case of the over-identified estimator, the inverse of the asymptotic variance of \( \sqrt{T} \hat{\theta}_T^I \) equals

\[
(D^T)^\top (S^T)^{-1} D^T = \frac{1 - \lambda}{\lambda} D_{0,1}^T S_{11}^{-1} D_{0,1} + D_0^T \begin{bmatrix} B_{21} \Sigma^{-1} B_{21} & -B_{21} \Sigma^{-1} \\ -\Sigma^{-1} B_{21} & \Sigma^{-1} \end{bmatrix} D_0,
\]

where \( \Sigma \) is defined by (5). This follows from Theorem 2.4 and Lemma C.3.

It follows from the distribution of the adjusted moment estimator (19) and Lemma C.3 that the inverse of the variance of \( \sqrt{T} \hat{\theta}_T^A \) equals

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

which equals (20). Thus the estimators are asymptotically equivalent. \( \Box \)

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 \( W^k \to_{a.s.} (S^k)^{-1} \), for \( k \in S, A, I \), the estimators \( \hat{\theta}_T^A \) and \( \hat{\theta}_T^S \) are asymptotically more efficient than \( \hat{\theta}_T^S \).

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

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

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

For any \( l \times 1 \) vector \( v = [v_1^T, v_2^T]^T \),

\[
v^T(S - S^A)v = (1 - \lambda) \left( v_1^T S_{11} v_1 + v_1^T S_{12} v_2 + v_2^T S_{21} v_1 + v_2^T S_{21}S_{11}^{-1}S_{12} v_2 \right)
\]

\[
= (1 - \lambda) \left( v_1^T S_{11}^{-1}S_{11} v_1 + v_1^T S_{11}S_{11}^{-1}S_{12} v_2 + v_2^T S_{12}S_{11}^{-1}S_{11} v_1 + v_2^T S_{12}S_{11}^{-1}S_{12} v_2 \right)
\]

\[
= (1 - \lambda) (S_{11} v_1 + S_{12} v_2)^T S_{11}^{-1} (S_{11} v_1 + S_{12} v_2) \geq 0
\]

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

To prove statement (2), define

\[
U = W^A D_0^A \left( (D_0^A)^T W^A D_0^A \right)^{-1}.
\]

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

\[
U = W^S D_0^S \left( (D_0^S)^T W^S D_0^S \right)^{-1}.
\]

By Theorem 2.3, proving (2) is equivalent to showing \( U^T S^A U - U^T S^S U \) is positive semi-definite. But for any vector \( v \),

\[
v^T(U^T S^A U - U^T S^S U)v = (Uv)^T (S - S^A) Uv > 0
\]

because \( S - S^A \) is positive semi-definite. Therefore \( \hat{\theta}_T^A \) is more efficient when \( W^S = W^A \).

Note that statement (2) of Theorem 3.2 does not make sense for the over-identified estimator \( \hat{\theta}_T^I \) because \( \hat{\theta}_T^I \) has \( l_1 \) more moment conditions than \( \hat{\theta}_T^A \) and \( \hat{\theta}_T^S \). 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 $\hat{\theta}_T^A$ with the estimator $\hat{\theta}_T^F$.

**Theorem 3.3** Assume 1–5 and 7–9 then

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

2. If $W_{XT}^k \to_{a.s.} W^k$ for $W^k$ positive definite, and such that $W^A = W^\mathcal{L}$ almost surely, $\hat{\theta}_T^A$ is more efficient than $\hat{\theta}_T^F$.

**Proof:** As in Theorem 3.2, it suffices to show that

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

is positive semi-definite. Note that for any vector $v = [v_1^\top, v_2^\top]^\top$,

$$v^\top(S^\mathcal{L} - S^A)v = (1 - \lambda)(S_{12}v_2)^\top S_{11}^{-1}S_{12}v_2 \geq 0$$

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

To show the second statement, define $U$ analogously to (21):

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

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

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

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

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

$$S^\mathcal{L} - S = (1 - \lambda) \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & 0 \end{bmatrix}$$
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 \( W_T^T = (S^T)^{-1}, W_T^A = (S^A)^{-1} \), and \( \hat{B}_{21,AT} = B_{21} \).

Differentiating the over-identified estimator with respect to \( \theta \) yields

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

Equation (22) is the first-order condition that determines the over-identified estimator \( \hat{\theta}_T^T \). By contrast, the first order condition associated with the adjusted-moment estimator is

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

which reduces to

\[
0 = \frac{1}{\lambda} g_{1,T}^T S_{11}^{-1} \frac{\partial g_{1,T}}{\partial \theta} + (B_{21} g_{1,T} - h_{2,T})^T \Sigma^{-1} \frac{\partial}{\partial \theta} (B_{21} g_{1,T} - h_{2,T}) + (B_{21} g_{1,AT} - g_{2,AT})^T \Sigma^{-1} \frac{\partial}{\partial \theta} (B_{21} g_{1,AT} - g_{2,AT}).
\]

Equation (23) is the first-order condition that determines the adjusted-moment estimator \( \hat{\theta}_T^A \). According to Theorem 3.1, these two first order conditions must be equivalent as \( T \to \infty \). Indeed they are, because

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

and

\[
\frac{1 - \lambda}{\lambda} g_{1,(1-\lambda)T}^T S_{11}^{-1} D_{0,1} + g_{1,\lambda T}^T S_{11}^{-1} D_{0,1} = \frac{1}{\lambda} \left( (1 - \lambda) g_{1,(1-\lambda)T} + \lambda g_{1,\lambda T} \right) S_{11}^{-1} D_{0,1} = \frac{1}{\lambda} g_{1,T}^T S_{11}^{-1} D_0.
\]
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 Quantifying the efficiency gains

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 \( \theta = [\theta_1^T \theta_2^T]^T \), where \( f_1 \) is a function of \( \theta_1 \) alone. 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.

Let \( l_1 \) be the length of \( \theta_1 \), and \( l_2 = q - l_1 \) the length of \( \theta_2 \). For convenience assume that \( W_T = (S^T)^{-1} \) 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 \( \Sigma^{-1} \frac{\partial}{\partial \theta} (B_{21}g_{1,\lambda T} - g_{2,\lambda T}) \) is invertible, the first order conditions for the over-identified estimator \( \hat{\theta}_{T} \) reduce to
\[ g_{2,\lambda T} - B_{21}g_{1,\lambda T} = 0 \] (24)
for \( \theta_2 \), and thus
\[ \frac{1 - \lambda}{\lambda} g_{1,(1-\lambda)T} S_{11}^{-1} \frac{\partial g_{1,(1-\lambda)T}}{\partial \theta_1} + g_{1,\lambda T} S_{11}^{-1} \frac{\partial g_{1,\lambda T}}{\partial \theta_1} = 0 \] (25)
for \( \theta_1 \). In comparison, the first order conditions for the adjusted-moment estimator \( \hat{\theta}_T^A \) reduce to (24) for \( \theta_2 \) but to \( g_{1,T} = 0 \) for \( \theta_1 \). 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 \( h_{2,T}^A \) equal to zero. When \( g_{1,T} = 0 \), \( h_{2,T}^A \) is equivalent to the left-hand side of (24).

We have shown that in the case considered here, the adjusted-moment estimator gives the same estimate for \( \theta_1 \) as simply using the full sample. The over-identified estimator gives a possibly different estimate, one that depends on the point in time in which the second series begins. While this dependence is possibly unattractive, (25) 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 \( \theta \). In the special case where the system is exactly identified and \( f_1 \) depends on \( \theta_1 \), the derivative matrix \( D_0 \) is invertible and takes the form

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

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 \( \theta_1 \), \( d_{21} \) gives the derivatives of \( f_2 \) with respect to \( \theta_1 \), and \( d_{22} \) gives the derivatives of \( f_2 \) with respect to \( \theta_2 \).

The inverse of \( D_0 \) takes the form

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

Therefore the first diagonal block of \( (D_0^\top S^{-1}D_0)^{-1} \) equals \( d_{11}^{-1}S_{11}(d_{11}^{-1})^\top \). Similarly, the first block of \( (D_0^\top (S^A)^{-1}D_0)^{-1} \) can be written as\(^9\)

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

This shows that asymptotic standard errors for the estimates of \( \theta_1 \) 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 \( \theta_2 \). The second diagonal block of \( (D_0^\top (S^A)^{-1}D_0)^{-1} \) reduces to

\[
(D_0^\top (S^A)^{-1}D_0)^{-1}_{22} = d_{22}^{-1} [d_{21}d_{11}^{-1}S_{11}^A - S_{21}^A] (S_{11}^A)^{-1} \left[ d_{21}d_{11}^{-1}S_{11}^A - S_{21}^A \right]^\top (d_{22}^{-1})^\top
\]

\[
+ d_{22}^{-1} \left[ S_{22}^A - S_{21}^A (S_{11}^A)^{-1} S_{12}^A \right] (d_{22}^{-1})^\top.
\]

(26)

Thus 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: \( S_{22}^A - S_{21}^A (S_{11}^A)^{-1} S_{12}^A \) is the variance-covariance matrix of the second set of moment conditions conditional on the first. The second diagonal block of \( (D_0^\top S D_0)^{-1}_{22} \), which gives the standard errors for \( \theta_2 \) under standard GMM, has an analogous decomposition.

\(^9\)Recall that \( D_0^A = D_0 \).
Adding the new data reduces the first term in (26), just as it reduces the asymptotic variance of $\theta_1$:

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

However the second term in (26) 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:

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

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,

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

and the standard errors for $\theta_2$ also shrink by a factor of $1 - \sqrt{\lambda}$. At the other extreme, suppose that $f_2$ tells you nothing about $\theta_2$, i.e. $d_{21} = 0$ ($\theta_1$ 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 $\theta_2$.

Of course, even if the two sets of moment conditions are independent ($S_{21} = S_{12}^\top = 0$), the sampling variance of $\theta_2$ may still fall because the sampling variance of $\theta_1$ is reduced. As long as $d_{21} \neq 0$, the first term in (26) is nonzero and there is an effect on the standard errors of $\theta_2$. Similarly, even if there is no impact of $\theta_1$ on the second set of moment conditions ($d_{21} = 0$) the first set of moment conditions help to estimate $\theta_2$ if the covariance between the two moment conditions is nonzero.

### 5 Applications to international value and growth portfolios

This section illustrates our methods in applications to international stock return data. Our first application investigates factor regressions; our second application investigates return predictability. In both cases, we focus attention on value and growth portfolios in the United States and internationally.

Fama and French (1992) show using post-war US data, that when stocks are sorted into deciles based on ratios of book value to market value, the returns increase in the decile, and that these returns are not explained by an increase in beta with respect to the market portfolio. Davis, Fama,
and French (2000) show that returns increase in the decile for US data going back to 1927. Moreover, as shown by Fama and French (1998), Chan, Hamao, and Lakonishok (1991), and Rouwenhorst (1999) among others, the outperformance of low book-to-market (value) portfolios relative to high book-to-market (growth) portfolios is a robust feature of the international data. Unlike data for the US, however, reliable international data begin much later than 1927. For example, Fama and French (1992) truncate all their data to begin in 1977, a date that is rarely used as a start date when international data is not part of the study. Our methods allow international data to be used at the same time as longer US data.\textsuperscript{10}

5.1 Data

Data on factors and on international book-to-market portfolios come from Ken French’s website. The raw data on international book-to-market portfolios come from Morgan Stanley’s Capital International Perspectives (MSCI). Value and growth portfolios for each country are formed using the book-to-market ratio. Portfolios include firms with book-to-market (B/M) ratios in the top 30\% for that country (value), and the bottom 30\% for that country (growth). Portfolios are formed at the end of December for each year, and monthly returns are constructed for the following twelve months. Monthly return data for each country begin in 1975 or later and end in 2003. Within each portfolio, firms are value-weighted. Value and growth portfolios for the EAFE (Europe, Australia, and the Far East) index are formed by weighting the country portfolios by their weight in the EAFE index. Monthly index data begin in 1975 and end in 2004. The high and low book-to-market portfolios in the US are constructed similarly, with the high book-to-market portfolio containing firms with ratios in the top 30\% and the low book-to-market portfolio containing firms with ratios in the bottom 30\%. Monthly US data begin in 1927 and end in 2004.

We estimate one and two-factor models. For the one-factor model, the factor is the value-weighted return on the CRSP index less the return on the one-month Treasury bill. We denote this factor $r_M$. For the two-factor model, the factors are $r_M$ and HML, the average return on the small and large value portfolio minus the average return on the small and large growth portfolio, as constructed by Fama and French (1993). For the predictability regressions, we use a single predictor variable, the log of the price-dividend ratio, to maintain simplicity. The price-dividend ratio is for the CRSP value-weighted index, and is constructed by dividing the price by the sum of

\textsuperscript{10}Our analysis is similar to that of Stambaugh (1997), who analyzes means and variances of international returns. Rather than examining means and variances, we study regression coefficients, a case that is not covered by the maximum likelihood analysis in Stambaugh.
the dividends over the past year.

5.2 Estimation

We first estimate factor models for value and growth portfolios in the US and internationally. Let $F_t$ denote a $K \times 1$ vector of returns on factor portfolios at time $t$. We estimate the equation

$$r_t = \alpha + \beta F_t + \epsilon_t,$$

where $\beta$ is $1 \times K$. This equation is first estimated jointly for excess returns $r_t$ on four portfolios: $H$ and $L$ for the US and $H$ and $L$ for the EAFE index. These four returns imply $4(K + 1)$ moment conditions. Let

$$f_1(x_{1t}, \theta) = \begin{bmatrix} H_{1t} - \alpha_{H1} - \beta_{H1} F_t \\ L_{1t} - \alpha_{L1} - \beta_{L1} F_t \end{bmatrix} \otimes \begin{bmatrix} 1 \\ F_t \end{bmatrix},$$

where $H_1$ and $L_1$ denote excess returns on the US high and low book-to-market portfolios respectively, and let

$$f_2(x_{2t}, \theta) = \begin{bmatrix} H_{2t} - \alpha_{H2} - \beta_{H2} F_t \\ L_{2t} - \alpha_{L2} - \beta_{L2} F_t \end{bmatrix} \otimes \begin{bmatrix} 1 \\ F_t \end{bmatrix},$$

where $H_2$ and $L_2$ excess dollar returns on, respectively, the high and low book-to-market portfolios for the EAFE index. Here, $x_{1t} = (H_{1t}, L_{1t}, F_t^T)$, $x_{2t} = (H_{2t}, L_{2t}, F_t^T)$, and

$$\theta = (\alpha_{H1}, \beta_{H1}^T, \alpha_{L1}, \beta_{L1}^T, \alpha_{H2}, \beta_{H2}^T, \alpha_{L2}, \beta_{L2}^T).$$

Then the regression coefficients identified are exactly by the conditions

$$E[f_1(x_{1t}, \theta)] = 0 \quad (29)$$
$$E[f_2(x_{2t}, \theta)] = 0. \quad (30)$$

Because $f_1$ and $f_2$ depend on different parameters, we are in the setting of Section 4.

We estimate (28) using the short, adjusted-moment, and over-identified methods. Short is simply standard GMM with (29) and (30) as moment conditions, and monthly data from 1975 to 2004. $\hat{S}_{XT}$ is constructed as the sample variance-covariance matrix of $f_1$ and $f_2$ using the short estimates and data from 1975 to 2004. A consistent estimate of $D_0$ is obtained as

$$\hat{D}_T = I_4 \otimes \frac{1}{T} Z^T Z,$$

\[11\]Following Fama and French (1998), excess returns are computed as raw dollar returns less the 1-month US Treasury bill return.
where

\[ Z = \begin{bmatrix}
1 & F_1^\top \\
\vdots & \vdots \\
1 & F_T^\top 
\end{bmatrix}, \]

and \( I_4 \) is the 4x4 identity matrix. The moments for the adjusted-moment estimator are then given by (8), with \( f_1, f_2 \) defined as above, and \( \hat{B}_{21,\lambda T} = \hat{S}_{11,\lambda T} \left( \hat{S}_{11,\lambda T} \right)^{-1} \). Similarly, the over-identified estimator is defined by (9), with \( S^T \) constructed using submatrices of \( \hat{S}_{\lambda T} \). For purposes of comparison, the standard errors for all three estimators are constructed using \( \hat{D}_T \) and adjusted-moment estimates.

Panel A of Table 1 reports estimation results for the one-factor model, where \( F_t = r_{M,t} \), the return on the CRSP value-weighted index less the return on the one-month Treasury bill. As Table 1 shows, when the short method is used, the high book-to-market portfolio has a positive \( \alpha \), both in US and in international data. The point estimate in the US is 0.37 (monthly), while for the EAFE, it is 0.58. The low book-to-market portfolio in the US has a negative \( \alpha \) (-0.14) and the EAFE low book-to-market portfolio has an \( \alpha \) that is essentially zero (-0.01). The value minus growth portfolio has a positive and significant \( \alpha \), both for the US (where the value is 0.52), and for the EAFE (where the value is 0.58).

The adjusted-moment method uses the longer sample, and produces an \( \alpha \) on the high book-to-market portfolio that is substantially lower (0.18) than the short estimate of \( \alpha \). The \( \alpha \) on the low book-to-market portfolio is closer to zero, leading to an \( \alpha \) on \( H_1 - L_1 \) that is half of what it is in the shorter sample. As shown in Section 4, the adjusted-moment estimator produces the same coefficients as if the US equations were estimated by themselves. The over-identified estimator produces slightly different estimates because the overlapping and non-overlapping parts of the sample receive slightly different weights than under the adjusted-moment estimator. However, the difference between the results from the over-identified and adjusted-moment estimators are small.

As discussed in Section 4, introducing the longer data on the US can also change the estimates for the short-history assets, in this case the high and low book-to-market portfolios for the EAFE. The estimate for \( \alpha \) on \( H_2 \) falls from 0.58 to 0.52 and the estimate for \( \beta \) rises from 0.57 to 0.73 when the adjusted-moment estimator is used (estimates for the over-identified estimator are similar). The intuition behind this increase lies in the formula (11). The adjusted-moment estimator takes the short moment condition for the EAFE, and adjusts it based on the difference between US sample averages over the full sample and the subsample. In this case, the differences between these sample averages indicates that the point estimate for \( \alpha \) is unusually high and the estimate for \( \beta \)
Table 1: CAPM and Two-Factor Regressions for Value and Growth Portfolios

The CAPM (Panel A) and the two-factor model (Panel B) are estimated for returns on US and EAFE portfolios formed on book-to-market equity (B/M). All returns are monthly, in dollars. Excess returns on the high and low B/M portfolios are denoted as $H$ and $L$ respectively. The US portfolios have a subscript 1; the EAFE portfolios have a subscript 2. The factors $r_M$ and $HML$ are constructed as described in Section 5.1. The regressions are each jointly estimated for $H_1, L_1, H_2, L_2$. “Short” denotes standard GMM; “AM” denotes the adjusted-moment method; “OI” denotes the over-identified method. Standard errors, in parentheses, are computed using AM estimates. Data for Short span the 1975–2004 period; data for AM and OI span the 1927–2004 period.

<table>
<thead>
<tr>
<th>Panel A: $r = \alpha + \beta r_M + \epsilon$</th>
<th>Short</th>
<th>AM</th>
<th>OI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r$</td>
<td>$\alpha$</td>
<td>$\beta$</td>
</tr>
<tr>
<td>US $H_1$</td>
<td>0.37</td>
<td>0.85</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>(0.16)</td>
<td>(0.05)</td>
<td>(0.10)</td>
</tr>
<tr>
<td>$L_1$</td>
<td>-0.14</td>
<td>1.07</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.02)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>$H_1 - L_1$</td>
<td>0.52</td>
<td>-0.21</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>(0.20)</td>
<td>(0.07)</td>
<td>(0.12)</td>
</tr>
<tr>
<td>EAFE $H_2$</td>
<td>0.58</td>
<td>0.57</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>(0.24)</td>
<td>(0.06)</td>
<td>(0.24)</td>
</tr>
<tr>
<td>$L_2$</td>
<td>-0.01</td>
<td>0.61</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>(0.22)</td>
<td>(0.06)</td>
<td>(0.22)</td>
</tr>
<tr>
<td>$H_2 - L_2$</td>
<td>0.58</td>
<td>-0.04</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>(0.14)</td>
<td>(0.04)</td>
<td>(0.14)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: $r = \alpha + \beta_M r_M + \beta_{HML} HML + \epsilon$</th>
<th>Short</th>
<th>AM</th>
<th>OI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r$</td>
<td>$\alpha$</td>
<td>$\beta_M$</td>
</tr>
<tr>
<td>US $H_1$</td>
<td>0.11</td>
<td>0.98</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>(0.09)</td>
<td>(0.02)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>$L_1$</td>
<td>-0.05</td>
<td>1.02</td>
<td>-0.18</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.01)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>$H_1 - L_1$</td>
<td>0.16</td>
<td>-0.04</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>(0.12)</td>
<td>(0.03)</td>
<td>(0.10)</td>
</tr>
<tr>
<td>EAFE $H_2$</td>
<td>0.44</td>
<td>0.64</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>(0.23)</td>
<td>(0.05)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>$L_2$</td>
<td>-0.00</td>
<td>0.61</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>(0.22)</td>
<td>(0.04)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>$H_2 - L_2$</td>
<td>0.44</td>
<td>0.03</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>(0.13)</td>
<td>(0.03)</td>
<td>(0.06)</td>
</tr>
</tbody>
</table>
is unusually low in post-1977 data: the positive correlation between the errors then indicates that
the α estimate for the EAFE should be adjusted downward, while the estimate for β should be
adjusted upward. Nonetheless, the coefficients on α for the EAFE value minus growth strategy are
significant at the 1% level.

Panel B reports analogous results for the two-factor model, where the factors are $r_M$ and HML,
constructed as described in Section 5.1. Including HML in the regression shrinks the αs on all the
portfolios toward zero. Using the longer data in this case has a smaller effect on the point estimates
for the US data. Moreover, the additional factor reduces the correlation between the US and the
EAFE moment conditions, leading to little change in the EAFE estimates between the short and
the AM and OI estimators.

Table 2 reports results of estimating the regression equation (28) for the portfolio that is long
the value portfolio and short the growth portfolio in the US and in each country. To focus on the
effect of the longer US data on the shorter country-specific data, we run (28) as a system of two
equations

$$H_{1t} - L_{1t} = \alpha_1 + \beta_1 r_{Mt} + \epsilon_{1t}$$
$$H_{2t} - L_{2t} = \alpha_2 + \beta_2 r_{Mt} + \epsilon_{2t}$$

separately for each country. The moment conditions and standard errors are calculated exactly as
above, except that here

$$f_1(x_{1t}, \theta) = (H_{1t} - L_{1t} - \alpha_1 - \beta_1 F_t) \otimes \left[ \begin{array}{c} 1 \\ F_t \end{array} \right],$$

$$f_2(x_{2t}, \theta) = (H_{2t} - L_{2t} - \alpha_2 - \beta_2 F_t) \otimes \left[ \begin{array}{c} 1 \\ F_t \end{array} \right],$$

and

$$\hat{D}_T = I_2 \otimes \frac{1}{T} Z^\top Z.$$  

Panels A, B, and C of Table 2 report the start year for each country and the coefficients $\alpha_2$ and $\beta_2$
under the short, adjusted-moment, and over-identified methods. The estimates of $\alpha_1$ and $\beta_1$ differ
only slightly for each estimation (and are close to their estimates in Table 1) and so are omitted.
For the remainder of the discussion, we drop the subscript 2 and refer only to $\alpha$ and $\beta$.  

24
Table 2: CAPM Regressions for Individual Country Value and Growth Portfolios

For each country, a one-factor model is estimated jointly for \( H_1 - L_1 \) and \( H_2 - L_2 \), where \( H_1 - L_1 \) is the return on the high book-to-market (B/M) portfolio minus the return on the low B/M portfolio in the US and \( H_2 - L_2 \) is the analogous portfolio in that country. All returns are monthly, in dollars. The constant (\( \alpha \)) and factor loading (\( \beta \)) are reported for \( H_2 - L_2 \); the one factor is \( r_M \), described in Section 5.1. The US data are from 1927-2003; the starting year for the international data is given in the column “Start Yr.”. “Short” denotes standard GMM; “AM” denotes the adjusted-moment method; “OI” denotes the over-identified method. Standard errors, in parentheses, are computed using AM estimates. The last panel reports coefficients averaged across all countries and cross-sectional \( t \)-statistics.

<table>
<thead>
<tr>
<th>Country</th>
<th>Start Yr</th>
<th>Short ( \alpha )</th>
<th>Short ( \beta )</th>
<th>AM ( \alpha )</th>
<th>AM ( \beta )</th>
<th>OI ( \alpha )</th>
<th>OI ( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>1987</td>
<td>0.83</td>
<td>0.12</td>
<td>1.08</td>
<td>0.05</td>
<td>1.09</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.40)</td>
<td>(0.08)</td>
<td>(0.40)</td>
<td>(0.08)</td>
<td>(0.40)</td>
<td>(0.08)</td>
</tr>
<tr>
<td>Belgium</td>
<td>1975</td>
<td>0.50</td>
<td>-0.05</td>
<td>0.32</td>
<td>0.02</td>
<td>0.30</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.23)</td>
<td>(0.05)</td>
<td>(0.23)</td>
<td>(0.05)</td>
<td>(0.23)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>Denmark</td>
<td>1989</td>
<td>0.42</td>
<td>-0.28</td>
<td>0.56</td>
<td>-0.31</td>
<td>0.56</td>
<td>-0.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.46)</td>
<td>(0.10)</td>
<td>(0.46)</td>
<td>(0.10)</td>
<td>(0.46)</td>
<td>(0.10)</td>
</tr>
<tr>
<td>Finland</td>
<td>1988</td>
<td>0.32</td>
<td>-0.74</td>
<td>0.18</td>
<td>-0.05</td>
<td>0.11</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.79)</td>
<td>(0.22)</td>
<td>(0.74)</td>
<td>(0.20)</td>
<td>(0.74)</td>
<td>(0.20)</td>
</tr>
<tr>
<td>France</td>
<td>1975</td>
<td>0.43</td>
<td>0.08</td>
<td>0.15</td>
<td>0.28</td>
<td>0.12</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.27)</td>
<td>(0.08)</td>
<td>(0.26)</td>
<td>(0.07)</td>
<td>(0.26)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>Germany</td>
<td>1975</td>
<td>0.55</td>
<td>-0.08</td>
<td>0.35</td>
<td>0.13</td>
<td>0.33</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.24)</td>
<td>(0.08)</td>
<td>(0.22)</td>
<td>(0.06)</td>
<td>(0.22)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>Great Britain</td>
<td>1975</td>
<td>0.29</td>
<td>-0.03</td>
<td>0.34</td>
<td>0.02</td>
<td>0.34</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.20)</td>
<td>(0.05)</td>
<td>(0.20)</td>
<td>(0.05)</td>
<td>(0.20)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>Ireland</td>
<td>1991</td>
<td>1.34</td>
<td>-0.17</td>
<td>1.98</td>
<td>0.00</td>
<td>2.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.72)</td>
<td>(0.17)</td>
<td>(0.72)</td>
<td>(0.17)</td>
<td>(0.72)</td>
<td>(0.17)</td>
</tr>
<tr>
<td>Italy</td>
<td>1975</td>
<td>-0.11</td>
<td>-0.01</td>
<td>-0.16</td>
<td>-0.03</td>
<td>-0.17</td>
<td>-0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.28)</td>
<td>(0.06)</td>
<td>(0.28)</td>
<td>(0.06)</td>
<td>(0.28)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>Netherlands</td>
<td>1975</td>
<td>0.19</td>
<td>0.20</td>
<td>0.40</td>
<td>0.19</td>
<td>0.43</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.30)</td>
<td>(0.07)</td>
<td>(0.30)</td>
<td>(0.07)</td>
<td>(0.30)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>Norway</td>
<td>1975</td>
<td>0.33</td>
<td>-0.16</td>
<td>0.16</td>
<td>-0.31</td>
<td>0.14</td>
<td>-0.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.45)</td>
<td>(0.11)</td>
<td>(0.44)</td>
<td>(0.11)</td>
<td>(0.44)</td>
<td>(0.11)</td>
</tr>
<tr>
<td>Spain</td>
<td>1975</td>
<td>0.33</td>
<td>-0.21</td>
<td>-0.02</td>
<td>-0.19</td>
<td>-0.06</td>
<td>-0.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.38)</td>
<td>(0.09)</td>
<td>(0.37)</td>
<td>(0.09)</td>
<td>(0.37)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>Sweden</td>
<td>1975</td>
<td>0.70</td>
<td>-0.16</td>
<td>0.40</td>
<td>0.11</td>
<td>0.37</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.35)</td>
<td>(0.08)</td>
<td>(0.33)</td>
<td>(0.07)</td>
<td>(0.33)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>Switzerland</td>
<td>1975</td>
<td>0.07</td>
<td>0.25</td>
<td>-0.06</td>
<td>0.26</td>
<td>-0.07</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.23)</td>
<td>(0.06)</td>
<td>(0.23)</td>
<td>(0.06)</td>
<td>(0.23)</td>
<td>(0.06)</td>
</tr>
</tbody>
</table>
Table 2: CAPM Regressions on Individual Country Value and Growth Portfolios (cont.)

<table>
<thead>
<tr>
<th>Country</th>
<th>Start Yr</th>
<th>Short</th>
<th>AM</th>
<th>OI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\alpha$</td>
<td>$\beta$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Canada</td>
<td>1977</td>
<td>0.32</td>
<td>-0.09</td>
<td>-0.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.31)</td>
<td>(0.09)</td>
<td>(0.29)</td>
</tr>
<tr>
<td>Australia</td>
<td>1975</td>
<td>0.80</td>
<td>-0.11</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.23)</td>
<td>(0.06)</td>
<td>(0.23)</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>1975</td>
<td>0.21</td>
<td>0.12</td>
<td>-0.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.32)</td>
<td>(0.08)</td>
<td>(0.32)</td>
</tr>
<tr>
<td>Japan</td>
<td>1975</td>
<td>1.00</td>
<td>-0.12</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.27)</td>
<td>(0.07)</td>
<td>(0.27)</td>
</tr>
<tr>
<td>New Zealand</td>
<td>1988</td>
<td>-1.47</td>
<td>0.15</td>
<td>-1.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.63)</td>
<td>(0.13)</td>
<td>(0.62)</td>
</tr>
<tr>
<td>Singapore</td>
<td>1975</td>
<td>0.52</td>
<td>0.20</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.32)</td>
<td>(0.08)</td>
<td>(0.31)</td>
</tr>
</tbody>
</table>

Panel D: Average Coefficients

<table>
<thead>
<tr>
<th></th>
<th>Short</th>
<th>AM</th>
<th>OI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha$</td>
<td>$\beta$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Average</td>
<td>0.38</td>
<td>-0.06</td>
<td>0.31</td>
</tr>
<tr>
<td>$t$-Statistic</td>
<td>3.09</td>
<td>-1.12</td>
<td>2.06</td>
</tr>
</tbody>
</table>

26
Panels A, B, and C show that for nearly all countries in the sample the $H - L$ portfolio has a positive $\alpha$ when the short method is used. Indeed, the $\alpha$ on this portfolio is negative only in Italy and New Zealand. Panel D shows the average $\alpha$ and $\beta$ computed by equal-weighting the countries. The average $\alpha$ is 0.38, and treating each $\alpha$ as a separate observation results in a $t$-statistic of 3.09. The average $\beta$ is -0.06 (with a $t$-statistic of -1.12) indicating that the average $\beta$ for the high book-to-market portfolio is slightly below the average $\beta$ for the low book-to-market portfolio. Under the adjusted-moment method, some $\alpha$s that were positive but small in magnitude under short become negative and small in magnitude (Spain, Switzerland, Canada, Hong Kong). This occurs for the same reason that the estimate of $\alpha$ on $H - L$ for the EAFE portfolio falls: including the first part of the sample results in a lower estimate of $\alpha$ in the US, and, because the moment conditions in the US and in the countries tend to be positively correlated, the estimate for $\alpha$ tends to fall for the other countries as well. Under the adjusted-moment estimator, the average $\alpha$ is equal to 0.31, with a $t$-statistic of 2.06. The average $\beta$ is slightly positive.

We also estimate the predictive regression
\[
\begin{equation}
r_{t+1} = \alpha + \beta(p_t - d_t) + \epsilon_{t+1}
\end{equation}
\]
jointly for monthly excess returns on $H$ and $L$ for the US and $H$ and $L$ for the EAFE index. The moment conditions take the same form as those for the factor regressions described above, with the factors replaced by the lagged price-dividend ratio $p_t - d_t$.

Table 3 reports the predictive regression results. As Campbell and Shiller (1988) and Fama and French (1989) show for the US aggregate market, the coefficient on the predictor variable is negative for value and growth, both for the US and EAFE portfolios. For the US, the coefficients on the value and the growth portfolios are approximately the same in the post-war sample. In the longer sample, the coefficient is larger in magnitude for the value portfolio than the growth portfolio, as shown by both the AM and OI estimates. The longer data leads this coefficient to become statistically significant for the value portfolio. In contrast, in international data, the point estimate for growth is larger in magnitude than that for value. This leads the price-dividend ratio to predict returns on the value-minus-growth portfolio with a positive sign. This result is attenuated when the longer data for the US is used.

A notable feature of this application is that the efficiency gains of using the longer data are apparent both for the US and the EAFE estimates. Even though no additional data are used for the EAFE portfolios, the standard error for the predictive coefficient falls by about 15% for the value portfolio and 20% for the growth portfolio. Thus introducing the longer data for the US
Table 3: Predictive Regressions for Value and Growth Portfolios

The regression

\[ r_{t+1} = \alpha + \beta(p_t - d_t) + \epsilon_{t+1} \]

is estimated for US and international portfolios formed on book-to-market equity (B/M), where \( p_t - d_t \) is the log of the price-dividend ratio on the CRSP value-weighted portfolio. All returns are monthly, in dollars. Excess returns on the high and low B/M portfolios are denoted as \( H \) and \( L \) respectively. The US portfolios have a subscript 1; the EAFE portfolios have a subscript 2. The regressions are each jointly estimated for \( H_1, L_1, H_2, L_2 \). “Short” denotes standard GMM; “AM” denotes the adjusted-moment method; “OI” denotes the over-identified method. Standard errors, in parentheses, are computed using AM estimates. Data for Short span the 1975–2004 period; data for AM and OI span the 1927–2004 period.

<table>
<thead>
<tr>
<th></th>
<th>Short</th>
<th>AM</th>
<th>OI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( r )</td>
<td>( \alpha )</td>
<td>( \beta )</td>
</tr>
<tr>
<td>US</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( H_1 )</td>
<td>3.01</td>
<td>-0.58</td>
<td>5.07</td>
</tr>
<tr>
<td></td>
<td>(2.13)</td>
<td>(0.67)</td>
<td>(1.32)</td>
</tr>
<tr>
<td>( L_1 )</td>
<td>2.93</td>
<td>-0.65</td>
<td>3.20</td>
</tr>
<tr>
<td></td>
<td>(2.62)</td>
<td>(0.82)</td>
<td>(1.62)</td>
</tr>
<tr>
<td>( H_1 - L_1 )</td>
<td>0.08</td>
<td>0.07</td>
<td>1.86</td>
</tr>
<tr>
<td></td>
<td>(1.98)</td>
<td>(0.62)</td>
<td>(1.22)</td>
</tr>
<tr>
<td>EAFE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( H_2 )</td>
<td>1.84</td>
<td>-0.25</td>
<td>2.87</td>
</tr>
<tr>
<td></td>
<td>(2.54)</td>
<td>(0.79)</td>
<td>(2.19)</td>
</tr>
<tr>
<td>( L_2 )</td>
<td>2.60</td>
<td>-0.62</td>
<td>2.89</td>
</tr>
<tr>
<td></td>
<td>(2.28)</td>
<td>(0.71)</td>
<td>(1.85)</td>
</tr>
<tr>
<td>( H_2 - L_2 )</td>
<td>-0.77</td>
<td>0.37</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>(1.61)</td>
<td>(0.51)</td>
<td>(1.51)</td>
</tr>
</tbody>
</table>
results in considerably greater precision in the EAFE estimation.

5.3 Results in repeated samples

Besides illustrating how our methods can be applied, the settings above also provides a laboratory for evaluating the finite sample properties of our estimators. For both the contemporaneous factor regressions and the predictability regression, we conduct a Monte Carlo experiment where we simulate 50,000 identical samples constructed to resemble the sample in the data. We first simulate from (28), where the returns are excess returns on $H_1$, $L_1$, $H_2$, and $L_2$, and the factors are either $r_M$, or $r_M$ and HML. Errors are assumed to be normally distributed and independent across time; estimates of $\alpha$, $\beta$, the premium on the factors, the variance-covariance matrix of $\epsilon$, and the variance-covariance matrix of the factors are obtained using adjusted-moment estimates. The sample size for US returns is 936 months; for EAFE returns it is 360 months.

All three methods give unbiased estimates for the coefficients $\alpha$ and $\beta$. For this reason, Table 4 reports the standard deviation of the estimates obtained from the Monte Carlo. For the US, the adjusted-moment and over-identified methods dramatically shrink the standard deviation of the estimates, for both the one and two-factor models. For EAFE there is little shrinkage in the standard deviations. However, this is not a reflection of poor small-sample properties, as the asymptotic standard errors are also about the same regardless of which method is used. Both the asymptotic and the small-sample results show that, in this application, most of the efficiency gains arise from superior estimation of the US parameters.

We then conduct a Monte Carlo experiment to evaluate the finite-sample properties of our predictive regressions. We simulate 50,000 identical samples constructed from (31), where the returns are excess returns on $H_1$, $L_1$, $H_2$, and $L_2$. The price-dividend ratio is assumed to follow a first-order autoregressive process. Estimates of the price-dividend ratio process, of $\alpha$, $\beta$, the variance-covariance matrix of $\epsilon$, and the covariance of shocks to the price-dividend ratio with $\epsilon$ are obtained using the adjusted-moment method. As in the previous experiment, the sample size for US returns is 936 months; for EAFE returns it is 360 months.

Panel A of Table 5.3 reports the standard deviation of the estimates obtained from the Monte Carlo. As is well-known (see Stambaugh (1999)), the estimates of $\beta$ in (31) are biased. Panel B reports this bias (the difference between the estimated and the true coefficient) for each method. As in the previous example, the adjusted-moment and over-identified methods dramatically shrink the standard deviation of the estimates for the US coefficients. The standard deviations for the
EAFE coefficients also shrink by a substantial amount: 14% for the value portfolio and 17% for the growth portfolio when the AM method is used, and more when the OI method is used. Thus the shrinkage reported earlier is not confined to asymptotic standard errors; it is also present in small-sample standard errors calculated using the Monte Carlo. Moreover, as Panel B shows, applying the AM and the OI methods leads to a substantially reduced bias in the coefficient on the predictive variable.

6 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 $\sum_{k=1}^{\eta_n} \eta_k = 1$. Let $n_0$ be the smallest integer such that the product with $\eta_j$ 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 $g$:

$$g_{\eta_1T}(\theta) = \frac{1}{\eta_1T} \sum_{t=1}^{\eta_1T} f(x_t, \theta)$$

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

**Theorem 6.1** Define $g_{\eta_jT}$ as in (32) and $g_{\eta_jT}$ as in (33) for $j = 2, \ldots, n$. Assumptions 1–5 imply

$$\sqrt{T} \left( \begin{array}{c} \sqrt{\eta_1}g_{\eta_1T}(\theta_0) \\ \sqrt{\eta_2}g_{\eta_2T}(\theta_0) \\ \vdots \\ \sqrt{\eta_n}g_{\eta_nT}(\theta_0) \end{array} \right) \to_d N \left( 0, \left[ \begin{array}{cccc} S & 0 & \cdots & 0 \\ 0 & S & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & S \end{array} \right] \right)$$

as $T \to \infty$.

**Proof:** See Appendix D.

6.1 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 6.1 gives the joint distribution of partial sums of $g$. We now use this result to extend the over-identified estimator.
Table 4: CAPM and Two-Factor Regressions in Repeated Samples

50,000 samples are simulated assuming joint normality of factors and returns. The number of assets and factors corresponds to those in Table 1 and the true moments are set equal to the adjusted-moment estimates. As in Table 1, the length of the US data is 936 months while the length of the international data is 360 months. The table reports standard deviations of the point estimates. The regressions are each jointly estimated for $H_1, L_1, H_2, L_2$. “Short” denotes standard GMM; “AM” denotes the adjusted-moment method; “OI” denotes the over-identified method.

### Panel A: $r = \alpha + \beta r_M + \epsilon$

<table>
<thead>
<tr>
<th></th>
<th>Short</th>
<th>AM</th>
<th>OI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r$</td>
<td>$\sigma(\hat{\alpha})$</td>
<td>$\sigma(\hat{\beta})$</td>
</tr>
<tr>
<td>US</td>
<td>$H_1$</td>
<td>0.116</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td>$L_1$</td>
<td>0.056</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>$H_1 - L_1$</td>
<td>0.157</td>
<td>0.028</td>
</tr>
<tr>
<td>EAFE</td>
<td>$H_2$</td>
<td>0.233</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>$L_2$</td>
<td>0.222</td>
<td>0.041</td>
</tr>
<tr>
<td></td>
<td>$H_2 - L_2$</td>
<td>0.139</td>
<td>0.026</td>
</tr>
</tbody>
</table>

### Panel B: $r = \alpha + \beta r_M + \beta_{H_{M,L}} r_{H_{M,L}} + \epsilon$

<table>
<thead>
<tr>
<th></th>
<th>Short</th>
<th>AM</th>
<th>OI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r$</td>
<td>$\sigma(\hat{\alpha})$</td>
<td>$\sigma(\hat{\beta}_{M})$</td>
</tr>
<tr>
<td>US</td>
<td>$H_1$</td>
<td>0.077</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>$L_1$</td>
<td>0.046</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>$H_1 - L_1$</td>
<td>0.101</td>
<td>0.019</td>
</tr>
<tr>
<td>EAFE</td>
<td>$H_2$</td>
<td>0.231</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>$L_2$</td>
<td>0.226</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>$H_2 - L_2$</td>
<td>0.133</td>
<td>0.025</td>
</tr>
</tbody>
</table>
Table 5: Predictive Regressions in Repeated Samples

50,000 samples are simulated assuming joint normality of returns and the price-dividend ratio. The number of assets corresponds to those in Table 3. The price-dividend ratio is assumed to follow a first-order autoregressive process. The true return and price-dividend ratio parameters are set equal to their adjusted-moment estimates. As in Table 3, the length of the US data is 936 months while the length of the international data is 360 months. Panel A reports standard deviations of the point estimates. Panel B reports the bias in the coefficient on the predictor variable. The regressions are each jointly estimated for $H_1, L_1, H_2, L_2$. “Short” denotes standard GMM; “AM” denotes the adjusted-moment method; “OI” denotes the over-identified method.

### Panel A: Standard deviations of coefficient estimates

<table>
<thead>
<tr>
<th></th>
<th>Short</th>
<th>AM</th>
<th>OI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma(\hat{\alpha})$</td>
<td>$\sigma(\hat{\beta})$</td>
<td>$\sigma(\hat{\alpha})$</td>
</tr>
<tr>
<td>US</td>
<td>H1</td>
<td>4.13</td>
<td>1.18</td>
</tr>
<tr>
<td></td>
<td>L1</td>
<td>4.91</td>
<td>1.40</td>
</tr>
<tr>
<td></td>
<td>$H_1 - L_1$</td>
<td>2.57</td>
<td>0.74</td>
</tr>
<tr>
<td>EAFE</td>
<td>H2</td>
<td>4.39</td>
<td>1.25</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td>4.34</td>
<td>1.24</td>
</tr>
<tr>
<td></td>
<td>$H_2 - L_2$</td>
<td>2.18</td>
<td>0.62</td>
</tr>
</tbody>
</table>

### Panel B: Bias in the coefficient on the predictor variable

<table>
<thead>
<tr>
<th></th>
<th>Short</th>
<th>AM</th>
<th>OI</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>H1</td>
<td>-1.03</td>
<td>-0.39</td>
</tr>
<tr>
<td></td>
<td>L1</td>
<td>-1.31</td>
<td>-0.49</td>
</tr>
<tr>
<td>EAFE</td>
<td>H2</td>
<td>-0.67</td>
<td>-0.26</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td>-0.70</td>
<td>-0.27</td>
</tr>
</tbody>
</table>
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 were data missing at the start of the sample.

Consider intervals of the data defined by points in time where at least one sample moment starts or ends. Say these points in time divide the sample up into disjoint intervals $1, \ldots, n$. These intervals can be weakly ordered by how many of the sample moments are observed over that interval. That is, $\pi_1$ sample moments are observed over the first segment, $\pi_2 \leq \pi_1$ over the second segment, etc. Let $\lambda_1$ 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, $\lambda_2$ 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 $\eta$s are labeled according to their order in the sample, the $\lambda$s 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 $\phi_i$ denote the set of data series that are observed in data segment $\lambda_i$. Define

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

where $\{i_1, \ldots, i_{\pi_j}\} \in \phi_j$ and $i_1 < \cdots < i_{\pi_j}$. Then $f_{\phi_j}$ are the components of $f$ observed over the segment $\lambda_j T$. Similarly, define the $\pi_j \times q$ matrix

$$D_{0,\phi_j} = E \left[ \frac{\partial f_{\phi_j}}{\partial \theta} \right]_{\theta_0} = \left( D_{0,i_{i_1}}^\top, \ldots, D_{0,i_{i_{\pi_j}}}^\top \right)^\top,$$
the $\pi_j \times 1$ vector
\[
g_{\phi_j, \lambda_j T}(\theta) = \frac{1}{\lambda_j T} \sum_{t=t_j}^{t_j+\lambda_j} f_{\phi_j}(x_t, \theta),
\]
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
\[
h_{T}^{I_n}(\theta) = \left[ g_{\phi_1, \lambda_1 T}(\theta)^\top, \ g_{\phi_2, \lambda_2 T}(\theta)^\top, \ldots, \ g_{\phi_n, \lambda_n T}(\theta)^\top \right]^\top,
\]
for $\theta \in \Theta$. The $I_n$ superscript refers to the fact that these are moment conditions for the over-identified estimator, and that there are $n$ non-overlapping intervals. The $T$ subscript refers to the fact that the data length is $T$.\footnote{This notation does not, of course, completely define the over-identified estimator. For that, one would need the points at which the data intervals begin, $t_1, \ldots, t_n$. These points in turn depend in a complicated way on $\lambda_1, \ldots, \lambda_n$ and $T$.} As in Section 2, $\sqrt{T} h_{T}^{I_n}(\theta)$ is asymptotically normally distributed. The following is analogous to Theorem 2.1.

**Theorem 6.2** Assumptions 1–5 imply
\[
\sqrt{T} h_{T}^{I_n}(\theta_0) \to_d N \left( 0, S_{I_n}^{T_n} \right),
\]
where
\[
S_{I_n}^{T_n} = \begin{bmatrix}
\frac{1}{\lambda_1} S_{\phi_1} & 0 & \ldots & 0 \\
0 & \frac{1}{\lambda_2} S_{\phi_2} & \ldots & 0 \\
0 & 0 & \ddots & 0 \\
0 & 0 & 0 & \frac{1}{\lambda_n} S_{\phi_n}
\end{bmatrix}
\]
as $T \to \infty$.

The extended over-identified estimator takes (35) 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 (36). Define
\[
\hat{\theta}_{T}^{I_n} = \arg\min_\theta h_{T}^{I_n}(\theta)^\top W_{T}^{I_n} h_{T}^{I_n}(\theta).
\]
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 $W_{Tn}^T$ converges almost surely to a positive-definite matrix $W_{Tn}$.

**Theorem 6.3** Assumptions 1-5 and 10 imply that as $T \to \infty$, $\hat{\theta}_{Tn}^T \to_{a.s.} \theta_0$.

Define

$$D_{0n}^T = \left[ D_{0,\phi_1}^T D_{0,\phi_2}^T \ldots D_{0,\phi_n}^T \right]^T.$$  

Note that Assumptions 1-5 imply that

$$D_{0n}^T = \lim_{T \to \infty} \frac{\partial h_{Tn}}{\partial \theta} \bigg|_{\theta_0}.$$  

**Theorem 6.4** Assumptions 1-5, and 7-10 imply that as $T \to \infty$,  

$$\sqrt{T}(\hat{\theta}_{Tn}^T - \theta_0) \to_d N\left(0, \left((D_{0n}^T)\top W_{Tn} D_{0n}^T\right)^{-1} \left((D_{0n}^T)\top W_{Tn} S_{Tn} W_{Tn} D_{0n}^T\right) \left((D_{0n}^T)\top W_{Tn} D_{0n}^T\right)^{-1}\right).$$

**Theorem 6.5** Suppose $W_{Tn}^T \to_{a.s.} W_{Tn} = (S_{Tn})^{-1}$. Assumptions 1-5 and 7-10 imply

$$\sqrt{T}(\hat{\theta}_{Tn}^T - \theta_0) \to_d N\left(0, \left((D_{0n}^T)\top (S_{Tn})^{-1} (D_{0n}^T)\right)^{-1}\right).$$ (38)

Moreover, this choice of $W_{Tn}$ is efficient given the moment conditions (35).

The extended over-identified estimator reduces to the over-identified estimator considered in Section 2 when there is a single block of data. Section 6.3 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 6.6** Assume $W_{Tn}^T \to (S_{Tn})^{-1}$ and $W_{(1-\lambda_n)T}^{T_{n-1}} \to (S_{Tn-1})^{-1}$. Assumptions 1-5 and 7-9 imply $\hat{\theta}_{Tn}^T$ is asymptotically more efficient than $\hat{\theta}_{T_{n-1}}^{T_{n-1}}$.  

**Proof:** See Appendix D. $\square$
6.2 Extending the Adjusted-Moment Estimator

This section shows that the adjusted-moment estimator can also be extended to the case where there is a 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 (35), 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 6.1. To simplify notation, we consider a slightly less general problem than in Section 6.1. 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 \( n - 1 \) intervals of data, and extend the adjusted moment estimator to all \( n \) intervals 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 6.3.

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:

\[
h_{\lambda_1 T}^{A_1} = g_{\lambda_1 T}.
\]

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 \( A \) refers to the number of blocks of data. It follows from standard arguments that

\[
E \left[ h_{\lambda_1 T}^{A_1}(\theta_0) \right] = E \left[ f(x_t, \theta_0) \right] = 0,
\]

36
and that

\[ h_{\lambda T}^{A_1}(\theta) \to_{a.s.} E[f(x_t, \theta)] \]

as \( T \to \infty \). Assume by induction that

\[ h_{(1-\lambda_n)T}^{A_{n-1}}(\theta) = E[f(x_t, \theta)], \]

and that

\[ E \left[ (1 - \lambda_n)T \left( h_{(1-\lambda_n)T}^{A_{n-1}}(\theta_0) \right) \left( h_{(1-\lambda_n)T}^{A_{n-1}}(\theta_0) \right)^\top \right] \to_{a.s.} S^{A_{n-1}}, \]

for some symmetric, positive-definite matrix \( S^{A_{n-1}} \). Finally, assume that \( h_{(1-\lambda_n)T}^{A_{n-1}}(\theta_0) \) 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

\[ h_{(1-\lambda_n)T}^{A_{n-1}}(\theta_0) = M_{n-1}[g_{\phi_1}(\theta_0)^\top, \ldots, g_{\phi_{n-1}}(\theta_0)^\top]^\top. \]

This allows Theorem 6.1 to be applied. Note that (40) implies that there is a one-to-one correspondence between moment conditions in the adjusted-moment estimator, and moment conditions \( f_i \).

Let \( h_{\phi_n,(1-\lambda_n)T}^{A_{n-1}}(\theta) \) denote the \( \pi_n \) components of \( h_{(1-\lambda_n)T}^{A_{n-1}}(\theta) \) that converge to \( E[f_{\phi_n}(x_t, \theta)] \). These are the elements of \( h_{\phi_n,(1-\lambda_n)T}^{A_{n-1}}(\theta) \) corresponding to moments observed over the new data segment. We define the adjusted-moment estimator for \( n \) intervals 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

\[ B^{A_{n-1}} = \lim_{T \to \infty} E \left[ T h_{\phi_n,(1-\lambda_n)T}^{A_{n-1}}(\theta_0) \left( h_{\phi_n,(1-\lambda_n)T}^{A_{n-1}}(\theta_0) - g_{\phi_n,(1-\lambda_n)T}(\theta_0) \right) \right] \times \]

\[ E \left[ T \left( h_{\phi_n,(1-\lambda_n)T}^{A_{n-1}}(\theta_0) - g_{\phi_n,(1-\lambda_n)T}(\theta_0) \right) \left( h_{\phi_n,(1-\lambda_n)T}^{A_{n-1}}(\theta_0) - g_{\phi_n,(1-\lambda_n)T}(\theta_0) \right)^\top \right]^{-1}. \]  

\( B^{A_{n-1}} \) is the \( l \times \pi_n \) matrix of asymptotic regression coefficients from a regression of the \( (n-1) \)st adjusted-moment estimator on \( h_{\phi_n,(1-\lambda_n)T}(\theta_0) - g_{\phi_n,\lambda_n}(\theta_0) \), appropriately scaled by the square root of the sample length. In practice, \( B^{A_j} \) can be replaced by a sample estimate \( \hat{B}^{A_j}_{T} \) such that \( \hat{B}^{A_j}_{T} \to_{a.s.} B^{A_j} \) as \( T \to \infty \). Finally define the \( n \)th adjusted-moment estimator as

\[ h_{T}^{A_n}(\theta) = h_{(1-\lambda_n)T}^{A_{n-1}}(\theta) - B^{A_{n-1}} \left( h_{\phi_n,(1-\lambda_n)T}(\theta_0) - g_{\phi_n,\lambda_n}(\theta_0) \right). \]

\( ^{13} \)In that case, (42) would be replaced by the requirement that

\[ h_{(1-\lambda_n)T}^{A_{n-1}}(\theta_0) = M_{n-1,T}[g_{\phi_1}(\theta_0)^\top, \ldots, g_{\phi_{n-1}}(\theta_0)^\top]^\top, \]

where \( \lim_{T \to \infty} M_{n-1,T} \to_{a.s.} M_{n-1} \). None of the arguments would change.
When evaluated at $\theta_0$, $h^{A_n}_T$ is a regression residual. This completes the definition of the adjusted-moment estimator.

We now verify that the induction hypotheses (40)–(42) are valid for $n$. These are necessary to define (44). Because

$$h^{A_{n-1}}_{\phi_n,(1-\lambda_n)T}(\theta) - g_{\phi_n,\lambda_n} T(\theta) \rightarrow_{a.s.} E[f_{\theta_n}(x_t, \theta)] - E[f_{\phi_n}(x_t, \theta)] = 0,$$

(40) is satisfied for $n$. To show (41) for $n$, note first that Theorem 6.1 implies

$$\lim_{T \to \infty} E[h^{A_{n-1}}_{\phi_n,\lambda_n} T(\theta_0) h^{A_{n-1}}_{\phi_n,(1-\lambda_n)T}(\theta_0)^\top] = 0.$$ 

Because $h^{A_n}_T(\theta_0)$ is a regression residual, (44) implies

$$\lim_{T \to \infty} E\left[T h^{A_n}_T(\theta_0) h^{A_n}_T(\theta_0)^\top\right] = \frac{1}{(1-\lambda_n)} S^{A_{n-1}} - B^{A_{n-1}} \left[\begin{array}{c} \frac{1}{\lambda_n} S^{\phi_n} + \frac{1}{(1-\lambda_n)} S^{\phi_n} \\ (1-\lambda_n) T \end{array}\right] (B^{A_{n-1}})^\top,$$

where

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

Clearly (45) is well-defined and symmetric. It is positive definite because $\phi_n$ is a strict subset of $\{1, \ldots, l\}$, so not all the variance in $h^{A_{n-1}}_{(1-\lambda_n)T}$ can be explained by $h^{A_{n-1}}_{\phi_n,(1-\lambda_n)T}$. Finally, (42) follows from the form of (44).

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 6.1, $\lambda_1 = \lambda$ 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:

$$h^{A_1}_{\lambda_1 T} = g_{\lambda T}.$$ 

It follows from (44) that

$$h^{A_2}_T = h^{A_1}_{\lambda_1 T} - B^{A_1} \left(h^{A_1}_{\lambda_1 T} - g_{\phi_2,\lambda_2 T}\right) = g_{\lambda T} - B^{A_1} \left(g_{1,\lambda T} - g_{1,(1-\lambda)T}\right).$$

---

14Equation (40) insures a one-to-one correspondence between moment conditions and components of $h^{A_n}$. Equation (42) insures that $\sqrt{T} h^{A_{n-1}}_{(1-\lambda_n)T}$ and $\sqrt{T} g_{\phi_n,(1-\lambda_n)}$ have an asymptotic distribution that is well-defined (by Theorem 6.1). This implies that $B^{A_{n-1}}$ is well-defined. Equation (41) will be useful later in determining the asymptotic distribution of the adjusted-moment estimator.
By (43),
\[
B^{A_1} = \lim_{T \to \infty} \frac{1}{\lambda} \begin{bmatrix}
S_{11} \\
S_{21}
\end{bmatrix} \left[ \left( \frac{1}{1 - \lambda} + \frac{1}{\lambda} \right) S_{11} \right]^{-1}
= \begin{bmatrix}
(1 - \lambda)I \\
(1 - \lambda)B_{21}
\end{bmatrix},
\]
where we have suppressed the argument \( \theta_0 \) in the first line. Therefore, the moment conditions for the adjusted-moment estimator equal
\[
h^{A_2}_T = \begin{bmatrix}
g_{1,\lambda T} + (1 - \lambda) \left( g_{1,(1 - \lambda)T} - g_{1,\lambda T} \right) \\
g_{2,\lambda T} + (1 - \lambda) B_{21} (g_{1,(1 - \lambda)T} - g_{1,\lambda T})
\end{bmatrix},
\]
which are the same moment conditions as those given in Section 2.\(^{15}\)

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

**Lemma 6.1** Assumptions 1–5 imply
\[
\sqrt{T} \begin{bmatrix}
h_{\lambda T}^{A_n-1}(\theta_0) \\
g_{\phi_n,\lambda T}(\theta_0)
\end{bmatrix} \rightarrow_d N \left( 0, \begin{bmatrix}
\frac{1}{1 - \lambda_n} S_{A_n-1} & 0 \\
0 & \frac{1}{\lambda_n} S_{\phi_n}
\end{bmatrix} \right).
\]

**Proof:** It follows from (42) and Theorem 6.1 that \( \sqrt{T} h_{\lambda T}^{A_n-1}(\theta_0) \) and \( \sqrt{T} g_{\phi_n,\lambda T}(\theta_0) \) are asymptotically independent, and that each are asymptotically normally distributed. The form of the asymptotic variance follows from (41) and Theorem 6.1. □

This lemma implies that the sample moment conditions for the \( n \)th adjusted-moment estimator, scaled by \( \sqrt{T} \) are asymptotically normally distributed. Equation (44) implies an inductive equation for the variance.

**Theorem 6.7** Assumptions 1–5 imply
\[
\sqrt{T} h_T^{A_n}(\theta_0) \rightarrow_d N \left( 0, S^{A_n} \right),
\]
where \( S^{A_n} \) is defined inductively as
\[
S^{A_n} = \frac{1}{1 - \lambda_n} S^{A_{n-1}} - B^{A_{n-1}} \left[ \frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}} \right] \left( B^{A_{n-1}} \right)^\top,
\]
\(^{15}\)Here we make use of the equation \( g_{1T} = \lambda g_{1,\lambda T} + (1 - \lambda)g_{1,(1 - \lambda)T} \).
with

\[ S^{A_1} = S. \] (48)

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

\[ \hat{\theta}^A_n = \text{argmin}_\theta h^A_n(\theta)^\top W^A_n h^A_n(\theta), \] (49)

where \( W^A_n \) satisfies Assumption 11:

**Assumption 11** The weighting matrix \( W^A_n \) converges almost surely to a positive-definite matrix \( W^{A_n} \).

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

\[ h^A_n(\theta) \to \text{a.s.} \ E[f(x_t, \theta)] \]

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

**Theorem 6.8** Assumptions 1-5, and 11 imply that as \( T \to \infty \), \( \hat{\theta}^A_n \to \text{a.s.} \theta_0 \).

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

**Theorem 6.9** Assumptions 1-5, 7-9 and 11 imply that as \( T \to \infty \),

\[
\sqrt{T}(\hat{\theta}^A_n - \theta_0) \overset{d}{\to} N \left( 0, \left( D_0^\top W^{A_n} D_0 \right)^{-1} \left( D_0^\top W^{A_n} S^{A_n} W^{A_n} D_0 \right) \left( D_0^\top W^{A_n} D_0 \right)^{-1} \right) .
\]

Lastly, given moments \( h^A_n \), the most efficient asymptotic weighting matrix is the inverse of the variance of these moments.

**Theorem 6.10** Suppose \( W^A_n \to \text{a.s.} W^{A_n} = (S^{A_n})^{-1} \). Assumptions 1-5 and 7-9 imply that

\[
\sqrt{T}(\hat{\theta}^A_n - \theta_0) \overset{d}{\to} N \left( 0, \left( D_0^\top (S^{A_n})^{-1} D_0 \right)^{-1} \right) .
\] (50)

Moreover, this choice of \( W^{A_n} \) is efficient given the moment conditions (35).

While the extended adjusted-moment estimator appears completely different from the extended over-identified estimator, they are asymptotically equivalent.
Theorem 6.11 Assume that $W_T^{A_n} \to (S^{A_n})^{-1}$ and $W_T^{I_n} \to (S^{I_n})^{-1}$. For any integer $n$, assumptions 1-5 and 7-9 imply that the extended adjusted-moment estimator (49) is asymptotically equivalent to the extended over-identiﬁed estimator (37).

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 6.11 shows that the extended adjusted moment estimator is asymptotically equivalent to the extended over-identiﬁed estimator. By Theorem 6.6, we can conclude that adding a block of data always increases eﬃciency for the adjusted-moment estimator.

Corollary 6.1 Assume that $W_T^{A_n} \to (S^{A_n})^{-1}$ and $W_T^{A_{n-1}}/(1-\lambda_n) \to (S^{A_{n-1}})^{-1}$. Assumptions 1-5 and 7-9 imply $\sqrt{T} \hat{\theta}_T^{A_n}$ is asymptotically more eﬃcient than $\sqrt{T} \hat{\theta}_T^{A_{n-1}}/(1-\lambda_n)$.

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.

6.3 Explicit calculations for several missing data patterns

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 $D_0^T$. 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 (47). 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.

6.3.1 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.\footnote{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.} Figure 2 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 $g_{1.}(\theta)$ to denote partial sums of $f_1(x_{1t}, \theta)$ and $g_{2.}(\theta)$ 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
Figure 2: 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 2, $\lambda_1$ is the length of the middle segment divided by the total data length. Without loss of generality, we assume $\lambda_2$ is the length of the first segment of missing data divided by the total length (we could also have set $\lambda_2$ 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 $h_{I3}^T$ are

$$ h_{I3}^T(\theta) = \begin{bmatrix} g_{1,\lambda_1 T}(\theta)^\top & g_{2,\lambda_2 T}(\theta)^\top & g_{1,\lambda_2 T}(\theta)^\top & g_{1,\lambda_3 T}(\theta)^\top \end{bmatrix}^\top. $$

Then the results in Section 6.1 imply that $\sqrt{T}h_{I3}^T(\theta)$ has asymptotic variance\(^\dagger\)

$$ S_{I3}^2 = \begin{bmatrix} \frac{1}{\lambda_1^2} S_{11} & \frac{1}{\lambda_1^2} S_{12} & 0 & 0 \\ \frac{1}{\lambda_1^2} S_{21} & \frac{1}{\lambda_1^2} S_{22} & 0 & 0 \\ 0 & 0 & \frac{1}{\lambda_2^2} S_{11} & 0 \\ 0 & 0 & 0 & \frac{1}{\lambda_3^2} S_{11} \end{bmatrix}. $$

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

$$ \hat{\theta}_{I3}^T = \arg\min_{\theta} h_{I3}^T(\theta)^\top W_{I3}^T h_{I3}^T(\theta), \quad W_{I3}^T \rightarrow a.s. \ (S_{I3}^2)^{-1} $$

for $W_{I3}^T$ positive definite. The asymptotic distribution is given by

$$ \sqrt{T}(\hat{\theta}_{I3}^T - \theta_0) \rightarrow_d N\left(0, \left(D_{I3}^T\left(S_{I3}^2\right)^{-1} D_{I3}^T\right)^{-1}\right), $$

where

$$ D_{I3}^T = \begin{bmatrix} D_{0,1}^T & 0 \\ 0 & D_{0,2}^T \\ D_{0,1}^T & D_{0,2}^T \end{bmatrix}. $$

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

$$ h_{A1}^{A1}_{T} = g_{\lambda_1 T}. \quad (51) $$

\(^\dagger\)The asymptotic variance does not take exactly the same form as $S_{I}^2$ in Section 2. The reason for the discrepancy is that $S_{I}^2$ was defined as the asymptotic variance of the moment conditions scaled by $\sqrt{T}$, while $S_{I3}^2$ is the asymptotic variance of the moment conditions scaled by $\sqrt{T}$.\[43]
Substituting (51) into (44) yields the adjusted-moment estimator that includes the \( \lambda_2 \) block:

\[
h_{(\lambda_1+\lambda_2)T}^A = g_{\lambda_1T} - B_{A_1}^A (g_{1,\lambda_1T} - g_{1,\lambda_2T}), \tag{52}
\]

where

\[
B_{A_1}^A = \lim_{T \to \infty} E \left[ (\lambda_1 + \lambda_2)T g_{\lambda_1T} (g_{1,\lambda_1T} - g_{1,\lambda_2T})^\top \right] \left( E \left[ (\lambda_1 + \lambda_2)T (g_{1,\lambda_1T} - g_{1,\lambda_2T}) (g_{1,\lambda_1T} - g_{1,\lambda_2T})^\top \right] \right)^{-1}
\]

\[
= \frac{\lambda_2}{\lambda_1 + \lambda_2} \begin{bmatrix} I \\ B_{21} \end{bmatrix},
\]

where we have suppressed the \( \theta_0 \) argument in the first line. Substituting into (52) produces

\[
h_{(\lambda_1+\lambda_2)T}^A = \begin{bmatrix} g_{2,\lambda_1T} + \frac{\lambda_2}{\lambda_1 + \lambda_2} B_{21} (g_{1,\lambda_2T} - g_{1,\lambda_1T}) \\ g_{1,\lambda_1T} \end{bmatrix}, \tag{53}
\]

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 \( T \).\(^{18}\)

To construct the full adjusted-moment estimator for this case, we apply (44) again:

\[
h_{T}^{A_1} = h_{(\lambda_1+\lambda_2)T}^A - B_{A_2}^A \left( h_{(\lambda_1+\lambda_2)T}^A - g_{1,\lambda_3T} \right), \tag{55}
\]

where

\[
B_{A_2}^A = \lim_{T \to \infty} E \left[ Th_{(\lambda_1+\lambda_2)T}^A \left( h_{(\lambda_1+\lambda_2)T}^A - g_{1,\lambda_3T} \right)^\top \right] \left( E \left[ T \left( h_{(\lambda_1+\lambda_2)T}^A - g_{1,\lambda_3T} \right) \left( h_{(\lambda_1+\lambda_2)T}^A - g_{1,\lambda_3T} \right)^\top \right] \right)^{-1}
\]

\[
= \lim_{T \to \infty} E \left[ Th_{(\lambda_1+\lambda_2)T}^A \left( g_{1,\lambda_1+\lambda_2T} - g_{1,\lambda_3T} \right)^\top \right] \left( E \left[ T \left( g_{1,\lambda_1+\lambda_2T} - g_{1,\lambda_3T} \right) \left( g_{1,\lambda_1+\lambda_2T} - g_{1,\lambda_3T} \right)^\top \right] \right)^{-1}. \tag{56}
\]

It follows from Theorem 6.1 that

\[
\lim_{T \to \infty} E \left[ T \left( g_{1,\lambda_1+\lambda_2T} - g_{1,\lambda_3T} \right) \left( g_{1,\lambda_1+\lambda_2T} - g_{1,\lambda_3T} \right)^\top \right] = \left( \frac{1}{\lambda_3} + \frac{1}{\lambda_1 + \lambda_2} \right) S_{11}. \tag{57}
\]

Using (53) and the same argument,

\[
\lim_{T \to \infty} E \left[ Th_{(\lambda_1+\lambda_2)T}^A \left( g_{1,\lambda_1+\lambda_2T} - g_{1,\lambda_3T} \right)^\top \right] = \frac{1}{\lambda_1 + \lambda_2} S_{11}. \tag{58}
\]

\(^{18}\)Here and in the following computations, we make use of the equation

\[
g_{1,\lambda_1+\lambda_2T} = \frac{\lambda_1}{\lambda_1 + \lambda_2} g_{1,\lambda_1T} + \frac{\lambda_2}{\lambda_1 + \lambda_2} g_{1,\lambda_2T}. \tag{54}
\]
Finally, Theorem 6.1 and the same reasoning used to show (16) that
\[
\lim_{T \to \infty} E \left[ T h^2_{A_2,\lambda_1+\lambda_2} \left( g_{1,\lambda_1+\lambda_2} - g_{1,\lambda_3} \right) \right] = \lim_{T \to \infty} E \left[ T h^2_{A_2,\lambda_1+\lambda_2} g_{1,\lambda_1+\lambda_2}^T \right]
\]
\[
= \frac{1}{\lambda_1 + \lambda_2} S_{21}, \tag{59}
\]
where we have continued to suppressed the argument \( \theta_0 \). Combining (57), (58), and (59), and rearranging,
\[
B^{A_2} = \lambda_3 \begin{bmatrix} I \\ B_{21} \end{bmatrix}.
\]
Substituting into (55) and rearranging produces\(^\text{19}\)
\[
h_T^{A_3} = \begin{bmatrix} g_{2,\lambda_1} + (\lambda_2 + \lambda_3)B_{21} \left( \frac{\lambda_2}{\lambda_2 + \lambda_3} g_{1,\lambda_2} + \frac{\lambda_3}{\lambda_2 + \lambda_3} g_{1,\lambda_3} \right) \end{bmatrix}.
\]
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 intervals \( \lambda_2 \) and \( \lambda_3 \) were contiguous rather than separated by \( \lambda_1 \). For our asymptotic results, it does not matter whether the blocks defined by starting and ending points are contiguous.

### 6.3.2 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 3 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.

\(^\text{19}\)Here and in the following example, we use the fact that \( \lambda_1 + \lambda_2 + \lambda_3 = 1 \), and that
\[
g_{1,T} = (\lambda_1 + \lambda_2)g_{1,\lambda_1+\lambda_2} + \lambda_3 g_{1,\lambda_3}.
\]
\[
= \lambda_1 g_{1,\lambda_1} + \lambda_2 g_{1,\lambda_2} + \lambda_3 g_{1,\lambda_3}.
\]
Figure 3: 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 3, $\lambda_1$ 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 $f_1(x_1 t; \mu)$, the moment conditions observed for the last two data intervals into a vector $f_2(x_1 t; x_2 t; \mu)$, and the moment conditions observed only for the last data segment into a vector $f_3(x_1 t; x_2 t; x_3 t; \mu)$. 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 $h_T^{I_3}$ are

\[
h_T^{I_3}(\theta) = \begin{bmatrix} g_1,\lambda_1 T(\theta)^T g_2,\lambda_1 T(\theta)^T g_3,\lambda_1 T(\theta)^T g_1,\lambda_2 T(\theta)^T g_2,\lambda_2 T(\theta)^T g_1,\lambda_3 T(\theta)^T \end{bmatrix}.
\]

The results of Section 6.1 imply that $\sqrt{T}h_T^{I_3}$ has asymptotic variance

\[
S_T^{I_3} = \begin{bmatrix}
\frac{1}{\lambda_1^2}S_{11} & \frac{1}{\lambda_1^2}S_{12} & \frac{1}{\lambda_1^2}S_{13} & 0 & 0 & 0 \\
\frac{1}{\lambda_2^2}S_{21} & \frac{1}{\lambda_2^2}S_{22} & \frac{1}{\lambda_2^2}S_{23} & 0 & 0 & 0 \\
\frac{1}{\lambda_3^2}S_{31} & \frac{1}{\lambda_3^2}S_{32} & \frac{1}{\lambda_3^2}S_{33} & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{\lambda_2^2}S_{11} & \frac{1}{\lambda_2^2}S_{12} & 0 \\
0 & 0 & 0 & \frac{1}{\lambda_2^2}S_{21} & \frac{1}{\lambda_2^2}S_{22} & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{1}{\lambda_3^2}S_{11}
\end{bmatrix}.
\]

In this example the extended over-identified estimator is therefore

\[
\hat{\theta}_T^{I_3} = \arg\min_{\theta} h_T^{I_3}(\theta)^T W_T^{I_3} h_T^{I_3}(\theta),
\]

for $W_T^{I_3}$ positive definite and $W_T^{I_3} \rightarrow_{a.s.} (S_T^{I_3})^{-1}$. The estimator has asymptotic distribution

\[
\sqrt{T}(\hat{\theta}_T^{I_3} - \theta_0) \rightarrow_d N \left(0, \left((D_0^{I_3})^T (S_T^{I_3})^{-1} (D_0^{I_3})\right)^{-1}\right),
\]

46
where

\[ D^{T3} = \begin{bmatrix} D_{0,1}^T & D_{0,2}^T & D_{0,3}^T & D_{0,1}^T & D_{0,2}^T & D_{0,1}^T \end{bmatrix}^T. \]

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

\[ h_{\lambda_1 T}^{A1} = \begin{bmatrix} g_{1,\lambda_1 T} \\ g_{2,\lambda_1 T} \\ g_{3,\lambda_1 T} \end{bmatrix}. \]

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

\[ B^{A1} = \frac{\lambda_2}{\lambda_1 + \lambda_2} \begin{bmatrix} I \\ B_{3-12} \end{bmatrix} \]

and

\[ h_{(\lambda_1+\lambda_2)T}^{A2} = \begin{bmatrix} g_{1,(\lambda_1+\lambda_2)T} \\ g_{2,(\lambda_1+\lambda_2)T} \\ g_{3,\lambda_1 T} + \frac{\lambda_2}{\lambda_1 + \lambda_2} B_{3-12} \begin{bmatrix} g_{1,\lambda_2 T} - g_{1,\lambda_1 T} \\ g_{2,\lambda_2 T} - g_{2,\lambda_1 T} \end{bmatrix} \end{bmatrix}, \]

(60)

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

\[ B_{3-12} = [S_{31} \ S_{32}] \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}^{-1}. \]

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

\[ h_T^{A3} = h_{(\lambda_1+\lambda_2)T}^{A2} - B^{A2} \left( h_{1,(\lambda_1+\lambda_2)T}^{A2} - g_{1,\lambda_3 T} \right), \]

(61)

where the expression for \( B^{A2} \) is given by (56). It follows from Theorem 6.1 that

\[
\lim_{T \to \infty} E \left[ T \left( h_{1,(\lambda_1+\lambda_2)T}^{A2} - g_{1,\lambda_3 T} \right) \left( h_{1,(\lambda_1+\lambda_2)T}^{A2} - g_{1,\lambda_3 T} \right)^\top \right]^{-1} = \left( \frac{1}{\lambda_3} + \frac{1}{\lambda_1 + \lambda_2} \right) S_{11}.
\]

Similar reasoning shows that

\[
\lim_{T \to \infty} E \left[ T h_{j,(\lambda_1+\lambda_2)T}^{A2} \left( h_{1,(\lambda_1+\lambda_2)T}^{A2} - g_{1,\lambda_3 T} \right)^\top \right] = \frac{1}{\lambda_1 + \lambda_2} S_{j1}, \quad j = 1, 2, 3,
\]

where we have made use of (16) for \( j = 3 \). Substituting into (61), and applying footnote 19 results in

\[ h_T^{A3} = \begin{bmatrix} g_{1T} \\ g_{2,(\lambda_1+\lambda_2)T} + \lambda_3 B_{21} (g_{1,\lambda_3 T} - g_{1,(\lambda_1+\lambda_2)T}) \\ g_{3,\lambda_1 T} + \frac{\lambda_2}{\lambda_1 + \lambda_2} B_{3-12} \begin{bmatrix} g_{1,\lambda_2 T} - g_{1,\lambda_1 T} \\ g_{2,\lambda_2 T} - g_{2,\lambda_1 T} \end{bmatrix} + \lambda_3 B_{31} (g_{1,\lambda_3 T} - g_{1,(\lambda_1+\lambda_2)T}) \end{bmatrix}, \]

47
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 $\lambda_1$ and $\lambda_2$ 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.

### 6.3.3 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 4 illustrates the pattern of missing data in this example.

![Figure 4: 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 $T$.](image)

Figure 4: 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 $T$.

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 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 intervals into a vector $f_1(x_{1t}, \theta)$ and the moment conditions observed for the last two intervals 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 $h_{I^3}^T$ are

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

Then the results in Section 6.1 imply that $\sqrt{T}h_{I^3}^T(\theta)$ has asymptotic variance

$$
S_{I^3} = \begin{bmatrix}
\frac{1}{\lambda_1} S_{11} & \frac{1}{\lambda_1} S_{12} & 0 & 0 \\
\frac{1}{\lambda_1} S_{21} & \frac{1}{\lambda_1} S_{22} & 0 & 0 \\
0 & 0 & \frac{1}{\lambda_2} S_{11} & 0 \\
0 & 0 & 0 & \frac{1}{\lambda_3} S_{22}
\end{bmatrix}.
$$

The extended over-identified estimator is therefore

$$
\hat{\theta}_{I^3}^T = \arg\min_\theta h_{I^3}^T(\theta)^\top W_{I^3}^T h_{I^3}^T(\theta)
$$

for $W_{I^3}^T$ positive definite and $W_{I^3}^T \to_{a.s.} (S_{I^3})^{-1}$. The asymptotic distribution is

$$
\sqrt{T}(\hat{\theta}_{I^3}^T - \theta_0) \to_d N \left(0, \left((D_{I^3}^T)^\top \ (S_{I^3})^{-1} \ (D_{I^3}^T)\right)^{-1}\right),
$$

where

$$
D_{I^3} = \begin{bmatrix}
D_{0,1}^T & D_{0,1}^T & D_{0,2}^T
\end{bmatrix}^\top.
$$

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

$$
h_{A^2}^{I^3 T} = \left[ g_{1,\lambda_1 + \lambda_2} T \
g_{2,\lambda_1 T} + \frac{\lambda_2}{\lambda_1 + \lambda_2} B_{21} (g_{1,\lambda_2 T} - g_{1,\lambda_1 T})\right].
$$

We have

$$
h_{A^3}^{I^3 T} = h_{A^2}^{I^3 T} - B^{A^2} \left( h_{A^2}^{I^3 T} - g_{2,\lambda_3 T} \right),
$$

where

$$
B^{A^2} = \lim_{T \to \infty} E \left[ T h_{A^2}^{I^3 T} \left( h_{A^2}^{I^3 T} - g_{2,\lambda_3 T} \right)^\top \right]
$$

$$
\times E \left( \left[ T \left( h_{A^2}^{I^3 T} - g_{2,\lambda_3 T} \right) \left( h_{A^2}^{I^3 T} - g_{2,\lambda_3 T} \right)^\top \right]^{-1} \right),
$$

where we have suppressed the argument $\theta_0$. Define

$$
\gamma = \frac{\lambda_2}{\lambda_1 + \lambda_2} \frac{\lambda_3}{\lambda_1 + \lambda_3}.
$$
Standard arguments (given in the Appendix) show that
\[
B^{A_2} = \frac{\lambda_3}{\lambda_1 + \lambda_3} \left[ S_{22} - \frac{\lambda_1}{\lambda_1 + \lambda_2} S_{12} \right] S_{22}^{-1} (I - \gamma B_{21} B_{12})^{-1}.
\] (64)

Given \(B^{A_2}\), (63) gives the moments for the adjusted-moment estimator. The first component is as follows:
\[
h_{1T}^{A_2} = g_{1, (\lambda_1 + \lambda_2)T} + \frac{\lambda_1}{\lambda_1 + \lambda_2} \left( I - \gamma B_{21} B_{12} \right)^{-1} \left( g_{2, \lambda_2 T} - g_{2, \lambda_1 T} - \frac{\lambda_2}{\lambda_1 + \lambda_2} B_{21} (g_{1, \lambda_2 T} - g_{1, \lambda_1 T}) \right),
\] (65)
while more extensive matrix algebra results in the following expression for the second component:
\[
h_{2T}^{A_2} = g_{2, (\lambda_1 + \lambda_3)T} + \frac{\lambda_1}{\lambda_1 + \lambda_3} \left( I - \gamma B_{21} B_{12} \right)^{-1} \left( g_{1, \lambda_2 T} - g_{1, \lambda_1 T} - \frac{\lambda_3}{\lambda_1 + \lambda_3} B_{12} (g_{2, \lambda_3 T} - g_{2, \lambda_1 T}) \right).
\] (66)

Because
\[
B_{12} \left( I - \gamma B_{21} B_{12} \right)^{-1} = (I - \gamma B_{12} B_{21})^{-1} B_{12},
\] (67)
these expressions are symmetric.\[^{20}\]

At first glance, the adjustments implicit in (65) and (66) do not seem as intuitive as their counterparts in Section 2, or, for that matter, in Sections 6.3.1 and 6.3.2. However, there is a reason for the apparently strange form. It follows from (68) and (54) that (65) can be rewritten as
\[
h_{1T}^{A_3} = \frac{\lambda_2}{\lambda_1 + \lambda_2} g_{1, \lambda_2 T} + \frac{\lambda_1}{\lambda_1 + \lambda_3} \left( \sum_{m=0}^{\infty} (\gamma B_{21} B_{12})^m \right) \left( g_{2, \lambda_2 T} - g_{2, \lambda_1 T} - \frac{\lambda_2}{\lambda_1 + \lambda_2} B_{21} (g_{1, \lambda_2 T} - g_{1, \lambda_1 T}) \right).
\] (69)

It is instructive to expand out the infinite sum explicitly:
\[
h_{1T}^{A_3} = \frac{\lambda_2}{\lambda_1 + \lambda_2} g_{1, \lambda_2 T} + \frac{\lambda_1}{\lambda_1 + \lambda_3} \left( \sum_{m=0}^{\infty} (\gamma B_{21} B_{12})^m \right) \left( g_{2, \lambda_2 T} - g_{2, \lambda_1 T} - \frac{\lambda_2}{\lambda_1 + \lambda_2} B_{21} (g_{1, \lambda_2 T} - g_{1, \lambda_1 T}) \right)
- \frac{\lambda_1}{\lambda_1 + \lambda_3} B_{12} \frac{\lambda_2}{\lambda_1 + \lambda_2} B_{21} (g_{1, \lambda_2 T} - g_{1, \lambda_1 T})
+ \frac{\lambda_1}{\lambda_1 + \lambda_3} B_{12} \frac{\lambda_2}{\lambda_1 + \lambda_2} B_{21} \frac{\lambda_3}{\lambda_1 + \lambda_3} B_{12} (g_{2, \lambda_3 T} - g_{2, \lambda_1 T}) - \ldots
\] (70)

\[^{20}\]Equation (67) can be shown by noting that
\[
(I - \gamma B_{21} B_{12})^{-1} = \sum_{m=0}^{\infty} \gamma^m (B_{21} B_{12})^m.
\] (68)
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, $\frac{\lambda_3}{\lambda_1 + \lambda_3} B_{12}(g_{2,\lambda_1 T} - g_{2,\lambda_1 T})$, 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 (70), which, by (68), converges to the extended adjusted-moment estimator (65). A symmetric explanation holds for (66). Thus even in this complicated problem, the adjusted-moment estimator produces moments that have intuitive appeal.

7 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, the impact of including the non-overlapping portion of the data is not limited to parameters which impact the longer data series. 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 will be an impact on all the parameters. This interaction can be through
covariances between the moment conditions, or through the fact that some parameters appear in both the long-sample and short-sample moment conditions. In an application of our methods to estimation of asset pricing models in international data, we show that this impact can be large.

Our two estimators are as straightforward to implement as standard GMM and have intuitive interpretations. The adjusted-moment estimator calculates moments using all the data available for each series, and then adjusts the moments available over the shorter series using 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  Proofs for Section 1

Proof of Lemma 1.1:

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

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

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

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

and

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

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

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

$$= \lim_{T \to \infty} P\left(\left[ \frac{1}{\sqrt{T}} \sum_{t=\lfloor \sqrt{T} \rfloor + 1}^{T} \mu w_t < c \right] P(F) \right)$$

$$= P_0 P(F),$$

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

Proof of Theorem 1.1:

Assumptions 1–4 imply that

$$\sqrt{(1 - \lambda)T} g_{(1 - \lambda)T}(\theta_0) \to_d N(0, S) \quad (71)$$

53
and
\[
\sqrt{\lambda T} g_{\lambda T}(\theta_0) \to_d N(0, S)
\]  
(72)
by Theorem 2.4 of White and Domowitz (1984). Stationarity of \(x_t\) (Assumption 1) implies that random variables \(f(x_{-(1-\lambda)T+1}, \theta), \ldots, f(x_{\lambda T}, \theta)\) have the same joint distribution as random variables \(f(x_1, \theta), \ldots, f(x_T, \theta)\). Thus partial sums taken over \(f(x_{-(1-\lambda)T+1}, \theta), \ldots, f(x_{\lambda T}, \theta)\) have the same distribution as the corresponding partial sums taken over \(f(x_1, \theta), \ldots, f(x_T, \theta)\). Define
\[
\tilde{g}_{\lambda T}(\theta) = \frac{1}{\lambda T} \sum_{t=1}^{\lambda T} f(x_t, \theta)
\]
\[
\tilde{g}_{(1-\lambda)T}(\theta) = \frac{1}{(1-\lambda)T} \sum_{t=0}^{(1-\lambda)T-1} f(x_{-t}, \theta).
\]
It suffices to prove the results for \(\tilde{g}_{\lambda T}\) and \(\tilde{g}_{(1-\lambda)T}\).

Let \(N(c)\) denote the cumulative distribution function of the standard normal distribution evaluated at \(c\). Let \(\mu_1\) and \(\mu_2\) be \(1 \times l\) vectors such that \(\mu_1 \mu_1^T = \mu_2 \mu_2^T = 1\). By Lemma 1.1,
\[
\lim_{T \to \infty} P \left( \mu_1 \sqrt{(1-\lambda)T} S^{-1} \tilde{g}_{(1-\lambda)T}(\theta_0) < c_1, \mu_2 \sqrt{\lambda T} S^{-1} \tilde{g}_{\lambda T}(\theta_0) < c_2 \right) =
\lim_{T \to \infty} P \left( \mu_1 \sqrt{(1-\lambda)T} S^{-1} \tilde{g}_{(1-\lambda)T}(\theta_0) < c_1 \right) \lim_{T \to \infty} \left( \mu_2 \sqrt{\lambda T} S^{-1} \tilde{g}_{\lambda T}(\theta_0) < c_2 \right) = N(c_1)N(c_2)
\]
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 (71) and (72).

\[\square\]

**B Proofs for Section 2**

\[\text{Proof of Theorem 2.2:}\]

White and Domowitz (1984) show that under these assumptions
\[
|g_{\lambda T}(\theta) - Ef(x_t, \theta)| \to_{a.s.} 0
\]
\[
|g_{(1-\lambda)T}(\theta) - Ef(x_t, \theta)| \to_{a.s.} 0
\]
as \(T \to \infty\) uniformly in \(\theta \in \Theta\). By the continuous mapping theorem,
\[
h_T^k(\theta)^\top W_T^k h_T^k(\theta) \to_{a.s.} E[f(x_t, \theta)]^\top W^k E[f(x_t, \theta)]
\]

54
for \( k \in \{S, L, A\} \), and

\[
h_T^k(\theta)^\top W_T^k h_T^k(\theta) \overset{\text{a.s.}}{\to} \mathbb{E} [f_1(x_{1t}, \theta)^\top f(x_t, \theta)^\top] W_T^E \begin{bmatrix} f_1(x_{1t}, \theta) \\ f(x_t, \theta) \end{bmatrix}
\]

uniformly in \( \theta \). The result then follows from Amemiya (1985, Theorem 4.1.1).

**Proof of Theorem 2.3:** Define

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

for \( \theta \) in the interior of \( \Theta \). For \( T \) sufficiently large, \( \hat{\theta}_T^k \) lies in the interior of \( \Theta \). By the mean value theorem, there exists a \( \tilde{\theta}^k \) in the segment between \( \theta_0 \) 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 \( D_T^k(\hat{\theta}_T^k)^\top W_T^k \):

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

By the first-order condition of the optimization problem,

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

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

\[
D_T^k(\theta) \overset{\text{a.s.}}{\to} \mathbb{E} \begin{bmatrix} \frac{\partial f_1}{\partial \theta}(x_{1t}, \theta) \\ \frac{\partial f}{\partial \theta}(x_t, \theta) \end{bmatrix}
\]

for \( k \in \{S, L, A\} \), and

\[
D_T^k(\theta) \overset{\text{a.s.}}{\to} \mathbb{E} \begin{bmatrix} \frac{\partial f_1}{\partial \theta}(x_{1t}, \theta) \\ \frac{\partial f}{\partial \theta}(x_t, \theta) \end{bmatrix}
\]

uniformly in \( \theta \). Therefore by Theorem 2.2 and Assumptions 7 and 8, Amemiya (1985, Theorem 4.1.5) implies

\[
D_T^k(\hat{\theta}_T^k) \overset{\text{a.s.}}{\to} D_0^k
\]

\[
D_T^k(\tilde{\theta}_T^k) \overset{\text{a.s.}}{\to} D_0^k
\]

\[
W_T^k \overset{\text{a.s.}}{\to} W^k
\]

The result follows from the Slutsky Theorem.
C Proofs for Section 3

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

Proof: See Goldberger (1964, Chapter 2.7). \(\Box\)

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

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

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

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

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

Lemma C.3 Let

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

be a symmetric invertible matrix. Then

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

where $\Sigma$ is defined by (5). 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

$$\bar{S}^{-1} = \begin{bmatrix}
\frac{1}{\lambda} S_{11}^{-1} + B_{21}^\top \Sigma^{-1} B_{12} & -B_{21}^\top \Sigma^{-1} \\
-\Sigma^{-1} B_{12} & \Sigma^{-1}
\end{bmatrix}. \tag{77}$$
\[-(\eta_1 + \cdots + \eta_{n-1})T + 1 \quad -(\eta_2 + \cdots + \eta_{n-1})T + 1 \quad 1 \quad \eta_nT\]

Figure 5: Numbering scheme used in the proof of Theorem 6.1

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

\[
\tilde{S}^{-1} = \begin{bmatrix}
\tilde{S}^{-1}_{11} + \tilde{B}_{21}^T\Sigma^{-1}\tilde{B}_{21} & -\tilde{B}_{21}^T\Sigma^{-1} \\
-\Sigma^{-1}\tilde{B}_{12} & \Sigma^{-1}
\end{bmatrix},
\]

where

\[
\tilde{B}_{21} = \tilde{S}_{21} (\tilde{S}_{11})^{-1} = S_{21}S_{11}^{-1} = B_{21},
\]

and

\[
\Sigma = \tilde{S}_{22} - \tilde{S}_{21}\tilde{S}_{11}^{-1}\tilde{S}_{12} = S_{22} - (1 - \lambda)S_{21}S_{11}^{-1}S_{12} - \lambda S_{21}S_{11}^{-1}S_{12} = \Sigma.
\]

Therefore (77) holds. \(\Box\)

## D Proofs for Section 6

**Proof of Theorem 6.1:** We proceed by induction. The case for \(n = 1\) follows from standard results (e.g. White and Domowitz (1984, Theorem 2.4)). Suppose (34) 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

\[
\tilde{g}_{nT}(\theta_0) = \frac{1}{\eta_nT} \sum_{t=1}^{\eta_nT} f(x_t, \theta);
\]

while

\[
\tilde{g}_{jT}(\theta_0) = \sum_{-(\eta_j + \cdots + \eta_{n-1})T+1}^{-(\eta_{j+1} + \cdots + \eta_{n-1})T} f(x_t, \theta), \quad j = 1, \ldots, n - 1.
\]

As Figure 5 illustrates, the start data of new “sample” is \(-(\eta_1 + \cdots + \eta_{n-1})T + 1\) while the end date is \(\eta_nT\). Then \(\tilde{g}_{nT}(\theta_0), \ldots, \tilde{g}_{nT}(\theta_0)\) have the same joint distribution as \(g_{nT}(\theta_0), \ldots, g_{nT}(\theta_0)\).
By Lemma 1.1, for any $1 \times l$ vectors $\mu_1, \ldots, \mu_n$ such that $\mu_j\mu_j^\top = 1$, and scalars $c_1, \ldots, c_n$,

$$\lim_{T \to \infty} P \left( \sqrt{\eta_nT} \mu_n S^{-1} \tilde{g}_{\eta_nT}(\theta_0) < c_n, \sqrt{\eta_{n-1}T} \mu_{n-1} S^{-1} \tilde{g}_{\eta_{n-1}T}(\theta_0) < c_{n-1}, \ldots, \sqrt{\eta_1T} \mu_1 S^{-1} \tilde{g}_{\eta_1T}(\theta_0) < c_1 \right) = \lim_{T \to \infty} P \left( \sqrt{\eta_nT} \mu_n S^{-1} \tilde{g}_{\eta_nT}(\theta_0) < c_n \right) \times \lim_{T \to \infty} P \left( \sqrt{\eta_{n-1}T} \mu_{n-1} S^{-1} \tilde{g}_{\eta_{n-1}T}(\theta_0) < c_{n-1}, \ldots, \sqrt{\eta_1T} \mu_1 S^{-1} \tilde{g}_{\eta_1T}(\theta_0) < c_1 \right).$$

The result then follows from the induction assumption and asymptotic normality of $\sqrt{\eta_nT} \tilde{g}_{\eta_nT}$. □

**Proof of Theorem 6.6:**

It suffices to compare the asymptotic variance of $\sqrt{T} \hat{\theta}_T^n$ with $\sqrt{T} \theta_{(1-\lambda_n)T}$. By Theorem 6.5,

$$E \left[ (1 - \lambda_n)T \left( \hat{\theta}_{(1-\lambda_n)T} - \theta_0 \right) \left( \hat{\theta}_{(1-\lambda_n)T} - \theta_0 \right)^\top \right] = \left( (D_{0n}^\top)^{-1} (S_{n-1}^T)^{-1} (D_{0n}^T) \right)^{-1},$$

where

$$S_{n-1}^T = \begin{bmatrix}
\frac{1}{1-\lambda_n} S_{\phi_1} & 0 & \cdots & 0 \\
0 & \frac{1}{1-\lambda_n} S_{\phi_2} & \cdots & 0 \\
0 & 0 & \ddots & 0 \\
0 & 0 & 0 & \frac{1-\lambda_n}{\lambda_n-1} S_{\phi_{n-1}}
\end{bmatrix},$$

because data segment $\lambda_j$ 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,

$$E \left[ T \left( \hat{\theta}_{(1-\lambda_n)T} - \theta_0 \right) \left( \hat{\theta}_{(1-\lambda_n)T} - \theta_0 \right)^\top \right] = \frac{1}{1 - \lambda_n} \left( (D_{0n}^T)^{-1} (S_{n-1}^T)^{-1} (D_{0n}^\top)^{-1} \right)$$

$$= \left( \begin{bmatrix}
\frac{1}{1-\lambda_n} S_{\phi_1} & 0 & \cdots & 0 \\
0 & \frac{1}{1-\lambda_n} S_{\phi_2} & \cdots & 0 \\
0 & 0 & \ddots & 0 \\
0 & 0 & 0 & \frac{1-\lambda_n}{\lambda_n-1} S_{\phi_{n-1}}
\end{bmatrix} \right)^{-1} \left( \begin{bmatrix}
0 \\
0 \\
0 \\
\frac{1-\lambda_n}{\lambda_n-1} S_{\phi_{n-1}}
\end{bmatrix}
\right)$$

By (78) and Lemma C.1, it suffices to show that

$$(D_{0n}^T)^{-1} (S_{n}^T)^{-1} (D_{0n}^\top) - \frac{1}{1 - \lambda_n} (D_{0n}^T)^{-1} (S_{n-1}^T)^{-1} (D_{0n}^\top)$$

is positive semi-definite. Applying (78), we have

$$S_n = \begin{bmatrix}
\frac{1}{1-\lambda_n} S_{n-1} & 0 \\
0 & \frac{1}{\lambda_n} S_{\phi_n}
\end{bmatrix}$$

and

$$D_n = \begin{bmatrix}
(D_n^T)^\top & D_{0, \phi_n}^\top
\end{bmatrix}^\top.$$
Therefore

\[(D_0^{T_n})^\top (S_t^{n-1})^{-1} (D_0^{T_n}) - (1 - \lambda_n)(D_0^{T_{n-1}})^\top (S_t^{n-1})^{-1} (D_0^{T_{n-1}}) = \lambda_n D_0^{\top \phi_n} S_{\phi_n}^{-1} D_0^{\phi_n},\]  

which is positive semi-definite.

**Proof of Theorem 6.9:**

We show by induction on \( n \) that

\[
\frac{\partial h_{A_n}^{\lambda}}{\partial \theta}(\theta_0) \to_{\text{a.s.}} E \left[ \frac{\partial f}{\partial \theta}(x_t, \theta) \right] = D_0. \tag{80}
\]

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

\[
\frac{\partial h_{A_1}^{A_1 T}}{\partial \theta}(\theta_0) = \frac{\partial g_{A_1 T}}{\partial \theta}(\theta_0) \to_{\text{a.s.}} D_0.
\]

Assume (80) holds for \( n - 1 \). By (44),

\[
h_{A_n}^{\lambda T}(\theta) = h_{A_{n-1}}^{A_{n-1} T}(\theta) - B_{A_{n-1}} \left( h_{A_{n-1}}^{A_{n-1} T}(\theta) - g_{\phi_n (1 - \lambda_n) T}(\theta) \right). \tag{81}
\]

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 (81) 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. \( \square \)

**Proof of Theorem 6.11:**

By Theorem 6.10, it suffices to show that the asymptotic variance of \( \sqrt{T} \hat{\theta}_T^{A_n} \) is the same as the asymptotic variance of \( \sqrt{T} \hat{\theta}_T^{A_n} \). The proof is by induction on \( n \). For \( n = 1 \),

\[
\hat{\theta}_T^{A_1} = \theta_{A_1 T}
\]

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

\[
D_0^\top (S_t^{A_{n-1}})^{-1} D_0 = \left( D_0^{T_{n-1}} \right)^\top (S_t^{n-1})^{-1} D_0^{T_{n-1}}.
\]

Without loss of generality, let \( \phi_n = \{1, \ldots, \pi_n\} \). That is, the first \( \pi_n \) moment conditions are observed over data region \( \lambda_n \). By (79) it suffices to show

\[
D_0^\top (S_t^{A_n})^{-1} D_0 - (1 - \lambda_n)D_0^\top (S_t^{A_{n-1}})^{-1} D_0 = \lambda_n D_0^\top \phi_n S_{\phi_n}^{-1} D_0^{\phi_n}.
\]
Equivalently, it suffices to show

\[(S\bar{A}_n)^{-1} = (1 - \lambda_n) (S\bar{A}_{n-1})^{-1} + \lambda_n \begin{bmatrix} S_{\bar{\phi}_n}^{-1} & 0 \\ 0 & 0 \end{bmatrix}. \] (82)

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

\[S\bar{A}_n = \begin{bmatrix} S_{\phi_n}^{\bar{A}_n} & S_{\phi_n, \bar{\phi}_n}^{\bar{A}_n} \\ S_{\bar{\phi}_n, \phi_n}^{\bar{A}_n} & S_{\bar{\phi}_n}^{\bar{A}_n} \end{bmatrix},\]

where

\[S_{\bar{\phi}_n}^{\bar{A}_n} = E \begin{bmatrix} T h_{\bar{\phi}_n, T} (\theta_0) h_{\bar{\phi}_n, T} (\theta_0)^\top \end{bmatrix},\]

\[S_{\bar{\phi}_n, \phi_n}^{\bar{A}_n} = E \begin{bmatrix} T h_{\bar{\phi}_n, T} (\theta_0) h_{\phi_n, T} (\theta_0)^\top \end{bmatrix},\]

and $S_{\phi_n, \bar{\phi}_n}^{\bar{A}_n} = (S_{\phi_n, \bar{\phi}_n}^{\bar{A}_n})^\top$. Note that under this ordering,

\[B_{n-1}^{A_n} = \frac{1}{1 - \lambda_n} \begin{bmatrix} S_{\phi_n}^{A_{n-1}} & S_{\bar{\phi}_n}^{A_{n-1}} \\ S_{\bar{\phi}_n, \phi_n}^{A_{n-1}} & S_{\phi_n}^{A_{n-1}} \end{bmatrix} \left[ \frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1 - \lambda_n} S_{\bar{\phi}_n}^{A_{n-1}} \right]^{-1}. \]

Analogously to $B_{21}$ in Section 2, define

\[B_{21}^{A_n} = S_{\phi_n, \bar{\phi}_n}^{\bar{A}_n} \left( S_{\phi_n}^{A_n} \right)^{-1}, \] (83)

\[B_{21}^{A_{n-1}} = S_{\phi_n, \bar{\phi}_n}^{\bar{A}_{n-1}} \left( S_{\phi_n}^{A_{n-1}} \right)^{-1}. \] (84)

Analogously to $\Sigma$ in Section 2, define

\[\Sigma_{\phi_n}^{A_n} = S_{-\phi_n, -\phi_n}^{\bar{A}_n} - S_{-\phi_n, \bar{\phi}_n}^{\bar{A}_n} \left( S_{\phi_n}^{A_n} \right)^{-1} S_{\phi_n, -\phi_n}^{\bar{A}_n}. \] (86)

\[\Sigma_{\phi_n}^{A_{n-1}} = S_{-\phi_n, -\phi_n}^{\bar{A}_{n-1}} - S_{-\phi_n, \bar{\phi}_n}^{\bar{A}_{n-1}} \left( S_{\phi_n}^{A_{n-1}} \right)^{-1} S_{\phi_n, -\phi_n}^{\bar{A}_{n-1}}. \] (87)

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

\[\left( S_{\phi_n}^{A_n} \right)^{-1} = (1 - \lambda_n) \left( S_{\phi_n}^{A_{n-1}} \right)^{-1} + \lambda_n (S_{\phi_n}^{-1}). \] (88)

\[B_{21}^{A_n} = B_{21}^{A_{n-1}}. \] (89)

\[\Sigma^{A_n} = \frac{1}{1 - \lambda_n} \Sigma^{A_{n-1}}. \] (90)
We first show (88). Equation (47) implies
\[
S_{\phi_n}^A = \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}} - \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}} \left[ \frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}} \right]^{-1} \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}}.
\]
Pre-multiplying by \((1 - \lambda_n) \left(S_{\phi_n}^{A_{n-1}}\right)^{-1}\) yields
\[
(1 - \lambda_n) \left(S_{\phi_n}^{A_{n-1}}\right)^{-1} S_{\phi_n}^A = I - \left[ \frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}} \right]^{-1} \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}} = \left[ \frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}} \right]^{-1} \frac{1}{\lambda_n} S_{\phi_n}.
\]
Taking inverses yields
\[
\frac{1}{1 - \lambda_n} \left(S_{\phi_n}^{A_{n-1}}\right)^{-1} S_{\phi_n}^A = \lambda_n (S_{\phi_n})^{-1} \left[ \frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}} \right] = I + \frac{\lambda_n}{1 - \lambda_n} (S_{\phi_n})^{-1} S_{\phi_n}^{A_{n-1}}.
\]
Post-multiplying by \((1 - \lambda_n) \left(S_{\phi_n}^{A_{n-1}}\right)^{-1}\) yields (88).

We now show (89). By (47),
\[
S_{\phi_n}^{A_{n-1}} = \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}} - \frac{1}{1 - \lambda_n} \left[ \frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}} \right]^{-1} \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}}.
\]
Post-multiplying by \(S_{\phi_n}^{A_{n-1}}\) and applying (88) produces
\[
B_{21}^{A_{n}} = S_{\phi_n}^{\phi_{n-1}} \left(S_{\phi_n}^{A_{n}}\right)^{-1} = \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}} \left[ (1 - \lambda_n) \left(S_{\phi_n}^{A_{n-1}}\right)^{-1} + \lambda_n S_{\phi_n}^{-1} \right] - \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}} \left[ \frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}} \right]^{-1} \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}} \left[ (1 - \lambda_n) \left(S_{\phi_n}^{A_{n-1}}\right)^{-1} + \lambda_n S_{\phi_n}^{-1} \right].
\]
Expanding out the first term on the right hand side and multiplying through by \(S_{\phi_n}^{A_{n-1}}\) in the second term produces
\[
B_{21}^{A_{n}} = B_{21}^{A_{n-1}} + \frac{\lambda_n}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}} S_{\phi_n}^{-1} - \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}} \left[ \frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}} \right]^{-1} \left[ I + \frac{\lambda_n}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}} S_{\phi_n}^{-1} \right].
\]
Factoring out \(\lambda_n S_{\phi_n}^{-1}\) in the last term yields (89).

Lastly, we show (90). Equation (47) implies
\[
S_{\phi_n}^{\phi_{n-1}} = \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}} - \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}} \left[ \frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}} \right]^{-1} \frac{1}{1 - \lambda_n} S_{\phi_n}^{A_{n-1}}. \tag{91}
\]
Therefore,\[\begin{align*}
S_{\phi_n, \phi_n}^{-1} S_{\phi_n, \phi_n}^{-1} - S_{\phi_n, \phi_n}^{-1} S_{\phi_n, \phi_n}^{-1} & = S_{\phi_n, \phi_n}^{-1} - S_{\phi_n, \phi_n}^{-1} \left[ (1 - \lambda_n) S_{\phi_n, \phi_n}^{-1} + \lambda_n S_{\phi_n, \phi_n}^{-1} \right]^{-1} S_{\phi_n, \phi_n}^{-1} S_{\phi_n, \phi_n}^{-1} \\
& = S_{\phi_n, \phi_n}^{-1} \left[ (1 - \lambda_n) S_{\phi_n, \phi_n}^{-1} + \lambda_n S_{\phi_n, \phi_n}^{-1} \right]^{-1} S_{\phi_n, \phi_n}^{-1} S_{\phi_n, \phi_n}^{-1}.
\end{align*}\]Substituting in (91) and (92) into (86),
\[\Sigma^{A_n} = \frac{1}{1 - \lambda_n} \sum_{\phi_n} - \frac{1}{1 - \lambda_n} S_{\phi_n, \phi_n}^{-1} \left[ \frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1 - \lambda_n} S_{\phi_n, \phi_n}^{-1} \right]^{-1} \left[ \frac{1}{1 - \lambda_n} I + \frac{1}{\lambda_n} S_{\phi_n} \left( S_{\phi_n, \phi_n}^{-1} \right)^{-1} S_{\phi_n, \phi_n}^{-1} \right].\]

Which implies
\[\Sigma^{A_n} = \frac{1}{1 - \lambda_n} \sum_{\phi_n} - \frac{1}{1 - \lambda_n} S_{\phi_n, \phi_n}^{-1} \left[ \frac{1}{\lambda_n} S_{\phi_n} + \frac{1}{1 - \lambda_n} S_{\phi_n, \phi_n}^{-1} \right]^{-1} \left[ \frac{1}{1 - \lambda_n} I + \frac{1}{\lambda_n} S_{\phi_n} \left( S_{\phi_n, \phi_n}^{-1} \right)^{-1} S_{\phi_n, \phi_n}^{-1} \right].\]

which shows (90). \(\square\)

**Derivation of (64):**

From Theorem 6.1 and (62), it follows that
\[\lim_{T \rightarrow \infty} E \left[ T \left( h_{A_2, (\lambda_1 + \lambda_2) T} - g_{2, \lambda_3 T} \right) \left( h_{A_2, (\lambda_1 + \lambda_2) T} - g_{2, \lambda_3 T} \right)^\top \right] = \left( \frac{1}{\lambda_3} + \frac{1}{\lambda_1} \right) S_{22} + \left( \frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^2 \left( \frac{1}{\lambda_2} + \frac{1}{\lambda_1} \right) B_{21} S_{11} B_{21}^\top - \frac{1}{\lambda_1 + \lambda_2} \left( S_{21} B_{21}^\top + B_{21} S_{12} \right)\]
\[= \left( \frac{1}{\lambda_3} + \frac{1}{\lambda_1} \right) S_{22} - \frac{\lambda_2}{\lambda_1 + \lambda_2} S_{21} S_{11} S_{12}^{-1} S_{12}, \]
and
\[\lim_{T \rightarrow \infty} E \left[ T h_{A_2, (\lambda_1 + \lambda_2) T} \left( h_{A_2, (\lambda_1 + \lambda_2) T} - g_{2, \lambda_3 T} \right) \left( h_{A_2, (\lambda_1 + \lambda_2) T} - g_{2, \lambda_3 T} \right)^\top \right] = \lim_{T \rightarrow \infty} E \left[ T g_{1, (\lambda_1 + \lambda_2) T} \left( h_{A_2, (\lambda_1 + \lambda_2) T} \right) \left( h_{A_2, (\lambda_1 + \lambda_2) T} \right)^\top \right] = \frac{1}{\lambda_1 + \lambda_2} S_{12}.\]
where we have applied the reasoning of (16). (94) multiplied by the inverse of (93) equals the first component of $B^A$. The second component equals

$$
\lim_{T \to \infty} E \left[ T h^A_{2, (\lambda_1 + \lambda_2) T} \left( h^A_{2, (\lambda_1 + \lambda_2) T} - g_2, \gamma T \right)^T \right] = \lim_{T \to \infty} E \left[ T h^A_{2, (\lambda_1 + \lambda_2) T} \left( h^A_{2, (\lambda_1 + \lambda_2) T} \right)^T \right] = \frac{1}{\lambda_1} \left( S_{22} - \frac{\lambda_2}{\lambda_1 + \lambda_2} S_{21} S_{11}^{-1} S_{12} \right), \quad (95)
$$

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

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

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

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

\qed
References


