Contents

1 Proof of Claim 2

Claim. *Given a fixed setting of the parameters θ there exists some $0 \leq \lambda < 1$ such that for each M-statistic the difference between the true value and the value as approximated by the ASOS procedure can be expressed as a linear function of the approximation error in the DAS whose operator norm is bounded above by ${ck_{lim}}^2 \lambda^{k_{lim}-1}$ for some constant c that doesn't depend on k_{lim}.*

Proof. (sketch) We will use the Δ symbol to denote the error in a given term. So for example, $\Delta(y, x^*)_{k_{lim}+1}$ will denote error due to in the first DAS.

By repeated application of the first 2nd-order recursion we have that $\Delta(y, x^*)_k = \Delta(y, x^*)_{k_{lim}+1}(H')^{k_{lim}+1-k}$. Then by repeated application of the second 2nd-order recursion we have that $\Delta(x^*, y)_k = H^k \Delta(y, x^*)'_0 = H^{k+k_{lim}+1} {\Delta(y, x^*)_{k_{lim}+1}}'$. We can already see a pattern starting to emerge here. The error for statistics of small k-value/time-lag is given by the approximation error in the DAS, multiplied by some large power of H. If H is "small" in some sense then the error will decay exponentially as k decreases. This will be made more rigorous later.

Repeatedly applying the third and fifth 2nd-order recursions we also have:

$$\Delta(x^*, x^*)_k = \Delta(x^*, x^*)_{k_{lim}}(H')^{k_{lim}-k} + \sum_{i=k}^{k_{lim}} \Delta(x^*, y)_i K'(H')^{i-k}$$

$$\Delta(x^T, y)_k = J^{k_{lim}-k} \Delta(x^T, y)_{k_{lim}} + \sum_{i=k}^{k_{lim}} J^{i-k} P \Delta(x^*, y)_i$$

If we plug in the previously derived error formulae into these (for the terms $\Delta(x^*, y)_i$) we note that at each term being summed is multiplied by H and/or J a total of $k_{lim} - k$ times or more. Repeated application of the sixth 2nd-order equation gives:

$$\Delta(x^T, x^*)_k = J^{k_{lim}-k}\Delta(x^T, x^*)_{k_{lim}} + \sum_{i=k}^{k_{lim}} J^{i-k} P \Delta(x^*, x^*)_i$$

The right hand side of this equation contains terms of the form $\Delta(x^*, x^*)_i$, multiplied by J^{i-k}. Thus as before, the combined power H and J in each term of the sum is $\geq k_{lim} - i + i - k = k_{lim} - k$, and there are roughly ${k_{lim}}^2$ such terms.

The spectrums of J and H are equal (basic LDS result) and their spectral radius λ (the maximum of the magnitudes of the eigenvalues) is less than 1. In practice we have found that λ is often significantly less than 1 even when the spectral radius of A is relatively close to 1. Intuitively J and H capture the strength of the dependency between the hidden states in consecutive time-steps. Smaller eigenvalues correspond to eigen-components with weaker dependencies that decay faster. Letting $\sigma(X)$ denote the spectral radius of an arbitrary matrix X and using the basic property that $\sigma(XY) \leq \sigma(X)\sigma(Y)$ and the identity $\|B\| \leq \dim(B)\sigma(B)$ we can estimate the 2-norms of various error matrices in terms of the 2-norms of the DAS errors. For example, we have that $\| \operatorname{vec}(\Delta(y, x^*)_k))\|_2 \leq N_x^2 \|\Delta(y, x^*)_{k_{lim}+1}\|_2 \lambda^{k_{lim}+1-k}$.

For harder cases such as $\Delta(x^T, y)_k$ that involve the sum over many terms we can apply triangle inequality for norms and then bound the norm of each term. This is the reason that the factor ${k_{lim}}^2$ appears in the claimed operator norm bound.

The most difficult case is $\Delta(x^T, x^T)_0$. The ASOS procedure estimates $(x^T, x^T)_0$ by solving the 2nd ASOS equation as a Lyapanov equation. The resultant error in $(x^T, x^T)_0$ is thus also the solution of a similar Lyapanov equation:

$$\Delta(x^T, x^T)_0 = J\Delta(x^T, x^T)_0 J' + J\Delta(x^T, x^*)_1 P' + P\Delta(x^T, x^*)_0{}'$$

While it is possible, although unlikely, that solving this equation could greatly amplify the error, this effect would be linear and constant (since the linear coefficients on $(x^T, x^T)_0$ do not depend on k_{lim}) and thus bounded in norm. □

2 Proof of Claim 3

Claim. *For* $i = 1, 2, 3$*:*

$$\lim_{T\to\infty} \mathrm{E}_\theta[\ \| \frac{1}{T} \operatorname{vec}(\phi_i) \|_2^2\ ] = 0$$

Proof. First we will consider the case $i = 1$.

Define residual prediction error γ_t by $\gamma_t = y_t - \mathrm{E}_\theta[\ y_t\ | y_{\leq t-1}\ ]$ and note that $\phi_1 \equiv (y, x^*)_{k_{lim}+1} - CA\left((x^*, x^*)_{k_{lim}} - x_T^* x_{T-k_{lim}}^*{}'\right)$ can be expressed as $\sum_{t=1}^{T-k-1} \gamma_{t+k+1} x_t{}^{*\prime}$.

Using this fact and the linearity of expectation we have that the expectation can be written as:

$$\begin{aligned}
\mathrm{E}_\theta[\ \| \frac{1}{T} \operatorname{vec}(\phi_1) \|_2^2\ ] &= \mathrm{E}_\theta[\ \frac{1}{T^2} \operatorname{tr}(\phi_1\ \phi_1')\ ] \\
&= \frac{1}{T^2} \operatorname{tr}(\sum_{t=1}^{T-k-1} \sum_{s=1}^{T-k-1} \mathrm{E}_\theta[\ \operatorname{vec}(\gamma_{t+k+1} x_t{}^{*\prime}) \operatorname{vec}(\gamma_{s+k+1} x_s{}^{*\prime})'\ ]) \\
&= \frac{1}{T^2} \operatorname{tr}(\sum_{t=1}^{T-k-1} \sum_{s=1}^{T-k-1} \mathrm{E}_\theta[\ x_t^* x_s^{*\prime} \otimes \gamma_{t+k+1} \gamma_{s+k+1}{}'\ ])
\end{aligned}$$

First we consider the terms of the inner sum where $t \neq s$. By symmetry we may assume, without loss of generality, that $s > t$. Then using the law of iterated expectation and the fact that $\forall i\ \ \mathrm{E}_\theta[\ \gamma_i\ ] = 0$ we have:

$$\begin{aligned}
\mathrm{E}_\theta[\ x_t^* x_s^{*\prime} \otimes \gamma_{t+k+1} \gamma_{s+k+1}'\ ] &= \mathrm{E}_\theta[\ \mathrm{E}_\theta[\ x_t^* x_s^{*\prime} \otimes \gamma_{t+k+1} \gamma_{s+k+1}' \mid y_{\leq s+k}\ ]\ ] \\
&= \mathrm{E}_\theta[\ x_t^* x_s^{*\prime} \otimes \gamma_{t+k+1} \otimes \mathrm{E}_\theta[\ \gamma_{s+k+1}' \mid y_{\leq s+k}\ ]\ ] \\
&= \mathrm{E}_\theta[\ x_t^* x_s^{*\prime} \otimes \gamma_{t+k+1} \otimes 0\ ] = 0
\end{aligned}$$

For terms where $t = s$ we have instead that:

$$\begin{aligned}
\mathrm{E}_\theta[\ x_t^* x_t^{*\prime} \otimes \mathrm{E}_\theta[\ \gamma_t \gamma_t' \mid y_{\leq t+k}\ ]\ ] &= \mathrm{E}_\theta[\ x_t^* x_t^{*\prime} \otimes S_{t+k+1}\ ] \\
&= \mathrm{E}_\theta[\ x_t^* x_t^{*\prime}\ ] \otimes S_{t+k+1}
\end{aligned}$$

where we recall that $S_i \equiv \mathrm{Cov}_\theta[\ \gamma_t \mid y_{\leq t+k}\ ] = \mathrm{E}_\theta[\ \gamma_t \gamma_t' \mid y_{\leq t+k}\ ]$

Our final equation for the expectation is then:

$$\frac{1}{T^2} \operatorname{tr}(\sum_{t=1}^{T-k-1} \mathrm{E}_\theta[\ x_t^* x_t^{*\prime}\ ] \otimes S_{t+k+1}) = \frac{1}{T^2} \sum_{t=1}^{T-k-1} \mathrm{E}_\theta[\ \|x_t^*\|_2^2\ ] \cdot \operatorname{tr}(S_{t+k+1})$$

Intuitively, the growth of the sum is linear in T, not quadratic, and thus the factor $\frac{1}{T^2}$ will cause the entire right-hand expression to go to zero in the limit. We can formalize this intuition. This first tool we will need as a basic result about the asymptotic behavior of the LDS: under the control-theoretic conditions necessary for steady-state we also have that distributions over x_t^* and y_t approach equilibrium as $t \to \infty$.

A simple consequence of this result is that various expectations over x_t^* and y_t will converge as $t \to \infty$. Thus we have $\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T-k-1} \mathrm{E}_\theta[\ x_t^* x_t^{*\prime}\ ] = X_0$ where $X_0 = \lim_{t\to\infty} \mathrm{E}_\theta[\ x_t^* x_t^{*\prime}\ ]$. We also know that S_t approaches its steady-state value S as as $t \to \infty$. These two facts allow us to evaluate the limit:

$$\lim_{T\to\infty} \frac{1}{T^2} \sum_{t=1}^{T-k-1} \mathrm{E}_\theta[\ \|x_t^*\|_2^2\ ] \cdot \mathrm{tr}(S_{t+k+1}) = \lim_{T\to\infty} \frac{1}{T} \mathrm{tr}(X_0)\mathrm{tr}(S) = 0$$

For the remain cases of i we can show, using a similar argument to the one given above, that:

$$\mathrm{E}_\theta[\ \|\frac{1}{T}\,\mathrm{vec}(\phi_2)\|_2^2\ ] = \frac{1}{T^2} \sum_{t=1}^{T-k-1} \mathrm{E}_\theta[\ \|x_t^*\|_2^2\ ] \cdot \mathrm{tr}(V_{t+k,t+k}^{t+k-1})$$

$$\mathrm{E}_\theta[\ \|\frac{1}{T}\,\mathrm{vec}(\phi_3)\|_2^2\ ] = \frac{1}{T^2} \sum_{t=1}^{T-k-1} \mathrm{E}_\theta[\ \|y_t\|_2^2\ ] \cdot \mathrm{tr}(V_{t+k,t+k}^{t+k-1})$$

and that these expectations also converge to 0 in the limit as $T \to \infty$. □

3 Proof of claim 4

Claim. *The approximation error in $\frac{1}{T}$-scaled 2nd-order statistics as estimated by the ASOS procedure converges to 0 in expected squared $\|\cdot\|_2$-norm as $T \to \infty$.*

Lemma 1. *Let X be a large vector formed by concatenating the vectorizations of the true values of all the 2nd-order statistics that are estimated at some point during the ASOS procedure (this includes the DAS, the M-statistics, and all of the intermediate quantities). Let $\hat{X}$ be the corresponding approximate estimate obtained from the ASOS procedure. Then we have:*

$$\lim_{T\to\infty} \mathrm{E}_\theta[\ \|\frac{1}{T}(X - \hat{X})\|_2^2\ ] = 0$$

Proof. The system of equations solved by the ASOS procedure consists of the ASOS equations, the ASOS approximations, and of a particular *subset* of the 2nd-order equations. All of these are linear in the 2nd-order statistics. Moreover, except for the ASOS approximations they are all satisfied by the exact values of the second-order statistics. We may thus write the system as:

$$\Upsilon X = \Gamma + \Phi$$

where Υ is a matrix of coefficients, Γ a vector that accounts for the constant terms in each equation (i.e. those only involving statistics of the form $(y, y)_k$ for some k) and Φ is a vector that accounts for errors in the ASOS approximations.

The ASOS procedure is simply a computationally efficient method for solving this system where the unknown Φ is replaced by the zero vector. So, $\hat{X} = \Upsilon^{-1}\Gamma$ while the true value is given by $X = \Upsilon^{-1}(\Gamma + \Phi)$. Thus $\frac{1}{T}(X - \hat{X}) = \Upsilon^{-1}(\frac{1}{T}\Phi)$ and hence the expectation in the claim can be re-written and then bounded:

$$\mathrm{E}_\theta[\ \|\Upsilon^{-1}(\frac{1}{T}\Phi)\|_2^2\ ] \leq \|\Upsilon^{-1}\|^2 \mathrm{E}_\theta[\ \|\frac{1}{T}\Phi\|_2^2\ ]$$

But by the previous claim $\lim_{T\to\infty} \mathrm{E}_\theta[\ \|\frac{1}{T}\Phi\|_2^2\ ] = 0$ and then noting that Υ doesn't depend on T, the result follows. □

4 Solving The "Primary Equation"

Lemma 1. *Let V be a vector space, $f : V \longrightarrow V$ be a continuous linear function such that $\rho(f) < 1$. Then a solution to the equation $x = f(x) + y$ is given by:*

$$x_0 = \sum_{i=0}^{\infty} f^i(y) \tag{1}$$

where the exponents denote function composition.

Proof. The condition $\rho(f) < 1$ ensures that the series converges (and determines the rate of convergence).

Then,

$$x_0 = \sum_{i=0}^{\infty} f^i(y) = \sum_{i=1}^{\infty} f^i(y) + f^0(y) = \sum_{i=0}^{\infty} f \circ f^i(y) + y = f(\sum_{i=0}^{\infty} f^i(y)) + y = f(x_0) + y$$

Algorithm 1 Algorithm solving the primary equation

1: **Input:** A, C, K, H, G
2: Initialize $X := 0$, $Y := G$.
3: **while** Y has not converged to 0 **do**
4: $\quad Y :=$ Solution for Z of $(Z = AZH' + Y)$
5: $\quad X := X + Y$
6: $\quad Y := H^{2k_{lim}+1} Y' A' C' K'$
7: **end while**

where we have used the fact that f is both continuous and linear so that it respect the infinite sum. □

Now let $f_1(X) = X - AXH'$, $f_2(X) = H^{2k_{lim}+1} X' A' C' K'$ and $y = G$ where A, H, C, K and G. These functions are clearly linear in X and continuous. Then the solution of $f_1(X) = f_2(X) + G$ is the solution of the primary equation. Taking f_1^{-1} of both sides yields $X = f_1^{-1} \circ f_2(X) + f_1^{-1}(G)$ which is the form of the equation solved in the previous lemma with $f = f_1^{-1} \circ f_2$ and $y = f_1^{-1}(G)$.

Conjecture 1. *For all $k_{lim} \geq 0$, $f_1^{-1} \circ f_2$ is a continuous linear function with $\rho(f_1^{-1} \circ f_2) < 1$*

In practice, $\rho(f_1^{-1} \circ f_2)$ will be a significantly less than 1 when k_{lim} is large enough (even when $\rho(H)$ is close to 1) which implies rapid convergence of the series defined in (1).

Algorithm 1 computes this series term-by-term and so by the previous lemmas and rapid converge property it is an efficient method for solving the primary equation. Note that Y can be computed easily on line 4 because $Z = AZH' + Y$ is a Sylvester equation, for which there are know efficient algorithms.

5 Psuedo-code for ASOS

Algorithm 2 The ASOS algorithm for computing the E-step. Note that for the purposes of implementation, symbols such as $(y,x)^\dagger_k$ can simply be interpreted as k-indexed matrix-valued variable names.

perform steady-state computations (Algorithm 3)
compute approximate first and last k_{lag} 1st-order statistics (Algorithm 4)
$(y,x^*)^\dagger_{k_{lim}+1} := -CAx^*_T {x^*_{T-k_{lim}}}'$
for $k = k_{lim}$ down to 0 **do**
    $(y,x^*)^\dagger_k := (y,x^*)^\dagger_{k+1}H' + ((y,y)_k - y_{1+k}y_1')\,K' + y_{1+k}{x^*_1}'$
end for
$(x^*,y)^\dagger_0 := {(y,x^*)^\dagger_0}'$
for $k = 1$ to k_{lim} **do**
    $(x^*,y)^\dagger_k := H((x^*,y)^\dagger_{k-1} - x^*_T y'_{T-k+1}) + K(y,y)_k$
end for
$G := (-Ax^*_T {x^*_{T-k_{lim}}}' + ((x^*,y)^\dagger_{k_{lim}} - x^*_{1+k_{lim}}y_1')K' + x^*_{1+k_{lim}}{x^*_1}'$
$(x^*,x^*)_{k_{lim}} :=$ SolvePrimaryEquation(A, C, K, H, G)
for $k \in \{0,\ 1,\ 2,\ ...,\ k_{lim}\}$ **do**
    $(y,x^*)_k := (y,x^*)^\dagger_k + CA(x^*,x^*)_{k_{lim}}{H^{k_{lim}+1-k}}'$
    $(x^*,y)_k := (x^*,y)^\dagger_k + H^{k_{lim}+1+k}(x^*,x^*)'_{k_{lim}}A'C'$
end for
for $k = k_{lim}$ down to 0 **do**
    $(x^*,x^*)_k := (x^*,x^*)_{k+1}H' + \left((x^*,y)_k - x^*_{1+k}y_1'\right)K' + x^*_{1+k}{x^*_1}'$
end for
$(x^T,x^*)_{k_{lim}} := (x^*,x^*)_{k_{lim}}$
for $k = k_{lim}-1$ down to 0 **do**
    $(x^T,x^*)_k := J(x^T,x^*)_{k+1} + P\left((x^*,x^*)_k - x^*_T{x^*_{T-k}}'\right) + x^T_T{x^*_{T-k}}'$
end for
$(x^*,x^T)_0 := (x^T,x^*)'_0$
$L := -Jx^T_1{x^T_1}'J' + J(x^T,x^*)_1P' + P((x^*,x^T)_0 - x^*_T{x^T_T}') + x^T_T{x^T_T}'$
$(x^T,x^T)_0 :=$ SolveLyapunov(J, J', L)
$(x^T,x^T)_1 := \left((x^T,x^T)_0 - x^T_1x^T_1\right)J' + (x^T,x^*)_1P'$
$(x^T,y)_{k_{lim}} := (x^*,y)_{k_{lim}}$
for $k = k_{lim}-1$ down to 0 **do**
    $(x^T,y)_k := J(x^T,y)_{k+1} + P\left((x^*,y)_k - x^*_Ty'_{T-k}\right) + x^T_Ty'_{T-k}$
end for
$(y,x^T)_0 := (x^T,y)'_0$

Algorithm 3 Steady-state Computations

$\Lambda_0^1 :=$ SolveDARE($A,\ C,\ Q,\ R$)
$S := C\Lambda_0^1 C' + R$
$K := \Lambda_0^1 C' S^{-1}$
$\Lambda_0^0 := \Lambda_0^1 - KC\Lambda_0^1$
$J := \Lambda_0^0 A' (\Lambda_0^1)^{-1}$
$\Lambda_0 :=$ SolveSylvester($J,\ J',\ \Lambda_0^0 - J\Lambda_0^1 J'$)
$V_1^T := V_0^T J'$
$H := A - KCA$
$P := I - JA$

Algorithm 4 Compute Approximate First and Last k_{lag} 1st-order Statistics

$x_1^* := \pi_1 + K(y_1 - C\pi_1)$
for $k = 2$ to k_{lag} **do**
    $x_t^* := Hx_{t-1}^* + Ky_t$
end for
$x_{k_{lag}}^T := x_{k_{lag}}^*$
for $k = k_{lag} - 1$ down to 1 **do**
    $x_t^T := Jx_{t+1}^T + Px_t^*$
end for
$x_{T-k_{lim}}^* := \mathbf{0}$
for $k = T - k_{lag} + 1$ to T **do**
    $x_t^* := Hx_{t-1}^* + Ky_t$
end for