Figures 3, 4, 6, 7, 10, 11 and 14
should be printed in color.

Adaptive Least Squares:
Recursive Least Squares with Constant Noise-to-Signal Ratio

J. Huston McCulloch
New York University

Aug. 9, 2024

The author is Adjunct Professor at New York University and Emeritus Professor at Ohio State University. He is indebted to James Bullard, Kan Chen, Chang-Jin Kim, Young-Il Kim, Lennart Ljung, Bruce McGough, Athanasios Orphanides, Thomas Sargent, and Yasushi Sugayama, as well as participants at the OSU Money/Macro Seminar, the 2005 conference on Computation in Economics and Finance, the Federal Reserve Bank of St. Louis, the Federal Reserve Bank of New York, and the NYU Econometrics Seminar for helpful comments and suggestions, and especially to James Durbin for his inspirational lectures at Ohio State in 1993.

Author's email: mcculloch.2@osu.edu. Author's address: Department of Economics, New York University, 19 W. 4th St, 6th Floor, New York NY 10012.

See http://www.asc.ohio-state.edu/mcculloch.2/papers/ALS/ for updates and MATLAB programs.

ABSTRACT

Adaptive Least Squares (ALS), a refinement of the Constant Gain Recursive Least Squares (CG-RLS) algorithm proposed by Ljung (1992) and Sargent (1993, 1999), is a parsimonious method of estimating linear regressions with time-varying coefficients and of proxying agents' time-evolving expectations. By holding the noise-to-signal ratio constant rather than the Kalman gain as in CG-RLS, the ALS filter nests the univariate Local Level Model (LLM), with its optimally declining but bounded gain, and permits the hyperparameters to be estimated by Maximum Likelihood from the time series itself. The algorithm is easily initialized with an uninformative prior on the regression coefficients. The ALS filter, which uses only past and current data, emulates agents' empirical expectations at each point in time. The ALS smoother, that also uses future data, is developed as well. A global test for coefficient significance at every point in time is developed, based on the smoother coefficients.

The ALS filter algorithm is illustrated with a univariate time series model of PCE inflation through Nov. 2023. The global coefficient test soundly rejects the LLM in favor of an AR(1) model, but AR(1) could not be rejected in favor of higher order AR models. The estimated noise-to-signal ratio implies an asymptotic gain of 1/21.8 mo^{-1}. Although the annualized one-month-ahead forecast from Nov. 2023 was 1.33%, the estimated "entrenched" or long-run inflation rate then was 3.30%, down considerably from its post-1980s high of 5.73% in March 2022.

Simulations are used to quantify the uncertainty in the model's implied forecasts, which is considerable. The Jarque-Bera statistic rejects i.i.d. normality in the inflation model.

Keywords: Adaptive Learning, Kalman Filter, Inflation, Time Varying Parameters

JEL Codes: C32 -- Time Series Models
E31 – Inflation

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Declarations of interest: none.

I. Introduction

Adaptive Least Squares, a refinement of the Constant Gain Recursive Least Squares (CG-RLS) proposed by Ljung (1992) and Sargent (1993, pp. 120-2), provides a method of estimating time-varying relationships that is more rigorous than rolling regression, yet is far more parsimonious than an unrestricted Time Varying Parameters (TVP) model. ALS and the more general concept of Adaptive Learning (AL) provide a means of proxying agents' expectations that incorporates learning, in a way that is far more realistic than the severe informational requirements of fully Equilibrium, or "Rational," Expectations. Sargent (1999), Bullard and Mitra (2002), Bullard and Duffy (2003), Evans and Honkapohja (2001, 2004), Orphanides and Williams (2003), Preston (2004), and Milani (2005) and are just a few of the many applications of the AL concept.

An early, but very restrictive, special case of RLS was Cagan's (1956) "Adaptive Expectations" (AE) model, in which m_t, the time t expectation of the future of a time series y_t (in Cagan's case inflation), was assumed to obey an equation of the form

$$m_t = m_{t-1} + \gamma(y_t - m_{t-1}) \tag{1}$$

In Cagan's original formulation, the gain coefficient γ was assumed to be an arbitrary subjective constant to be inferred indirectly from agents' expectationally motivated behavior, viz. their demand for money balances.

Shortly after Cagan's original paper, Muth (1960) and Kalman (1960) independently demonstrated that (1) in fact gives the long-run behavior of the optimal signal-extraction forecast of future y_t, but only provided the process is generated by the *Local Level Model* (LLM), i.e. if y_t is the sum of an unobserved Gaussian random walk plus independent Gaussian white noise, and provided the long-run gain coefficient is computed as a certain function of the empirical noise/signal ratio. The gain coefficient is therefore not an arbitrary subjective learning parameter akin to a demand elasticity, but rather should take on a specific value determined by the behavior of the process in question.

Although Muth (1960) developed only the constant long-run gain coefficient, Kalman's more rigorous treatment (1960; see also Harvey 1989, p. 107 and Appendix A.1 below) demonstrated that in finite samples with a constant ratio of noise to signal, the optimal gain is not constant, but in fact declines rapidly at the beginning of the sample. Kalman's approach also allows the noise/signal ratio and therefore the gain coefficients and their limiting value to be estimated by Maximum Likelihood (ML).

The Kalman Filter solution of the elementary LLM has been generalized to solve a Time Varying Parameter (TVP) model in which all the coefficients of a linear regression relation are allowed to change randomly over time, as exposited, for example, by Ljung and Söderström (1983) or Harvey (1989, Ch. 3). Ljung (1992) and Sargent (1993, 1999, Ch. 8) have proposed a parsimonious restriction on the covariance matrix of the random coefficient changes that leads, by this Extended Kalman Filter (EKF), to CG-

RLS. However, because their gain is constant throughout, their model does not nest the rigorous declining-gain solution of the LLM when it is restricted to a simple time-varying intercept term with no time-varying slope coefficients.

The present study introduces a new specification of the TVP covariance matrix that does nest the rigorous LLM with its declining, yet bounded, gain, by imposing a constant ratio of noise to signal, appropriately defined. The resulting *Adaptive Least Squares* (ALS) algorithm may be easily initialized with a diffuse prior on the regression coefficients that makes no presupposition of their values. The noise variance and the noise/signal ratio may then be rigorously estimated by Maximum Likelihood (ML), rather than simply postulated or estimated by ad hoc means as in the previous literature.

The ALS algorithm is used to estimate a univariate autoregressive (AR) model of monthly Personal Consumption Expenditures inflation. The proposed global significance test soundly rejects the LLM in favor of one with AR(1) transients. Howeer, AR(1) could not be rejected in favor of higher-order AR models. The ML-estimated noise-to-signal ratio of 21.2 mo. implies an asymptotic gain of 1/21.8 mo^{-1}. Although the annualized one-month-ahead forecast from Nov. 2023 was 1.33%, the estimated "entrenched" or long-run inflation rate then was 3.30%, down considerably from its post-1980s high of 5.73% in March 2022.

Section II below reviews and restates the rigorous Kalman solution of the LLM, in terms of the key concept of *Effective Sample Size*. This motivates Section III, which states the ALS model in the context of the general TVP and EKF model. Section IV relates the ALS model to the previous TVP and RLS literature. Section V develops a test for the hypothesis that one of the regression coefficients is globally zero. Section VI applies the ALS filter algorithm to US PCE inflation data, while Section VII uses simulations to quantify the uncertainty of these inflation forecasts. Section VIII concludes with mention of potential future applications and extensions. The Appendix provides mathematical details, and corrects an error in a critical equation in the pioneering work of Ljung (1992) and Sargent (1999).

II. The Local Level Model

Before presenting Adaptive Least Squares, we first review and restate the Kalman solution of the elementary *Local Level Model* (LLM) in terms of the useful concept of *Effective Sample Size*.

In the LLM, an observed process y_t is the sum of an unobserved Gaussian random walk μ_t plus independent Gaussian white noise:

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{NID}(0, \sigma_\varepsilon^2), \quad t = 1, \ldots N \\ \mu_t &= \mu_{t-1} + \eta_t, \quad \eta_t \sim \mathrm{NID}\left(0, \sigma_\eta^2\right). \end{aligned} \tag{2}$$

The *signal/noise variance ratio* is defined to be

$$\rho = \sigma_\eta^2 / \sigma_\varepsilon^2,$$

so that the two "hyperparameters" σ_ε^2 and ρ completely determine the system. Let the vector $\mathbf{y}_t = (y_1, \ldots y_t)'$ represent the observations up to and including y_t.

As reviewed in Appendix A.1, the classic Kalman Filter solution of the LLM may be expressed as follows:

$$\mu_t|\mathbf{y}_t \sim \mathrm{N}(m_t, \sigma_t^2). \tag{3}$$

with

$$m_t = m_{t-1} + \gamma_t(y_t - m_{t-1}), \tag{4}$$

$$\sigma_t^2 = \gamma_t \sigma_\varepsilon^2. \tag{5}$$

The *Kalman Gain* γ_t is the reciprocal of the *Effective Sample Size* N_t, determined by

$$N_t = (1 + \rho N_{t-1})^{-1} N_{t-1} + 1, \tag{6}$$

with the uninformative initialization

$$N_0 = 0. \tag{7}$$

The time t expectation of $\mu_{t'}$ and therefore of $y_{t'}$ for all $t' > t$ is then m_t. The precision, or reciprocal variance, of this estimate is directly proportional to N_t:

$$\sigma_t^{-2} = N_t / \sigma_\varepsilon^2.$$

In the limiting case $\rho = 0$, so that $\mu_t = \mu$, a constant, the effective sample size N_t equals the true sample size t. When $\rho > 0$, the effective sample size still behaves much like t initially, but is strictly less than t for $t > 1$, and is bounded above by its long-run value

$$N_{\mathrm{LR}} = \lim_{t \uparrow \infty} N_t = 1/2 \; + \; \sqrt{1/4 + 1/\rho}, \tag{8}$$

which is the unique positive root of the quadratic equation

$$\rho N_{\mathrm{LR}}^2 - \rho N_{\mathrm{LR}} - 1 = 0$$

that determines the fixed points of (6). The Cagan/Muth constant gain AE formula (1) is therefore valid only in this limit, with the limiting long run gain $\gamma_{\mathrm{LR}} = 1/N_{\mathrm{LR}}$. The gain in fact should be $\gamma_t = 1/N_t$.

The predictive error decomposition gives the distribution of the one-period-ahead forecasts:

$$y_t \mid \boldsymbol{y}_{t-1} \sim \mathrm{N}(m_{t-1}, \sigma_{t-1}^2 + \rho\sigma_\varepsilon^2 + \sigma_\varepsilon^2). \tag{9}$$

The product of these densities for $t = 2, \ldots N$ gives the joint probability of $y_2, \ldots y_N$ conditional on y_1 as a function of σ_ε^2 and ρ, and therefore the likelihood of σ_ε^2 and ρ conditioned on $y_1, \ldots, y_N$. The noise variance σ_ε^2 may be concentrated out of the log likelihood function, so that a numerical search is required only over the single parameter ρ.

Although it is convenient mathematically to develop the LLM in terms of the signal/noise variance ratio ρ, this ratio has the unnatural units [time^{-2}], and often is a very small number. It is more natural to report empirical results in terms of the equivalent *Noise/Signal standard deviation Ratio,*

$$NSR = \sigma_\varepsilon / \sigma_\eta = \rho^{-1/2}.$$

The *NSR* has the natural units [time], and equals the number of time units it takes for the variance of the average of *NSR* ε_t shocks to equal that of the sum of *NSR* η_t shocks. In other words, it is the typical length of time it takes for changes in the level of the state variable to begin to be empirically detectable. Furthermore, (8) implies that N_{LR} is only slightly larger than *NSR*:

$$NSR + .5 < N_{LR} < NSR + 1.$$

Figure 1 plots N_t versus t, using $NSR = 21.3$, the empirical value obtained in Section VI below for monthly PCE inflation with an ALS/AR(1) specification. N_t is virtually indistinguishable from t until $t = 8$, but then it grows more slowly and is virtually indistinguishable from its asymptotic value of $N_{LR} = 21.8$ after $t = 70$.

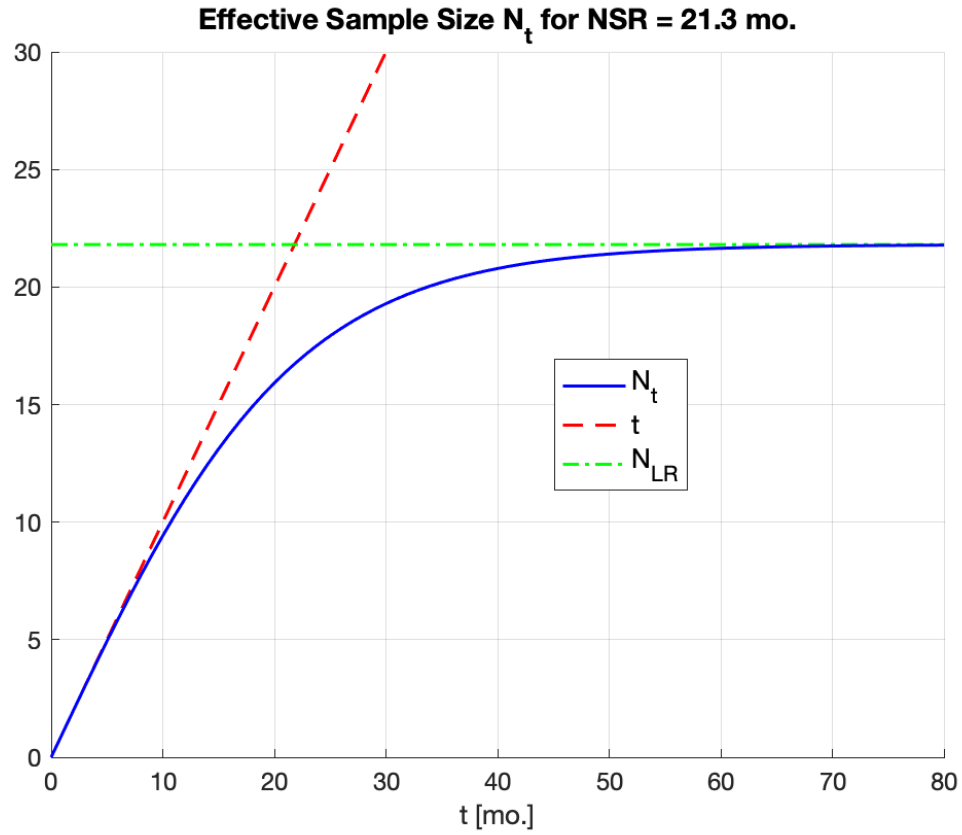

Figure 1
Effective sample size N_t with *NSR* = 21.3 mo., N_{LR} = 21.8 mo.

Equation (4) implies that each m_t is a linear combination of current and past values of y_t

$$m_t = \sum_{s=1}^{t} \varphi_{t,s} y_s \qquad (10)$$

with weights $\varphi_{t,s}$ that diminish as the distance into the past increases, according to the recursion

$$\varphi_{t,t} = \gamma_t,$$
$$\varphi_{t,s} = \varphi_{t,s+1}\frac{(1-\gamma_{s+1})\gamma_s}{\gamma_{s+1}}, \quad s < t.$$

For large t and s, as each γ_s approaches γ_{LR}, this implies that the weights decay approximately geometrically with distance into the past, according to

$$\varphi_{t,s} \approx \gamma_{LR}(1-\gamma_{LR})^{t-s},$$

as recognized in the title, "Optimal Properties of Exponentially Weighted Forecasts," of Muth (1960). The average lag implied by these limiting weights, as measured from the first forecast date t+1, is $N_{LR} = 1/\gamma_{LR}$.

It happens that (10) is equivalent to the Weighted Least Squares (WLS) estimate of m_t as if the random variables

$$v_{t,s} = y_s - m_t = \varepsilon_s - \sum_{\tau=s+1}^{t}\eta_\tau, \quad s \le t$$

were serially uncorrelated across s and had variances that grew with $t-s$ in proportion to $1/\varphi_{t,s}$, and therefore approximately geometrically when t and s are both large. However, that in fact is *not* their structure in the LLM. In fact, the LLM (2) implies that their variances increase *arithmetically* with $t-s$:

$$\mathrm{var}(v_{t,s}) = \sigma_\varepsilon^2 + (t-s)\sigma_\eta^2$$

and that they are *positively correlated* across s:

$$\mathrm{cov}(v_{t,s}, v_{t,s'}) = (t - \max(s,s'))\sigma_\eta^2, \quad s < t \text{ or } s' < t.$$

Equation (10) is therefore justified because it is the Generalized Least Squares (GLS) or Aitken's formula solution to the LLM problem, and not because it also happens to solve an unrelated, and observationally non-equivalent, WLS problem. The Kalman Filter is simply a recursive, computationally efficient way to solve the LLM problem without the massive matrix operations that GLS would require.

III. Adaptive Least Squares

The simplistic LLM allows the observed dependent variable y_t to depend only on a (time-varying) mean. A much more general and useful framework is the Time-Varying Parameter (TVP) linear regression model,

$$y_t = \mathbf{x}_t\boldsymbol{\beta}_t + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{NID}(0, \sigma_\varepsilon^2), \tag{11}$$
$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \mathrm{NID}(\mathbf{0}_{k\times 1}, \mathbf{Q}_t), \tag{12}$$

in which $\mathbf{x}_t$ is a $1\times k$ *row* vector[1] of explanatory variables, $\boldsymbol{\beta}_t$ is a $k\times 1$ column vector of time-varying coefficients $\beta_{j,t}$, and $\boldsymbol{\eta}_t$ is a $k\times 1$ column vector of permanent coefficient changes $\eta_{j,t}$ that are independent of the observation errors ε_t. Let $\mathbf{y}_t$ be the $t\times 1$ vector of dependent variables observed up to and including time t, and $\mathbf{X}_t$ be the $t\times k$ matrix of ideally exogenous explanatory variables up to and including time t. $\mathbf{Q}_t$ is the possibly time-dependent $k\times k$ covariance matrix of the transition errors $\boldsymbol{\eta}_t$. We assume that the first column of $\mathbf{X}_t$ is a vector of units, so that the first element of $\boldsymbol{\beta}_t$ is the intercept. When k = 1, the TVP model therefore reduces to the LLM when the single element of $\mathbf{Q}_t$ and

[1] We let $\mathbf{x}_t$ be a row vector rather than a column vector, since $\mathbf{x}_t$ is the t-th row of the regressor matrix $\mathbf{X}_N$.

therefore the noise/signal ratio is constant. For simplicity, we assume here that $\mathbf{X}_k$ is of full rank, although this restriction can be worked around.

As reviewed in Appendix A.2 below, system (11)-(12) may be solved by means of the well-known Extended Kalman Filter (EKF), which provides a recursive Bayesian rule giving the posterior filter distribution

$$\boldsymbol{\beta}_t|\mathbf{y}_t \sim \mathrm{N}(\mathbf{b}_t, \mathbf{P}_t) \tag{13}$$

for a $k\times1$ mean vector $\mathbf{b}_t$ and a $k\times k$ covariance matrix $\mathbf{P}_t$. The system may easily be initialized with an uninformative diffuse prior on the regression coefficients that makes no presupposition of their values. However, the full-blown TVP model (11)-(12) is much too general for most econometric purposes, since if even if $\mathbf{Q}_t$ is made time-invariant, it still introduces $k(k+1)/2$ ill-conditioned and incidental time-variation hyperparameters to be estimated, in addition to the observation variance σ_ε^2.

Ljung (1992) and Sargent (1999, p. 117) ingeniously postulate that $\mathbf{Q}_t$ *is directly proportional to* $\mathbf{P}_{t-1}$. This assumption, which leads to *Generalized Recursive Least Squares* (G-RLS), allows every component of $\boldsymbol{\eta}_t$ to be stochastic, yet effectively reduces $\mathbf{Q}_t$ to a single unknown parameter, while still being invariant in its implications to changes in the basis of the regressors. It also has the benefit of greatly simplifying the filter computations, by eliminating two $k\times k$ matrix inversions at each step. Nevertheless, the constant of proportionality that Ljung and Sargent propose must be modified in order for G-RLS to nest the elementary LLM when $k = 1$, as required, and to allow the likelihood to be computed.

The one element of the signal shock $\boldsymbol{\eta}_t$ that contributes directly to y_t, on a one-for-one basis and whose variance is therefore potentially comparable to σ_ε^2, is its first element, corresponding to the regression intercept term. However, the magnitude of this component, and therefore the implied variance of the first component of $\boldsymbol{\eta}_t$, is sensitive to the arbitrary manner in which the time $t-1$ variable regressors have been centered. In order to eliminate this ambiguity, we consider time-specific alternative bases in which the variable regressors have all been recentered in such a way that the covariance matrix of the transformed time $t-1$ coefficients and therefore the covariance matrix of the similarly transformed time t transition errors are block-diagonal, and then assume, just as in the LLM, that the variance of the transformed intercept coefficient is some scalar constant ρ times the noise variance σ_ε^2. In Appendix A.2 below, it is shown that this assumption implies that

$$\mathbf{Q}_t = \rho\, N_{t-1}\, \mathbf{P}_{t-1}, \tag{14}$$

where N_t is computed from ρ exactly as in (6) and (7) in the LLM.

The resulting "Adaptive Least Squares" (ALS) filter may then be written as

$$\mathbf{b}_t = \mathbf{W}_t^{-1}\mathbf{z}_t, \tag{15}$$

$$\mathbf{P}_t = \sigma_\varepsilon^2\mathbf{W}_t^{-1}, \tag{16}$$

where

$$\mathbf{z}_t = (1 + \rho N_{t-1})^{-1}\mathbf{z}_{t-1} + \mathbf{x}_t' y_t, \tag{17}$$
$$\mathbf{W}_t = (1 + \rho N_{t-1})^{-1}\mathbf{W}_{t-1} + \mathbf{x}_t'\mathbf{x}_t, \tag{18}$$

and N_t is computed as in (6) and (7). When there is no prior information about the coefficients at time 0, the algorithm is initialized with

$$\mathbf{W}_0 = \mathbf{0}_{k\times k}, \quad \mathbf{z}_0 = \mathbf{0}_{k\times 1}. \tag{19}$$

Note that in the fixed coefficient case $\rho = 0$, $\mathbf{z}_t$ becomes $\mathbf{X}_t'\mathbf{y}_t$, $\mathbf{W}_t$ becomes $\mathbf{X}_t'\mathbf{X}_t$, and (15) becomes the familiar expanding-window OLS formula $\mathbf{b}_t = (\mathbf{X}_t'\mathbf{X}_t)^{-1}\mathbf{X}_t'\mathbf{y}_t$.

Having thus initialized and updated the ALS filter, the predictive error decomposition becomes

$$y_t|\mathbf{y}_{t-1} \sim \mathrm{N}(\mathbf{x}_t\mathbf{b}_{t-1}, \sigma_\varepsilon^2 s_t^2), \quad t > k, \tag{20}$$

where

$$s_t^2 = (1 + \rho N_{t-1})\mathbf{x}_t\mathbf{W}_{t-1}^{-1}\mathbf{x}_t' + 1.$$

The log likelihood is then

$$\begin{aligned}\mathrm{L}(\rho, \sigma_\varepsilon^2|\mathbf{y}_N) &= \textstyle\sum_{t=k+1}^{N} \log \mathrm{p}(y_t|\mathbf{y}_{t-1}) \\ &= -\frac{N-k}{2}\log(2\pi\sigma_\varepsilon^2) - \textstyle\sum_{t=k+1}^{N}\log s_t - \frac{1}{2\sigma_\varepsilon^2}\sum_{t=k+1}^{N} u_t^2,\end{aligned} \tag{21}$$

where the scale-adjusted residuals,

$$u_t = e_t/s_t \tag{22}$$

equal the actual predictive errors,

$$e_t = y_t - \mathbf{x}_t\mathbf{b}_{t-1},$$

adjusted by their time-varying standard deviations s_t. Since $\mathbf{W}_t$ is of rank t for $t \le k$, the predictive density and therefore the likelihood contribution may only be computed for $t > k$.

Under the maintained assumptions, and given the two hyperparameters, these adjusted residuals are homoskedastic with variance σ_ε^2, even though the predictive errors themselves will in general be heteroskedastic. As in the LLM, the observation variance σ_ε^2 may be concentrated out of the log likelihood function in such a way that for any value of ρ, the likelihood is maximized with σ_ε^2 estimated in closed form by

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-k}\sum_{t=k+1}^{N} u_t^2 \tag{23}$$

A numerical search over the remaining hyperparameter ρ then provides its ML estimate.

If the model is well-specified and ρ equal to its true value, the adjusted residuals u_t must be iid $\mathrm{N}(0, \sigma_\varepsilon^2)$. Since the hyperparameter ρ and is consistently estimated by ML, routine large-sample specification tests such as the Jarque-Bera test for i.i.d. normality may therefore be applied to these residuals, as noted by Durbin and Koopman (2001, Ch. 5).

ALS may be written in terms of the Kalman gain $\gamma_t = 1/N_t$ by defining $\bar{\mathbf{z}}_t = \mathbf{z}_t/N_t$ and $\bar{\mathbf{W}}_t = \mathbf{W}_t/N_t$. Equations (17) and (18) then become

$$\bar{\mathbf{z}}_t = (1 - \gamma_t)\,\bar{\mathbf{z}}_{t-1} + \gamma_t\,\mathbf{x}_t' y_t,$$

$$\bar{\mathbf{W}}_t = (1-\gamma_t)\,\bar{\mathbf{W}}_{t-1} + \gamma_t\,\mathbf{x}_t'\mathbf{x}_t. \tag{24}$$

This formulation provides the intuition that the updated average moment vector and matrix are equal to a weighted average of their received values and the new data, with the Kalman gain serving as the weight on the new data. It also shows that as in the LLM, the ALS estimate of $\boldsymbol{\beta}_t$ happens to be equivalent to the WLS solution to a problem in which the weights on observation s decay geometrically with $t - s$ for large t and s, but again the structure of the errors is observably different than that assumed by WLS.

ALS may also be written without $\mathbf{z}_t$ in the *error-correction form*

$$\mathbf{b}_t = \mathbf{b}_{t-1} + \mathbf{W}_t^{-1}\mathbf{x}_t'(y_t - \mathbf{x}_t\mathbf{b}_{t-1}),\ t > k, \tag{25}$$

with $\mathbf{W}_t$ updated as in (18) or (24). This is essentially the form of RLS preferred by Evans and Honkapohja (2001, Eq. (2.9)), where their $\mathbf{R}_t$ is like our $\bar{\mathbf{W}}_t$, but computed with a constant gain. This form provides the intuition that the forecasting errors drive the revisions of the coefficient vector, but unfortunately it does not work for $t \le k$, since $\mathbf{b}_{t-1}$ is undefined then. It is also not obvious how to initialize it with a diffuse prior.

If one is estimating an autoregression by ALS, it is important to remember that, as in OLS, the inverse AR roots are biased away from unity. In the usual fixed-coefficients OLS environment, this bias disappears in large samples, but this consistency is absent in the ALS case, because the effective sample size never rises above N_{LR}.[2]

The Kalman *Filter* $(\mathbf{b}_t, \mathbf{P}_t)$, presented above, provides the posterior distribution of the coefficient vector conditional on the *past and current history* of the data up to time t. This is the appropriate question to ask if one is interested in simulating empirical expectations as of time t. However, if one instead wanted to retrospectively estimate the time-varying regression coefficients at a given point in time t, conditional on *both prior and subsequent experience* up to time $N > t$, the Kalman *Smoother*

$$\boldsymbol{\beta}_t|\mathbf{y}_N \sim \text{N}(\mathbf{b}_t^S, \mathbf{P}_t^S) \tag{26}$$

(also known as the 2-sided filter) becomes the appropriate tool. This is straightforward, but is somewhat more complicated. The pertinent smoother equations for both the general TVP case and the special ALS case are given in Appendix A.3. The ALS smoother, like the ALS filter, is unidentified for $t < k$. At $t = N$, the two are equivalent.

IV. Relation of ALS to other TVP approaches

There have been numerous attempts to make the general TVP model (11)-(12) more tractable through restrictions on the signal variance $\mathbf{Q}_t$. Early on, Cooley and Prescott (1973) were able to reduce $\mathbf{Q}_t$ to a single parameter, but only by permitting only the intercept to change, so that only the (1, 1) element of $\mathbf{Q}_t = \mathbf{Q}$ is non-zero. Their model does nest the LLM, and is invariant to recenterings of the regressors, but is unnecessarily restrictive.

---

[2] McCulloch (2016) corrects for this finite sample autoregressive bias in the fixed coefficient case by replacing OLS with a *Moment Ratio Estimator*. The ALS estimator could perhaps be modified in a similar manner, but the present paper makes no attempt to do this.

Sims (1988) and Kim and Nelson (2004) use (11) with a time-invariant covariance matrix **Q**, but assume that **Q** is diagonal in order to keep the problem tractable. This assumption still introduces *k* hyperparameters, yet is not particularly natural, since if a slope coefficient in a regression were to change, we would ordinarily expect to see compensating changes in the intercept and the slopes of correlated regressors, *ceteris paribus*. Furthermore, a change of basis for the regressors should leave the story told by a regression unchanged, yet this will not be the case under this assumption, since the implications of a zero correlation between the regressors will depend upon the arbitrary choice of basis. Like the Cooley-Prescott model, this diagonality assumption does nest the LLM, but is unnecessarily restrictive.

McGough (2003) uses a diagonal covariance matrix that is a (time-varying) constant times the identity matrix. Although this model is adequate for the theoretical point he was making, it is empirically unsatisfactory, since it forces all the coefficients to have the same transition variance (at each point in time), even though their units depend upon the often arbitrary units in which the regressors happen to be measured.

As noted above, Ljung (1992) and Sargent (1999) observe that if $\mathbf{Q}_t$ is restricted to be proportional to $\mathbf{P}_{t-1}$, not only are there far fewer parameters to estimate, but the filter also simplifies greatly.[3] They then set

$$\mathbf{Q}_t = \frac{\gamma}{1-\gamma}\mathbf{P}_{t-1}, \tag{26}$$

resulting in *Constant Gain Recursive Least Squares* (CG-RLS) which is similar to ALS, but with constant gain γ. Cp. also Sargent (1993, eq. (10)) and Evans and Honkapohja (2001, eq. (2.9). However, this CG-RLS does not nest the rigorous declining-gain Kalman solution of the LLM that justifies (1) as an asymptotic approximation, and that permits ML estimation of the noise/signal parameter that determines the long run gain itself. Furthermore, it is not obvious how to initialize CG-RLS with a diffuse prior.

Sargent (1999, Ch. 8) goes on to recommend initializing CG-RLS with the unconditional expected values of the coefficient vector and covariance matrix. However, by his maintained assumption, the coefficients are nonstationary, and therefore have an undefined unconditional mean, and infinite unconditional variances. Although full sample OLS coefficients can be computed from $\mathbf{X}_N$ and $\mathbf{y}_N$, they are in no sense "prior" information or "unconditional" values for $t < N$. Ljung (1992, p. 100) unhelpfully instructs his reader to initialize the covariance matrix with an unspecified $\mathbf{P}_0$.

In Sargent's empirical Chapter 9, he provides estimates of two quarterly macroeconomic models with CG-RLS. However, rather than estimate his constant gain

[3] This insight is valid despite the error in Ljung (1992) and Sargent (1999) discussed in Appendix A.4. The approximation invoked by Ljung (1992, p. 100) is in fact unnecessary.

from his data, he arbitrarily sets it to 0.015, which corresponds to a long-run effective sample size of 66.67 quarters, or 16.67 years.

Unfortunately, there is an error in the Kalman Filter as presented by Ljung (1992) and Sargent (1999), leading them to conclude that their CG-RLS is only an approximate implication of it under their assumption (26), when in fact it is exact. This error is explained and corrected in the Appendix.

Stock and Watson (1996) and Sargent and Williams (2003) assume, in place of either (14) or (26), that

$$\mathbf{Q}_t = \mathbf{Q} = \rho\sigma_{\varepsilon}^{2}(\mathbb{E}\mathbf{x}_t'\mathbf{x}_t)^{-1}. \tag{27}$$

If the relevant expectation exists, this is equivalent in an expectational sense to (14), since then

$$\mathbb{E}\mathbf{W}_t = N_t\mathbb{E}\mathbf{x}_t'\mathbf{x}_t.$$

However, it is not necessarily true that the required moments do exist, and even if they did, it would impose a great informational burden on agents to require them to know what they were. Equation (14), on the other hand, does not even require these moments to be finite, and only requires that agents observe $\mathbf{X}_t$, and $\mathbf{y}_t$, and know ρ.[4] Assumption (27) does nest the LLM, since then the required expectation is just a unit scalar. For $k > 1$, however, it only approximates ALS. It also lacks the computational simplicity of ALS or CG-RLS, since it requires the more general EKF described in Appendix A.2.

Stock and Watson (1996) calibrate the coefficient ρ in (27) (their λ^2) for several macroeconomic time series and relationships by minimizing the sum of squared forecasting errors. This will give results similar to ours, but is by no means equivalent, even apart from the often subtle difference between our (14) and their (27). For one thing, the initial errors have much larger variance than the later errors, simply because the coefficient vector is still highly uncertain. Equation (20) correctly takes this into account and enables the full permissible sample ($N–k$ observations) to be incorporated into the log likelihood. Stock and Watson, on the other hand, only roughly take this factor into account, by discarding their first 60 monthly observations *a priori*. This is wasteful if the signal/noise ratio is large, and inadequate if the signal/noise ratio is small. Furthermore, it is obvious from (20), which is similar to the formula for the conditional distribution that would result from (27), that even asymptotically the squared forecasting errors e_t^2 are greater in expectation than σ_{ε}^2 by an amount that depends on ρ, so that minimizing their sum of squares will give a biased estimate of σ_{ε}^2. In addition, even after the warm-up period they are not homoskedastic, and hence they should not be given equal weight.

Orphanides and Williams (2004) calibrate their CG-RLS gain coefficient both by minimizing a sum of squared forecast errors as in Stock and Watson (1996), and by matching simulated forecasts of inflation, unemployment, and the fed funds rate as closely as possible to the mean forecasts of the Survey of Professional Forecasters.

---

[4] The observation variance σ_ε^2 is required to compute $\mathbf{P}_t$, but not $\mathbf{b}_t$.

However, if one's objective is to construct one's own expert forecast of these variables, one should use actual experience, and not the forecasts of other, perhaps less sophisticated, experts, to calibrate one's own forecasts.

Milani (2005) calibrates his CG-RLS gain parameter by optimizing the fit of an ancillary, fixed coefficient New Keynesian Phillips Curve equation, rather than to the behavior of observed inflation. This repeats Cagan's (1956) mistake of treating the gain like a subjective learning parameter to be inferred from agents' expectationally-motivated behavior, specifically their demand for money in his case, instead of estimating ρ from the inflation series in question using (9) and then computing the long run gain according to (8).

Cogley and Sargent (2005) estimate a three equation VAR(2) TVP model with 21 coefficients. However, rather than impose the Ljung-Sargent parsimonious RLS restriction, they estimate all 231 elements of the unrestricted 21×21 covariance matrix, subject to reflecting boundaries that prevent nonstationary autoregressive roots. With fixed coefficients there might be a case for imposing such restrictions, since an explosive process would have long-since blown up and would never have been observed. With time-varying coefficients, however, there is no reason one could not drift into such a situation if called for by sufficient evidence of instability, as is all too often the case with inflation data. They also incorporate stochastic volatilities as additional state variables, and estimate the system with Monte Carlo Markov chain methods rather than the Kalman filter.

V. Hypothesis Testing

A *local* test on the null hypothesis

$$\beta_{t,j} = 0$$

for a single value of *t* may be performed either using the filter coefficients with test statistic

$$z_{t,j} = b_{t,j}/{p_{t,j,j}}^{1/2},$$

where $b_{t,j}$ is the *j*-th element of $\mathbf{b}_t$ and $p_{t,j,j}$ is the (*j, j*) element of $\mathbf{P}_t$, or else using the smoother coefficients with

$$z^S_{t,j} = b^S_{t,j}/{p^S_{t,j,j}}^{1/2}$$

and analogous notation. These test statistics may be given a frequentist interpretation with a N(0, 1) distribution under the null, conditional on the signal and noise variances, since under a uniform prior for the coefficients, the filter and smoother values are normally distributed about β_t with the estimated covariance matrix. The present paper takes the consistent ML estimates of the two hyperparameters as their true values. In practice, their estimation errors add some uncertainty that should be investigated in future research.

Global linear restrictions on coefficients of the type

$$\beta_{t,j} = 0, \ t = k, \ldots n \tag{28}$$

are very important, but much more difficult. The first problem is that they require access to the 4-dimensional covariance array $\mathbf{C} = \left(c_{t_1,t_2,j_1,j_2}\right)$, where

$$c_{t_1,t_2,j_1,j_2} = \mathrm{cov}\left(b^S_{t_1,j_1}, b^S_{t_2,j_2}\right). \tag{29}$$

This covariance array contains the N-k+1 smoother covariance matrices $\mathbf{P}_t^S$, but also the intertemporal covariances for $t_1 \neq t_2$. Unfortunately, there appears to be no way to compute the required elements of **C** recursively as in the Kalman filter and smoother. Instead, **C** must be found within the covariance matrix of a quite large Generalized Least Squares (GLS) problem. This GLS problem is specified and solved in Appendix A.4 below. It runs much more slowly than the ALS filter and smoother themselves, but fortunately is only required for occasional specification tests and not for hyperparameter estimation or routine forecasting.

Let $\mathbf{C_{T,j}}$ be the n_T x n_T matrix of coefficients from **C** with t_1 and t_2 in subset T of the integers k ... n, and with $j_1 = j_2 = j$, where n_T = o(T). Let $\mathbf{b^S_{j,T}}$ be the 1 x n_T vector of smoother coefficients $b^S_{j,t}$ with t in T. Then, given the two hyperparameters, the global test statistic

$$G_{\mathrm{T,j}} \;=\; \mathbf{b^S_{j,T}} \mathbf{C^{-1}_{T,j}} \; \mathbf{b^S_{j,T}}' \tag{30}$$

has a χ^2 distribution with n_T degrees of freedom under the null.

Simulations show that such a test has correct size under the null, as expected. However, when, as in ALS, the correlation of nearby observations is very close to unity, its power to detect non-zero values of $\beta_{t,j}$ actually declines as n_T approaches its greatest possible value, N–k+1, as shown in Appendix A.4 below. A good balance between number of observations used and independence is achieved when the observations used are equally spaced approximately 2·*NSR* apart. Accordingly, we define

$$n_\mathrm{T} = \mathrm{round}\left(\frac{(N-k+1)}{2NSR}\right),$$

$$\mathrm{T}(h) = k - 1 + \mathrm{round}\left(\frac{(h-.5)(N-k+1)}{n_\mathrm{T}}\right), h = 1, \ldots\, n_\mathrm{T},$$

and set

$$\mathrm{T} = \{\mathrm{T}(h), h = 1, \ldots\, n_\mathrm{T}\}$$

in (30).

Joint local or global hypotheses on more than one coefficient may be tested in the analogous fashion using the appropriate submatrix of $\mathbf{P}_t, \mathbf{P}_t^S$ **P** or **C.**

Because the null hypothesis of no parameter change, i.e. $\rho = 0$, is on the boundary of the permissible parameter space $\rho \geq 0$, the usual regularity conditions for the χ^2 limiting distribution of the Lagrange Multiplier (LM) and Likelihood Ratio (LR) statistics are not met (Moran 1971a, 1971b). Nevertheless, Tanaka (1983) has shown that the LM statistic is still useful and informative in the LLM case, provided the critical values are appropriately adjusted. Preliminary simulations with the LLM indicate that the 5% critical value is approximately 2.3, which is far less than the value of 3.84 from the chi-squared distribution with one degree of freedom.

VI. Application to US Inflation

As noted early on by Klein (1979), the time series behavior of US inflation has evolved over time: In the 19th century, the price level itself appeared to be stationary. In the early 20th century, the price level underwent permanent shifts, but the inflation rate appeared to be stationary with mean near 0. But then in the later 20th century, the inflation rate itself became more and more persistent. Writing in 1971, Sargent (1971) was still able to argue that inflation was clearly a stationary process, but within a few years, a unit root in CPI inflation could no longer be rejected. A univariate time series model of the US inflation rate is therefore a natural application of the ALS method.

Figure 2 shows the chained Personal Consumption Expenditures Deflator (PCE) inflation rate, seasonally adjusted, computed as $\pi_t = 1200(\ln(P_t / P_{t-1}))$, for Feb. 1959 through Nov. 2023. A series of high-inflation months had left continuously compounded year-over-year PCE inflation at 6.88% in June 2022, followed by a marked decline to 2.61% in Nov. 2023. These swings set off a vigorous debate over whether these inflation rates should be regarded as "transitory" or "entrenched."

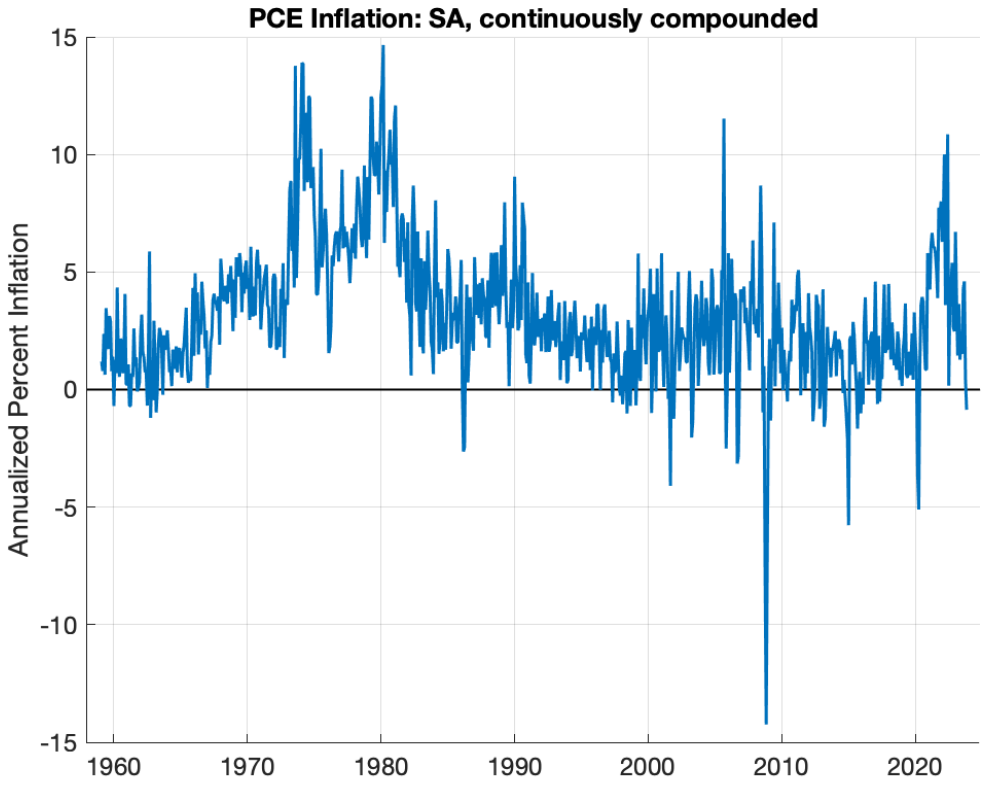

Figure 2
Monthly PCE inflation, seasonally adjusted, Feb. 1959 – Nov. 2023.
Source: NIPA via FRED

Table 1 presents the outcome of ALS estimation of time-varying AR(p) models of PCE inflation for $p = 0, ...4$, where the AR(0) model is simply the LLM, with no time-varying autoregressive parameters:

$$\pi_t = \beta_{1,t} + \sum_{j=2}^{p+1} \beta_{j,t}\pi_{t+1-j} + \varepsilon_t$$

In each model, there are $k = p+1$ time-varying parameters including the intercept. The sample for each model, allowing for up to 4 lags, is June 1959 – Nov. 2023, for $N = 774$ months.

Table 1
ALS AR(p) model of PCE Inflation
6/59 – 11/23, $N = 774$

p	0	1	2	3	4
NSR [mo.]	2.88	21.2	29.5	38.8	51.0
95% CI	(2.13, 3.87)	(14.2, 31.6)	(20.7, 42.7)	(27.5, 56.1)	(35.5, 79.4)
N_{LR} [mo.]	3.43	21.8	30.0	39.3	51.5
ρ [mo.$^{-2}$]	0.120	2.21e–3	1.15e–3	6.65e–4	3.85e–4
σ_ε^2 [(%/yr.)^2]	3.04	3.72	3.69	3.67	3.72
LR: $\rho = 0$	566.31	89.47	72.22	52.18	29.99
G: AR(p)=0	---	163.6	13.35	10.64	5.74
χ^2 DOF	---	18	13	10	8
$p(G)$	---	1.6e–25	0.421	0.386	0.676
Jarque-Bera	562.1	220.2	325.3	347.4	309.2
$p(JB)$	8.8e–123	1.5e–48	2.3e–71	3.7e–76	7.3e–68
Forecasts from 11/23:					
1 mo.	1.53	1.33	0.98	1.93	1.84
1 yr. avg.	1.53	2.98	2.65	2.11	2.04
1 yr. marg.	1.53	3.30	3.11	2.56	2.43
Long Run	1.53	3.30	3.12	2.84	2.71

The estimated noise-to-signal standard deviation ratio *NSR* increases sharply with the autoregressive order p, beginning with an absurdly low value of 2.9 months for $p = 0$, and then growing from 21.2 months for $p = 1$ to 51.0 months at $p = 4$. Evidently, the LLM interprets as permanent fluctuations that can more easily be explained as transitory low-order AR components. In general, the more serially correlated regressors that are present, the less ALS must rely on the random walk in the regression coefficients to explain local persistence, and the higher the *NSR*. The long-run effective sample size or reciprocal long-run gain, N_{LR} is, as required, slightly higher than *NSR*.

The Likelihood Ratio (LR) statistics for the hypothesis of fixed coefficients ($\rho = 0$, or equivalently, $NSR = \rho^{-1/2} = \infty$), are all are well above 2.3, the approximate 5% critical value in the case of the LLM, so that we are justified in rejecting fixed coefficients. The strength of the case against fixed coefficients declines with the autoregressive order, but remains very strong even at $p = 4$.

The global test statistic G for the null that the AR(p) coefficient in the AR(p) model is 0 for all t is distributed χ^2 with the indicated degrees of freedom DOF under the null with the maintained assumptions. For this purpose, the two hyperparameters were estimated under the alternative hypothesis and treated as known. The results for $p = 1$ overwhelmingly reject that the AR(1) coefficient is globally 0 in the AR(1) model, which is to say that it overwhelmingly rejects the LLM. However, we cannot reject that the AR(2) coefficient is globally 0 in the AR(2) model, or that the AR(3) coefficient is globally 0 in the AR(3) model, or that the AR(4) coefficient is globally 0 in the AR(4) model. We therefore accept the AR(1) model as being the most parsimonious model consistent with long-run experience.

Figure 3 shows the ALS filter estimates of the two coefficients in the preferred AR(1) model. Since the initial effective sample size N_t is very small, it is to be expected that the first several months will have very erratic point estimates, as well as very large standard deviations. These first several estimates may therefore safely be disregarded.

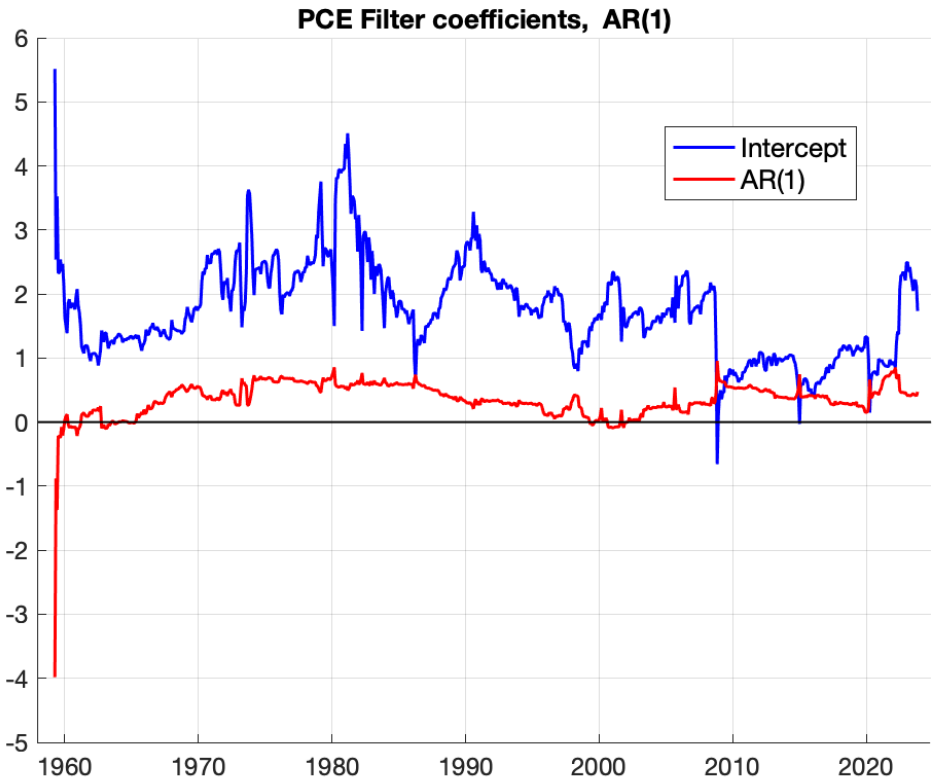

Figure 3
AR(1) model coefficients

Figure 4 plots the filter AR(1) coefficients from Figure 3 (heavy blue line), along with the smoother estimates (heavy red line) and the 95% credible intervals (CIs, thin lines) for each The CI for the filter is naturally very wide in the early years because of the small effective sample size there, and so is truncated in the graph. It may be seen that the filter coefficient is locally significantly different from zero throughout 1974-88, 2009-16, and again since 2020. However, it is essentially zero before 1965 and throughout 2020, so that there has been substantial change in the persistence of inflation. Although the filter point estimate never quite reaches unity, its 95% CI either includes or almost includes this value throughout 1967-87 and during 2021, indicating near-unit-root transitory behavior there, on top of the unit root already implied by the TVP model itself. (As noted by above, ALS estimates of AR models are prone to the same small-sample bias away from a unit root as are OLS estimates, per Dickey and Fuller (1979), with the added problem that the effective sample size is bounded above by N_{LR}. The present paper makes no attempt to correct for this bias.)

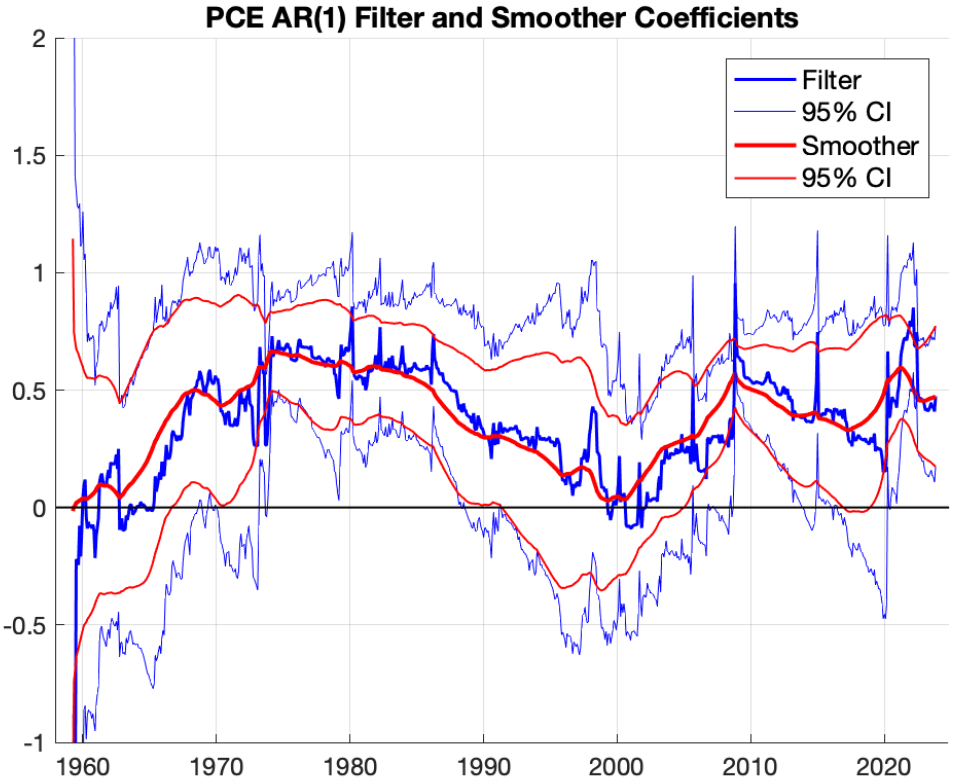

Figure 4
Filter (blue) and smoother (red) estimates of the AR(1) coefficient in the AR(1) model, with 95% credible intervals (thin lines). Vertical axis truncated to [-1, 2].

The filter estimates shown in Figures 3 and 4 are those that might have been used by a contemporary observer to forecast inflation. However, the smoother estimates, shown in red for the AR(1) coefficient, give more accurate retrospective estimates of the state variables conditional on the entire data set. At the last observation, where $t = N$, the filter and smoother necessarily coincide, as do their confidence intervals.

Figure 5 below shows the smoother z-statistic for the hypothesis that the AR(1) coefficient is zero (red line), derived from the point estimate and standard error used to construct the red line in Figure 4. Each is distributed N(0, 1) under the null, conditional on the model and the two estimated hyperparameters. However, they are highly serially correlated, so that a χ^2 test for global significance must take these correlations into account. Although a test based on all $N–k+1$ identified observations on the smoother has correct size under the null hypothesis, the serial correlation severely reduces the power of the test when all observations are used. Accordingly, as explained above, the time period after the first $k-1$ observations was divided into DOF = 18 subperiods, each of length

approximately $2 \cdot NSR = 42.4$ months, and only the midpoints of each of these subperiods (indicaatd by blue dots in the figure) were used. The resulting χ^2 statistic, as reported in Table 1, was 163.6, which overwhelmingly rejects the hypothesis that the coefficient is globally 0, despite periods like 1989–2005 when it cannot be locally rejected.

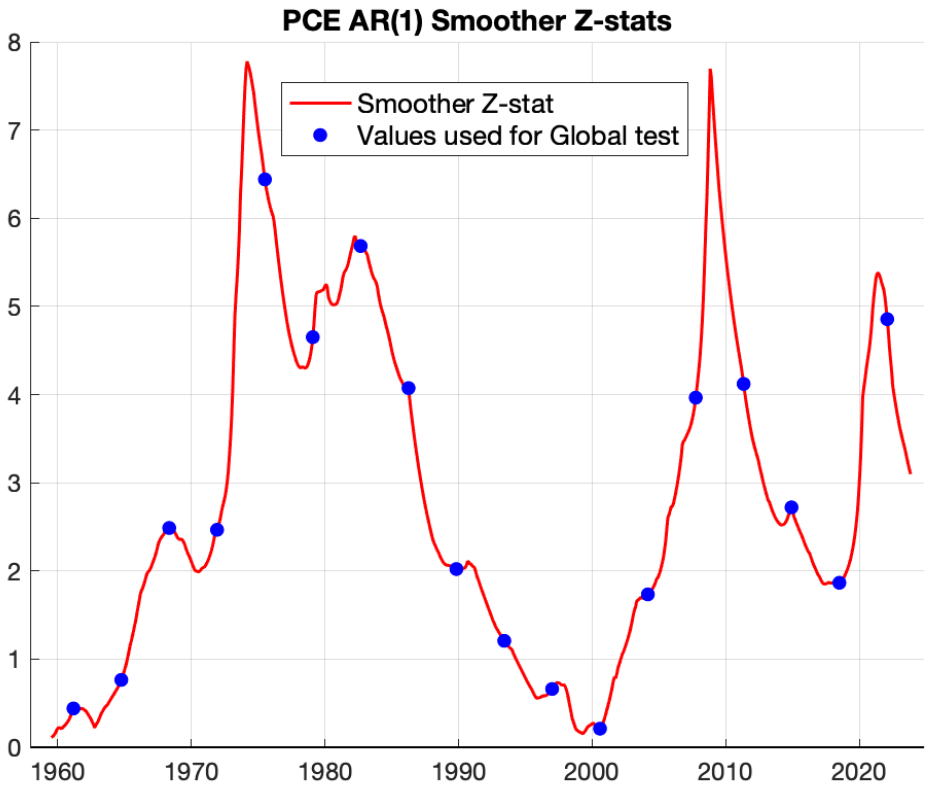

Figure 5
Smoother z-statistics for the AR(1) coefficient (red line), with subset used for the global *G*-statistic (blue dots).

Even though AR(2) effects have not been globally statistically significant in the past, they may become so in the future under the Fed's Aug. 27, 2020 announcement that it would henceforth deliberately attempt to overshoot (or undershoot) its long-run inflation goal of 2.00% if it has fallen short (or exceeded) that goal in the recent past, albeit not by so much as to make the price level trend-stationary as under a Wicksell rule (Powell 2020). The significance of the AR(2) coefficient should therefore be periodically reviewed in the future.

For each time t, the model can be used to predict annualized monthly inflation 1, 2, or h months into the future, and these marginal inflation predictions can be averaged to obtain an average inflation forecast for horizon h. These forecasts simulate the expectations of agents whose information set consists only of their experience of past inflation. Some agents may take a broader set of variables into account, and a limited number of such variables such as unemployment or nominal rates could easily be included in an ALS vector autoregression (ALS-VAR), but the present paper focusses on the univariate case. It is an empirical matter whether such additional variables would actually be globally significant.

Figure 6 plots the marginal and average ALS/AR(1) inflation forecasts for the last month in our observation set, Nov. 2023, when the economy happened to have just experienced an unusually low monthly inflation observation. The last annualized observation, taken directly into account by the AR(1) forecasts, is plotted as a green star, at –0.86%. The red line gives the marginal inflation forecasts, which start at 1.33% for h = 1 month, but then quickly ascend and stabilize at an asymptotic value of 3.29%. The marginal forecasts are within 0.01% of this value by h = 9 months.

The blue line gives the average inflation forecasts, which begin at the same 1.33% at 1 month, and then ascend more slowly toward the same asymptotic value of 3.29. The 1-year average inflation forecast is still only 2.98%. With the simplistic LLM, by way of contrast, the forecast of inflation as given in Table 1 for p = 0 is 1.53% at all horizons.

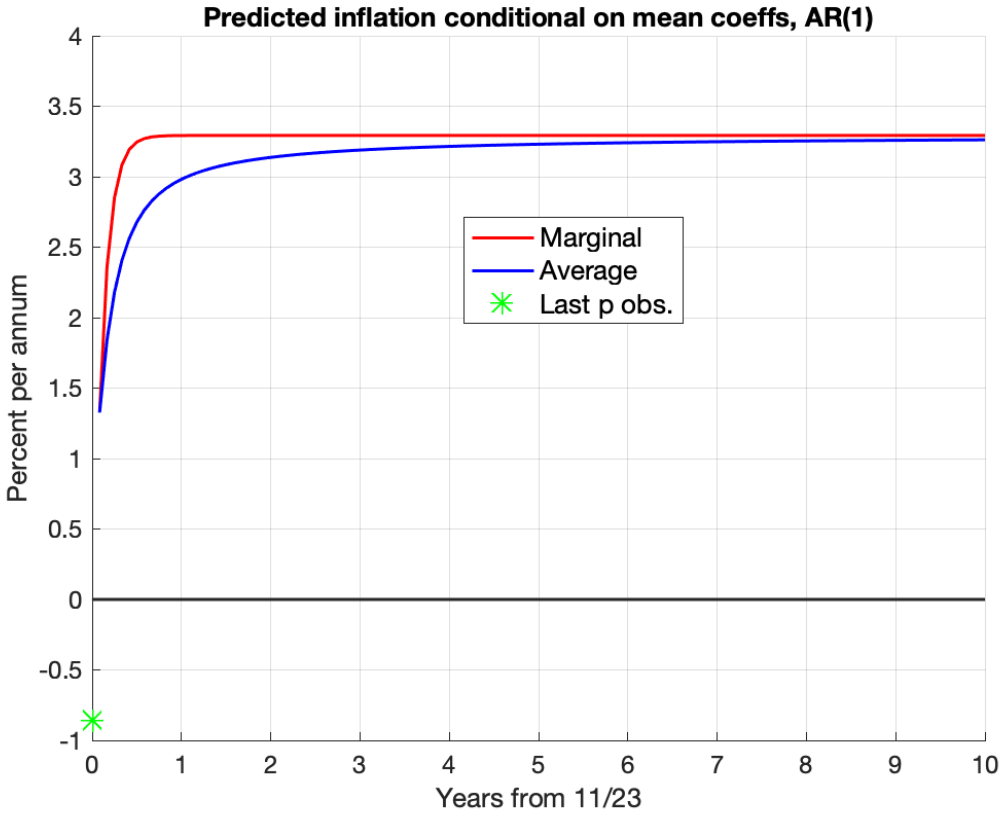

Figure 6
Marginal (red) and average (blue) predicted inflation, Nov. 2023. Green star is observed Nov. 2023 annualized inflation.

The asymptotic, long-run expected inflation rate implied by an AR(p) process at time t may be obtained directly, without recursively forecasting monthly inflation, as

$$\pi_{\mathrm{LR}}(t) = \begin{cases} \beta_{1,t}/\left(1-\sum_{j=2}^{k}\beta_{j,t}\right), & \text{if stationary,} \\ \mathrm{sgn}\left(\beta_{1,t}\right)\cdot\infty & \text{otherwise.} \end{cases}$$

This may be estimated, in the stationary case, by

$$\hat{\pi}_{\mathrm{LR}}(t) = b_{1,t}/\left(1-\sum_{j=2}^{k} b_{j,t}\right). \tag{31}$$

Since this value abstracts from the transitory AR components, it corresponds best to what is popularly meant by "entrenched inflationary expectations," and would be the appropriate univariate experience-based "inflation" variable to include in a Taylor-type rule. It is graphed for our AR(1) estimates in Figure 7 below as the blue line, along with the one month ahead forecast as the red line. In the simplistic Cagan Adaptive Expectations model (1) and its LLM rationalization (2), short-run and long-run inflation forecasts are necessarily one and the same thing. It may be seen from Figure 7 that with the much richer ALS model, there are often substantial differences between the two.

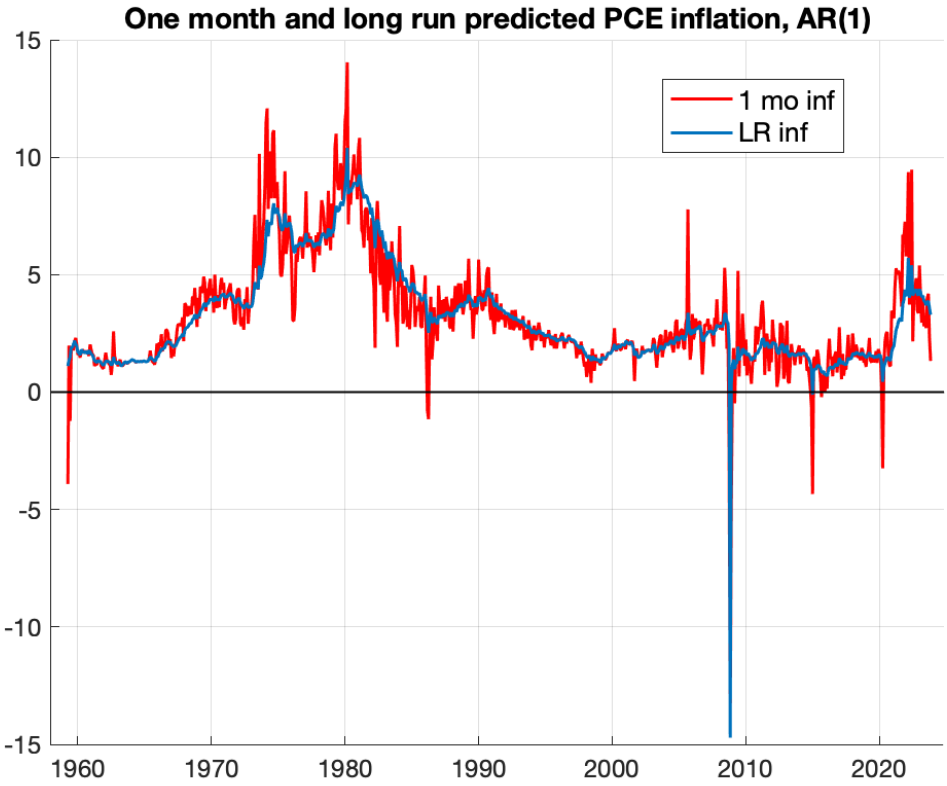

Figure 7
One-month (red) and long-run (blue) predicted PCE inflation

Long-run "entrenched" inflation was only 1.49% in Aug. 2020, when the FOMC announced its new policy of briefly permitting a little more than 2% PCE inflation in order to speed the growth of entrenched expectations up to its 2% target (Powell 2020). Entrenched inflation reached this target value already in March 2021, but then continued to rise, exceeding 4.00% throughout Dec. 2021 – March 2023, and briefly reaching 5.72% in March 2022. This more than accomplished the FOMC's new policy.

VII. Forecast uncertainty

The uncertainty of forecasts made with an ALS/AR(p) model from time $t = N$ forward has at least five components:

1. Future "noise" errors $\varepsilon_t, t > N$. Although these affect the accuracy of forecasts, they have mean zero, and therefore to not enter the forecasts themselves, and do not affect the accuracy of our estimates of the forecasts per se.

2. Initial parameter uncertainty, as reflected in $\mathbf{P}_n$.

3. Future parameter drift, as reflected in $\mathbf{Q}_t$, $t > N$.

4. Hyperparameter uncertainty with respect to NSR and σ_ε^2.

5. Model uncertainty, in this case with respect to its assumption of a low-order AR process, with independent and normal errors. The Jarque-Bera JB statistics in Table 1 have an asymptotically χ^2 distribution with 2 degrees of freedom under the null of i.i.d. normality. The *p*-values in the last row of Table 1 overwhelmingly reject this hypothesis, indicating conditional non-normality and/or conditional heteroskedasticity of the errors. Furthermore, a low-order AR process may not adequately approximate a "long-memory" (fractionally integrated) error structure, even if the innovations are Gaussian.

In Figures 8 through 11 below, we investigate the second and third of these sources of forecast uncertainty by means of Monte Carlo simulations. Figure 8 illustrates the average inflation forecasts from Nov. 2023 that would have been made with the AR(1) model, using 20 draws from the N($\mathbf{b}_N$, $\mathbf{P}_N$) posterior distribution in place of $\mathbf{b}_N$ itself, and no further coefficient drift. Each simulation is initialized with the observed annualized monthly inflation for Nov. 2023, represented by the green star at –0.86%. The heavier blue line is the point forecast from Figure 6. Each of these draws happens to be stationary.

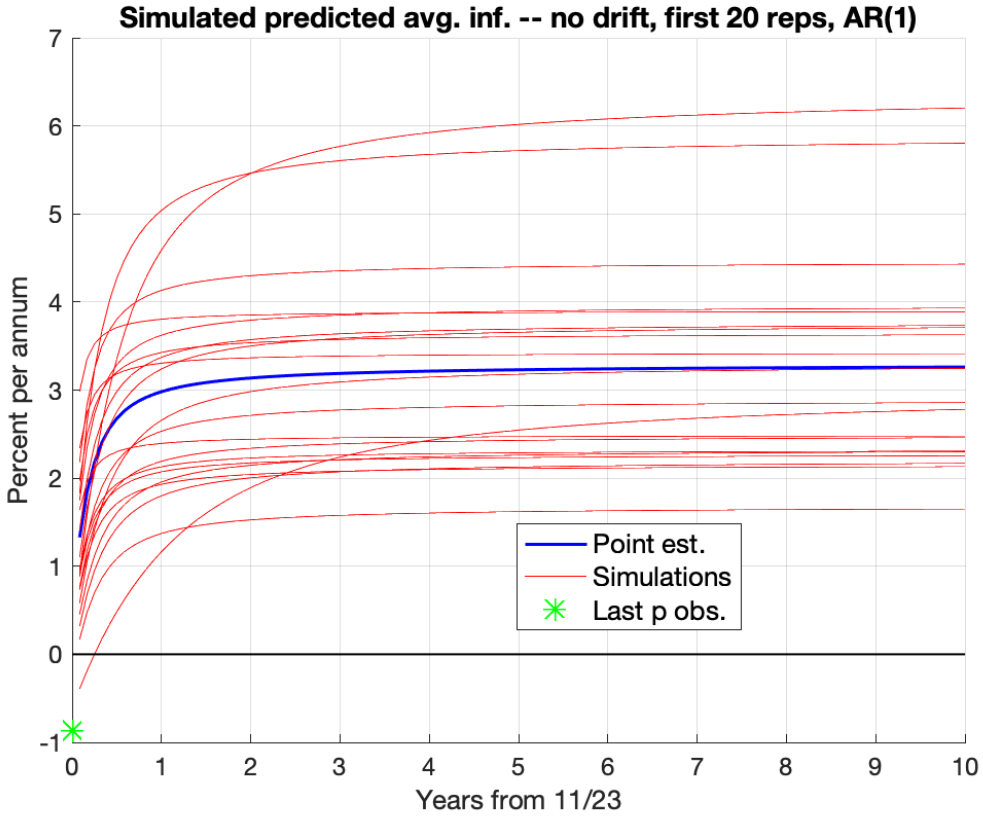

Figure 8
Simulated predicted average PCE inflation from Nov. 2023, AR(1) model with no coefficient drift, first 20 replications. Blue line is point estimate from Figure 7. Green star is observed Nov. 2023 annualized inflation.

The thin red lines in Figure 9 plot the posterior median and 50% and 95% credible intervals for predicted average inflation in the AR(1) model with no coefficient drift, using 1000 such simulations, starting with the 20 of Figure 8. Again, the heavier blue line is the point estimate from Figure 6. The posterior median coincides with this point estimate so closely that it is almost entirely hidden beneath it. The 50% CI is converging to approximately (2.7, 3.8), while the 95% CI is approximately (1.0, 5.1) at the 10-year horizon and is continuing to grow, so that there is considerable uncertainty in the 10-year forecast, even assuming away parameter drift and hyperparameter and model uncertainty.

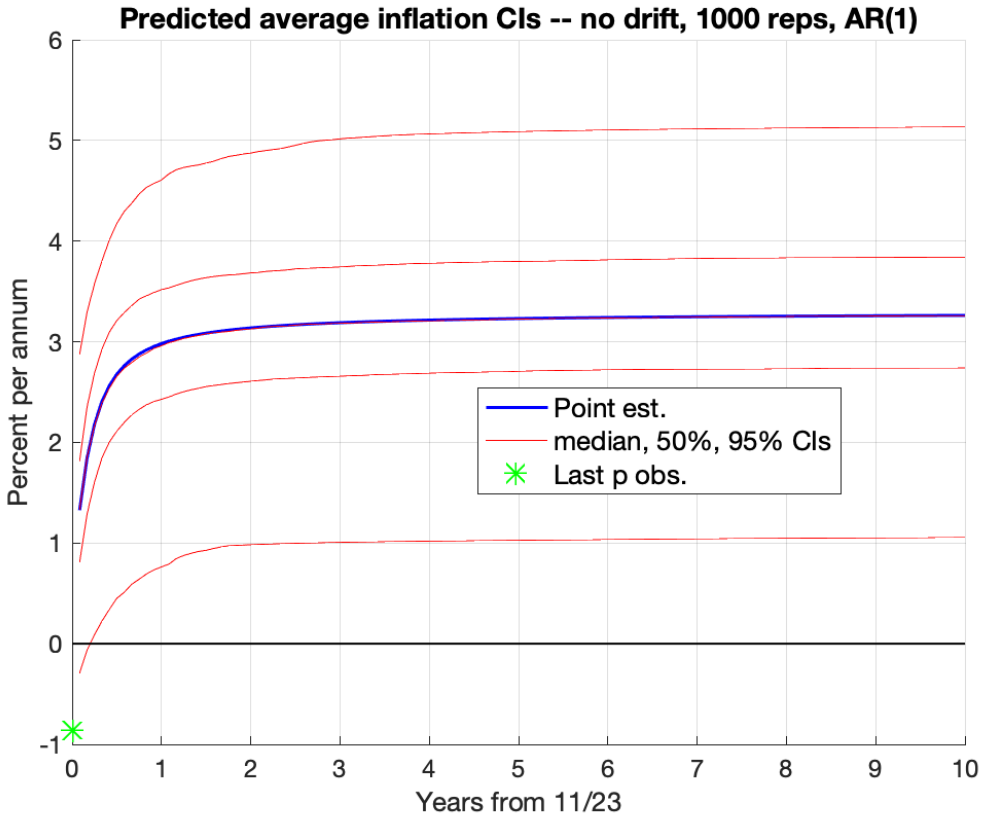

Figure 9
50% and 95% credible intervals for predicted average PCE inflation, AR(1) model with no drift, 1000 replications, Nov. 2023

Although $\mathbf{b}_N$ is precisely N($\mathbf{b}_N$, $\mathbf{P}_N$), given the model, the estimated hyperparameters, and setting aside the AR bias, the forecast uncertainty beyond $t = N+1$ is no longer Gaussian due to the nonlinearity of the dependence on the initial state variable. Thus the CIs in Figure 9 are increasingly leptokurtic and skewed as the horizon increases. At any finite horizon, this distribution is proper, with all quantiles finite, even if not all parameter draws are in the stationary region. However, the limiting distribution, computed from (31), will be improper, since there is always some posterior probability that the coefficients are non-stationary and therefore mass at $\pm\infty$ in the limiting distribution.

The thinner blue lines in Figure 10 add the effect of coefficient drift to the 20 illustrative simulations of Figure 8, which are now represented by thin dashed red lines for reference. For this purpose, $\mathbf{Q}_{N+1}$, based on $\mathbf{P}_N$, was used as the signal covariance matrix for all $t > N$. These simulations therefore abstract from the continuing changes in $\mathbf{Q}_{t+1}$ via $\mathbf{P}_t$ implied by the ALS model. Even though all 20 simulations started off in the

stationary region, several of them now drift into the nonstationary zone, with geometric explosions.

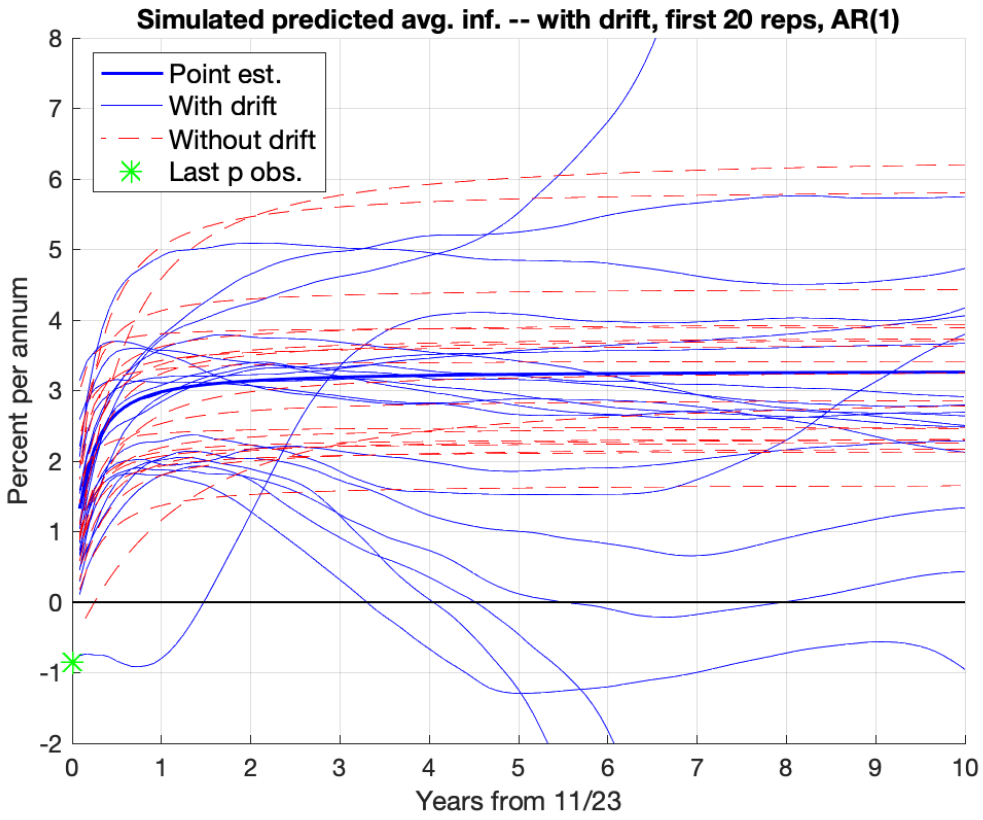

Figure 10
Simulated predicted average inflation, AR(1) model with drift (thin blue lines) and without drift (dashed red lines), first 20 replications.

The thinner solid blue lines in Figure 11 plot the posterior median, as well as the 50% and 95% credible intervals for predicted average inflation, with coefficient drift, using 1000 such replications. The thinner red dashed lines are the CIs without drift from Figure 9, for comparison. The darker blue line is the point forecast, and once again the posterior median is virtually indistinguishable from it. The 50% CI at 10 years is (2.1, 4.2) and still growing, and the 95% CI is already completely off scale within 8 years.

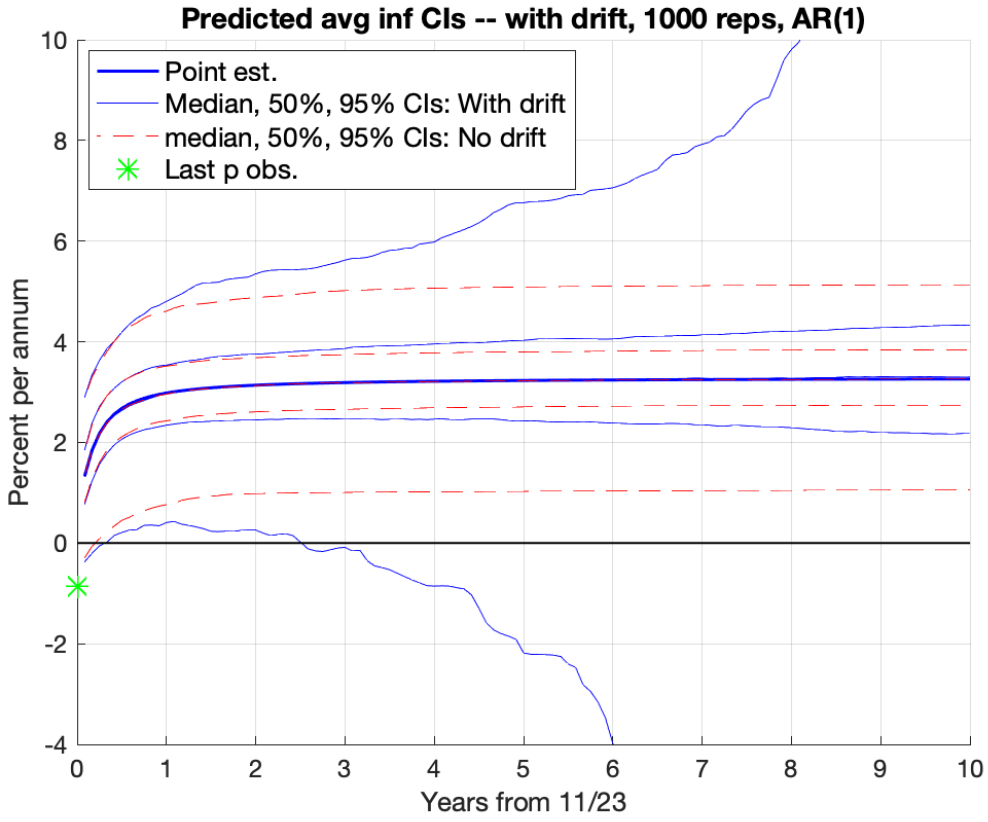

Figure 11
50% and 95% credible intervals for predicted average inflation, AR(1) model with drift (blue lines) and without drift (red dashed lines), 1000 replications.

Given the noise-to-signal ratio, the estimate of the noise variance $\hat{\sigma}_\varepsilon^2$ in (23) is governed by a chi-squared distribution with $N–k = 772$ degrees of freedom, and therefore is not a major source of uncertainty. However, Figure 12 plots the log likelihood, already maximized over the noise variance, versus the noise-to-signal standard deviation ratio *NSR* on a log scale. The vertical green bar is positioned at the ML estimate of 21.3. The horizontal red line is 1.92 units below the maximized likelihood, where the likelihood ratio (LR) statistic, twice the change in log likelihood, is just 3.84, the 5% critical value of the chi-square distribution with one degree of freedom. The LR-based 95% confidence interval for *NSR* is therefore a very considerable (14.2, 31.6). The present paper makes no attempt to quantify the effect of this uncertainty on the already large forecast uncertainty manifested in Figure 11 above.[5]

[5] Despite the quadratic appearance of Figure 12, the log likelihood is bounded below at $NSR = 0$ and ∞. The likelihood itself therefore integrates to ∞ over $\log(NSR)$ in $(-\infty, \infty)$. This would present a problem for a fully Bayesian analysis with a uniform prior on $\log(NSR)$.

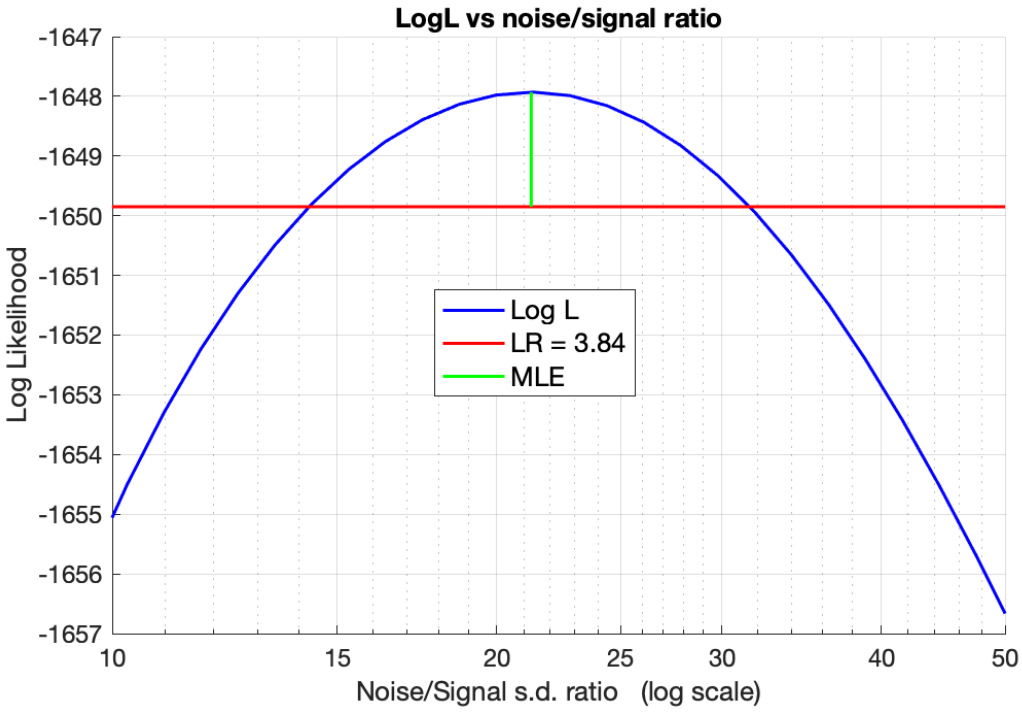

Figure 12
Log likelihood versus Noise/Signal standard deviation ratio *NSR* in AR(1) model (log scale).

Figure 13 below shows the standardized scale-adjusted residuals $u_t/\hat{\sigma}_\varepsilon$ for the preferred AR(1) model. Under the assumptions of the model and given the two hyperparameters, these should be i.i.d. N(0, 1), but the Jarque-Bera JB statistics in Table 1 soundly reject this hypothesis. There is some visual evidence of volatility clustering, but the several outliers in excess of 3 that appear without warning suggest that conditional non-normality is a larger problem than conditional heteroskedasticity.

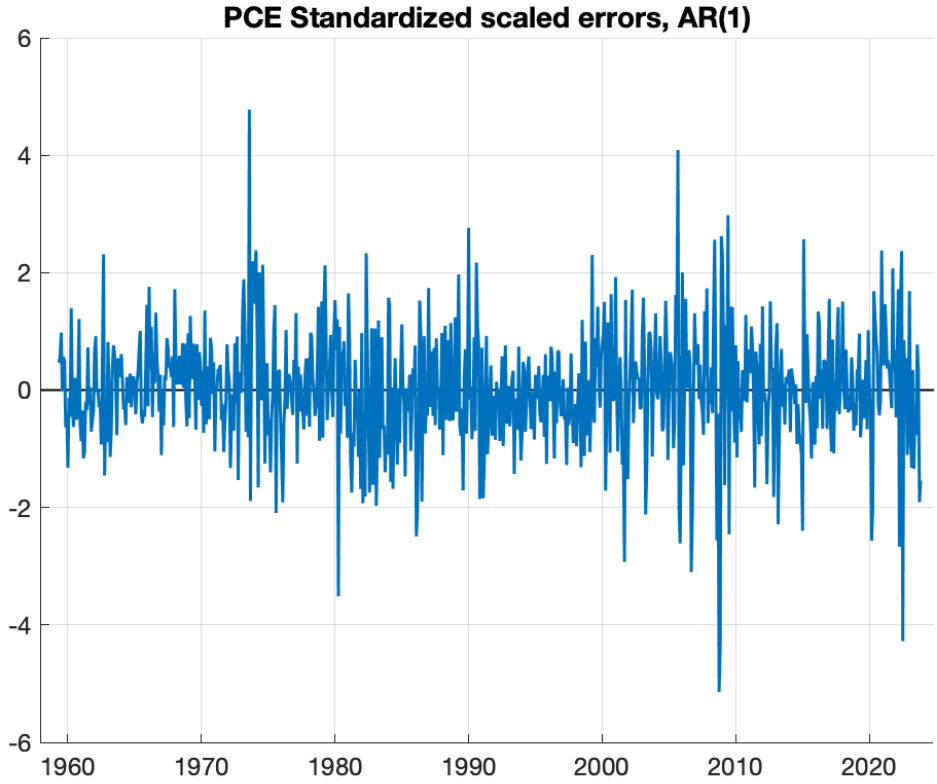

Figure 13
Scaled and standardized errors u_t in AR(1) model

VIII. Potential extensions and future applications.

A Vector Autoregression such as that run by Cogley and Sargent (2005) could easily be implemented by ALS with a common *NSR* for the entire system, simply by maximizing the sum of the log likelihoods across equations after orthogonalization of their residuals. The equation residuals may be orthogonalized by including, say, the residuals of the first equation as an additional regressor in the second equation, then the residuals of the first two equations in the third equation, and so on. This would essentially have the same effect as a Cholesky decomposition of the errors, but with coefficients that time-vary by the same rule that governs the lag coefficients. Alternatively, each equation could be given its own *NSR*.

In any ALS equation, care should be taken that the number of regressors be less than the long-run effective sample size N_{LR}, lest the regression coefficients be effectively unidentified. This is particularly of concern in a VAR, since the number of regressors increases with both the number of lags and the number of variables.

Clarida, Galí and Gertler (2000), Orphanides and Williams (2003), Kim and Nelson (2004), and others have found time variation in the "Taylor Rule" monetary policy response function. McCulloch (2007) makes a preliminary application of ALS to this problem, using the ALS filter to simulate real-time expectations of inflation and the unemployment gap, and then the ALS smoother to estimate the effective Taylor Rule in retrospect.

The Jarque-Bera JB statistics in Table 1 overwhelmingly reject the null of i.i.d. normality of the scaled forecasting errors, and therefore of the underlying noise and signal errors. Bidarkota and McCulloch (1998) estimate a Local Level Model of US inflation using heavy-tailed stable distributions in place of the Gaussian assumtion of the Kalman Filter, as well as a GARCH model of volatility clustering, but the numerical integrals required would quickly become intractable in the general TVP case. McCulloch (2021) develops a particle filter for the LLM with heavy-tailed stable errors. It is anticipated that such a particle filter can be generalized to an ALS model with stable errors and a constant noise/signal scale ratio.

Appendix A

A.1. The Local Level Model

The Local Level Model (2) implies

$$\mu_1 = y_1 - \varepsilon_1,$$

so that the distribution of μ_1 given y_1 and an uninformative prior may be written

$$\mu_1 | \, y_1 \sim \mathrm{N}(m_1, \sigma_1^2),$$

where

$$m_1 = y_1,$$
$$\sigma_1^2 = \sigma_\varepsilon^2.$$

Assume, as we now know to be the case for $t = 2$, that the distribution of the state variable μ_{t-1} given the observations $\mathbf{y}_{t-1} = (y_1, \ldots y_{t-1})'$ up to and including y_{t-1}, is likewise normal, with parameters

$$\mu_{t-1} \mid \mathbf{y}_{t-1} \sim \mathrm{N}(m_{t-1}, \sigma_{t-1}^2).$$

It follows that

$$\mu_t \mid \mathbf{y}_{t-1} \sim \mathrm{N}\left(m_{t-1}, \sigma_{t-1}^2 + \sigma_\eta^2\right) = \mathrm{N}(m_{t-1}, \sigma_{t-1}^2 + \rho\sigma_\varepsilon^2). \tag{32}$$

We also know that

$$y_t \mid \mu_t \sim \mathrm{N}(\mu_t, \sigma_\varepsilon^2).$$

Using Bayes' Rule as in Eqn. (3.7.24a) of Harvey (1989, p. 163), and completing the square with the appropriate constant term, we then have

$$\begin{aligned} p(\mu_t \mid \mathbf{y}_t) &= p(y_t \mid \mu_t, \mathbf{y}_{t-1})\, p(\mu_t \mid \mathbf{y}_{t-1}) / (const.) \\ &= p(y_t \mid \mu_t)\, p(\mu_t \mid \mathbf{y}_{t-1}) / (const.) \\ &= \exp\left(-\frac{1}{2} \frac{(y_t - \mu_t)^2}{\sigma_\varepsilon^2} \right) \exp\left(-\frac{1}{2} \frac{(\mu_t - m_{t-1})^2}{\sigma_{t-1}^2 + \rho\sigma_\varepsilon^2} \right) / (const.) \\ &= \exp\left(-\frac{1}{2} \frac{(\mu_t - m_t)^2}{\sigma_t^2} \right) / (const.), \end{aligned} \tag{33}$$

so that (3) is valid with

$$m_t = \frac{\sigma_t^2}{\sigma_\varepsilon^2} y_t + \frac{\sigma_t^2}{\sigma_{t-1}^2 + \rho\sigma_\varepsilon^2} m_{t-1} \tag{34}$$

and

$$\frac{1}{\sigma_t^2} = \frac{1}{\sigma_{t-1}^2 + \rho\sigma_\varepsilon^2} + \frac{1}{\sigma_\varepsilon^2}. \tag{35}$$

Defining $N_t = \sigma_\varepsilon^2 / \sigma_t^2$, (34) becomes (4) and (35) becomes (6), which may be initialized either with $N_0 = 0$ or $N_1 = 1$.

A.2. The TVP, Generalized RLS, and ALS Filters

The general TVP system (11) may similarly be solved recursively by means of the well-known Extended Kalman Filter (EKF). Assume that we have found a rule according to which

$$\boldsymbol{\beta}_{t-1} \mid \mathbf{y}_{t-1} \sim \mathrm{N}(\mathbf{b}_{t-1}, \mathbf{P}_{t-1}) \tag{36}$$

for some $k \times k$ covariance matrix $\mathbf{P}_{t-1}$ that may depend on $\mathbf{X}_{t-1}$, but not $\mathbf{y}_{t-1}$ or $\boldsymbol{\varepsilon}_{t-1}$. Then by Harvey (1989, pp. 105-6), or equivalently, Ljung and Söderström (1983, p. 420), and simplifying to the univariate random walk case (11) of interest,

$$\boldsymbol{\beta}_t \mid \mathbf{y}_t \sim \mathrm{N}(\mathbf{b}_t, \mathbf{P}_t),$$

where

$$\mathbf{b}_t = \mathbf{b}_{t-1} + f_t^{-1}\left(\mathbf{P}_{t-1} + \mathbf{Q}_t\right)\mathbf{x}_t'\left(y_t - \mathbf{x}_t\mathbf{b}_{t-1}\right), \tag{37}$$

$$\mathbf{P}_t = \left(\mathbf{P}_{t-1} + \mathbf{Q}_t\right)\left(\mathbf{I}_{k\times k} - f_t^{-1}\mathbf{x}_t'\mathbf{x}_t\left(\mathbf{P}_{t-1} + \mathbf{Q}_t\right)\right), \tag{38}$$

$$f_t = \mathbf{x}_t'\left(\mathbf{P}_{t-1} + \mathbf{Q}_t\right)\mathbf{x}_t' + \sigma_\varepsilon^2 \tag{39}$$

The textbook EKF equations (37) and (38) above may be rearranged to eliminate f_t and to look more like Recursive LS (RLS), as follows: Post-multiply (38) by $\mathbf{x}_t'$ and combine with (39) to obtain

$$\begin{aligned}\mathbf{P}_t\mathbf{x}_t' &= \left(\mathbf{P}_{t-1} + \mathbf{Q}_t\right)\left(\mathbf{x}_t' - f_t^{-1}\mathbf{x}_t'\left(f_t - \sigma_\varepsilon^2\right)\right)\\ &= \sigma_\varepsilon^2 f_t^{-1}\left(\mathbf{P}_{t-1} + \mathbf{Q}_t\right)\mathbf{x}_t',\end{aligned}$$

so that (37) becomes

$$\mathbf{b}_t = \mathbf{b}_{t-1} + \left(1/\sigma_\varepsilon^2\right)\mathbf{P}_t\mathbf{x}_t'\left(y_t - \mathbf{x}_t\mathbf{b}_{t-1}\right), \tag{40}$$

and (38) becomes

$$\mathbf{P}_t = \left(\mathbf{P}_{t-1} + \mathbf{Q}_t\right) - \left(1/\sigma_\varepsilon^2\right)\mathbf{P}_t\mathbf{x}_t'\mathbf{x}_t\left(\mathbf{P}_{t-1} + \mathbf{Q}_t\right).$$

Then multiply the last equation on the left by $\mathbf{P}_t^{-1}$ and on the right by $\left(\mathbf{P}_{t-1} + \mathbf{Q}_t\right)^{-1}$ and rearrange to obtain

$$\mathbf{P}_t^{-1} = \left(\mathbf{P}_{t-1} + \mathbf{Q}_t\right)^{-1} + \left(1/\sigma_\varepsilon^2\right)\mathbf{x}_t'\mathbf{x}_t. \tag{41}$$

The rearranged TVP filter (40), (41) may be placed in the even more convenient "information" or precision form, mentioned but not developed by Harvey (1989, p. 108), in terms of the scaled precision matrix

$$\mathbf{W}_t = \sigma_\varepsilon^2\mathbf{P}_t^{-1},$$

and the scaled signal covariance matrix

$$\mathbf{V}_t = (1/\sigma_\varepsilon^2)\mathbf{Q}_t,$$

as follows:

$$\mathbf{z}_t = \left(\mathbf{I} + \mathbf{W}_{t-1}\mathbf{V}_t\right)^{-1}\mathbf{z}_{t-1} + \mathbf{x}_t'y_t, \tag{42}$$

$$\mathbf{W}_t = \left(\mathbf{I} + \mathbf{W}_{t-1}\mathbf{V}_t\right)^{-1}\mathbf{W}_{t-1} + \mathbf{x}_t'\mathbf{x}_t, \tag{43}$$

whence $\mathbf{b}_t$ and $\mathbf{P}_t$ may be recovered by (15) and (16). Bullard (1992) uses a similar approach. Note that the observation error variance σ_ε^2 cannot be estimated until after the filter has been run, so that it must in any event be factored out of $\mathbf{P}_t$ and $\mathbf{Q}_t$ in order to run the filter.

In the absence of prior information about $\boldsymbol{\beta}_0$, the above information form filter may easily be initialized with a diffuse prior by taking the limit of $\mathbf{P}_0$ as all its eigenvalues go to infinity, or equivalently, by letting all the eigenvalues of the initial precision matrix $\mathbf{P}_0^{-1}$ go to zero, which in turn implies

$$\mathbf{W}_0 = \mathbf{0}_{k\times k}$$

as in (19). For any choice of $\mathbf{b}_0$, $\mathbf{z}_0 = \mathbf{W}_0\mathbf{b}_0$ then implies

$$\mathbf{z}_0 = \mathbf{0}_{k\times 1}.$$

It is not so obvious how to impose a diffuse prior on either (37) – (39) or (40) – (41), however.

With this diffuse prior, $\mathbf{W}_t$ is ordinarily of rank t for $t < k$, and hence $\mathbf{b}_t$ and $\mathbf{P}_t$ may not be computed by (15) and (16) until $t \geq k$. Note that in the fixed coefficient case $\mathbf{Q}_t = \mathbf{V}_t = \mathbf{0}_{k\times k}$, $\mathbf{z}_t$ becomes $\mathbf{X}_t'\mathbf{y}_t$, $\mathbf{W}_t$ becomes $\mathbf{X}_t'\mathbf{X}_t$, and (15) then becomes the familiar OLS formula, so that our diffuse prior is therefore implicit in OLS. (If any of the regressors is discrete, there is a chance that $\mathbf{X}_t$ and therefore $\mathbf{W}_t$ may still be singular for some $t \geq k$, but we have assumed here that this is never the case.)

Ljung (1992) and Sargent (193, 1999) observe that if $\mathbf{Q}_t$ is restricted to be proportional to $\mathbf{P}_{t-1}$, not only are there far fewer parameters to estimate, but the filter also simplifies greatly. If we set

$$\mathbf{Q}_t = k_t \mathbf{P}_{t-1} \tag{44}$$

for some constant k_t, then $\mathbf{V}_t = k_t \mathbf{W}_{t-1}^{-1}$, and (42) and (43) become following *Generalized* version of *Recursive Least Squares* (G-RLS):

$$\mathbf{z}_t = \frac{1}{1+k_t}\mathbf{z}_{t-1} + \mathbf{x}'_t\mathbf{y}_t, \tag{45}$$

$$\mathbf{W}_t = \frac{1}{1+k_t}\mathbf{W}_{t-1} + \mathbf{x}'_t\mathbf{x}_t. \tag{46}$$

Thus, the G-RLS class of restrictions reduce the matrix inversions in (42) and (43) to a single scalar inversion.

In order for G-RLS to nest the LLM, we need to choose k_t in such a way that the variance of the noise is in fixed proportion, in an appropriate sense, to that of the signal impacting the intercept term $\beta_{1,t}$. As in OLS, however, the magnitude and uncertainty of the intercept term will depend on the arbitrary manner in which the variable regressors $x_{2,t} \ldots x_{k,t}$ have been centered. In order to eliminate this arbitrariness, and at the same time to eliminate the effect of the slope coefficients on the intercept, we must center the variable regressors for each t in such a way that the covariance matrix of the transformed coefficients is block-diagonal on its first row and column. To this end, we define

$$\mathbf{x}_t^{\perp} = \mathbf{x}_t\mathbf{A}_t,$$

for

$$\mathbf{A}_t = \begin{pmatrix} 1 & \mathbf{p}_{t,1,2}\mathbf{P}_{t,2,2}^{-1} \\ \mathbf{0}_{k-1\times 1} & \mathbf{I}_{k-1} \end{pmatrix},$$

where

$$\mathbf{P}_t = \begin{pmatrix} p_{t,1,1} & \mathbf{p}_{t,1,2} \\ \mathbf{p}_{t,2,1} & \mathbf{P}_{t,2,2} \end{pmatrix},$$

so that the transformed coefficients are

$$\mathbf{b}_t^{\perp} = \mathbf{A}_t^{-1}\mathbf{b}_t,$$

with block-diagonal covariance matrix

$$\mathbf{P}_t^{\perp} = \mathrm{Cov}(\mathbf{b}_t^{\perp}) = \mathbf{A}_t^{-1}\mathbf{P}_t\mathbf{A}_t^{-1\prime} = \begin{pmatrix} \mathbf{p}_{t,1,2}\mathbf{P}_{t,2,2}^{-1}\mathbf{p}_{t,2,1} & \mathbf{0}_{1\times k-1} \\ \mathbf{0}_{k-1\times 1} & \mathbf{P}_{t,2,2} \end{pmatrix}.$$

Setting

$$\mathbf{W}_t = \sigma_\varepsilon^2\mathbf{P}_t^{-1} = \begin{pmatrix} w_{t,1,1} & \mathbf{w}_{t,1,2} \\ \mathbf{w}_{t,2,1} & \mathbf{W}_{t,2,2} \end{pmatrix},$$

and applying the block matrix inversion formula to $\mathbf{P}_t$, we have

$$\mathbf{P}_t^{\perp} = \begin{pmatrix} \sigma_\varepsilon^2 w_{t,1,1} & \mathbf{0}_{1\times k-1} \\ \mathbf{0}_{k-1\times 1} & \mathbf{P}_{t,2,2} \end{pmatrix}.$$

Now if $\mathbf{Q}_t = k_t \mathbf{P}_{t-1}$, we will also have $\mathbf{Q}_t^{\perp} = k_t\mathbf{P}_{t-1}^{\perp}$, where $\mathbf{Q}_t^{\perp} = \mathbf{A}_{t-1}^{-1}\mathbf{Q}_t\mathbf{A}_{t-1}^{-1\prime}$ is the covariance matrix of the appropriately transformed time t transition errors $\boldsymbol{\eta}_t^{\perp} = \mathrm{A}_{\boldsymbol{t-1}}^{-\mathbf{1}}\boldsymbol{\eta}_t^{\perp}$. Therefore if, as in the LLM, the variance of the shock to the orthogonalized intercept is a constant ρ times the noise variance,

$$q_{t,1,1}^{\perp} = \rho\sigma_\varepsilon^2,$$

it follows that

$$k_t = \rho w_{t-1,1,1}^{\perp} = \rho w_{t-1,1,1},$$

whence (46) implies that $w_{t,1,1}$ obeys the same recursion as the LLM's effective sample size N_t in (6). Furthermore, the diffuse prior (19) implies $w_{0,1,1} = 0$, just as $N_0 = 0$ in the LLM. Setting $w_{t,1,1} = N_t$, we obtain the ALS updating equations (17) and (18) with $k_t = \rho N_{t-1}$, as claimed in (14) in the text.

A.3. The TVP and ALS Smoother

In order to obtain "smoother," or "two-sided filter," estimates of the coefficients, conditional on the *entire* data set, we first run the Information Filter backwards from the end of the data set, so as to obtain estimates $\mathbf{b}_t^*$ of $\boldsymbol{\beta}_t$ conditional on y_t, ... y_N and no other information, with variances $\mathbf{P}_t^*$. In the general TVP case, this backward filter may be computed by:

$$\mathbf{z}_{n+1}^* = \mathbf{0}_{k\times 1}$$

$$\mathbf{W}_{n+1}^* = \mathbf{0}_{k\times k}$$

$$\mathbf{z}_t^* = \left(\mathbf{I} + \mathbf{W}_{t+1}^*\mathbf{V}_{t+1}\right)^{-1}\mathbf{z}_{t+1}^* + \mathbf{x}_t' y_t, \tag{47}$$

$$\mathbf{W}_t^* = \left(\mathbf{I} + \mathbf{W}_{t+1}^*\mathbf{V}_{t+1}\right)^{-1}\mathbf{W}_{t+1}^* + \mathbf{x}_t'\mathbf{x}_t, \tag{48}$$

$$\mathbf{b}_t^* = \mathbf{W}_t^{*-1}\mathbf{z}_t^*, \quad t \le N-k, \tag{49}$$

$$\mathbf{P}_t^* = \sigma_\varepsilon^2 \mathbf{W}_t^{*-1}, \quad t \le N-k. \tag{50}$$

The backward filter $\mathbf{b}_{t+1}^*$ obtained in this manner estimates $\boldsymbol{\beta}_{t+1}$, conditional on $y_{t+1}, \ldots y_N$, with variance $\mathbf{P}_{t+1}^*$, but it also provides an estimate of $\boldsymbol{\beta}_t$, conditional on the same values, with the somewhat larger variance $\mathbf{P}_{t+1}^* + \mathbf{Q}_{t+1} = \sigma_\varepsilon^2 \mathbf{W}_{t+1}^{*-1}\left(\mathbf{I} + \mathbf{W}_{t+1}^* \mathbf{V}_{t+1}\right)$. Since $\mathbf{b}_{t+1}^*$ as an estimate of $\boldsymbol{\beta}_t$ is independent of the filter estimate $\mathbf{b}_t$, the two estimates may be averaged in proportion to their respective precision matrices to form the smoother estimate $\mathbf{b}_t^S$ of $\boldsymbol{\beta}_t$, conditional on the entire sample, as follows:

$$\begin{aligned}\mathbf{b}_t^S &= \mathbf{W}_t^{S-1}\left(\mathbf{W}_t \mathbf{b}_t + \left(\left(\mathbf{I} + \mathbf{W}_{t+1}^* \mathbf{V}_{t+1}\right)^{-1} \mathbf{W}_{t+1}^*\right)\mathbf{b}_{t+1}^*\right) \\ &= \mathbf{W}_t^{S-1}\mathbf{z}_t^S,\end{aligned} \tag{51}$$

where

$$\mathbf{W}_t^S = \mathbf{W}_t + \left(\mathbf{I} + \mathbf{W}_{t+1}^* \mathbf{V}_{t+1}\right)^{-1} \mathbf{W}_{t+1}^*, \tag{52}$$

$$\mathbf{z}_t^S = \mathbf{z}_t + \left(\mathbf{I} + \mathbf{W}_{t+1}^* \mathbf{V}_{t+1}\right)^{-1} \mathbf{z}_{t+1}^*. \tag{53}$$

This smoother estimate has variance

$$\mathbf{P}_t^S = \sigma_\varepsilon^2 \mathbf{W}_t^{S-1}. \tag{54}$$

Note that it is not necessary to actually compute the backward filter $\mathbf{b}_t^*$, $\mathbf{P}_t^*$ itself, however, since $\mathbf{z}_t^*$ and $\mathbf{W}_t^*$ suffice to obtain the smoother and its variance using (51) – (54). The smoother and its variance may therefore be computed even for $t > N–k$, even though the backward filter is not defined there.

In order to compute the smoother, it is necessary to save $\mathbf{z}_t$ and $\mathbf{W}_t$ for all t on the forward filter pass. However, since the smoother is not needed to compute the likelihood and estimate the two hyperparameters, there is no point in computing it except on a final pass.

To obtain the smoother in the ALS case, we must set

$$\mathbf{V}_{t+1} = \rho N_t \mathbf{W}_t^{-1}$$

in (47) - (54), in order to be consistent with the forward filter. If desired, the term $(\mathbf{I} + \rho N_t \mathbf{W}_{t+1}^* \mathbf{W}_t^{-1})^{-1}$ may then be replaced by $\mathbf{W}_t(\mathbf{W}_t + \rho N_t \mathbf{W}_{t+1}^*)^{-1}$ to avoid having to invert $\mathbf{W}_t$. In the general TVP case, where the transition covariance matrix $\mathbf{Q}_t$ is non-singular for all t, we may compute the smoother clear back to $t = 1$. In the ALS case, however, $\mathbf{W}_t^S$, which may be written, using the above substitution, as $\mathbf{W}_t(\mathbf{I} + (\mathbf{W}_t + \rho N_t \mathbf{W}_{t+1}^*)^{-1}\mathbf{W}_{t+1}^*)$, is proportional to $\mathbf{W}_t$ and therefore singular for $t < k$. The ALS smoother, like its filter, is therefore defined only for $t \ge$ k. Unfortunately, the serendipitous cancellation that occurs in the filter equations is no longer present, so that the ALS smoother will run somewhat slower than the ALS filter.

A.4 The Global Test of Significance

The (N–k) x (N–k) x k x k intertemporal covariance matrix **C** defined in (29) may be found as part of a large Generalized Least Squares (GLS) problem that solves directly for the smoother coefficients without using the Kalman recursions:

For t = 1, ..., k–1, the $\boldsymbol{\beta}_t$ are unidentified under the ALS specification. Nevertheless, the ancillary variables

$$\zeta_t = \mathbf{x}_t \beta_t$$

are identified, with observation equation

$$y_t = \zeta_t + \varepsilon_t. \tag{55}$$

Set $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_k)'$.

For t = k, ..., N, we have the standard observation equation

$$y_t = \mathbf{x}_t \beta_t + \varepsilon_t. \tag{11}$$

For t = 1, ..., k–1, the transition equation (12) relates the ancillary variable ζ_t to β_k via

$$\mathbf{x}_t \beta_k = \zeta_t + \delta_t, \tag{56}$$

where

$$\delta_t = \mathbf{x}_t \textstyle\sum_{s=t+1}^{k} \boldsymbol{\eta}_s ,$$

so that for $t' \geq t$,

$$\begin{aligned} \operatorname{cov}(\delta_t, \delta_{t'}) &= \sigma_\varepsilon^2 \, \mathbf{x}_t \left(\textstyle\sum_{s=t'+1}^{k} \mathbf{V}_t \right) \mathbf{x}_{t'}' \\ &= \sigma_\varepsilon^2 \, \xi_{t,t'} . \end{aligned}$$

Set $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{k-1})'$. Define the ($k$–1) x ($k$–1) matrix $\boldsymbol{\Xi} = \left(\xi_{t,t'}\right)$, imposing symmetry. (In the empirical examples given in the text, it was found that the off-diagonal elements of $\boldsymbol{\Xi}$ were all on the order of e–18, or within computational error of zero, but this was not imposed.)

For t = k+1, ..., N, we have the standard transition equation

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \;\; \boldsymbol{\eta}_t \sim \text{NID}(\mathbf{0}_{k\times 1}, \boldsymbol{\sigma}_\varepsilon^2 \mathbf{V}_t) \; . \tag{12}$$

The above N+k–1+k(N–k) equations in 2k–1+k(N–k) unknowns may be stacked into the matrix equation

$$\boldsymbol{\phi} = \boldsymbol{\Psi}\boldsymbol{\theta} + \boldsymbol{\upsilon} \; ,$$

where

$$\boldsymbol{\phi} = \begin{pmatrix} \mathbf{y}_N \\ \mathbf{0}_{k-1+k(N-k)\times 1} \end{pmatrix}$$

$$\boldsymbol{\Psi} = \begin{pmatrix} \mathbf{0}_{(k-1)\times(k(N-k+1))} & & \mathbf{I}_{k-1} \\ \mathbf{x}_k & \cdots & \mathbf{0}_{k\times k} & \\ \vdots & \ddots & \vdots & \mathbf{0}_{(N-k+1)\times(k-1)} \\ \mathbf{0}_{k\times k} & \cdots & \mathbf{x}_N & \\ -\mathbf{X}_{\boldsymbol{k-1}} & \mathbf{0}_{(k-1)\times(k(N-k))} & \mathrm{I}_{k-1} & \\ \mathrm{I}_k & -\mathrm{I}_k & \cdots & \mathbf{0}_{k\times k} \\ \vdots & \ddots & \ddots & \vdots & \mathbf{0}_{k(N-k)\times(k-1)} \\ \mathbf{0}_{k\times k} & \cdots & \mathrm{I}_k & -\mathrm{I}_k \end{pmatrix},$$

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\beta}_k \\ \vdots \\ \boldsymbol{\beta}_N \\ \boldsymbol{\zeta} \end{pmatrix},$$

$$\boldsymbol{\upsilon} = \begin{pmatrix} \boldsymbol{\varepsilon} \\ \boldsymbol{\delta} \\ \boldsymbol{\eta}_{k+1} \\ \vdots \\ \boldsymbol{\eta}_N \end{pmatrix}.$$

The covariance matrix of the error vector is

$$\mathrm{Cov}(\boldsymbol{\upsilon}) = \sigma_\varepsilon^2 \boldsymbol{\Omega}\,,$$

where

$$\boldsymbol{\Omega} = \begin{pmatrix} \mathbf{I}_N & \mathbf{0}_{N\times k-1} & \mathbf{0}_{N\times k(N-k)} \\ \mathbf{0}_{k-1\times N} & \boldsymbol{\Xi} & \mathbf{0}_{k-1\times k(N-k)} \\ & & \mathbf{V}_{k+1} \quad \cdots \quad \mathbf{0}_{k\times k} \\ \mathbf{0}_{k(N-k)\times N} & \mathbf{0}_{k(N-k)\times k-1} & \vdots \quad \ddots \quad \vdots \\ & & \mathbf{0}_{k\times k} \quad \cdots \quad \mathbf{V}_N \end{pmatrix}.$$

The GLS estimator of the coefficient vector is then

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{\Psi}'\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}'\boldsymbol{\Omega}^{-1}\boldsymbol{\phi}\,,$$

with covariance matrix

$$\boldsymbol{\Gamma} = \mathrm{Cov}\left(\widehat{\boldsymbol{\theta}}\right) = \sigma_\varepsilon^2(\boldsymbol{\Psi}'\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi})^{-1}.$$

The elements of **C** required for the global significance tests may easily be extracted from **Γ.** These include redundant values of the smoother covariance matrices $\mathbf{P}_t^S$. It was found that these differ from the values already obtained recursively by at most 2.2e–13. Likewise, the redundant values of the smoother coefficients $\mathbf{b}_t^S$ contained in $\widehat{\boldsymbol{\theta}}$ differ from the recursive values by at most 1.9e–12.

In order to investigate the power of the global test of significance proposed in Section V in the text, 1000 autocorrelated sequences $\langle x_i \rangle$ of length 1000 were generated with autocovariance structure

$$\mathrm{cov}\left(x_i, x_j\right) = 0.5^{(|i-j|/20)^{1.5}}$$

and mean μ. The factor of 20 implies that at lag 20, which may be thought of as the "bandwidth" of the process, the autocorrelation will be 0.5. This is similar to what would be expected of the estimation errors in an ALS model with *NSR* = 20. The power of 1.5

makes the simulated process have smoothness similar to that of the ALS smoother, while retaining full rank in the covariance matrix. Each sample was subdivided into $n_T = 1000/\Delta t$ intervals of size Δt for Δt = 1, 2, 5, 10, ... 1000, and a subset T of the full sample was selected using the approximate midpoints of these intervals, so that the χ^2 test for $\mu = 0$ using subset T has n_T = 1000, 500, 200, 100, ..., 1 degrees of freedom.

When $\mu = 0$, the test was found to have the correct size, and power equal to its size, regardless of Δt. However, when $\mu = 1$ as shown in Figure 14, the power is highest for step size $\Delta t = 50$, with 20 and 100 close behind. (Power = 0.719, 0.613 and 0.544 for test size 0.05, resp.) For test sizes below 0.33, using the full sample ($\Delta t = 1, n_T = 1000$) is actually worse than using only one, centrally located smoother value ($\Delta t = 1000, n_T = 1$). (Power = 0.142 and 0.182 for test size 0.05, resp.)

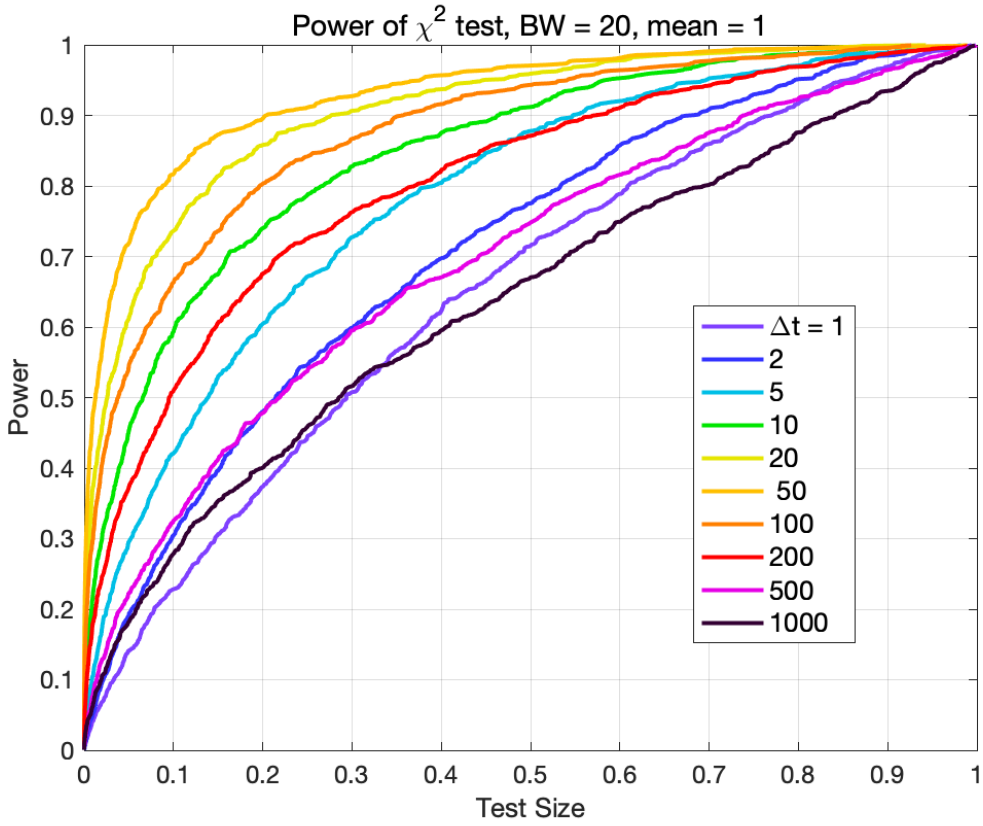

Figure 14
Power as a function of test size for the global test of significance, with bandwidth = 20 and true mean = 1.

In the absence of serial correlation, it is of course optimal to use as many degrees of freedom as possible in global significance tests. However, when there is serial correlation and the true mean is non-zero, the transformation that whitens the observations simultaneously moves the mean of the series closer to zero, making the non-

zero mean harder to detect. For example, the mean of the transformed data is 0.148 in the above example when the full sample is used, even though the mean of the raw data is 1.0. There therefore is a tradeoff between degrees of freedom and independence. A sampling interval approximately twice the bandwidth, and therefore approximately twice the *NSR*, is therefore recommended for global significance tests.

A.5. The error in Ljung (1992) and Sargent (1999)

As mentioned in the text, there is an error in the Kalman Filter as presented in Sargent's (1999) equation (94). This error led Sargent to wrongly assert that RLS is only an approximate implication of his basic assumption that $\mathbf{Q}_t$ is directly proportional to $\mathbf{P}_{t-1}$, when in fact it is an exact implication. To correct this error, $\mathbf{P}_{t-1}$ in Sargent's (94b) and in the term after the minus sign in (94c) should be replaced with $\mathbf{P}_{t-1} + \mathbf{R}_{1t}$ in Sargent's notation, i.e. by $\mathbf{P}_{t-1} + \mathbf{Q}_t$ in ours and Harvey's.

The same error appears in the source Sargent cites, namely Ljung (1992), equations (36) – (39). Nevertheless, Ljung's own source, Ljung and Söderström (LS, 1983) is correct.

LS consider a more general case of the KF than is used here or in Sargent or Ljung, one which permits the coefficient vector to follow a stationary matrix AR(1) process with a driving process, rather than just a random walk as in (11) of the present paper. Harvey treats a similarly general case. In this more general case, it is expedient to introduce, as Harvey does, a notation like $\mathbf{b}_{t|t-1}$ to indicate the expectation of $\boldsymbol{\beta}_t$ conditional on $\mathbf{y}_{t-1}$, and $\mathbf{P}_{t|t-1}$ for its covariance matrix, in addition to $\mathbf{b}_t$, $\mathbf{b}_{t-1}$, $\mathbf{P}_t$, and $\mathbf{P}_{t-1}$.

In terms of the Harvey conditional subscripts, but our symbols otherwise, Ljung and Söderström's (1.C.14) – (1.C.16) on p. 420 become, in the special case of interest,

$$\mathbf{b}_{t+1|t} = \mathbf{b}_{t|t-1} + \mathbf{K}(t)(y_t - \mathbf{x}_t \mathbf{b}_{t|t-1}) \tag{57}$$

$$\mathbf{K}(t) = \mathbf{P}_{t|t-1}\mathbf{x}_t'\left(\mathbf{x}_t \mathbf{P}_{t|t-1}\mathbf{x}_t' + \sigma_\varepsilon^2\right)^{-1} \tag{58}$$

$$\mathbf{P}_{t+1|t} = \mathbf{P}_{t|t-1} + \mathbf{Q}_{[t+1]} - \mathbf{P}_{t|t-1}\mathbf{x}_t'\mathbf{x}_t\mathbf{P}_{t|t-1}\left(\mathbf{x}_t \mathbf{P}_{t|t-1}\mathbf{x}_t' + \sigma_\varepsilon^2\right)^{-1}. \tag{59}$$

Since in the random walk case, $\mathbf{b}_{t+1|t}$ becomes our $\mathbf{b}_t$ and $\mathbf{P}_{t|t-1}$ becomes our $\mathbf{P}_{t-1} + \mathbf{Q}_t$, and (57) – (59) are equivalent to (37) – (39) above, which in turn derive from Harvey's (3.2.3a) – (3.2.3c). Thus, Harvey and LS are in agreement.

However, LS do not use Harvey's conditional subscript notation, but instead refer to the expectation of their time *t* coefficient vector "$\mathbf{x}_t$," conditional on information up to and including *t*-1 (i.e. $\mathbf{b}_{t|t-1}$ above), simply as "$\hat{\mathbf{x}}(t)$," and to its covariance matrix ($\mathbf{P}_{t|t-1}$ above) simply as "$\mathbf{P}(t)$," etc. The source of the error in Ljung (1992) and thence Sargent (1999) is that when Ljung simplified (1.C.14) – (1.C.16) in LS to the random walk case, he redefined "$\hat{\mathbf{x}}(t)$" to be the expectation of the time *t* coefficient vector conditional on information up to and including time *t*, i.e. our $\mathbf{b}_t$, and "$\mathbf{P}(t)$" to be *its* covariance matrix, i.e. our $\mathbf{P}_t$. In making this notational revision, however, he simply replaced "$\mathbf{P}(t)$" in his

former notation, at all but one point, with "$\mathbf{P}(t-1)$", instead of with $\mathbf{P}_{t|t-1} = \mathbf{P}_{t-1} + \mathbf{Q}_t$, i.e. "$\mathbf{P}(t-1) + \mathbf{R}_1(t)$" in terms of his new notation, as he should have.[6]

In order to correct equations (36) – (39) in Ljung (1992), therefore, "$\mathbf{P}(t-1)$" in his (38) and in the expression after the minus sign in his (39) should be replaced with "$\mathbf{P}(t-1) + \mathbf{R}_1(t)$." Corresponding replacements should be made in Sargent's (1999) equation (94), as noted above.

In correspondence, Lennart Ljung has kindly indicated that he in fact intended the "$\mathbf{P}(t-1)$" of his 1992 book to be $\mathbf{P}_{t|t-1}$, despite the apparently contrary definition given in his text which led Sargent (1999) to interpret it as $\mathbf{P}_{t-1|t-1}$. However, he points out that even with this interpretation there is an error, since then the $\mathbf{R}_1(t)$ in the first part of (39) on his p. 99 should not be present.

[6] Note that whereas Ljung (1992) associates subscript t with the change in the coefficient vector between times $t-1$ and t, this subscript is $t-1$ in LS. Although LS do not explicitly date the covariance $\mathbf{R}_1$ of this change, if they had, the "$\mathbf{R}_1(t)$" of Ljung (1992) would therefore have been "$\mathbf{R}_1(t-1)$" in the LS notation.

REFERENCES

Bidarkota, Prasad V., and J. Huston McCulloch, "Optimal Univariate Inflation Forecasting with Symmetric Stable Shocks," *Journal of Applied Econometrics* **13** (1998), 659-70.

Bullard, James. "Time-Varying Parameters and Nonconvergence to Rational Expectations under Least Squares Learning," *Economics Letters* **40** (1992): 159-66.

Bullard, James, and John Duffy. "Learning and Structural Change in Macroeconomic Data," St. Louis Fed and University of Pittsburgh, 2003. Online at <http://research/stlouisfed.org/econ/bullard/ltdmd2002march23.pdf>.

Bullard, James, and Kaushik Mitra. "Learning about Monetary Policy Rules," *Journal of Monetary Economics* **49** (2002), 1105-1129.

Cagan, Phillip. "The Monetary Dynamics of Hyperinflation," in Friedman, ed., *Studies in the Quantity Theory of Money,* University of Chicago Press, 1956.

Clarida, Richard, Jordi Galí and Mark Gertler. "Monetary Policy Rules and Macroeconomic Stability: Evidence and Some Theory," *Quarterly Journal of Economics* **115** (2000), 147-180.

Cogley, Timothy, and Thomas J. Sargent. "Drifts and Volatilities: Monetary Policies and Outcomes in the Post WWII U.S.," *Review of Economic Dynamics* **8** (2005): 262-302.

Cooley, Thomas F., and Edward C. Prescott. "An Adaptive Regression Model," *International Economic Review* **14** (1973), 364-71. .

Dickey, D.A., and W.A. Fuller, "Distribution of the Estimators for Autoregressive Time Series with a Unit Root," *Journal of the American Statistical Association* **74** (366) (1979): 427-31.

Durbin, James, and S.J. Koopman, *Time Series Analysis by State Space Methods*. Oxford University Press, 2001.

Evans, George W., and Seppo Honkapohja. *Learning and Expectations in Macroeconomics*. Princeton University Press, 2001.

Harvey, Andrew C. *Forecasting, Structural Times Series Models and the Kalman Filter*. Cambridge University Press, 1989.

Kalman, R.E. "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering, Transactions ASME Series D* **82** (1960): 35-45.

Kim, Chang-Jin, and Charles Nelson. "Estimation of a Forward-Looking Monetary Policy Rule: A Time-Varying Parameter Model using Ex-Post Data," Korea University and University of Washington, 2004.

Klein, Benjamin. "The Measurement of Long- and Short-Term Price Uncertainty: A Moving Regression Time Series Analysis," *Economic Inquiry* **16** (1978), 438-52.

Ljung, Lennart. "Applications to Adaptive Algorithms," in L. Ljung, Georg Pflug, and Harro Walk, *Stochastic Approximations and Optimization of Random Systems*, Birkhäuser, 1992, pp. 95-113.

Ljung, Lennart, and Torsten Söderström. *Theory and Practice of Recursive Identification*, MIT Press, 1983.

McCulloch, J. Huston. "Adaptive Least Squares Estimation of the Time-Varying Taylor Rule," June 6, 2007, <www.asc.ohio-state.edu/mcculloch.2/papers/TaylorALS.pdf>

McCulloch, J. Huston. "Moment Ratio Estimation of Autoregressive/Unit Root Processes and Autocorrelation-Consistent Standard Errors," *Computational Statistics and Data Analysis* 100 (Aug. 2016): 712-733. DOI: 10.1016/j.csda.2015.07.003

McCulloch, J. Huston. "State Estimation with Stable Errors and Whisker Particles," March 10, 2021, <www.asc.ohio-state.edu/mcculloch.2/papers/ParticleFilter/StableParticleFilter.pdf>

McGough, Bruce. "Statistical Learning with Time-Varying Parameters," *Macroeconomic Dynamics* **7** (2003): 119-139.

Milani, Fabio, "Adaptive Learning and Inflation Persistence," Princeton University, 2005, <econwpa.wustl.edu/eps/mac/papers/0506/0506013.pdf>.

Moran, P. A. P. "The Uniform Consistency of Maximum-Likelihood Estimators," *Proceedings of the Cambridge Philosophical Society* **17** (1971a), 435-39.

Moran, P. A. P. "Maximum-Likelihood Estimation in Non-Standard Conditions," *Proceedings of the Cambridge Philosophical Society* **17** (1971b) 441-50.

Muth, John F. "Optimal Properties of Exponentially Weighted Forecasts," *J. of the American Statistical Assn. (JASA)* 1960, 299-306.

Orphanides, Athanasios, and John C. Williams, "The Decline of Activist Stabilization Policy: Natural Rate Misperceptions, Learning, and Expectations," Federal Reserve Board WP 2004-804, dated 12/2003, <http://www.federalreserve.gov/pubs/ifdp/2004/804/default.htm>.

Powell, Jerome H. "New Economic Challenges and the Fed's Monetary Policy Review," Aug. 27, 2020, <https://www.federalreserve.gov/newsevents/speech/powell20200827a.htm>m

Preston, Bruce. "Adaptive Learning, Forecast-Based Instrument Rules and Monetary Policy," Columbia University, 2004, online at <http://www.columbia.edu/~bp2121/targetrules.pdf>.

Sargent, Thomas J. "A Note on the 'Accelerationist' Controversy," *J. of Money, Credit and Banking* **3** (1971): 721-5.

Sargent, Thomas J. *Bounded Rationality in Macroeconomics.* Clarendon Press, Oxford, 1993.

Sargent, Thomas J. *The Conquest of American Inflation*. Princeton University Press, 1999.

Sargent, Thomas J., and Noah Williams. "Impacts of Priors on Convergence and Escapes from Nash Inflation," *Review of Economic Dynamics* **8** (2005): 360-91.

Sims, Christopher. "Projecting Policy Effects with Statistical Models," *Revista de Analisis Economico* **3** (1988), 3-20.

Stock, James H., and Mark W. Watson. "Evidence on Structural Instability in Macroeconomic Time Series Relations," *Journal of Business & Economic Statistics* **14** (1996): 11-30.

Tanaka, Katsuto. "Non-Normality of the Lagrange Multiplier Statistic for Testing the Constancy of Regression Coefficient," *Econometrica* **51** (1983): 1577-82.