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

J. Huston McCulloch

Jan. 26, 2022

The author is Professor Emeritus at Ohio State University and Adjunct Professor at New York University. He is indebted to James Bullard, Kan Chen, Chang-Jin Kim, Young-Il Kim, Lennart Ljung, Bruce McGough, Athanasios Orphanides, Thomas Sargent, Yasushi Sugayama, and participants at the OSU Money/Macro Seminar and the 2005 conference on Computation in Economics and Finance 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.

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 method of estimating time-varying relationships and of proxying agents’ time-evolving expectations. By holding the noise-to-signal ratio constant rather than the Kalman gain as in CG-RLS, it nests the univariate Local Level Model, 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 algorithm is illustrated with a univariate time series model of PCE inflation through Nov. 2021, when there had just been a run of high inflation months. With the preferred AR(2) model, the estimated noise-to-signal standard deviation ratio is 28.4 months, implying an asymptotic gain of $1/28.9 \text{ mo}^{-1}$. The sum of the AR coefficients varies over the decades from near-unit root behavior to virtually zero, with often substantial difference between short-run and long-run inflation forecasts. Although the one-month-ahead forecast from Nov. 2021 was 5.68%, the estimated "entrenched" long-run inflation rate then was only 3.10%.

An ALS smoother, that uses both past and future data, is developed. 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, suggesting that a particle filter with non-Gaussian errors would be an appropriate extension.

Keywords: Adaptive Learning, Kalman Filter, Inflation

JEL Codes: C32 -- Time Series Models
            E31 -- Inflation
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. Bullard and Mitra (2002), Bullard and Duffy (2003), Evans and Honkapohja (2001, 2004), Milani (2005), Orphanides and Williams (2003), Preston (2004) and Sargent (1999) 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

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

In Cagan’s original formulation, the gain coefficient $\gamma$ was assumed to be an arbitrary subjective constant to be inferred indirectly from agents’ expectationally motivated behavior, e.g. 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 $\gamma$ is computed as a certain function of the noise/signal ratio. The gain coefficient is therefore not an arbitrary subjective learning parameter akin to a demand elasticity, but rather takes 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 rigorous gain is not constant, but in fact declines rapidly at the beginning of the sample. Kalman’s analysis 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 exposito, for example, by Harvey (1989, Ch. 3). Ljung (1992) and Sargent (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 declining-yet-bounded-gain LLM. 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.

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 an approximate 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 briefly mentions potential future applications and extensions. The Appendix provides mathematical details, develops an ALS smoother algorithm, and corrects an error in a critical equation in 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 concept of Effective Sample Size.

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

$$y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, \sigma^2_\varepsilon),$$

$$\mu_t = \mu_{t-1} + \eta_t, \quad \eta_t \sim \text{NID}(0, \sigma^2_\eta).$$

The signal/noise variance ratio is defined to be

$$\rho = \frac{\sigma^2_\eta}{\sigma^2_\varepsilon},$$

so that the two "hyperparameters" $\sigma^2_\varepsilon$ and $\rho$ completely determine the system. Let the vector $y_t = (y_1, \ldots, y_t)'$ 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:

$$\mu_t | y_t \sim \text{N}(m_t, \sigma^2_t).$$

with

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

$$\sigma^2_t = \gamma_t \sigma^2_\varepsilon,$$
and where the gain $\gamma$ is the reciprocal of the effective sample size $N_t$, determined by
\[
N_t = (1 + \rho N_{t-1})^{-1}N_{t-1} + 1, \tag{5}
\]
with uninformative initialization
\[
N_0 = 0. \tag{6}
\]
The expectation of $y_t'$ for all $t' > t$ is then $m_t$.

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
\[
N_{\text{LR}} = \lim_{t \to \infty} N_t = 1/2 + \sqrt{1/4 + 1/\rho} \tag{7}
\]
the unique positive root of the quadratic equation
\[
\rho T^2 - \rho T - 1 = 0
\]
that determines the fixed points of (5). The constant gain AE formula (1) is therefore valid only in this limit, with the limiting long run gain $\gamma_{\text{LR}} = 1/N_{\text{LR}}$. The gain in fact should be $1/N_t$, which behaves much like $1/t$ for smaller values of $t$.

The predictive error decomposition gives the distribution of the one-period-ahead forecasts:
\[
y_t \mid y_1 \ldots y_{t-1} \sim N(m_{t-1}, \sigma_{t-1}^2 + \rho \sigma_\epsilon^2 + \sigma_\epsilon^2). \tag{8}
\]
The product of these densities for $t = 2, \ldots, n$ gives log joint probability of $y_2, \ldots, y_n$ conditional on $y_1$ as a function of $\sigma_\epsilon^2$ and $\rho$, and therefore the log likelihood of $\sigma_\epsilon^2$ and $\rho$ conditioned on $y_1, \ldots, y_n$. The observation variance $\sigma_\epsilon^2$ may be concentrated out of the log likelihood function, so that a numerical maximization search is required only over the single parameter $\rho$.

Although it is common to develop the LLM in terms of the signal/noise variance ratio $\rho$, 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,
\[
\text{NSR} = \sigma_\epsilon / \sigma_\eta = \rho^{1/2}.
\]
The NSR ratio has the natural units [time], and equals the number of time units it takes for the average of NSR $\varepsilon_t$ shocks to equal the sum of NSR $\eta_t$ shocks, or, in other words, the length of time it takes for typical changes in the level of the state variable to begin to be empirically detectable. Furthermore, (7) implies that $N_{\text{LR}}$ is closely approximated by NSR:
\[
\text{NSR} < N_{\text{LR}} < \text{NSR} + 1.
\]

Figure 1 plots $N_t$ versus $t$, using $\text{NSR} = 28.4$, the empirical value obtained in Section VI below for monthly PCE inflation with an ALS/AR(2) specification. $N_t$ is virtually indistinguishable from $t$ until $t = 10$, but then it grows more slowly and it is virtually indistinguishable from its asymptotic value of $N_{\text{LR}} = 28.9$ after $t = 100$. 
III. Adaptive Least Squares

The simplistic LLM allows the 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,

$$y_t = x_t \beta_t + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, \sigma^2_{\varepsilon}),$$

$$\beta_t = \beta_{t-1} + \eta_t, \quad \eta_t \sim \text{NID}(0_{k \times 1}, Q_t),$$

in which $x_t$ is a $1 \times k$ row vector\(^1\) of ideally exogenous explanatory variables, $\beta_t$ is a $k \times 1$ column vector of time-varying coefficients $\beta_{t,t}$, and $\eta_t$ is a $k \times 1$ column vector of transition errors $\eta_{t,t}$ that are independent of the observation errors $\varepsilon_t$. 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 $\eta_t$. We assume that the first column of $\mathbf{X}_t$ is a vector of units, so that the first element of $\beta_t$ is the intercept. When $k =\]

\(^1\) We make $x_t$ a row vector rather than a column vector, since $x_t$ is simply the $t$-th row of the regressor matrix $\mathbf{X}_t$. 

---

**Figure 1**

Effective sample size $N_t$ with $NSR = 28.4$ mo., $N_{LR} = 28.9$ mo.
1, the TVP model reduces to the LLM if the single element of $Q_t$ and therefore the signal/noise ratio is time-invariant.

As reviewed in Appendix A.2 below, system (9)-(10) may be solved by means of the well-known Extended Kalman Filter, which provides a recursive rule giving

$$
\beta_t | y_t \sim N(b_t, P_t)
$$

(11)

for some $k \times k$ covariance matrix $P_t$. 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 (9)-(10) is much too general for most econometric purposes, since if even if $Q_t$ 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 $\sigma_e^2$.

Ljung (1992) and Sargent (1999, p. 117) ingeniously postulate that $Q_t$ is directly proportional to $P_{t-1}$. This assumption, which leads to Recursive Least Squares (RLS), allows every component of $\eta_t$ to be stochastic, yet effectively reduces $Q_t$ to a single unknown parameter, while being invariant in its implications to changes of basis. It also has the benefit of greatly simplifying the computations, by eliminating two matrix inversions. Nevertheless, the proportionality that Ljung and Sargent propose must be modified in order to nest the elementary LLM when $k = 1$ and to allow the likelihood to be computed.

The one element of the signal shock $\eta_t$ that contributes directly to $y_t$, on a one-for-one basis and whose variance is therefore potentially comparable to $\sigma_e^2$, 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 $\eta_t$, 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 $\rho$ times the noise variance $\sigma_e^2$. In Appendix A.2 below, it is shown that this assumption implies that

$$
Q_t = \rho N_t \cdot P_t,
$$

(12)

where $N_t$ is computed exactly as in (5) in the LLM.

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

$$
b_t = W_t^{-1} z_t,
$$

(13)

$$
P_t = \sigma_e^2 W_t^{-1},
$$

(14)

where

$$
z_t = (1 + \rho N_{t-1})^{-1} z_{t-1} + x_t'y_t,
$$

(15)

$$
W_t = (1 + \rho N_{t-1})^{-1} W_{t-1} + x_t' x_t.
$$

(16)
With no prior information about the coefficients at time 0, the algorithm is initialized with
\[
\mathbf{W}_0 = \mathbf{0}_{k \times k}, \quad \mathbf{z}_0 = \mathbf{0}_{k \times 1}.
\]
Note that in the fixed coefficient case \( \rho = 0 \), \( \mathbf{z}_t \) becomes \( \mathbf{X}' \mathbf{y}_t \), \( \mathbf{W}_t \) becomes \( \mathbf{X}' \mathbf{X}_t \), and (13) becomes the familiar expanding-window OLS formula \( \mathbf{b}_t = (\mathbf{X}' \mathbf{X}_t)^{-1} \mathbf{X}' \mathbf{y}_t \).

Having thus initialized and updated the ALS filter, the predictive error decomposition becomes
\[
y_t | y_{t-1} \sim N(\mathbf{x}_t \mathbf{b}_{t-1}, \sigma^2 \varepsilon^2 s_t^2), \quad t > k,
\]
where
\[
s_t^2 = (1 + \rho T_{t-1}) \mathbf{x}_t \mathbf{W}_{t-1}^{-1} \mathbf{x}_t' + 1.
\]
The log likelihood is then
\[
L(\rho, \sigma^2_\epsilon | y_t) = \sum_{t=k+1}^n \log p(y_t | y_{t-1})
= -\frac{n-k}{2} \log(2\pi) - \frac{n-k}{2} \log \sigma^2_\epsilon
- \sum_{t=k+1}^n \log s_t
- \frac{1}{2 \sigma^2_\epsilon} \sum_{t=k+1}^n u_t^2,
\]
where the scale-adjusted residuals
\[
u_t = e_t / s_t
\]
equal the actual predictive errors,
\[
e_t = y_t - \mathbf{x}_t \mathbf{b}_{t-1},
\]
adjusted by their time-varying scales \( s_t \). (Since \( \mathbf{W}_t \) is of rank \( t \) for \( t \leq k \), the predictive density and therefore the likelihood contribution may only be computed for \( t > k \).)

Under the maintained assumptions, these adjusted residuals are homoskedastic with variance \( \sigma^2_\epsilon \), even though the predictive errors themselves are highly heteroskedastic. As in the LLM, the observation variance \( \sigma^2_\epsilon \) may be concentrated out of the log likelihood function in such a way that for any value of \( \rho \), the likelihood is maximized with \( \sigma^2_\epsilon \) estimated in closed form by
\[
\hat{\sigma}^2_\epsilon = \frac{1}{n-k} \sum_{t=k+1}^n u_t^2
\]
A numerical search over the remaining hyperparameter \( \rho \) then provides its ML estimate.

If the model is well-specified and \( \rho \) equal to its true value, the adjusted residuals \( u_t \) must be iid \( \text{N}(0, \sigma^2_\epsilon) \). Since the hyperparameter \( \rho \) 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, as noted by Durbin and Koopman (2001, Ch. 5).

The Kalman Filter for the ALS model, presented above, provides the posterior distribution of the coefficient vector conditional on the past and current history of the data. This is the appropriate question to ask if one is interested in simulating expectations as of each point in time. However, if one instead wanted to estimate the time-varying
regression coefficients at a given point in time conditional on both prior and subsequent experience, the Kalman Smoother (also known as the 2-sided filter) becomes the appropriate tool. This is straightforward, but somewhat more complicated. The pertinent smoother equations for both the general TVP case and the special ALS case are given in Appendix A.3.

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 sample, but this consistency is absent in the ALS case, because the effective sample size never rises above \( N_{LR} \).

**IV. Relation of ALS to other approaches**

There have been numerous attempts to make the general TVP model (9)-(10) 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 nests the LLM, but only by being unnecessarily restrictive.

Sims (1988) and Kim and Nelson (2004) use (9) with a time-invariant covariance matrix \( \mathbf{Q} \), but assume that \( \mathbf{Q} \) 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.

Cogley and Sargent (2004) ambitiously estimate an autoregressive TVP model in which the coefficients take a random walk with unrestricted covariance matrix, subject to reflecting boundaries that prevent nonstationary autoregressive roots. With fixed coefficients there is a case for imposing such restrictions, since such 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 acceleration, as is all too often the case with inflation data. Nevertheless, binding inequality restrictions may easily be imposed on
ALS estimates after unrestricted estimation, if desired, by using the standard conditional multivariate normal formula given, e.g., by Harvey (1989, p. 165).

Ljung (1992) and Sargent (1999) observe that if $Q_t$ is restricted to be proportional to $P_{t-1}$, not only are there far fewer parameters to estimate, but the filter also simplifies greatly. They then set

$$Q_t = \frac{\gamma}{1-\gamma} P_{t-1}, \quad (21)$$

resulting in *Constant Gain Recursive Least Squares* (CG-RLS) with constant gain $\gamma$: In this case the Extended Kalman Filter (35 - 36) in Appendix A.2 may be written

$$b_t = b_{t-1} + \gamma R_t^{-1} x_t'(y_t - x_t b_{t-1}), \quad (22)$$

$$R_t = R_{t-1} + \gamma (x_t x_t' - R_{t-1}), \quad (23)$$

$$P_t = \gamma \sigma^2. \quad (24)$$

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 signal/noise parameter that determines the long run gain itself. Furthermore, it is not immediately obvious how to initialize CG-RLS in the absence of prior information about $P_0$ and $b_0$.

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 eventually be computed from $X_n$ and $y_n$, 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 $P_0$.

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

Stock and Watson (1996) and Sargent and Williams (2003) assume, in place of either (12) or (21), that

$$Q_t = Q = \rho \sigma^2 \mathbb{E} x_t' x_t^{-1} \quad (25)$$

If the relevant expectation exists, this is equivalent in an expectational sense to (12), since then

$$\mathbb{E} W_t = N_t \mathbb{E} x_t' x_t.$$

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

---

2 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.
they were. Equation (12), on the other hand, does not require these moments to be finite, and only requires that agents know $X_t, Y_t, \text{ and } \rho$. Assumption (25) 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 RLS, since it requires the more general EKF described in Appendix A.2.

Stock and Watson (1996) calibrate the coefficient $\rho$ in (25) (their $\lambda^2$) 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 (12) and their (25). For one thing, the initial errors have much larger variance than the later errors, simply because the coefficient vector is still highly uncertain. Equation (18) 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 (18), which is similar to the formula for the conditional distribution that would result from (25), that even asymptotically the squared forecasting errors $e_t^2$ are greater in expectation than $\sigma^2_e$ by an amount that depends on $\rho$, so that minimizing their sum of squares will give a biased estimate of $\sigma^2_e$. In addition, even after the warm-up period they are not homoskedastic, and hence 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. However, if one’s objective is to construct one’s own expert forecast of these variables, one should use actual experience, not the forecasts of other, perhaps less sophisticated, experts, to calibrate one’s own procedures.

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 his gain like a subjective learning parameter to be inferred from agents’ expectationally-motivated behavior, specifically their demand for money, instead of estimating $\rho$ from the inflation series in question using (8) and then computing the long run gain according to (7).

IV. Hypothesis testing

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 $\chi^2$ limiting distribution of the Lagrange Multiplier (LM) and Likelihood Ratio (LR) statistics

---

3 The observation variance $\sigma^2_e$ is required to compute $P_t$, but not $b_t$. 
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.

Local tests on coefficient restrictions for an individual value of \( t \) may be performed, using either the filter or smoother estimates, by standard \( z \)- or Chi-square tests with the appropriate degrees of freedom, since the coefficients are normally distributed with the estimated covariance matrix, at least if the two variances are known. In practice, the variances must be estimated, but since their ML estimates are consistent, this should not be a problem in large samples.

Global linear restrictions on coefficients are more difficult. The author has been unable to construct a precise test, but offers the following approximate test: For \( t \geq k \), define

\[
\theta_{j,t} = \beta_{j,t} / \sqrt{p_{t,j,j'}}.
\]

Under the null \( \beta_{t,j} = 0 \) for all \( t \) and given the two hyperparameters, these are distributed \( \text{N}(0,1) \) for each \( t \), but are highly serially correlated. In order to obtain an at least approximately valid global test statistic, we assume that they are \( \text{AR}(1) \), with persistence governed by the ALS gains \( \gamma \):

\[
\theta_{j,t} = (1 - \gamma_t) \theta_{j,t-1} + v_{j,t},
\]

and with serially independent innovations that have variance

\[
\text{var}(v_{j,t}) = 1 - (1 - \gamma_t)^2.
\]

Then under our assumption, the normalized innovations

\[
\zeta_{j,t} = \frac{v_{j,t}}{\sqrt{1 - (1 - \gamma_t)^2}}
\]

will be i.i.d. \( \text{N}(0,1) \), so that

\[
G_j = \sum_{t=k}^n \zeta_{j,t}^2
\]

will be distributed \( \chi^2 \) with \( n-k+1 \) degrees of freedom.

**VI. Application to US Inflation**

As Klein (1978) pointed out, the time series behavior of US inflation has not been constant 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 – Nov. 2021. A series of high-inflation months had left the year-over-year PCE inflation at 5.58% in Nov. 2021, and set off a vigorous debate over whether this elevated inflation rate should be regarded as "transitory" or "entrenched."

---

**Figure 2**
Monthly PCE inflation, seasonally adjusted
Table 1 presents the outcome of ALS estimation of time-varying AR(p) models of PCE inflation for \( p = 0, \ldots, 4 \), where the AR(0) model is simply the LLM, with no time-varying autoregressive parameters. 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. 2021, for \( n = 750 \) months.

### Table 1  
**ALS AR(p) model of PCE Inflation**  
6/59 - 11/21, \( n = 750 \)

<table>
<thead>
<tr>
<th>( p )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( NSR ) (mo.)</td>
<td>2.90</td>
<td>21.9</td>
<td>28.4</td>
<td>38.8</td>
<td>52.4</td>
</tr>
<tr>
<td>( N_{LR} ) (mo.)</td>
<td>3.44</td>
<td>22.3</td>
<td>28.9</td>
<td>39.2</td>
<td>52.9</td>
</tr>
<tr>
<td>( \rho ) (mo.(^{-2}))</td>
<td>0.119</td>
<td>0.00209</td>
<td>0.00124</td>
<td>0.000666</td>
<td>0.000364</td>
</tr>
<tr>
<td>LR: ( \rho = 0 )</td>
<td>552.36</td>
<td>91.73</td>
<td>79.26</td>
<td>56.68</td>
<td>32.13</td>
</tr>
<tr>
<td>G: AR(p)=0</td>
<td>---</td>
<td>5455.2</td>
<td>846.9</td>
<td>614.8</td>
<td>553.6</td>
</tr>
<tr>
<td>( p(G) )</td>
<td>---</td>
<td>0.0000</td>
<td>0.0068</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Jarque-Bera ( p(JB) )</td>
<td>579.6</td>
<td>194.5</td>
<td>301.4</td>
<td>351.8</td>
<td>305.8</td>
</tr>
<tr>
<td>Forecasts from 11/21:</td>
<td>1.4e-126</td>
<td>5.9e-43</td>
<td>3.6e-66</td>
<td>4.1e-77</td>
<td>3.9e-67</td>
</tr>
<tr>
<td>1 mo.</td>
<td>6.23</td>
<td>6.17</td>
<td>5.68</td>
<td>5.19</td>
<td>5.16</td>
</tr>
<tr>
<td>1 yr. marg.</td>
<td>6.23</td>
<td>3.69</td>
<td>3.11</td>
<td>2.94</td>
<td>2.73</td>
</tr>
<tr>
<td>1 yr. avg.</td>
<td>6.23</td>
<td>4.32</td>
<td>3.59</td>
<td>3.64</td>
<td>3.47</td>
</tr>
<tr>
<td>Long Run</td>
<td>6.23</td>
<td>3.64</td>
<td>3.10</td>
<td>2.84</td>
<td>2.56</td>
</tr>
</tbody>
</table>

The estimated noise-to-signal standard deviation ratio \( NSR \) increases sharply with the autoregressive order \( p \), beginning with an absurdly low value of 2.90 months for \( p = 0 \). Evidently, the LLM interprets as permanent fluctuations that can more easily be explained as transitory low-order AR components. The long-run effective sample size or reciprocal 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 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 approximately distributed \( \chi^2 \) with \( n-p = 750, \ldots, 746 \) degrees of freedom under the null with the maintained assumptions. Perhaps a more precise statistic will be developed in the future, but in the meanwhile, this provides no evidence against omitting the AR(4) term in the AR(4) model, or against omitting the AR(3) term in the AR(3) model. However, it does reject omitting the AR(2) term in the AR(2) model, and
even more emphatically, omitting the AR(1) term in the AR(1) model. We therefore provisionally accept the AR(2) model as being most consistent with long-run experience.

Figure 3 shows the ALS estimates of the three coefficients in the preferred AR(2) model. Since the initial effective sample size is very small, it is to be expected that the first several months will have very erratic predictions. The AR(2) coefficient is generally much smaller than the AR(1) coefficient, but it is their sum that determines the total persistence of the transitory AR components of the process.

![Figure 3](image.png)

**Figure 3**

AR(2) model coefficients
Figure 4 plots the sum of the two AR coefficients from Figure 3, along with its 95% Credible Interval (CI). The CI is naturally very wide in the early years because of the small effective sample size there. It may be seen that the sum is locally significantly different from 0 throughout 1968-90, 2008-16, and again in 2021. However, it is essentially 0 before 1967 and throughout 1999-2008, so that there has been substantial change in the persistence of inflation. Although the point estimate never quite reaches unity, the 95% CI either includes or almost includes this value throughout 1967-98, suggesting near-unit-root transitory behavior there. As noted above, ALS estimates of AR models are prone to the same small-sample bias away from a unit root as are OLS estimates, with the added problem that the effective sample size is bounded above by $N_{LR}$. However, this problem may be offset by the unit roots in all coefficients already postulated by the TVP model itself. The present paper makes no attempt to correct for it.

![Sum of lag coefficients in AR(2) model](image-url)

**Figure 4**

*Sum of lag coefficients in AR(2) model*
The AR(2) coefficient from Figure 3 is shown by itself in Figure 5, with its 95% CI. The CI excludes 0 only briefly, in the mid-'70s and late '00s, and then not as emphatically as in Figure 4. The global G statistic in Table 1 for the null that the AR(2) coefficient is always zero is 846.9. This is approximately $\chi^2$ with 748 degrees of freedom, which yields a $p$ value of 0.0068, so we reject AR(1) in favor of AR(2). Even if AR(2) effects have not been quantitatively large in the past, they may become so in the future under the Fed's Aug. 27, 2020 announcement that it will 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).
For each time $t$, the model can be used to predict 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 could easily be included in an ALS regression, but it is an empirical matter just how much explanatory power such other variables would actually have.
Figure 6 plots the marginal and average ALS/AR(2) inflation forecasts for the last month in our observation set, Nov. 2021, when the economy happened to have just experienced a short run of unusually high monthly inflation observations, using the final point estimates of the coefficients. The last two of these observations, taken directly into account by the AR(2) forecasts, are plotted as green stars, at 8.12% and 7.29%. The red line gives the marginal inflation forecasts, which start at 5.68% for \( h = 1 \) month, but then quickly descend and stabilize at an asymptotic value within rounding error of 3.10% for all horizons beyond 10 months. The blue line gives the average inflation forecasts, which begin at the same 5.68% at 1 month, and then descend more slowly toward the same asymptotic value. The 1-year average inflation forecast is 3.59%. The average forecast is still 3.15%. at 10 years, because of the high transitory forecasts in the first year, but will eventually reach 3.10%. With the simplistic LLM, by way of contrast, the forecast of inflation as given in Table 1 for \( p = 0 \) is 6.23% at all horizons.
The asymptotic, long-run inflation rate implied by the AR($p$) process at time $t$ may be obtained directly without recursively forecasting monthly inflation as

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

This may be estimated, in the stationary case, by

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

Since this value abstracts from the transitory AR components, it corresponds best to what the Fed has in mind by "entrenched inflationary expectations," and would be the appropriate experience-based "inflation" variable to include in a Taylor-type rule. It is graphed for our AR(2) 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. As indicated in Table 1 above, the 1-month forecast as of Nov. 2021 is 5.68%, while the long-run predicted rate is only 3.10%. Figure 4 indicates that in March of 1974 the sum of the lag coefficients briefly almost reached unity, with the result that the long-run forecast was nearly undefined at that date.
Long-run predicted inflation was only 1.43% in Nov. 2020 and 1.55% one year before that. Although the ALS/AR(2) model does indicate that the Nov. 2021 year-over-year figure of 5.58% was mostly transitory, it did push the entrenched long-run prediction up to 3.10%, well above the Fed's announced 2% target. This already more than accomplished the FOMC's new policy (Powell 2020) of briefly permitting a little more than 2% PCE inflation to speed the growth of entrenched inflationary expectations up to its target.

VII. Forecast uncertainty

The uncertainty of forecasts made with the ALS model from time $n$ forward has at least five components:

1. Future "noise" errors $\varepsilon_t, t > n$. However, these average out to a value approaching zero over long forecast horizons, and therefore are not directly an important source of long-run forecast uncertainty.

2. Initial parameter uncertainty, as reflected in $P_n$.

3. Future parameter drift, as reflected in $Q_n, t > n$.

4. Hyperparameter uncertainty with respect to NSR and $\sigma^2$.

5. Model uncertainty 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 $\chi^2$ 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 forecasts from Nov. 2021 that would have been made with the AR(2) model, using 20 draws from the $N(b_n, P_n)$ distribution in place of $b_n$ itself, and no further coefficient drift. Each simulation is initialized with the observed monthly inflation for Oct. and Nov. 2021, represented by the green stars at 8.12% and 7.29%. The heavier blue line is the point forecast from Figure 6. Each of these draws happens to be stationary and a few are oscillatory with damped oscillations.

![Simulated predicted average inflation, AR(2) model with no coefficient drift, first 20 replications.](image)

**Figure 8**
Simulated predicted average inflation, AR(2) model with no coefficient drift, first 20 replications.
The thin red lines in Figure 9 plot the posterior median and 50% and 95% credible intervals for predicted average inflation in the AR(2) 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 (2.5, 3.8), while the 95% CI is approximately (0.7, 7.8) at the 10-year horizon and is still growing, so that there is considerable uncertainty in the 10-year forecast, even assuming away parameter drift.

![Predicted average inflation CIs -- no drift, 1000 reps, AR(2)](image)

**Figure 9**
Credible intervals for predicted average inflation, AR(2) model with no drift, 1000 replications.
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 this purpose, $Q_{n+1}$, based on $P_n$, was used as the signal covariance matrix for all $t > n$. These simulations therefore abstract from the continuing changes in $Q_{t+1}$ via $P_t$ 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 or even oscillating explosions.

![Simulated predicted avg. inf. -- with drift, first 20 reps, AR(2)](image)

**Figure 10**

Simulated predicted average inflation, AR(2) model with and without drift, 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, 4.6) and growing, and the 95% CI is already completely off scale within 6 years.

![Predicted avg inf CIs -- with drift, 1000 reps, AR(2)](image)

**Figure 11**

**Credible intervals for predicted average inflation, AR(2) model with and without drift, 1000 replications.**

Similar simulations using the AR(3) and AR(4) models of Table 1 produce tighter forecast CIs as the autoregressive degree increases. This is due to the fact that \( N_{LR} \) increases with \( p \), with the result that \( P_n \) tends to be smaller, which in turn implies that \( Q_n \) will be smaller as well. However, since in any TVP model each coefficient is taking a random walk, the CIs with drift will necessarily be unbounded as the forecast horizon increases.
Given the noise-to-signal ratio, the estimate of the noise variance $\hat{\sigma}_\varepsilon^2$ in (21) is governed by a chi-squared distribution with $n-k = 747$ 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 $NSR$. The vertical green bar is positioned at the ML estimate of 28.4. The horizontal red line is 1.92 units below the maximized likelihood, where the likelihood ratio 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 95% CI for $NSR$ is therefore a very considerable $(19.9, 41.1)$. The present paper makes no attempt to quantify the effect of this uncertainty on the already large forecast uncertainty manifested in Figure 11 above.

**Figure 12**
Log likelihood versus $NSR$ in AR(2) model
Figure 13 below shows the scale-adjusted residuals $u_t$ for the preferred AR(2) model. Under the assumptions of the model, these should be i.i.d. normal, but the Jarque-Bera JB statistics in Table 1 soundly reject this hypothesis. There is some visual evidence of volatility clustering, but the several extreme outliers that appear without warning suggest that conditional non-normality is a larger problem than conditional heteroskedasticity.

![Figure 13](image)

**Figure 13**
Scaled errors $u_t$ in AR(2) model

**VIII. Potential extensions and future applications.**

Clarida, Gali 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) applies ALS to this problem, using the ALS filter to simulate real-time expectations of inflation and the natural unemployment rate, and 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 assumption of the Kalman Filter, but the numerical integrals required 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 this particle filter can be generalized to an ALS model with constant signal/noise ratio and stable errors.
Appendix

A.1. The Local Level Model

The Local Level Model (2) implies
\[ \mu_t = y_t - \varepsilon_t, \]
so that the distribution of \( \mu_t \) given \( y_t \) and an uninformative prior may be written
\[ \mu_t \mid y_t \sim N(m_t, \sigma^2_t), \]
where
\[ m_t = y_t, \]
\[ \sigma^2_t = \sigma^2_e. \]

Assume now, as we know to be the case for \( t = 2 \), that the distribution of the state variable \( \mu_{t-1} \) given the observations \( y_{t-1} = (y_1, \ldots, y_{t-1})' \) up to and including \( y_{t-1} \), is likewise normal, with parameters
\[ \mu_{t-1} \mid y_{t-1} \sim N(m_{t-1}, \sigma^2_{t-1}), \]
It follows that
\[ \mu_t \mid y_{t-1} \sim N(m_{t-1}, \sigma^2_{t-1} + \sigma^2_q) = N(m_{t-1}, \sigma^2_{t-1} + \rho \sigma^2_e). \quad (27) \]
We also know that
\[ y_t \mid \mu_t \sim N(\mu_t, \sigma^2_e). \]

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
\[
p(\mu_t \mid y_t) = p(y_t \mid \mu_t, y_{t-1}) p(\mu_t \mid y_{t-1}) / (\text{const.})
= p(y_t \mid \mu_t) p(\mu_t \mid y_{t-1}) / (\text{const.})
= \exp \left( -\frac{1}{2} \frac{(y_t - \mu_t)^2}{\sigma^2_e} \right) \exp \left( -\frac{1}{2} \frac{(\mu_t - m_{t-1})^2}{\sigma^2_{t-1} + \rho \sigma^2_e} \right) / (\text{const.})
= \exp \left( -\frac{1}{2} \frac{(\mu_t - m_{t-1})^2}{\sigma^2_{t-1} + \rho \sigma^2_e} \right) / (\text{const.}),
\]
so that (3) is valid with
\[ m_t = \frac{\sigma^2_t}{\sigma^2_e} y_t + \frac{\sigma^2_e}{\sigma^2_{t-1} + \rho \sigma^2_e} m_{t-1} \quad (29) \]
and
\[ \frac{1}{\sigma^2_t} = \frac{1}{\sigma^2_{t-1} + \rho \sigma^2_e} + \frac{1}{\sigma^2_e}. \quad (30) \]
Defining \( N_t = \frac{\sigma^2_e}{\sigma^2_t} \), (29) becomes (4) and (30) becomes (5), which may be initialized either with \( N_0 = 0 \) or \( N_1 = 1 \).
A.2. Time-Varying Parameters and Adaptive Least Squares

The general TVP system (9) 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,

$$\beta_{t-1} | \gamma_t \sim N(b_{t-1}, P_{t-1})$$

for some $k \times k$ covariance matrix $P_{t-1}$ that may depend on $X_{t-1}$, but not $\gamma_{t-1}$ or $e_{t-1}$. Then by Harvey (1989, pp. 105-6), or equivalently, Ljung and Söderström (1983, p. 420), and simplifying to the univariate random walk case (9) of interest,

$$\beta_t | \gamma_t \sim N(b_t, P_t),$$

where

$$b_t = b_{t-1} + f_t^{-1}(P_{t-1} + Q_t)x_t(y_t - x_t b_{t-1}),$$

$$P_t = (P_{t-1} + Q_t)(I_{k \times k} - f_t^{-1}x_t x_t'(P_{t-1} + Q_t)),$$

$$f_t = x_t'(P_{t-1} + Q_t)x_t + \sigma_e^2.$$

The textbook EKF equations (32) and (33) above may be rearranged to eliminate $f_t$ and to look more like RLS, as follows: Post-multiply (33) by $x_t'$ and combine with (34) to obtain

$$P_t x_t' = (P_{t-1} + Q_t)(x_t' - f_t^{-1}x_t'(f_t - \sigma_e^2))$$

$$= \sigma_e^2 f_t^{-1}(P_{t-1} + Q_t)x_t',$$

so that (32) becomes

$$b_t = b_{t-1} + (1/\sigma_e^2)P_t x_t'(y_t - x_t b_{t-1}),$$

and (33) becomes

$$P_t = (P_{t-1} + Q_t) - (1/\sigma_e^2)P_t x_t x_t'(P_{t-1} + Q_t).$$

Then multiply the last equation on the left by $P_t^{-1}$ and on the right by $(P_{t-1} + Q_t)^{-1}$ and rearrange to obtain

$$P_t^{-1} = (P_{t-1} + Q_t)^{-1} + (1/\sigma_e^2)x_t'x_t.$$

The rearranged filter (35), (36) may be placed in the even more convenient “Information” form, mentioned but not developed by Harvey (1989, p. 108), in terms of the scaled precision matrix $W_t = \sigma_e^2 P_t^{-1}$, the scaled transition covariance matrix $V_t = (1/\sigma_e^2)Q_t$, and what might be called the “cumulative evidence” vector $z_t = W_t b_t$, as follows:

$$z_t = (I + W_{t-1}V_t)^{-1}z_{t-1} + x_t' y_t,$$

$$W_t = (I + W_{t-1}V_t)^{-1} W_{t-1} + x_t'x_t.$$
whence $b_t$ and $P_t$ may be recovered by (13) and (14).\footnote{The observation error variance $\sigma^2_\varepsilon$ cannot be estimated until after the filter has been run, so it must in any event be factored out of $P_t$ and $Q_t$ in order to run the filter.} Cp. also Bullard (1992).

In the absence of prior information about $b_0$, the above Information form filter may easily be initialized with a diffuse prior by taking the limit of $P_0$ as all its eigenvalues go to infinity, or equivalently, by letting the initial precision matrix $P_0^{-1}$ go to zero, which in turn implies

$$W_0 = 0_{k \times k}$$

as in (17). For any choice of $b_0$, $z_0 = W_0 b_0$ then implies

$$z_0 = 0_{k \times 1}.$$

It is not so obvious how to impose a diffuse prior on either $(32) - (34)$ or $(35) - (36)$, however.

With this diffuse prior, $W_t$ is of rank $t$ for $t \leq k$, and hence $b_t$ and $P_t$ may not be computed by (13) and (14) until $t \geq k$. Note that in the fixed coefficient case

$Q_t = V_t = 0_{k \times k}$, $z_t$ becomes $x'_t y_t$, $W_t$ becomes $x'_t x_t$, and (13) 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 $W_t$ may still be singular for some $t \geq k$, but we assume here that the regressors all vary continuously and that this is never the case.)

Ljung (1992) and Sargent (1999) observe that if $Q_t$ is restricted to be proportional to $P_{t-1}$, not only are there far fewer parameters to estimate, but the filter also simplifies greatly. If we set $Q_t = k_t P_{t-1}$ for some constant $k_t$, then $V_t = k_t W_{t-1}^{-1}$, and (37) and (38) become following generalized version of Restricted Least Squares (RLS):

$$z_t = \frac{1}{1 + k_t} z_{t-1} + x'_t y_t,$$  \hspace{1cm} \text{(39)}

$$W_t = \frac{1}{1 + k_t} W_{t-1} + x'_t x_t.$$  \hspace{1cm} \text{(40)}

Thus, the RLS class of restrictions reduces the matrix inversions in (37) and (38) to a single scalar inversion.

In order for 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 RLS intercept term $b_{1,t}$. 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. To this end, we define

$$x^*_t = x_t A_t,$$

for
\[ A_t = \begin{pmatrix} 1 & p_{t,1,2} p_{t-2,2} \\ 0_{k-1,1} & I_{k-1 \times k-1} \end{pmatrix}, \]

where

\[ p_t = \begin{pmatrix} p_{t,1,1} & p_{t,1,2} \\ p_{t,2,1} & p_{t,2,2} \end{pmatrix}, \]

so that the transformed coefficients are

\[ b_t^* = A_t^{-1} b_t, \]

with block-diagonal covariance matrix

\[ P_t^* = \text{Cov}(b_t^*) = A_t^{-1} P_t A_t^{-1}' = \begin{pmatrix} p_{t,1,2} p_{t,2,2} & 0_{1 \times k-1} \\ 0_{k-1 \times 1} & P_{t,2,2} \end{pmatrix}. \]

Setting

\[ W_t = \sigma_e^2 P_t^{-1} = \begin{pmatrix} w_{t,1,1} & w_{t,1,2} \\ w_{t,2,1} & w_{t,2,2} \end{pmatrix}, \]

and applying the block matrix inversion formula to \( P_t \), we have

\[ P_t^* = \begin{pmatrix} \sigma_e^2 w_{t,1,1} & 0_{1 \times k-1} \\ 0_{k-1 \times 1} & P_{t,2,2} \end{pmatrix}. \]

Now if \( Q_t = k_t P_{t-1} \), we will also have \( Q_t^* = k_t P_{t-1}^* \), where \( Q_t^* = A_t^{-1} Q_t A_t^{-1}' \) is the covariance matrix of the appropriately transformed time \( t \) transition errors \( \eta_t^* = A_t^{-1} \eta_t \). Therefore if, as in the LLM, the variance of the shock to the (orthogonalized) intercept is a constant \( \rho \) times the noise variance,

\[ q_{t,1,1}^* = \rho \sigma_e^2, \]

it follows that

\[ k_t = \rho w_{t-1,1,1}, \]

whence (40) implies that \( w_{t,1,1} \) obeys the same recursion as the LLM’s effective sample size \( N_t \) in (5). Furthermore, the diffuse prior (17) implies \( w_{0,1,1} = 0 \), just as \( N_0 = 0 \) in the LLM. Setting \( N_t = w_{t,1,1} \), we obtain the ALS updating equations (15) and (16) with \( k_t = \rho N_t \), as claimed in the text.

It may easily be shown, using (35) and (36) and setting \( R_t = W_t / N_t \), that the ALS filter (13) – (16) is equivalent to the generalized variable-gain RLS formula

\[ b_t = b_{t-1} + \gamma_t R_t^{-1} x_t' (y_t - x_t b_{t-1}), \]

\[ R_t = R_{t-1} + \gamma_t (x_t' x_t - R_{t-1}), \]

\[ P_t = \gamma_t \sigma_e^2 R_t^{-1}, \]

where the gain \( \gamma_t \) equals \( 1/N_t \).

### A.3. The 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 \( b_t^* \) of \( b_t \) conditional on \( y_1, ... y_t \) and no other
information, with variances \( P_t^* \). In the general TVP case, this backward filter may be computed by:

\[
\begin{align*}
  z_{n+1}^* &= 0_{k \times 1} \\
  W_{n+1}^* &= 0_{k \times k} \\
  z_t^* &= (I + W^*_{t+1} V_{t+1})^{-1} z_{t+1}^* + x_t^* y_t^*, \quad \text{(44)} \\
  W_t^* &= (I + W_{t+1}^* V_{t+1})^{-1} W_{t+1}^* + x_t^* x_t^*, \quad \text{(45)} \\
  b_t^* &= W_{t+1}^* z_t^*, \quad t \leq n-k, \quad \text{(46)} \\
  P_t^* &= \sigma_e^2 W_t^{*1}, \quad t \leq n-k. \quad \text{(47)}
\end{align*}
\]

The backward filter \( b_{t+1}^* \) obtained in this manner estimates \( \beta_{t+1} \), conditional on \( y_{t+1}, \ldots, y_n \), with variance \( P_{t+1}^* \), but it also provides an estimate of \( \beta_t \), conditional on the same values, with the somewhat larger variance \( P_{t+1}^* + Q_{t+1} = \sigma_e^2 W_{t+1}^* (I + W_{t+1}^* V_{t+1}) \). Since \( b_{t+1}^* \) as an estimate of \( \beta_t \) is independent of the filter estimate \( b_t \), the two estimates may be averaged in inverse proportion to their respective covariance matrices to form the smoother estimate \( \hat{b}_t^S \) of \( \beta_t \), conditional on the entire sample, as follows:

\[
\begin{align*}
  \hat{b}_t^S &= W_t^{S-1} \left( W_t b_t + \left( (I + W_{t+1}^* V_{t+1})^{-1} W_{t+1}^* \right) b_{t+1}^* \right) \\
  &= W_t^{S-1} z_t^S, \quad \text{(48)}
\end{align*}
\]

where

\[
\begin{align*}
  W_t^S &= W_t + (I + W_{t+1}^* V_{t+1})^{-1} W_{t+1}^* \quad \text{(49)} \\
  z_t^S &= z_t + (I + W_{t+1}^* V_{t+1})^{-1} z_{t+1}^* \quad \text{(50)}
\end{align*}
\]

This smoother estimate has variance

\[
P_t^S = \sigma_e^2 W_t^{S-1}. \quad \text{(51)}
\]

Note that it is not necessary to actually compute the backward filter and variance \( b_t^* \) and \( P_t^* \) themselves, however, since \( z_t^* \) and \( W_t^* \) suffice to obtain the smoother and its variance using (48) – (51). 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 \( z_t \) and \( W_t \) 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 simply set

\[
V_{t+1} = \rho N_t W_t^{-1}
\]

in (44) - (51). If desired, the term \( (I + \rho N_t W_{t+1}^* W_t^{-1})^{-1} \) may then be replaced by \( W_t (W_t + \rho N_t W_{t+1}^*)^{-1} \) to avoid having to invert \( W_t \). In the general TVP case, where the transition covariance matrix \( Q_t \) is well defined for all \( t \), we may compute the smoother
clear back to $t = 1$. In the ALS case, however, $W_t^S$, which may be written, using the
above substitution, as $W_t(I + (W_t + \rho N_t W_{t+1})^{-1}W_{t+1}^*)$, is proportional to $W_t$ and
therefore singular for $t < k$. The ALS smoother, like its filter, is therefore defined only
for $t \geq k$. Unfortunately, the serendipitous cancellation that occurs in the filter equations
is no longer present, so that the ALS smoother runs a little slower than the ALS filter.
A.4. The error in Ljung (1992) and Sargent (1999)

As mentioned above, there is an error in the Kalman Filter as presented in Sargent’s (1999) equation (94). To correct this error, \( P_{t-1} \) in Sargent’s (94b) and in the term after the minus sign in (94c) should be replaced with \( P_{t-1} + R_{t} \) in Sargent’s notation, i.e. by \( P_{t-1} + Q_{t} \) 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 (1983, LS) 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 a just random walk as in (9) 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 \( b_{n(t)} \) to indicate the expectation of \( \beta_{i} \) conditional on \( y_{t-1} \) and \( P_{n(t)} \) for its covariance matrix, in addition to \( b_{n}, b_{t-1}, P_{t}, \) and \( P_{t-1} \).

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,

\[
b_{t+1:t} = b_{t|t-1} + K(t)(y_{t} - x_{t} b_{t|t-1}) \tag{A.1}
\]

\[
K(t) = P_{t|t-1} x'_{t}(P_{t|t-1} x'_{t} + \sigma_{e}^{2})^{-1} \tag{A.2}
\]

\[
P_{t+1:t} = P_{t|t-1} + Q_{t} - P_{t|t-1} x'_{t} P_{t|t-1}(x_{t} P_{t|t-1} x'_{t} + \sigma_{e}^{2})^{-1} \tag{A.3}
\]

Since in the random walk case, \( b_{n(t)} \) becomes our \( b_{n} \) and \( P_{n(t)} \) becomes our \( P_{n} \) (A.1) – (A.3) are equivalent to (32) – (34) 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 “\( x_{t} \)” conditional on information up to and including \( t-1 \) (i.e. \( b_{t|t-1} \) above), simply as “\( \hat{x}(t) \)” and to its covariance matrix (\( P_{t|t-1} \) above) simply as “\( \hat{P}(t) \)” 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 “\( \hat{x}(t) \)” to be the expectation of the time \( t \) coefficient vector conditional on information up to and including \( t \), i.e. our \( b_{n} \) and “\( \hat{P}(t) \)” to be its covariance matrix, i.e. our \( P_{n} \). In making this notational revision, however, he simply replaced “\( \hat{P}(t) \)” in his former notation, at all but one point, with “\( \hat{P}(t-1) \)” instead of with \( P_{t|t-1} = P_{t-1} + Q_{t} \), i.e. “\( \hat{P}(t-1) + R_{t}(t) \)” in terms of his new notation, as he should have.5

---

5 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 \( R_{t} \) of this change, if they had, the “\( R_{t}(t-1) \)” of Ljung (1992) would therefore have been “\( R_{t}(t-1) \)” in the LS notation.
In order to correct equations (36) – (39) in Ljung (1992), therefore, “P(t-1)” in (38) and in the expression after the minus sign in (39) should be replaced with “P(t-1) + 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 “P(t-1)” of his 1992 book to be P_{g-1}, despite the apparently contrary definition given in his text which led Sargent (1999) to interpret it as P_{t-1|t-1}. However, he points out that even with this interpretation there is an error, since then the R_1(t) in the first part of (39) on p. 99 should not be present.
REFERENCES


