The Kalman Foundations of Adaptive Least Squares With Applications to Unemployment and Inflation

J. Huston McCulloch
Ohio State University

Jan. 31, 2005

PRELIMINARY AND INCOMPLETE

Adaptive Least Squares (ALS), i.e. recursive regression with asymptotically constant gain, as proposed by Ljung (1992), Sargent (1993, 1999), and Evans and Honkapohja (2001), is an increasingly widely-used method of estimating time-varying relationships and of proxying agents' time-evolving expectations. This paper provides theoretical foundations for ALS as a special case of the generalized Kalman solution of a Time Varying Parameter (TVP) model. This approach is in the spirit of that proposed by Ljung (1992) and Sargent (1999), but unlike theirs, nests the rigorous Kalman solution of the elementary Local Level Model, and employs a very simple, yet rigorous, initialization. Unlike other approaches, the proposed method allows the asymptotic gain to be estimated by maximum likelihood (ML).

The ALS algorithm is illustrated with univariate time series models of U.S. unemployment and inflation. Because the null hypothesis that the coefficients are in fact constant lies on the boundary of the permissible parameter space, the usual regularity conditions for the chi-square limiting distribution of likelihood-based test statistics are not met. Consequently, critical values of the Likelihood Ratio test statistics will be established by Monte Carlo means and used to test the constancy of the parameters in the estimated models.

See http://econ.ohio-state.edu/jhm/papers/KalmanAL.pdf for updates.

I. Introduction

Adaptive Least Squares (ALS), i.e. recursive regression with asymptotically constant gain, as proposed by Ljung (1992), Sargent (1993, pp. 120-2; 1999, ch. 8-9) and Evans and Honkapohja (2001, Ch. 3.3), provides a method of estimating time-varying relationships that is more elegant than rolling regression, yet is far more parsimonious than an unrestricted Time Varying Paramters (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 Equilibrious, or "Rational," Expectations.[1] Bullard and Mitra (2002), Bullard and Duffy (2003), Evans and Honkapohja (2004), Orphanides and Williams (2003), and Preston (2004) are just a few of the many recent applications of the AL concept. Giannitsarou (2004) provides an on-line bibliography of this burgeoning literature.

An early, but very restrictive, special case of ALS was Cagan's (1956) "Adaptive Expectations" (AE) model, in which m_t, the time t expectation of a future variable y_{t+1} (in Cagan's case inflation), was assumed to obey an equation of the form

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

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, e.g. their demand for money balances.

Shortly after Cagan's original paper, however, Muth (1960) and Kalman (1960) independently demonstrated that (1) in fact gives the long-run behavior of the optimal signal extraction forecast of y_{t+1}, if the process is generated by a *Local Level Model* (LLM), i.e. if y_t is the sum of an unobserved Gaussian random walk plus independent Gaussian white noise, provided the long-run gain coefficient γ is computed as a certain function of the constant signal/noise 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 section II below) demonstrated that *in finite samples the ideal gain is not constant,* and in fact declines rapidly at the beginning of the sample. *Kalman's rigorous analysis also allows the signal/noise ratio and therefore the gain coefficients and their limiting value to be estimated by Maximum Likelihood (ML).*

[1] "Adaptive Learning" is often construed to incorporate approaches such as Neural Networks and Genetic Algorithms, in addition to ALS and more general Random Coefficient Models.

The Kalman Filter solution of the elementary LLM has since been generalized to allow a Time Varying Parameter (TVP) model in which all *k* coefficients of a linear regression relation are allowed to change randomly over time, as exposited, for example, by Harvey (1989, Ch. 3). Sargent (1999, Ch. 8), following Ljung (1992), has proposed a restriction on the covariance matrix of the random coefficient changes that leads, by this Extended Kalman Filter (EKF), to ALS with a constant gain. However, because Sargent's gain is constant throughout, his 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. His model in fact incorporates an LLM with a non-constant signal/noise ratio.

Sargent (1999, Ch. 8) goes on to recommend initializing his constant gain ALS recursion with the unconditional expected values of the coefficient vector and covariance matrix. However, by his maintained assumption, the coefficients are nonstationary, and therefore have no unconditional mean, and infinite unconditional variances. In the absence of any truly prior information, the coefficients are in fact underidentified until $t = k$, at which time they have a precise initialization, developed below. [2]

In Sargent's empirical Chapter 9, he provides estimates of two quarterly macroeconomic models with Adaptive Least Squares. 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 (see below) of 66.67 quarters, or 16.67 years.

The present study introduces an alternative specification of the covariance matrix in question, (24) below, that does nest the rigorous declining-gain LLM. The corresponding ALS recursion may be validly initialized with (31) below. Equation (32) below then determines the log-likelihood for the corresponding AL recursion, and permits the signal/noise ratio to be actually estimated by ML rather than simply hypothesized as by Sargent (1999, Ch. 9), or by ad hoc means as in Stock and Watson (1996) and Orphanides and Williams (2004).[3] This algorithm has been implemented in GAUSS as program ALS, and is available on the PI's homepage (McCulloch 2005).

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 develops a parsimonious TVP model that nests the LLM yet at the same time leads in the long run to constant-gain ALS, and discusses related models that have been employed by others. Section IV outlines planned technical extensions of the model. Section V applies

[2] Although the full sample OLS coefficients can easily be computed, they are in no sense "prior" information or "unconditional" values. Ljung (1992, p. 100) unhelpfully instructs his reader to initialize the covariance matrix with an unspecified $\mathbf{P}_0$. Durbin and Koopman (2001, ch. 5) provide an "exact initialization" for the general KF, which may be equivalent to that provided below, although this is not obvious to me at present.

[3] Orphanides and Williams (2004) calibrate their gain coefficient 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. The procedure advocated here is instead to match the likelihood of the realized values as closely as possible. This is what the Professional Forecasters themselves should be doing to calibrate their own forecasting equations.

the ALS algorithm to US unemployment data, while Section VI develops an ALS model of US inflation. Section VII discusses potential future applications.

II. The Local Level Model

Before examining Adaptive Least Squares, we first review and restate the Kalman solution of the elementary Local Level Model in terms of the concepts *Effective Sample Size*, *Cumulative Sample Mass*, and *Cumulative Evidence*.

In the *Local Level Model* (LLM), an observed process y_t is the sum of an unobserved Gaussian random walk μ_t plus independent Gaussian white noise:

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

The *signal/noise ratio* is defined to be

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

so that σ_ε^2 and ρ completely describe the system.

Equation (2) implies

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

so that the distribution of μ_1 given y_1 may be written

$$\mu_1 \mid y_1 \sim N(m_1, \sigma_1^2),$$

where

$$\begin{aligned} m_1 &= y_1, \\ \sigma_1^2 &= \sigma_\varepsilon^2. \end{aligned}$$

Assume now, as we 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

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

It follows that

$$\mu_t \mid \mathbf{y}_{t-1} \sim N(m_{t-1}, \sigma_{t-1}^2 + \sigma_\eta^2) = N(m_{t-1}, \sigma_{t-1}^2 + \rho\sigma_\varepsilon^2). \tag{3}$$

We also know that

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

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

$$
\begin{aligned}
p(\mu_t \mid \mathbf{y}_t) &= p(y_t \mid \mu_t, \mathbf{y}_{t-1})\, p(\mu_t \mid \mathbf{y}_{t-1}) / (const.) \\
&= p(y_t \mid \mu_t)\, p(\mu_t \mid \mathbf{y}_{t-1}) / (const.) \\
&= \exp\left(-\frac{1}{2} \frac{(y_t - \mu_t)^2}{\sigma_\varepsilon^2} \right) \exp\left(-\frac{1}{2} \frac{(\mu_t - m_{t-1})^2}{\sigma_{t-1}^2 + \rho\sigma_\varepsilon^2} \right) / (const.) \qquad (4) \\
&= \exp\left(-\frac{1}{2} \frac{(\mu_t - m_t)^2}{\sigma_t^2} \right) / (const.),
\end{aligned}
$$

where

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

and

$$\frac{1}{\sigma_t^2} = \frac{1}{\sigma_{t-1}^2 + \rho\sigma_\varepsilon^2} + \frac{1}{\sigma_\varepsilon^2}. \qquad (5)$$

In other words,

$$\mu_t \mid \mathbf{y}_t \sim N(m_t, \sigma_t^2)$$

with

$$m_t = \gamma_t y_t + (1 - \gamma_t) m_{t-1} = m_{t-1} + \gamma_t (y_t - m_{t-1}), \qquad (6)$$

where

$$\gamma_t = \sigma_t^2 / \sigma_\varepsilon^2 .$$

In the special case $\rho = 0$, so that $\mu_t = \mu$, a constant, we have

$$\gamma_t = \sigma_t^2 / \sigma_\varepsilon^2 = 1/t .$$

When $\rho > 0$, the gain is somewhat larger than the reciprocal of the true sample size, as is the ratio $\sigma_t^2 / \sigma_\varepsilon^2$. In other words, the *Effective Sample Size*, T_t, which we define by

$$T_t = \sigma_\varepsilon^2 / \sigma_t^2 = 1/\gamma_t ,$$

is less than the true sample size t. In terms of this T_t, (3) becomes

$$\mu_t \mid \mathbf{y}_{t-1} \sim N\left(m_{t-1}, (1 + \rho T_{t-1}) \sigma_{t-1}^2\right), \qquad (7)$$

and (5) becomes

$$T_t = (1 + \rho T_{t-1})^{-1} T_{t-1} + 1, \qquad (8)$$

with

$$T_0 = 0 .$$

Note that so long as $\rho > 0$, $T_t > T_{t-1}$, yet $T_t < T_{t-1} + 1$, so that the effective sample size grows with t, but more slowly than the true sample size. Furthermore,

$$\lim_{t \uparrow \infty} T_t = T ,$$

where

$$T = 1/2 + \sqrt{1/4 + 1/\rho} \qquad (9)$$

is the unique positive root of the quadratic equation

$$\rho T^2 - \rho T - 1 = 0$$

that defines the fixed points of (8). The constant gain AE formula (1) is therefore strictly valid only in this limit, with the limiting gain $\gamma = 1/T$.[4]

Equation (7) implies

$$y_t \mid \mathbf{y}_{t-1} \sim N\left(m_{t-1}, (1+\rho T_{t-1})\sigma_{t-1}^2 + \sigma_\varepsilon^2\right),$$

which can be used to compute the log joint probability of y_2, ... y_n conditional on y_1 as a function of σ_ε^2 and ρ, and therefore the log likelihood of σ_ε^2 and ρ given y_1 as a function of y_2, ... y_n. The observation variance σ_ε^2 may be concentrated out of the log likelihood function, so that a numerical maximization search is only required over the single parameter ρ.

Given the effective sample sizes T_t as determined by (8), the LLM Kalman Filter may equivalently be computed in terms of what we may call the *cumulative evidence* z_t and the *cumulative sample mass* w_t as

$$m_t = z_t / w_t,$$

by means of the recursions

$$z_t = (1+\rho T_{t-1})^{-1} z_{t-1} + y_t, \quad (10)$$

$$w_t = (1+\rho T_{t-1})^{-1} w_{t-1} + 1, \quad (11)$$

together with the initial conditions

$$z_0 = 0,$$

$$w_0 = 0,$$

$$T_0 = 0.$$

Equations (8), (10) and (11) show that the passage of one period of time diminishes the effective sample size, the evidence, and the sample mass accumulated to date by the common factor

$$(1+\rho T_{t-1})^{-1} = (T_t - 1)/T_{t-1} < 1 \quad (11)$$

before the new time unit, evidence and sample mass (1, y_t, and 1, respectively) are added onto them. Since the LLM just has a (time-varying) level and no regressors, the cumulative sample mass w_t and the effective sample size T_t are here one and the same thing.

III. Adaptive Least Squares

Consider now the standard fixed coefficient regression equation,

$$y_t = \mathbf{x}_t \boldsymbol{\beta} + \varepsilon_t, \quad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2), \quad (12)$$

---

[4] As noted above, Muth (1960) developed the limiting gain γ, but not the exact finite sample gain γ_t required for ML estimation of ρ.

where $\mathbf{x}_t$ is a $1{\times}k$ *row* vector[5] of ideally exogenous explanatory variables, and $\boldsymbol{\beta}$ is a $k{\times}1$ column vector of coefficients. 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 explanatory variables up to and including time t. Ordinarily the first column of $\mathbf{X}_t$ is a vector of units, so that β_1 is the intercept.

It is well known (e.g. (10) of Sargent 1993 or (2.9) of Evans and Honkapohja 2001) that the OLS estimator $\mathbf{b}_t = \left(\mathbf{X}_t'\mathbf{X}_t\right)^{-1}\mathbf{X}_t'\mathbf{y}_t$ of $\boldsymbol{\beta}$ given the data up to and including time t can be expressed in terms of the *Recursive Least Squares* (RLS) formula[6]

$$\mathbf{b}_t = \mathbf{b}_{t-1} + \gamma_t \mathbf{R}_t^{-1}\mathbf{x}_t\left(y_t - \mathbf{x}_t\mathbf{b}_{t-1}\right), \tag{13}$$

where $\gamma_t = 1/t$ and $\mathbf{R}_t = (1/t)\mathbf{X}_t'\mathbf{X}_t$, the sample average value of the regressor outer product $\mathbf{x}_t'\mathbf{x}_t$, may be updated by

$$\mathbf{R}_t = \mathbf{R}_{t-1} + \gamma_t(\mathbf{x}_t'\mathbf{x}_t - \mathbf{R}_{t-1}). \tag{14}$$

The variance of $\mathbf{b}_t$ is then given by

$$\mathbf{P}_t = \gamma_t\sigma_\varepsilon^2\mathbf{R}_t^{-1}. \tag{15}$$

The ALS literature commonly replaces $\gamma_t = 1/t$ in (13) – (15) by a constant γ as in Cagan's original Adaptive Expectations formulation (1). However, this constant-gain ALS 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 parameter determining the long run gain itself.

The simplistic LLM that leads asymptotically to (1) allows the dependent variable y_t to depend only on a simple (time-varying) mean. A much more general framework is the Random-Coefficients linear regression Model,

$$\begin{aligned} y_t &= \mathbf{x}_t\boldsymbol{\beta}_t + \varepsilon_t, \quad \varepsilon_t \sim NID(0,\sigma_\varepsilon^2), \\ \boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim NID(\mathbf{0}_{k\times 1},\mathbf{Q}_t), \end{aligned} \tag{16}$$

where $\boldsymbol{\eta}_t$ is a $k{\times}1$ column vector of transition errors independent of the observation errors ε_t, and $\mathbf{Q}_t$ is a $k{\times}k$ covariance matrix.

System (16) may be solved recursively by means of the well-known Extended Kalman Filter. Assume that we have found a rule according to which, for $t > k$,

$$\boldsymbol{\beta}_{t-1} \mid \mathbf{y}_{t-1} \sim N(\mathbf{b}_{t-1},\mathbf{P}_{t-1}) \tag{17}$$

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),

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

where

$$\mathbf{b}_t = \mathbf{b}_{t-1} + f_t^{-1}\left(\mathbf{P}_{t-1} + \mathbf{Q}_t\right)\mathbf{x}_t'\left(y_t - \mathbf{x}_t\mathbf{b}_{t-1}\right), \tag{18}$$

---

[5] We make $\mathbf{x}_t$ a row vector rather than a column vector, since $\mathbf{x}_t$ is simply the t-th row of the regressor matrix $\mathbf{X}_n$.

[6] RLS follows immediately from the matrix identity $\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{B}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{A}^{-1}$.

$$\mathbf{P}_t = (\mathbf{P}_{t-1} + \mathbf{Q}_t)(\mathbf{I}_{k\times k} - f_t^{-1}\mathbf{x}_t'\mathbf{x}_t(\mathbf{P}_{t-1} + \mathbf{Q}_t)), \tag{19}$$

$$f_t = \mathbf{x}_t'(\mathbf{P}_{t-1} + \mathbf{Q}_t)\mathbf{x}_t' + \sigma_\varepsilon^2 .^7 \tag{20}$$

The textbook Kalman Filter equations (18) and (19) above may be rearranged to eliminate f_t and to look more like RLS, as follows: Post-multiply (19) by $\mathbf{x}_t'$ to obtain

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

so that (18) becomes

$$\mathbf{b}_t = \mathbf{b}_{t-1} + (1/\sigma_\varepsilon^2)\mathbf{P}_t\mathbf{x}_t'(y_t - \mathbf{x}_t\mathbf{b}_{t-1}), \tag{21}$$

and (19) becomes

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

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

$$\mathbf{P}_t^{-1} = (\mathbf{P}_{t-1} + \mathbf{Q}_t)^{-1} + (1/\sigma_\varepsilon^2)\mathbf{x}_t'\mathbf{x}_t . \tag{22}$$

The full-blown random coefficients model (16) is much too general for our purposes, however, since if even if $\mathbf{Q}_t$ is made time-invariant, it still introduces $k(k$-1)/2 incidental time-variation hyperparameters in addition to the observation variance σ_ε^2.

Cooley and Prescott (1973) were able to reduce $\mathbf{Q}_t$ to a single parameter, but only by allowing only the intercept to change, so that $\mathbf{Q}_t$ has only a single non-zero element. Their model nests the LLM, but not ALS.

More generally, Sims (1988) and Kim and Nelson (2004) use (**16**) 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 of 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 not ALS.

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, even aside from the above considerations, since it forces all the coefficients to have the same transition variance (at

---

[7] There is an error in Sargent's (1999) equation (94), which does not match Harvey's (3.2.3) unless $\mathbf{P}_{t-1}$ in (94b) and in the term after the minus sign in (94c) is replaced with $\mathbf{P}_{t-1} + \mathbf{R}_{1t}$ in Sargent's and Ljung's notation, i.e. $\mathbf{P}_{t-1} + \mathbf{Q}_t$ in ours (and Harvey's). The same error appears in Sargent's source, Ljung (1992), equations (36)-(39). However, Ljung's own source, Ljung and Söderström (1983), is correct. See Appendix below for details.

each point in time), even though their units depend upon the often arbitrary units in which the regressors happen to be measured.

In order to obtain a rigorous foundation for long-run fixed-gain ALS, however, it is sufficient and natural simply to postulate, following Ljung (1992) and Sargent (1999, p. 117), that $\mathbf{Q}_t$ *is directly proportional to* $\mathbf{P}_{t-1}$. It is apparent from (22) why one would want $\mathbf{Q}_t$ to be proportional to $\mathbf{P}_{t-1}$ and not $\mathbf{P}_t$, say. Nevertheless, the proportionality that Ljung and Sargent propose must be modified in order to reduce to the elementary LLM when $k = 1$.

Let ρ be an index of the uncertainty of the transition errors relative to the observation errors, such that T_t as computed from ρ as in (8) measures the *Effective Sample Size*. Recall that in the LLM, the variance of the "noise", i.e. the observation errors, is related to that of the estimation errors at time t - 1 by

$$\sigma_\varepsilon^2 \equiv T_{t-1}\sigma_{t-1}^2,$$

so that the variance of the "signal", i.e. the transition error η_t, is given by

$$\sigma_\eta^2 \equiv \rho T_{t-1}\sigma_{t-1}^2. \tag{23}$$

In the same spirit, we assume that $T_{t-1}\mathbf{P}_{t-1}$ measures the *measurement error per effective observation* as of time t - 1, just as does in the LLM, and thus that the transition covariance matrix $\mathbf{Q}_t$ of $\boldsymbol{\eta}_t$ in (16) is given by

$$\mathbf{Q}_t = \rho T_{t-1}\mathbf{P}_{t-1}. \tag{24}$$

When the random coefficients model (16) contains only an intercept term and no regressors, (24) becomes (23). Hence, the proposed covariance specification exactly nests the LLM.

Defining $\mathbf{R}_t = \left(\sigma_\varepsilon^2 / T_t\right)\mathbf{P}_t^{-1}$, under (24) the restated Kalman Filter equations (21) and (22) immediately become the variable-gain RLS equations (13) and (14), with gain

$$\gamma_t = 1/T_t,$$

exactly as in the LLM. As t becomes large, the gain takes on the fixed value $\gamma = 1/T$, where T is as in (9).

However, the required recursion is in fact much simpler, and the entire derivation of the filter quite transparent, if we dispense with $\mathbf{R}_t$ altogether and consider instead the *cumulative sample mass matrix* $\mathbf{W}_{t-1} = \sigma_\varepsilon^2\mathbf{P}_{t-1}^{-1} = T_{t-1}\mathbf{R}_{t-1}$ and the *cumulative evidence vector* $\mathbf{z}_{t-1} = \mathbf{W}_{t-1}\mathbf{b}_{t-1}$, as follows: Equations (17) and (24) imply, analogously to (7),

$$\boldsymbol{\beta}_t \mid \mathbf{X}_{t-1}, \mathbf{y}_{t-1} \sim N\left(\mathbf{b}_{t-1}, (1+\rho T_{t-1})\mathbf{P}_{t-1}\right)$$

Furthermore,

$$y_t \mid \boldsymbol{\beta}_t, \mathbf{x}_t \sim N(\mathbf{x}_t\boldsymbol{\beta}_t, \sigma_\varepsilon^2).$$

Noting that $(\mathbf{x}_t\boldsymbol{\beta}_t)^2 = (\boldsymbol{\beta}_t'\mathbf{x}_t')(\mathbf{x}_t\boldsymbol{\beta}_t) = \boldsymbol{\beta}_t'(\mathbf{x}_t'\mathbf{x}_t)\boldsymbol{\beta}_t,$ we have, as in (4),

$$
\begin{aligned}
p(\boldsymbol{\beta}_t \mid \mathbf{X}_t, \mathbf{y}_t) &= \mathrm{p}(\mathrm{y}_t \mid \boldsymbol{\beta}_t, \mathbf{x}_t)\, p(\boldsymbol{\beta}_t \mid \mathbf{X}_{t-1}, \mathbf{y}_{t-1}) / (const.) \\
&= \exp\left(-\frac{(y_t - \mathbf{x}_t \boldsymbol{\beta}_t)^2}{2\sigma_\varepsilon^2} \right) \exp\left(-\frac{(1+\rho T_{t-1})^{-1} (\boldsymbol{\beta}_t - \mathbf{b}_{t-1})' \mathbf{W}_{t-1} (\boldsymbol{\beta}_t - \mathbf{b}_{t-1})}{2\sigma_\varepsilon^2} \right) / (const.) \\
&= \exp\left(-\frac{\boldsymbol{\beta}_t' \left((1+\rho T_{t-1})^{-1} \mathbf{W}_{t-1} + \mathbf{x}_t' \mathbf{x}_t \right) \boldsymbol{\beta}_t - 2\boldsymbol{\beta}_t' \left((1+\rho T_{t-1})^{-1} \mathbf{W}_{t-1} \mathbf{b}_{t-1} + \mathbf{x}_t' y_t \right)}{2\sigma_\varepsilon^2} \right) / (const.) \\
&= \exp\left(-\frac{(\boldsymbol{\beta}_t - \mathbf{b}_t)' \mathbf{W}_t (\boldsymbol{\beta}_t - \mathbf{b}_t)}{2\sigma_\varepsilon^2} \right) / (const.),
\end{aligned}
$$

where

$$\mathbf{b}_t = \mathbf{W}_t^{-1} \mathbf{z}_t, \tag{25}$$

$$\mathbf{z}_t = (1+\rho T_{t-1})^{-1} \mathbf{z}_{t-1} + \mathbf{x}_t' y_t, \tag{26}$$

and

$$\mathbf{W}_t = (1+\rho T_{t-1})^{-1} \mathbf{W}_{t-1} + \mathbf{x}_t' \mathbf{x}_t . \tag{27}$$

The variance of $\mathbf{b}_t$ is then

$$\mathbf{P}_t = \sigma_\varepsilon^2 \mathbf{W}_t^{-1}. \tag{28}$$

Equations (26) and (27) are essentially the "information" filter mentioned by Harvey (1989, p. 108), cp. also Bullard (1992), since our $\mathbf{W}_t$ is just a scaled version of the information matrix $\mathbf{P}_t^{-1}$. They are equivalent to (13) and (14), but make clear how, as in the LLM, the accumulated evidence $\mathbf{z}_t$ and sample mass $\mathbf{W}_t$ diminish in the same proportion as T_t before accruing the new evidence and sample mass. When $\rho = 0$, $\mathbf{z}_t$ becomes $\mathbf{X}_t'\mathbf{y}_t$, $\mathbf{W}_t$ becomes $\mathbf{X}_t'\mathbf{X}_t$, and (25) becomes the familiar OLS formula.

We now consider initialization of the above recursion. For $t < k$, the distribution $\boldsymbol{\beta}_t|\mathbf{y}_t$ is defective, since the $t \times k$ matrix $\mathbf{X}_t$ has rank $t < k$. Nevertheless, $\mathbf{X}_t\boldsymbol{\beta}_t$ has a proper t-dimensional distribution that characterizes what we can say about $\boldsymbol{\beta}_t$. For $t = 1$,

$$\mathbf{X}_1 \boldsymbol{\beta}_1 = y_1 - \varepsilon_1 \sim N(\mathbf{y}_1, \boldsymbol{\Sigma}_1),$$

where $\boldsymbol{\Sigma}_1 = \left(\sigma_\varepsilon^2\right)$ is a 1×1 matrix. Now suppose that we have found that

$$\mathbf{X}_{t-1} \boldsymbol{\beta}_{t-1} \sim N(\mathbf{y}_{t-1}, \boldsymbol{\Sigma}_{t-1})$$

for some (t-1)×(t-1) matrix $\boldsymbol{\Sigma}_{t-1}$. We have

$$\mathbf{X}_{t-1} \boldsymbol{\beta}_t = \mathbf{X}_{t-1} \boldsymbol{\beta}_{t-1} + \mathbf{X}_{t-1} \boldsymbol{\eta}_t$$

If, in the spirit of (24), we also assume

$$\mathbf{Cov}\left(\mathbf{X}_{t-1} \boldsymbol{\eta}_t\right) = \rho T_{t-1} \mathbf{Cov}\left(\mathbf{X}_{t-1} \boldsymbol{\beta}_{t-1}\right) = \rho T_{t-1} \boldsymbol{\Sigma}_{t-1},$$

we have

$$\mathbf{X}_{t-1} \boldsymbol{\beta}_t \sim N\left(\mathbf{y}_{t-1}, (1+\rho T_{t-1}) \boldsymbol{\Sigma}_{t-1}\right).$$

We also know that

$$\mathbf{x}_t \boldsymbol{\beta}_t = y_t - \varepsilon_t,$$

so that

$$\mathbf{X}_t \boldsymbol{\beta}_t \sim N(\mathbf{y}_t, \boldsymbol{\Sigma}_t),$$

where

$$\boldsymbol{\Sigma}_t = \begin{pmatrix} (1+\rho T_{t-1})\boldsymbol{\Sigma}_{t-1} & \mathbf{0}_{k\times 1} \\ \mathbf{0}_{1\times k} & \sigma_\varepsilon^2 \end{pmatrix}.^{8} \tag{29}$$

After using (29) recursively to find Σ_k, we therefore may, indeed *must* in the absence of true prior information, initialize our recursion at $t = k$ with

$$\begin{aligned} \mathbf{b}_k &= \mathbf{X}_k^{-1}\mathbf{y}_k, \\ \mathbf{P}_k &= \mathbf{X}_k^{-1}\boldsymbol{\Sigma}_k\mathbf{X}_k'^{-1}. \end{aligned} \tag{30}$$

Equivalently, and more elegantly, we may simply initialize (26) and (27) with

$$\begin{aligned} \mathbf{z}_0 &= \mathbf{0}_{k\times 1} \\ \mathbf{W}_0 &= \mathbf{0}_{k\times k}, \end{aligned} \tag{31}$$

and use (25) and (28) for all $t \geq k$, since then for $t \leq k$,

$$\begin{aligned} \mathbf{z}_t &= \mathbf{X}_t'\boldsymbol{\Omega}_t^{-1}\mathbf{y}_t, \\ \mathbf{W}_t &= \mathbf{X}_t'\boldsymbol{\Omega}_t^{-1}\mathbf{X}_t, \end{aligned}$$

where

$$\boldsymbol{\Omega}_t = \frac{1}{\sigma_\varepsilon^2}\boldsymbol{\Sigma}_t,$$

so that for $t = k$,

$$\begin{aligned} \mathbf{W}_k^{-1}\mathbf{z}_k &= \left(\mathbf{X}_k'\boldsymbol{\Omega}_k^{-1}\mathbf{X}_k\right)^{-1}\mathbf{X}_k'\boldsymbol{\Omega}_k^{-1}\mathbf{y}_k \\ &= \mathbf{X}_k^{-1}\boldsymbol{\Omega}_k\mathbf{X}_k'^{-1}\mathbf{X}_k'\boldsymbol{\Omega}_k^{-1}\mathbf{y}_k \\ &= \mathbf{X}_k^{-1}\mathbf{y}_k, \end{aligned}$$

and (28) holds for $t = k$.

Having thus initialized and updated the filter, the log likelihood of the two hyperparameters σ_ε^2 and ρ may be found by adding together the log densities for $y_{k+1}|\mathbf{y}_k$, ... $y_n|\mathbf{y}_{n-1}$, where

$$\begin{aligned} y_t \mid \mathbf{y}_{t-1} &\sim N\left(\mathbf{x}_t\mathbf{b}_{t-1}, \sigma_\varepsilon^2 s_t^2\right), \\ s_t^2 &= (1+\rho T_t)\mathbf{x}_t\mathbf{W}_{t-1}^{-1}\mathbf{x}_t' + 1 \end{aligned} \tag{32}$$

As in the LLM, the observation variance σ_ε^2 may be concentrated out of the log likelihood function, so that a numerical maximization search is required only over the single parameter ρ.

---

[8] The recursion (29) could, if desired, be continued beyond $t = k$, and then used to find $\mathbf{b}_t$ by Generalized Least Squares (GLS). This would give the same numerical answer as the Kalman Filter, but without its computational efficiency.

Note that although $\boldsymbol{\Sigma}_t$ is diagonal for $t \leq k+1$, it is not diagonal for $t > k+1$. Thus although ALS is numerically equivalent to the solution of a Weighted Least Squares (WLS) problem, the problem it solves is not a WLS problem. The same is true of the LLM, which in this respect is a special case of ALS.

For (16) with general $\mathbf{Q}_t$, the same argument leads to (29) with the upper left element of the RHS replaced by $\boldsymbol{\Sigma}_{t-1} + \mathbf{X}_{t-1}\mathbf{Q}_t\mathbf{X}_{t-1}'$. (30) may then be used at $t = k$ to initialize the efficient Kalman filter. As noted above, it is possible that this would be equivalent to the "exact initialization" developed by Durban and Koopman (2001, ch. 5), though this is not obvious to the author at present.

If the model is well-specified, the variance-equalized forecast errors

$$u_t = \left(y_t - \mathbf{x}_t \mathbf{b}_{t-1}\right)/ s_t \tag{33}$$

should be iid N(0, σ^2). Routine specification tests such as Q statistics, the Jarque-Bera test, etc., may be applied to these. (See Durbin and Koopman, Ch. 5).

The Kalman *Filter* for the ALS model provides the best estimate of the coefficient vector given the *past 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 measure the ex post value of the regression coefficients given *both prior and subsequent experience*, the Kalman *Smoother* becomes the appropriate tool. This is straightforward, but requires some care because of the asymmetrical (backward- rather than forward-looking) nature of our assumption about the transition matarix covariance matrix. The smoother and its covariance matrix have already been developed and incorporated into program ALS (McCulloch 2005), and an illustration of its use is given below. Details will be forthcoming.

Ljung (1992) and Sargent (1999, Ch. 8) in fact assume, in place of (24), that

$$\mathbf{Q}_t = \frac{\gamma}{1-\gamma} \mathbf{P}_{t-1}, \tag{33}$$

with the result that (13) holds with a constant γ in place of $\gamma_t = 1/T_t$.[9] Under this Ljung-Sargent assumption, the initial observations are given too little weight. This underweighting makes little difference for the final estimates of the regression coefficients or the long-run behavior of the system if ρ is known, but will distort the early estimates and will cause the ML estimate of ρ and therefore the computed asymptotic gain to be biased in a finite sample. Note that the Ljung-Sargent specification, unlike the present model, does not nest the LLM, since it in fact implies a time-varying signal/noise ratio.

Stock and Watson (1996) and Sargent and Williams (2003) assume, in place of either (24) or (33), that

$$\mathbf{Q}_t = \mathbf{Q} = \rho \sigma_\varepsilon^2 \left(E \mathbf{x}_t' \mathbf{x}_t\right)^{-1}. \tag{34}$$

If the relevant expectation exists, this is equivalent in an expectational sense to (24), since then

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

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 are. Our equation (24), on the other hand, does not require these moments to be finite, and only requires agents to know $\mathbf{X}_t$, $\mathbf{y}_t$, and ρ.[10] Assumption (34) does nest the LLM, since then the required expectation is just a unit scalar. For $k > 1$, however, it only approximates ALS with gain $1/T_t$.

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

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

Stock and Watson (1996) calibrate the gain coefficient ρ in (34) (their λ^2) by minimizing the sum of squared forecasting errors. This will give similar to ours, but by no means equivalent, even apart from the difference between our (24) and their (34). For one thing, the initial errors have much larger variance than the later errors simply because the coefficient vector is still highly uncertain. Equation (32) 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 grossly take this factor into account, by discarding the 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 (32), which is similar to the formula for the conditional distribution that would result from (34), that even asymptotically the errors are not homoskedastic, and hence should not be given equal weight.

Orphanides and Williams (2004) calibrate their gain coefficient both by minimizing a sum of squared 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.

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. Their procedure is far more computation-intensive than ALS, however.

IV. Technical extensions

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-square 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.

The author plans in the near future to determine Monte Carlo critical values for the LR statistic under the null of no change. These Monte Carlo critical values will be adjusted for multiple-test Monte Carlo sampling error using the methodology introduced by McCulloch (1997, p. 79). However, in that paper all critical values were corrected for 100 independent tests of the null hypothesis in question. A more appropriate correction would be to correct the p-critical value for 1/p independent tests of the null hypothesis. This modification is easily implemented.

It is conjectured that these critical values will not depend asymptotically on either the sample size or the number of regressors (k) in the model, let alone on the numerical values taken by the regressors.[11] 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.

The Kalman *Filter* for the ALS model provides the best estimate of the coefficient vector given the *past 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 measure the ex post value of the regression coefficients given *both prior and subsequent experience*, the Kalman *Smoother* becomes the appropriate tool. This is straightforward, but requires some care because of the asymmetrical (backward- rather than forward-looking) nature of the assumption (24). The smoother has already been implemented in program ALS (McCulloch 2005), but has not yet been written up here.

Omitting a regressor from the model will reduce calculated likelihood. However, it is not clear at this point to the author just what justification there would be for using the resulting likelihood ratio statistic to test the hypothesis that the coefficient on this regressor is uniformly zero, since this time-varying coefficient is not a hyperparameter of the model.

If one is estimating an autoregression by ALS, it is important to remember that, as in OLS, the inverse AR roots are biased downwards, particularly as they approach 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 T. It therefore may be important to mean- or median-unbias the AR coefficients according to the effective sample size before using them to simulate forecasts. Such a correction is proposed by Fuller and Roy (2001) and has been implemented, using US inflation data with expanding window regression, by McCulloch and Stec (2000). See also Harvey (1989, ch. 7) concerning endogenous regressors.

If the coefficients change over time, it is possible that the observation variance (and therefore also, holding the signal/noise ratio constant, the transition variance) likewise changes over time. This can incorporated into the present ALS model as a GARCH effect. Such GARCH effects were first introduced into an econometric model by McCulloch (1985), were applied to a non-Gaussian signal extraction problem by Bidarkota and McCulloch (1998), and have been incorporated into a random-coefficients model by Kim and Nelson (2004). However, the incorporation of such effects into ALS have not yet been implemented.

V. Application to US Unemployment Rate.

Orphanides and Williams (2004) find, using constant-gain ALS, that perceptions of the "natural unemployment rate" have not remained constant over the last 5 decades,

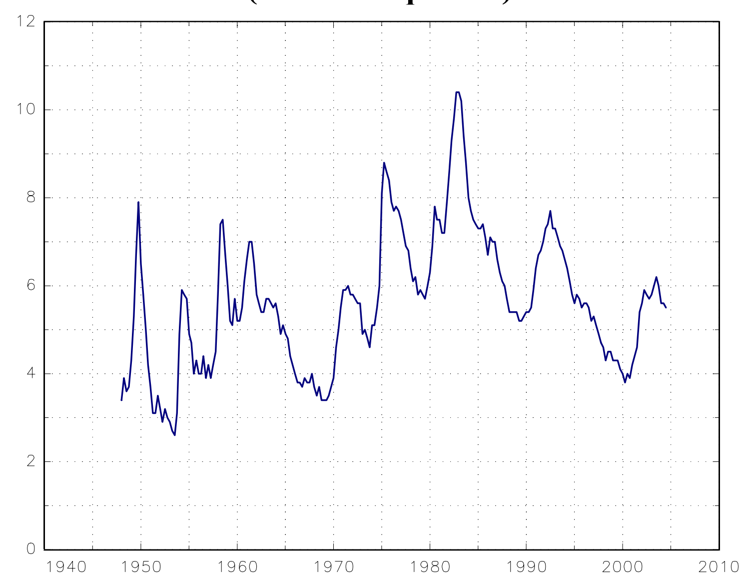
---

[11] Note that in order to be directly comparable to the likelihood under the alternative, the likelihood under the null should be computed as $\sum_{t=k+1}^{n} p(y_t \mid \mathbf{y}_{t-1})$, rather than as $\sum_{t=1}^{n} p(y_t \mid \mathbf{y}_n)$ as in OLS.

but rather have drifted up and back down again. If so, perhaps the autoregressive coefficients of its dynamics have also undergone changes. Rigorous declining-gain ALS allows this relationship to be estimated as far back as there are reliable data, and to test for constancy of the parameters as discussed above.

Figure 1 below shows the US Civilian unemployment rate for 1948Q1 – 2004Q3, as seasonally adjusted by the BLS. The raw monthly data has been converted to a quarterly basis, using the first month of each quarter as a proxy for the entire quarter. Note that there is an uptrend in unemployment to about 1983, followed by a downtrend. Furthermore, the swings in unemployment are more peristent in the later period than in the earlier period, as evidenced by the fact that the first four major peaks in unemployment span 12 years, while the last four major peaks span 28 years. It may be hoped that ALS will pick up these effects.

Figure 1.
Civilian Unemployment Rate, SA, 48Q1 – 04Q3
(1st mo. of quarter)



The following AR(2) model was fit by ALS to the data for 1948Q3 – 2004Q3:

$$U_t = \beta_{1t} + \beta_{2t}\, U_{t-1} + \beta_{3t}\, U_{t-2} + \varepsilon_t$$

The ALS ML estimates of the signal/noise parameter and its implied long-run gain and effective sample size are:

$\rho = 0.000397$ (s.e. = 0.00352)

$\gamma = 0.01974$

$T = 1/\gamma = 50.66$ qtr. = 12.67 yr.

Because 3 regression coefficients are being estimated, the first predictive density that can be computed is for the 4th observation, counting 1948Q3 as the first after reserving two lags. The log likelihood is therefore

$$\log L(\rho, \sigma^2) = p(y_4, \ldots y_n | y_{-1}, \ldots y_3) = -110.32.$$

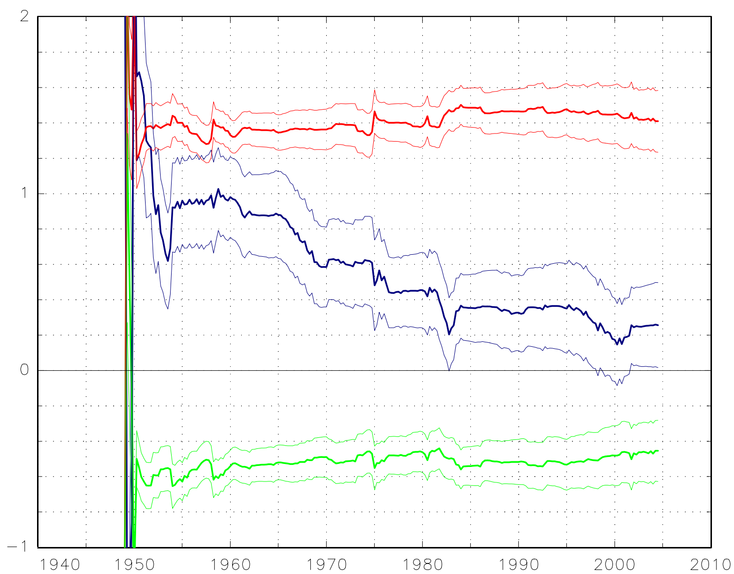
When ρ is constrained to be 0, the joint density for the same observations with the same conditioning declines, and the likelihood ratio statistic LR, i.e. twice the change log likelihood, is

$$LR\ (\rho = 0) = 4.31.$$

As noted above, this statistic does not have its customary chi-squared distribution with 1 degree of freedom because the null lies on the boundary of the permissible parameter space. Nevertheless, preliminary Monte Carlo simulations (using only 100 replications) indicate that the 5% critical value is approximately 2.3. We may therefore tentatively reject constancy at the 5% level, despite the large standard error of ρ, as estimated from the Hessian of the likelihood function at the point estimate.

Figure 2 shows the time-varying filter coefficient estimates, together with ±1 s.e. bands, with the intercept in blue (in black and white the central, darkest group), the AR(1) term in red (the upper group), and the AR(2) term in green (the lower, lightest group).

Figure 2
Filter Coefficient Estimates
$b_{1,t}$ (blue), $b_{2,t}$ (red), $b_{3,t}$ (green), ± 1 s.e.



The AR(2) coefficient is always at least 2 standard errors below 0. When the second lag of unemployment is excluded from the ALS regression altogether, the equivalently conditioned joint density of the same observations declines to

$$\log L = p(y_4, \ldots y_n | y_{-1}, \ldots y_3;\ \beta_{3,t} = 0) = -139.87$$

and the corresponding LR statistic is

$$LR\ (\beta_{3,t} = 0) = 59.10$$

This value is dramatically in excess of the customary 5% chi-squared critical value of 3.84, but as noted above, it is not clear why this is meaningful, since this does not correspond to a restriction on either of the hyperparameters of the model. We may tentatively conclude, however, that the process is at least AR(2) at this quarterly frequency, with a negative coefficient. The state of the labor market at any point in time therefore depends both on the level of unemployment and the direction of change of this level from the quarter before. In particular, the negative AR(2) coefficient implies that the direction of change of the unemployment rate has pronounced positive inertia.

Figure 3 shows the standard errors by themselves for the filter coefficient estimates of Figure 2. At the beginning of the sample, these are offscale because of the very small effective sample size that has accumulated. However, by T = 12.67 years, if not before, they stabilize with only minor subsequent fluctuations, caused by the evolving weighted moment matrix. Because the moments of the first and second lags are almost identical, their coefficients have virtually the same standard errors.

Figure 3
ALS Filter Standard Errors
$b_{1,t}$ (blue), $b_{2,t}$ (red), $b_{3,t}$ (green).

Figure 4 below shows the intercept term again from Figure 2, this time in red, along with the sum of the two AR coefficients, in blue, together with their 1 s.e. bands as computed from the coefficient covariance matrix. There is a pronounced secular decline in the intercept, but at the same time a pronounced increase in the sum of the lag coefficients, and therefore in the persistence of unemployment.

Figure 4
$b_{1,t}$ (red), $b_{2,t} + b_{3,t}$ (blue), $\pm$ 1 s.e.

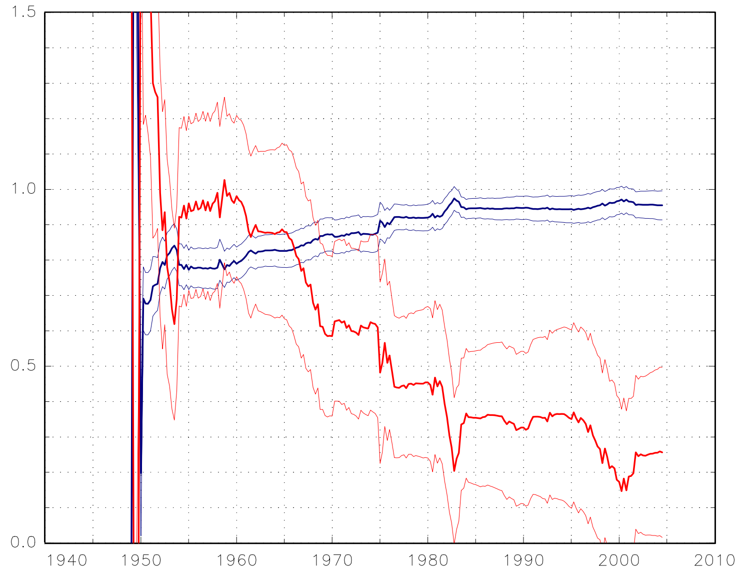
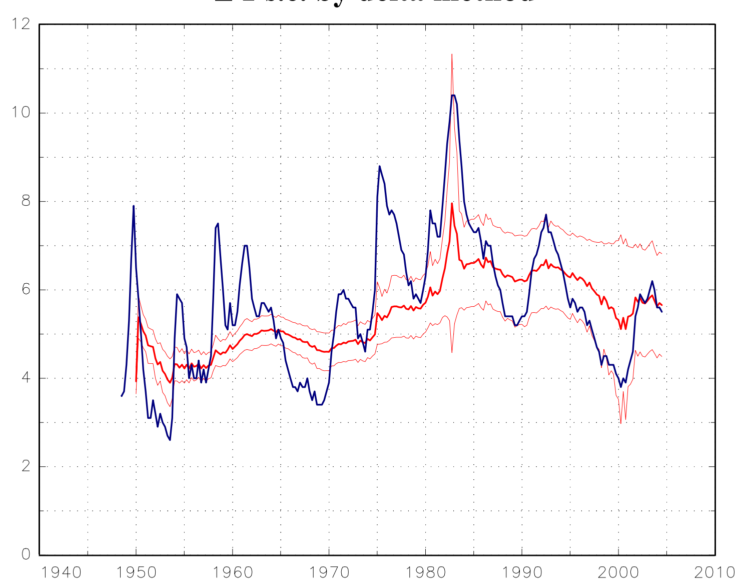


Figure 5 belows show filter estimates of the equilibrium or natural unemployment rate, computed as

$$U^N_t = \beta_{1t} / (1 - (\beta_{2t} + \beta_{3t})),$$

together with its 1 s.e. bands as approximated by the delta method, along with the unemployment rate itself. A multivariate model of the natural rate might also take inflation surprises into account, but would probably not yield substantially different estimates of the natural rate.

Figure 5
U (blue), ALS Filter Estimates of U^N (red),
± 1 s.e. by delta method



Although the sum of the lag coefficients is biased downwards, as noted above, this bias is offset by a compensating bias in the intercept, so that it is not obvious that the point estimate of the natural rate itself is biased at all. Nevertheless, the validity of the delta method depends on the standard errors of the numerator and denominator of the required coefficients being small in comparison to their respective point estimates, which is not at all the case here. The estimated standard errors should therefore be used with extreme caution. The tendency of the standard errors to grow with time is related to the fact that the numerator and denominator of the expression evaluated both become closer to zero as time proceeds.

It may be seen that the secular decline in the intercept from 1950 to 1980 was more than offset by the secular rise in the persistence of unemployment, as measured by the sum of the lag coefficients, with the net result that the estimated natural rate rose during this period. After 1980 this trend was reversed, although the subsequent decline in the natural rate has been small in comparison with the approximate standard errors. The actual unemployment rate was not significantly above the filter-estimated natural rate even in the early 1980s, and since 1980, the actual rate has always been within 1 standard error of the point estimate of the natural rate. The terminal estimate of the natural rate, for 2004Q3, is 5.66%, but with a standard error of 1.16%.

The filter estimates of the Natural Rate in Figure 5 simulate real-time estimates that could have been made by policy makers or agents at the time in question.

Figure 6 shows the variance-equalized ALS forecast errors, as computed from (33). Clearly some volatility clustering is present, so that more efficient point estimates and more accurate standard errors could be obtained by GARCH-ALS. This is still under development, however. It appears that some kurtosis will be present even after the volatility clustering is removed.

Figure 6
Variance-equalized ALS forecast errors

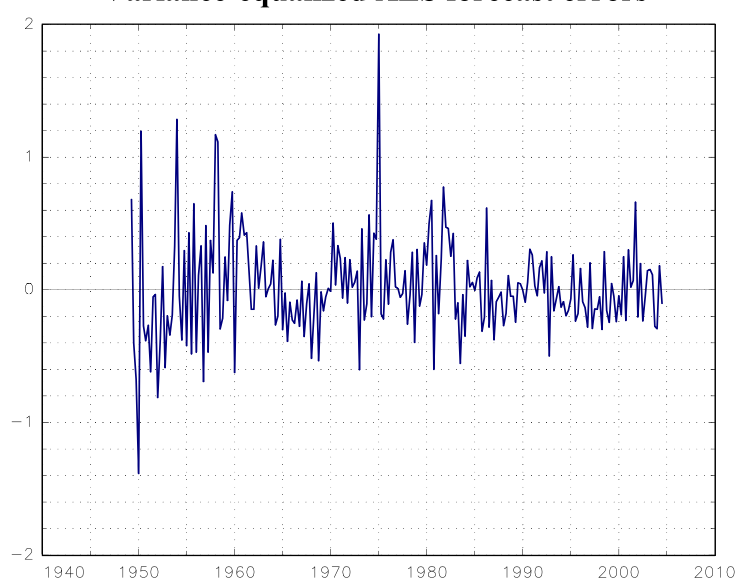


Figure 7 below shows the ALS smoother coefficient estimates, computed using the subsequent data as well as the preceding data. Except near the ends of the data set, the smoother standard errors should be about 71% as large as the filter standard errors, since they are based on almost twice as much data.

Figure 7
ALS Smoother Coefficient Estimates
$b^S_{1,t}$ (blue), $b^S_{2,t}$ (red), $b^S_{3,t}$ (green), ± 1 s.e.

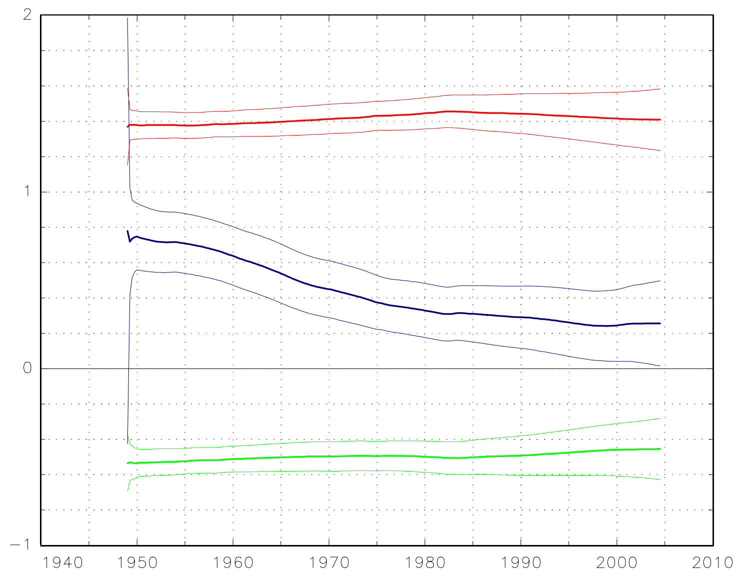


Figure 8 shows the smoother estimates of the natural unemployment rate, along with the unemployment rate itself. While these estimates are more precise than the filter estimates, they could not have been known to policy makers in real time. Note that actual unemployment is significantly greater than the smoother estimate of the natural rate, if not the filter estimate, in the early 80s, and is occasionally more than 1 standard error from the natural rate in the early 90s and around 2000.

Figure 8
U, ALS Smoother Estimates of U_N, ± 1 s.e.

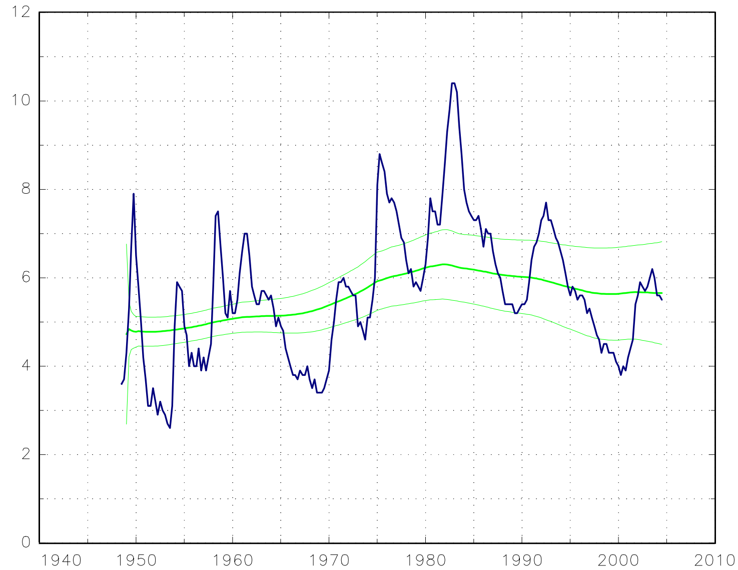
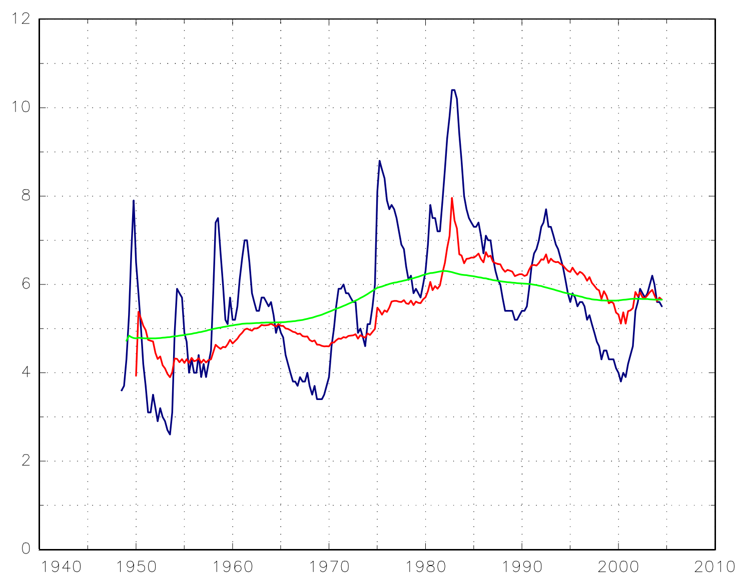


Figure 9 shows the smoother (green) and filter (red) point estimates of the natural rate, along with the raw unemployment rate (blue). It may be seen that the Fed could reasonably have underestimated the natural rate before 1992, and overestimated it afterwards, as suggested by Orphanides and Williams (2004). However, the difference between these two estimates is small in comparison with their standard errors. Note that the smoother and filter necessarily coincide at the end of the sample.

Figure 9
U (blue) with ALS filter (red), smoother (green) estimates of U_N



VI. Application to US CPI Inflation

As Klein (1978) pointed out early on, 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 became more and more persistent. Writing in 1971, Sargent (1971) was still able to argue that inflation was clearly a stationary process, but by 1974, a unit root in CPI inflation could no longer be rejected using an expanding window regression with fixed coefficients, as demonstrated by McCulloch and Stec (2000). A univariate time series model of the US inflation is therefore a natural application of the method. Monthly CPI inflation has strong seasonality that itself varies from decade to decade. This is easily accommodated with ALS, since it automatically permits such variation.

Figure 10 shows the U.S. CPI-U (n.s.a.) from 1913.1 to 2004.8. In order to reduce rounding error, the 1967 base year was employed. For 1967.1 to 1983.8, the BLS published a CPI-X, which retroactively computed the housing component using the rental equivalent basis adopted in 1983. This was spliced into the CPI-U to obtain what may be called the CPI-UX (blue).

Figure 10
CPI-U (red), CPI-UX (blue)

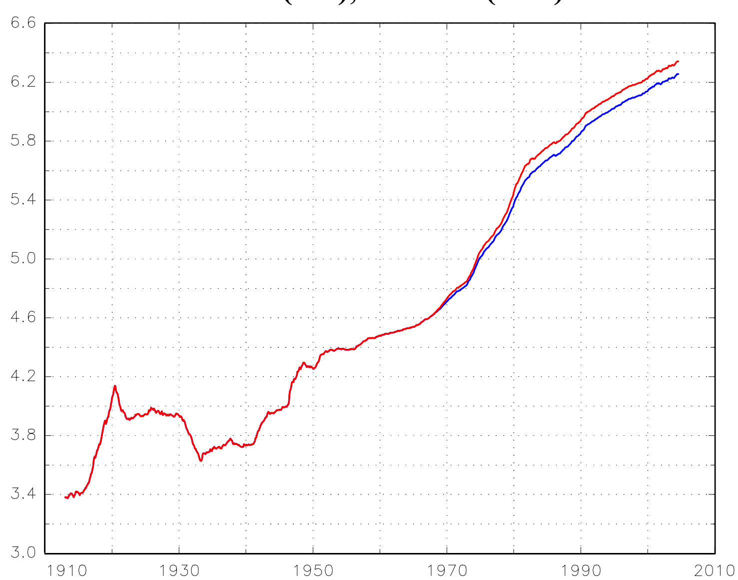
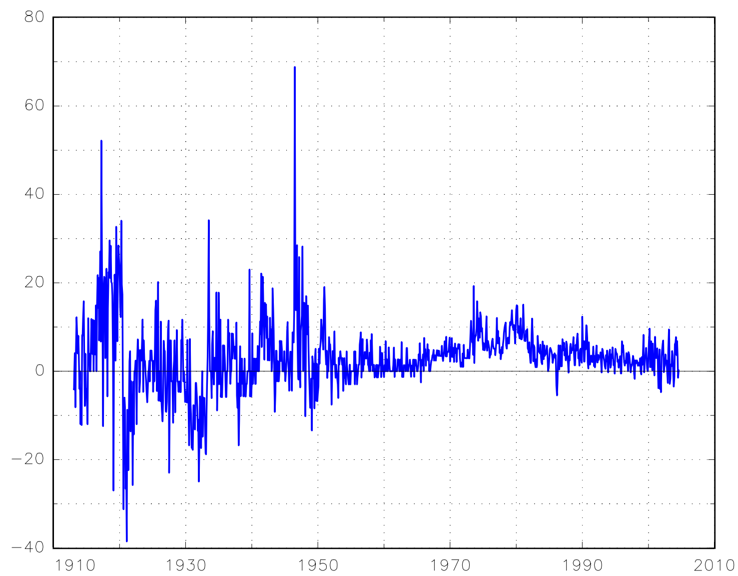


Figure 11 shows the annualized percentage logarithmic inflation rate computed from the data of Figure 10, for 1913.2-2004.8.

Figure 11
CPI-UX inflation (annualized, percent)



A restricted AR(24) model was fit by ALS to this data. In order to reduce the number of free AR parameters to 5, the AR coefficients were constrained by a linear spline, with knot points at 2, 4, 7, 13, and 25 months, with a zero restriction at the last knot. Since monthly CPI inflation has strong seasonals, 12 seasonal dummy variables were included in place of a single intercept, for a total of $k = 17$ parameters with $n = 1100$ observations. Despite the huge number of computations required, ML estimation of the model, complete with smoother estimates, takes only about 75 seconds on a PC. The ALS ML estimates were

$\rho = 0.000025$ (s.e. = .000009)

$\gamma = 0.00499$; $T = 1/\gamma = 200.27$ mo. = 16.69 yr.

LR ($\rho = 0$) = 29.262

Once again, constancy will likely be easily rejected, even after the extreme GARCH effects (not shown for reason of space) are taken into account.

Figure 12 shows the estimated seasonal contribution to inflation, computed as the time-varying coefficient on the respective seasonal dummy variable, minus their average, so that at any point in time these seasonals identically average to 0. The vertical scale is annualized percent inflation. It may be seen that seasonals were far more pronounced prior to 1960 than afterwards. In the 1920s-40s, April had the strongest positive seasonal, while February had the strongest negative seasonal. The seasonals were much smaller in size during the 1970s and 80s, but since 1990 have grown in size again, with January now being the most positive and December the most negative.

Figure 12
Seasonals

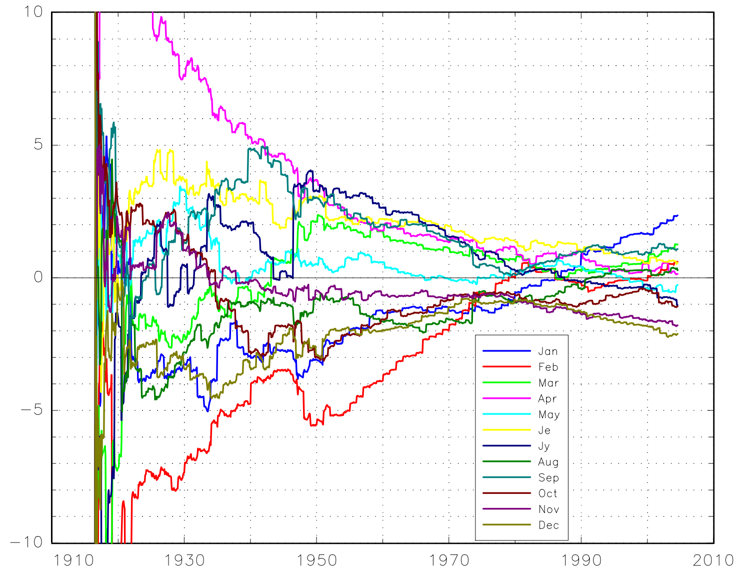


Figure 13 shows the time-varying coefficients on selected weighted averages of inflation that serve as a basis for the set of first degree splines with knots at 2, 4, 7, 13, and 25 months and a zero long-end restriction. INF1 is simply 1 lag of inflation. INF3 is a weighted average of 1 to 3 lags of inflation, with linearly decaying weights declining to 0 when projected to month 4. INF6 is a similar linearly weighted average of 1 to 6 months of inflation, etc. Such weighted averages were in fact first proposed by Irving Fisher in 1930, who gave them the name “Distributed Lags”. When combined in this manner, they generate first degree splines. The fact that the coefficient on INF24 is quite small in comparison to the others suggests that inflation lagged by more 12 months has little if any marginal predictive power. This hypothesis is examined further below.

Figure 13
Coefficients on Spline Basis Functions of Inflation

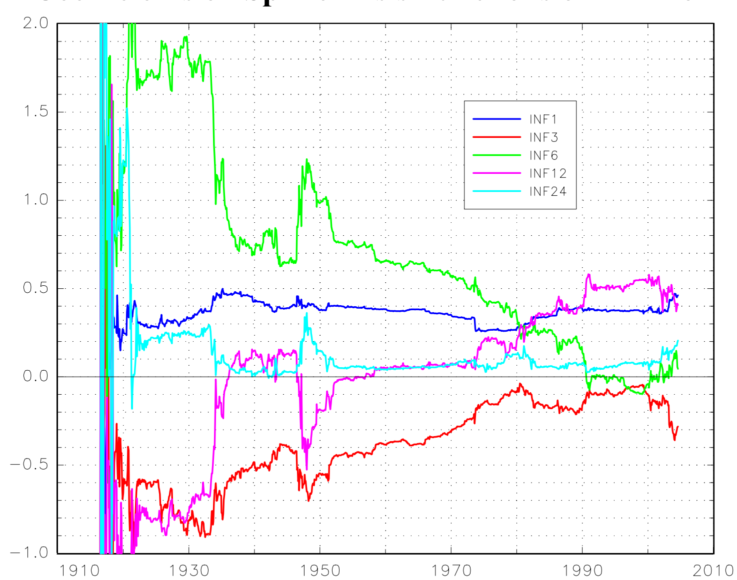


Figure 14 shows the net lag structures implied by the coefficients in Figure 13 for selected dates, in roughly 10-year intervals. These dates are January of the year shown, except for 2004, which is August. The lag structures are remarkably stable over time, at least qualitatively. The first lag is always near 0.4, the second lag dips to 0.1 or smaller, and then there is a hump in lags 3-6. In lags 7-13, the coefficients decline to near 0. In fact, prior to 1985, the coefficients are nearly 0 already by lag 7.

Figure 14
Net Lag Coefficients, Selected Dates

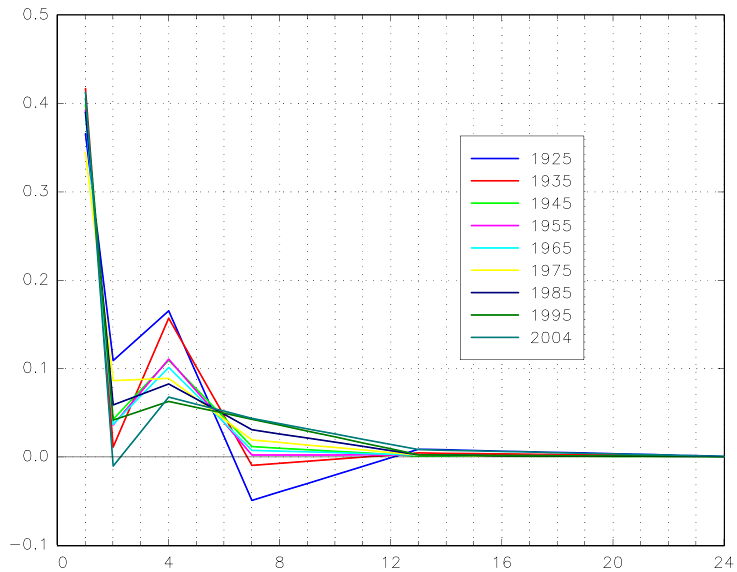


Figure 15 shows the net lag coefficients again, but now for only a few selected lag lengths versus all calendar dates. Since the lag lengths shown are the spline knots, the coefficients for the omitted lag lengths are just linear interpolations of their two adjoining maturities.

Figure 15
Net Lag Coefficients, Selected Lags

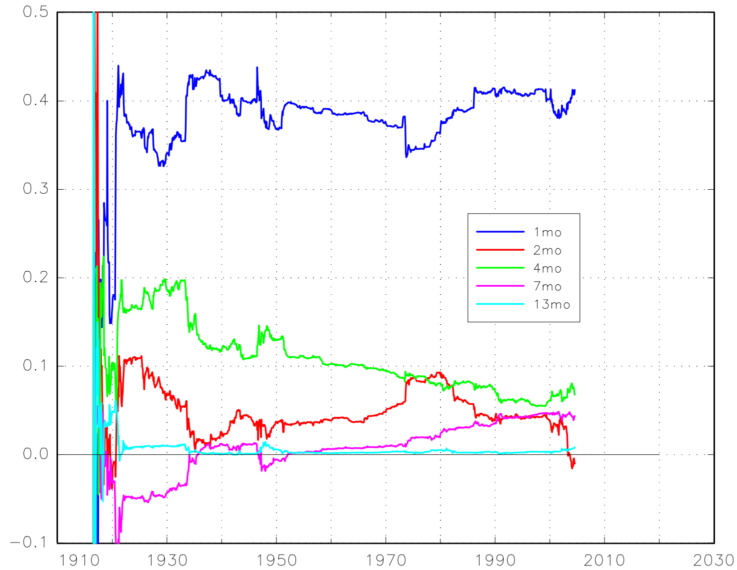


Figure 16 shows the coefficient on INF24 from Figure 13, plus and minus 1 standard error. Apart from vertical scale, this is the same line as the 13 month net lag coefficient in Figure 16, since only INF24 contains the 13th lag of inflation. Indeed, the coefficient is never significantly different from 0, so that omitting INF24 would probably not hurt the predictive power of the equation. As noted above, however, it is not clear to me at this point how test such a hypothesis formally in the present framework.

Figure 16
Coefficient on INF24, ± 1 s.e.

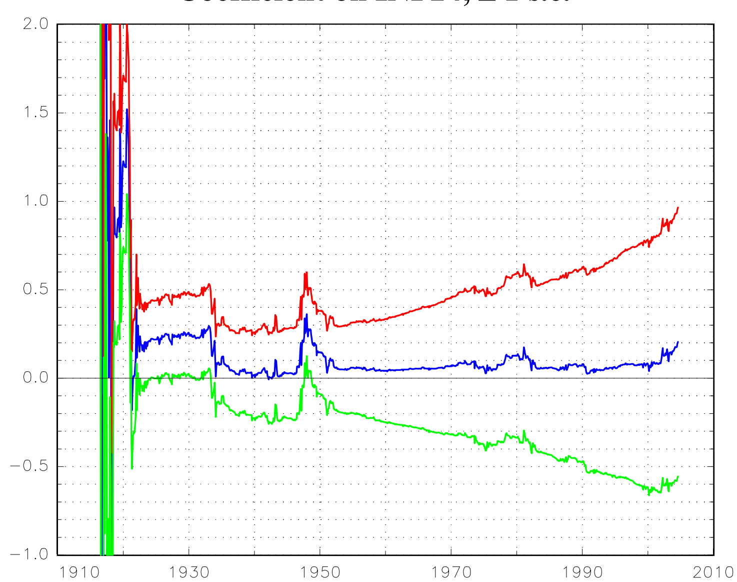


Figure 17 shows the seasonally adjusted intercept in blue, computed as the average of the time-varying coefficients on the seasonal dummies, along with the sum of the 24 lag coefficients, in red.

Figure 17
S.A. Intercept (blue), Sum of Lag Coefficients (red)

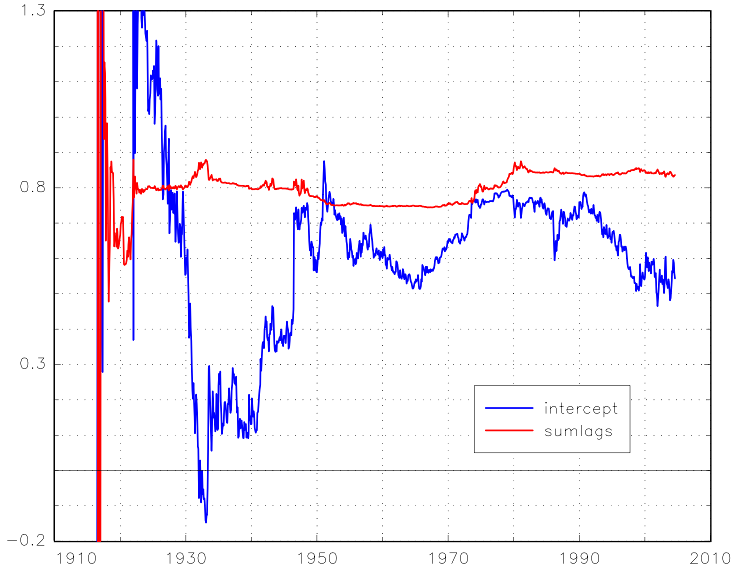
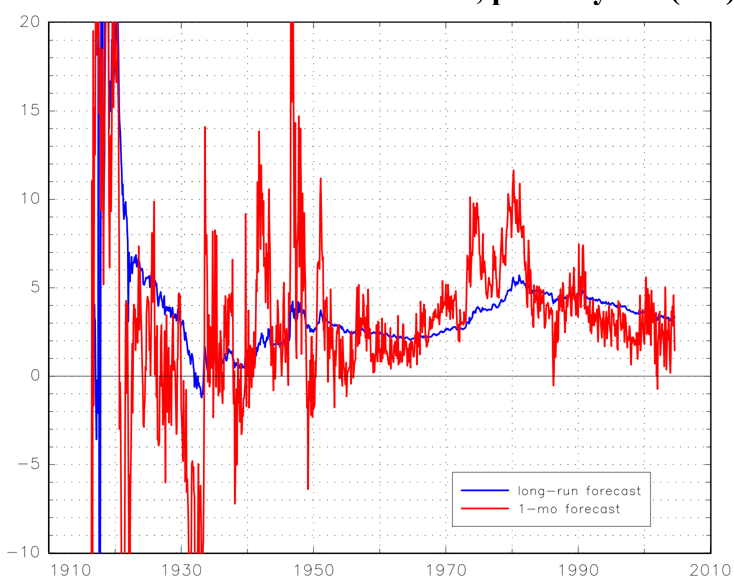


Figure 18 shows the long-run inflation forecast, in blue, along with the 1-month-ahead inflation forecast, partially seasonally adjusted, in red. The long-run forecast is computed from the data in Figure 17 as the seasonally adjusted intercept divided by one minus the sum of the lag coefficients. This is the inflation rate that would eventually prevail if the coefficients all remained constant at their current levels and no shocks to inflation itself occurred.

Figure 18
Long-Run Inflation Forecast (blue),
1-Month-Ahead Inflation Forecast, partially s.a. (red)

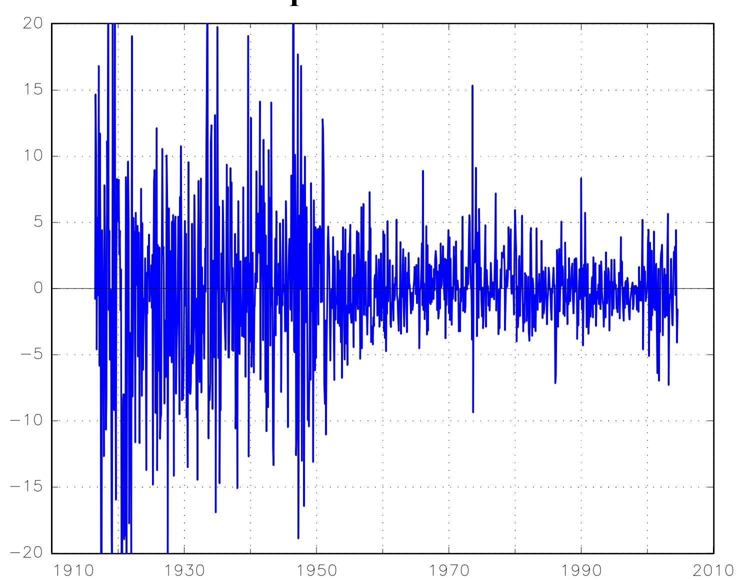


The 1-month-ahead forecasts in Figure 18 have been only partially seasonally adjusted, by subtracting the relevant seasonal from Figure 12 from the raw 1-month-ahead forecast. Full seasonal adjustment would require also seasonally adjusting the lagged inflation rates from which this forecast is computed, though I have not yet had a chance to implement this. Such complete seasonal adjustment would considerably reduce the volatility of the 1-month-ahead forecasts. Inflation forecasts could easily be computed for any horizon intermediate between 1 month and infinity.

In the simplistic Cagan Adaptive Expectations model of (1) and its LLM rationalization (2), short-run and long-run inflation forecasts are one and the same thing. It may be seen from Figure 18 that with the present much richer ALS model, there are often substantial differences between the two. In 1932, for example, simulated short-run annualized inflation forecasts were running around -8%, while the long-run forecast never fell below -1%. In 1980-81, short-run forecasts of CPI-X-adjusted inflation were about 9%, while long-run forecasts were barely 5%. In August of 2004, the (only partially s.a.) 1-month ahead forecast was 1.73%, while the long-run forecast was 3.32%.

Figure 19 shows the variance-equalized forecast errors, computed as in (33). There is even more pronounced evidence of volatility clustering here than in our model of unemployment. Whether this is due to the nature of inflation before 1950, or just due to improvements in data collection since then, this volatility clustering should be removed with a GARCH-ALS model in order to validly compute standard errors and test parameter constancy . As noted, such a model is still under development.

Figure 19
Variance-Equalized Forecast Errors



VII. Potential Future Applications

According to Perron (1983), real GDP growth apparently underwent a permanent but unforeseen decline in the vicinity of 1973. A univariate time series model of log real GDP will therefore also be of interest, using as long a history of quarterly real GDP as possible, and annual real GDP as far back as the data are reliable.

Clarida, Galí and Gertler (2000), Orphanides and Williams (2003), Kim and Nelson (2004), and others have found time variation in the "Taylor Equation" monetary policy response function and/or in policy makers' simulated forecasts of the variables that go into the policy response function. ALS provides a rigorous method of estimating such a time-varying policy equation. If the response function is "forward-looking" in the sense of responding to forecasts of the variables in question, this is a two-stage procedure, in which forecasts are first simulated using a ALS filter, and then a time-varying response function is estimated using the simulated forecasts, using the ALS smoother.

A long literature, going back to Goldfeld (1976), argues that US money demand parameters occasionally undergo permanent shifts. The current "New-Keynesian" conventional wisdom (e.g. Woodford 2003) is that such shifts render money aggregates irrelevant for monetary policy. However, an ALS framework accommodates such shifts, and at the same time allows money demand to be forecasted meaningfully (if not precisely) into the future.

Among other approaches, I will use the equation

$$\pi_t = E^*_{t-1}(\pi_t) + \lambda(m_{t-1} - m^D_{t-1}) + \varepsilon_t ,$$

developed in McCulloch (1980), to estimate log real money demand m_t^D, where π_t is inflation from period t-1 to t, and E^*_{t-1} indicates the public's expectations as of time t-1. ALS filter estimates will be used to proxy these expectations, while the ALS smoother will be used to estimate the adjustment coefficient λ along with the parameters of money demand. My previous attempts to implement this equation empirically were frustrated by lack of a rigorous way to proxy the public's expectations that would differ from the econometrician's. ALS now allows the former to be constructed from past experience, while the latter is constructed from past *and* future experience.

Appendix

As mentioned in footnote 6 above, there is an error in the Kalman Filter as presented in Sargent's (1999) equation (94). To correct this error, $\mathbf{P}_{t\text{-}1}$ in Sargent's (94b) and in the term after the minus sign in (94c) should be replaced with $\mathbf{P}_{t\text{-}1} + \mathbf{R}_{1t}$ in Sargent's notation, i.e. by $\mathbf{P}_{t\text{-}1} + \mathbf{Q}_t$ 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 (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 (16) 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 $\mathbf{b}_{t|t\text{-}1}$ to indicate the expectation of $\boldsymbol{\beta}_t$ conditional on $\mathbf{y}_{t\text{-}1}$, and $\mathbf{P}_{t|t\text{-}1}$ for its covariance matrix, in addition to $\mathbf{b}_t$, $\mathbf{b}_{t\text{-}1}$, $\mathbf{P}_t$, and $\mathbf{P}_{t\text{-}1}$.

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,

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

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

$$\mathbf{P}_{t+1|t} = \mathbf{P}_{t|t-1} + \mathbf{Q}_{[t+1]} - \mathbf{P}_{t|t-1}\mathbf{x}_t'\mathbf{x}_t\mathbf{P}_{t|t-1}\left(\mathbf{x}_t \mathbf{P}_{t|t-1}\mathbf{x}_t' + \sigma_\varepsilon^2\right)^{-1}. \tag{A.3}$$

Since in the random walk case, $\mathbf{b}_{t+1|t}$ becomes our $\mathbf{b}_t$ and $\mathbf{P}_{t|t\text{-}1}$ becomes our $\mathbf{P}_{t\text{-}1} + \mathbf{Q}_t$, (A.1) – (A.3) are equivalent to (18) – (20) 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. $\mathbf{b}_{t|t\text{-}1}$ above), simply as "$\hat{\mathbf{x}}(t)$," and to its covariance matrix ($\mathbf{P}_{t|t\text{-}1}$ 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 "$\mathbf{P}(t\text{-}1)$", instead of with $\mathbf{P}_{t|t\text{-}1} = \mathbf{P}_{t\text{-}1} + \mathbf{Q}_t$, i.e. "$\mathbf{P}(t\text{-}1) + \mathbf{R}_1(t)$" in terms of his new notation, as he should have.[12]

In order to correct equations (36) – (39) in Ljung (1992), therefore, "$\mathbf{P}(t\text{-}1)$" in (38) and in the expression after the minus sign in (39) should be replaced with "$\mathbf{P}(t\text{-}1)$ +

[12] 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 "$\mathbf{R}_1(t\text{-}1)$" in the LS notation.

$\mathbf{R}_1(t)$." Corresponding replacements should be made in Sargent's (1999) (94), as noted above.

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

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.

Cho, In-Koo, Noah Williams, and Thomas J. Sargent. "Escaping Nash Inflation," *Review of Economic Studies* **69** (2002), 1-40.

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.," University of California, Davis, and New York University, June 2004. <http://www.econ.ucdavis.edu/faculty/twcogley/D&V21.pdf>.

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

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.

Evans, George W., and Seppo Honkapohja. "Adaptive Learning and Monetary Policy Design," *Journal of Money, Credit, and Banking* **3**5 (2003): 1045-72.

Fuller, Wayne A., and Anindya Roy, "Estimation in Autoregressive Time Series with a Root Near 1," J*ournal of Business and Economic Statistics* **19** (2001): 482-93.

Giannitsarou, Chryssi. “Adaptive Learning in Macroeconomics” (2004), online bibliography at <http://docentes.fe.unl.pt/~chryssi/Adaptive Learning in Macroeconomics.html>

Goldfeld, Stephen. "The Case of the Missing Money," *Brooking Papers on Economic Activity* **3** (1976): 683-730.

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. “The Microfoundations of the Moderate Quantity Theory,” Ohio State University, 1980.

McCulloch, J. Huston. “Interest-Risk Sensitive Deposit Insurance Premia: Stable ACH Estimates,” *Journal of Banking and Finance* **9** (1985): 137-156.

McCulloch, J. Huston. “Measuring Tail Thickness to Estimate the Stable Index α: A Critique,” *Journal of Business & Economic Statistics* **15** (1997), 74-81.

McCulloch, J. Huston. GAUSS program ALS. Fits a linear regression by Adaptive Least Squares, estimating the signal/noise parameter by Maximum Likelihood, 2005, <http://www.econ.ohio-state.edu/jhm/programs/ALS>

McCulloch, J. Huston, and Jeffery A. Stec. “Generating Serially Uncorrelated Forecasts of Inflation by Estimating the Order of Integration Directly,” Ohio State University Working paper, 2000, <http://www.econ.ohio-state.edu/jhm/inflexp.doc>.

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

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, <http://www.federalreserve.gov/pubs/ifdp/2004/804/default.htm>.

Perron, Pierre. "The Great Crash, the Oil Price Shock and the Unit Root Hypothesis," *Econometrica* **58**, 1989, 1361-1401.

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," NYU and Princeton University, April 2003, at <www.princeton.edu/~noahw>

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: 11-30.

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

Woodford, Michael. *Interest and Prices*, Princeton University Press 2003.