The Kalman Foundations of Adaptive Least Squares, With Application to U.S. Inflation

J. Huston McCulloch
Ohio State University

August 2005
PRELIMINARY AND INCOMPLETE
(corr. 5/7/06, 8/1/07)

The author is indebted to James Bullard, Kan Chen, Chang-Jin 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.

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

ABSTRACT

Adaptive Least Squares (ALS), a refinement of the Constant Gain Recursive Least Squares (CGRLS) algorithm proposed by Ljung (1992) and Sargent (1993, 1999), is a 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. The approach is in the spirit of that proposed by Ljung and Sargent, but unlike theirs, nests the rigorous Kalman solution of the elementary Local Level Model, and employs a very simple, yet rigorous, initialization. Unlike CGRLS and related approaches, ALS permits estimation of the asymptotic gain by maximum likelihood (ML).

The ALS algorithm is illustrated with a univariate time series model of monthly U.S. inflation, using data for 1913 - 2005. The estimated asymptotic gain is 0.00779, for a limiting effective sample size of 128.4 months or 10.7 years. Volatility clustering is pronounced, and is modeled as GARCH (1,1).

Because the null hypothesis that the coefficients are constant lies on the boundary of the permissible parameter space, the regularity conditions for the chi-square limiting distribution of likelihood-based test statistics are not met. Consequently, critical values of the Likelihood Ratio (LR) test statistics are established by Monte Carlo means and used to test and overwhelmingly reject the constancy of the parameters in the estimated model.

Time-specific hypotheses on coefficients may be tested with a customary chi-square test, using either the ALS filter or smoother. Global hypotheses are tested with the Variance Ratio (VR) statistic, which extends the customary LR statistic. Using it, a restricted time-varying AR(6) model of inflation is rejected in favor of a restricted AR(12) specification, while AR(12) cannot be rejected in favor of AR(24). Seasonality is globally significant, although locally insignificant from 1950 – 1973.

Keywords: Adaptive Learning, Kalman Filter, Inflation

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

I. Introduction

Adaptive Least Squares (ALS), a refinement of the Constant Gain Recursive Least Squares (CGRLS) proposed by Ljung (1992) and Sargent (1993, pp. 120-2), provides a method of estimating time-varying relationships that is more elegant than rolling regression, yet 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.[1] 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 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 CGRLS 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

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

In Cagan's original formulation, the gain coefficient γ was assumed to be an arbitrary subjective constant to be inferred indirectly from agents' expectationally motivated behavior, 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 y_{t+1}, but only provided 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, and provided the long-run gain coefficient γ is computed as a certain function of the 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 Appendix I 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 sometimes construed to incorporate approaches such as Neural Networks and Genetic Algorithms, in addition to ALS and the general TVP model.

The Kalman Filter solution of the elementary LLM has since been generalized to solve 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). Ljung (1992) and Sargent (1999, Ch. 8) have proposed a restriction on the covariance matrix of the random coefficient changes that leads, by this Extended Kalman Filter (EKF), to CGRLS, i.e. ALS with a constant gain. 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. Their model in fact nests an LLM with a non-constant signal/noise ratio.

Sargent (1999, Ch. 8) goes on to recommend initializing CGRLS 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 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 CGRLS. 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 a new specification of the covariance matrix in question, (12) below, that does nest the rigorous declining- yet bounded-gain LLM. The corresponding ALS recursion may be naturally and validly initialized with (17) and (18) below. Equation (19) below then determines the log-likelihood, and permits the signal/noise ratio to be actually estimated by ML rather than simply postulated as by Sargent (1999, Ch. 9), or estimated 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 author's homepage (McCulloch 2005a).

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 develops hypothesis tests. Section V extends the model to incorporate GARCH(1,1) errors. Section VI applies the ALS-GARCH algorithm to US CPI inflation

[2] Although 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.

data, while Section VII briefly mentions potential future applications. Appendix I provides derivations, while Appendix II corrects an error in a critical equation in Ljung (1992) and Sargent (1999).

II. The Local Level Model

Before examining 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 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. Let the vector $\mathbf{y}_t = (y_1, \ldots y_t)'$ represent the observations up to and including y_t.

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

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

with

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

$$\sigma_t^2 = \frac{1}{T_t}\sigma_\varepsilon^2,$$

where the *Effective Sample Size*, T_t, is determined by

$$T_t = \left(1 + \rho T_{t-1}\right)^{-1} T_{t-1} + 1, \tag{5}$$

with initialization

$$T_0 = 0\,. \tag{6}$$

In the special case $\rho = 0$, so that $\mu_t = \mu$, a constant, the effective sample size T_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

$$\lim_{t \uparrow \infty} T_t = T = 1/2 + \sqrt{1/4 + 1/\rho}\,, \tag{7}$$

the unique positive root of the quadratic equation

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

that defines the fixed points of (5). The constant gain AE formula (1) is therefore strictly valid only in this limit, with the limiting gain $\gamma = 1/T$.[4] The gain in fact should be $1/T_t$, which behaves much like $1/t$ for small values of t.

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

$$y_t \mid y_1 \ldots y_{t-1} \sim N\left(m_{t-1}, \sigma_{t-1}^2 + \rho\sigma_\varepsilon^2 + \sigma_\varepsilon^2\right). \tag{8}$$

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 σ_ε^2 and ρ, and therefore the log likelihood of σ_ε^2 and ρ conditioned on $y_1, \ldots, y_n$. 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.

III. Adaptive Least Squares

Cagan's Adaptive Expectations formula (1) is thus the rigorous long-run solution to a well-specified statistical model. Cagan, however, erroneously treated his gain like a subjective learning parameter to be inferred from agents' expectationally-motivated behavior, specifically their demand for money, instead of estimating ρ from the series in question using (8) and then computing the limiting gain according to (7).[5]

Furthermore, the simplistic LLM that leads asymptotically to Cagan's formula (1) allows the dependent variable y_t to depend only on a simple (time-varying) mean. A much more general framework is the Time-Varying Parameter (TVP) linear regression model,

$$\begin{aligned} y_t &= \mathbf{x}_t\boldsymbol{\beta}_t + \varepsilon_t, \quad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2), \\ \boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim NID(\mathbf{0}_{k\times 1}, \mathbf{Q}_t), \end{aligned} \tag{9}$$

in which $\mathbf{x}_t$ is a $1\times k$ *row* vector[6] of ideally exogenous explanatory variables, $\boldsymbol{\beta}_t$ is a $k\times 1$ column vector of time-varying coefficients, and $\boldsymbol{\eta}_t$ is a $k\times 1$ column vector of transition errors that are independent of the observation errors ε_t. Let $\mathbf{y}_t$ be the $t\times 1$ vector of dependent variables observed up to and including time t, and $\mathbf{X}_t$ be the $t\times k$ matrix of explanatory variables up to and including time t. Ordinarily the first column of $\mathbf{X}_t$ is a vector of units, so that the first element of $\boldsymbol{\beta}_t$ is the intercept. $\mathbf{Q}_t$ is the possibly time-dependent $k\times k$ covariance matrix of the transition errors $\boldsymbol{\eta}_t$.

System (9) may be solved by means of the well-known Extended Kalman Filter, which provides a recursive rule giving

---

[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.

[5] Milani (2005) likewise makes the mistake of calibrating his CGRLS gain parameter by optimizing the fit of an ancillary, fixed coefficient New Keynesian Phillips Curve equation.

[6] 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 $\mathbf{X}_n$.

$$\boldsymbol{\beta}_t \mid \mathbf{y}_t \sim N(\mathbf{b}_t, \mathbf{P}_t) \tag{10}$$

for some $k \times k$ covariance matrix $\mathbf{P}_t$. (See Appendix I for details.) The full-blown TVP model (9) 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 to be estimated, 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 only the (1, 1) element of $\mathbf{Q}_t = \mathbf{Q}$ is non-zero. Their model nests the LLM, but does not yield ALS.

More generally, 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 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 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* $\mathbf{P}_{t\text{-}1}$. 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 a scalar index of the uncertainty of the transition errors relative to the observation errors, such that T_t as computed from ρ as in (5) 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

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

In the LLM, $\mathbf{Q}_t$ is simply the 1×1 matrix (σ_η^2), while $\mathbf{P}_{t\text{-}1}$ is simply (σ_{t-1}^2), so that the variance of the "signal", i.e. the scalar transition error η_t, is given by

$$\mathbf{Q}_t = \left(\sigma_\eta^2\right) \equiv \rho\left(\sigma_\varepsilon^2\right) = \rho T_{t-1}\left(\sigma_{t-1}^2\right) = \rho T_{t-1}\mathbf{P}_{t-1} . \tag{11}$$

In the same spirit, we assume in our TVP model 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 (9) is given by

$$\mathbf{Q}_t = \rho T_{t-1}\mathbf{P}_{t-1}. \tag{12}$$

When the random coefficients model (9) contains only an intercept term and no regressors, (12) becomes (11). Hence, the proposed covariance specification (12) exactly nests the LLM.

Under specification (12), the extended Kalman filter greatly simplifies, and may be written in may be written in "Information" form (see Appendix I) as

$$\mathbf{b}_t = \mathbf{W}_t^{-1}\mathbf{z}_t, \tag{13}$$

$$\mathbf{P}_t = \sigma_\varepsilon^2 \mathbf{W}_t^{-1}, \tag{14}$$

where

$$\mathbf{z}_t = (1+\rho T_{t-1})^{-1}\mathbf{z}_{t-1} + \mathbf{x}_t' y_t, \tag{15}$$

$$\mathbf{W}_t = (1+\rho T_{t-1})^{-1}\mathbf{W}_{t-1} + \mathbf{x}_t'\mathbf{x}_t, \tag{16}$$

and T_t is updated as in (5).

If there is a diffuse prior about the coefficients at time 0, the initial covariance matrix $\mathbf{P}_0$ is infinite in all its eigenvalues, or equivalently, the "precision" or "information" matrix $\mathbf{P}_0^{-1}$ is all zeroes, whence

$$\mathbf{W}_0 = \mathbf{0}_{k\times k}. \tag{17}$$

For any choice of $\mathbf{b}_0$, (13) then implies

$$\mathbf{z}_0 = \mathbf{0}_{k\times 1}. \tag{18}$$

With such a diffuse prior, $\mathbf{W}_t$ is of rank t for $t \le k$, and hence $\mathbf{b}_t$ and $\mathbf{P}_t$ may only be computed for $t \ge k$. Note that in the fixed coefficient case $\rho = 0$, $\mathbf{z}_t$ then becomes $\mathbf{X}_t'\mathbf{y}_t$, $\mathbf{W}_t$ becomes $\mathbf{X}_t'\mathbf{X}_t$, and (13) becomes the familiar OLS formula $\mathbf{b}_t = (\mathbf{X}_t'\mathbf{X}_t)^{-1}\mathbf{X}_t'\mathbf{y}_t$.

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

$$y_t \mid \mathbf{y}_{t-1} \sim N\left(\mathbf{x}_t\mathbf{b}_{t-1},\ \sigma_\varepsilon^2 s_t^2\right), \tag{19}$$

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

$$\begin{aligned} \mathrm{L}\left(\rho, \sigma_\varepsilon^2 \mid \mathbf{y}_t\right) &= \sum_{t=k+1}^{n} \log p(y_t \mid \mathbf{y}_{t-1}) \\ &= -\frac{n-k}{2}\log(2\pi) - \frac{n-k}{2}\log\sigma_\varepsilon^2 - \sum_{t=k+1}^{n}\log s_t - \frac{1}{2\sigma_\varepsilon^2}\sum_{t=k+1}^{n} u_t^2, \end{aligned} \tag{20}$$

where the *scale-adjusted residuals*

$$u_t = e_t / s_t \tag{21}$$

equal the actual predictive errors,

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

adjusted by their time-varying scales s_t. Under the maintained assumptions, these adjusted residuals are homoskedastic with variance σ_ε^2, even though the predictive errors themselves are highly heteroskedastic. As in the LLM, the observation variance σ_ε^2 may be concentrated out of the log likelihood function in such a way that for any value of ρ, the likelihood is maximized with σ_ε^2 estimated in closed form by

$$s_\varepsilon^2 = \frac{1}{n-k} \sum_{t=k+1}^{n} u_t^2 \ . \tag{22}$$

A numerical maximization search is therefore required only over the single parameter ρ.

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

It may easily be shown, setting $\mathbf{R}_t = \mathbf{W}_t / T_t$, that the ALS filter (13) – (16) is equivalent to the variable-gain *Recursive Least Squares* (RLS) formula

$$\mathbf{b}_t = \mathbf{b}_{t-1} + \gamma_t \mathbf{R}_t^{-1} \mathbf{x}_t \left(y_t - \mathbf{x}_t \mathbf{b}_{t-1} \right), \tag{23}$$

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

$$\mathbf{P}_t = \gamma_t \sigma_\varepsilon^2 \mathbf{R}_t^{-1}. \tag{25}$$

See Appendix I for details. In the fixed coefficients case that is equivalent to OLS, the gain γ_t in (23) – (25) is 1/t. The previous AL literature (e.g. Sargent (1993, eq. (10)) or Evans and Honkapohja (2001, eq. (2.9)) commonly replaces this gain by a constant γ as in Cagan's original Adaptive Expectations formulation (1). However, this Constant Gain RLS (CGRLS) 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. Furthermore, it is not obvious how to initialize RLS, in the absence of prior information, with the required $\mathbf{R}_0 = \mathbf{W}_0 / T_0 = \mathbf{0}_{k \times k} / 0$.

Ljung (1992) and Sargent (1999, Ch. 8) assume, in place of (12), that

$$\mathbf{Q}_t = \frac{\gamma}{1-\gamma} \mathbf{P}_{t-1}, \tag{25}$$

with the result that (23) holds with a constant γ in place of $\gamma_t = 1/T_t$.[7] 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

[7] 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.

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 (12) or (25), that

$$\mathbf{Q}_t = \mathbf{Q} = \rho\sigma_\varepsilon^2\left(E\mathbf{x}_t'\mathbf{x}_t\right)^{-1}. \tag{26}$$

If the relevant expectation exists, this is equivalent in an expectational sense to (12), 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 were. Equation (12), 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.[8] Assumption (26) does nest the LLM, since then the required expectation is just a unit scalar. For $k > 1$, however, it only approximates ALS with gain $1/T_t$,. It also lacks the computational simplicity of ALS, since it requires the more general EKF described in Appendix I.

Stock and Watson (1996) calibrate the gain coefficient ρ in (26) (their λ^2) for several macroeconomic time series and relationships by minimizing the sum of squared forecasting errors. This will give results similar to ours, but by no means equivalent, even apart from the often subtle difference between our (12) and their (26). For one thing, the initial errors have much larger variance than the later errors simply because the coefficient vector is still highly uncertain. Equation (19) 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 (19), which is similar to the formula for the conditional distribution that would result from (26), that even asymptotically the squared forecasting errors e_t^2 are greater in expectation than σ_ε^2 by an amount that depends on ρ, so that minimizing their sum of squares will give a biased estimate of ρ 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 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.

Cogley and Sargent (2004) ambitiously estimate an autoregressive TVP model in which the coefficients take a random walk with unrestricted covariance matrix, subject to

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

reflecting boundaries that prevent nonstationary autoregressive roots. Their procedure is far more computation-intensive than ALS, because of both these complications. Binding inequality restrictions may easily be imposed on ALS estimates after unrestricted estimation, if desired, however, simply by using the standard conditional multivariate normal formula given, e.g., by Harvey (1989, p. 165).[9]

The Kalman *Filter* for the ALS model, discussed 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 regression coefficients given *both prior and subsequent experience*, the Kalman *Smoother* (also known as the 2-sided filter) 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 matrix covariance matrix. The pertinent equations for both the general TVP case and the special ALS case are given in Appendix I.

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.

IV. Hypothesis testing

[Section under Construction]

Because the null hypothesis of no parameter change, i.e. $\rho = 0$, is on the boundary of the permissible parameter space $\rho \geq 0$, the usual regularity conditions for the χ^2 limiting distribution of the Lagrange Multiplier (LM) and Likelihood Ratio (LR) statistics are not met (Moran 1971a, 1971b). Nevertheless, Tanaka (1983) has shown that the LM statistic is still useful and informative in the LLM case, provided the critical values are appropriately adjusted.

---

[9] Two candidates for such an adjustment would be 1973 and 1980 in the inflation equation estimated below, when the sum of the AR coefficients briefly slightly exceeds unity, even without a correction for AR bias as discussed below. With fixed coefficients there is a case for overriding such estimates, since such an explosive process would have long-since blown up and would never be 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. Getting out of such a situation would presumably call for an even larger Taylor Rule coefficient on past inflation than is normally called for.

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).

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.[10] Preliminary simulations 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 ρ is known. In practice, ρ is estimated, but since its ML estimate is consistent, this should not be a problem in large samples.

Global linear restrictions on coefficients, i.e. for all t, are a little more difficult. With no loss of generality, we may assume that these are zero restrictions on q of the coefficients, since the regressors can always be redefined in such a way that this is true. It is then easy enough to run unrestricted (UR) ALS with all k potential regressors and restricted (R) ALS using only the k-q non-restricted regressors. Unfortunately, the likelihoods of these two regressions are not directly comparable as in the fixed coefficients case, since in the predictive error decomposition framework of (20), they actually contain different numbers of terms.

Nevertheless, in the fixed coefficient case $\rho = 0$, it turns out that the sum of squared adjusted residuals

$$\text{SSU} = \sum_{t=k+1}^{n} u_t^2$$

in (22) is exactly equal to the traditional sum of squared errors SSE computed from OLS, despite the fact that OLS is adding n terms, while ALS is adding only n-k. (The author has found this to be true for numerous examples, but does not have an analytical proof.) In the fixed coefficients case, the likelihood ratio LR can be expressed in terms of the restricted and unrestricted sum of squared errors SSE_R and SSE_UR as

$$\text{LR} = n \log(\text{SSE}_\text{R}/\text{SSE}_\text{UR}).$$

A test based on the statistic

$$\text{VR} = n \log(\text{SSU}_\text{R}/\text{SSU}_\text{UR})$$

will therefore become the familiar LR test in the fixed coefficient case. However, since this statistic is not based directly on the likelihoods in (22), it could be confusing to call this an "LR" statistic. Since it is nevertheless based on the sums of squared errors that

[10] 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.

are used to estimate the variances, we instead call this the "Variance Ratio" (VR) statistic.

For $\rho = 0$, the VR statistic will have the asymptotic $\chi^2_{(q)}$ distribution of the familiar LR statistic. For $\rho > 0$, however, the critical values may be considerably higher, since in effect q(n/T) independent restrictions are being imposed. The author therefore plans to tabulate, by Monte Carlo means, critical values for VR for various values of n, k, q, and ρ in a future version of this paper. Some illustrative tests are employed below.

V. GARCH-ALS

It is often the case that the squared ALS adjusted residuals (and therefore the squared observation errors and/or transition errors) are positively serially correlated. Such volatility clustering can easily be incorporated into the present ALS model as a GARCH(1,1) effect. 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 TVP models by Kim and Nelson (2004) and Cogley and Sargent (2004).

Although in ALS the observation errors ε_t and transition errors η_t are not directly observed or even consistently estimated, the adjusted residuals u_t defined by (21) are observed and hence may be used to drive the GARCH process governing ε_t and thence, indirectly, η_{t+1}. Hence, we modify our basic model (9) by postulating that

$$y_t = \mathbf{x}_t \boldsymbol{\beta}_t + \varepsilon_t, \quad \varepsilon_t \sim NID(0, h_t^2), \tag{27}$$

where

$$h_t^2 = \omega + \phi h_{t-1}^2 + \theta u_{t-1}^2, \; \omega > 0, \; \phi \geq 0, \; \theta \geq 0, \; \phi + \theta \leq 1.^{11} \tag{28}$$

This specification implies that

$$y_t^* = \mathbf{x}_t^* \boldsymbol{\beta}_t + \varepsilon_t^* \tag{29}$$

is a homoskedastic ALS model with the original coefficients, where

$$\begin{aligned} y_t^* &= y_t / h_t, \\ \mathbf{x}_t^* &= \mathbf{x}_t / h_t, \\ \varepsilon_t^* &= \varepsilon_t / h_t \sim N(0, \sigma_*^2), \end{aligned} \tag{30}$$

with $\sigma_*^2 \equiv 1$.[12] Given ρ, we may therefore estimate $\boldsymbol{\beta}_t$ and hence

$$u_t^* = u_t / h_t$$

by running ALS on (28). The heteroskedastic u_t to use in (**Error! Bookmark not defined.**) may then be reconstructed with

$$u_t = h_t u_t^* .$$

---

[11] The customary restriction $\theta + \phi \leq 1$ is sufficient, though not quite necessary, to guarantee strict stationarity of the process. In McCulloch (1985), the author imposed the "IGARCH" restriction $\theta + \phi = 1$, and set $\omega = 0$ in the erroneous belief that this was necessary for strict stationarity in the IGARCH case.

[12] It is important to remember that the unitary first element of $\mathbf{x}_t$ when an intercept is included in the model requires that the first element of x_t^* be $1/h_t$.

The log likelihood, conditional on the initial value of h_t, is then

$$\begin{aligned} \mathrm{L}\left(\rho, \sigma_*^2 \mid \mathbf{y}_t, h_1\right) &= \sum_{t=k+1}^{n} \log p(y_t \mid \mathbf{y}_{t-1}, \rho, \sigma_*^2, h_1) \\ &= -\frac{n-k}{2}\log(2\pi) - \frac{n-k}{2}\log\sigma_*^2 - \sum_{t=k+1}^{n}\log h_t s_t^* - \frac{1}{2\sigma_*^2}\sum_{t=k+1}^{n}\left(u_t^*\right)^2 , \\ &= \mathrm{L}^* - \sum_{t=k+1}^{n}\log h_t , \end{aligned}$$

where L* is the maximized log likelihood returned by ALS on the transformed problem (28). Since the h_t's depend on the u_t's, and vice-versa, it is necessary to compute them both iteratively for any choice of GARCH parameters, until the likelihood (or the pseudo-likelihood computed under whatever ad hoc initialization is employed) stabilizes to the desired precision.

In order to estimate such a GARCH model by true ML, one must first compute the unconditional density f(h) implied by (28). This may be found by performing the appropriate convolution integral iteratively to convergence. The unconditional likelihood is then the expectation, over f(h_1), of the likelihood conditional on h_1. However, this would be a lot of computation for very little payoff, since the GARCH effects are of only secondary interest, and since the initial value of the variance is even less critical for most questions of interest. It is therefore customary in GARCH estimation to replace true ML with conditional ML, using any of a variety of ad hoc initializations for h_1.

Many of the initialization procedures in use have the undesirable property that they leave the initial observation(s) with a different variance than the subsequent ones in the supposedly homoskedastic case $\theta = 0$.[13] The author's GAUSS program ALSGARCH,[14] which implements ALS with GARCH effects, gets around this and other problems by backcasting h_1 from the u_t's, starting with u_n, and using the current values of the GARCH parameters. When $\theta = 0$ this automatically yields $h_1^2 = h_t^2 = \omega$, as would true ML, while when $\theta > 0$, it is very similar in effect to the much more laborious procedure of finding the value of h_1^2 that maximizes the conditional log likelihood, and thus treating it like an additional hyperparameter to be estimated by ML.

If the GARCH parameters were estimated by true ML as described above, the estimate

$$s_*^2 = \frac{1}{n-k}\sum_{t=k+1}^{n}\left(u_t^*\right)^2$$

of σ_*^2 from ALS on (29) would be identically unity at the GARCH parameters that maximize L* with $\sigma_*^2 = 1$, so that estimation of both ω and σ_*^2 would be redundant. However, when true ML is replaced with any pseudo-ML procedure based on an ad hoc

[13] When $\theta = 0$, f(h) collapses to a unit mass point on ω, so that $h_t = \sqrt{\omega}$ for all t, and ϕ becomes irrelevant.
[14] A preliminary "beta" version of ALSGARCH is informally available at <http://www.econ.ohio-state.edu/jhm/programs/ALSGARCH>.

initialization, the computed s_*^2 on which L* is based need not be exactly unity. Nevertheless, it should be close. It is therefore prudent to check that this is in fact the case.

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 ALS 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 1 shows the U.S. CPI-U (not seasonally adjusted) from 1913.1 to 2005.4. In order to reduce rounding error in computing inflation rates, 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. It may be seen that inflation has been persistent, but that the nature of this persistence has changed considerably over time.

Figure 1
CPI-U, CPI-UX

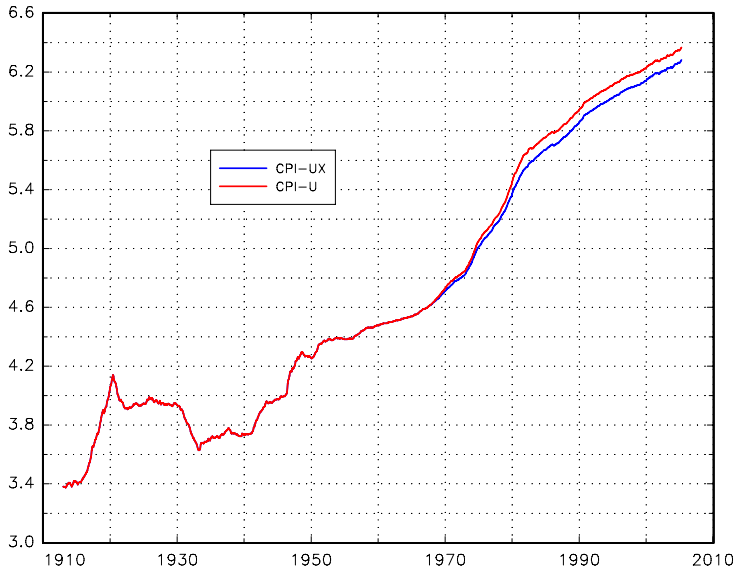
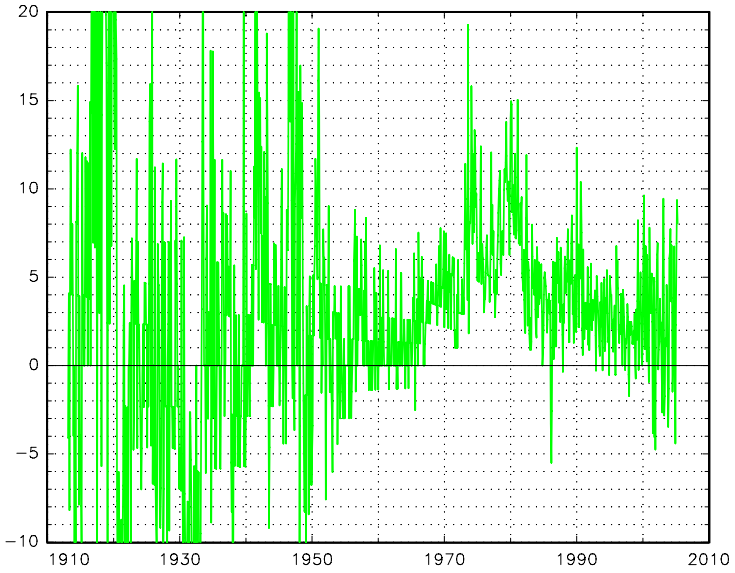


Figure 2 shows the annualized percentage logarithmic CPI-UX inflation rate, computed as $\pi_t = 1200(\ln(P_t / P_{t-1}))$, for 1913.2-2005.4. It is obvious from this diagram that measured inflation volatility was much higher prior to the computational improvements of 1953 than after.

Figure 2
CPI-UX inflation (annualized, percent)
Not seasonally adjusted



A restricted AR model was fit to this data by ALS, with up to $L = 24$ monthly lags and 12 seasonal intercepts, or conceptually,

$$\pi_t = \sum_{j=1}^{12} \alpha_{tj} d_{tj} + \sum_{l=1}^{L} \delta_{tl} \pi_{t-l} + \varepsilon_t \ .$$

The d_{tj} are seasonal dummies, with $d_{t1} = 1$ if month t is a January, etc. In order to reduce the number of free AR parameters to at most $p = 5$, however, the AR coefficients were constrained to be piecewise linear and to decay to 0 at lag L+1, by means of the following specification:

$$\pi_t = \sum_{j=1}^{12} \alpha_{tj} d_{tj} + \sum_{j=1}^{p} \gamma_{tj} \overline{\pi}_{tj} + \varepsilon_t \ ,$$

where the $\overline{\pi}_{tj}$ are weighted averages of past inflation with linearly declining weights, as follows: [15]

[15] Such linearly declining weighted averages were in fact first proposed by Irving Fisher (1930: 419-20), who dubbed them "Distributed Lags".

$$\overline{\pi}_{1t} = INF1_t = \pi_{t-1}$$

$$\overline{\pi}_{2t} = INF3_t = \sum_{l=1}^{3}(4-l)\pi_{t-l}/6$$

$$\overline{\pi}_{3t} = INF6_t = \sum_{l=1}^{6}(7-l)\pi_{t-l}/21$$

$$\overline{\pi}_{4t} = INF12_t = \sum_{l=1}^{12}(13-l)\pi_{t-l}/78$$

$$\overline{\pi}_{5t} = INF24_t = \sum_{l=1}^{24}(25-l)\pi_{t-l}/3600$$

These may in general be expressed as

$$\overline{\pi}_{jt} = \sum_{l=1}^{L} w_{lj}\pi_{t-l}$$

The net lag coefficients δ_{tl} may then be recovered from the γ_{tj} by

$$\delta_{tl} = \sum_{j=1}^{p} w_{lj}\gamma_{tj}$$

Significant volatility clustering was modeled as GARCH(1,1), as discussed below.

It was found that the hypothesis that the parameter γ_{t5} governing the net coefficients on lags 13-24 was zero could not be rejected, either locally or globally, as discussed at greater length below. Hence the results discussed here use only $L = 12$ lags, governed by $p = 4$ parameters, with net lags decaying to 0 at lag $L + 1 = 13$. This gives a total of $k = 12 + p = 16$ time-varying parameters, with $n = 1083$ observations.

The ALS-GARCH ML estimates of ρ and therefore T and γ with $p = 4$ are:

$\hat{\rho} = 0.00006155$

$\hat{T} = 128.4$ mo. $= 10.7$ yrs.

$\hat{\gamma} = 1/\hat{T} = 0.007788$

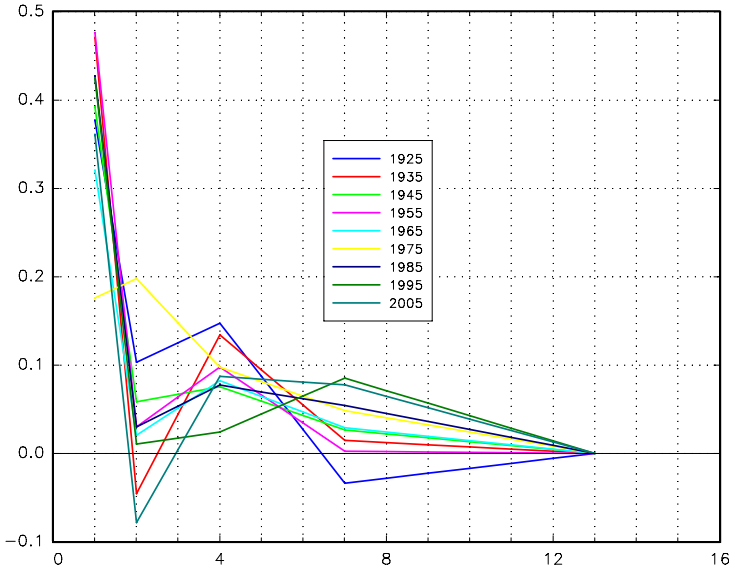
LR ($\rho = 0$) = 62.98

Since the 5% critical value for the LR statistic is approximately 2.3, the hypothesis of time-invariant parameters is overwhelmingly rejected. Because of the considerable number of computations required, ML estimation of the model, complete with GARCH effects and smoother estimates, takes about 8 minutes on a PC (or 75 seconds if GARCH is suppressed).

Figure 3 shows the filter estimates of the piecewise linear net lag coefficients δ_{jl} implied by the time-varying regression parameters γ_{tj}, in 10-year intervals. The illustrative dates are January of the year indicated. The lag structures are qualitatively remarkably stable over time, despite their quantitative variation. The first lag is almost always near 0.4, the second lag usually dips to 0.1 or smaller, and then there is a hump in lags 3-6. In lags 7-13, the coefficients are constrained to decline linearly toward 0 at lag 13. In fact, prior to 1985, the coefficients are already nearly 0 or even negative by lag 7.

The restricted AR(12) coefficients in Figure 3 may be described as a *first-degree spline*, with *knot points*, i.e. discontinuities in the first derivative, at lags 1, 2, 4, 7, and 13.

Figure 3
Piecewise linear net lag coefficients δ_{jl}, selected Januaries.
(filter estimates)



Lag *l* (months)

Figure 4 depicts the net lag coefficients δ_{jl} at the key lags $l = 1, 2, 4$, and 7, for all $t \geq k$. The implicit coefficient at lag 13 is always zero and hence not shown. Since the lags illustrated are the spline knots, the coefficients for the intervening lags are just linear interpolations of the two adjoining knot values, and hence have been omitted to avoid unnecessary clutter. It may be seen that the first lag coefficient consistently lies between 0.3 and 0.5, except for dropping to near 0.2 during the 1960s and 70s. However, there was a compensating increase in the strength of lag 2 (and therefore 3) during this period. The 4th lag (and therefore adjacent lags) is usually positive, except for the 1990s when it was near zero. The 7th lag, which directly governs all lags out to 12 with $p = 4$, is near 0 until about 1960, and thereafter becomes positive. It picks up the temporary weakness in the 4th lag during the 1990s. All coefficient point estimates are highly erratic in the first few years of the sample because of the small effective sample size, and accordingly have very high standard errors (not plotted to avoid clutter). This erratic "startup" behavior is normal, particularly for $t < T$, and should be no cause for concern.

Figure 4
Net lag coefficients δ_{jl} at knot points $l = 1, 2, 4, 7$
(filter estimates)

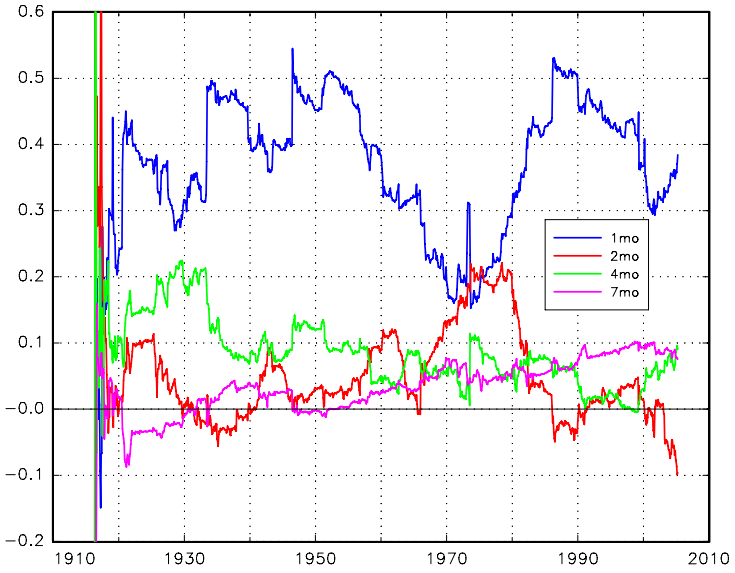


Figure 5 depicts the filter estimate of γ_{t4}, the coefficient on $INF12_t$, with a ± 1 s.e. band as computed from the filter covariance matrices. Except for vertical scale, this point estimate is the same line as the 7-month lag coefficient in Figure 4, since with $p = 4$, this coefficient governs all lags from 7 through 12. It may be seen that this coefficient is never locally significant with either sign at the 5% level before 1986, but that it has been significantly positive since 1990. Since the filter is conditioned on only past experience, forecasters might not have realized that lags 7-12 were important until after these dates.

Figure 5
Filter estimate of γ_{t4}, the coefficient on $INF12_t$, ± 1 s.e.

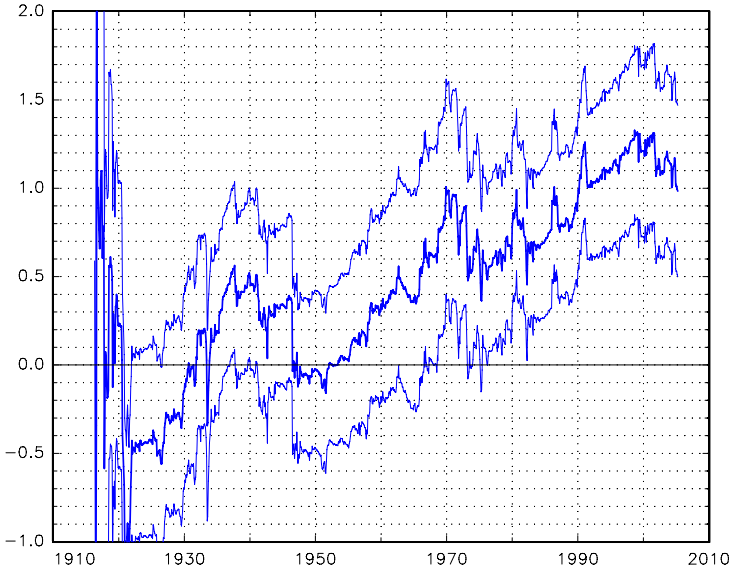
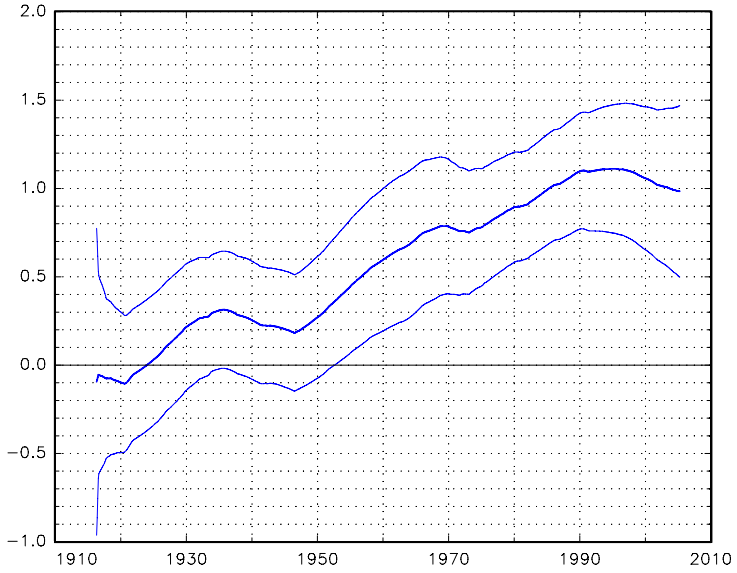


Figure 6 shows the corresponding smoother, or "two-sided filter," estimate of γ_{t4}, conditioned on the entire data set, with $p = 4$. The smoother and filter estimates and their standard errors necessarily coincide exactly at the end of the sample. In the interior of the sample, however, the smoother standard errors are about 30% smaller than the filter standard errors, since they are based on almost twice as much data, looking both forwards and backwards. The filter estimates are locally significantly positive at the 5% level after about 1970, so that with the benefit of hindsight we may say that γ_{t4} and therefore δ_{t7} - δ_{t12} have in fact been non-zero and positive since that time.

Figure 6
Smoother estimate of γ_{t4}, the coefficient on $INF12_t$, ± 1 s.e.



The Variance Ratio statistic for the hypothesis that γ_{t4} is globally zero, i.e. 0 for all t, is 19.71. The 5% Monte Carlo critical value for $q = 1$ restriction with $k = 16$ parameters in the unrestricted model, $n = 1083$ observations, and $T = 128.4$ is 9.08, using $r = 99$ Monte Carlo replications, while the 1% critical value is 13.32. We may therefore reject the hypothesis that lags 7-12 have always been zero with a high degree of confidence, when $INF24_t$ has already been excluded from the regression.

When $INF24_t$ was initially included, so that $p = 5$, $k = 17$, and $L = 24$, the filter and smoother estimates of its coefficient γ_{t5} (not illustrated) was rarely more than even 1 standard error from 0, so that 0 could never be locally rejected at the 5% level using either estimate. The Variance Ratio for the hypothesis that it was always zero is 2.085, which has a Monte Carlo p-value of 85% using $r = 99$ replications. We therefore may not reject the hypothesis that lags 13-24 are always zero, and hence, under the maintained piecewise linear assumption, that lags 7-12 always decay toward zero when extrapolated to lag 13. In the interest of parsimony, $INF24_t$ was therefore excluded from the results presented here, including the above tests on γ_{t4}. However, since, having excluded $INF24_t$, the coefficient on $INF12_t$ is globally significant, the shorter and therefore *a priori* even

more plausible average lag variables $INF1_t$ – $INF6_t$ are included, regardless of their local or global significance.

Figure 7 shows the filter estimates of the average intercept

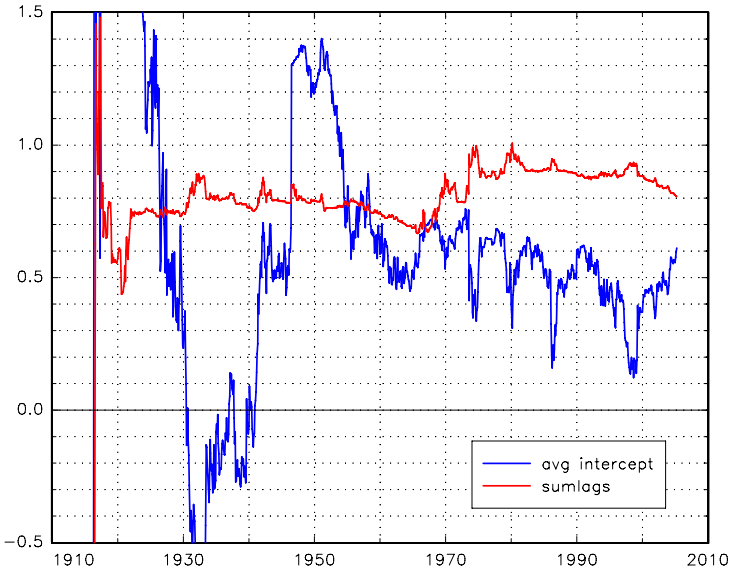
$$\overline{\alpha}_t = \sum_{j=1}^{12} \alpha_j / 12 ,$$

along with the sum of the $L = 12$ non-zero net lag coefficients,

$$\Delta_t = \sum_{l=1}^{L} \delta_{tl} .$$

The latter is a measure of the short-run persistence of inflation. Except for the initial warm-up period, Δ_t is always 0.7 or more. It was generally about 0.8 until the 1960s, when it fell to around 0.7. After 1970, it quickly rose to 0.9, where it remained until 2000, contrary to the popular belief that inflation has been less persistent in recent decades than earlier. During the soaring inflation years 1974 and 1980, Δ_t briefly even touched unity. As in a fixed-coefficient autoregression, the estimate of Δ_t will be biased downwards, so that the true persistence was generally even greater than suggested by Figure 7. The present paper makes no attempt to correct for this effect, however.[16]

Figure 7
Average intercept $\overline{\alpha}_t$, sum of lag coefficients Δ_t
(filter estimates)



It is interesting to note that while Sargent (1971) was able to claim that the sum of the lag coefficients in a U.S. inflation autoregression clearly pointed to stationarity, this

[16] Any correction of this bias or Dickey-Fuller-type test for $\Delta_t = 1$ should be based on the effective sample size T for the filter, rather than the true sample size n, however, since the filter coefficients are in effect estimated anew every T periods. For the smoother estimates, the sample size for this purposes is effectively $2T$-1.

property of the data was quickly being reversed, even as he wrote. It should be noted, however, that in ALS, as in any TVP model of the type (9), all the coefficients, including the intercept, are assumed to be taking random walks. Unless $\rho = 0$, the inflation process therefore exhibits long-run non-stationarity, by assumption. If Δ_t also turns out to be unity, that would only add short-run non-stationarity on top of the assumed long-run non-stationarity. This would make the process in fact I(2) in the long run.

Figure 8 shows the long-run inflation forecasts

$$\hat{\pi}_t^{LR} = \overline{\alpha}_t / (1 - \Delta_t)$$

implied by the data in Figure 7. It may be seen that the long-run forecast undergoes substantial change, dipping to near 2% in the mid-1960s, rising to 6% or higher during 1973-82, and then slowly falling back to around 3% since 2000. During 1974 and 1980, when the filter estimate of Δ_t was briefly touching or even exceeding 1 while the average intercept was still clearly positive, inflation seemed to be drifting upwards without bound, and $\hat{\pi}_t^{LR}$ is plotted as + infinity. The terminal value of the long-run forecast, in 2005.4, is 3.15%.

Figure 8
Long-run and short-run inflation forecasts
(filter estimates)

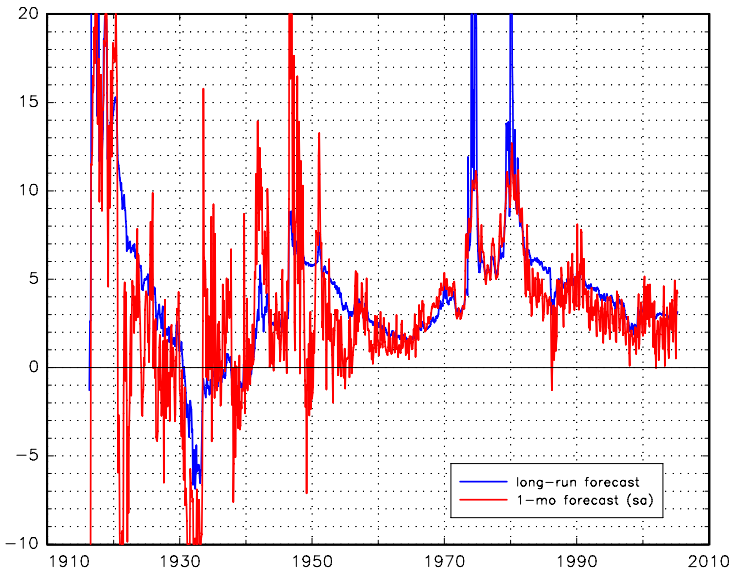


Figure 8 also plots the seasonally-adjusted 1-month-ahead forecasts

$$\hat{\pi}_t^{sa} = \overline{\alpha}_t + \sum_{j=1}^{4} \gamma_{tj} \overline{\pi}_{tj} \ .$$

The short-run inflation forecast likewise shows substantial time-variation, and often differs from the long-run forecast. Although it exceeded the long-run forecast in the late 1930s and the late 1960s, it falls short of it throughout the 1940s, 40s, and early 60s, the mid-1980s, the 1990s, and for the past few years. On the heels of a 9.36% (n.s.a.)

annualized actual inflation rate for 2005.3, however, the final short-run forecast for 2005.4 was up to 4.21% (s.a.). The actual rate for 2005.4 was 8.06%, n.s.a.

In the simplistic Cagan Adaptive Expectations model of (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 8 that with the present much richer ALS model, there are often substantial differences between the two. The net AR coefficients δ_{tj} can straightforwardly be used to construct one-month forecasts to any horizon, and these could easily be averaged to long-run forecasts of any desired horizon, to generate an entire term structure of inflation forecasts at each point in time.

Figure 9 shows the filter-estimated seasonal adjustments,

$$seas_t = \alpha_{tj} - \overline{\alpha}_t .$$

The estimated seasonals are clearly much larger before c. 1960 than before. While there really could have been much larger seasonals in the more agricultural earlier economy, this phenomenon may in part just be due to worse data-collection before 1953, or even to larger GARCH-induced estimation error in the earlier period. The seasonals were quite small during the 1970s and 80s, but have grown considerably since 1990. At the end of the sample, in 2005.4, the largest seasonal is for January, at +2.55% per annum (s.e. = 0.70), while the smallest is for December, at -2.60 (0.67), with November a close second at -2.15 (0.65).

Figure 9
Seasonals
(%/yr, filter estimates)

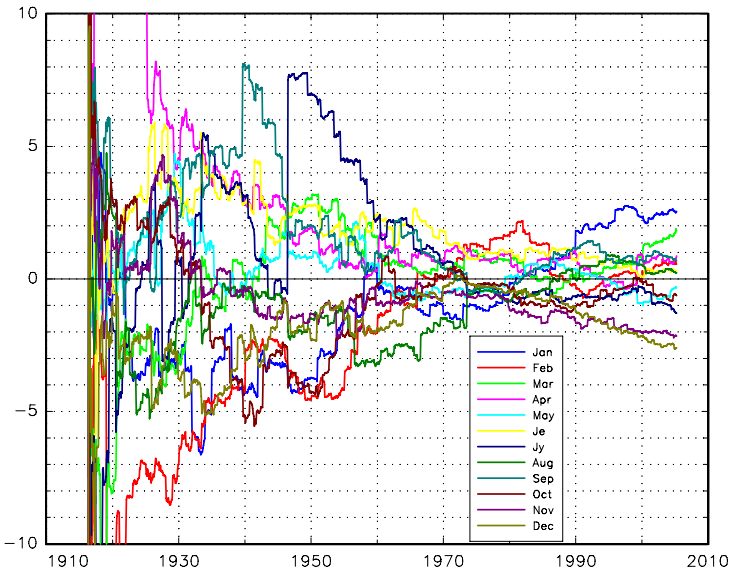


Figure 10 shows the χ^2 test statistic for the hypothesis that all 12 of the filter seasonals of Figure 9 are locally 0. Since this hypothesis imposes 11 restrictions, the distribution has 11 degrees of freedom, and the 5% critical value is 19.68. Although the

test statistic dips to insignificance during the 1970s and 80s, it is on-and-off significant before 1970, and clearly significant after 1990. Although forecasters might legitimately have dispensed with seasonal adjustment between 1970 and 1990, it cannot be ignored in today's economy.

Figure 10
χ^2 test statistic for local significance of filter seasonals
(11 DOF; 5% critical value = 19.68)

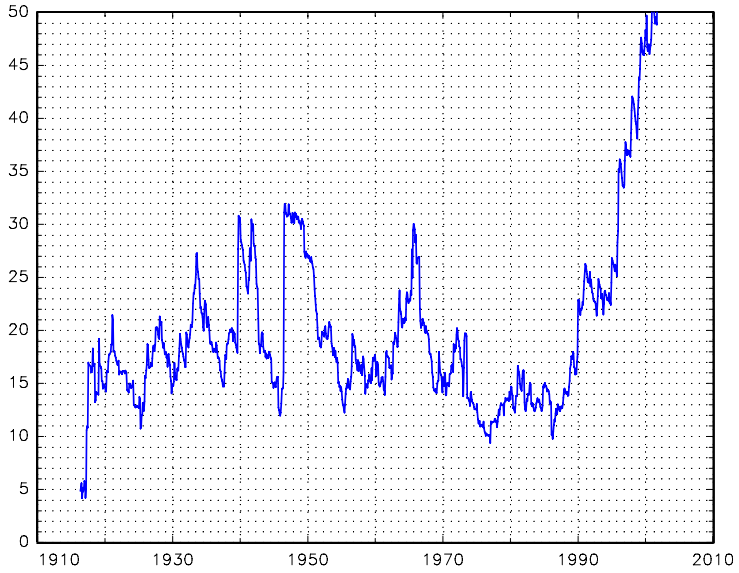
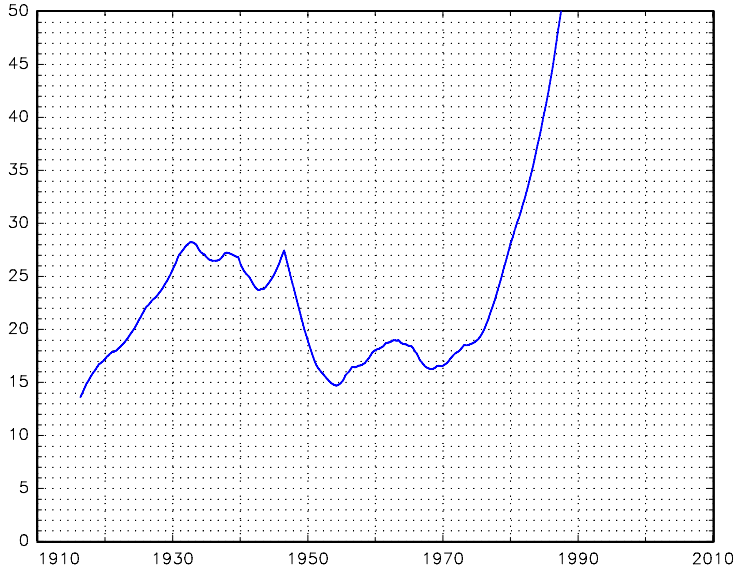


Figure 11 shows the χ^2 test statistic for the hypothesis that all 12 of the smoother seasonals (not shown) are locally 0. In hindsight, there was therefore unambiguous seasonality throughout the 1930s and 40s and again since about 1976. However, the temporary abatement of seasonality during the 1960s and early 70s may in fact have begun as early as 1950.

Figure 11
χ^2 test statistic for local significance of smoother seasonals
(11 DOF; 5% critical value = 19.68)



The Variance Ratio statistic for the global hypothesis of no seasonality ever is 206.11. The Monte Carlo 1% critical value of this statistic, using r = 99 replications with n = 1083, k = 16, q = 11, and T = 128.4, is 65.60, so we may overwhelmingly reject the total absence of seasonality.

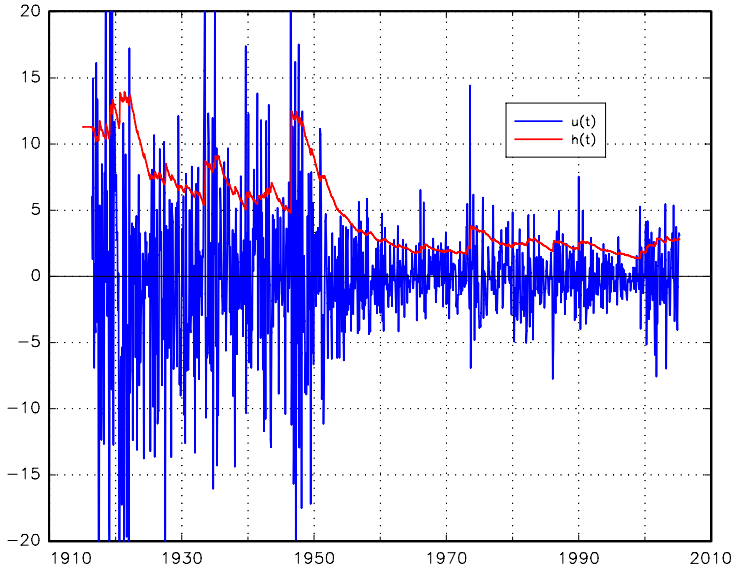
Figure 12 shows the restored scale-adjusted forecast errors u_t (noisy series), along with the GARCH(1,1)-estimated conditional standard deviation h_t. The ML estimates of the GARCH model are

$$h_t^2 = .03949 + .9566h_{t-1}^2 + .04124u_{t-1}^2,$$

LR (no GARCH) = 722.73.

Although the LR statistic has a non-standard distribution and this paper makes no attempt to quantify critical values for it, there is clearly overwhelming evidence against homoskedasticity. Since the coefficients on the lagged variance and lagged squared error sum virtually to unity, this is a highly persistent, nearly IGARCH, volatility series.

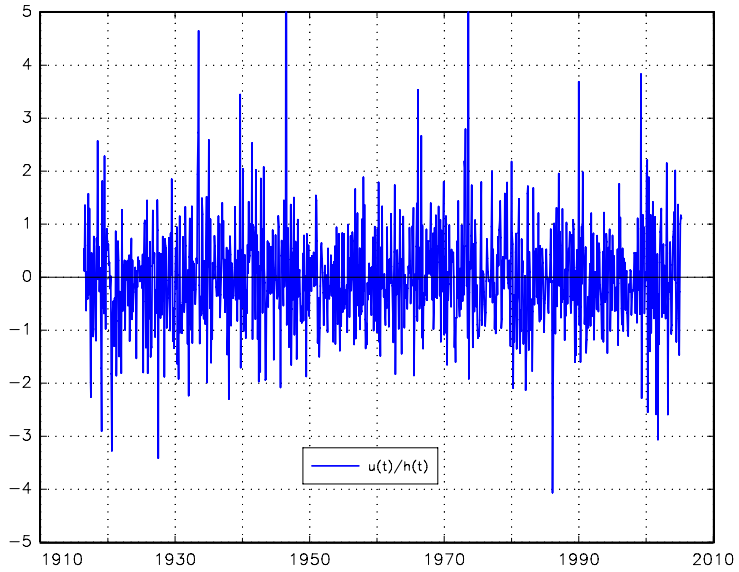
Figure 12
Variance-equalized forecast errors u_t,
with GARCH(1,1) standard deviations h_t



If these GARCH effects were ignored, the point estimates of the coefficients, and therefore the long-run and short-run inflation forecasts, would not be greatly altered except in a few episodes. However, the estimated standard errors would then be biased in inverse proportion to h_t, and hence far too small in the early period, and far too large in the later period. Our specification tests for the significance of INF24, INF12, and seasonality would then be grossly invalid. With the GARCH specification, the estimated coefficient standard errors, as for example in Figures 5 and 6 above, are more nearly constant over time.

Figure 13 depicts the GARCH-standardized scale-adjusted forecast errors, $u^*_t = u_t/h_t$. If the model is correctly specified (and ρ equal to its true value), these are NID(0, 1). The striking evidence of volatility clustering in Figure 12 is essentially gone, so that a higher-order GARCH process would have little, if any payoff. There is also no strong visual evidence of unexplained serial correlation. Their estimated variance, $s^2_* = 1.020$, is close to unity, as required. There is, however, strong evidence of non-normality in these residuals, with a raw Kurtosis of 21.63 and a Skewness of 1.79. Future work should therefore explore adapting ALS to non-Gaussian errors.

Figure 13
GARCH-standardized, scale-adjusted forecast errors, u_t/h_t



The present univariate model of inflation is merely intended to provide an illustration of ALS, rather than a definitive forecasting model for inflation. It could easily be extended to incorporate less conspicuous variables such as unemployment, or even the stance of monetary policy, as measured by the money stock relative to money demand variables, or interest rates relative to their natural real level. Before mining the proverbial wheelbarrow of additional variables, however, it is valuable to first ascertain just how parsimonious a univariate model of inflation can be constructed, as above.

VII. Potential future Applications.

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 over the desired horizon are first simulated using a ALS filter, and then a time-varying response function is estimated using the simulated forecasts, by means of 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 plan to use the "Moderate Quantity Theory" expectations-augmented price-adjustment 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^D_t, 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 using the ALS filter, while the latter is constructed from past *and* future experience by means of the ALS smoother.

Appendix I

The Local Level Model (2) implies

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

so that the distribution of μ_1 given y_1 and no other information may be written

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

where

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

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

$$\mu_t \mid \mathbf{y}_{t-1} \sim N(m_{t-1}, \sigma_{t-1}^2 + \sigma_\eta^2) = N(m_{t-1}, \sigma_{t-1}^2 + \rho\sigma_\varepsilon^2). \tag{31}$$

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

$$\begin{aligned} p(\mu_t \mid \mathbf{y}_t) &= p(y_t \mid \mu_t, \mathbf{y}_{t-1})\, p(\mu_t \mid \mathbf{y}_{t-1}) / (const.) \\ &= p(y_t \mid \mu_t)\, p(\mu_t \mid \mathbf{y}_{t-1}) / (const.) \\ &= \exp\left(-\frac{1}{2} \frac{(y_t - \mu_t)^2}{\sigma_\varepsilon^2} \right) \exp\left(-\frac{1}{2} \frac{(\mu_t - m_{t-1})^2}{\sigma_{t-1}^2 + \rho\sigma_\varepsilon^2} \right) / (const.) \\ &= \exp\left(-\frac{1}{2} \frac{(\mu_t - m_t)^2}{\sigma_t^2} \right) / (const.), \end{aligned} \tag{32}$$

so that (3) is valid with

$$m_t = \frac{\sigma_t^2}{\sigma_\varepsilon^2} y_t + \frac{\sigma_t^2}{\sigma_{t-1}^2 + \rho\sigma_\varepsilon^2} m_{t-1} \tag{33}$$

and

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

Defining $T_t = \sigma_\varepsilon^2 / \sigma_t^2$, (33) becomes (4) and (34) becomes (5), which may be initialized either with $T_0 = 0$ or $T_1 = 1$.

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,

$$\boldsymbol{\beta}_{t-1} \mid \mathbf{y}_{t-1} \sim N(\mathbf{b}_{t-1}, \mathbf{P}_{t-1}) \tag{35}$$

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

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

where

$$\mathbf{b}_t = \mathbf{b}_{t-1} + f_t^{-1}(\mathbf{P}_{t-1} + \mathbf{Q}_t)\mathbf{x}_t'(y_t - \mathbf{x}_t\mathbf{b}_{t-1}), \tag{36}$$

$$\mathbf{P}_t = (\mathbf{P}_{t-1} + \mathbf{Q}_t)(\mathbf{I}_{k\times k} - f_t^{-1}\mathbf{x}_t'\mathbf{x}_t(\mathbf{P}_{t-1} + \mathbf{Q}_t)), \tag{37}$$

$$f_t = \mathbf{x}_t'(\mathbf{P}_{t-1} + \mathbf{Q}_t)\mathbf{x}_t' + \sigma_\varepsilon^2 .^{17} \tag{38}$$

The textbook EKF equations (36) and (37) above may be rearranged to eliminate f_t and to look more like RLS, as follows: Post-multiply (37) by $\mathbf{x}_t'$ and combine with (38) to obtain

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

so that (36) becomes

$$\mathbf{b}_t = \mathbf{b}_{t-1} + (1/\sigma_\varepsilon^2)\mathbf{P}_t\mathbf{x}_t'(y_t - \mathbf{x}_t\mathbf{b}_{t-1}), \tag{39}$$

and (37) 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

$$\mathbf{P}_t^{-1} = (\mathbf{P}_{t-1} + \mathbf{Q}_t)^{-1} + (1/\sigma_\varepsilon^2)\mathbf{x}_t'\mathbf{x}_t . \tag{40}$$

The rearranged filter (39), (40) may be placed in the even more convenient "Information" form, mentioned but not developed by Harvey (1989, p. 108), in terms of the scaled Information Matrix $\mathbf{W}_t = \sigma_\varepsilon^2\mathbf{P}_t^{-1}$, the scaled transition covariance matrix $\mathbf{V}_t = (1/\sigma_\varepsilon^2)\mathbf{Q}_t$, and what might be called the "cumulative evidence" vector $\mathbf{z}_t = \mathbf{W}_t\mathbf{b}_t$, as follows:

$$\mathbf{z}_t = (\mathbf{I} + \mathbf{W}_{t-1}\mathbf{V}_t)^{-1}\mathbf{z}_{t-1} + \mathbf{x}_t'y_t , \tag{41}$$

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

whence $\mathbf{b}_t$ and $\mathbf{P}_t$ may be recovered by (13) and (14).[18] Cp. also Bullard (1992).

---

[17] 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 II below for details.

[18] The observation error variance σ_ε^2 cannot be estimated until after the filter has been run, so it must in any event be factored out of $\mathbf{P}_t$ and $\mathbf{Q}_t$ in order to run the filter.

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

$$\mathbf{W}_0 = \mathbf{0}_{k\times k}\,. \tag{17}$$

For any choice of $\mathbf{b}_0$, $\mathbf{z}_0 = \mathbf{W}_0\mathbf{b}_0$ then implies

$$\mathbf{z}_0 = \mathbf{0}_{k\times 1}\,. \tag{18}$$

Nothing could be simpler than nothing itself! It is not so obvious how to impose a diffuse prior on either (36) – (38) or (39) - 40, however.

With the diffuse prior, $\mathbf{W}_t$ is of rank t for $t \le k$, and hence $\mathbf{b}_t$ and $\mathbf{P}_t$ may not be computed by (13) and (14) until $t \ge k$. Note that in the fixed coefficient case $\mathbf{Q}_t = \mathbf{V}_t = \mathbf{0}_{k\times k}$, $\mathbf{z}_t$ becomes $\mathbf{X}_t'\mathbf{y}_t$, $\mathbf{W}_t$ becomes $\mathbf{X}_t'\mathbf{X}_t$, and (13) then becomes the familiar OLS formula.

In the LLM special case (2) of the general TVP, $\mathbf{V}_t = (\rho)$, $\mathbf{x}_t'\mathbf{x}_t = (1)$, $\mathbf{x}_t' y_t = (y_t)$, (42) becomes (5) with $\mathbf{W}_t = (T_t)$, and (39) becomes (4) with $\gamma_t = 1/T_t$.

Under the ALS specification of $\mathbf{Q}_t$ in (12), $\mathbf{W}_{t-1}\mathbf{V}_t = \rho T_{t-1}\mathbf{I}$, and the Information filter (41), (42) becomes the ALS filter (15), (16). It is obvious from (41) and (42) that if $\mathbf{Q}_t$ is specified as any scalar times $\mathbf{P}_{t-1}$, so that $\mathbf{V}_t$ is the same scalar times $\mathbf{W}_{t-1}^{-1}$, the matrix inversions will no longer be required and the calculations enormously simplified. As originally noted by Ljung (1992), the filter in this case simply becomes RLS, with gain determined by the scalar chosen. However, if the scalar chosen is time-invariant, as specified by Ljung, the RLS gain will inappropriately also be time-invariant.

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, as if to obtain estimates $\mathbf{b}_t^*$ of $\boldsymbol{\beta}_t$ conditional on y_t, ... y_n and no other information, with variances $\mathbf{P}_t^*$. In the general TVP case, this backward filter may be computed by:

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

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

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

$$\mathbf{W}_t^* = \left(\mathbf{I} + \mathbf{W}_{t+1}^*\mathbf{V}_{t+1}\right)^{-1}\mathbf{W}_{t+1}^* + \mathbf{x}_t'\mathbf{x}_t\,, \tag{44}$$

$$\mathbf{b}_t^* = \mathbf{W}_t^{*-1}\mathbf{z}_t^*, \quad t \le n-k\,, \tag{45}$$

$$\mathbf{P}_t^* = \sigma_\varepsilon^2\mathbf{W}_t^{*-1}, \quad t \le n-k\,. \tag{46}$$

The backward filter $\mathbf{b}_{t+1}^*$ obtained in this manner estimates $\boldsymbol{\beta}_{t+1}$, conditional on y_{t+1}, ... y_n, with variance $\mathbf{P}_{t+1}^*$, but it also provides an estimate of $\boldsymbol{\beta}_t$, conditional on the same values, with the somewhat larger variance $\mathbf{P}_{t+1}^* + \mathbf{Q}_{t+1} = \sigma_\varepsilon^2\mathbf{W}_{t+1}^{*-1}\left(\mathbf{I} + \mathbf{W}_{t+1}^*\mathbf{V}_{t+1}\right)$. Since $\mathbf{b}_{t+1}^*$ as an

estimate of $\boldsymbol{\beta}_t$ is independent of the filter estimate $\mathbf{b}_t$, the two estimates may be averaged in inverse proportion to their respective covariance matrices to form the smoother estimate $\mathbf{b}_t^S$ of $\boldsymbol{\beta}_t$, conditional on the entire sample, as follows:

$$\begin{aligned}\mathbf{b}_t^S &= \mathbf{W}_t^{S^{-1}}\left(\mathbf{W}_t\mathbf{b}_t + \left(\left(\mathbf{I} + \mathbf{W}_{t+1}^*\mathbf{V}_{t+1}\right)^{-1}\mathbf{W}_{t+1}^*\right)\mathbf{b}_{t+1}^*\right)\\ &= \mathbf{W}_t^{S^{-1}}\mathbf{z}_t^S, \end{aligned} \tag{47}$$

where

$$\mathbf{W}_t^S = \mathbf{W}_t + \left(\mathbf{I} + \mathbf{W}_{t+1}^*\mathbf{V}_{t+1}\right)^{-1}\mathbf{W}_{t+1}^*, \tag{48}$$

$$\mathbf{z}_t^S = \mathbf{z}_t + \left(\mathbf{I} + \mathbf{W}_{t+1}^*\mathbf{V}_{t+1}\right)^{-1}\mathbf{z}_{t+1}^*. \tag{49}$$

This smoother estimate has variance

$$\mathbf{P}_t^S = \sigma_\varepsilon^2\mathbf{W}_t^{S^{-1}}. \tag{50}$$

Note that it is not necessary to actually compute the backward filter and variance $\mathbf{b}_t^*$ and $\mathbf{P}_t^*$ themselves, however, since $\mathbf{z}_t^*$ and $\mathbf{W}_t^*$ suffice to obtain the smoother and its variance using (47) – (50). The smoother and its variance may therefore be computed even for $t > n\text{-}k$, even though the backward filter is not defined there.

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

To obtain the smoother in the ALS case (12), we simply set

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

in (43) – (50). If desired, the term $\left(\mathbf{I} + \rho T_t\mathbf{W}_{t+1}^*\mathbf{W}_t^{-1}\right)^{-1}$ may then be replaced by $\mathbf{W}_t\left(\mathbf{W}_t + \rho T_t\mathbf{W}_{t+1}^*\right)^{-1}$ to avoid having to invert $\mathbf{W}_t$. In the general TVP case, where the transition covariance matrix $\mathbf{Q}_t$ is well defined for all t, we may compute the smoother clear back to $t = 1$. In the ALS case, however, $\mathbf{W}_t^S$, which may be written, using the above substitution, as $\mathbf{W}_t\left(\mathbf{I} + \left(\mathbf{W}_t + \rho T_t\mathbf{W}_{t+1}^*\right)^{-1}\mathbf{W}_{t+1}^*\right)$, is proportional to $\mathbf{W}_t$ 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.

Appendix II

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 (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 $\mathbf{b}_{t|t\text{-}1}$ 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 (36) – (38) 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.[19]

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 "$\mathbf{P}(t\text{-}1)$ +

[19] 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) equation (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 which led Sargent (1999) to interpret it as $\mathbf{P}_{t\text{-}1|t\text{-}1}$. However, he points out that even with this interpretation there is an error, since then the $\mathbf{R}_1(t)$ in the first part of (39) on 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.," *Review of Economic Dynamics* **8** (2005): 262-302.

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

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.

Fisher, Irving. *The Theory of Interest*. Augustus M. Kelley, New York, 1965 reprint of 1930 edition.

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. Supported by PI's earlier NSF grant SES 80-05994. <http://www.econ.ohio-state.edu/jhm/papers/MQT80.pdf>.

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, 2005a, <http://www.econ.ohio-state.edu/jhm/programs/ALS>

McCulloch, J. Huston. GAUSS program ALSGARCH. Fits a linear regression by ALS with GARCH(1,1) effects, estimating signal/noise ratio and GARCH paramters

by ML, 2005b. Developmental version available to reviewers at <http://www.econ.ohio-state.edu/jhm/programs/ALSGARCH>.

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.

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

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

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

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

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

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," *Review of Economic Dynamics* **8** (2005): 360-91.

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

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

Stock, James H., and Mark W. Watson. "Median Unbiased Estimation of Coefficient Variance in a Time-Varying Parameter Model," *Journal of the American Statistical Association* **93** (1998): 349-358.

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.