Adaptive Models and Heavy Tails

Davide Delle Monache*[Job Market Paper]
Queen Mary, University of London

Ivan Petrella†
Birkbeck, University of London

November, 2013

Abstract

This paper proposes a novel and flexible framework to estimate autoregressive models with time-varying parameters. Our framework nests various adaptive algorithms that are commonly used in the macroeconometric literature, such as learning-expectations and forgetting-factor algorithms. These are generalized along several directions: specifically, we allow for both Student-t distributed innovations as well as time-varying volatility. Meaningful restrictions are imposed to the model parameters, so as to attain local stationarity and bounded mean values. The model is applied to the analysis of inflation dynamics. Allowing for heavy-tails leads to a significant improvement in terms of fit and forecast. Moreover, it proves to be crucial in order to obtain well-calibrated density forecasts. The analysis provides new insights into a recent debate surrounding the role of trend-inflation, as well as time-variation in the persistence and volatility of the inflation rate.


Keywords: Time-Varying Parameters, Score-driven Models, Heavy-Tails, Adaptive Algorithms, Inflation.

*E-mail: d.dellemonache@qmul.ac.uk; Phone: +44 (0)2078825873.
†E-mail: i.petrella@bbk.ac.uk; Phone: +44 (0)2076316418.
1 Introduction

Since the seminal work of Cogley and Sargent (2001), time-varying parameter models have been widely regarded as a flexible tool for investigating the dynamics of key macroeconomic aggregates and changes in the statistical and structural laws that drive their joint behavior.\(^1\) This paper proposes a new adaptive algorithm for time-varying autoregressive models that builds on recent insights of Creal et al. (2012) and Harvey (2013) on score-driven models. To enhance the framework’s flexibility, both Gaussian and Student-t innovations are considered. As a matter of fact, these conditional distributions lead to substantially different updating mechanisms that prove to be more appropriate depending on the specific economic problem we tackle.

The discount regression model has been extensively used in the adaptive control literature (see Brown, 1963; Montgomery and Johnson, 1976; Abraham and Ledolter, 1983). Engineering has a long tradition of exploiting adaptive algorithms based on forgetting factors (see Fagin, 1964 and Jazwinski, 1970). Ljung and Soderstrom (1985) derive recursive formulations of a variety of least squares algorithms for quadratic criterion functions and interpret them as a stochastic approximation of the Gauss-Newton algorithm. The adaptive model developed in this paper extends these traditional adaptive algorithms along various dimensions, making three distinct contributions. First, it considers how the existing algorithms are to be modified in the presence of heavy tails, focussing on Student-t innovations. Second, it introduces time-variation in volatility, emphasizing when and how this interacts with the coefficients’ updating rule. Last, it shows how to impose restrictions on the time-varying parameters, so that the model is locally stationary and has a bounded mean. An obvious advantage of the adaptive algorithm proposed in this paper is to derive from an observation-driven model,\(^2\) so that no simulation techniques are involved and estimation may be pursued by standard methods. Furthermore, restrictions to the model can be conveniently imposed.\(^3\)

Given the versatility of the adaptive algorithm we develop, one potential application relates to the analysis of learning expectations. Since the seminal work of Marcet and Sargent (1989), adaptive algorithms have been extensively used in macroeconomics to describe the learning mechanism of expectation formation (see, e.g., Sargent, 1999 and Evans and Honkapohja, 2001). It is well known that, under certain conditions, learning algorithms can be obtained as a restriction on the Kalman filter (Sargent and Williams, 2005; Evans et al., 2010). We show that the most of commonly used learning algorithms can be derived as a special case of the algorithm developed in this paper. As consequence, we open the route to the analysis of

---

\(^1\)Recent contributions have also suggested that – when used in real time– these models provide good economic forecasts (see, e.g., D’Agostino et al., 2013).

\(^2\)Cox (1981) categorizes time series models with time-varying parameters into parameter-driven and observation-driven models. In the former class of models the parameters are stochastic processes which are subject to their own source of error. In the observation-driven approach the parameters are functions of the observed variables. Although the parameters are stochastic, they are perfectly predictable given past information.

\(^3\)In contrast, parameters-driven models typically rely on simulation techniques and they can be computational demanding when restrictions are imposed (see, e.g., Koop and Potter, 2011 and Chan et al., 2013).
learning dynamics in the presence of time-variation in the volatility of the structural innovations (see, e.g., Justiniano and Primiceri, 2008) and/or in a context where rare events are introduced into a structural macroeconomic model (see Curdia et al., 2013). Furthermore, we discuss a convenient way to implement the projection facility used in the learning context.4

Our work also relates to a number of studies on forecasting in the presence of structural changes. In this context, Stock and Watson (1996) have pioneered the use of time-varying regression models that imply an exponentially weighting scheme. Giraitis et al. (2011) consider deterministic time-varying coefficient models and discuss the properties of the non-parametric estimation approach for an autoregressive model with a stochastic attractor. Related work by Pesaran and Timmerman (2007), Pesaran and Pick (2011) and Pesaran et al (2013) considers the issue of how choosing optimal weights in the presence of structural breaks. In strict connection with these works, we show that with Gaussian innovations our model reduces to an exponential weighted algorithm, while with Student-t innovations we obtain a double weighting scheme over time and across-realizations. Koop and Korobilis (2012) propose the use of an exponential weighted algorithm (obtained by ad-hoc restrictions on the Kalman Filter) to model time-variation in both the coefficients and volatility. We show how their algorithm is nested as special case of the adaptive model we put forward.5

We implement the adaptive model in the analysis of inflation dynamics. Specifically, we show how the time-varying autoregressive model improves upon the evidence of Cogley (2002), whose forecast relies on a simple exponential weighting moving average. Imposing bounds to the long-run mean is relatively convenient in our framework. In this respect, we report results in line with recent work of Chan et al. (2013), in that the imposed discipline (marginally) improves the forecast of the time-varying model. Most importantly, we highlight how allowing for heavy-tails is of great importance to obtain well-calibrated density forecasts. In line with the literature (e.g., Cogley and Sargent, 2005), we report significant changes in the persistence of inflation volatility. Interestingly, the model points to a lower degree of persistence when we allow for a larger number of lags (in line with Pivetta and Reis, 2007, who consider an AR(3) specification). However, this model tends to be (marginally) dominated in terms of out-of-sample forecasts by specifications with a smaller number of lags. The inclusion of heavy tails induces more high frequency variation in the volatility of inflation above the well documented low frequency variation over the sample.

The paper is organized as follows. Section 2 introduces the score-driven autoregressive model with Gaussian innovations and discusses its relationship with well-established adaptive algorithms; Section 3 extends the model to the case of Student-innovations; Section 4 shows how to impose restrictions to the model parameters; Section 5 reports an application to inflation dynamics; Section 6 concludes.

4The projection facility is a procedure that constrains the time-varying parameters in the neighborhood of a particular solution, such as the Rational Expectations (RE) equilibrium; see Timmermann (1996) and Evans and Honkapohja (1998). In the context of adaptive algorithms, the parameters are restricted so that the model produces stable predictions; see Ljung and Soderstrom (1985, Section 3.4.4).

5Koop and Korobilis (2012) consider a more general multivariate specification with possible time-varying dimensions. It is clear that the approach discussed in this paper generalizes to the multivariate case.
2  Autoregressive model with time varying parameters

An autoregressive model of order $p$ with time-varying parameters is defined as

$$y_t = \phi_{0,t} + \phi_{1,t}y_{t-1} + \ldots + \phi_{t,p}y_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim IID \left(0, \sigma^2_t\right), \quad t = 1, \ldots, n, \quad (1)$$

where $IID \left(0, \sigma^2_t\right)$ stands for independently and identically distributed random variable with time-varying variance $\sigma^2_t$. The model is typically augmented with an updating rule describing the dynamics of the parameters. Specifically, let $f_t$ define the vector of time-varying parameters, its variation is described by a dynamic model, e.g. a first order Markov process

$$f_{t+1} = \omega + Af_t + \eta_t, \quad \eta_t \sim IID \left(0, Q_t\right), \quad (2)$$

where $\omega$, $A$ and $B$ are matrices of appropriate dimension containing the static parameters, and $\eta_t$ is a vector of stochastic shocks driving the parameters’ variation. Equations (1)-(2) denote the typical specification of a parameter-driven model. In particular, given past and concurrent observations, the filtered estimates of $f_t$ is not perfectly predictable and in fact it is considered as unobserved state vector with an associated covariance matrix which also is recursively estimated.\(^6\)

The alternative avenue to model the time-variation of the parameters, which is followed in this paper, is represented by observation-driven models. In line with Creal et al (2012) and Harvey (2013), the dynamics of the time-varying parameters is driven by the scaled score of the conditional distribution. Denote with $f_{t|t-1} = (\phi_{t|t-1}^0, \sigma_{t|t-1}^2)'$ the filter estimate of $f_t$ given information up to time $t - 1$, the updating rule for the parameters’ variation is

$$f_{t+1|t} = \omega + Af_{t|t-1} + Bs_t, \quad (3)$$

where $\omega$, $A$ and $B$ are matrices of appropriate dimension containing the static parameters. The driving mechanism is equal to the scaled score vector, $s_t = \mathcal{I}_t^{-1}\nabla_t$, which is computed as follows

$$\nabla_t = \frac{\partial [\ell_t(\gamma_t | \mathcal{F}_t, \theta)]}{\partial f_{t|t-1}} \quad \text{and} \quad \mathcal{I}_t = -\mathbb{E} \left[ \frac{\partial^2 [\ell_t(\gamma_t | \mathcal{F}_t, \theta)]}{\partial f_{t|t-1} \partial f_{t|t-1}'} \right], \quad (4)$$

where $\ell_t(\gamma_t | \mathcal{F}_t, \theta) = \log p(\gamma_t | \mathcal{F}_t, \theta)$ is the predictive log-likelihood for the $t-$th observation which is conditioned to the information set $\mathcal{F}_t = \{F_t, Y_{t-1}\}$. Specifically, $F_t = \{f_{t|t-1}, f_{t-1|t-2}, \ldots, f_{1|0}\}$ denotes present and past values of the estimated parameters and $Y_{t-1} = \{y_{t-1}, y_{t-2}, \ldots, y_1\}$ are the past observations.

Note that $\nabla_t$ is known as the score vector and the scaling matrix $\mathcal{I}_t^{-1}$ is the inverse Fisher information matrix. As a result, the scaled score vector has the conditional mean $\mathbb{E}(s_t | \mathcal{F}_t) = 0$ and variance $\mathbb{E}(s_ts_t' | \mathcal{F}_t) = \mathcal{I}_t$, thus the updating rule (3) can be seen as a Gauss–Newton type al-

\(^6\)For linear and Gaussian models, the likelihood function can be computed in closed form using the Kalman filter (KF) (see Harvey, 1989 and Kim and Nelson, 1999). In non-linear and non-Gaussian models, conditional density is instead generally evaluated via simulation methods (see e.g. Durbin and Koopman, 2001).
gorithm for estimating the time-varying parameters.\textsuperscript{7} Clearly, in the observation-driven framework the vector $f_{t+1|t}$, although stochastic, is perfectly predictable at time $t$. The observation-driven models can be estimated by maximum likelihood. Thus, the vector of static parameters is estimated as
\[
\hat{\theta} = \arg \max \mathcal{L} = \arg \max \sum_{t=1}^{n} \ell_t (y_t | \mathcal{F}_t, \theta).
\]
The evaluation of the log-likelihood is straightforward and the maximization can be obtained using recursive formulae for the Gradient and the Hessian of $\mathcal{L}$ with respect to the static parameter $\theta$. Alternatively, those derivatives can be obtained numerically. In line with Creal et al (2012, sec. 2.3) we conjecture that $p_n(b) \sim \mathcal{N}(0, \Omega)$, where $\Omega$ is evaluated by numerical derivative at the optimum.

Model (1) can be rewritten in vector form as
\[
y_t = x_t' \phi_{t|t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{IID}(0, \sigma^2_{t|t-1}), \quad t = 1, \ldots, n,
\]
where $x_t = (1, y_{t-1}, \ldots, y_{t-p})'$ and $\phi_{t|t-1} = (\phi_{0,t|t-1}, \phi_{1,t|t-1}, \ldots, \phi_{p,t|t-1})'$. Under Gaussian distribution $\varepsilon_t \sim \mathcal{N}(0, \sigma^2_{t|t-1})$, the predictive log-likelihood at time $t$ is equal to
\[
\ell_t (y_t | \mathcal{F}_t, \theta) = -\frac{1}{2} \log (2\pi) - \frac{1}{2} \log \sigma^2_{t|t-1} - \frac{\varepsilon_t^2}{2\sigma^2_{t|t-1}},
\]
where $\varepsilon_t = y_t - x_t' \phi_{t|t-1}$ is the prediction error and $\sigma^2_{t|t-1}$ is the conditional variance.\textsuperscript{8} It can be shown that $\mathcal{I}_t$ is block diagonal so that the scaled score vector $s_t$ can be specialized in two parts: the vector $s_{s\phi t}$ driving the coefficients
\[
s_{s\phi t} = (x_t \sigma^{-2}_{t|t-1} x_t')^{-1} x_t \sigma^{-2}_{t|t-1} \varepsilon_t,
\]
and the scalar $s_{s\sigma t}$ driving the volatility
\[
s_{s\sigma t} = (\varepsilon_t^2 - \sigma^2_{t|t-1}).
\]
In accordance with the literature on time-varying parameters models, we choose a random walk specification for the updating rule, i.e. we set $\omega = 0$ and $A = I$ in (3). Furthermore, $B$ is restricted to depend only upon two scalar parameters leading to the following recursions:\textsuperscript{9}
\[
\phi_{t+1|t} = \phi_{t|t-1} + \kappa_\phi (x_t \sigma^{-2}_{t|t-1} x_t')^{-1} x_t \sigma^{-2}_{t|t-1} \varepsilon_t,
\]
and
\[
\sigma^2_{t+1|t} = \sigma^2_{t|t-1} + \kappa_\sigma (\varepsilon_t^2 - \sigma^2_{t|t-1}).
\]
\textsuperscript{7}Different scaling matrix can be used as discussed in Creal et al (2012, sec. 2.2).
\textsuperscript{8}When the model is written in vector form it becomes evident that the results derived in this paper generalize to any univariate model with exogenous and/or predetermined regressors.
\textsuperscript{9}One could relax those restrictions allowing a more general specification of $\omega$, $A$ and $B$. 
Note that the scaling matrix $x_t \sigma_{t|t-1}^{-2} x_t'$ is not invertible, so we can use the Moore-Penrose pseudo-inverse. In order to avoid swift changes in the parameters it is customary to replace the Hessian matrix $(x_t \sigma_{t|t-1}^{-2} x_t')$ with its smoothed version\(^{10}\)

$$R_t = (1 - \kappa_h) R_{t-1} + \kappa_h x_t \sigma_{t|t-1}^{-2} x_t' = R_{t-1} + \kappa_h (x_t \sigma_{t|t-1}^{-2} x_t' - R_{t-1}), \quad (11)$$

and the updating rule for the coefficients (9) becomes now equal to

$$\phi_{t+1|t} = \phi_{t|t-1} + \kappa \phi R_t^{-1} x_t \sigma_{t|t-1}^{-2} (y_t - x_t' \phi_{t|t-1}). \quad (12)$$

Note that under Gaussianity the time-varying variance is simply a scalar factor and thus it does not directly affect the coefficients’s filtering in (9), even though it will have an impact on the estimation of the static parameters. However, this is not anymore the case when the Hessian matrix is replaced with a smoothed version, i.e. when $\kappa_h \neq 1$ in (11).

Equations (10)-(12) describe the dynamics of the parameters in an observation-driven model. As opposed to the parameter-driven approach in (2), both the signal (5) and the parameters (3) are driven by the prediction error, as such the model is similar to the single-source of error model of Casalas et al (2002) and Hyndman et al (2008).\(^{11}\)

### 2.1 Relation with the adaptive algorithms

This section highlights the relation between the score-driven model with Gaussian innovations and various adaptive algorithms widely used in the literature. To make this parallel it is convenient to start with a model with constant variance, so that the derivations in the previous section can be viewed as a generalization to the time-varying variance case. With constant variance, and setting $\kappa_\phi = \kappa_h = \kappa$, the score-driven filter (11)-(12) collapses to

$$R_t = R_{t-1} + \kappa (x_t \sigma_{t|t-1}^{-2} x_t' - R_{t-1}), \quad (13)$$

$$\phi_{t+1|t} = \phi_{t|t-1} + \kappa R_t^{-1} x_t \sigma_{t|t-1}^{-2} (y_t - x_t' \phi_{t|t-1}).$$

this is exactly the so-called Constant Gain Learning (CGL) algorithm, widely used in the learning expectations literature.\(^{12}\)

**Lemma 1** The CGL algorithm weights the observations $y_{t-j}$ with the exponential rate $(1 - \kappa)^j$, where $0 < \kappa < 1$, and the parameter $\kappa$ gives a trade-off between the tracking capability and the noise sensitivity. Moreover, the CGL is a forgetting factor algorithm and it can be also derived from an off-line method, i.e. the discounted least squares principle. See details in Appendix.

---

\(^{10}\)For some extreme observation at time $t$, the the second moment matrix can be very large or very small and this might lead to instability (see Creal et al., 2012). Ljung and Soderstrom (1985) justify the smoothing of the Hessian matrix appealing to the stochastic Gauss-Newton principle as it is discussed in the next section.

\(^{11}\)In the single source of error specification, the state space model has perfectly correlated disturbances, the MSE of the state vector converges to zero and the filter is equal to the smoother.

The discounted regression model, has been extensively used in adaptive control literature (see Brown, 1963, Montgomery and Johnson, 1976, and Abraham and Ledolter, 1983). Similarly, in the engineering literature the same algorithm is known as forgetting factor algorithm. Fagin (1964) notes that a given linear state space model might be adequate for a time period but may not be for long time intervals, and therefore proposes to robustify the KF using an exponentially decay forgetting factor labelled as fading memory (or limited memory) filter (see Jazwinski, 1970, p. 255).

The CGL algorithm is often derived from a parameter-driven model (2) with specific restrictions. In this respect it is useful to point out the result of the following Lemma.

**Lemma 2** Given the parameter-driven model

\[ y_t = x_t' \phi_t + \varepsilon_t, \varepsilon_t \sim N(0, \frac{\sigma^2}{1 - \kappa}), \tag{14} \]

\[ \phi_{t+1} = \phi_t + \eta_t, \eta_t \sim N(0, P_t[I - \frac{\kappa}{1 - \kappa}]), \]

where \( P_{t|t} = \text{E}[(\phi_{t|t} - \phi_t)(\phi_{t|t} - \phi_t)'] \) and \( \kappa \) is the gain parameter. The implied KF is exactly equal to the CGL algorithm. It is worth noticing that the restrictions on (14) imply that the shock \( \eta_t \) is driven by the prediction error and thus the model (14) collapses to an observation-driven model. See Appendix for details.

Note that Koop and Korobilis (2012) propose to estimate a large VAR with time-varying parameters using the restrictions in the previous Lemma. Thus they use the CGL algorithm which is nested within the score-driven framework. Koop and Korobilis (2012) also allow for time-varying covariance matrix which they estimate by an exponential smoothing and also this feature is nested in our framework.

It is worth discussing another widely used specifications of (14), which assumes \( \varepsilon_t \sim N(0, \sigma^2) \) and \( \eta_t \sim N(0, \kappa^2 \Sigma) \), with \( \Sigma = \sigma^2[\text{E}(x_t x_t')^{-1} \) (see Stock and Watson, 1996, Sargent and William, 2005, Branch and Evans, 2006 and Li, 2008). Evans et al. (2010) named this specification Stochastic Gradient algorithm,\(^{13}\) whereas Slobodyan and Wouters (2012) refer to it as Kalman filter learning.

**Lemma 3** Setting \( \eta_t \sim N(0, \kappa^2 \Sigma) \) implies that the parameter-driven model (14) collapses to an observation-driven model and the KF converges to the score-driven filter (9), where the time-varying scaling matrix is replaced by its unconditional expectation \( \sigma^{-2} \text{E}(x_t x_t') \). Similarly, setting \( \eta_t \sim N(0, \kappa^2 \Sigma^{-1}) \) leads to the score-driven filter (9) where the scaling matrix replaced by the identity matrix. See details in the Appendix.

It is instructive at this point to recall that all of the recursive algorithms discussed in this sub-section can be seen as particular cases of the adaptive algorithms popularized by Ljung\(^{13}\) Note that this specification is really an approximation of the Stochastic Gradient Algorithm; see details of Lemma 3 discussed in the Appendix.
Remark 4 Ljung and Soderstrom (1985) show how the CGL algorithm can be seen as a recursive solution of quadratic loss function criterion. In particular, given a sequence of random IID random variables \( \epsilon = \{ \epsilon_1, ..., \epsilon_T \} \), the optimal choice of coefficients \( \phi = \{ \phi_1, ..., \phi_T \} \) can be obtained by solving a quadratic criterion \( \mathbb{E}[\epsilon' \sigma^{-2} \epsilon] \), and it leads to a recursive stochastic Gradient method which is the stochastic analog of a Gauss-Newton search direction

\[
\hat{\phi}_{t+1|t} = \hat{\phi}_{t|t-1} + \kappa_t [H(\hat{\phi}_{t|t-1}, \epsilon_t)]^{-1} G(\hat{\phi}_{t|t-1}, \epsilon_t),
\]

where \( G(\hat{\phi}_{t|t-1}, \epsilon_t) \) and \( H(\hat{\phi}_{t|t-1}, \epsilon_t) \) are the Gradient vector and the Hessian matrix respectively, and \( \kappa_t \) is a sequence of gain parameters appropriately chosen. Under Gaussian distribution the minimization of the quadratic loss function is equivalent to the maximization of the log-likelihood function.

Therefore, the most commonly used adaptive algorithms can be derived from a score-driven model under Gaussian innovation. In fact, the score-driven models mimics the principle underline in Remark 4. The score-driven model (5), with (10)-(12), extends the adaptive algorithms discussed in this sub-section to allow for changing in the volatility, and its updating rule has been derived following exactly the same reasoning of the stochastic Gauss-Newton mechanism. The estimated variance obtained by (10) implies an exponentially smoothing of the squared prediction errors

\[
\sigma^2_{t+1|t} = \kappa_\sigma \sum_{j=0}^{\infty} (1 - \kappa_\sigma)^j \epsilon^2_{t-j}.
\]

It is worth noticing that Ljung and Soderstrom (1985, sec. 3.4.3) and Koop and Korobilis (2012) use this model to capture the variation in the volatility. Their filtering is however chosen in a rather heuristic way and not derived from the Gauss-Newton principle.

Next section extends the adaptive algorithms to the case of Student-t distribution, i.e. to a framework with heavy tails. This can be seen as recursive solution with non-quadratic loss function mentioned in Ljung and Soderstrom (1985, sec. 3.5).

3 Student-t Distribution

The score-driven model can be easily extended to the case of non-Gaussian distributions. The Student-t has higher mass probability on the tails of the distribution, therefore it can be considered for the cases where rare events become relevant. Harvey and Luati (2012) highlight that a score-driven filter for the Student-t innovations leads to an algorithm which is robust to
few large errors.\footnote{A score-driven model with non-Gaussian innovations, not only modifies the likelihood function (as, e.g., standard t-GARCH of Bollerslev, 1987) but it also implies a different filtering process for the time–varying parameters.} Model (5) becomes

\[ y_t = x_t' \phi_{t|t-1} + \varepsilon_t, \quad \varepsilon_t \sim t_v(0, \sigma^2_{t|t-1}), \]  

where \( \sigma^2_{t|t-1} \) is the conditional variance and \( v \) is the degrees of freedom parameter regulating the heavy-tails. The predicted log-likelihood can be written as

\[ \ell_t(y_t|\mathcal{F}_t, \theta) = c(\eta) - \frac{1}{2} \ln \sigma^2_{t|t-1} - \left( \frac{\eta + 1}{2\eta} \right) \log \left[ 1 + \frac{\eta}{1 - 2\eta} \frac{\varepsilon^2_t}{\sigma^2_{t|t-1}} \right], \]  

where

\[ c(\eta) = \log \left[ \Gamma \left( \frac{\eta + 1}{2\eta} \right) \right] - \log \left[ \Gamma \left( \frac{1}{2\eta} \right) \right] - \frac{1}{2} \log \left( \frac{1 - 2\eta}{\eta} \right) - \frac{1}{2} \log \pi, \]

\( \eta = 1/v \) and \( \Gamma \) is the Gamma function. The model (15) generalizes the setting in Harvey and Luati (2012) to the case of additional regressors and time-varying variance. It can be shown that the scaled-score driving the coefficients and the variance are equal to

\[ s_{\phi t} = \frac{(1 - 2\eta)(1 + 3\eta)}{(1 + \eta)} (x_t \sigma^{-2}_{t|t-1} x_t')^{-1} x_t \sigma^{-2}_{t|t-1} w_t \varepsilon_t, \]  

and

\[ s_{\sigma t} = (1 + 3\eta) (w_t \varepsilon^2_t - \sigma^2_{t|t-1}). \]  

Notice that both depend upon the scalar weights

\[ w_t = \frac{(1 + \eta)}{(1 - 2\eta + \eta \zeta_t)}, \]  

where \( \zeta_t = \varepsilon^2_t / \sigma^2_{t|t-1} \) and they nest the Gaussian case for \( \eta = 0 (v \to \infty) \); see Appendix for details. Clearly, the resulting adaptive algorithm is affected by the distributional assumption. Furthermore, while in a Gaussian setting the score driving the dynamic of the coefficients is not affected by the variance, when we allow for Student-t the time-varying volatility has a direct impact on the updating mechanism for the time-varying coefficients.

The crucial role played by the weights (19) is visualized by the plots in Figure 1. Specifically, the former show the magnitude of the weights \( w_t \), while the latter shows the weighted realizations \( w_t \sqrt{\zeta_t} \) which is known as influence function in the robust literature (see Maronna et al., 2006). Note that large innovations are categorized as being part of the tails of the distribution, as such they are downweighted and have a small effect on the dynamic of the time-varying parameters.

Under Student-t distribution the score-driven algorithm leads to a robust filter and gener-
alizes the CGL algorithm (13). This is formalized in the following proposition.

**Proposition 5** Under Student-t distribution, the score-driven model leads to the following adaptive algorithm for the time-varying coefficients

\[ R_t = R_{t-1} + \kappa_\phi \theta w_t (x_t \sigma_{t-1}^{-2} x'_t - R_{t-1}), \]

\[ \phi_{t+1|t} = \phi_{t|t-1} + \kappa_\phi R_{t}^{-1} x_t \sigma_{t-1}^{-2} [\theta w_t (y_t - x'_t \phi_{t|t-1})], \]

with \( \theta = \frac{(1-2\eta)(1+3\eta)}{(1+\eta)} \) and \( w_t \) are expressed in (19). The algorithm weights the observations with rate \( \gamma_j w_t \), where \( \gamma_j = \prod_{k=t-j+1}^t (1 - \kappa_\theta w_k) \) and \( \gamma_0 = 1 \), denote the weights assigned to the observations \( y_{t-j} \) in estimating the coefficients at time \( t \), and \( \kappa_\theta = \kappa_\phi \theta \). Thus, the algorithm implies a double weighting schema, where \( w_t \) weights the realization across the support of the distribution and \( \gamma_j \) are the weights across time. Clearly, for the Gaussian case \( w_t = \theta = 1 \ \forall t \), and \( \gamma_j = (1 - \kappa_\phi)^j \), recovering the usual exponentially decaying pattern. The implied adaptive algorithm for the variance is

\[ \sigma^2_{t+1|t} = \sigma^2_{t|t-1} + \kappa_\sigma [(1 + 3\eta) (w_t \epsilon_t^2 - \sigma^2_{t|t-1})], \]

where the weights across the support of the distribution are still \( w_t \), while the weights across time decay exponentially, i.e. \( \gamma_j = [1 - \kappa_\sigma (1 + 3\eta)]^j \). See details in the Appendix.

Proposition 5 underlines how the recursive estimation of the parameters is obtained from a double weighting of the observations across realizations and time. Specifically, the magnitude of the weights \( w_t \) depends on how close the actual observation is to the center of the distribution. Large deviations are downweighted: a small \( w_t \) is more likely with lower degree of freedom and lower dispersion of the distribution. In fact, \( w_t \) reweights each of the observations so that the algorithm is robust to extreme events. Moreover, it is worth emphasizing that, for the recursive estimation of the coefficients, also the weights across time \( \gamma_j \) are affected by \( w_t \), therefore they depend upon a time-varying forgetting factor and cannot be separated from \( w_t \). In contrast, the recursive algorithm for the variance follows a standard exponential decaying pattern once the squared observations are weighted by \( w_t \).

**Remark 6** With a Student-t distribution the score-driven algorithm (20)-(21) is a recursive stochastic Gauss-Newton algorithm which cannot be derived as a solution of quadratic loss function with re-weighted observations of the type discussed in Ljung and Soderstrom (1985, sec. 2.2).

To clarify the latter point it is worth considering a simplified version of (15) with \( x_t = 1 \) and therefore \( \phi_{0,t} = \mu_t \). This specification\(^{15}\) implies an IMA(1,1) model with time-varying moving

\(^{15}\)Harvey and Luati (2013) consider this specification with constant variance.
average coefficient \((1 - \kappa_\theta w_t)\) and variance \(\sigma^2_{t|t-1}\). The estimated level is obtained as follows

\[
\mu_{t+1|t} = \kappa_\theta \sum_{j=0}^{\infty} \gamma_j \tilde{y}_{t-j} = \frac{\kappa_\theta}{1 - (1 - \kappa_\theta w_t)L} \tilde{y}_t, \tag{22}
\]

with \(\tilde{y}_{t-j} = w_{t-j} y_{t-j}\). Equation (22) implies that the time-varying mean is obtained by a double weighting schema so that observations are: (i) weighted to be robust to the impact of extreme events, i.e. \(\tilde{y}_t = w_t y_t\) and (ii) smoothed across time by a one-sided low-pass filter with time-varying coefficients, that is \(\kappa_\theta/[1 - (1 - \kappa_\theta w_t)L]\), and this implies a time-varying transfer function \(G(\lambda) = \kappa_\theta [1 + (1 - \kappa_\theta w_t)^2 - 2(1 - \kappa_\theta w_t) \cos(\lambda)]^{-1/2}\), where \(0 < \lambda < \pi\).\(^{16}\) In contrast, the variance can be obtained by

\[
\sigma^2_{t+1|t} = \kappa_\zeta \sum_{j=0}^{\infty} (1 - \kappa_\zeta)^2 \tilde{\varepsilon}^2_{t-j} = \frac{\kappa_\zeta}{1 - (1 - \kappa_\zeta)L} \tilde{\varepsilon}^2_t, \]

where \(\tilde{\varepsilon}^2_{t-j} = w_{t-j} \varepsilon^2_{t-j}\) is the weighting across realizations, and \(\kappa_\zeta/[1 - (1 - \kappa_\zeta)L]\) is the standard one-sided low-pass filter, with \(\kappa_\zeta = \kappa_\sigma (1 + 3\eta)\).

### 4 Model restrictions

Applications of time-varying parameters models often require to impose restrictions on the parameters space. For instance, in the autoregressive model (1) it is customary to impose restrictions on the autoregressive coefficients so that the implied roots are always within the unit circle, i.e. restrictions implying a locally stationary model.\(^{17}\)

General non-linear restrictions can be accommodated within the score-driven model. This requires to re-parameterized the model with respect to a new vector of unconstrained parameters. Define the following transformation \(\tilde{f}_t = g(\tilde{f}_t)\), where \(\tilde{f}_t\) is the original vector of parameters, \(\tilde{f}_t\) is the new parameterization and \(g(\cdot)\) is a continuous and twice differentiable transformation function, often known as link function, which maps the new vector of unconstrained parameters into the space of constrained parameters. Following Creal et al (2012) and Harvey (2013), the score-driven model (3) is re-parameterized with respect to the new vector of parameters

\[
\tilde{f}_{t+1|t} = \tilde{\omega} + \tilde{A} \tilde{f}_{t|t-1} + \tilde{B} \tilde{s}_t, \tag{23}
\]

where \(\tilde{s}_t = \tilde{T}_t^{-1} \tilde{\varphi}_t\) is the scaled score computed with respect \(\tilde{f} = g^{-1}(f)\), where \(g^{-1}(\cdot)\) is the inverse function of \(g(\cdot)\). For a given continuous and differentiable function \(g(\cdot)\), the new score

---

\(^{16}\)See Dahlhaus (2012) for details on stationary processes with time-varying spectral density.

\(^{17}\)In the Bayesian framework those constrained are usually imposed by rejection sampling (see e.g. Cogley and Sargent, 2005, and Koop and Potter, 2012). Eventually, the approach proposed here can be adopted in a Bayesian framework as well. In practice, the model should be re-parameterized in order to replace the inequality constraints by equality constraints.
vector is then
\[
\nabla_t = \frac{\partial \ell_t}{\partial \mathbf{f}_{t|t-1}} = \left[ \frac{\partial \ell_t}{\partial \mathbf{f}_{t|t-1}'} \frac{\partial \mathbf{f}_{t|t-1}}{\partial \mathbf{f}_{t|t-1}'} \right] = \Psi_t' \nabla_t,
\]
where \(\Psi_t = \partial \mathbf{f}_{t|t-1}/\partial \mathbf{f}_{t|t-1}'\) is the Jacobian of \(g(\cdot)\) and is deterministic given the past information. Therefore, the transformed scaling matrix is equal to \(\tilde{\Psi}_t = \Psi_t' \mathcal{I} \Psi_t\) and the new scaled score is then equal to
\[
\begin{align*}
\tilde{s}_t &= (\Psi_t' \mathcal{I} \Psi_t)^{-1} \Psi_t' \nabla_t.
\end{align*}
\]

At this stage it is worth discussing how the introduction of restrictions is going to affect the adaptive algorithm. The transformation function \(g(\cdot)\) imposes (possibly) non-linear restrictions on the time-varying parameters. However, the dynamic of the unrestricted time-varying parameters (23) is still conditional linear. In practice, the resulting algorithm is solving a non-linear filtering problem by a linear approximation. This intuition is formalized in the Theorem below.

**Proposition 7** Consider the Gaussian model (14) and let impose a non-linear transformation on the coefficients \(\phi_t = g(\alpha_t)\). The model can be solved by the Extended Kalman filter of Anderson and Moore (1979, sec. 8.2) and the implied algorithm is exactly the score-driven filter (23).

Proof in the Appendix.

Ljung and Soderstrom (1985, sec. 6.6) propose to implement a constrained adaptive algorithm by means of a projection facility (see also Timmermann, 1996, and Evans and Honkapohja, 1998). Specifically, they use a constant parameter weighting the driving process such that the incremental step is progressively shrunken until the restriction is satisfied.\(^{18}\) The adaptive model (5), with (23)-(24), automatically achieves the same objective. In fact, the matrix \(\Psi_t\) re-weights the Gauss-Newton search direction so that the restrictions are always satisfied. With respect to the standard projection facility, the re-weighting of our adaptive model varies at different points of the recursion and, most importantly, shrinks the search in the optimal way as opposed to the usual scalar shrinkage.

In the next sub-sections we illustrate how to implement specific restrictions which are commonly imposed to an autoregressive model with time-varying parameters. Specifically, the restrictions on the autoregressive coefficients and bounds on the long-run mean are discussed.

### 4.1 Imposing stationarity

In this section we consider restrictions to the parameters space implying the model is locally stationary: this exploits the mapping between the coefficients of an autoregressive model and its partial autocorrelations. Stationarity is then imposed by simply restricting the latter in the interval \((-1, 1)\). To simplify the notation we start with model (1) without the intercept and then consider how the algorithm changes for the general model.

\(^{18}\)In practice this is often implemented by skipping the updating each time the restrictions are violated.
Proposition 8  For each point in time $t$, let $\phi_t = (\phi_{1,t}, \ldots, \phi_{p,t})'$ denote the vector of coefficients, $\pi_t = (\pi_{1,t}, \ldots, \pi_{p,t})'$ the corresponding vector of partial autocorrelations and $\alpha_t = (\alpha_{1,t}, \ldots, \alpha_{p,t})'$ the vector of unrestricted coefficients. A locally stationary model has $\phi_t \in S^p$, where $S^p$ is the hyperplane with all roots, $z_i$, inside the unit circle, i.e. $\phi_t(z_i) = 0$, $z_i \in \mathbb{C}^p$ and $|z_{j,t}| < 1$ for $j = 1, \ldots, p$. It is possible to show that $\phi_t \in S^p$ if and only if $\pi_t \in \mathbb{R}^p$ and $|\pi_{j,t}| < 1$. Therefore, let $\phi_t = \Phi(\pi_t)$ defines the function mapping the coefficients to the partial autocorrelations and $\pi_t = \Upsilon(\alpha_t)$ a function that restricts the partial autocorrelations to lie in the region $(-1,1)$. The function $\phi_t = \Phi(\pi_t)$ is uniquely obtained by the last recursion of the Durbin-Levinson algorithm

$$
\phi_t^{j,k} = \phi_t^{j,k-1} - \pi_{k,t}\phi_t^{j-k,k-1} \quad \text{for} \quad j = 1, \ldots, k - 1 \quad \text{and} \quad k = 2, \ldots, p,
$$

with $\phi_t^{1,1} = \pi_{1,t}$ and $\phi_t^{k,k} = \pi_{k,t}$. The function $\pi_t = \Upsilon(\alpha_t)$ is any monotonic and differentiable function

$$
\pi_{j,t} = \Upsilon(\alpha_{j,t}), \quad \text{such that} \quad \pi_{j,t} \in (-1,1), \quad j = 1, \ldots, p.
$$

The composite function $g(\cdot) = \Phi[\Upsilon(\cdot)]$ maps the restricted stationary coefficients into the unrestricted parameters, i.e. $\phi_t = g(\alpha_t)$ with $\alpha_t \in (-\infty, \infty)$ and $\phi_t \in S^p$.

The Proof follows from Bandorff-Nielsen and Schou (1973) and Monahan (1984).

The functions $\Phi(\cdot)$ and $\Upsilon(\cdot)$ are continuous and differentiable and the Jacobian matrix is

$$
\frac{\partial g(\alpha_t)}{\partial \alpha_t'} = \frac{\partial \Phi(\pi_t)}{\partial \pi_t'} \frac{\partial \Upsilon(\alpha_t)}{\partial \alpha_t'}.
$$

It is straightforward to show that $\partial \Upsilon(\alpha_t)/\partial \alpha_t'$ is diagonal matrix containing $\partial \Upsilon(\alpha_{j,t})/\partial \alpha_{j,t}$ with $j = 1, \ldots, p$, while the analytic expression for $\Gamma_t = \partial \phi_t/\partial \pi_t'$ is obtained in the theorem below.

Theorem 9  The Jacobian matrix $\Gamma_t$ is obtained from the last iteration of the recursion

$$
\Gamma_{k,t} = \begin{bmatrix}
\bar{\Gamma}_{k-1,t} & b_{k-1,t} \\
0_{k-1}' & 1
\end{bmatrix},
$$

$$
\bar{\Gamma}_{k-1,t} = J_{k-1,t}\Gamma_{k-1,t}, \quad k = 2, \ldots, p,
$$

with

$$
b_{k-1,t} = \begin{bmatrix}
\phi_{k-1,k-1}^{1,1} \\
\phi_{k-1,k-1}^{1,2} \\
\vdots \\
\phi_{k-1,k-1}^{2,1} \\
\phi_{k-1,k-1}
\end{bmatrix}, \quad J_{k-1,t} = \begin{bmatrix}
1 & 0 & \cdots & 0 & -\pi_{k,t} \\
0 & 1 & 0 & -\pi_{k,t} & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & -\pi_{k,t} & 0 & 1 & 0 \\
-\pi_{k,t} & 0 & \cdots & 0 & 1
\end{bmatrix}.
$$

Note that if $k$ is even the central element of $J_{k-1,t}$ is equal to $(1 - \pi_{k,t})$. The recursion is initialized with $J_{1,t} = (1 - \pi_{2,t})$ and $\Gamma_{1,t} = 1$.  

13
Proof in the Appendix.

Given the scaled score vector \( s_{\phi t} = \mathcal{I}_{\phi t}^{-1}\nabla_{\phi t} \) computed in (17), the adaptive algorithm for the transformed coefficients is\(^{19}\)

\[
\alpha_{t+1|t} = \alpha_{t|t-1} + \kappa_{\alpha}(\Psi'_{t}\mathcal{I}_{\phi t}\Psi_{t})^{-1}\Psi'_{t}\nabla_{\phi t},
\]

where \( \phi_{t} = g(\alpha_{t}) \) and \( \Psi_{t} = \Psi(\alpha_{t}) \) are computed as outlined in Proposition 8 and Theorem 9, respectively. When the time-varying intercept is included without any restrictions, i.e. \( \phi_{0,t} = \alpha_{0,t} \), the Jacobian matrix is modified as follows

\[
\Psi_{t} = \frac{\partial \phi_{t}}{\partial \alpha_{t}} = \begin{bmatrix}
1 & 0' \\
0 & \Psi_{22,t}
\end{bmatrix},
\]

where \( \Psi_{22,t} = \partial(\phi_{1,t}, ..., \phi_{p,t})'/\partial(\alpha_{1,t}, ..., \alpha_{p,t}) \) is computed in the Theorem 9.

4.2 Bounded trend

However, it is often the case that the informed researcher wants to discipline the algorithm so to have a bounded conditional mean. Following Beveridge and Nelson (1981), a stochastic trend can be expressed in terms of long-horizon forecasts. For a driftless random variable, the Beveridge-Nelson trend is defined as the value to which the series is expected to converge once the transitory component dies out (see e.g. Benati, 2007 and Cogley et al, 2010). Specifically, for a stationary time-varying autoregressive process, local-to-date approximation implies that the unconditional time-varying mean is equal to \( \mu_{t} = \phi_{0,t} / (1 - \sum_{j=1}^{p} \phi_{j,t}) \). In line with Cogley et al (2010), our specification implies that the detrended component, that is \( \tilde{y}_{t} = (y_{t} - \mu_{t}) \), follows a locally stationary time-varying AR(p) model. Following Chan et al (2013), we want to restrict \( \mu_{t} \in [b, \overline{b}] \).

**Proposition 10** Let \( h(\cdot) \) be any continuous and differential function so that \( h(\cdot) \in [b, \overline{b}] \). The restriction on \( \mu_{t} \in [b, \overline{b}] \) can achieved by the following transformation

\[
\phi_{0,t} = h(\alpha_{0,t}) \left( 1 - \sum_{j=1}^{p} \phi_{j,t} \right).
\]

The Jacobian matrix is then equal to

\[
\Psi_{t} = \frac{\partial \phi_{t}}{\partial \alpha_{t}} = \begin{bmatrix}
\psi_{11,t} & \psi_{12,t}' \\
\psi_{12,t} & \psi_{22,t}
\end{bmatrix},
\]

\(^{19}\)In order to restrict the partial autocorrelations we use the inverse Fisher transformation \( \Upsilon(\alpha_{j,t}) = \frac{\exp(2\alpha_{j,t}) - 1}{\exp(2\alpha_{j,t}) + 1} \), so that \( \alpha_{j,t} = \arctan(\pi_{j,t}) \), with \( j = 1, ..., p \).
where $\Psi_{22,t}$ has been computed in Theorem 9, while $\psi_{11,t}$ and $\psi'_{12,t}$ are

$$
\psi_{11,t} = \frac{\partial h(\alpha_{0,t})}{\partial \alpha_{0,t}} \left( 1 - \sum_{j=1}^{p} \phi_{j,t} \right), \quad \psi'_{12,t} = -h(\alpha_{0,t}) \lambda' \Psi_{22,t},
$$

where $\lambda = (1, 1, ..., 1)'$.

The Proof is straightforward.

To summarize, for each time $t$, the recursion (30) is implemented as follows: first, the stationary AR coefficients are computed following Proposition 8; second, the constrained intercept\(^{20}\) and the Jacobian matrix $\Psi_t$ are computed as described in Proposition 10; and finally, all the necessary elements to update $\alpha_{t|t-1}$ are available.

## 5 Application to the inflation dynamics

Following Cogley and Sargent (2005) and Pivetta and Reis (2007), we estimate the following $p$-th order autoregressive representation for inflation:

$$
\pi_t = \phi_{0,t} + \sum_{j=1}^{p} \phi_{j,t} \pi_{t-j} + \varepsilon_t, \quad \varepsilon_t \sim IID \left(0, \sigma^2_t\right), \quad t = 1, ..., n. \quad (34)
$$

This specification is flexible enough to capture changes in the long-run trend as well as changes in the persistence of the deviation around the trend. In addition, it allows for variation in the volatility which has been proved to be particularly important to understand the dynamic of inflation (see e.g. Pivetta and Reis, 2007 and Clark and Doh, 2011). Those features are of foremost importance to understand the changes in inflation dynamics over the post-WWII sample. The literature has mainly focussed on the parameter-driven approach, estimated by Bayesian methods, to estimate these models.\(^{21}\)

The change in the persistence of the inflation has been strongly advocated by Cogley and Sargent (2001).\(^{22}\) Specifically, they find that the persistence of inflation in the United States rose in the early 1970s and remained high during this decade, before starting a gradual decline from the early 1980s until the present. Pivetta and Reis (2007) challenge these findings presenting evidence of a stable level of persistence throughout the sample. It is therefore interesting to examine those issues in the light of our model.

A related issue that has received much attention in the recent years is related to the presence of a time-varying level of trend-inflation. Specifically, trend-inflation is generally thought as a measure of the public’s perception of the credibility of the central bank inflation targeting. (Kozicki and Tinsley, 2001, and Faust and Wright, 2011). Furthermore Clark and Doh (2011) and Chan et al. (2013) highlight how trend-inflation might play an important role in disciplining

---

\(^{20}\)In the application we use $h(\alpha_{0,t}) = \frac{\exp(\alpha_{0,t}) + \lambda}{\exp(\alpha_{0,t}) + 1}$, so that $\alpha_{0,t} = \ln \left( \frac{\text{Exp}_t}{\text{Exp}_t+\lambda} \right)$.

\(^{21}\)A noticeable exception is the work of Pivetta and Reis (2007).

\(^{22}\)Similar results are provided by Taylor (2000) and Brainard and Perry (2000).
inflation forecasts at a long-horizon. Let $\pi_t$ denote trend-inflation, following Beveridge and Nelson (1981) the stochastic trend is defined as the value to which the series is expected to converge in the long-run, i.e. $\lim_{h \to \infty} E_t (\pi_{t+h}) = \pi_t$. Furthermore, defining the inflation-gap as the deviation of current inflation from the trend, i.e. $\pi_t = \pi_t - \pi_t$, we follow the literature and have that the inflation gap follows a locally stationary AR(p) process with $\Pr \{ \lim_{h \to \infty} E_t (\pi_{t+h}) = 0 \} = 1$.

Stock and Watson (2007, SW hereafter) documents that when correctly specified a model featuring a time-varying trend-inflation is the best performing model for producing point forecasts. Given the prominence of the SW benchmark, it is worth discussing how this model is related to the score-driven model (5) that does not include autoregressive terms. In SW the conditional mean and the measurement error are driven by two independent shocks with stochastic volatility, so that the model implies that inflation follows a reduced form IMA(1,1) with time-varying MA coefficient and time-varying variance, where both parameters are driven by a convolution of the two independent stochastic volatilities. The observation-driven model also implies an IMA(1,1) which has time-varying variance but constant coefficient under Gaussian innovations. Yet, as was pointed out in Section 3, when the Student-t distribution is considered, the score-driven model produces an IMA(1,1) with both time-varying coefficients and time-varying variance. Cogley (2002) proposed a simple exponential smoothing which is equal to (22) under Gaussian distribution and constant variance. It is worth noticing that Cogley (2002) proposes a simple exponential smoothing which is nested in (22) under Gaussian distribution and constant variance.

In the application we allow for various specifications of (34). Specifically, we first consider a model with time-varying trend-only and then we allow for various specifications of the autoregressive parameters, i.e. $p = 1, 2, 4$. Chan et al. (2013) forcefully argue for imposing bounds on the long-run trend on the grounds that level too low (or too high) of inflation is inconsistent with the clear mandate of the central bank inflation stability. Therefore, for every specification we also include a counterpart derived with a bound (between 0 and 5) on the long-run trend.\textsuperscript{23} Furthermore, we consider all specification under Gaussian and Student-t innovations. Finally, note that partial autocorrelations are always bounded so to impose local stationarity of the model and the variance is reparametrized so that is always positive.\textsuperscript{24}

Table 1 reports the estimated for the various specifications for annualized quarterly CPI inflation $\pi_t = 400 \Delta \ln p_t$. Besides the estimates of the parameters and their associated standard error, we also report the value of the log likelihood function and the Akaike (AIC) and

\[ \text{Table 1} \]

\[ \text{Insert Table 1} \]

\textsuperscript{23}Those are the upper and lower bound in the posterior in Chen et al. (2013). They highlight that is difficult to identify exactly those bounds, but also that their results are not affected by those. In fact, they show that once the bounds are imposed to the autoregressive specification variations in the estimated long-run trend tends to be very limited. We also obtain a stable estimate for the long-run trend and this is typically not affected by the choice of the upper and lower bound.

\textsuperscript{24}In particular, the exponential link function is used so that $\sigma_t^2 = \exp(2\lambda_t)$ and the unrestricted time-varying parameter is $\lambda_t = \ln(\sigma_t)$.
Bayesian Information Criterion (BIC). A number of results neatly emerge from this table. First, perhaps not surprisingly, adding the autoregressive components, regardless the specification of the model, always shows a substantially smaller estimate of the smoothing parameter associated to the coefficient of the model. In contrast, the smoothing parameter associated to the variance equation is instead rather stable and typically higher than the one associated to the coefficients, lending support to the idea that changes in the variance are particularly relevant in our sample (see also Pivetta and Reis, 2007). Second, variations in the variances tends to be more pronounced (i.e. $\kappa_\nu$ is always higher) in the specifications with the Student-t innovations. In fact, with Student-t innovations the variance is less affected by the outliers and therefore can better adjust to accommodate changes in the dispersion of the central part of the distribution. Finally, the specifications with Student-t distribution always outperforms, and quite considerably, the ones with Gaussian innovation. This is the case both in terms of likelihood values and information criteria. Moreover, the estimated small number of degrees of freedom $v$ depicts a marked difference between the Gaussian and Student-t specification. This small number suggests that there might be pronounced variations at the quarterly frequency, either arising from measurement issues or related to the presence of rare events that structural macroeconomic should explicitly account for (as recently advocated by Curdia et al., 2013). Notice also that $v = 5$ has been considered to produce reasonable density forecasts from autoregressive models of inflation (see e.g. Corradi and Swanson, 2006).

It is also interesting to focus on the trend-only specification with Gaussian innovations. As shown in Section 3, this model implies that the trend is equal to an exponential smoothing of past observations, a model which has been used by Cogley (2002) as a measure of core inflation. The estimate of the smoothing parameter is fairly high, 0.3367, implying a an half life estimate of approximately 2 quarter. This estimate is substantially higher than the one used in Cogley (2002) which instead implies a much slower learning process (half life of 5.5 quarters).

5.1 Forecasting Evaluation

Overall the results in Table 1 do not suggest that one of the specification clearly dominates the others, as the ranking would depend on the chosen criterion. Furthermore, simply looking at the in-sample fit it is not clear whether bounding the long-run trend substantially improves the fit of the model. Therefore, we also assess the forecasting performance of the various specification. Specifically, we evaluate the forecasts over the period 1973Q1–2012Q4, with the model re-estimated recursively over an expanding window. Consistent with a long-standing tradition in the learning literature (referred to as anticipated-utility by Kreps, 1998), we update the coefficients period by period and then treat the updated values as if they would remain constant going forward in time. We first consider point forecast and use both root mean squared

\[25\] Notice that with respect to the model in Cogley (2002) the specification used as benchmark also allows for the time-variation in the variance. The latter does not affect the way the trend component is computed, nevertheless it is likely to affect the estimate of the smoothing parameter.
forecast error (RMSFE) and the absolute mean forecast error (MAFE). The specification with trend-only and Gaussian innovation is taken as the benchmark model, as this is the closest specification to the one of SW. Moreover, this model corresponds to an exponential smoothing of the inflation whose predictive virtues are highlighted by Cogley (2002).

Tables 2 report the results. Despite the well-know performance of the benchmark it is worth noticing that most of the alternative models tend to have lower RMSFE or MAFE. The gain become substantial at longer forecast horizons and for some cases this difference is also statistically significant as documented by the test of Giacomini and White (2006, GW hereafter). A comparison between the Gaussian and Student-t models does reveal a marginal improvement under Student-t innovations. Whereas when it comes to the number of the lags to be included, specifications with one or two lags typically tend to outperform the ones with more lags. Imposing the bounds enhances the performance of the AR specifications only marginally (in particular for models with one or two lags), and more markedly for the specification with Student-t innovations. In fact, the marginal improvement is such that the difference in performance with respect to the benchmark often becomes statistically significative. One of the specification that tends to be always outperformed by the alternative is the trend-bound, however it is important to stress that the relative performance of this specification is severely biased by the inclusions of the great inflation period (mid 70s-80s), as the model is bounded by construction to a maximum forecast at the upper bound of 5%.27

Table 2 reports the results. Despite the well-know performance of the benchmark it is worth noticing that most of the alternative models tend to have lower RMSFE or MAFE. The gain become substantial at longer forecast horizons and for some cases this difference is also statistically significant as documented by the test of Giacomini and White (2006, GW hereafter). A comparison between the Gaussian and Student-t models does reveal a marginal improvement under Student-t innovations. Whereas when it comes to the number of the lags to be included, specifications with one or two lags typically tend to outperform the ones with more lags. Imposing the bounds enhances the performance of the AR specifications only marginally (in particular for models with one or two lags), and more markedly for the specification with Student-t innovations. In fact, the marginal improvement is such that the difference in performance with respect to the benchmark often becomes statistically significative. One of the specification that tends to be always outperformed by the alternative is the trend-bound, however it is important to stress that the relative performance of this specification is severely biased by the inclusions of the great inflation period (mid 70s-80s), as the model is bounded by construction to a maximum forecast at the upper bound of 5%.27

Table 4 reports the results from a density forecast exercise where we focus on the one-step-ahead forecast. A comparison of the average log score reveals that the models with Student-t innovations substantially improves with respect to ones with Gaussian innovations, regardless of the model.28 Furthermore, the table report two test for the calibration of the densities. One is the LR test on the inverse transformation in the PITs (Berkowitz, 2001) and the other is the nonparametric test of Rossi and Sekhposyan (2013, RS hereafter). The latter test remains valid also in presence of parameter estimation error. Clearly, the specifications with Gaussian innovations proves to be not well calibrated.

26Despite the expanding window it is possible to apply the GW test as the model implicitly discount the observations, so that the earlier observations tend to have limited or no relevance to the estimates in the late part of the sample that are used for the forecast. To put it differently the model implies an implicitly discounting of the data so that the conditions for the GW test are met.

27When the exercise is repeated with forecasting window starting at 1984:Q1 also the trend-bound model becomes competitive in terms of point forecast, whereas no noticeable differences are present for the other specifications as reported in Table 3.

28Clark and Ravazzolo (2012) document the gains of allowing for fatter tails. However, they found much smaller improvement.
Figure 2 plots the empirical distribution of the PITs. In addition to the PITs, we also provide the 95% confidence interval (broken lines) using a normal approximation to a binomial distribution similar to Diebold et al. (1998).

Figure 3 displays the cumulative distribution of the PITs for each realization, under the null hypothesis the PITs should be uniformly distributed; thus the CDF of the PITs should be the 45° line. The figure also reports the critical values based on the RS test. If the empirical CDF of the PITs is outside the critical value lines, we conclude that the density forecast is misspecified.

From both figures it is evident that the models with Gaussian innovations tend to produce densities which are too tight, too many realizations fall in the middle of the forecast densities relative to what we would expect if the data were really Normally distributed.

In Table 5, for each pair of model we report the p-values for the test of difference in the average log predictive score, as outlined in Amisano and Giacomini (2007), using uniform weights. The results confirm that the substantial difference between the Normal and Student-t are indeed significant. At the same time the p-values confirm that the difference across the various specifications with Student-t innovations are instead never significative.

5.2 Estimates of Trend Inflation, Persistence, and Volatility

Figure 4 presents estimates of the long-run trend in inflation for the various specifications considered in this paper. It can be seen that large differences exist between the unbounded models and the others. The former are much more erratic and yield more extreme results than the latter. Conversely, if one excludes the trend-bound, which almost by construction is destined to reach the bounds during the great inflation period. The different autoregressive specifications imply a similar patter for the long-run trend when bounded. The trend in this case is consistent with the idea of a central bank anchoring the expectations of trend-inflation to a fairly stable level over the sample. In fact, imposing the bounds implies a qualitatively similar picture for the trend-inflation across the specifications. All specifications attribute the high inflationary periods to persistent deviations from the trend. Trend-inflation rises toward the late 60s to a value of around 3% and then decreases back to a slightly lower level in the late 80s early 90s and is anchored at approximately 2% in the last part of the sample. It is quite interesting to note that the pattern in the long-run trend is very similar to the one found by Chan et al. (2012), despite the fact that they use a different model’s specification and estimation techniques. When the long-run trend is unrestricted, the specifications with 2 or 4
lags tend to be somewhat instable.

[Insert Figure 4]

Given the analogy with the exponential smoothing often used as a proxy for the long-run trend, it is worth discussing the difference between the unrestricted trend-only specification under Normal and Student-t distribution. Both specifications track the behavior of inflation very closely, emphasizing the low frequency variations just as the standard exponential smoothing specification (see Cogley, 2002). The trend-inflation under Student-t innovation seem to be less affected by the high frequency variations of the inflation and is especially robust to the large spikes in the late part of the sample, yet this follows the inflation series more closely during the 70s when the changes are more persistent.

Moving to the analysis of the persistence in inflation, for $p > 1$ we follow Pivetta and Reis (2007) and compute both the sum of the AR coefficients and the largest root as proxy of the overall persistence; those are shown in Figure 5 and 6. Similar to Cogley and Sargent (2001), most of our specifications tend to suggest that the persistence of inflation in the US rose in the early part of the sample to reach the pick during the great inflation of the 1970s, before starting a gradual decline from mid to late 1980s. Yet it is also interesting to note that allowing for a larger number of lags tends to flatten up the persistency’s measure. This finding reconciles the different results obtained by Pivetta and Reis (2007) who focus on time-varying AR model with three lags. However, the results reported in the previous section suggest that models with a low, typically one, number of lags tend to generally outperform those with more lags.

[Insert Figure 5 and Figure 6]

Figure 7 reports measures of the change in volatility. Some interesting issues emerge. All specifications show that the variance of inflation was substantially high in the 50s, in the 70s and then again in the last decade. As in Chan et al. (2013), the trend-only specifications feature substantial difference between bound versus unrestricted trend. Clearly, the bounded specifications overstate the level of volatility in the period when the bound is binding. Interestingly, comparing Gaussian versus Student-t distribution one can see that they share similar low frequency variation, and the specifications with Student-t innovation display substantially more variation in the volatility. This latter result is particularly important in the light with the considerable evidence in favor of the Student-t specification reported in the previous subsection. In fact, most of the macroeconomic literature, who has manly focused on Gaussian distribution, has reported and emphasized the importance of the low frequency variation in the volatility. Furthermore, it is also worth mentioning that the measures based on the Student-t is also more robust to single outliers, indeed it is clear that under Gaussianity the volatility in the last part of the sample seems to be disproportionately affected by very few observations.

[Insert Figure 7]
6 Conclusion

In this paper we derive an adaptive algorithm for time-varying autoregressive models, both under Gaussianity and with heavy tails using a Student-t distribution. Following Creal et al. (2012) and Harvey (2013), the score of the conditional distribution is the driving process for the evolution of the parameters. This approach extends the least squares algorithms popularized by Ljung and Soderstrom (1985) - which are the building block of the learning expectation literature - to non-quadratic criterion functions. Furthermore, the algorithm is extended to incorporate restrictions which are popular in the empirical literature. Specifically, the model is allowed to have a bounded long-run mean and the coefficients are restricted so that the model is locally stationary. Specifically, the adaptive algorithm is extended to an environment with changes in volatility and non-Gaussian distribution. The latter extension robustifies the standard adaptive algorithms to the presence of tail events.

We apply the algorithm to the study of inflation dynamic. Various alternative specifications are shown to track the data very well, so that they give a parsimonious characterization of the inflation dynamic and at the same time produce good forecasts. Allowing for heavy tails is found to be a key ingredient to obtain well calibrated density forecasts over the analyzed sample. The empirical results confirm the well known pattern in inflation persistence documented in the literature and find support for the importance of incorporating theoretically consistent bounds in the long-run trend of inflation in line with Chan et al. (2013). When a model with Student-t innovations is considered we find more higher frequency variations in the volatility on top of the well documented low frequency changes. With respect to the more popular parameter-driven models the root taken in this paper allows for a direct interpretation of the time-varying parameters and their updating rules in terms of the observables.

The results of this paper can be extended along various directions. The convergence properties of the algorithm are to be explored, so that the algorithm could be directly applicable to study of the convergence to equilibrium under learning expectations (in an environment with changes in volatility or/and heavy tails). Furthermore, the model can be extended (along the lines of Koop and Korobilis, 2012) to the multivariate case where the dimensions of the model might be so large that the use of MCMC methods is infeasible and imposing the stationarity is problematic.
References


Table 1: Estimation of the various AR models for the US CPI 1955Q1-2012Q1. “Trend” denotes the time-varying mean-only and “B” denotes the restricted long-run mean.

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>Student-t</th>
<th></th>
<th>Normal</th>
<th>Student-t</th>
<th></th>
<th>Normal</th>
<th>Student-t</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Trend</td>
<td>Trend-B</td>
<td>AR(1)</td>
<td>Trend</td>
<td>Trend-B</td>
<td>AR(1)</td>
<td>Trend</td>
<td>Trend-B</td>
<td>AR(1)</td>
</tr>
<tr>
<td>(\kappa_c)</td>
<td>0.3367</td>
<td>0.3547</td>
<td>0.0407</td>
<td>0.0387</td>
<td>0.0479</td>
<td>0.0286</td>
<td>0.0325</td>
<td>0.0239</td>
<td>0.5415</td>
</tr>
<tr>
<td></td>
<td>(0.0480)</td>
<td>(0.0190)</td>
<td>(0.0048)</td>
<td>(0.0039)</td>
<td>(0.0027)</td>
<td>(0.0029)</td>
<td>(0.0028)</td>
<td>(0.0021)</td>
<td>(0.0115)</td>
</tr>
<tr>
<td>(\kappa_\sigma)</td>
<td>0.1479</td>
<td>0.1910</td>
<td>0.1127</td>
<td>0.1341</td>
<td>0.1044</td>
<td>0.1484</td>
<td>0.0806</td>
<td>0.1036</td>
<td>0.1479</td>
</tr>
<tr>
<td></td>
<td>(0.0277)</td>
<td>(0.0189)</td>
<td>(0.0180)</td>
<td>(0.0225)</td>
<td>(0.0128)</td>
<td>(0.0293)</td>
<td>(0.0064)</td>
<td>(0.0142)</td>
<td>(0.0189)</td>
</tr>
<tr>
<td>AIC</td>
<td>1131.9071</td>
<td>1154.7142</td>
<td>1102.1009</td>
<td>1080.1277</td>
<td>1121.2306</td>
<td>1083.8516</td>
<td>1127.9050</td>
<td>1094.4123</td>
<td>1146.1498</td>
</tr>
</tbody>
</table>
### Table 2: Point forecast. Forecast sample 1973Q1–2012Q4. RMSFE (root mean square forecast error), MAFE (mean absolute forecast error), “h” denotes the forecast horizon, in brackets the p-values for the GW test

<table>
<thead>
<tr>
<th></th>
<th>RMSFE</th>
<th></th>
<th>MAFE</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>h=1</td>
<td>h=4</td>
<td>h=8</td>
<td>h=1</td>
<td>h=4</td>
<td>h=8</td>
</tr>
<tr>
<td>Trend</td>
<td>2.2917</td>
<td>2.8929</td>
<td>3.3511</td>
<td>1.5294</td>
<td>2.1347</td>
<td>2.4657</td>
</tr>
<tr>
<td>Trend-B</td>
<td>1.2722</td>
<td>1.0300</td>
<td>0.8428</td>
<td>1.2628</td>
<td>0.9536</td>
<td>0.8123</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.0179)</td>
<td>(0.7780)</td>
<td>(0.2242)</td>
</tr>
<tr>
<td>AR(1)</td>
<td>0.9460</td>
<td>0.9062</td>
<td>0.8936</td>
<td>0.9807</td>
<td>0.9027</td>
<td>0.8608</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.3398)</td>
<td>(0.0820)</td>
<td>(0.2297)</td>
</tr>
<tr>
<td></td>
<td>AR(1)</td>
<td>0.9404</td>
<td>0.9042</td>
<td>0.9140</td>
<td>0.9658</td>
<td>0.8636</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.2949)</td>
<td>(0.0518)</td>
<td>(0.2936)</td>
</tr>
<tr>
<td></td>
<td>AR(2)</td>
<td>0.9687</td>
<td>0.9145</td>
<td>0.9204</td>
<td>1.0113</td>
<td>0.8717</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.2601)</td>
<td>(0.1495)</td>
<td>(0.4549)</td>
</tr>
<tr>
<td></td>
<td>AR(2)-B</td>
<td>0.9832</td>
<td>0.9308</td>
<td>0.9107</td>
<td>1.0130</td>
<td>0.8841</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.6893)</td>
<td>(0.2610)</td>
<td>(0.4697)</td>
</tr>
<tr>
<td>AR(4)</td>
<td>0.9733</td>
<td>0.9491</td>
<td>0.9340</td>
<td>1.0000</td>
<td>0.8897</td>
<td>0.8606</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.951)</td>
<td>(0.4949)</td>
<td>(0.6398)</td>
</tr>
<tr>
<td></td>
<td>AR(4)-B</td>
<td>0.9611</td>
<td>1.0002</td>
<td>0.9705</td>
<td>0.9880</td>
<td>0.9359</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.2330)</td>
<td>(0.9977)</td>
<td>(0.8420)</td>
</tr>
</tbody>
</table>

**Student-t**

<table>
<thead>
<tr>
<th></th>
<th>RMSFE</th>
<th></th>
<th>MAFE</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>h=1</td>
<td>h=4</td>
<td>h=8</td>
<td>h=1</td>
<td>h=4</td>
<td>h=8</td>
</tr>
<tr>
<td>Trend</td>
<td>0.9879</td>
<td>0.9824</td>
<td>0.9803</td>
<td>0.9570</td>
<td>0.9680</td>
<td>0.9497</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.7942)</td>
<td>(0.4422)</td>
<td>(0.6184)</td>
</tr>
<tr>
<td>Trend-B</td>
<td>1.3203</td>
<td>1.0492</td>
<td>0.8597</td>
<td>1.3071</td>
<td>0.9854</td>
<td>0.8270</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.0051)</td>
<td>(0.6370)</td>
<td>(0.2852)</td>
</tr>
<tr>
<td>AR(1)</td>
<td>0.9175</td>
<td>0.9058</td>
<td>0.8924</td>
<td>0.9597</td>
<td>0.8903</td>
<td>0.8447</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.0733)</td>
<td>(0.0912)</td>
<td>(0.2432)</td>
</tr>
<tr>
<td></td>
<td>AR(1)-B</td>
<td>0.9128</td>
<td>0.8972</td>
<td>0.9004</td>
<td>0.9349</td>
<td>0.8759</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.1171)</td>
<td>(0.0307)</td>
<td>(0.1571)</td>
</tr>
<tr>
<td>AR(2)</td>
<td>0.9821</td>
<td>0.9792</td>
<td>0.9885</td>
<td>1.0056</td>
<td>0.9367</td>
<td>0.9143</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.5159)</td>
<td>(0.6933)</td>
<td>(0.9057)</td>
</tr>
<tr>
<td></td>
<td>AR(2)-B</td>
<td>0.9201</td>
<td>0.8551</td>
<td>0.8557</td>
<td>0.9580</td>
<td>0.8286</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.1210)</td>
<td>(0.0013)</td>
<td>(0.0853)</td>
</tr>
<tr>
<td>AR(4)</td>
<td>0.9862</td>
<td>1.0117</td>
<td>0.9462</td>
<td>1.0168</td>
<td>0.9378</td>
<td>0.8720</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.7415)</td>
<td>(0.8912)</td>
<td>(0.6869)</td>
</tr>
<tr>
<td>AR(4)-B</td>
<td>0.9858</td>
<td>1.0727</td>
<td>1.0037</td>
<td>1.0036</td>
<td>0.9667</td>
<td>0.9190</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.7420)</td>
<td>(0.5114)</td>
<td>(0.9777)</td>
</tr>
</tbody>
</table>
Table 3: Point forecast. Forecast sample 1984Q1–2012Q4. RMSFE (root mean square forecast error), MAFE (mean absolute forecast error), “h” denotes the forecast horizon. in brackets the p-values for the GW test

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSFE</th>
<th></th>
<th></th>
<th>MAFE</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>h=1</td>
<td>h=4</td>
<td>h=8</td>
<td>h=1</td>
<td>h=4</td>
<td>h=8</td>
</tr>
<tr>
<td>Normal</td>
<td>2.2259</td>
<td>2.4902</td>
<td>2.4748</td>
<td>1.3787</td>
<td>1.6961</td>
<td>1.8652</td>
</tr>
<tr>
<td>Trend-B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.0314)</td>
<td>(0.0071)</td>
<td>(0.1668)</td>
<td>(0.0677)</td>
<td>(0.0396)</td>
<td>(0.0649)</td>
<td></td>
</tr>
<tr>
<td>AR(1)</td>
<td>0.9294</td>
<td>0.8547</td>
<td>0.9097</td>
<td>0.9648</td>
<td>0.8921</td>
<td>0.8703</td>
</tr>
<tr>
<td>(0.3795)</td>
<td>(0.0423)</td>
<td>(0.4582)</td>
<td>(0.6312)</td>
<td>(0.1934)</td>
<td>(0.2954)</td>
<td></td>
</tr>
<tr>
<td>AR(1)</td>
<td>0.9131</td>
<td>0.8052</td>
<td>0.8137</td>
<td>0.9413</td>
<td>0.7723</td>
<td>0.7381</td>
</tr>
<tr>
<td>(0.2868)</td>
<td>(0.0048)</td>
<td>(0.1117)</td>
<td>(0.4127)</td>
<td>(0.0014)</td>
<td>(0.0152)</td>
<td></td>
</tr>
<tr>
<td>AR(2)</td>
<td>0.9446</td>
<td>0.8224</td>
<td>0.7996</td>
<td>0.9629</td>
<td>0.8040</td>
<td>0.7589</td>
</tr>
<tr>
<td>(0.1138)</td>
<td>(0.0040)</td>
<td>(0.0613)</td>
<td>(0.4258)</td>
<td>(0.0018)</td>
<td>(0.0142)</td>
<td></td>
</tr>
<tr>
<td>AR(2)-B</td>
<td>0.9603</td>
<td>0.8426</td>
<td>0.7949</td>
<td>0.9535</td>
<td>0.8031</td>
<td>0.7388</td>
</tr>
<tr>
<td>(0.4839)</td>
<td>(0.0037)</td>
<td>(0.0620)</td>
<td>(0.4545)</td>
<td>(0.0013)</td>
<td>(0.0113)</td>
<td></td>
</tr>
<tr>
<td>AR(4)</td>
<td>0.9627</td>
<td>0.8466</td>
<td>0.8116</td>
<td>0.9368</td>
<td>0.8147</td>
<td>0.7536</td>
</tr>
<tr>
<td>(0.4609)</td>
<td>(0.0054)</td>
<td>(0.0539)</td>
<td>(0.2372)</td>
<td>(0.0054)</td>
<td>(0.0073)</td>
<td></td>
</tr>
<tr>
<td>AR(4)-B</td>
<td>0.9307</td>
<td>0.8562</td>
<td>0.8095</td>
<td>0.9042</td>
<td>0.8319</td>
<td>0.7599</td>
</tr>
<tr>
<td>(0.0745)</td>
<td>(0.0053)</td>
<td>(0.0650)</td>
<td>(0.0818)</td>
<td>(0.0065)</td>
<td>(0.0114)</td>
<td></td>
</tr>
<tr>
<td>Student-t</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trend</td>
<td>0.9687</td>
<td>0.9345</td>
<td>0.9174</td>
<td>0.9331</td>
<td>0.9290</td>
<td>0.8977</td>
</tr>
<tr>
<td>(0.6436)</td>
<td>(0.0742)</td>
<td>(0.3665)</td>
<td>(0.3487)</td>
<td>(0.0999)</td>
<td>(0.2410)</td>
<td></td>
</tr>
<tr>
<td>Trend-B</td>
<td>0.9276</td>
<td>0.8849</td>
<td>0.9056</td>
<td>0.9338</td>
<td>0.9064</td>
<td>0.8650</td>
</tr>
<tr>
<td>(0.2899)</td>
<td>(0.0306)</td>
<td>(0.3999)</td>
<td>(0.4150)</td>
<td>(0.1929)</td>
<td>(0.2106)</td>
<td></td>
</tr>
<tr>
<td>AR(1)</td>
<td>0.8846</td>
<td>0.8422</td>
<td>0.8971</td>
<td>0.9288</td>
<td>0.8657</td>
<td>0.8367</td>
</tr>
<tr>
<td>(0.0729)</td>
<td>(0.0238)</td>
<td>(0.4062)</td>
<td>(0.2997)</td>
<td>(0.1379)</td>
<td>(0.1853)</td>
<td></td>
</tr>
<tr>
<td>AR(1)-B</td>
<td>0.8722</td>
<td>0.8108</td>
<td>0.8387</td>
<td>0.8943</td>
<td>0.8056</td>
<td>0.7684</td>
</tr>
<tr>
<td>(0.1122)</td>
<td>(0.0068)</td>
<td>(0.1779)</td>
<td>(0.1587)</td>
<td>(0.0099)</td>
<td>(0.0372)</td>
<td></td>
</tr>
<tr>
<td>AR(2)</td>
<td>0.9451</td>
<td>0.8893</td>
<td>0.9293</td>
<td>0.9442</td>
<td>0.8584</td>
<td>0.8482</td>
</tr>
<tr>
<td>(0.1081)</td>
<td>(0.0174)</td>
<td>(0.5327)</td>
<td>(0.2302)</td>
<td>(0.0125)</td>
<td>(0.1081)</td>
<td></td>
</tr>
<tr>
<td>AR(2)-B</td>
<td>0.8712</td>
<td>0.7970</td>
<td>0.8079</td>
<td>0.9147</td>
<td>0.7835</td>
<td>0.7468</td>
</tr>
<tr>
<td>(0.0813)</td>
<td>(0.0035)</td>
<td>(0.1154)</td>
<td>(0.1790)</td>
<td>(0.0021)</td>
<td>(0.0238)</td>
<td></td>
</tr>
<tr>
<td>AR(4)</td>
<td>0.9435</td>
<td>0.8429</td>
<td>0.8398</td>
<td>0.9369</td>
<td>0.8355</td>
<td>0.7815</td>
</tr>
<tr>
<td>(0.1996)</td>
<td>(0.0104)</td>
<td>(0.1052)</td>
<td>(0.2771)</td>
<td>(0.0322)</td>
<td>(0.0088)</td>
<td></td>
</tr>
<tr>
<td>AR(4)-B</td>
<td>0.9413</td>
<td>0.8480</td>
<td>0.8270</td>
<td>0.9239</td>
<td>0.8258</td>
<td>0.7733</td>
</tr>
<tr>
<td>(0.1603)</td>
<td>(0.0092)</td>
<td>(0.0634)</td>
<td>(0.1878)</td>
<td>(0.0173)</td>
<td>(0.0053)</td>
<td></td>
</tr>
</tbody>
</table>
Table 4: Density Forecast. Forecast sample 1973Q1-2012Q4. “LR” denotes the p-values associated to the Likelihood Ratio (LR) test of Berkowitz (2001), $\kappa_{\alpha,P}^{CS}$, corresponds to Rossi and Sekhposyan (2013) with critical values 2.25 (1%), 1.51 (5%), 1.1 (10%)

<table>
<thead>
<tr>
<th>Model</th>
<th>Normal Av Log Score</th>
<th>LR</th>
<th>$\kappa_{\alpha,P}^{CS}$</th>
<th>Student-t Av Log Score</th>
<th>LR</th>
<th>$\kappa_{\alpha,P}^{CS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trend</td>
<td>-2.8237</td>
<td>0.0001</td>
<td>5.7760</td>
<td>-1.5999</td>
<td>0.5694</td>
<td>0.9923</td>
</tr>
<tr>
<td>Trend-B</td>
<td>-3.0188</td>
<td>0.0001</td>
<td>6.6422</td>
<td>-1.6353</td>
<td>0.0124</td>
<td>1.5210</td>
</tr>
<tr>
<td>AR(1)</td>
<td>-2.7127</td>
<td>0.0055</td>
<td>4.0960</td>
<td>-1.6065</td>
<td>0.6715</td>
<td>0.1322</td>
</tr>
<tr>
<td>AR(1)-B</td>
<td>-2.6537</td>
<td>0.3831</td>
<td>4.7610</td>
<td>-1.6223</td>
<td>0.5172</td>
<td>0.5760</td>
</tr>
<tr>
<td>AR(2)</td>
<td>-2.7784</td>
<td>0.0129</td>
<td>4.7610</td>
<td>-1.6145</td>
<td>0.1988</td>
<td>1.1560</td>
</tr>
<tr>
<td>AR(2)-B</td>
<td>-2.6932</td>
<td>0.0121</td>
<td>4.7610</td>
<td>-1.6146</td>
<td>0.3501</td>
<td>0.4623</td>
</tr>
<tr>
<td>AR(2)</td>
<td>-2.9495</td>
<td>0.1794</td>
<td>4.4223</td>
<td>-1.6313</td>
<td>0.2424</td>
<td>0.9923</td>
</tr>
<tr>
<td>AR(4)-B</td>
<td>-2.7859</td>
<td>0.0822</td>
<td>4.0960</td>
<td>-1.6603</td>
<td>0.1826</td>
<td>0.9923</td>
</tr>
</tbody>
</table>
Table 5: Pairwise comparison for difference in the average log score. Forecasting sample 1973Q1–2012Q4, p-values of Amisano and Giacomini (2007) test computed with uniform weights.

<table>
<thead>
<tr>
<th>Normal</th>
<th>Trend</th>
<th>Trend-B</th>
<th>AR(1)</th>
<th>AR(1)-B</th>
<th>AR(2)</th>
<th>AR(2)-B</th>
<th>AR(4)</th>
<th>AR(4)-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trend-B</td>
<td>0.003</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR(1)</td>
<td>0.018</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR(1)-B</td>
<td>0.033</td>
<td>0.000</td>
<td>0.167</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR(2)</td>
<td>0.158</td>
<td>0.000</td>
<td>0.183</td>
<td>0.101</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR(2)-B</td>
<td>0.024</td>
<td>0.000</td>
<td>0.741</td>
<td>0.605</td>
<td>0.025</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR(4)</td>
<td>0.077</td>
<td>0.460</td>
<td>0.008</td>
<td>0.010</td>
<td>0.021</td>
<td>0.008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR(4)-B</td>
<td>0.548</td>
<td>0.003</td>
<td>0.298</td>
<td>0.148</td>
<td>0.893</td>
<td>0.141</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>Student-t</td>
<td>Trend</td>
<td>Trend-B</td>
<td>AR(1)</td>
<td>AR(1)-B</td>
<td>AR(2)</td>
<td>AR(2)-B</td>
<td>AR(4)</td>
<td>AR(4)-B</td>
</tr>
<tr>
<td>Trend</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Trend-B</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>AR(1)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>AR(1)-B</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>AR(2)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>AR(2)-B</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>AR(4)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>AR(4)-B</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Figure 1: Blue (thick) line $\nu = 5$, pink (dash) $\nu = 10$, green (dots) $\nu = 30$ and black (thin) $\nu = \infty$ (Gaussian), $w_t$ are the weights and $\sqrt{\xi_t} = \varepsilon_t/\sigma_{t|t-1}$ are the standardized prediction error.
Figure 2: The Figures show the pdf of the PITs (normalized) and the 95% critical values approximated under Diebold et al. (1998)’s binomial distribution (dashed lines), constructed using a normal approximation.
Figure 3: The plots show the empirical CDF of the PITs (solid line), the CDF of the PITs under the null hypothesis (the 45 degree line) and the critical values based KCS test of Rossi and Sekhposyan (2013).
Figure 4: Implied trend-inflation. Thin (dark) line is the actual inflation. Thick (blue) unrestricted trend, light (green) bounded trend. The plots also include the 68% confidence interval accounting for parameters uncertainty (Hamilton, 1986).
Figure 5: Sum of ARs coefficients. Dark (blue) line long-run trend, light (green) bounded long-run trend. Figures also include 68% confidence interval accounting for parameters uncertainty (Hamilton, 1986).
Figure 6: Largest Root. Dark (blue) line long-run trend, light (green) line bounded long-run trend. Figures also include 68% confidence interval accounting for parameters uncertainty (Hamilton, 1986).
Figure 7: Volatilities computed as $\ln(\sigma_{t|t-1})$. 
Appendix

Lemma 1 Following Ljung and Soderstrom (1985, section 2.6.2), the recursive estimation of the CGL can be obtained from an off-line identification approach that minimizes the weighted sum of squared errors

$$S_t(\phi_t) = \sum_{j=1}^{t} \gamma_j (y_{t-j} - x'_{t-j}\phi_t)^2,$$

where $\gamma_j = \prod_{k=j+1}^{t} \delta_k$ is a sequence of weights assign to the observation $y_{t-j}$. Setting $\delta = (1-\kappa)$, where $\delta \leq 1$ is known as the forgetting factor, the observations are weighted exponentially, i.e. $\gamma_j = (1-\kappa)^j$, and the gain parameter is equal to $\left[\sum_{j=1}^{t} \gamma_j\right]^{-1} \rightarrow \kappa$.

Thus, the CGL can be seen as a recursive estimation of the discounted least squares and it generalizes the exponential smoothing of Hyndman et al (2008) when explanatory variables are included. Under time-varying parameters model the constant gain $\kappa$ regulates the tracking ability (large $\kappa$) and the noise insensitivity (small $\kappa$). On the other hand, for $\kappa = 1/t$ we obtain the recursive least squares and the parameters variation vanishes asymptotically.

Lemma 2 Ljung (1992, p. 99) and Sargent (1999, p. 115) show how to obtain the CGL algorithm from the KF applied to the restricted state space model. It is worth to show that the restrictions imply that $\eta_t = c(\phi_{t|t} - \phi_t)$, where $c = [\kappa/(1-\kappa)]^{1/2}$. Consequently, the transition equation in (14) is equal to $\phi_{t+1} = (1-c)\phi_t + c\phi_{t|t}$ and the true state vector can be expressed as exponential weighted average of past filter estimates

$$\phi_{t+1} = c \sum_{j=0}^{t-1} (1-c)^j \phi_{t-j|t-j}.$$ 

Moreover, the filter estimate can be expressed as

$$\phi_{t|t} = L_t \phi_{t-1|t-1} + K_t y_t = \sum_{j=0}^{t-1} \left( \prod_{i=0}^{j-1} L_{t-i} \right) K_{t-j} y_{t-j}$$

where

$$L_t = (I - K_t x'_t), \quad K_t = P_{t|t-1} x_t \left( x'_t P_{t|t-1} x_t + \frac{\sigma^2}{1-\kappa} \right)^{-1}.$$

Thus, differently from the parameter-driven model, the Kalman gain does not depend on any unobserved shock and it rather obtained from past observations only. Therefore, those restrictions leads to have time-varying coefficients that are driven by past observations only.

Lemma 3 Setting $Q_t := \kappa^2 \Sigma$, with $\Sigma = \sigma^2 E[(x_i x'_i)^{-1}$, we have that the shock driving the time-
varying coefficients is
\[ \eta_t = \kappa(x_t'x_t')^{-1}x_t'\varepsilon_t = \kappa(x_t'x_t')^{-1}x_t(y_t - x_t'\phi_{t|t-1}). \]

Therefore, the parameter-driven model collapses to an observation-driven model. Moreover, up to a scalar factor, the shock \( \eta_t \) is equal to the driving process of our score driven model. However, under the parameter-driven framework the vector of coefficients is considered as an unobserved state vector which is optimally estimated by the mean of KF which leads to
\[
\begin{align*}
\phi_{t+1|t} &= \phi_{t|t-1} + P_{t|t-1}x_t(x_t'P_{t|t-1}x_t + \sigma^2)^{-1}(y_t - x_t'\phi_{t|t-1}) \\
P_{t+1|t} &= P_{t|t-1} - P_{t|t-1}x_t(x_t'P_{t|t-1}x_t + \sigma^2)^{-1}x_t'P_{t|t-1} + \kappa^2\Sigma.
\end{align*}
\]

Following Benveniste et al (1990, p. 139), for \( \kappa^2 \ll \sigma^2 \) meaning that the variance drifting parameters is much smaller than the variance model disturbances, for \( t > \hat{t} \), where \( \hat{t} \) is a given large value of \( t \), one has the approximation \( (x_t'P_{t|t-1}x_t + \sigma^2) \approx \sigma^2 \), this implies that the conditional variance of the forecast error converges to the variance of model disturbances. For \( t \) large enough, the variation of \( P_{t|t-1} \) is small with respect to \( x_t \), and \( x_t'P_{t|t-1}x_t \) can be neglected with respect to \( \sigma^2 \). Using these approximations, we obtain
\[
\begin{align*}
\phi_{t+1|t} &= \phi_{t|t-1} + P_{t|t-1}x_t\sigma^{-2}(y_t - x_t'\phi_{t|t-1}) \\
P_{t+1|t} &= P_{t|t-1} - P_{t|t-1}x_t\sigma^{-2}x_t'P_{t|t-1} + \kappa^2\Sigma.
\end{align*}
\]

Replacing \( x_t'x_t'/\sigma^2 \) with its expected value \( \Sigma^{-1} \) we obtain \( P_{t+1|t} = P_{t|t-1} - P_{t|t-1}\Sigma^{-1}P_{t|t-1} + \kappa^2\Sigma \). When \( P_{t|t-1} \) is set to its steady-state value \( P \) as in Harvey (1989, p. 118), one has \( \Sigma^{-1}P = \Lambda\Sigma\Lambda \Rightarrow \kappa^2\Sigma^{-1}P = \Sigma \Rightarrow \kappa^{-1}P = \Sigma \). Using last expression the recursion for the vector of coefficients is
\[
\phi_{t+1|t} = \phi_{t|t-1} + \kappa\Sigma x_t\sigma^{-2}(y_t - x_t'\phi_{t|t-1}),
\]
which has the same asymptotic behavior of the CGL; see Sargent and William (2005) and Evans et al (2010). Similarly, setting \( Q_t := \kappa^2\Sigma^{-1} \), we have that \( \eta_t = \kappa x_t\varepsilon_t \) and the parameter-driven model collapses to an observation-driven model. In the steady-state \( \kappa^{-1}P = I \) and the recursion for the coefficients is
\[
\phi_{t+1|t} = \phi_{t|t-1} + \kappa x_t\sigma^{-2}(y_t - x_t'\phi_{t|t-1}).
\]
which is a score based algorithm without the use of scaling matrix.

**Scaled Score under Student-t distribution** We re-write the predictive log-likelihood (16) as follows
\[ \ell_t(F_t, \theta) = c + d_t + g_t \]
Following Fiorentini et al. (2003), recalling that 
\[ \varepsilon_i = \frac{\epsilon_i^2}{\sigma^2_{t|t-1}} \]
and 
\[ \eta = \log \left( \frac{1}{2} \right), \quad \eta = \log \left( \frac{1 - 2\eta}{\eta} \right), \quad \frac{1}{2} \log \pi, \]
where \( \varepsilon_i \) is uniformly distributed on the unit set, \( \eta \) is a chi-squared
\[ g_t = -\left( \frac{\eta + 1}{2\eta} \right) \log \left[ 1 + \frac{\eta}{1 - 2\eta} \zeta_t \right], \]
where \( \zeta_t \) is uniformly distributed on the unit set, \( \eta \) is a chi-squared
\[ d_t = -\frac{1}{2} \log \sigma^2_{t|t-1}, \quad g_t = -\left( \frac{\eta + 1}{2\eta} \right) \log \left[ 1 + \frac{\eta}{1 - 2\eta} \zeta_t \right], \]
where \( \zeta_t \) is uniformly distributed on the unit set, \( \eta \) is a chi-squared
\[ \text{The score for the coefficients of the model is then equal to} \]
\[ \nabla_{\phi} = \left( \frac{d_t}{d_{t_t}} \right) \cdot \left( \frac{g_t}{g_{t_t}} \right) \cdot \left( \frac{\zeta_t}{\zeta_{t_t}} \right), \]
\[ \text{The gradient for the variance component is} \]
\[ \nabla_{\sigma} = \left( \frac{d_t}{d_{t_t}} \right) \cdot \left( \frac{g_t}{g_{t_t}} \right) \cdot \left( \frac{\zeta_t}{\zeta_{t_t}} \right), \]
\[ \text{where} \ \frac{\partial \zeta_t}{\partial \sigma_{t|t-1}} = -\frac{\varepsilon_i^2}{\sigma^2_{t|t-1}} \]
and thus we obtain
\[ \nabla_{\sigma} = \left( \frac{d_t}{d_{t_t}} \right) \cdot \left( \frac{g_t}{g_{t_t}} \right) \cdot \left( \frac{\zeta_t}{\zeta_{t_t}} \right), \]
\[ \text{We compute the information matrix as} \ I_t = -E_t(H_t), \text{where} \ H_t, \text{the Hessian matrix and} \]
\[ \text{it can be partitioned in four blocks} \]
\[ H_t = \begin{bmatrix} H_{\phi\phi,t} & H_{\phi\sigma,t} \\ H_{\sigma\phi,t} & H_{\sigma\sigma,t} \end{bmatrix}. \]
\[ \text{The first block} \ H_{\phi\phi,t} \text{can be calculated as} \]
\[ H_{\phi\phi,t} = \left( \frac{1 + \eta}{(1 + 2\eta - 1)} \right) \cdot \left( \frac{x_t x_t^t}{\sigma^2_{t|t-1}} \right). \]
random variable with 1 degree of freedom, $\xi_t$ is a gamma variate with mean $\nu > 2$ variance $2\nu$, and $u_t$, $\zeta_t$ and $\xi_t$ are mutually independent. Therefore, it is possible to show that

$$I_{\phi\phi,t} = -E(H_{\phi\phi,t}) = \frac{(1 + \eta)}{(1 - 2\eta)(1 + 3\eta)} x_t x_t'. $$

The cross-derivative term in the Hessian is $H_{\sigma\sigma,t} = -\frac{x_t x_t'}{\sigma_{t|t-1}^2}$ and therefore $I_{\phi\sigma,t} = -E(H_{\phi\sigma,t}) = 0$.

$$H_{\sigma\sigma,t} = \frac{\partial^2 \ell_t}{\partial \sigma_{t|t-1}^2} = \frac{\partial \nabla_{\sigma}}{\partial \sigma_{t|t-1}^2} = \frac{1}{2\sigma_{t|t-1}^4} \left[ \frac{2(1 - 2\eta) + \eta \varepsilon_t^2 / \sigma_{t|t-1}^2}{2[1 - 2\eta + \eta \varepsilon_t^2 / \sigma_{t|t-1}^2]} \right].$$

it is possible to show that

$$I_{\sigma\sigma,t} = -E_t(H_{\sigma\sigma,t}) = \frac{(1 + \eta)}{2(3 + \eta)} \sigma_{t|t-1}^4 - \frac{\eta}{2(3 + \eta)} \sigma_{t|t-1}^4 = \frac{1}{2(1 + 3\eta)} \sigma_{t|t-1}^4.$$

Finally, the information matrix is equal to

$$I_t = \begin{bmatrix} \frac{(1 + \eta)}{(1 - 2\eta)(1 + 3\eta)} x_t \sigma_{t|t-1}^{-2} x_t' & \mathbf{0} \\ \frac{1}{2(1 + 3\eta)} \sigma_{t|t-1}^{-2} & \mathbf{0} \end{bmatrix}.$$

and the final expression for the scaled score vector is

$$s_t = I_t^{-1} \nabla_t = \begin{bmatrix} s_{\phi t} \\ s_{\sigma t} \end{bmatrix} = \begin{bmatrix} \frac{(1 - 2\eta)(1 + 3\eta)}{(1 - 2\eta + \eta \varepsilon_t^2)} (x_t \sigma_{t|t-1}^{-2} x_t')^{-1} x_t \sigma_{t|t-1}^{-2} \varepsilon_t \\ (1 + 3\eta) \left[ \frac{(1 + \eta)}{(1 - 2\eta + \eta \varepsilon_t^2)} \varepsilon_t^2 - \sigma_{t|t-1}^2 \right] \end{bmatrix}.$$

**Proposition 5** Under Student-t distribution the driving process is (17)-(19) and the coefficients’ updating rule is

$$\phi_{t+1|t} = \phi_{t|t-1} + \kappa_\phi \frac{(1 - 2\eta)(1 + 3\eta)}{(1 + \eta)} (x_t \sigma_{t|t-1}^{-2} x_t')^{-1} x_t \sigma_{t|t-1}^{-2} [w_t(y_t - x_t' \phi_{t|t-1})],$$

and smoothing the scaling matrix (incorporation $w_t$) we obtain (20). If we consider the example with time varying mean only, we have that

$$y_t = \mu_{t|t-1} + \varepsilon_t, \quad \varepsilon_t \sim \nu(0, \sigma_{t|t-1}^2)$$
and the estimated level is

\[ \mu_{t+1|t} = \mu_{t|t-1} + \kappa_{\theta}w_t(\gamma_t - \mu_{t|t-1}) \]
\[ = (1 - \kappa_{\theta}w_t)\mu_{t|t-1} + \kappa_{\theta}w_t \gamma_t \]
\[ = \frac{\kappa_{\theta}}{1 - \kappa_{\theta}w_t} w_t \gamma_t \]
\[ = \kappa_{\theta} \sum_{j=0}^{\infty} \gamma_j w_{t-j} \gamma_{t-j}, \]

with \( \kappa_{\theta} = \frac{1}{1 + \kappa_{\theta}} \). After a bit of algebra, we can obtain explicit expression the weights across time that is

\[ \gamma_0 = 1 \text{ and } \gamma_j = \prod_{k=t-j+1}^{t} (1 - \kappa_{\theta}w_k). \]

The same weighting pattern is obtained when regressors are included. Since the weights across time are affected by the cross sectional weights \( w_t \), we can not obtained the robust filter (20) as solution of a re-weighted quadratic criterion function as Ljung and Sostrestrom (1985, sec. 2.6.2). In general, when we depart from Gaussianity the stochastic Newton-Gradient algorithm cannot be obtained as a recursive solution of a quadratic criterion function. For the variance is straightforward to obtain (21) and the implied weighting pattern.

**Proposition 7** Given the non-linear state space model

\[ y_t = x_t^0 g(\alpha_t) + \varepsilon_t, \varepsilon_t \sim N(0, \sigma^2), \]
\[ \phi_{t+1} = \phi_t + \eta_t, \eta_t \sim N(0, Q_t), \]

we can solve it by the mean of the Extended Kalman filter

\[ v_t = y_t - x_t^0 \phi_{t|t-1}, \]
\[ K_t = P_{t|t-1} x_t F_t^{-1}, \]
\[ F_t = x_t^0 P_{t|t-1} x_t + \sigma^2 \]
\[ \alpha_{t+1|t} = \alpha_{t|t-1} + K_t v_t, \]
\[ P_{t+1|t} = P_{t|t-1} - P_{t|t-1} x_t F_t^{-1} x_t^0 P_{t|t-1} + Q_t, \]

where \( x_t^0 = x_t^0 \frac{\partial g(\alpha_t)}{\partial \alpha}, \alpha = \alpha_{t|t-1} = x_t^0 \Psi_t \). Setting \( \sigma^2 = \frac{\sigma^2}{1 - \kappa} \) and \( Q_t = P_{t|t-1} \frac{\kappa}{1 - \kappa} \) and following same approch in Ljung (1992, p. 99) and Sargent (1999, p. 115), we obtain the modified version of the CGL algorithm

\[ \phi_{t+1} = \phi_{t|t-1} + \kappa R_{t-1}^{-1} \Psi_t x_t \sigma^{-2} (y_t - x_t^0 \phi_{t|t-1}), \]
\[ R_t = (1 - \kappa)R_{t-1} + \kappa (\Psi_t x_t \sigma^{-2} x_t^0 \Psi_t). \]
This is exactly the algorithm implied by the the score-driven model (30) with the information matrix \( \Psi_t x_t \sigma^{-2} x_t' \Psi_t \) is replaced by its smoothed version.

**Theorem 9** For semplicity we drop the temporal subscript \( t \) such that the \( p \times p \) Jacobian matrix is

\[
\Gamma = \frac{\partial \Phi}{\partial \pi'} = \frac{\partial \Phi(\pi)}{\partial \pi'}.
\]

The first \((p - 1)\) coefficients are obtained from last recursion in (25), and the last coefficients is equal to the last partial autocorrelation \( \pi_p \). We denote the final vector of coefficients as \( \Phi_p = (\phi_1^p, ..., \phi_{p-1}^p, \phi_p^p)' = (a'_p, \pi_p) \), where \( a_p = (\phi_1^p, ..., \phi_{p-1}^p) \) and \( \phi_p^p = \pi_p \).

Therefore, we can express the last iteration of (25) in matrix form \( a_p = J_{p-1} \Phi_{p-1} \); where \( \Phi_{p-1} = (\phi_1^{p-1}, ..., \phi_{p-2}^{p-1}, \phi_{p-1}^{p-1})' = (a_{p-1}', \pi_{p-1})' \) and

\[
J_{p-1} = \begin{bmatrix}
1 & 0 & \cdots & 0 & -\pi_p \\
0 & \ddots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
0 & \ddots & 0 \\
-\pi_p & 0 & \cdots & 0 & 1
\end{bmatrix}
\]

Note that if \( p \) is even the central element of \( J_{p-1} \) is \( 1 - \pi_p \). Moreover, the vector \( \tilde{\Phi}_p = (\phi_{p-1}', \pi_p)' \) contains all the partial autocorrelations, i.e. \( \tilde{\Phi}_p = (a_{p-1}', \pi_{p-1}, \pi_p) \) and keep substituting we obtain \( \tilde{\Phi}_p = \pi_p = (\pi_1, ..., \pi_{p-1}, \pi_p) \). The Jacobian matrix can be expressed as follows

\[
\Gamma = \Gamma_p = \begin{bmatrix}
\frac{\partial a_p}{\partial \phi_{p-1}} & \frac{\partial a_p}{\partial \pi_p} \\
\frac{\partial a_p}{\partial \phi_{p-1}} & \frac{\partial a_p}{\partial \pi_p} \\
\frac{\partial a_p}{\partial \phi_{p-1}} & \frac{\partial a_p}{\partial \pi_p}
\end{bmatrix}
\]

The upper-left block is a \((p - 1) \times (p - 1)\) matrix and it can be computed using the definition \( a_p = J_{p-1} \Phi_{p-1} \); since \( J_{p-1} \) contains the last partial correlation \( \pi_p \) we have the recursive formulation

\[
\frac{\partial a_p}{\partial \phi_{p-1}} = J_{p-1} \Gamma_{p-1}
\]

where \( \Gamma_{p-1} = \partial \phi_{p-1}/\partial \pi_{p-1} \) is the Jacobian of the first \( p - 1 \) coefficients with respect to the first \( p - 1 \) partial autocorrelations. Finally, we have that the other three blocks are

\[
\frac{\partial \pi_p}{\partial a_{p-1}'} = 0', \quad \frac{\partial \pi_p}{\partial \pi_{p-1}} = 1 \quad \text{and} \quad \frac{\partial a_p}{\partial \pi_p} = \frac{\partial J_{p-1}}{\partial \pi_p} \phi_{p-1} = \begin{bmatrix}
-\phi_1^{p-1, p-1} \\
-\phi_2^{p-2, p-1} \\
\vdots \\
-\phi_1^{1, p-1}
\end{bmatrix}
\]

Note that \( \phi_{p-1} \) is a given and \( \frac{\partial J_{p-1}}{\partial \pi_p} = \text{antidiag}(-1, ..., -1) \) inverts the order of elements in \( \phi_{p-1} = (\phi_1^{p-1}, ..., \phi_{p-2}^{p-1}, \phi_{p-1}^{p-1})' \) with opposite sign.
Figure 8: Restricted long-run trends, “N” denotes Normal distribution and Student-t is denoted by “T”.
Figure 9: Implied long-run trend (without bounds) for the trend-only model. Blue (thick) line under Gaussian distribution, green (dash) under Student-t, and black (thin) is the Inflation.