ISSN 1745-8587

Department of Economics, Mathematics and Statistics

BWPEF 1409

Adaptive Models and Heavy Tails

Davide Delle Monache
Queen Mary, University of London

Ivan Petrella
Birkbeck, University of London

July 2014

Birkbeck Working Papers in Economics & Finance

Adaptive Models and Heavy Tails*

Davide Delle Monache† Ivan Petrella‡

July, 2014

Abstract

This paper proposes a novel and flexible framework to estimate autoregressive models with time-varying parameters. Our setup 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.

JEL classification: C22, C51, C53, E31.

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

*We are very grateful to Michele Caivano, Anthony Garratt, Emmanuel Guerre, Dennis Kristensen, Haroon Mumtaz, Zacharias Psaradakis, Emiliano Santoro, Ron Smith and Fabrizio Venditti for their useful suggestions; and to the participants of the "Econometric Reading Group" in QMUL, seminar participant at the Bank of England, Bank of Italy, the workshop "Economic Modelling and Forecasting" in WBS, the EABCN Conference "Inflation Developments after the Great Recession" in Eltville, the "7th International Conference on Computational and Financial Econometrics (CFE 2013)", the workshop on "Dynamic Models driven by the Score of Predictive Likelihoods" in Tenerife and the "IAAE 2014 Annual Conference" in London, for their comments.

†School of Economics and Finance. Queen Mary, University of London. *E-mail*: d.dellemonache@qmul.ac.uk; *Phone*: +44 (0)2078825873.

‡Department of Economics, Mathematics and Statistics. Birkbeck, University of London and CEPR. *E-mail*: i.petrella@bbk.ac.uk; *Phone*: +44 (0)2076316418.

1 Introduction

Since the seminal work of Cogley and Sargent (2002) and Primiceri (2005) time-varying parameter (TVP) 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. In particular, the importance of accounting for time-variation in the coefficients as well as in the volatilities has been emphasized in a stream of papers that: (i) document changes in the predictability and the persistence of key macro variables (Benati and Mumtaz, 2007, Cogley, Sargent and Primiceri, 2010); (ii) link the Great Moderation to changes in monetary policy regimes (Canova and Gambetti, 2009, Primiceri 2005, Cogley and Sargent, 2005); and (iii) stress the relative gains in terms of forecast accuracy achieved by this framework compared to the traditional constant parameter models (D'Agostino et al., 2013).

Notice that all these papers are framed in a Bayesian setup that presents some shortcomings: (i) it is computational demanding (ii) when restrictions are imposed to achieve a stationary representation of the VAR a large number of draws need to be discarded, therefore leading to potentially large inefficiency. Furthermore, most of these studies assume a Normal distribution of the errors, a convenient assumption that however limits their ability to capture the tails behavior that characterizes a number of macro variables in turbulent periods.[1]

Building on recent insights of Creal et al. (2012) and Harvey (2013), in this paper we propose a new adaptive algorithm for time-varying autoregressive models that addresses simultaneously all these issues. First, the resulting model is an observation-driven model[2] that can be estimated by traditional maximum likelihood methods, rather than by simulation based methods. Second, we show how restrictions can be easily imposed *ex-ante* rather than being checked *ex-post*, therefore increasing computational efficiency.[3] Third, it can accommodate various assumptions on the distribution of the error terms. In particular, in our application we stress the importance of considering Student-t innovations. The different distributions lead to substantially different updating mechanisms that prove to be more appropriate depending on the specific economic problem we tackle.

Our model resembles the discount regression model that has been extensively used in the engineering literature (Fagin, 1964, Jazwinski, 1970, Ljung and Soderstrom, 1985). The adaptive model developed in this paper extends 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 coeffi-

[1] A noticeable exception is the recent paper by Chiu, Mumtaz and Pinter (2014).

[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, parameter-driven models which typically rely on simulation techniques can be particularly computational demanding when restrictions are imposed (see e.g. Koop and Potter, 2011, and Chan et al., 2013).

cients' 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.

On a more theoretical side, our work relates to the analysis of learning expectations. Since the seminal work of Marcet and Sargent (1989) adaptive algorithms have in fact 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 rules can be obtained from the Kalman filter (KF) with appropriate restrictions (Sargent and Williams, 2005; Evans et al., 2010). We show that most of the commonly used learning algorithms can be derived as a special case of the one developed in this paper. As a consequence, we open the route to the analysis of 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]

Moreover, our work speaks to the literature on forecasting in the presence of structural changes. In this context, Cooley and Prescott (1973, 1976) have pioneered the use of adaptive models to deal with the structural instability in economic relationships. Stock and Watson (1996) have highlighted the usefulness in economic forecasting of time-varying regressions 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 the optimal weights in the presence of structural breaks. Koop and Korobilis (2012) propose the use of an exponential weighted algorithm (obtained by ad-hoc restrictions on the KF) to model time-variation in both the coefficients and volatility. Some of these models are nested as a special case of the adaptive model we put forward.[5]

The empirical application applies our setup to the analysis of U.S. inflation dynamics in the past 60 years. We find that, when confronted with the data, our model produces reasonable patterns for the long-run trend of inflation and the underlying volatility as well as describing accurately the changes in inflation persistence and predictability highlighted by most of the literature. Most importantly, we show that by introducing the Student-t distribution we make model estimates more robust to short lived spikes in inflation (especially in the last part of the sample), a feature that leads to better in sample fit and out of sample forecasting performance. The latter is particularly striking when we try to characterize the density of the data, since well calibrated density forecasts are obtained only when we allow for heavy tails.

The paper is organized as follows. Section 2 introduces the score-driven autoregressive

[4]The 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 e.g. 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).

[5]Koop and Korobilis (2012) consider a multivariate specification with possible time-varying dimensions. It is clear that the approach discussed in this paper generalizes to the multivariate case.

model with Gaussian innovations and Section 3 discusses the relationship with the adaptive algorithms used in the literature. Section 4 extends the model to the case of Student-innovations and Section 5 shows how to impose restrictions to the model parameters. Section 6 reports an application to inflation dynamics and Section 7 concludes the article.

2 Autoregressive model with time varying parameters

An autoregressive model of order p with time-varying parameters and Gaussian residuals is defined as

$$y_t = \phi_{0,t} + \phi_{1,t} y_{t-1} + ... + \phi_{t,p} y_{t-p} + \varepsilon_t, \qquad \varepsilon_t \sim N\left(0, \sigma_t^2\right), \qquad t = 1, ..., n. \tag{1}$$

The model is typically augmented with an updating rule describing the dynamics of the parameters. Specifically, the variation of the vector of time-varying parameters, $\mathbf{f}_t = (\boldsymbol{\phi}_t', \sigma_t^2)'$ with $\boldsymbol{\phi}_t = (\phi_{0,t}, \phi_{1,t}, ..., \phi_{p,t})'$, is described by a dynamic model, e.g. a first order Markov process

$$\mathbf{f}_{t+1} = \boldsymbol{\omega} + \mathbf{A}\mathbf{f}_t + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim N\left(\mathbf{0}, \mathbf{Q}_t\right), \tag{2}$$

where $\boldsymbol{\omega}$, $\mathbf{A}$ and $\mathbf{Q}_t$ are matrices of appropriate dimension containing the hyper-parameters, and $\boldsymbol{\eta}_t$ 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 $\mathbf{f}_t$ are not perfectly predictable. In fact the unobserved state vector has an associated covariance matrix which is also 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 likelihood. The updating rule for filter estimate of $\mathbf{f}_t$ given information up to time $t-1$, $\mathbf{f}_{t|t-1} = (\boldsymbol{\phi}_{t|t-1}', \sigma_{t|t-1}^2)'$, is

$$\mathbf{f}_{t+1|t} = \boldsymbol{\omega} + \mathbf{A}\mathbf{f}_{t|t-1} + \mathbf{B}\mathbf{s}_t, \tag{3}$$

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

$$\nabla_t = \frac{\partial \left[\ell_t\left(y_t|\mathcal{F}_t, \boldsymbol{\theta}\right)\right]}{\partial \mathbf{f}_{t|t-1}} \text{ and } \mathcal{I}_t = -\mathbb{E}\left[\frac{\partial^2 \left[\ell_t\left(y_t|\mathcal{F}_t, \boldsymbol{\theta}\right)\right]}{\partial \mathbf{f}_{t|t-1} \partial \mathbf{f}_{t|t-1}'}\right], \tag{4}$$

where $\ell_t\left(y_t|\mathcal{F}_t, \boldsymbol{\theta}\right) = \log p(y_t|\mathcal{F}_t, \boldsymbol{\theta})$ is the predictive log-likelihood for the $t-$th observation which is conditioned to the information set $\mathcal{F}_t = \{F_t, Y_{t-1}\}$ and the vector of static parameters $\boldsymbol{\theta}$. Specifically, $F_t = \{\mathbf{f}_{t|t-1}, \mathbf{f}_{t-1|t-2}, ..., \mathbf{f}_{1|0}\}$ denotes present and past values of the estimated

[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, the conditional density is instead generally evaluated via simulation methods (see e.g. Durbin and Koopman, 2001).

parameters and $Y_{t-1} = \{y_{t-1}, y_{t-2}, ..., y_1\}$ are the past observations.

Note that ∇_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 $\mathbb{E}(\mathbf{s}_t|\mathcal{F}_t) = 0$ and variance $\mathbb{E}(\mathbf{s}_t\mathbf{s}_t'|\mathcal{F}_t) = \mathcal{I}_t^{-1}$: the updating rule (3) takes a step in the direction that maximizes the predictive likelihood given the past information, therefore it can be rationalized as a stochastic analog of the Gauss–Newton search direction for estimating the time-varying parameters.[7] Clearly, in the observation-driven framework the vector $\mathbf{f}_{t+1|t}$, 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

$$\widehat{\boldsymbol{\theta}} = \arg\max \mathcal{L} = \arg\max \sum_{t=1}^{n} \ell_t \left(y_t|\mathcal{F}_t, \boldsymbol{\theta}\right).$$

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 $\boldsymbol{\theta}$. Alternatively, those derivatives can be obtained numerically. In line with Creal et al (2012, sec. 2.3) we conjecture that $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \rightarrow N(0, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ is evaluated by numerical derivative at the optimum. The observation-driven counterpart of (1) can be expressed as follows

$$y_t = \mathbf{x}_t'\boldsymbol{\phi}_{t|t-1} + \varepsilon_t, \qquad \varepsilon_t|Y_{t-1} \sim N(0, \sigma^2_{t|t-1}), \qquad t = 1, ..., n, \tag{5}$$

where $\mathbf{x}_t = (1, y_{t-1}, ..., y_{t-p})'$ and $\boldsymbol{\phi}_{t|t-1} = (\phi_{0,t|t-1}, \phi_{1,t|t-1}, ..., \phi_{p,t|t-1})'$. Under Gaussian distribution, the predictive log-likelihood at time t is equal to

$$\ell_t \left(y_t|\mathcal{F}_t, \boldsymbol{\theta}\right) = -\frac{1}{2}\log\left(2\pi\right) - \frac{1}{2}\log\sigma^2_{t|t-1} - \frac{\varepsilon_t^2}{2\sigma^2_{t|t-1}}, \tag{6}$$

where $\varepsilon_t = (y_t - \mathbf{x}_t'\boldsymbol{\phi}_{t|t-1})$ is the prediction error and $\sigma^2_{t|t-1}$ is the conditional variance.[8] It can be shown that $\mathcal{I}_t$ is block diagonal so that the scaled score vector $\mathbf{s}_t$ can be specialized in two parts: the vector $\mathbf{s}_{\phi t}$ driving the coefficients[9]

$$\mathbf{s}_{\phi t} = (\mathbf{x}_t\sigma^{-2}_{t|t-1}\mathbf{x}_t')^{-1}\mathbf{x}_t\sigma^{-2}_{t|t-1}\varepsilon_t, \tag{7}$$

and the scalar $s_{\sigma t}$ driving the volatility

$$s_{\sigma t} = (\varepsilon_t^2 - \sigma^2_{t|t-1}). \tag{8}$$

In accordance with the literature on time-varying parameters models, we opt for a random walk specification and the matrix $\mathbf{B}$ is restricted to depend only upon two scalar parameters.[10]

[7] In principle one could also use a different scaling matrix as discussed in Creal et al (2012, sec. 2.2).

[8] When the model is written is vector form it becomes evident that the results derived in this paper generalize to any univariate model with exogenous and/or predetermined regressors.

[9] Note that the scaling matrix $(\mathbf{x}_t\sigma^{-2}_{t|t-1}\mathbf{x}_t')$ is not invertible, we therefore use the Moore-Penrose pseudo-inverse.

[10] Lucas (1973) first noted that most policy changes will cause changes in the decision rules that are perma-

The implied filter is then equal to

$$\boldsymbol{\phi}_{t+1|t} = \boldsymbol{\phi}_{t|t-1} + \kappa_\phi(\mathbf{x}_t\sigma^{-2}_{t|t-1}\mathbf{x}'_t)^{-1}\mathbf{x}_t\sigma^{-2}_{t|t-1}\varepsilon_t, \tag{9}$$

and

$$\sigma^2_{t+1|t} = \sigma^2_{t|t-1} + \kappa_\sigma(\varepsilon^2_t - \sigma^2_{t|t-1}). \tag{10}$$

Equation (9) resembles the Kalman filter, in fact the updated parameters react to the prediction error ε_t scaled by a gain which depends on $x_t\sigma^{-2}_{t|t-1}$. Moreover, (9) also resembles the recursive least squares where $1/t$ is replaced by the constant parameter κ_ϕ. Equation (10) is the same as the integrated GARCH model. Note that the time-varying volatility cancel out from the coefficients' dynamics and it does not directly affect the coefficients' filtering in (9).[11] In order to avoid swift changes in the parameters, it is customary to replace it with its smoothed version[12]

$$\mathbf{R}_t = (1-\kappa_h)\mathbf{R}_{t-1} + \kappa_h\mathbf{x}_t\sigma^{-2}_{t|t-1}\mathbf{x}'_t = \mathbf{R}_{t-1} + \kappa_h(\mathbf{x}_t\sigma^{-2}_{t|t-1}\mathbf{x}'_t - \mathbf{R}_{t-1}), \tag{11}$$

where κ_h is a smoothing parameter to be estimated. As a result the updating rule for the coefficients (9) is equal to

$$\boldsymbol{\phi}_{t+1|t} = \boldsymbol{\phi}_{t|t-1} + \kappa_\phi\mathbf{R}^{-1}_t\mathbf{x}_t\sigma^{-2}_{t|t-1}\varepsilon_t. \tag{12}$$

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. The model is therefore similar to the single-source of error model of Casalas et al (2002) and Hyndman et al (2008).[13] Blasques et al. (2014) focus on the AR(1) specification with constant variance showing that the implied reduced form model follows a nonlinear ARMA and show that this class of models is optimal in terms of the Kullback-Leibler criterium.

3 Relation with the adaptive algorithms

This section highlights the relation between the score-driven model and various adaptive algorithms widely used in the literature. We illustrate that our setup is very general and nests some important model used in macroeconomics as well as in econometrics. In particular, the

nent. According to this view we assume that the parameters of the model will drift systematically over time away from their initial value with no tendency to return to a mean value (see also Cooley and Prescott, 1976). In practice we restrict $\boldsymbol{\omega} = \mathbf{0}$, $\mathbf{A} = \mathbf{I}$ and $\mathbf{B} = \begin{bmatrix} \kappa_\phi\mathbf{I}_{p+1} & 0 \\ 0 & \kappa_\sigma \end{bmatrix}$. One could relax those restrictions allowing a more general specification of $\boldsymbol{\omega}$, $\mathbf{A}$ and $\mathbf{B}$. However, by doing so the model would not resemble a stochastic version of the Gauss-Newton algorithm (see Remark 1).

[11] Note that this is no longer the case when the Hessian matrix is replaced with a smoothed version as described later on.

[12] For some extreme observation at time t, 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.

[13] 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.

algorithms widely used to model the learning expectations, the large TVP-VAR of Koop and Korobilis (2012) and the TVP model of Stock and Watson (1996) can be all derived as a special case of our score-driven model. To facilitate the comparison 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 case of time-varying variance. With constant variance, and setting $\kappa_\phi = \kappa_h = \kappa$, the score-driven filter (11)-(12) collapses to

$$\begin{aligned} \mathbf{R}_t &= \mathbf{R}_{t-1} + \kappa(\mathbf{x}_t \sigma^{-2} \mathbf{x}_t' - \mathbf{R}_{t-1}), \\ \boldsymbol{\phi}_{t+1|t} &= \boldsymbol{\phi}_{t|t-1} + \kappa \mathbf{R}_t^{-1} \mathbf{x}_t \sigma^{-2} \varepsilon_t. \end{aligned} \tag{13}$$

The recursive algorithm in (13) is exactly the Constant Gain Learning (CGL) widely used in the learning expectations literature.[14]

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 κ gives a trade-off between the tracking capability and the smoothness. Moreover, the CGL is a forgetting factor algorithm and can also be derived from an off-line method, i.e. the discounted least squares principle. See details in the Appendix A.*

The discounted regression model has been extensively used in the 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 following parameter-driven model*

$$\begin{aligned} y_t &= \mathbf{x}_t' \boldsymbol{\phi}_t + \varepsilon_t, \ \varepsilon_t \sim N\left(0, \frac{\sigma^2}{1-\kappa}\right), \\ \boldsymbol{\phi}_{t+1} &= \boldsymbol{\phi}_t + \boldsymbol{\eta}_t, \ \boldsymbol{\eta}_t \sim N\left(\mathbf{0}, \mathbf{P}_{t|t} \frac{\kappa}{1-\kappa}\right), \end{aligned} \tag{14}$$

where $\mathbf{P}_{t|t} = \mathbb{E}[(\boldsymbol{\phi}_{t|t} - \boldsymbol{\phi}_t)(\boldsymbol{\phi}_{t|t} - \boldsymbol{\phi}_t)']$ and $\boldsymbol{\phi}_{t|t} = \mathbb{E}(\boldsymbol{\phi}_t | Y_t)$ are the estimated quantities from the KF, and κ is the gain parameter. The KF delivers the estimated state vector $\boldsymbol{\phi}_{t+1|t} = \mathbb{E}(\boldsymbol{\phi}_{t+1}|Y_t)$ which is exactly equal to the CGL algorithm and thus it is a score-driven filter. It is worth noticing that the restrictions on (14) imply that the shock $\boldsymbol{\eta}_t$ is driven by the prediction error and thus the parameter-driven model collapses to an observation-driven model. See Appendix A for details.

[14] See, among others, Evan and Honkaphoja (2001), Sargent and William (2005), Branch and Evans (2006) and Carceles-Poveda and Giannitsarou (2007).

Koop and Korobilis (2012) propose to estimate a large TVP-VAR using the specification described in the previous Lemma. Therefore, they use the CGL algorithm which is nested within the score-driven framework. Koop and Korobilis (2012) also allow for a time-varying covariance matrix estimated by an exponential smoothing; later on we show that also this feature is nested in our framework.

Another widely used specification of the parameter-driven model (14) assumes that $\varepsilon_t \sim N(0, \sigma^2)$ and $\boldsymbol{\eta}_t \sim N(\mathbf{0}, \kappa^2\Sigma)$ with $\Sigma = \sigma^2[\mathbb{E}(\mathbf{x}_t\mathbf{x}_t')]^{-1}$ (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,[15] whereas Slobodyan and Wouters (2012) refer to it as KF learning.

Lemma 3 *Setting $\boldsymbol{\eta}_t \sim N(\mathbf{0}, \kappa^2\Sigma)$ 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 $\sigma^{-2}\mathbb{E}(\mathbf{x}_t\mathbf{x}_t')$. Similarly, setting $\boldsymbol{\eta}_t \sim N(\mathbf{0}, \kappa^2\Sigma^{-1})$ leads to the score-driven filter (9) where the scaling matrix is replaced by the identity matrix. See details in the Appendix A.*

In fact, all the recursive algorithms discussed in this sub-section can be seen as particular cases of the adaptive algorithms popularized by Ljung and Soderstrom (1985), which are the building blocks of the learning expectations literature in macroeconomics.

Remark 1 *Following Ljung and Soderstrom (1985), the CGL can be obtained from a recursive solution of a quadratic loss function. In particular, given a sequence of random IID random variables $\epsilon = \{\varepsilon_1, ..., \varepsilon_T\}$, the optimal choice of the full coefficients' path across time, that is $\boldsymbol{\phi} = \{\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_T\}$, can be obtained from a quadratic criterion function and it leads to the stochastic analog of a Gauss-Newton search direction method*

$$\widehat{\boldsymbol{\phi}}_{t+1|t} = \widehat{\boldsymbol{\phi}}_{t|t-1} + \kappa_t[\mathbf{H}(\widehat{\boldsymbol{\phi}}_{t|t-1}, \varepsilon_t)]^{-1}\mathbf{G}(\widehat{\boldsymbol{\phi}}_{t|t-1}, \varepsilon_t),$$

where $\mathbf{G}(\boldsymbol{\phi}_{t|t-1}, \varepsilon_t)$ and $\mathbf{H}(\boldsymbol{\phi}_{t|t-1}, \varepsilon_t)$ are the Gradient vector and the Hessian matrix respectively, and κ_t is a sequence of gain parameters appropriately chosen. Under Gaussian distribution, the recursive Gauss-Newton solution for a quadratic criterion function is equivalent to the score-driven model proposed in this paper.

Remark 1 highlights how the score-driven model (5)-(10)-(12) extends the adaptive algorithms allowing for non-Gaussian distribution as well as for changes in volatility. In fact, the estimated volatility (10) is obtained following exactly the same criterion and the implied filter is an exponentially smoothing of the squared prediction errors

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

[15] Note that this specification is an approximation of the Stochastic Gradient Algorithm; see details of Lemma 3 discussed in the Appendix.

Ljung and Soderstrom (1985, sec. 3.4.3) and Koop and Korobilis (2012) use exactly the same model to capture the variation in the volatility. However, they propose this model in a rather heuristic way without a derivation from the Gauss-Newton principle.

The next section extends the adaptive algorithms to the case of non-Gaussian distribution, i.e. the Student-t distribution. This can be considered as a recursive algorithm for a non-quadratic loss function (see Ljung and Soderstrom, 1985, sec. 3.5).

4 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, it can therefore be considered for cases where rare events become relevant. In light of the recent turbulent time the departure from Gaussianity become very relevant in both applied and theoretical works (see Curdia et al. 2013, Chiu et al., 2014).

Harvey and Luati (2012) highlight that a score-driven model with Student-t innovations leads to a filter which is robust to a few large errors. Thus model (5) becomes[16]

$$y_t = \mathbf{x}_t'\boldsymbol{\phi}_{t|t-1} + \varepsilon_t, \qquad \varepsilon_t|Y_{t-1} \sim t_\upsilon(0, \sigma^2_{t|t-1}), \tag{15}$$

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

$$\ell_t\left(y_t|\mathcal{F}_t, \boldsymbol{\theta}\right) = c\left(\eta\right) - \frac{1}{2}\ln\sigma^2_{t|t-1} - \left(\frac{\eta+1}{2\eta}\right)\log\left[1 + \frac{\eta}{1-2\eta}\frac{\varepsilon_t^2}{\sigma^2_{t|t-1}}\right], \tag{16}$$

where

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

$\eta = 1/\upsilon$ and $\Gamma(\cdot)$ is the Gamma function. It can be shown that the scaled-score driving the coefficients and the variance are equal to

$$\mathbf{s}_{\phi t} = \frac{\left(1-2\eta\right)\left(1+3\eta\right)}{\left(1+\eta\right)}(\mathbf{x}_t\sigma^{-2}_{t|t-1}\mathbf{x}_t')^{-1}\mathbf{x}_t\sigma^{-2}_{t|t-1}w_t\varepsilon_t, \tag{17}$$

and

$$s_{\sigma t} = \left(1+3\eta\right)\left(w_t\varepsilon_t^2 - \sigma^2_{t|t-1}\right). \tag{18}$$

Notice that both depend upon scalar weights

$$w_t = \frac{\left(1+\eta\right)}{\left(1-2\eta+\eta\zeta_t\right)}, \tag{19}$$

[16] Model (15) generalizes the setting in Harvey and Luati (2012) to the case of additional regressors and time-varying variance. A score-driven model with non-Gaussian innovations, not only modifies the likelihood function (as in the t-GARCH of Bollerslev, 1987) but it also implies a different filtering process for the time–varying parameters.

where $\zeta_t = \varepsilon_t^2/\sigma_{t|t-1}^2$ and nest the Gaussian case for $\eta = 0$ ($\upsilon \to \infty$); see the Appendix A 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 Figure 1. The left panel shows the magnitude of the weights w_t as a function of the standardized prediction errors, while the right one shows the weighted realizations $w_t\sqrt{\zeta_t}$ which is known as *influence function* in robust statistics (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.

[insert Figure 1]

Under Student-t distribution the score-driven algorithm leads to a robust filter and generalizes the CGL algorithm (13).

Proposition 1 *Under Student-t distribution the score-driven model leads to the following adaptive algorithm for the time-varying parameters*

$$\begin{aligned} \mathbf{R}_t &= \mathbf{R}_{t-1} + \kappa_h[\alpha w_t(\mathbf{x}_t\sigma_{t|t-1}^{-2}\mathbf{x}_t' - \mathbf{R}_{t-1})], \\ \boldsymbol{\phi}_{t+1|t} &= \boldsymbol{\phi}_{t|t-1} + \kappa_\phi\mathbf{R}_t^{-1}\mathbf{x}_t\sigma_{t|t-1}^{-2}[\alpha w_t(y_t - \mathbf{x}_t'\boldsymbol{\phi}_{t|t-1})], \\ \sigma_{t+1|t}^2 &= \sigma_{t|t-1}^2 + \kappa_\sigma[(1+3\eta)\,(w_t\varepsilon_t^2 - \sigma_{t|t-1}^2)], \end{aligned} \tag{20}$$

with $\alpha = [(1-2\eta)\,(1+3\eta)\,/(1+\eta)]$, w_t *defined in (19) and* $\boldsymbol{\theta} = (\upsilon, \kappa_h, \kappa_\phi, \kappa_\sigma)'$ *is the corresponding vector of static parameters. The magnitude of the weights* w_t *depends on how close the actual observation is to the center of the distribution of* ε_t*: large deviations are downweighted and a small value of* w_t *is more likely with lower degree of freedom and lower dispersion of the distribution. Therefore, the recursions above imply a double weighting scheme, i.e. the observations are weighted both across time and realizations, and the estimated time-varying parameters are robust to extreme events.*

A simplified version of model (15) helps clarify the impact of the double weighting. Assume that $\mathbf{x}_t = 1$ and w_t is exogenously given. This specification leads to an IMA(1,1) model with time-varying moving average coefficient $(1-\kappa_\theta w_t)$, and time-varying variance. The time-varying mean can be expressed as follows

$$\mu_{t+1|t} = \mu_{t|t-1} + \kappa_\theta w_t(y_t - \mu_{t|t-1}) = \kappa_\theta\sum_{j=0}^{\infty}\gamma_j\widetilde{y}_{t-j} = \frac{\kappa_\theta}{1-(1-\kappa_\theta w_t)L}\widetilde{y}_t, \tag{21}$$

with $\widetilde{y}_{t-j} = w_{t-j}y_{t-j}$. Specifically, equation (21) shows that the observations are: (i) weighted to be robust to the impact of extreme events, i.e. $\widetilde{y}_t = w_t y_t$, and (ii) they are also smoothed

across time with weights $\gamma_j = \prod_{k=t-j+1}^{t}(1-\kappa_\theta w_k)$, $\gamma_0 = 1$ and $\kappa_\theta = \kappa_\phi \alpha$. This is equivalent to a one-sided low-pass filter with time-varying coefficients, that is $\kappa_\theta/[1-(1-\kappa_\theta w_t)L]$, and it implies a time-varying transfer function.[17] Similarly, in order to estimate the variance σ_t^2, the squared prediction errors ε_{t-j}^2 are weighted by $\gamma_j w_t$, where the weights across time are $\gamma_j = [1-\kappa_\sigma(1+3\eta)]^j$, namely

$$\sigma_{t+1|t}^2 = \sigma_{t|t-1}^2 + \kappa_\zeta(w_t\varepsilon_t^2 - \sigma_{t|t-1}^2) = \kappa_\zeta \sum_{j=0}^{\infty}(1-\kappa_\zeta)^j \widetilde{\varepsilon}_{t-j}^2 = \frac{\kappa_\zeta}{1-(1-\kappa_\zeta)L}\widetilde{\varepsilon}_t^2, \tag{22}$$

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

Remark 2 *In practice the weights w_t depend (non-linearly) on the current observations and the past parameters' estimation through $\zeta_t = \varepsilon_t^2/\sigma_{t|t-1}^2$. Therefore, the score-driven model under Student-t distribution solves a recursive stochastic Gauss-Newton algorithm for a non-quadratic loss function and it leads to a non-linear filter. Therefore, it 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).*

5 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. In the Bayesian framework constraints are usually imposed by rejection sampling (see e.g. Cogley and Sargent, 2005, and Koop and Potter, 2012). Thus, however, leads to heavy inefficiencies.

General non-linear restrictions can be accommodated within the score-driven model. This requires to reparameterize the model with respect to a new vector of unconstrained parameters. Define the following transformation $\mathbf{f}_t = g(\widetilde{\mathbf{f}}_t)$, where $\mathbf{f}_t$ is the original vector of parameters, $\widetilde{\mathbf{f}}_t$ is the new parametrization 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) can be expressed with respect to the new vector of parameters

$$\widetilde{\mathbf{f}}_{t+1|t} = \widetilde{\boldsymbol{\omega}} + \widetilde{\mathbf{A}}\widetilde{\mathbf{f}}_{t|t-1} + \widetilde{\mathbf{B}}\widetilde{\mathbf{s}}_t, \tag{23}$$

where $\widetilde{\mathbf{s}}_t = \widetilde{\mathcal{I}}_t^{-1}\widetilde{\nabla}_t$ is the scaled score computed with respect to $\widetilde{\mathbf{f}}_t = g^{-1}(\mathbf{f}_t)$, where $g^{-1}(\cdot)$ is the

[17] The transfer function ca be expressed as follows $G(\lambda) = \kappa_\theta\left[1+(1-\kappa_\theta w_t)^2 - 2(1-\kappa_\theta w_t)\cos(\lambda)\right]^{-1/2}$, where $0 < \lambda < \pi$ is the radian frequancy. See Dahlhaus (2012) for details on stationary processes with time-varying spectral density.

inverse function of $g(\cdot)$. For a given continuous and differentiable function $g(\cdot)$, the new score vector is then

$$\widetilde{\nabla}_t = \frac{\partial \ell_t}{\partial \widetilde{\mathbf{f}}_{t|t-1}} = \left[\frac{\partial \ell_t}{\partial \mathbf{f}'_{t|t-1}} \frac{\partial \mathbf{f}_{t|t-1}}{\partial \widetilde{\mathbf{f}}'_{t|t-1}} \right]' = \Psi'_t \nabla_t,$$

where $\Psi_t = \partial \mathbf{f}_{t|t-1} / \partial \widetilde{\mathbf{f}}'_{t|t-1}$ is the Jacobian of $g(\cdot)$ and is deterministic given past information. Therefore, the transformed scaling matrix is equal to $\widetilde{\mathcal{I}}_t = \Psi'_t \mathcal{I}_t \Psi_t$ and the new scaled score is then equal to

$$\widetilde{\mathbf{s}}_t = (\Psi'_t \mathcal{I}_t \Psi_t)^{-1} \Psi'_t \nabla_t. \tag{24}$$

The transformation function $g(.)$ imposes (possibly) non-linear restrictions on the time-varying parameters. It is worth noticing that under Gaussian distribution, the non-linear filtering problem can be solved by first order Taylor approximation. This argument is formalized in the Theorem below. Also in this case we can replace the scaling matrix $\widetilde{\mathcal{I}}_t$ with its smoothed version $\mathbf{R}_t = (1 - \kappa_h)\mathbf{R}_{t-1} + \kappa_h \widetilde{\mathcal{I}}_t$.

Theorem 1 *Consider the Gaussian model (14) and impose a non-linear transformation on the coefficients $\boldsymbol{\phi}_t = g(\boldsymbol{\alpha}_t)$. The model can be solved by the Extended KF of Anderson and Moore (1979, sec. 8.2) and the implied algorithm is exactly equal to the score-driven filter (23).*

(Proof in the Appendix A.)

The constrained algorithm has been commonly implemented in the literature by means of the *projection facility* (see Ljung and Soderstrom, 1985, sec. 6.6, Timmermann, 1996, and Evans and Honkapohja, 1998). Specifically, they use a constant parameter weighting the driving process such that the incremental step is progressively shrunk until the restriction is satisfied.[18] The adaptive model (5), with (23)-(24), automatically achieves the same objective. In fact, the matrix Ψ_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.

5.1 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 restricting the latter in the interval $(-1, 1)$. To simplify the notation we start with model (1) without the intercept and then we consider the general model.

[18] In practice this is often implemented by skipping the updating each time the restrictions are violated.

Proposition 2 *For each point in time t, let* $\boldsymbol{\phi}_t = (\phi_{1,t}, ..., \phi_{p,t})'$ *denote the vector of coefficients,* $\boldsymbol{\pi}_t = (\pi_{1,t}, ..., \pi_{p,t})'$ *the corresponding vector of partial autocorrelations and* $\boldsymbol{\alpha}_t = (\alpha_{1,t}, ..., \alpha_{p,t})'$ *the vector of unrestricted coefficients. A locally stationary model has* $\boldsymbol{\phi}_t \in \mathbf{S}^p$, *where* $\mathbf{S}^p$ *is the hyperplane with all roots,* $\mathbf{z}_t$, *inside the unit circle, i.e.* $\boldsymbol{\phi}_t(\mathbf{z}_t) = 0$, $\mathbf{z}_t \in \mathbf{C}^p$ *and* $|z_{j,t}| < 1$ *for* $j = 1, ..., p$. *It is possible to show that* $\boldsymbol{\phi}_t \in \mathbf{S}^p$ *if and only if* $\boldsymbol{\pi}_t \in \mathbf{R}^p$ *and* $|\pi_{j,t}| < 1$. *Therefore, let* $\boldsymbol{\phi}_t = \Phi(\boldsymbol{\pi}_t)$ *define the function mapping the coefficients to the partial autocorrelations and* $\boldsymbol{\pi}_t = \Upsilon(\boldsymbol{\alpha}_t)$ *a function that restricts the partial autocorrelations to lie in the region* $(-1, 1)$. *The function* $\boldsymbol{\phi}_t = \Phi(\boldsymbol{\pi}_t)$ *is uniquely obtained by the last recursion of the Durbin-Levinson algorithm*

$$\phi_t^{j,k} = \phi_t^{j,k-1} - \pi_{k,t}\phi_t^{k-j,k-1} \quad \text{for } j = 1, ..., k-1 \text{ and } k = 2, ..., p, \tag{25}$$

with $\phi_t^{1,1} = \pi_{1,t}$ *and* $\phi_t^{k,k} = \pi_{k,t}$. *The function* $\boldsymbol{\pi}_t = \Upsilon(\boldsymbol{\alpha}_t)$ *is any monotonic and differentiable function*

$$\pi_{j,t} = \Upsilon(\alpha_{j,t}), \text{ such that } \pi_{j,t} \in (-1,1), \quad j = 1, ..., p. \tag{26}$$

The composite function $g(\cdot) = \Phi[\Upsilon(\cdot)]$ *maps the restricted stationary coefficients into the unrestricted parameters, i.e.* $\boldsymbol{\phi}_t = g(\boldsymbol{\alpha}_t)$ *with* $\boldsymbol{\alpha}_t \in (-\infty, \infty)$ *and* $\boldsymbol{\phi}_t \in \mathbf{S}^p$.

(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

$$\Psi_t = \frac{\partial g(\boldsymbol{\alpha}_t)}{\partial \boldsymbol{\alpha}_t'} = \frac{\partial \Phi(\boldsymbol{\pi}_t)}{\partial \boldsymbol{\pi}_t'} \frac{\partial \Upsilon(\boldsymbol{\alpha}_t)}{\partial \boldsymbol{\alpha}_t'}, \tag{27}$$

where $\partial \Upsilon(\boldsymbol{\alpha}_t)/\partial \boldsymbol{\alpha}_t'$ is diagonal matrix containing $\partial \Upsilon(\alpha_{jt})/\partial \alpha_{j,t}$ with $j = 1, ..., p$, while the analytic expression for $\partial \Phi(\boldsymbol{\pi}_t)/\partial \boldsymbol{\pi}_t' = \Gamma_t$ is obtained in the theorem below.

Theorem 2 *The Jacobian matrix* Γ_t *is obtained from the last iteration of the recursion*

$$\begin{aligned} \Gamma_{k,t} &= \begin{bmatrix} \widetilde{\Gamma}_{k-1,t} & \mathbf{b}_{k-1,t} \\ \mathbf{0}_{k-1}' & 1 \end{bmatrix}, \\ \widetilde{\Gamma}_{k-1,t} &= \mathbf{J}_{k-1,t}\Gamma_{k-1,t}, \qquad k = 2, ..., p, \end{aligned} \tag{28}$$

with

$$\mathbf{b}_{k-1,t} = -\begin{bmatrix} \phi_t^{k-1,k-1} \\ \phi_t^{k-2,k-1} \\ \vdots \\ \phi_t^{2,k-1} \\ \phi_t^{1,k-1} \end{bmatrix}, \quad \mathbf{J}_{k-1,t} = \begin{bmatrix} 1 & 0 & \cdots & 0 & -\pi_{k,t} \\ 0 & 1 & 0 & -\pi_{k,t} & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & -\pi_{k,t} & 0 & 1 & 0 \\ -\pi_{k,t} & 0 & \cdots & 0 & 1 \end{bmatrix}. \tag{29}$$

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

(Proof in the Appendix A).

Given the elements of the scaled score vector $\mathbf{s}_{\phi t} = \mathcal{I}_{\phi t}^{-1} \nabla_{\phi t}$ (computed with respect to $\boldsymbol{\phi}_t$), the adaptive algorithm for the transformed coefficients $\boldsymbol{\alpha}_t$ is equal to

$$\boldsymbol{\alpha}_{t+1|t} = \boldsymbol{\alpha}_{t|t-1} + \kappa_\alpha (\Psi_t' \mathcal{I}_{\phi t} \Psi_t)^{-1} \Psi_t' \nabla_{\phi t}, \tag{30}$$

where $\boldsymbol{\phi}_t = g(\boldsymbol{\alpha}_t)$ and $\Psi_t = \Psi(\boldsymbol{\alpha}_t)$ are computed as outlined in Proposition 2 and Theorem 2, 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

$$\Psi_t = \frac{\partial \boldsymbol{\phi}_t}{\partial \boldsymbol{\alpha}_t'} = \begin{bmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \Psi_{22,t} \end{bmatrix}, \tag{31}$$

where $\Psi_{22,t} = \partial(\phi_{1,t}, ..., \phi_{p,t})' / \partial(\alpha_{1,t}, ..., \alpha_{p,t})$ as computed in Theorem 2.

5.2 Bounded trend

It is also often the case that in practice one wants to discipline the model so as 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), i.e. $\lim_{h \to \infty} \mathbb{E}_t (y_{t+h}) = \mu_t$. . Specifically, for a stationary time-varying autoregressive process, local-to-date t 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 $\widetilde{y}_t = (y_t - \mu_t)$, follows a locally stationary time-varying AR(p) model, i.e. $\Pr \{\lim_{h \to \infty} \mathbb{E}_t (\widetilde{y}_{t+h}) = 0\} = 1$. Following Chan et al (2013), we want to restrict $\mu_t \in [\underline{b}, \overline{b}]$.

Proposition 3 *Let $h(\cdot)$ be any continuous and differential function so that $h(\cdot) \in [\underline{b}, \overline{b}]$. The restriction on $\mu_t \in [\underline{b}, \overline{b}]$ can be achieved with the following transformation*

$$\phi_{0,t} = h(\alpha_{0,t}) \left(1 - \sum_{j=1}^{p} \phi_{j,t} \right). \tag{32}$$

The Jacobian matrix is then equal to

$$\Psi_t = \frac{\partial \boldsymbol{\phi}_t}{\partial \boldsymbol{\alpha}_t'} = \begin{bmatrix} \psi_{11,t} & \boldsymbol{\psi}_{12,t}' \\ \mathbf{0} & \Psi_{22,t} \end{bmatrix}, \tag{33}$$

where $\Psi_{22,t}$ has been computed in Theorem 2, while $\psi_{11,t}$ and $\boldsymbol{\psi}_{12,t}'$ are

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

where $\boldsymbol{\iota} = (1, 1, ..., 1)'$.

To summarize, for each time t, the recursion (30) is implemented as follows: first, the stationary AR coefficients are computed following Proposition 2; second, the constrained intercept and the Jacobian matrix Ψ_t are computed as described in Proposition 3 so that all the necessary elements to update $\boldsymbol{\alpha}_{t|t-1}$ are then available. In this section we have shown how to implement some popular restrictions in a score-driven setup and this leads to a non-linear filter that can be implement in the Classical framework without incurring in the computational demanding simulation methods of Koop and Potter (2011) and Chan el al (2013).

6 Application to the inflation dynamics

We implement the adaptive model in the analysis of inflation dynamics. The change in the persistence of the inflation has been strongly supported by Cogley and Sargent (2001).[19] 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. Another issue that has received much attention in recent years is related to the presence of a time-varying level of trend-inflation (Cogley, 2002, and Stock and Watson, 2006). Specifically, trend-inflation is generally thought of as a measure of the public's perception of the credibility of the central bank inflation targeting, (see Kozicki and Tinsley, 2001, and Faust and Wright, 2011). Furthermore, Clark and Doh (2011) and Chan et al. (2013) highlight how accurate estimates of trend-inflation can improve the inflation forecasts at a long-horizon.

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

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

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 proven 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 dynamic over the post-WWII sample. The literature has mainly focussed on the parameter-driven models, estimated by Bayesian methods.[20]

In the application we allow for various specifications of (34). Specifically, we first consider a model with time-varying trend-only ($p = 0$), then we allow for various specifications of

[19] Similar results are provided by Taylor (2000) and Brainard and Perry (2000).

[20] A noticeable exception is the work of Pivetta and Reis (2007).

the autoregressive components ($p = 1, 2$ and 4), and the time-varying mean $\phi_{0,t}$ is always included. Chan et al. (2013) forcefully argue for imposing bounds on the long-run trend on the grounds that a level of the trend inflation that is too low (or too high) 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.[21] Furthermore, we consider all specifications under Gaussian and Student-t innovations. Finally, partial autocorrelations are always bounded so as to impose local stationarity of the model and the variance is reparameterized so that it is always positive.

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) without the autoregressive terms. In SW the conditional mean and the measurement error are driven by two independent shocks with stochastic volatility. The model then 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 4, 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.

[Insert Table 1]

Table 1 reports the estimates for the various specifications for the annualized quarterly US-CPI inflation over the period 1955Q1–2012Q4. 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 Bayesian Information Criterion (BIC). The trend-only specification with Gaussian innovations implies that the trend is estimated by the exponential smoothing as in Cogley (2002).[22] This model features a high estimation of the smoothing parameter which implies a faster learning process. This is also true for all the specifications without autoregressive coefficients. Adding the autoregressive components shows a substantially smaller estimate of the smoothing parameter as some of the persistence of inflation is now captured by the autoregressive terms. In contrast, the smoothing parameter associated to the variance equation is instead stable and typically higher than the one associated with the coefficients, lending support to the idea that changes in the variance are particularly relevant in our sample (see also Pivetta and Reis, 2007). Noticeably, the specifications with Student-t distribution

[21] The bounds correspond to the upper and lower bounds in the posterior in Chen et al. (2013). They highlight that it is difficult to identify exactly those bounds. They also 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. This is typically not affected by the choice of the upper and lower bound.

[22] Notice that with respect to the model in Cogley (2002) the specification used as benchmark allows for the time-variation in the variance. The latter does not affect the way the trend component is computed. Nevertheless, it does affect the estimate of the smoothing parameter.

always considerably outperform the ones with Gaussian innovations, as for the likelihood values and information criteria. In fact, the estimated low value of degrees of freedom υ depicts a remarkable difference between the Gaussian and the Student-t specification. The low value of υ suggests that there might be pronounced variations of inflation at the quarterly frequency. Those variations either arise from measurement issues or are related to the presence of rare events that structural macroeconomics should explicitly account for (as recently advocated by Curdia et al., 2013). Notice that $\upsilon = 5$ is also consistent with the calibrated density forecast in Corradi and Swanson (2006). Furthermore, the AR(1) specification without bounds on the long-run mean slightly outperforms all the others in terms of fitting.

6.1 Estimates of Trend Inflation, Persistence, and Volatility

Figure 2 presents estimates of the long-run trend in inflation for the various specifications considered in this paper. The long-run trend, when is not bounded, tends to follow the underlying inflation, smoothing away the transitory fluctuations. Some differences can be appreciated when comparing the different specifications. The trend-only specification follows inflation very closely trough the ups and downs. When we add autoregressive terms to the model, few differences can be appreciated across various specifications. The inclusion of lags delivers a smoother long-run trend, suggesting that the high inflation in the early part of the sample and in the 70s is to be attributed to deviations from the trend. All specifications suggest that since the mid 90s, the long-run trend is stable between 2-3%, going slightly over 3% on the run up to the recent recession. Also, it is worth noticing that the specification with Student-t are less affected by the sharp transitory movements in inflation, in particular in the last part of the sample. Imposing the upper bound on the long-run mean implies a qualitatively similar picture for the trend-inflation across the specifications.[23] The trend is consistent with the idea of a central bank anchoring the expectations of trend-inflation to a fairly stable level over the sample. Trend-inflation rises above 3% in the early 70s and then decreases back to a slightly lower level only in the mid 90s. It is interesting to note that the pattern in the long-run trend is quite similar to the one found by Chan et al. (2012), although they use a different model specification and estimation techniques.

[Insert Figure 2]

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 Figures 3 and 4. 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

[23] Figure 2 excludes the trend-bound specification which is destined to reach the bounds during the great inflation period by construction.

for a large number of lags tends to decrease the estimated persistence. This finding reconciles the different results obtained by Pivetta and Reis (2007), who focus on time-varying AR model with three lags. It reports evidence of little variation in inflation persistence. Interestingly, the specifications with Student-t innovations are more robust to sharp variations which are due to the short lived spikes in the late part of the sample.

[Insert Figure 3]

[Insert Figure 4]

Figure 5 reports measures of the change in volatility. Some interesting issues emerge. All specifications show that the variance of inflation was substantially higher 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 differences between bound versus unrestricted trend. Clearly, the bounded specifications overstate the level of volatility in the period when the bound is binding. Interestingly, if we compare Gaussian and Student-t distribution, they share similar low-frequency variation and the specifications with Student-t innovation display substantially more variation in the volatility. Consequently, with Student-t innovations the variance is less affected by the outliers and it can better adjust to accommodate changes in the dispersion of the central part of the distribution. This latter result is particularly important in light of the considerable evidence in favor of the Student-t specification reported in the previous sub-section. In fact, most of the macroeconomic literature, which has mainly 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 are 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 5]

6.2 Forecasting Evaluation

In this section we assess the forecasting performance of the various specifications. 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 remained constant going forward in time. We first consider point forecast and use both root mean squared error (RMSE) and the absolute mean error (MAE). 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 and it very close to the model of Cogley (2002).

Table 2 reports the results. Despite the well-known performance of the benchmark model, many of the alternative models tend to have lower RMSE or MAE. This improvement becomes

substantial at longer forecast horizons, although in most of the cases the difference in forecasting performance is not statistically significant.[24] A comparison between the Gaussian and Student-t models reveals little differences in terms of point forecast. Imposing bounds on the long-run mean marginally enhances the performance of the various specifications, and in particular for the specification with Student-t innovations.[25]

[Insert Table 2]

Table 3 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 improve in performance with respect to the ones with Gaussian innovations, regardless of the model.[26] Furthermore, the table reports two tests for the calibration of the densities. One is the LR test on the inverse transformation of 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 the presence of parameter estimation error. The specifications with Gaussian innovations prove to be not well calibrated. In order to understand why this is the case Figure 6 plots the empirical distribution function (p.d.f.) 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 as in Diebold et al. (1998). Figure 7 displays the cumulative distribution function (c.d.f.) of the PITs for each realization, under the null hypothesis the PITs should be uniformly distributed. Therefore the c.d.f. of the PITs should be the 45^o line. The figure also reports the critical values based on the RS test. If the c.d.f. of the PITs is outside the critical value lines, we conclude that the density forecast is not well calibrated.

[Insert Table 3]

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

[Insert Figure 6]

[Insert Figure 7]

In Table 4, for each pair of models, we report the p-values of the test of difference in the average log predictive score using uniform weights, as outlined in Amisano and Giacomini

[24] Despite the expanding window, it is possible to apply the Giacomini and White (2006) test as the models 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 is used to forecast.

[25] The trend-only specification with restricted long-run mean is always outperformed by the alternative ones, in particular for the short horizon. Anyway, the relative performance of this specification is severely biased by the inclusions of the great inflation period (mid 70s-80s), as the model has an upper boundeat 5%.

[26] Clark and Ravazzolo (2012) document the gains of allowing for fatter tails. However, they found much smaller improvement.

(2007). The results confirm that the substantial differences between the Normal and Student-t are indeed significant. At the same time, the p-values confirm that some of the differences across the various specifications with Student-t innovations are significative, but none of the various specifications clearly outperforms the others.

[Insert Table 4]

The adaptive model developed in this paper delivers a model-consistent algorithm in presence of heavy tails distribution. Appendix B explores the importance of using a law of motion for the parameters consistent with the score-driven model as opposed to some ad-hoc specifications. We show that the score-driven specification outperforms the alternative ones: in particular, both the low degree of freedom and the score-driven law of motion, are important to achieve a well calibrated density forecasts.

Concluding, the empirical exercise shows that the model with Student-t distribution produces time variation in the parameters which are robust to the presence of heavy tails. Furthermore, the volatility is less affected by the behavior in the tail of the distribution so that it can better reflect the changes in the spread of the central part of the density. These aspects of the model are key in order to retrieve well calibrated density forecast for inflation over the sample analyzed.

7 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. Moreover, 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. With regards to the parameter-driven models, the route taken in this paper does not require the use of simulation techniques and thus has a clear computational advantage especially when restrictions on the parameters are imposed.

We apply the algorithm to the study of inflation dynamics. Several alternative specifications are shown to track the data very well, so that they give a parsimonious characterization of the inflation dynamics and producing good forecasts. Allowing for heavy-tails is found to be a key ingredient to obtain well calibrated density forecasts over the analyzed sample. The dynamics of the parameters under Student-t innovations are more robust to short lived variations in

inflation, especially in the last decade. Furthermore, the use of heavy-tails highlights the presence of high-frequency variations in the volatility on top of the well documented low-frequency variations.

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 the 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 stationarity is problematic.

References

Abraham, B. and J. Ledolter (1983). *Statistical Model for forecasting.* Wiley.

Anderson, B.D.O. and J.B. Moore (1979). *Optimal Filtering.* Prentice Hall Englewood.

Bandorff-Nielsen, O. and G. Shou (1973). On the Parametrization of Autoregressive Models by Partial Autocorrelations. *Journal of Multivariate Analysis, 3, 408-419.*

Benati, L. (2007). Drift and breaks in labor productivity, *Journal of Economic Dynamics and Control, 31(8), 2847-2877.*

Benati L. and H. Mumtaz (2007) The U.S. evolving macroeconomic dynamics: a structural investigation. *ECB- Working Paper Series 0746.*

Benveniste, A., Metivier, M. and P. Priouret (1990). *Adaptive algorithms and stochastic approximations.* Springer-Verlag.

Berkowitz, J. (2001). Testing Density Forecasts with Applications to Risk Management. *Journal of Business and Economic Statistics, 19, 465-474.*

Beveridge, S. and C.R. Nelson (1981). A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the business cycle. *Journal of Monetary Economics, 7, 151-174.*

Blasques F., Koopman S.J. and A. Lucas (2014). Time-Varying Temporal Dependence in Autoregressive Models: An Observation Driven Approach. Tinbergen Institute Discussion Papers.

Brainard, W. and G. Perry (2000). Making policy in a changing world. *In: Perry, G., Tobin, J. (Eds.), Economic Events, Ideas, and Policies: and Policies: The 1960s and After. Brookings Institution Press, Washington.*

Bollerslev, T. (1987). A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return. *Review of Economics and Statistics, 69, 542-547.*

Branch, W.A. and G.W. Evans (2006). Simple Recursive Forecasting model. *Economic Letter, 91, 158-166.*

Brown, R.G. (1963). *Smoothing, forecasting and prediction.* Prentice Hall.

Canova, F. and L. Gambetti (2009). Structural changes in the US economy: Is there a role for monetary policy? *Journal of Economic Dynamics and Control, Elsevier, 33(2), 477-490.*

Carceles-Poveda, E. and C. Giannitsarou (2007). Adaptive learning in practice. *Journal of Economic Dynamics and Control, 3, 2659-2669.*

Casals, J., Jerez, M. and S. Sotoca (2002). An Exact Multivariate Model-Based Structural Decomposition. *Journal of the American Statistical Association, 97, 553-564.*

Chan, J.C.C., Koop, G. and S.M. Potter (2013). A New Model of Trend Inflation. *Journal of Business and Economic Statistics, 31(1), 94-106.*

Chiu, Ching-Wai, H. Mumtaz and G. Pinter (2014). Fat-tails in VAR Models. *Working Papers 714, Queen Mary, University of London, School of Economics and Finance.*

Clark, T.E. and XX. Doh (2011). A Bayesian Evaluation of Alternative Models of Trend Inflation. *Kansas City FED WP2011-16.*

Clark, T.E. and F. Ravazzolo (2012). The macroeconomic forecasting performance of autoregressive models with alternative specifications of time-varying volatility. *WP-1218, Cleveland FED.*

Cogley, T. (2002). A Simple Adaptive Measure of Core Inflation. *Journal of Money, Credit and Banking, 34(1), 94-113.*

Cogley, T. and T.J. Sargent (2001). Evolving Post World War II U.S. In Inflation Dynamics. *NBER Macroeconomics Annual,16.*

Cogley, T. and T.J. Sargent (2005). Drifts and Volatilities: Monetary Policies and Outcomes in the Post WWII US. *Review of Economic Dynamics, 8(2), 262-302.*

Cogley, T., Primiceri, G. and T.J. Sargent (2010). Inflation-Gap persistence in the US. *American Economic Journal: Macroeconomics, 2(1), 43-69.*

Cooley, T. F. and E.C. Prescott (1973). An Adaptive Regression Model. *International Economic Review, 14(2), 364-71.*

Cooley, T. F. and E.C. Prescott (1976). Estimation in the Presence of Stochastic Parameter Variation. *Econometrica, 44(1), 167-84.*

Corradi, V. and N.R. Swanson (2006). Predictive Density and Conditional Confidence Interval Accuracy Tests. *Journal of Econometrics, 135, (1-2),187-228.*

Creal, D., Kopman, S.J. and A. Lucas (2012). Generalised Autoregressive Score Models with Applications. *Journal of Applied Econometrics, 28, 777-795.*

Cox, D.R. (1981). Statistical analysis of time series: some recent developments. *Scandinavian Journal of Statistics, 8(2), 93-115.*

Cúrdia, V., Del Negro, M. and D.L. Greenwald (2013). Rare shocks, Great Recessions. *San Francisco FED, WP 2013-01.*

D'Agostino, A., Gambetti, L. and D. Giannone (2013). Macroeconomic forecasting and structural change. *Journal of Applied Econometrics, 28, 82-101 .*

Dahlhaus, R. (2012). Locally stationary processes. *Handbook of Statistics, Time Series Analysis: Methods and Applications. Ed. T.S. Rao, S.S. Rao and C.R. Rao, (30), 351-413.*

Diebold, F.X , Gunther, T.A. and S.A. Tay (1998). Evaluating Density Forecasts with Applications to Financial Risk Management. *International Economic Review, 39(4), 863-83.*

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

Evans, G.W. and S. Honkapohja (1998). Convergence of learning algorithms without a projection facility. *Journal of Mathematical Economics, 30, 59-86.*

Evans, G.W. and S. Honkapohja (2001). *Learning and Expectations in Macroeconomics.* Princeton University Press.

Evans, G.W., Honkapohja, S. and N. Williams (2010). Generalized Stochastic Gradient Learning. *International Economic Review, 51, 237-262.*

Fagin, S.L. (1964). Recursive linear regression theory, optimal filter theory, and error analysis of optimal systems. *IEEE International Conv. Record, 12, 216-240.*

Faust, J. and J.H Wright (2013). Forecasting Inflation. *Handbook of Economic Forecasting, forthcoming, Vol 2A, Elliott G. and A. Timmermann Eds.* North Holland.

Fiorentini, G., Calzolari, G. and L. Panattoni (1996). Analytical Derivatives and the Computation of GARCH Models. *Journal of Applied Econometrics, 11, 399-417.*

Fiorentini, G., Sentana, E. and G. Calzolari (2003). Maximum Likelihood Estimation and Inference in Multivariate Conditionally Heteroskedastic Dynamic Regression Models with Student t Innovations. *Journal of Business and Economic Statistics, 21(4), 532-46.*

Giraitis, L., Kapetanios, G. and T. Yates (2011). Inference on stochastic time-varying coefficient models. *Queen Mary Univeristy of London, WP540.*

Giacomini, R. and H. White (2006). Tests of Conditional Predictive Ability. *Econometrica, 74(6), 1545-1578.*

Hamilton, J.D. (1986). A standard error for the estimated state vector of a state-space model. *Journal of Econometrics, 33(3), 387-397.*

Harvey, A.C. (1989). *Forecasting, structural time series models and Kalman filter.* Cambridge University Press.

Harvey, A.C. (2013). *Dynamic Models for Volatility and Heavy Tails. With Applications to Financial and Economic Time Series.* Cambridge University Press.

Harvey, A.C. and A. Luati (2012). Filtering with Heavy-Tails. *Cambridge WP1255.*

Hyndman, R.J., Koehler, A.B., Ord, J.K. and R.D. Snyder (2008). F*orecasting with Exponential Smoothing: The State Space Approach.* Springer.

Jazwinski, A.H. (1970). *Stochastic Processes and Filtering Theory.* Academic Press, San Diego.

Justiniano, A. and G. Primiceri (2008). The Time-Varying Volatility of Macroeconomic Fluctuations. *American Economic Review, 98(3), 604-641.*

Kim, C-J. and C. Nelson (1999). *State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications.* MIT Press.

Koop, G. and D. Korobilis (2009). Bayesian Multivariate Time Series Methods for Empirical Macroeconomics. *Foundations and Trends in Econometrics, 3 (4), 267–358.*

Koop, G. and D. Korobilis (2012). Large Time Varying Parameter VARs. *Journal of Econometrics, forthcoming.*

Koop, G. and S.M. Potter (2011). Time varying VARs with inequality restrictions. *Journal of Economic Dynamics and Control, 35(7), 1126-1138.*

Kozicki, S. and P.A. Tinsley (2001). Term structure views of monetary policy under alternative models of agent expectations. *Journal of Economic Dynamics and Control, 25(1-2), 149-184.*

Kreps, D. (1998). Anticipated Utility and Dynamic Choice. 1*997 Schwartz Lecture, in Frontiers of Research in Economic Theory, Edited by D.P. Jacobs, E. Kalai, and M. Kamien.* Cambridge University Press.

Li, H. (2008). Estimation and testing of Euler equation models with time-varying reduced-form coefficients. *Journal of Econometrics, 142, 425-448.*

Ljung, L. (1992). *Stochastic approximation and optimization of random system.* Georg Pflug, Harro Walk - Basel [etc.]. Birkhauser.

Ljung, L. (1999) *System Identification. Theory for the User.* Prentice Hall. 2nd Eds.

Ljung, L. and T. Soderstrom (1985). *Theory and Practice of Recursive Identification.* MIT Press.

Lucas, R. Jr (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy, 1(1),19-46.*

Marcet, A and T. Sargent (1989). Convergence of least squares learning mechanisms in self-referential linear stochastic models. *Journal of Economic Theory, 48(2), 337-368.*

Maronna, R.A., Martin, R. D. and V.J. Yohai (2006). *Robust Statistics: Theory and Methods.* Wiley Series in Probability and Statistics.

Monahan, J.F. (1984). A note on enforcing stationarity in autoregressive-moving average models. *Biometrika, 71(2), 403-404.*

Montgomery, D.C. and L.A. Johnson (1976). *Forecasting and time series analysis.* Mc Graw-Hill.

Pesaran, M.H. and A. Timmermann (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics, Elsevier, 137(1), 134-161.*

Pesaran, M.H. and A. Pick (2011). Forecast Combination across Estimation Windows. *Journal of Business Economics and Statistics, 29, 307-318.*

Pesaran, M.H., Pick, A. and M. Pranovich (2013). Optimal Forecasts in the Presence of Structural Breaks. *Journal of Econometrics (forthcoming).*

Pivetta, F. and R. Reis (2007). The persistence of inflation in the United States. *Journal of Economic Dynamics and Control, 31(4), 1326-1358.*

Primiceri, G. (2005). Time Varying Structural Vector Autoregressions and Monetary Policy. *The Review of Economic Studies, 72, 821-852.*

Rossi, B. and T. Sekhposyan (2013). Conditional Predictive Density Evaluation in the Presence of Instabilities. *WP688, Barcelona Graduate School of Economics.*

Sargent, T.J. (1999). *The Conquest of American Inflation.* Princeton University Press.

Sargent, T.J., and N. William (2005). Impacts of priors on convergence and escapes from Nash inflation. *Review of Economic Dynamics, 8, 360-391.*

Sargent, T.J., William, N. and T. Zha (2006). Shocks and Government Beliefs: The Rise and Fall of American Inflation, *American Economic Review, 96(4), 1193-1224.*

Slobodyan S. and R. Wouters (2012). Learning in an estimated medium-scale DSGE model. *Journal of Economic Dynamics and Control, 36(1), 26-46.*

Stock, J. and M. Watson (1996). Evidence on Structural Instability in Macroeconomic Time Series Relations. *Journal of Business and Economic Statistics, 14(1), 14-30.*

Stock, J. and M. Watson (2007). Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking 39(F), 3-33.*

Taylor, J. (2000). Low inflation, pass-through, and the pricing power of firms. *European Economic Review 44 (7), 1389–1408.*

Timmermann, A. (1996). Excess volatility and predictability of stock prices in autoregressive dividend models with learning. *The Review of Economic Studies, 63, 523-557.*

	Trend	Trend-B	AR(1)	AR(1)-B	AR(2)	AR(2)-B	AR(4)	AR(4)-B
				Normal				
κ_c	0.5309	0.2055	0.1483	0.0861	0.1168	0.0976	0.1360	0.0960
	(0.0214)	(0.0221)	(0.0227)	(0.0201)	(0.0215)	(0.0209)	(0.0224)	(0.0206)
κ_σ	0.1467	0.4870	0.1830	0.2831	0.1661	0.2669	0.2144	0.2729
	(0.0225)	(0.0203)	(0.0233)	(0.0232)	(0.0229)	(0.0233)	(0.0233)	(0.0230)
LogLik	-549.1139	-604.3270	-541.1469	-535.9191	-551.3741	-535.4122	-544.2799	-545.2302
AIC	1102.2277	1212.6541	1086.2937	1075.8381	1106.7481	1074.8245	1092.5598	1094.4603
BIC	1109.3491	1219.7754	1093.4151	1082.9595	1113.8695	1081.9458	1099.6811	1101.5817
				Student-t				
κ_c	0.4707	0.8197	0.1704	0.0735	0.0978	0.0683	0.1381	0.0867
	(0.0700)	(0.1649)	(0.0003)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
κ_σ	0.2116	0.2461	0.1697	0.2624	0.2171	0.2527	0.2598	0.2952
	(0.0847)	(0.1137)	(0.0003)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
υ	5.3309	5.8753	5.1371	4.2080	5.7377	4.7426	6.2070	5.6393
	(1.2107)	(1.4062)	(0.0019)	(0.0001)	(0.0005)	(0.0001)	(0.0006)	(0.0003)
LogLik	-523.1822	-561.7474	-519.5975	-520.5671	-526.5179	-520.7939	-520.5114	-521.3150
AIC	1052.3645	1129.4949	1045.1951	1047.1342	1059.0359	1047.5879	1047.0227	1048.6299
BIC	1063.0465	1140.1769	1055.8771	1057.8163	1069.7179	1058.2699	1057.7048	1059.3120

Table 1: Estimation of the annualized quarterly US-CPI inflation, $\pi_t = 400\Delta \log p_t$, sample 1955Q1-2012Q4. "Trend" denotes the specification without ARs coefficients ($p = 0$), "B" denotes the specifications with restricted long-run mean, and κ_c, κ_σ and υ are the static parameters (s.e. in brackets).

	RMSE			MAE		
	h=1	h=4	h=8	h=1	h=4	h=8
			Normal			
Trend	2.1829	2.8050	3.3144	1.4529	2.0313	2.3960
	—	—	—	—	—	—
Trend-B	1.4379	1.1448	0.9086	1.5986	1.1854	0.9780
	(0.0001)	(0.1685)	(0.4930)	(0.0000)	(0.0810)	(0.8676)
AR(1)	0.9652	0.9438	0.8658	0.9973	0.9497	0.8668
	(0.3171)	(0.3479)	(0.1694)	(0.9473)	(0.4926)	(0.1681)
AR(1)-B	1.0422	0.9616	0.8830	1.0922	0.9485	0.8857
	(0.4135)	(0.5723)	(0.3359)	(0.0809)	(0.5019)	(0.2873)
AR(2)	0.9787	0.9271	0.9158	0.9957	0.9451	0.9116
	(0.4128)	(0.0501)	(0.2053)	(0.8922)	(0.2811)	(0.2167)
AR(2)-B	1.0255	0.9372	0.8764	1.0596	0.9339	0.8587
	(0.2646)	(0.2053)	(0.2189)	(0.0608)	(0.2691)	(0.1013)
AR(4)	0.9492	0.9061	0.8892	0.9693	0.9287	0.8820
	(0.0966)	(0.0318)	(0.0726)	(0.4023)	(0.1703)	(0.0616)
AR(4)-B	1.1479	0.9617	0.8916	1.0828	0.9261	0.8600
	(0.3362)	(0.5566)	(0.3774)	(0.3023)	(0.2810)	(0.1437)
			Student-t			
Trend	1.0191	0.9878	0.9898	0.9856	0.9792	0.9735
	(0.6677)	(0.4429)	(0.7113)	(0.7342)	(0.2304)	(0.3427)
Trend-B	1.4013	1.1109	0.8800	1.4728	1.0932	0.9236
	(0.0007)	(0.2944)	(0.3666)	(0.0000)	(0.3389)	(0.5336)
AR(1)	0.9668	0.9597	0.8838	0.9889	0.9703	0.8740
	(0.2731)	(0.4611)	(0.2059)	(0.7687)	(0.6700)	(0.1760)
AR(1)-B	0.9570	0.9397	0.8644	0.9917	0.9017	0.8204
	(0.2325)	(0.4179)	(0.2843)	(0.8383)	(0.2188)	(0.0966)
AR(2)	1.0104	0.9562	0.9537	1.0187	0.9705	0.9364
	(0.6512)	(0.1769)	(0.4184)	(0.5512)	(0.5405)	(0.4004)
AR(2)-B	1.0148	0.9566	0.9127	1.0684	0.9345	0.8710
	(0.4743)	(0.4413)	(0.3643)	(0.0119)	(0.2692)	(0.1424)
AR(4)	0.9561	0.9083	0.8849	0.9714	0.9240	0.8696
	(0.1460)	(0.0171)	(0.0866)	(0.4586)	(0.1523)	(0.0818)
AR(4)-B	1.0229	0.9611	0.9290	1.0521	0.9432	0.8983
	(0.6411)	(0.5034)	(0.5418)	(0.2431)	(0.3545)	(0.2638)

Table 2: Point forecast 1973Q1–2012Q4. The RMSE and the MAE are expressed in relative term with respect to the benchmark model "Trend". "h" is the forecast horizon, in brackets the p-values of the Giacomini and White (2006) test.

	ALogS	LR	RS	ALogS	LR	RS
	Normal			Student-t		
Trend	-2.5591	0.1445	4.4223	-1.5897	0.7000	0.7023
Trend-B	-3.1073	0.0581	7.8323	-1.5688	0.0048	0.2723
AR(1)	-2.4857	0.0041	3.7823	-1.6325	0.9976	0.1322
AR(1)-B	-2.4543	0.0046	3.1923	-1.6766	0.9946	0.0423
AR(2)	-2.5357	0.0058	4.6922	-1.6249	0.5005	0.5522
AR(2)-B	-2.6275	0.4572	3.7210	-1.6671	0.7590	0.3610
AR(4)	-2.4663	0.1455	3.4810	-1.5976	0.7914	0.2560
AR(4)-B	-2.6022	0.9784	4.0960	-1.6272	0.7549	1.3323

Table 3: Density Forecast 1973Q1-2012Q4. The average log-score (AlogS), the p-values of the Likelihood Ratio (LR) test of Berkowitz (2001), and RS corresponds to the test of Rossi and Sekhposyan (2013) with critical values 2.25 (1%), 1.51 (5%), 1.1 (10%).

	Normal								Student-t						
	Trend	Trend-B	AR(1)	AR(1)-B	AR(2)	AR(2)-B	AR(4)	AR(4)-B	Trend	Trend-B	AR(1)	AR(1)-B	AR(2)	AR(2)-B	AR(4)
Normal															
Trend-B	0.000														
AR(1)	0.0458	0.0000													
AR(1)-B	0.0431	0.0000	0.4027												
AR(2)	0.4218	0.0000	0.0025	0.0490											
AR(2)-B	0.4587	0.0001	0.2042	0.1411	0.3829										
AR(4)	0.0555	0.0000	0.5724	0.7961	0.0536	0.2042									
AR(4)-B	0.4238	0.0000	0.0786	0.0289	0.2972	0.8402	0.0135								
Student-t															
Trend	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000							
Trend-B	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.6531						
AR(1)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2808	0.1554					
AR(1)-B	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0767	0.0209	0.0853				
AR(2)	0.0000	0.0000	0.000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2544	0.3086	0.8184	0.2434			
AR(2)-B	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0385	0.0386	0.2610	0.7604	0.1609		
AR(4)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.8341	0.5756	0.3295	0.0596	0.3486	0.0592	
AR(4)-B	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.3494	0.2119	0.8971	0.2708	0.9568	0.2987	0.3476

Table 4: Pairwise comparison: p-values for the difference in the average log-scores (Amisano and Giacomini, 2007).

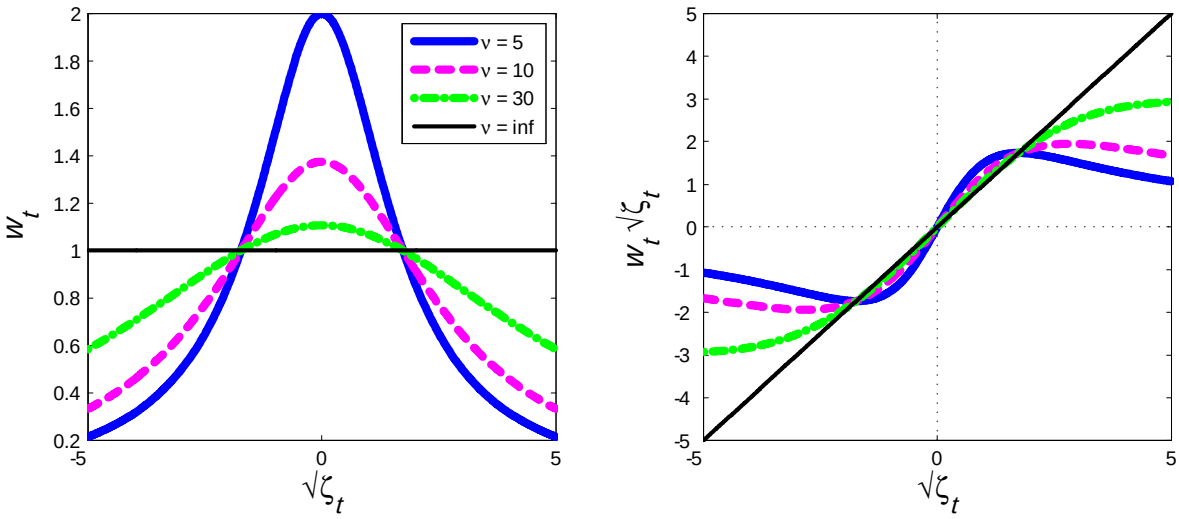

Figure 1: Blue (thick) line $\upsilon = 5$, pink (dash) $\upsilon = 10$, green (dots) $\upsilon = 30$ and black (thin) $\upsilon = \infty$ (Gaussian), w_t are the weights and $\sqrt{\zeta_t} = \varepsilon_t/\sigma_{t|t-1}$ are the standardized prediction error.

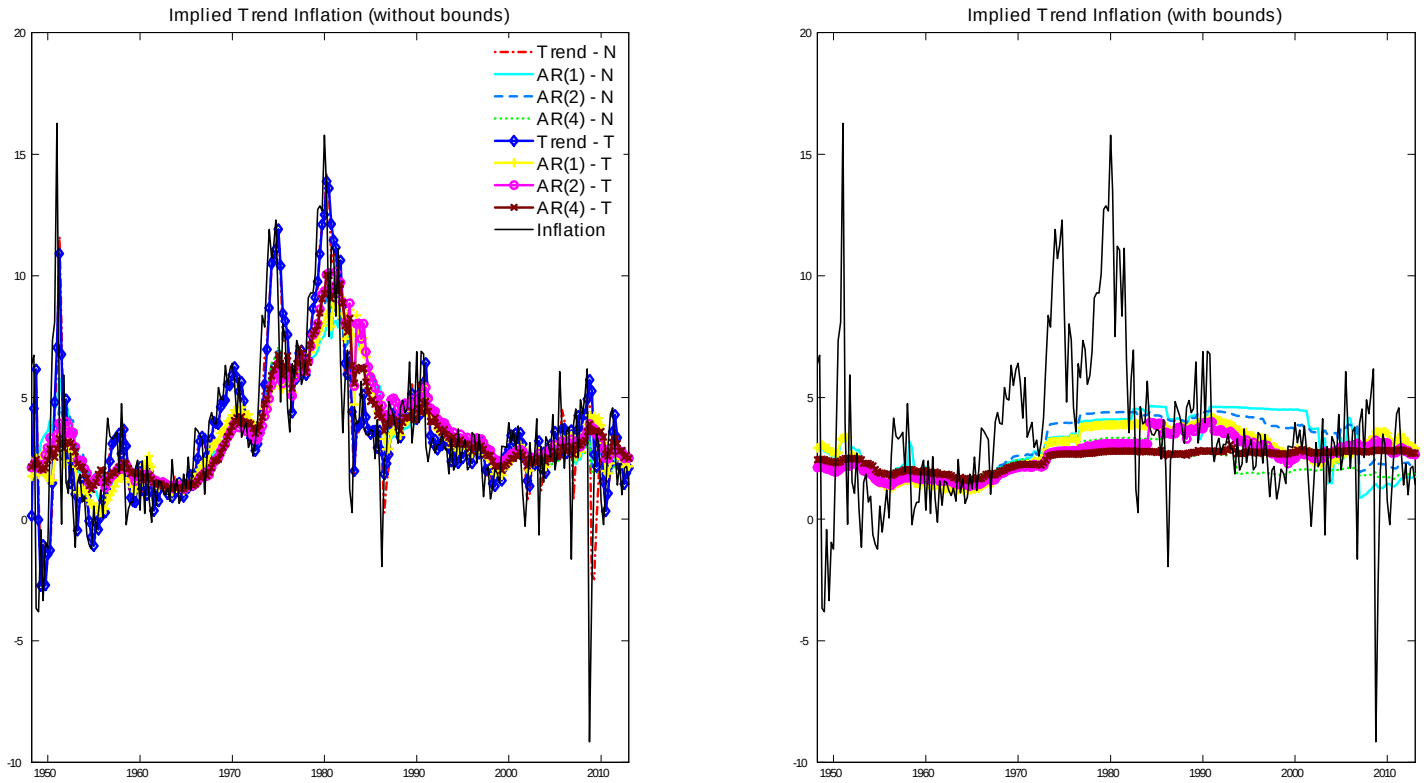

Figure 2: Implied trend-inflation and realized inflation. "N" denotes Gaussian distribution while "T" Student-t distribution. On the right panel we exclude the trend-only specification which reaches the upper bound during the great inflation.

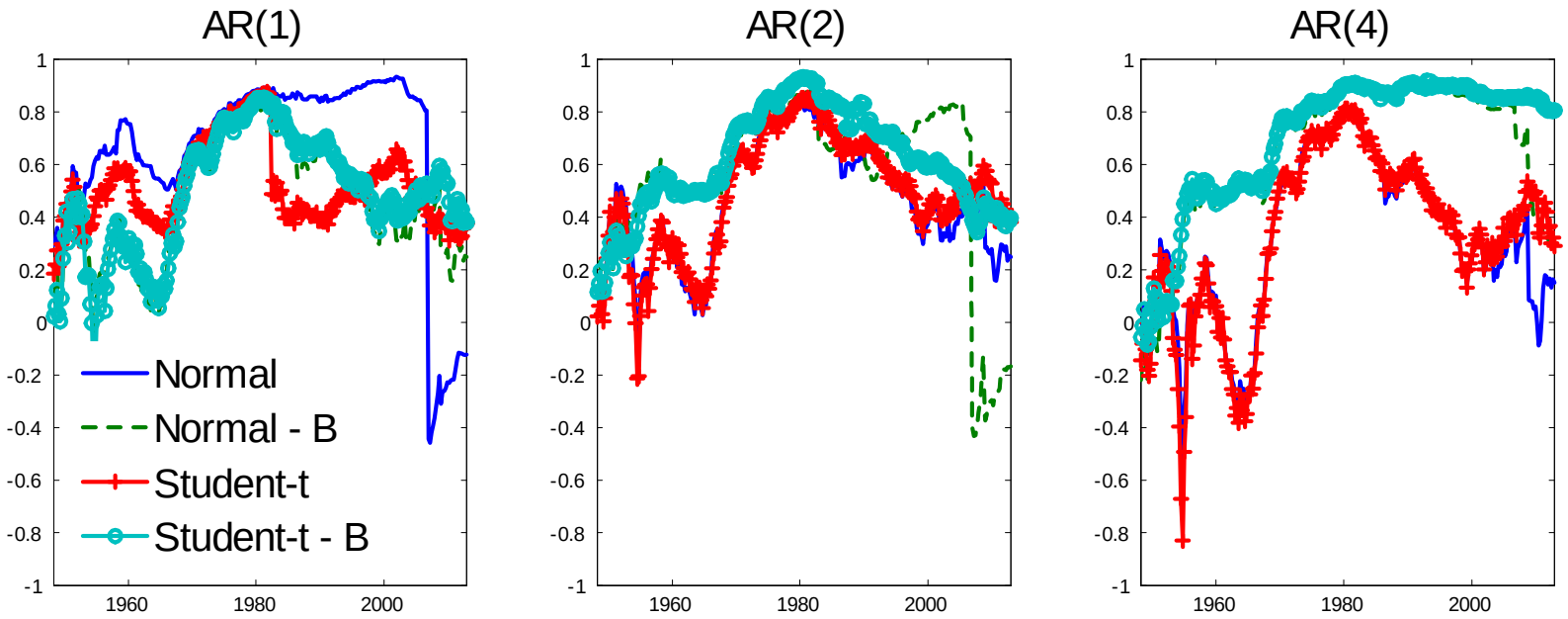

Figure 3: Sum of ARs coefficients, "B" denotes the specifications with bounded long-run mean.

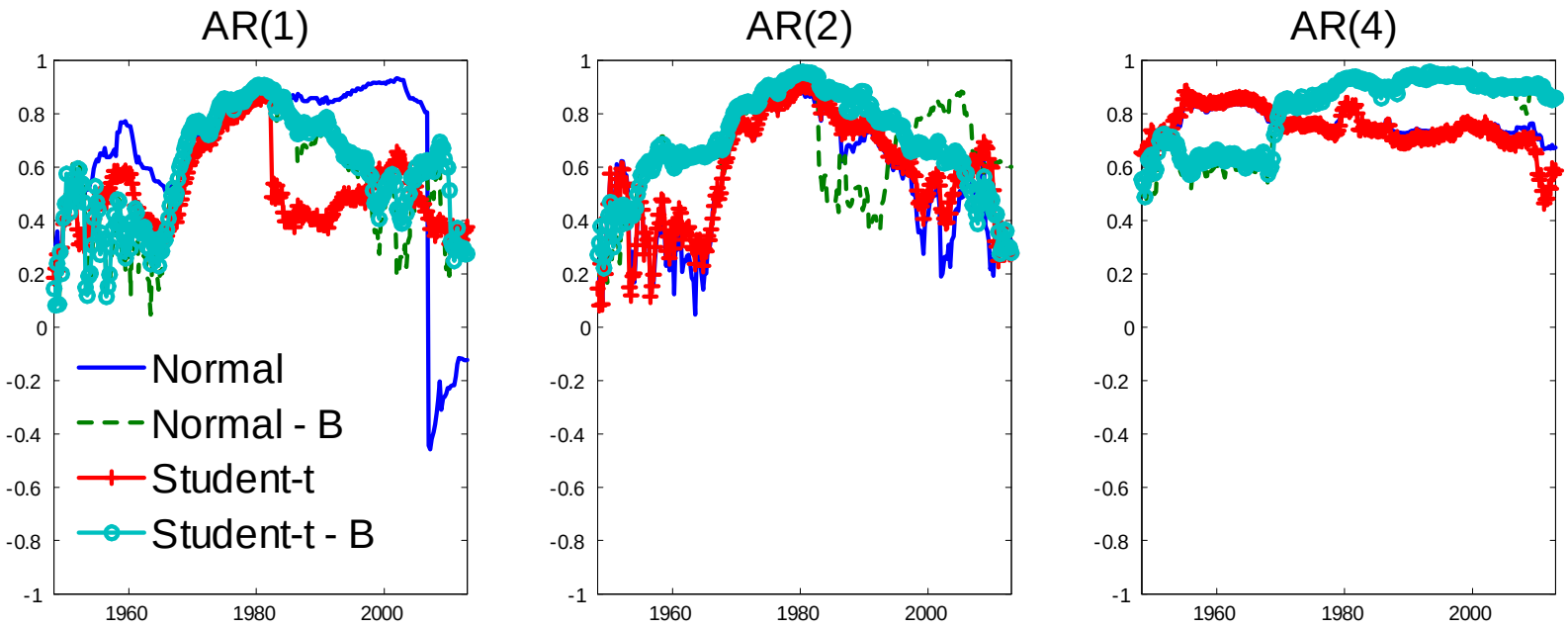

Figure 4: Largest eigenvalue, "B" denotes the specifications with bounded long-run mean.

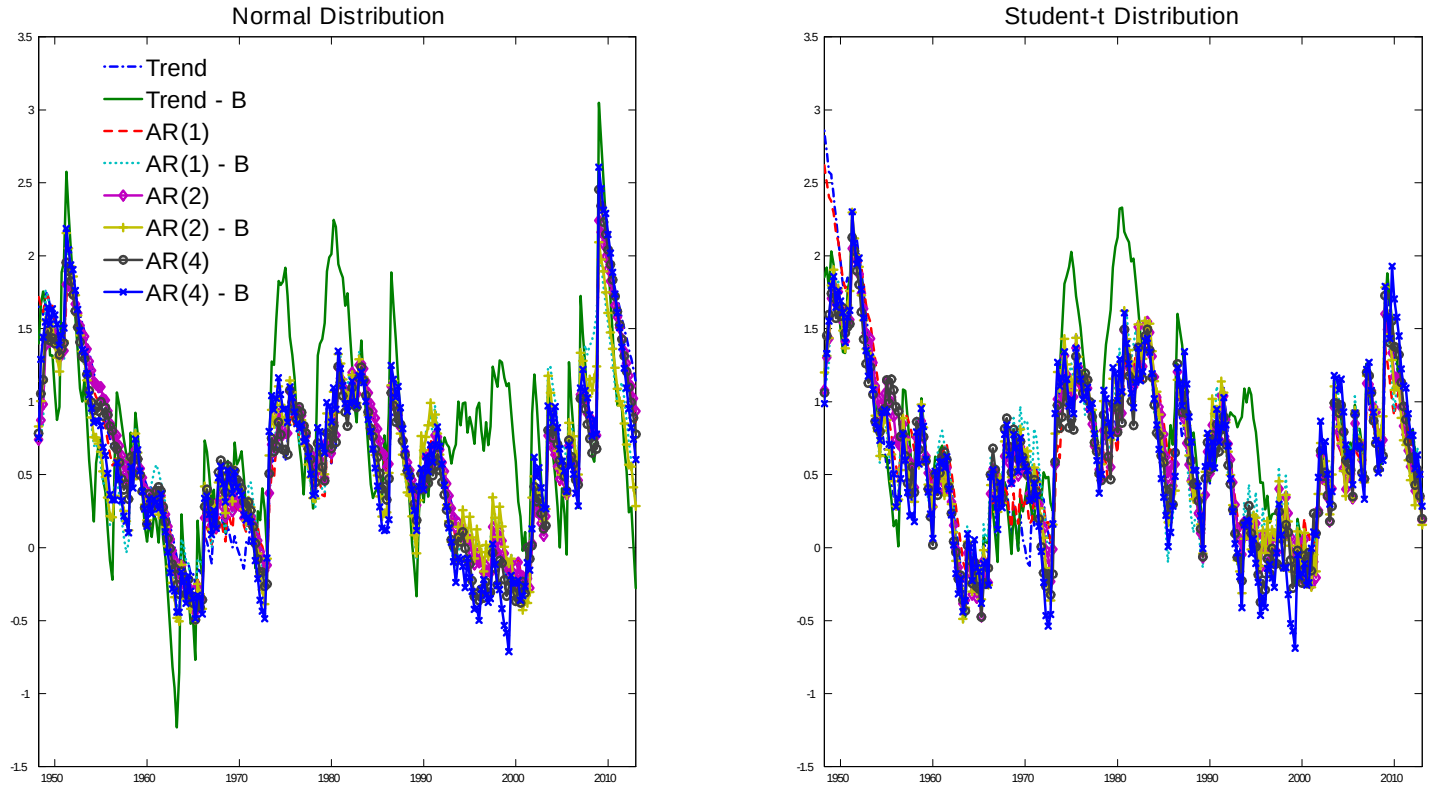

Figure 5: Volatility $\log \sigma_{t|t-1}$ for different specifications.

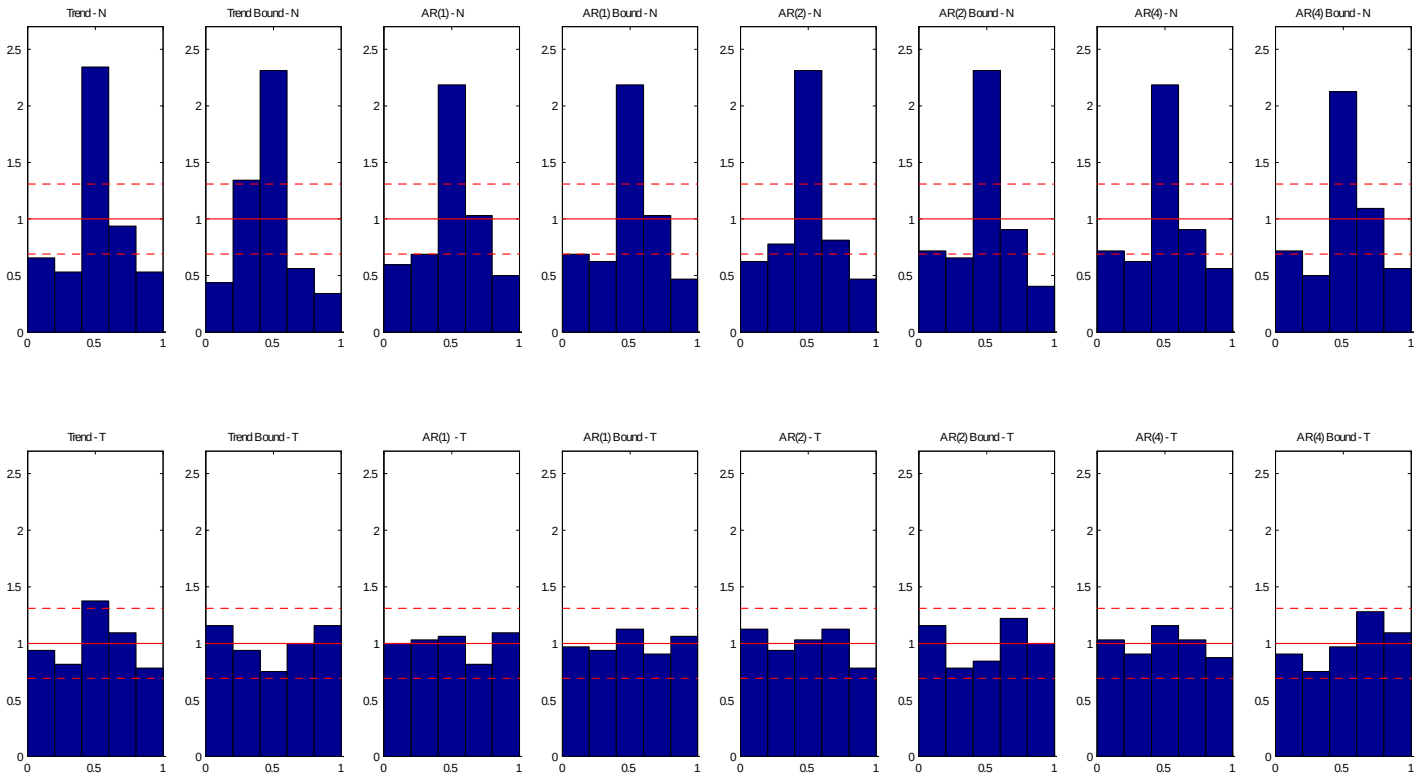

Figure 6: The p.d.f. of the PITs (normalized) and the 95% critical values (dashed lines) approximated by binomial distribution, constructed using a normal approximation as in Diebold et al. (1998).

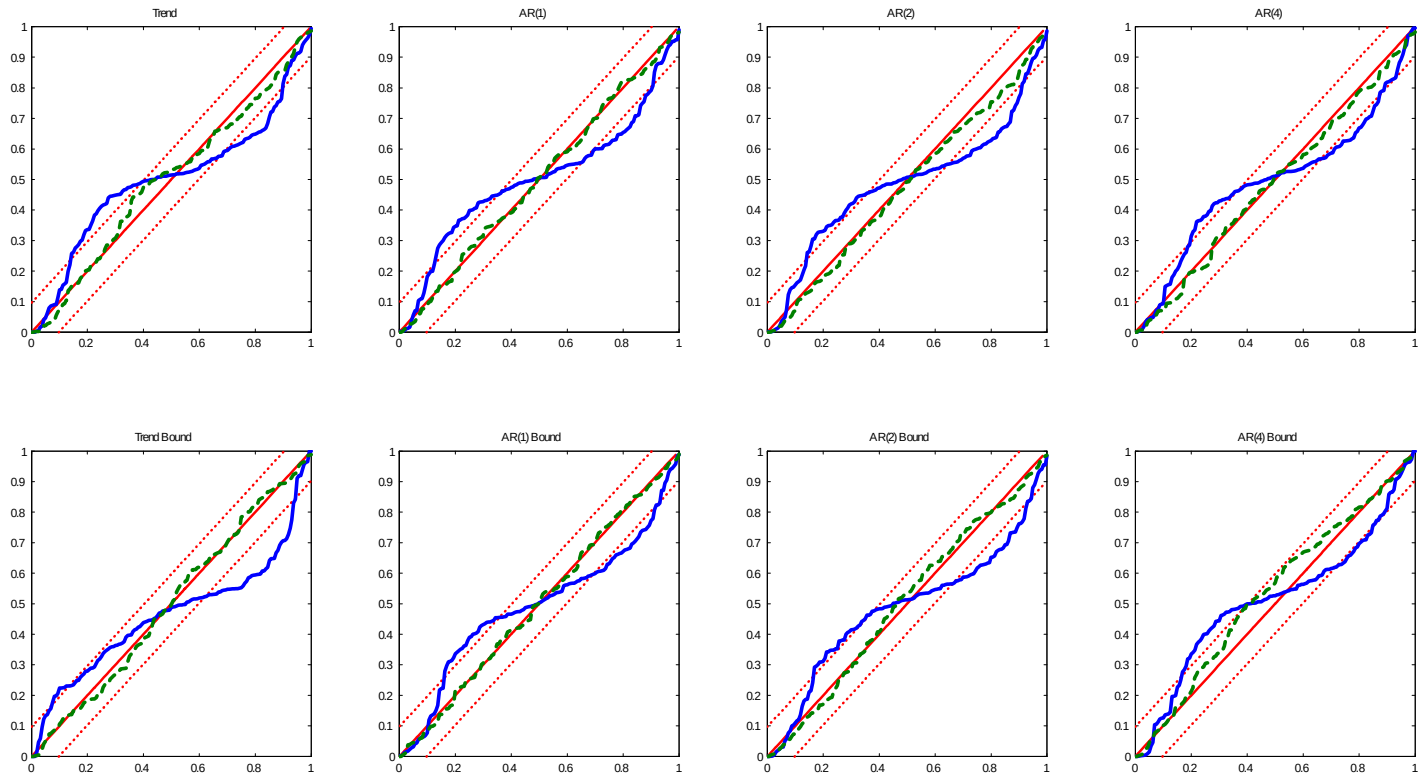

Figure 7: The c.d.f. of the PITs and the 95% critical values based on Rossi and Sekhposyan (2013). Solid (blue) line for Gaussian distribution and dashed (green) line for Student-t distribution.

Appendix A: Proofs

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

$$S_t(\boldsymbol{\phi}_t) = \sum_{j=1}^{t} \gamma_j \left(y_{t-j} - \mathbf{x}'_{t-j}\boldsymbol{\phi}_t\right)^2,$$

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 $\left[\sum_{j=1}^{t} \gamma_j\right]^{-1} \to \kappa$. 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 κ regulates the tracking ability (large κ) and the noise insensitivity (small κ). 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 $\boldsymbol{\eta}_t = c(\boldsymbol{\phi}_{t|t} - \boldsymbol{\phi}_t)$, where $c = [\kappa/(1-\kappa)]^{1/2}$. Consequently, the transition equation in (14) is equal to $\boldsymbol{\phi}_{t+1} = (1-c)\boldsymbol{\phi}_t + c\boldsymbol{\phi}_{t|t}$ and the true state vector can be expressed as exponential weighted average of past filter estimates

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

Moreover, the filter estimate can be expressed as

$$\boldsymbol{\phi}_{t|t} = \mathbf{L}_t\boldsymbol{\phi}_{t-1|t-1} + \mathbf{K}_t y_t = \sum_{j=0}^{t-1}\left(\prod_{i=0}^{j-1}\mathbf{L}_{t-i}\right)\mathbf{K}_{t-j}y_{t-j}$$

where

$$\mathbf{L}_t = (\mathbf{I} - \mathbf{K}_t\mathbf{x}'_t), \quad \mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{x}_t\left(\mathbf{x}'_t\mathbf{P}_{t|t-1}\mathbf{x}_t + \frac{\sigma^2}{1-\kappa}\right)^{-1}.$$

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 $\mathbf{Q}_t := \kappa^2\boldsymbol{\Sigma}$, with $\boldsymbol{\Sigma} = \sigma^2\mathbb{E}[(\mathbf{x}_t\mathbf{x}'_t)]^{-1}$, we have that the shock driving the time-

varying coefficients is

$$\boldsymbol{\eta}_t = \kappa(\mathbf{x}_t\mathbf{x}_t')^{-1}\mathbf{x}_t\varepsilon_t = \kappa(\mathbf{x}_t\mathbf{x}_t')^{-1}\mathbf{x}_t\varepsilon_t.$$

Therefore, the parameter-driven model collapses to an observation-driven model. Moreover, up to a scalar factor, the shock $\boldsymbol{\eta}_t$ is equal to the driving process of our score driven model. However, under the parameter driven framework the vector of coefficients is considered as unobserved state vector which is optimally estimated by the mean of KF which leads to

$$\begin{aligned}
\boldsymbol{\phi}_{t+1|t} &= \boldsymbol{\phi}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{x}_t(\mathbf{x}_t'\mathbf{P}_{t|t-1}\mathbf{x}_t + \sigma^2)^{-1}(y_t - \mathbf{x}_t'\boldsymbol{\phi}_{t|t-1}) \\
\mathbf{P}_{t+1|t} &= \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{x}_t(\mathbf{x}_t'\mathbf{P}_{t|t-1}\mathbf{x}_t + \sigma^2)^{-1}\mathbf{x}_t'\mathbf{P}_{t|t-1} + \kappa^2\boldsymbol{\Sigma}.
\end{aligned}$$

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 $t > \mathring{t}$, where $\mathring{t}$ is a given large value of t, one has the approximation $(\mathbf{x}_t'\mathbf{P}_{t|t-1}\mathbf{x}_t + \sigma^2) \approx \sigma^2$, this implies that the conditional variance of the forecast error converges to the variance of model disturbances. For t large enough, the variation of $\mathbf{P}_{t|t-1}$ is small with respect to $\mathbf{x}_t$, and $\mathbf{x}_t'\mathbf{P}_{t|t-1}\mathbf{x}_t$ can be neglected with respect to σ^2. Using these approximations, we obtain

$$\begin{aligned}
\boldsymbol{\phi}_{t+1|t} &= \boldsymbol{\phi}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{x}_t\sigma^{-2}(y_t - \mathbf{x}_t'\boldsymbol{\phi}_{t|t-1}) \\
\mathbf{P}_{t+1|t} &= \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{x}_t\sigma^{-2}\mathbf{x}_t'\mathbf{P}_{t|t-1} + \kappa^2\boldsymbol{\Sigma}.
\end{aligned}$$

Replacing $\mathbf{x}_t\mathbf{x}_t'/\sigma^2$ with its expected value $\boldsymbol{\Sigma}^{-1}$ we obtain $\mathbf{P}_{t+1|t} = \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\boldsymbol{\Sigma}^{-1}\mathbf{P}_{t|t-1} + \kappa^2\boldsymbol{\Sigma}$. When $\mathbf{P}_{t|t-1}$ is set to its steady-state value $\mathbf{P}$ as in Harvey (1989, p. 118), one has $\mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P} = \boldsymbol{\Lambda}\boldsymbol{\Sigma}\boldsymbol{\Lambda} \Rightarrow \kappa^{-2}\mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P} = \boldsymbol{\Sigma} \Rightarrow \kappa^{-1}\mathbf{P} = \boldsymbol{\Sigma}$. Using last expression the recursion for the vector of coefficients is

$$\boldsymbol{\phi}_{t+1|t} = \boldsymbol{\phi}_{t|t-1} + \kappa\boldsymbol{\Sigma}\mathbf{x}_t\sigma^{-2}(y_t - \mathbf{x}_t'\boldsymbol{\phi}_{t|t-1}),$$

which has the same asymptotic behavior of the CGL; see Sargent and William (2005) and Evans et al (2010). Similarly, setting $\mathbf{Q}_t := \kappa^2\boldsymbol{\Sigma}^{-1}$, we have that $\boldsymbol{\eta}_t = \kappa\mathbf{x}_t\varepsilon_t$ and the parameter-driven model collapses to an observation-driven model. In the steady-state $\kappa^{-1}\mathbf{P} = \mathbf{I}$ and the recursion for the coefficients is

$$\boldsymbol{\phi}_{t+1|t} = \boldsymbol{\phi}_{t|t-1} + \kappa\mathbf{x}_t\sigma^{-2}(y_t - \mathbf{x}_t'\boldsymbol{\phi}_{t|t-1}).$$

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

$$\ell_t(\mathcal{F}_t, \boldsymbol{\theta}) = c + d_t + g_t$$

with

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

and

$$d_t = -\frac{1}{2}\log\sigma^2_{t|t-1},\ g_t = -\left(\frac{\eta+1}{2\eta}\right)\log\left[1+\frac{\eta}{1-2\eta}\zeta_t\right],$$

where $\zeta_t = \varepsilon_t^2/\sigma^2_{t|t-1}$ and Γ is the Euler's gamma function. Let $\nabla_t = \partial\ell_t(\mathcal{F}_t, \boldsymbol{\theta})/\partial\mathbf{f}_{t|t-1}$ denote the gradient function and partition it in two blocks, $\nabla_{\boldsymbol{\phi}}$ and ∇_σ, the first one depend upon g_t and ζ_t, while the second upon d_t, g_t and ζ_t. We have to compute $\frac{\partial g_t}{\partial \boldsymbol{\phi}_t'} = \frac{\partial g_t}{\partial \zeta_t}\frac{\partial \zeta_t}{\partial \boldsymbol{\phi}_t'}$, where

$$\begin{aligned}\frac{\partial g_t}{\partial \zeta_t} &= -\left(\frac{\eta+1}{2\eta}\right)\frac{\partial \ln\left[1+\frac{\eta}{1-2\eta}\zeta_t\right]}{\partial\zeta_t} = -\left(\frac{\eta+1}{2\eta}\right)\left[1+\frac{\eta}{1-2\eta}\zeta_t\right]^{-1}\left(\frac{\eta}{1-2\eta}\right)\\ &= -\frac{\eta+1}{2\left(1-2\eta\right)}\left[\frac{1-2\eta+\eta\zeta_t}{1-2\eta}\right]^{-1} = -\frac{\eta+1}{2(1-2\eta+\eta\zeta_t)}\end{aligned}$$

and $\frac{\partial \zeta_t}{\partial \boldsymbol{\phi}_t'} = -\frac{2\mathbf{x}_t'\varepsilon_t}{\sigma^2_{t|t-1}}$. The score for the coefficients of the model is then equal to

$$\nabla_{\boldsymbol{\phi}} = \frac{\partial g_t}{\partial \zeta_t}\frac{\partial \zeta_t}{\partial \boldsymbol{\phi}_t} = \mathbf{x}_t\frac{(\eta+1)\,\varepsilon_t/\sigma^2_{t|t-1}}{(1-2\eta+\eta\varepsilon_t^2/\sigma^2_{t|t-1})}.$$

The gradient for the variance component is

$$\nabla_\sigma = \frac{\partial d_t}{\partial \sigma_t^2} + \frac{\partial g_t}{\partial \sigma_t^2} = \frac{\partial d_t}{\partial \sigma_t^2} + \frac{\partial g_t}{\partial \zeta_t}\frac{\partial \zeta_t}{\partial \sigma_t^2},$$

where $\frac{\partial \zeta_t}{\partial \sigma_t^2} = -\frac{\varepsilon_t^2}{\sigma^4_{t|t-1}}$ and thus we obtain

$$\nabla_\sigma = -\frac{1}{2\sigma^2_{t|t-1}} + \frac{(\eta+1)\varepsilon_t^2/\sigma^4_{t|t-1}}{2(1-2\eta+\eta\zeta_t)} = \frac{1}{2\sigma^4_{t|t-1}}\left[\frac{(\eta+1)}{(1-2\eta+\eta\zeta_t)}\varepsilon_t^2 - \sigma^2_{t|t-1}\right].$$

We compute the information matrix as $\mathcal{I}_t = -E_t(\mathbf{H}_t)$, where $\mathbf{H}_t$ the Hessian matrix and it can be partitioned in four blocks

$$\mathbf{H}_t = \begin{bmatrix}\mathbf{H}_{\phi\phi,t} & \mathbf{H}_{\phi\sigma,t}\\ \mathbf{H}_{\phi\sigma,t}' & \mathbf{H}_{\sigma\sigma,t}\end{bmatrix}.$$

The first block $\mathbf{H}_{\phi\phi,t}$ can be calculated as

$$\mathbf{H}_{\phi\phi,t} = \frac{\partial \nabla_{\boldsymbol{\phi},t}}{\partial \boldsymbol{\phi}'_{t|t-1}} = \frac{(1+\eta)\left[\eta\zeta_t + 2\eta - 1\right]}{(1-2\eta+\eta\zeta_t)^2}\frac{\mathbf{x}_t\mathbf{x}_t'}{\sigma^2_{t|t-1}}.$$

Following Fiorentini et al. (2003), recalling that $\varepsilon_t/\sigma_{t|t-1} = \zeta_t^{1/2} \sim t_\upsilon(0,1)$ implies that $\zeta_t^{1/2} = \sqrt{\frac{(\upsilon-2)\varsigma_t}{\xi_t}}u_t$, where u_t is uniformly distributed on the unit set, ς_t is a chi-squared

random variable with 1 degree of freedom, ξ_t is a gamma variate with mean $\upsilon > 2$ variance 2υ, and u_t, ς_t and ξ_t are mutually independent. Therefore, it is possible to show that

$$\mathcal{I}_{\phi\phi,t} = -E(\mathbf{H}_{\phi\phi,t}) = \frac{(1+\eta)}{(1-2\eta)(1+3\eta)} \frac{\mathbf{x}_t \mathbf{x}_t'}{\sigma^2_{t|t-1}}.$$

The cross-derivative term in the Hessian is $\mathbf{H}_{\sigma\phi,t} = -\frac{\mathbf{x}_t \varepsilon_t}{\sigma^4_{t|t-1}}$ and therefore $\mathcal{I}_{\phi\sigma,t} = -E(\mathbf{H}_{\sigma\phi,t}) = 0$.

$$\mathbf{H}_{\sigma\sigma,t} = \frac{\partial^2 \ell_t}{\partial^2 \sigma^2_{t|t-1}} = \frac{\partial \nabla_\sigma}{\partial \sigma^2_{t|t-1}} = \frac{1}{2\sigma^4_{t|t-1}} - \frac{\left[2(1-2\eta) + \eta\varepsilon_t^2/\sigma^2_{t|t-1}\right](\eta+1)\,\varepsilon_t^2/\sigma^6_{t|t-1}}{2[1-2\eta+\eta\varepsilon_t^2/\sigma^2_{t|t-1}]^2},$$

it is possible to show that

$$\mathcal{I}_{\sigma\sigma,t} = -E_t(\mathbf{H}_{\sigma\sigma,t}) = \frac{(1+\eta)}{2(3+\eta)\,\sigma^4_{t|t-1}} - \frac{\eta}{2(3+\eta)\,\sigma^4_{t|t-1}} = \frac{1}{2(1+3\eta)\,\sigma^4_{t|t-1}}.$$

Finally, the information matrix is equal to

$$\mathcal{I}_t = \begin{bmatrix} \frac{(1+\eta)}{(1-2\eta)(1+3\eta)} \mathbf{x}_t \sigma^{-2}_{t|t-1} \mathbf{x}_t' & \mathbf{0} \\ \mathbf{0}' & \frac{1}{2(1+3\eta)\sigma^4_{t|t-1}} \end{bmatrix},$$

and the final expression for the scaled score vector is

$$\mathbf{s}_t = \mathcal{I}_t^{-1} \nabla_t = \begin{bmatrix} \mathbf{s}_{\phi t} \\ s_{\sigma t} \end{bmatrix} = \begin{bmatrix} \frac{(1-2\eta)(1+3\eta)}{(1-2\eta+\eta\zeta_t)} (\mathbf{x}_t \sigma^{-2}_{t|t-1} \mathbf{x}_t')^{-1} \mathbf{x}_t \sigma^{-2}_{t|t-1} \varepsilon_t \\ (1+3\eta)\left[\frac{(1+\eta)}{(1-2\eta+\eta\zeta_t)} \varepsilon_t^2 - \sigma^2_{t|t-1}\right] \end{bmatrix}.$$

Proposition 1 Under Student-t distribution the driving process is (17)-(19) and the coefficients' updating rule is

$$\boldsymbol{\phi}_{t+1|t} = \boldsymbol{\phi}_{t|t-1} + \kappa_\phi \frac{(1-2\eta)(1+3\eta)}{(1+\eta)} (\mathbf{x}_t \sigma^{-2}_{t|t-1} \mathbf{x}_t')^{-1} \mathbf{x}_t \sigma^{-2}_{t|t-1} [w_t (y_t - \mathbf{x}_t' \boldsymbol{\phi}_{t|t-1})],$$

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

$$y_t = \mu_{t|t-1} + \varepsilon_t, \quad \varepsilon_t \sim t_\upsilon(0, \sigma^2_{t|t-1})$$

and the estimated level is

$$\begin{aligned}
\mu_{t+1|t} &= \mu_{t|t-1} + \kappa_\theta w_t(y_t - \mu_{t|t-1}) \\
&= (1 - \kappa_\theta w_t)\mu_{t|t-1} + \kappa_\theta w_t y_t \\
&= \frac{\kappa_\theta}{1 - \kappa_\theta w_t L} w_t y_t \\
&= \kappa_\theta \sum_{j=0}^{\infty} \gamma_j w_{t-j} y_{t-j},
\end{aligned}$$

with $\kappa_\theta = \kappa_\phi \frac{(1-2\eta)(1+3\eta)}{(1+\eta)}$. 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 (21) 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 (22) and the implied weighting pattern.

Theorem 1 Given the non-linear state space model

$$\begin{aligned}
y_t &= \mathbf{x}_t' \boldsymbol{\phi}_t + \varepsilon_t,\ \varepsilon_t \sim N\left(0, \sigma^2\right), \\
\boldsymbol{\alpha}_{t+1} &= \boldsymbol{\alpha}_t + \boldsymbol{\eta}_t,\ \boldsymbol{\eta}_t \sim N\left(\mathbf{0}, \mathbf{Q}_t\right),
\end{aligned} \tag{35}$$

with $\boldsymbol{\phi}_t = g(\boldsymbol{\alpha}_t)$. We can solve it by the mean of the Extended Kalman filter

$$\begin{aligned}
v_t &= y_t - \mathbf{x}_t' \boldsymbol{\phi}_{t|t-1}, \\
\mathbf{K}_t &= \mathbf{P}_{t|t-1} \widetilde{\mathbf{x}}_t \mathbf{F}_t^{-1}, \\
\mathbf{F}_t &= \widetilde{\mathbf{x}}_t' \mathbf{P}_{t|t-1} \widetilde{\mathbf{x}}_t + \sigma^2 \\
\boldsymbol{\alpha}_{t+1|t} &= \boldsymbol{\alpha}_{t|t-1} + \mathbf{K}_t v_t, \\
\mathbf{P}_{t+1|t} &= \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1} \widetilde{\mathbf{x}}_t \mathbf{F}_t^{-1} \widetilde{\mathbf{x}}_t' \mathbf{P}_{t|t-1} + \mathbf{Q}_t,
\end{aligned}$$

where $\widetilde{\mathbf{x}}_t' = \mathbf{x}_t' \frac{\partial g(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}'}|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_{t|t-1}} = \mathbf{x}_t' \Psi_t$. Setting $\sigma^2 = \frac{\sigma^2}{1-\kappa}$and $\mathbf{Q}_t = \mathbf{P}_{t|t} \frac{\kappa}{1-\kappa}$ and following same approach in Ljung (1992, p. 99) and Sargent (1999, p. 115), we obtain the modified version of the CGL algorithm

$$\begin{aligned}
\boldsymbol{\alpha}_{t+1|t} &= \boldsymbol{\alpha}_{t|t-1} + \kappa \mathbf{R}_t^{-1} \Psi_t' \mathbf{x}_t \sigma^{-2} (y_t - \mathbf{x}_t' \boldsymbol{\phi}_{t|t-1}), \\
\mathbf{R}_t &= (1 - \kappa)\mathbf{R}_{t-1} + \kappa(\Psi_t' \mathbf{x}_t \sigma^{-2} \mathbf{x}_t' \Psi_t).
\end{aligned}$$

This is exactly the score-driven filter (30), where the information matrix $\Psi_t' \mathbf{x}_t \sigma^{-2} \mathbf{x}_t' \Psi_t$ is replaced by its smoothed version $\mathbf{R}_t$.

Theorem 2 For simplicity we drop the temporal subscript t such that the $p \times p$ Jacobian matrix is

$$\Gamma = \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\pi}'} = \frac{\partial \Phi(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}'}.$$

The first $(p-1)$ coefficients are obtained from last recursion in (25), and the last coefficients is equal to the last partial autocorrelation π_p. We denote the final vector of coefficients as $\boldsymbol{\phi}_p = (\phi^{1,p}, ..., \phi^{p-1,p}, \phi^{p,p})' = (\mathbf{a}_p', \pi_p)$, where $\mathbf{a}_p = (\phi^{1,p}, ..., \phi^{p-1,p})$ and $\phi^{p,p} = \pi_p$. Therefore, we can express the last iteration of (25) in matrix form $\mathbf{a}_p = \mathbf{J}_{p-1} \boldsymbol{\phi}_{p-1}$, where $\boldsymbol{\phi}_{p-1} = (\phi^{1,p-1}, ..., \phi^{p-2,p-1}, \phi^{p-1,p-1})' = (\mathbf{a}_{p-1}', \pi_{p-1})'$ and

$$\mathbf{J}_{p-1} = \begin{bmatrix} 1 & 0 & \cdots & 0 & -\pi_p \\ 0 & \ddots & & & 0 \\ \vdots & & . & & \vdots \\ 0 & & & \ddots & 0 \\ -\pi_p & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Note that if p is even the central element of $\mathbf{J}_{p-1}$ is $1 - \pi_p$. Moreover, the vector $\widetilde{\boldsymbol{\phi}}_p = (\boldsymbol{\phi}_{p-1}', \pi_p)'$ contains all the partial autocorrelations, i.e. $\widetilde{\boldsymbol{\phi}}_p = (\mathbf{a}_{p-1}', \pi_{p-1}, \pi_p)$ and keep substituting we obtain $\widetilde{\boldsymbol{\phi}}_p = \boldsymbol{\pi}_p = (\pi_1, ..., \pi_{p-1}, \pi_p)$. The Jacobian matrix can be expressed as follows

$$\Gamma = \Gamma_p = \begin{bmatrix} \frac{\partial \mathbf{a}_p}{\partial \boldsymbol{\phi}_{p-1}'} & \frac{\partial \mathbf{a}_p}{\partial \pi_p} \\ \frac{\partial \pi_p}{\partial \boldsymbol{\phi}_{p-1}'} & \frac{\partial \pi_p}{\partial \pi_p} \end{bmatrix}.$$

The upper-left block is a $(p-1) \times (p-1)$ matrix and it can be computed using the definition $\mathbf{a}_p = \mathbf{J}_{p-1} \boldsymbol{\phi}_{p-1}$; since $\mathbf{J}_{p-1}$ contains the last partial correlation π_p we have the recursive formulation

$$\frac{\partial \mathbf{a}_p}{\partial \boldsymbol{\phi}_{p-1}'} = \mathbf{J}_{p-1} \Gamma_{p-1}$$

where $\Gamma_{p-1} = \partial \boldsymbol{\phi}_{p-1} / \partial \boldsymbol{\pi}_{p-1}$ 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

$$\frac{\partial \pi_p}{\partial \mathbf{a}_{p-1}'} = \mathbf{0}', \ \frac{\partial \pi_p}{\partial \pi_p} = 1 \text{ and } \frac{\partial \mathbf{a}_p}{\partial \pi_p} = \frac{\partial \mathbf{J}_{p-1}}{\partial \pi_p} \boldsymbol{\phi}_{p-1} = \begin{bmatrix} -\phi^{p-1,p-1} \\ -\phi^{p-2,p-1} \\ \vdots \\ -\phi^{1,p-1} \end{bmatrix}.$$

Note that $\boldsymbol{\phi}_{p-1}$ is a given and $\frac{\partial \mathbf{J}_{p-1}}{\partial \pi_p} = antidiag(-1, ..., -1)$ inverts the order of elements in $\boldsymbol{\phi}_{p-1} = (\phi^{1,p-1}, ..., \phi^{p-2,p-1}, \phi^{p-1,p-1})'$ with opposite sign.

Appendix B: Robustness

Section 4 shows that, in presence of heavy tails, the adaptive algorithm developed in this paper delivers a model-consistent penalization of the outliers. In fact, the estimated time variation in the parameters is such that the observations are downweighted when they are too large. In this appendix we assess the importance of using the law of motion of the parameters consistent with the score-driven model in presence of heavy-tails. In order to achieve this goal, we compare the density forecast of the specifications under Student-t innovations to two 'misspecified' cases. Firstly, we consider the case where the dynamic of the parameters is driven by the law of motion under Normal distribution (10)-(12) but we assume that the appropriate density is the Student-t; this is similar in spirit to the t-GARCH model of Bollerslev (1987) and it is labelled "Miss1". Secondly, we use the estimated time varying parameters obtained under Gaussian distribution and produce the density using a Student-t with calibrated degrees of freedom. Following Corradi and Swanson (2006) we choose $\upsilon = 5$. This second case is labelled "Miss2".

Table 5 reports the average log-scores for the above two specifications together with the benchmark Student-t specifications. Figures 8 and 9 report the empirical distribution of the PITs as in Diebold et al. (1998), and its cumulative distribution as in Rossi and Sekhposyan (2013). In both cases, we report the 95% confidence interval. Miss1 model delivers average log-scores which are comparable with the baseline Student-t specifications. However, an inspection of the PITs suggests that the densities from this model tend to be not well calibrated, slightly overstating the probability mass at the center of the density. Conversely, Miss2 model produces much better calibrated densities, but they perform rather poorly compared to the benchmark models as documented in the lower panel of Table 5. Those results suggest that both the low degree of freedom and the score-driven law of motion of the time-varying parameters, are important to achieve well calibrated density forecasts.

			Student-t							
			Trend	Trend-B	AR(1)	AR(1)-B	AR(2)	AR(2)-B	AR(4)	AR(4)-B
		ALogS	-1.5897	-1.5688	-1.6325	-1.6766	-1.6249	-1.6671	-1.5976	-1.6272
Miss1	Trend	-1.6546	0.0530	0.1148	0.5386	0.6230	0.3557	0.6763	0.1611	0.5172
	Trend-B	-1.4760	0.0493	0.0244	0.0019	0.0001	0.0116	0.0001	0.0270	0.0071
	AR(1)	-1.5713	0.6151	0.9601	0.0177	0.0043	0.1338	0.0036	0.5013	0.1975
	AR(1)-B	-1.5354	0.2296	0.4723	0.0007	0.0001	0.0322	0.0005	0.1258	0.0357
	AR(2)	-1.5248	0.0606	0.4343	0.0024	0.0016	0.0000	0.0001	0.0210	0.0209
	AR(2)-B	-1.5902	0.9907	0.6932	0.3241	0.0835	0.3895	0.0384	0.8839	0.4795
	AR(4)	-1.5266	0.1377	0.4760	0.0149	0.0045	0.0032	0.0026	0.0007	0.0144
	AR(4)-B	-1.5453	0.2534	0.6111	0.0464	0.0103	0.0749	0.0060	0.1534	0.0047
Miss2	Trend	-1.7339	0.0000	0.0054	0.0072	0.2327	0.0006	0.0505	0.0012	0.0162
	Trend-B	-1.8480	0.0001	0.0000	0.0000	0.0005	0.0005	0.0007	0.0001	0.0002
	AR(1)	-1.7260	0.0041	0.0033	0.0000	0.1052	0.0097	0.1071	0.0029	0.0430
	AR(1)-B	-1.7896	0.0001	0.0000	0.0000	0.0011	0.0007	0.0011	0.0001	0.0009
	AR(2)	-1.7171	0.0017	0.0119	0.0115	0.3333	0.0003	0.1270	0.0012	0.0556
	AR(2)-B	-1.7747	0.0001	0.0005	0.0003	0.0256	0.0002	0.0012	0.0003	0.0048
	AR(4)	-1.7354	0.0020	0.0055	0.0099	0.2026	0.0025	0.0979	0.0000	0.0048
	AR(4)-B	-1.7703	0.0000	0.0002	0.0022	0.0605	0.0016	0.0161	0.0000	0.0000

Table 5: Average log-scores (ALogS) in the first row and column. All the other entries correspond to the p-values for the difference in the ALogS (Amisano and Giacomini, 2007).

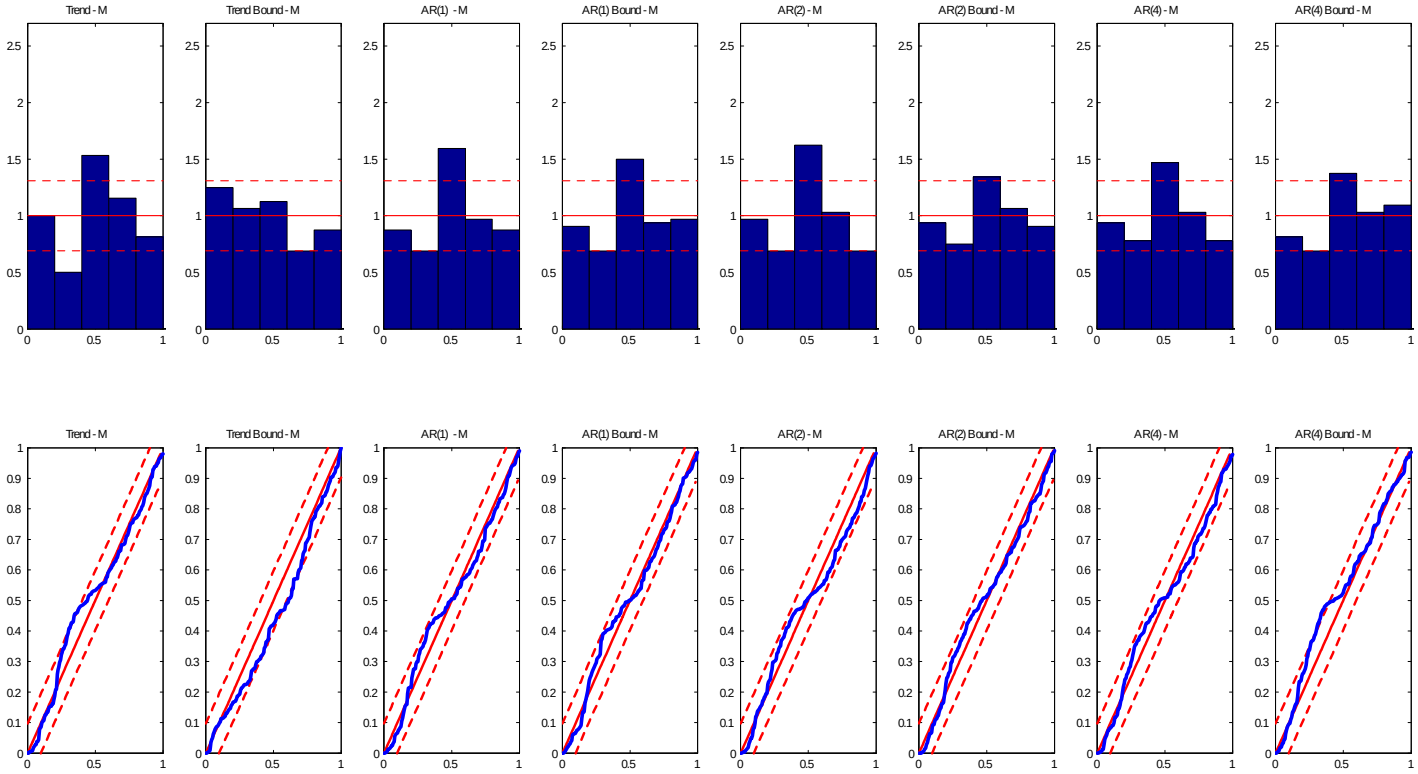

Figure 8: Density forecast-Miss1: in the upper panel, the p.d.f. of the PITs (normalized) and the 95% critical values (dashed lines) approximated by binomial distribution, constructed using a normal approximation as in Diebold et al. (1998). In the lower panel, the c.d.f. of the PITs with critical values based on Rossi and Sekhposyan (2013).

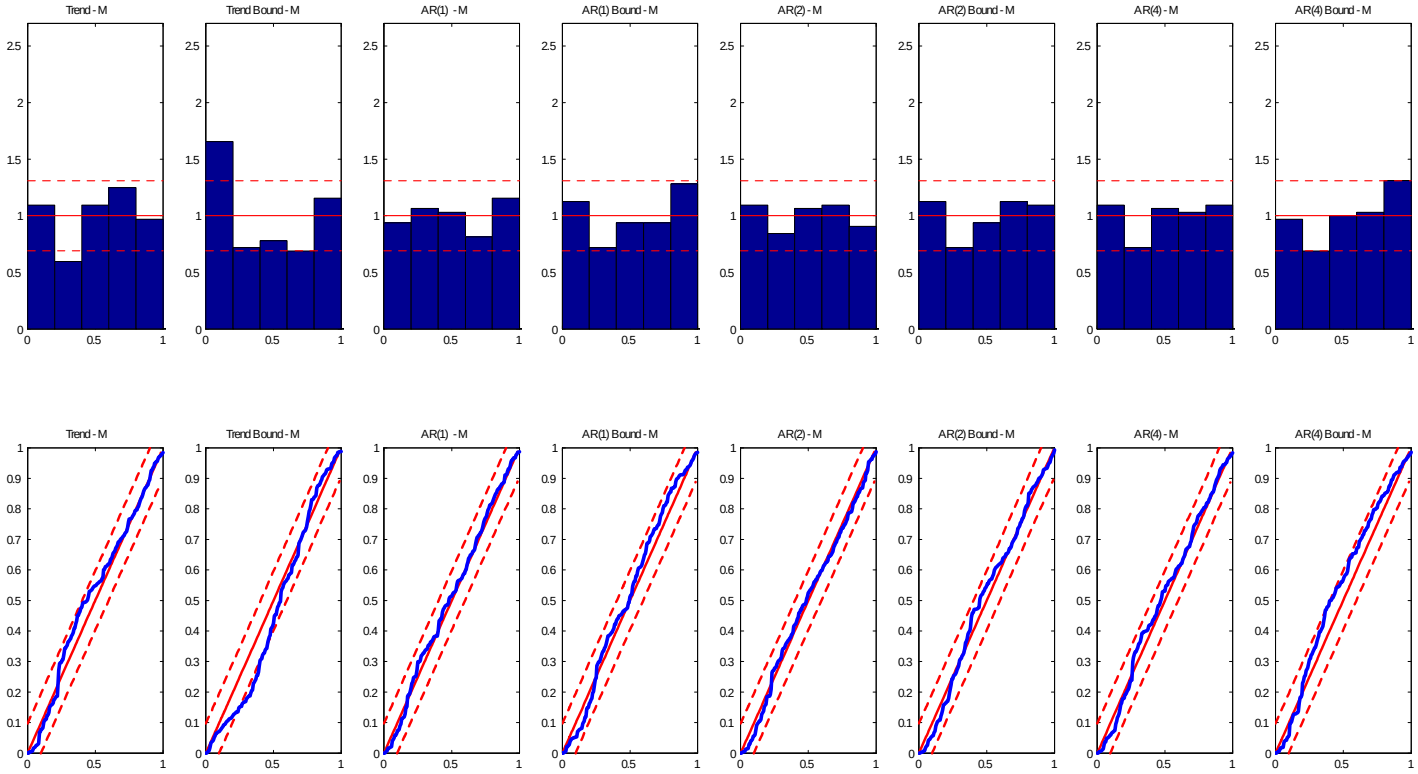

Figure 9: Density forecast-Miss2: in the upper panel, the p.d.f. of the PITs (normalized) and the 95% critical values (dashed lines) approximated by binomial distribution, constructed using a normal approximation as in Diebold et al. (1998). In the lower panel, the c.d.f. of the PITs with critical values based on Rossi and Sekhposyan (2013).