Learning and Loss Functions: Comparing Optimal and Operational Monetary Policy Rules

Eric Gaus*  Srikanth Ramamurthy†

February 2013

Abstract

Modern Bayesian tools aided by MCMC techniques allow researchers to estimate models with increasingly intricate dynamics. This paper highlights the application of these tools with an empirical assessment of optimal versus operational monetary policy rules within a standard New Keynesian macroeconomic model with adaptive learning. The question of interest is which of the two policy rules - contemporaneous data or expectations of current variables - better describes the policy undertaken by the U.S. central bank. Results for the data period 1954:III to 2007:I indicate that the data strongly favors contemporaneous expectations over real time data.

Keywords: Adaptive Learning, Rational Expectations, Bayesian Econometrics, MCMC.

*Gaus: Ursinus College, 601 East Main St., Collegeville, PA 19426-1000 (e-mail: egaus@ursinus.edu)
†Ramamurthy: Sellinger School of Business, Loyola University Maryland, 4501 N. Charles Street, Baltimore, MD 21210 (e-mail: sramamurthy@loyola.edu).
1 Introduction

There is a growing literature on the optimality of monetary policy rules. As highlighted by Bernanke (2007), the framework underlying much of this literature incorporates learning on the part of agents. Duffy and Xiao (2007), Evans and Honkapohja (2009) and Gaus (2013) use adaptive learning to analyze several monetary policy rules. Adaptive learning models suppose that agents use some form of recursive least squares (RLS) to forecast the economic variables of interest. The two most common type of gains associated with the RLS algorithm are decreasing gain and constant gain. In particular, a special case of the former scheme is the inverse of the sample length, in which case the agents weight all the observations equally. In contrast, the older observations are discounted more heavily under constant gain. Duffy and Xiao (2007) finds that a class of optimal monetary policy rules is stable - in the sense that agents will learn the rational expectations solution - under decreasing gain learning. Evans and Honkapohja (2009) operationalizes these optimal rules and shows that they may not be stable under constant gain learning. Gaus (2013) reconciles these two results by providing an in-depth theoretical analysis of the differences between optimal and operational rules. Theoretical work by Bullard and Mitra (2002) investigates the stability properties of the model under learning and determinacy of these two rules, but does not take a stance on which best describes current policy. In this paper we provide an empirical assessment of optimal versus operational monetary policy rules. Specifically, we investigate which of the two rules - contemporaneous data or expectations of current variables - better describes the policy undertaken by the U.S. central bank in the past few decades.

A key aspect in the analysis of optimal policy is the the loss function of the policy maker. Following Duffy and Xiao (2007), a well studied loss function in the literature is the expected value of the discounted value of all future deviations of inflation $\pi_t$, output gap $x_t$ and the interest rate $i_t$ from their target value:

$$E_0\sum_{t=0}^{\infty} \beta^t [\pi_t^2 + \alpha_x x_t^2 + \alpha_i i_t^2]$$ (1)

where $\beta$ is the discount rate, $\alpha_x$ and $\alpha_i$ are the relative weight policy makers place on the output gap and interest rate, respectively. The policy maker’s objective is to minimize this loss function. The remaining components of the economy are specified by the standard New Keynesian IS and Phillips curves

$$x_t = x_{t+1}^c - \varphi(i_t - \pi_{t+1}^c) + \varepsilon_{x,t}$$ (2)
$$\pi_t = \beta \pi_{t+1}^c + \lambda x_t + \varepsilon_{\pi,t}$$ (3)

where the shocks $\varepsilon_{x,t}$ and $\varepsilon_{\pi,t}$ follow exogenous stationary AR(1) processes

$$\varepsilon_{x,t} = \rho_x \varepsilon_{x,t-1} + u_t, \quad \varepsilon_{\pi,t} = \rho_{\pi} \varepsilon_{\pi,t} + v_t$$

with $u_t \sim N(0, \sigma_u^2)$ and $v_t \sim N(0, \sigma_v^2)$. Finally, $x_{t+1}^c$ is the one period ahead conditional

---

For simplicity we have set all targets to zero.
expectation of $x$ based on the information available at time $t$. It is worth emphasizing here that these expectations need not be rational. For a derivation of this New Keynesian model see Woodford (2003).

Based on this model, Duffy and Xiao derive an optimal rule, which can be characterized as a Taylor rule

$$i_t = \theta_x x_t + \theta_\pi \pi_t,$$

(4)

where $\theta_x = \varphi_\lambda \alpha_x^{-1}$ and $\theta_\pi = \varphi \alpha_x \alpha_i^{-1}$. Evans and Honkapohja (2009) assume that agents do not have access to the contemporaneous endogenous variables ($\pi_t$ and $x_t$) and therefore examine an operational version, in the sense of McCallum (1999), of (4),

$$i_t = \theta_x x_t^e + \theta_\pi \pi_t^e,$$

(5)

One might also consider using lagged values of the output gap and inflation, but in the context of adaptive learning that would be a naive form of expectations. The seemingly minor difference between (4) and (5) has a profound bearing on the dynamics of the endogenous variables as well as the parameter estimates.

In the context of this setup, we seek to answer two questions. First, do policy makers use real time data or do they only form expectations of current variables? Second, what weight do policy makers place on output and interest rates, i.e. $\alpha_x$ and $\alpha_i$? To answer these questions, we rely on Bayesian techniques aided by Markov Chain Monte Carlo (MCMC) methods. Bayesian estimation of macroeconomic learning models originated in Milani (2007, 2008) that estimate simple linearized Dynamic Stochastic General Equilibrium (DSGE) models. These papers show that the adaptive learning assumption can lead to a better fit to the data compared to rational expectations. Slobodyan and Wouters (2012) demonstrates this in a medium scale DSGE model. In a recent paper, Gaus and Ramamurthy (2012) provide a systematic and fairly general procedure for estimating constant gain learning models. Our estimation methodology here closely follows that paper and showcases the application of modern Bayesian techniques to a macroeconomic learning model. As in their paper, the entire empirical analysis here is performed using the Tailored Randomized Block Metropolis Hastings (TaRB-MH) algorithm of Chib and Ramamurthy (2010).

Results from a model comparison exercise for quarterly data from 1954:III to 2007:I indicates that the data strongly favors expectations over real time data. In conformity with the literature, we estimate the parameters $\theta_x$ and $\theta_\pi$ instead of $\alpha_x$ and $\alpha_i$, and then back out the implied values of the latter. Estimation results for the model parameters for the same data period reveals that the implied weights are smaller than the calibrated values of Woodford (2003).

The rest of the paper is organized as follows. The next section describes the estimation and model comparison techniques in detail, paying careful attention to the systematic development of the steps involved in the process. Section three presents the empirical results and section four concludes.
2 Estimation Methodology

The estimation process involves a sequence of steps that is common to a broad class of adaptive learning models. The first of these steps involves writing the model in the canonical form. From this one derives the agents’ Perceived Law of Motion (PLM) and the Actual Law of Motion (ALM). These are the critical components that distinguish this class of models from rational expectations. The second step involves the agents’ learning algorithm that we assume here to be Constant Gain Least Squares (CGLS). Associated with the type of learning algorithm are the local stability conditions of the model that ensure convergence of the learning process to the rational expectations equilibrium (REE). In the third step one constructs the empirical state space model (SSM) by linking the ALM to the set of available measurements. As will be evident from the discussion below, the likelihood function that emanates from the SSM is a rather complicated object. Consequently, we rely on a well tuned MCMC scheme to summarize the posterior distribution of the model parameters. Likewise, the procedure that we use to compute the marginal likelihood of the models is also closely tied to the MCMC algorithm. We provide a practical user’s guide to both in this section.

2.1 Model and Stability Conditions

Evans and Honkapohja (2001) provides an in-depth and comprehensive treatment of the theoretical foundations of adaptive learning. The notations used here closely follow theirs. Also, for consistency of notation, elements of a row vector are separated by commas (,) and those of a column are separated by semicolons (;). For instance, \((x_1, x_2)\) refers to a row vector, whereas \((x_1; x_2)\) refers to a column vector. This notation is also extended to matrices. Finally, for the remainder of this paper we refer to the model in equations (2)-(4) as \(M_1\) and that in equations (2)-(3) and (5) as \(M_2\).

We begin by substituting the monetary policy rules in the IS curve and then casting both models in the following canonical form.

\[
\begin{align*}
\xi_t &= M_1 \xi_{t-1}^e + M_2 \xi_{t+1}^e + P w_t + Q \varepsilon_{t,t} \\
\varepsilon_t &= F w_{t-1} + e_t
\end{align*}
\]

Here \(\xi_t = (x_t, \pi_t)\) is the vector of endogenous variables and \(w_t = (\varepsilon_{x,t}, \varepsilon_{\pi,t})\) is the vector of exogenous variables. Accordingly, \(F = \text{diag}(\rho_x, \rho_\pi)\) and \(e_t\) is normally distributed with mean zero and covariance matrix \(\text{diag}(\sigma_x^2, \sigma_\pi^2)\). The remaining matrices in this canonical form are, for model \(M_1\)

\[
M_1 = 0 \quad M_2 = \begin{pmatrix} \tau & \tau \psi(1 - \theta_\pi \beta) \\ \tau \lambda & \beta + \tau \lambda \psi(1 - \theta_x \beta) \end{pmatrix} \quad P = \begin{pmatrix} \tau & -\tau \psi \theta_\pi \\ \tau \lambda & 1 - \lambda \tau \psi \theta_\pi \end{pmatrix} \quad Q = \begin{pmatrix} -\tau \psi \\ -\tau \lambda \psi \end{pmatrix}
\]

with \(\tau = 1/(1 + \psi(\theta_\pi \lambda + \theta_x))\), and for model \(M_2\)

\[
M_1 = \begin{pmatrix} -\psi \theta_x & -\psi \theta_\pi \\ -\lambda \psi \theta_x & -\lambda \psi \theta_\pi \end{pmatrix} \quad M_2 = \begin{pmatrix} 1 & \psi \\ \lambda & \beta + \lambda \psi \end{pmatrix} \quad P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad Q = \begin{pmatrix} -\psi \\ -\lambda \psi \end{pmatrix}
\]

For future reference, we call the parameters in \(M_1\) and \(M_2\) the structural parameters.
and those associated with the shocks, \( w_t \) and \( \varepsilon_{i,t} \), the shock parameters.

As mentioned earlier, superscript \( e \) above represents the agents’ expectation of \( \xi \) given their information set \( I \) at time \( t - 1 \): \( \xi_t^e := E(\xi_t|I_{t-1}) \). In this regard, what differentiates our learning agents from their rational expectations counterpart is their information set. In the latter case, the information set at time \( t \) includes the entire history of past realization of all the variables in the model, the structure of the model as well as all the parameters of the model. What this means is that the rational agent knows the exact model (6)-(7), taking into account the fact that expectations of future outcomes feedback into the evolution of the endogenous variables of the model. In contrast, our learning agents work with limited information. Specifically, in accordance with the learning literature, the agents are limited in their knowledge of the model as well as the structural parameters. One might then ask whether the agents eventually learn the rational expectations (RE) solution.

In this context, the model that the agents’ use to form expectations is referred to as their perceived law of motion (PLM). A natural candidate for the PLM is the minimum state variable (MSV) RE solution (McCallum 1983)

\[
\xi_t = a + bw_{t-1} + Pe_t + Q\varepsilon_{i,t} \tag{8}
\]

Based on this PLM the agents estimate \( a \) and \( b \) in real time by recursive least squares. Here we focus on constant gain least squares (CGLS). Denoting \( \phi_{i,t} = (a_{i,t}', \text{vec}(b_{i,t})')' \), the estimates are updated as

\[
\begin{align*}
\phi_{i,t} &= \phi_{i,t-1} + \gamma R_{t-1}^{-1}z_{t-1}(y_{i,t} - z_{t-1}'\phi_{i,t-1}) \\
R_t &= R_{t-1} + \gamma(z_{t-1}z_{t-1}' - R_{t-1}) \tag{9,10}
\end{align*}
\]

where \( i \) refers to the \( i \)th equation in the system, \( z_t = (1, w_t')' \), \( \gamma \) is the gain parameter and \( R_{t-1}^{-1} \) is the covariance matrix. Note that, the PLM evolves in real time with \( a_{t-1} \) and \( b_{t-1} \) replacing \( a \) and \( b \) in (8). One can now calculate the agents’ expectations as \( \xi_t^e = a + bw_{t-1} \) and \( \xi_{t+1}^e = a + bFw_{t-1} \), which, upon substituting into (6), yields the ALM

\[
\xi_t = (M_1 + M_2)a + (M_1b + M_2bF + PF)w_{t-1} + Pe_t + Q\varepsilon_{i,t} \tag{11}
\]

The mapping from the PLM to the ALM, which is referred to as the T-map, is instrumental to deriving the stability of the model. Here the T-map is of the form

\[
T(a, b) = ((M_1 + M_2)a, M_1b + M_2bF + PF) \tag{12}
\]

As derived in Evans and Honkapohja (2009), the stability condition for constant gain learning is that the eigenvalues of the derivatives of the T-map

\[
\begin{align*}
DT(a) &= M_1 + M_2 \tag{13} \\
DT(b) &= I_n \otimes M_1 + F' \otimes M_2 \tag{14}
\end{align*}
\]

lie within a circle of radius \( 1/\gamma \) with origin \((1 - 1/\gamma, 0)\). This is referred to as the E-stability condition that we enforce in our empirical analysis below. It should be noted that E-stability need not ensure determinacy. Thus, the econometrician may want to
separately include the determinacy constraint. Alternatively, one might impose only the determinacy constraint and allow for local instability under learning.

Before turning to the state space formulation, we note that the REE is obtained by equating terms in (8) and (11). In our model, this results in the closed form solution

\[ a^{RE} = 0 \]  
\[ \text{vec}(b^{RE}) = (I_{n \times n} - I_n \otimes M_1 - F' \otimes M_2)^{-1}(F' \otimes I_n)\text{vec}(P) \]

where \( I_n \) is the \( n \)-dimensional identity matrix and \( \text{vec} \) denotes vectorization by column.

### 2.2 State Space Model and Likelihood Function

Our interest centers around the model parameters \( \eta = \{ \psi, \lambda, \theta_x, \theta_{\pi}, \rho_x, \rho_{\pi}, \sigma_{x}^2, \sigma_{\pi}^2, \gamma \} \). In addition, we also require initial values of the learning parameter \( \phi_0 \). This is important because the PLM and ALM, and, consequently, the parameter estimates, are particularly sensitive to \( \phi_0 \) (Murray 2008). As discussed in Milani (2007) and Slobodyan and Wouters (2012), among others, one can estimate \( \phi_0 \) either based on a training sample or treat them as additional parameters. Gaus and Ramamurthy (2012) provide a methodical account of the two approaches and conclude that both approaches yield comparable parameter estimates. Here we take the latter approach of estimating \( \phi_0 \) alongside \( \eta \). This allows for a larger sample to be available for estimating the model parameters. Further, the marginal likelihood calculation is free of the effect of the training sample.

With the goal of estimating \( \eta \) and \( \phi_0 \), we turn to the state space formulation and derivation of the likelihood function. Our sample data includes quarterly measurements on \( \xi_t \) and \( i_t \) from 1954:III to 2007:I. Let \( y_t = (\xi_t; i_t) \) and \( Y_T = (y_1, \ldots, y_T) \). We combine the ALM (11), the monetary policy rule (4) or (5) and the exogenous process (7), to construct the state space setup

\[ s_t = \mu_{t-1} + G_{t-1}s_{t-1} + Hu_t \]  
\[ y_t = Bs_t \]

where \( s_t = (\xi_t; i_t; w_t) \), \( u_t = (e_t; \varepsilon_{i,t}) \sim \mathcal{N}(0, \Omega) \), \( B = (I_3, 0) \), and the remaining matrices are, for \( M_1 \)

\[
\begin{pmatrix}
M_2a_{t-1} \\
NM_2a_{t-1} \\
0
\end{pmatrix},
G_{t-1} = \begin{pmatrix}
0 & 0 & M_2b_{t-1}F + PF \\
0 & 0 & N(M_2b_{t-1}F + PF) \\
0 & 0 & F
\end{pmatrix},
H = \begin{pmatrix}
P & Q \\
NP & NQ + 1 \\
I_2 & 0
\end{pmatrix}
\]

While \( R_0 \) also needs initialization, it does not so much affect the ensuing PLM and ALM, and, consequently, the parameter estimates, as does \( \phi_0 \). Further, estimating the elements of \( R_0 \) introduces 21 additional parameters. For parsimonious considerations, therefore, it seems sufficient in practice to initialize \( R_t \) as an arbitrary symmetric, positive-definite matrix with large diagonal elements.
and for $M_2$

$$
\mu_{t-1} = \begin{pmatrix} (M_1 + M_2)a_{t-1} \\ Na_{t-1} \\ 0 \end{pmatrix},
G_{t-1} = \begin{pmatrix} 0 & 0 & M_1b_{t-1} + M_2b_{t-1}F + PF \\ 0 & 0 & Nb_{t-1} \\ 0 & 0 & F \end{pmatrix},
H = \begin{pmatrix} P & Q \\ 0 & 1 \\ I_2 & 0 \end{pmatrix}
$$

with 0 denoting vectors or matrices of appropriate dimensions. It is worth noting that the measurement equation simply extracts $\xi_t$ from the state vector $s_t$. More importantly, the fact that the latent exogenous process $w_t$ evolves independently from the rest of the system is pivotal in the likelihood evaluation. We refer the interested reader to Gaus and Ramamurthy (2012) for further details. As also pointed out in that paper, the presence of the endogenous time varying parameters $\phi_{t-1}$ in the SSM poses a technical challenge to the likelihood evaluation. The literature has sought to circumvent this problem by calculating $\phi_{t-1}$ based on the mean of the filtered density of the state given data up to $t - 2$, thus treating $\phi_{t-1}$ as a predetermined variable conditioned on $I_{t-2}$. We take the same approach here. Subsequently, the joint density of the data $f(Y_T|\eta, \phi_0)$ can be evaluated by the standard Kalman filter. For the reader’s convenience, we summarize the steps involved in the likelihood evaluation below. These calculations are valid under the prior assumption that $s_0|Y_0 \sim N(\psi_0, \Psi_0)$. For notational convenience, henceforth we refer to the collection of parameters in $\eta$ and $\phi_0$ as $\theta$.

For a given value of $\theta$:

- Construct the matrices in the SSM, initialize $\psi_0 = (I_5 - G_0)^{-1}\mu_0$, $vec(\Psi_0) = (I_{25} - G_0 \otimes G_0)^{-1}vec(H\Omega H')$ and $R_0 = k \times I_3$, $k$ large
- For $t = 1, \ldots, T$, iterate
  1. $\psi_{t|t-1} = \mu + G\psi_{t-1}$, $\Psi_{t|t-1} = G\Psi_{t-1}G' + H\Omega H'$, $\Gamma_t = B\Psi_{t|t-1}B'$
  2. Evaluate the log of the normal density with mean $B\psi_{t|t-1}$ and covariance matrix $\Gamma_t$ at $y_t$ and store the value
  3. Update $\phi_t$ and $R_t$ as in (9)-(10)
  4. Update $\psi_t = \psi_{t|t-1} + \Psi_{t|t-1}B'\Gamma_t^{-1}(y_t - B\psi_{t|t-1})$, $\Psi_t = \Psi_{t|t-1} - \Psi_{t|t-1}B'\Gamma_t^{-1}B\Psi_{t|t-1}$
  5. Update the matrices in the SSM with $\phi_t$
- Sum up the stored values in 2 to return the log of $f(Y_T|\theta)$

From a practical standpoint, the likelihood is subject to certain parameter constraints. In our applications, we impose the stationarity restriction on $s_t$ as well as the aforementioned E-stability conditions. Additional parameter bounds are handled through the prior distribution. The E-stability restriction, in particular, can be quite problematic, causing the likelihood function to behave erratically.

### 2.3 Prior Distribution

Before discussing the prior distribution of the model parameters, we note that two of the parameters, namely $\beta$ and $\lambda$, are held fixed. $\beta$ is rarely estimated in the literature
as the discount factor that is consistent with the data usually exceeds its upper limit of 1. Additionally, $\lambda$ is pinned down by the cost adjustment parameter in the underlying structural model that is absent in the reduced form linearized model that we deal with here. In our estimation we fix both parameters to the calibrated values in Woodford (2003), which are $\beta = 0.99$ and $\lambda = 0.024$. When specifying the marginal prior distribution over the remaining parameters we wanted to ensure that (a) it handled the appropriate parameter bounds, and (b) it was well dispersed to cover a wide range of potential parameter values. These marginal prior distributions are summarized in Table 2. As can be seen in this table, the prior standard deviations are all quite large. Particularly noteworthy are the prior assumptions on the autocorrelation parameters $\rho_x$ and $\rho_\pi$, each of which is assumed to be uniformly distributed between $-1$ and $1$. Similarly, the gain parameter $\gamma$ is a priori uniformly distributed in the interval $(0,1)$. The inverse gamma distributions over the variance parameters are all synchronized with a prior mean of 0.75 and a prior standard deviation of 2.0. Turning to the output gap and inflation coefficients in the monetary policy rules, we assume that these parameters are a priori normally distributed with respective means 0.25 and 1.0, and a standard deviation of 1. Lastly, the normal prior with zero-mean and standard deviation of 5 over the learning parameters reflects the little knowledge we have about them.

2.4 Posterior and MCMC Sampling

Once the likelihood function is calculated as above, it is in principle straightforward to construct the posterior distribution $\pi(\theta|Y_T) \propto f(Y_T|\theta) \pi(\theta)$, where $\pi(\theta)$ is the joint prior distribution of the parameters. However, as can be surmised in this problem, the intractable posterior that emerges from the combination of the irregular likelihood function and the aforementioned prior distribution requires the implementation of a carefully tuned MCMC scheme. For this purpose we employ the TaRB-MH algorithm.

The fundamental idea behind this algorithm roots back to Chib and Greenberg (1994,1995) that introduced the concept of tailored multiple block sampling. The key idea is to first divide the parameters into several blocks. Then the blocks are sampled in sequence by the MH algorithm with the first two moments of the proposal density for each block matched to it’s conditional posterior. The blocks themselves remain fixed through the course of the MCMC iterations. The TaRB-MH algorithm further extends this block sampling technique by randomizing over both the number of blocks and the components of each block in every iteration. The motivation for randomized blocks stems from the fact that the map from the structural model to the reduced form is highly nonlinear making it is difficult to group parameters efficiently based on some logical a priori correlation structure between them. The other aspect of the TaRB-MH algorithm is that, to deal with the potential irregularities in the posterior surface, the moments of the proposal density for each block are found by simulated annealing - a stochastic optimization algorithm. We would like to stress that the randomized blocks comprise elements drawn simultaneously from both $\eta$ and $\phi_0$ and not separately.

To illustrate this procedure for an MCMC iteration, suppose that $\theta$ is divided into $B$ blocks as $(\theta_1, \ldots, \theta_B)$. Consider the update of the $l$th block, $l = 1, \ldots, B$ with current value $\theta_l$. Denoting the current value of the preceding blocks as $\theta_{-l} = (\theta_1, \ldots, \theta_{l-1})$ and
the following blocks as \( \theta_{+l} = (\theta_{l+1}, \ldots, \theta_B) \)

1. Calculate

\[
\theta^*_l = \arg \max_{\theta_l} f(Y_T|\theta_{-l}, \theta_l, \theta_{+l}) \pi(\theta_l)
\]

using a suitable numerical optimization procedure

2. Calculate the variance \( V^*_l \) as the negative inverse of the Hessian evaluated at \( \theta^*_l \)

3. Generate \( \theta^*_l \sim t(\theta^*_l, V^*_l, \nu_l) \) where \( \nu_l \) denotes the degrees of freedom

4. Calculate the MH acceptance probability

\[
\alpha(\theta_l, \theta^*_l | Y_T, \theta^*_{-l}, \theta_{+l}) = \min \left\{ \frac{f(Y_T|\theta_{-l}, \theta^*_l, \theta_{+l}) \pi(\theta^*_l) t(\theta_l|\theta^*_l, V^*_l, \nu_l)}{f(Y_T|\theta_{-l}, \theta_l, \theta_{+l}) \pi(\theta_l) t(\theta^*_l|\theta^*_l, V^*_l, \nu_l)}, 1 \right\}
\]

5. Accept \( \theta^*_l \) if \( u \sim U(0,1) < \alpha \), else retain \( \theta_l \) and repeat for the next block.

It is worth mentioning that, in terms of the standard metrics of evaluating the efficiency of MCMC algorithms (see, for instance, Chib (2001)), the TaRB-MH algorithm is very efficient, particularly in the context of DSGE models. (Chib and Ramamurthy 2010) provide extensive comparisons to the RW-MH algorithm with the aid of several examples. As shown in their paper, the efficiency gains from the TaRB-MH algorithm increases significantly with the size of the model and the parameter space.

2.5 Model Comparison

We conclude this section with a discussion of the methodology that we use to compare models \( \mathcal{M}_1 \) and \( \mathcal{M}_2 \). As is well known, model comparison in the Bayesian framework is facilitated through marginal likelihood and Bayes factor. To compute the marginal likelihood for each of our models, we follow the approach of Chib (1995) and Chib and Jeliazkov (2001). The approach is simple and based on what is called the basic marginal likelihood identity

\[
m_i(Y_T) = \frac{f(Y_T|\theta) \pi(\theta)}{\pi(\theta|Y_T)}
\]

(19)

where \( m_i(Y_T) \) is the marginal likelihood for model \( \mathcal{M}_i \). According to this identity, all we need is to evaluate the likelihood, the prior and the posterior ordinate at a single value of \( \theta \), preferably a high density point in the support of the posterior that we denote \( \theta^1 \). Clearly, the only unknown here is the last quantity. To estimate the posterior ordinate one begins by decomposing the posterior into marginal-conditional distributions as follows

\[
\pi(\theta^1|Y_T) = \pi(\theta^1_1|Y_T) \pi(\theta^1_2|Y_T, \theta^1_1) \ldots \pi(\theta^1_B|Y_T, \theta^1_1, \ldots, \theta^1_{B-1}),
\]

\( ^3 \)In terms of the notations here, \( \alpha(\theta_l, \theta^*_l | Y_T, \theta^*_{-l}, \theta_{+l}) \) denotes the probability of moving from the value \( \theta_l \) to \( \theta^*_l \) for block \( l \), conditional on the values of the remaining blocks. See Chib and Greenberg (1994) for this interpretation.
Here $\theta_1, \ldots, \theta_B$ denote the (random) blocks that $\theta$ is partitioned into. Denoting the $l$th component above as $\pi(\theta_l^1 | Y_T, \theta^1_1, \ldots, \theta^1_{l-1})$, each of these distributions can be estimated following Chib and Jeliazkov (2001) as

\[
\hat{\pi}(\theta_l^1 | Y_T, \theta^1_1, \ldots, \theta^1_{l-1}) = \frac{n - 1}{n} \sum_{j=1}^n \alpha(\theta_l^1 | Y_T, \theta_j^1 | \theta^1_1, \ldots, \theta^1_{l-1}, \theta^{g+1}_j) q_l(\theta_l^1 | \theta^1_1, \ldots, \theta^1_{l-1}, \theta^{g+1}_j)
\]

where $\theta_{-l}, \theta_{+l}$ and $\alpha$ were defined in section (2.4). In this expression, therefore, the numerator $\alpha$ is interpreted as the probability of moving from the value $\theta^{g+1}_j$ to $\theta_l^1$ conditional on the values $\theta^1_1, \ldots, \theta^1_{l-1}$ for the remaining parameters, whereas $q$ denotes the proposal density evaluated at $\theta_l^1$: $t(\theta_l^1 | \theta^1_1, V_l, \nu_l)$. Thus, the numerator is simply the average over the product of $\alpha$ and $q$ for $n_1$ draws from the conditional posterior $\pi(\theta_l, \theta_{+l} | Y_T, \theta^1_{-l})$. In contrast, the denominator average over the $\alpha$s simply requires draws $\theta_{-l}$ from the proposal density $t(\theta^1_l | V^*_l, \nu_l)$. This is because the draws $\theta^1_l$ can be utilized from the numerator of the $(l + 1)$th component.

It is important to note that the marginal-conditional decomposition is valid only for fixed blocks. Hence, the $B$ randomized blocks are constructed only once at the outset of the marginal likelihood calculations. They remain unchanged thereafter. However, when drawing the $n_1$ samples $\{(\theta^2_g, \theta^2_{g+1}, \ldots, \theta^2_B)\}_{g=1}^{n_1}$ we use the TaRB-MH algorithm, randomizing only over the parameters in $\theta_{+l}$. Then for each draw one calculates the numerator as just explained, noting the conditioning on $\theta^1_{-l}$ when calculating the moments of the proposal density. Subsequently, in each iteration, one simply supplements the numerator draws by an additional draw of $\theta^1_l \sim t(\theta^1_l | V^*_l, \nu_l)$ to evaluate the denominator $\alpha$. For efficient implementation of this procedure, one should evaluate the denominator term of the $(l - 1)$th distribution when computing the numerator of the $l$th distribution.

3 Results

In this section we present the results of the two monetary policy rules. In addition, we discuss the implied parameters of the monetary policy maker’s loss function. As mentioned earlier, our sample data ranges from 1954:III to 2007:I.

3.1 Parameter Estimates

Table 2 reports the prior and posterior summaries for both models. For the prior, G, N, U and IG denote the gamma, normal, uniform and inverse gamma distributions, respectively. Posterior summaries include the .025, .5 and .975 quantiles. These results are based on 10,000 iterations of the TaRB-MH algorithm beyond an initial burn-in of 1000 iterations. Based on the comments from a referee, we also extended the sampler to 25,000 iterations but that made no tangible difference to the posterior estimates. As mentioned earlier, the TaRB-MH sampler is orders of magnitude more efficient than the RW-MH algorithm. To illustrate this efficiency aspect we note that the inefficiency
for all the parameters were well under 20 in the case of model $M_1$ and under 70 for model $M_2$. It is important to understand that the multi modality of the posterior in the latter case exaggerates these numbers. Overall, these inefficiency factors indicate a sampler that made frequent large jumps across the support of the posterior, resulting in draws for which the sample autocorrelation decays to zero within a few iterations. Thus, these many iterations prove sufficient for efficient global exploration of the posterior distribution. Finally, for both models, the sampler was initialized at the prior mean. The results are however insensitive to the choice of the starting values.

The first point to note is that the posterior median values are distinct from the prior means in both models for all the parameters. More interestingly, the posterior estimates also differ across the two models. As mentioned earlier, this is not surprising given that the policy makers’ information set is distinct for the two models. Under the optimal rule, the policy makers have access to contemporaneous output gap and inflation, whereas in the operational version they work with their expectations. As shown below, this distinction also has a significant impact on the marginal likelihood calculation. Further inspection of the posterior distribution for model $M_2$ reveals a bimodal distribution. It is difficult to intuitively reason the bimodality in the operational version of the monetary policy rule. The distinction in the two versions of the policy rule is the information set of the central bank. Whereas under the optimal policy rule they are assumed to know the contemporaneous values of output gap and inflation when setting the interest rate, under the operational rule their information set is the same as the other agents in the economy. Regardless of the cause of this bimodality, this feature can be seen clearly in Figure 2 that plots the kernel smoothed histograms of the draws from the posterior distribution for the two models. The red curve in this figure represents the prior whereas the green and blue curves represent the marginal posterior distributions for models $M_1$ and $M_2$, respectively. Interestingly, the posterior ordinates at the two modes are quite close. However, one of the modes appears dominant as the sampler spends relatively more time exploring the region around it. That the data is informative beyond the prior is also easily discernible in this figure.

From a macroeconomic perspective, the parameter estimates generally fall in line with macroeconomic theory. Recall that there are no structural breaks assumed by the model, which explains the relatively low value of the policy response to inflation. The median constant gain parameters, $\gamma$, implies that agents are using approximately 30 years of data in forming their expectations. This result is similar to previous estimates. Given previous theoretical research (Evans and Honkapohja 2009, Gaus 2013) these small values are particularly encouraging, because structural parameter values result that result in potential instability will need to be offset by small constant gain values.

Finally, we turn to issue of which of these two competing models better fits the data. As reported in Table 3, the log-marginal likelihood values for the two models are -1214.10 and -1077.77. This presents substantial evidence in favor of model $M_2$. We note that the reported marginal likelihood values are based on a two-block scheme ($B = 2$). For robustness, we also computed the marginal likelihood using a three block

---

4Defined as the reciprocal of the relative numerical efficiency, the inefficiency factor measures the cost of working with a correlated sample compared to hypothetical i.i.d. draws. Smaller values indicate a lower cost and thus a more efficient sample. See Chib (2001) for further details.
scheme. The differences, however, were only in the decimal units. The reliability of the estimates are also further indicated by the small numerical standard errors.

### 3.2 Discussion of $\alpha$’s

Using the posterior distributions from above, one can obtain the implied posterior distribution of both of the loss function parameters. Table 1 presents these implied distributions, which are much smaller than the calibrated values in Woodford (2003). This is likely due to the sample period under consideration. It is well known that monetary policy was neutral toward inflation prior to 1982 and has recently been more aggressive.

It also appears that the contemporaneous expectations model favors higher values of both the values. However, note that the bimodal distribution implies two likely parameter values for each parameter. We cannot say definitively whether these two modes are a product of the structural change that is likely present in the underlying data.

<table>
<thead>
<tr>
<th>param</th>
<th>median</th>
<th>.025</th>
<th>.975</th>
<th>median</th>
<th>.025</th>
<th>.975</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_i$</td>
<td>0.0027</td>
<td>0.0012</td>
<td>0.0070</td>
<td>0.0185</td>
<td>0.0114</td>
<td>0.0510</td>
</tr>
<tr>
<td>$\alpha_x$</td>
<td>0.0016</td>
<td>-0.0016</td>
<td>0.0069</td>
<td>0.0161</td>
<td>0.0087</td>
<td>0.0197</td>
</tr>
</tbody>
</table>

Figure 1: Green: posterior in model M1, Blue: posterior in model M2.

The results in Evans and Honkapohja (2009) are partially driven by the small values of the $\alpha$’s. Because the estimates here are even smaller than the calibrated values in Woodford (2003), one might have concerns about instability. However, the eigenvalues of the T-map at the median parameter values lie within the unit circle, implying stability under constant gain. More generally, because the posterior draws are constrained to the stable region of the parameter space, we are ensured that the parameter values do not lead to an explosive model. This in turn suggests that the agents learn the rational expectations solution.

---

One could sample the $\alpha$ directly instead of the $\theta$s. We chose to sample the latter for the sake of conformity with the vast literature in this area.
4 Conclusion

In this paper we compare optimal vs. operational monetary policy rules within a prototypical New Keynesian model with learning. We rely on modern MCMC based Bayesian techniques for the empirical analysis. For replicability and to aid in the application of these techniques to a broader class of learning models, we provide a user friendly guide to the various steps involved in the estimation process. Results for the period 1954:III to 2007:I indicate that the data prefers the model with contemporaneous expectations over contemporaneous data. This result is interesting because it highlights the role of expectations based monetary policy rule examined in Evans and Honkapohja (2009).

References


Table 2: Prior-Posterior Summary of Parameters in the NK Model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Prior</th>
<th>Posterior</th>
<th>Mean</th>
<th>Median</th>
<th>0.025</th>
<th>0.975</th>
<th>Log-posterior ordinate (unnormalized)</th>
<th>Log-marginal likelihood</th>
<th>(numerical standard error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_1$</td>
<td>0.00</td>
<td>0.50</td>
<td>0.25</td>
<td>0.30</td>
<td>0.15</td>
<td>0.40</td>
<td>-1196.62</td>
<td>-1214.10</td>
<td>(0.1412)</td>
</tr>
<tr>
<td>$M_2$</td>
<td>0.00</td>
<td>0.50</td>
<td>0.25</td>
<td>0.30</td>
<td>0.15</td>
<td>0.40</td>
<td>-1196.62</td>
<td>-1214.10</td>
<td>(0.1412)</td>
</tr>
<tr>
<td>$\psi$</td>
<td>4.00</td>
<td>2.00</td>
<td>0.50</td>
<td>0.50</td>
<td>0.30</td>
<td>0.70</td>
<td>-1196.62</td>
<td>-1214.10</td>
<td>(0.1412)</td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>0.00</td>
<td>0.50</td>
<td>0.25</td>
<td>0.30</td>
<td>0.15</td>
<td>0.40</td>
<td>-1196.62</td>
<td>-1214.10</td>
<td>(0.1412)</td>
</tr>
<tr>
<td>$\rho$</td>
<td>-1.0</td>
<td>1.00</td>
<td>0.50</td>
<td>0.50</td>
<td>0.30</td>
<td>0.70</td>
<td>-1196.62</td>
<td>-1214.10</td>
<td>(0.1412)</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.75</td>
<td>2.00</td>
<td>0.50</td>
<td>0.50</td>
<td>0.30</td>
<td>0.70</td>
<td>-1196.62</td>
<td>-1214.10</td>
<td>(0.1412)</td>
</tr>
</tbody>
</table>

Note: 0.025 and 0.975 denote the quantiles. Sample data: 1954:III to 2007:I.
Figure 2: Marginal prior-posterior plots of the parameters in the NK models. Red: prior, Green: posterior in model M1, Blue: posterior in model M2.