Quantitative Economics 12 (2021), 41–76 1759-7331/20210041

The discretization filter: A simple way to estimate nonlinear state space models

Leland E. Farmer
Department of Economics, University of Virginia

Existing methods for estimating nonlinear dynamic models are either highly computationally costly or rely on local approximations which often fail adequately to capture the nonlinear features of interest. I develop a new method, the discretization filter, for approximating the likelihood of nonlinear, non-Gaussian state space models. I establish that the associated maximum likelihood estimator is strongly consistent, asymptotically normal, and asymptotically efficient. Through simulations, I show that the discretization filter is orders of magnitude faster than alternative nonlinear techniques for the same level of approximation error in low-dimensional settings and I provide practical guidelines for applied researchers. It is my hope that the method's simplicity will make the quantitative study of nonlinear models easier for and more accessible to applied researchers. I apply my approach to estimate a New Keynesian model with a zero lower bound on the nominal interest rate. After accounting for the zero lower bound, I find that the slope of the Phillips Curve is 0.076, which is less than 1/3 of typical estimates from linearized models. This suggests a strong decoupling of inflation from the output gap and larger real effects of unanticipated changes in interest rates in post Great Recession.

Keywords. Nonlinear filtering, discretization, regime switching, state space models, DSGE models, zero lower bound.

JEL classification. C11, C13, E40, E50.

1. Introduction

Economists increasingly use nonlinear methods to confront their theories with data. The switch from linear to nonlinear methods is driven in part by increased computing power, but also by a desire to understand economic phenomena that cannot easily be

Leland E. Farmer: lefarmer@virginia.edu

I am especially indebted to my thesis advisors, James D. Hamilton and Allan Timmermann, for their invaluable input, support, and meaningful discussions. I am also grateful for helpful feedback and comments from Roy Allen, Brendan Beare, Darrell Duffie, Roger E. A. Farmer, Jesús Fernández-Villaverde, Patrick Gagliardini, Émilien Gouin-Bonenfant, Lars Peter Hansen, Michael Levy, Nelson Lind, Hanno Lustig, Eric Renault, Juan Rubio-Ramirez, Andres Santos, Lawrence Schmidt, Frank Schorfheide, Neil Shephard, Kenneth J. Singleton, Alexis Toda, Harald Uhlig, Daniel Waggoner, Cynthia Wu, Fan Dora Xia, and participants at several seminars and conferences. I would also like to thank the two anonymous referees for helpful feedback and suggestions. An earlier version of the current paper appeared with the title "Markov Chain Approximation and Estimation of Nonlinear, Non-Gaussian State Space Models" (Farmer (2014)). All errors are my own.

 https://doi.org/10.3982/QE1353

captured by linear models. Examples include models which incorporate the zero lower bound on interest rates (ZLB), stochastic volatility, time-varying risk premia, Poisson jumps, credit constraints, borrowing constraints, nonconvex adjustment costs, Markov-switching dynamics, and default.

Existing methods for estimating nonlinear dynamic models are either highly computationally costly or rely on local approximations which fail adequately to capture the nonlinear features of interest. In this paper, I develop a new method, the discretization filter, for approximating the likelihood of nonlinear, non-Gaussian state space models, which is especially appropriate for models with low-dimensional state spaces.

The major difficulty that arises when studying nonlinear state space models is that the likelihood cannot be evaluated recursively in closed form as it can in linear models with the Kalman filter. The discretization filter solves this problem by constructing a discrete-valued Markov chain that approximates the dynamics of the state variables. The dynamics of the system are summarized by a transition matrix as opposed to an infinite dimensional transition kernel.

When there are finitely many states, the likelihood can once again be evaluated recursively in closed form with an algorithm analogous to the Kalman filter. This computation involves a sequence of matrix multiplications which is fast and simple to implement. The discretization filter generates an approximation to the likelihood of any nonlinear, non-Gaussian state space model that can be used to estimate the model's parameters using classical or Bayesian methods.

I apply results from the statistics literature on uniformly ergodic Markov chains to establish that the associated maximum likelihood estimator is strongly consistent, asymptotically normal, and asymptotically efficient. I demonstrate through simulations that the discretization filter is orders of magnitude faster than alternative nonlinear techniques for the same level of approximation error in low-dimensional settings, and I provide practical guidelines for applied researchers. It is my hope that the method's simplicity will make the quantitative study of nonlinear models easier for and more accessible to applied researchers. A limitation of the approach presented in this paper is that it is subject to a curse of dimensionality because it relies on tensor-product integration rules. Nonproduct integration rules such as quasi-Monte Carlo methods provide a promising avenue to extend the discretization filter to handle high-dimensional models.

To demonstrate the applicability of my technique, I use the discretization filter to estimate a New Keynesian dynamic stochastic general equilibrium (DSGE) model with a zero-lower bound on nominal interest rates.[1] In this model, the central bank operates monetary policy according to a Taylor rule but cannot set the short-term nominal interest rate below zero. The model's nonlinearity is critical for understanding the propagation of shocks and the conduct of monetary policy at the ZLB, and cannot be consistently estimated with linear methods. Relative to the linearized model, the Taylor rule coefficients on inflation and the output gap are 36% larger and 75% smaller, respectively. I find that the slope of the Phillips curve is more than 3 times smaller in the nonlinear

[1]This model is similar to the one studied in Aruoba, Cuba-Borda, and Schorfheide (2018).

model, which suggests a strong decoupling of inflation from the output gap and larger real effects of unanticipated changes in interest rates in post Great Recession data.[2]

The rest of the paper is organized as follows. Section 2 reviews related literature. Section 3 explains the discretization filter and how it can be used to construct an approximation to the likelihood. Section 4 provides practical implementation advice for applied researchers. Section 5 establishes the strong consistency, asymptotic normality, and asymptotic efficiency of the approximate maximum likelihood estimator. Section 6 provides Monte Carlo comparisons with existing methods in the case of a stochastic volatility model and a a linear measurement error model. In Section 7, I estimate a New Keynesian DSGE model with a zero lower bound on nominal interest rates. Section 8 concludes.

2. Related literature

This paper is related to the literatures on the discretization of stochastic processes, filtering algorithms for nonlinear state space models, statistical properties of maximum likelihood estimators for state space models, and the estimation of nonlinear DSGE models.

Tauchen (1986) proposed the first method for discretizing stochastic processes with an application to first-order vector autoregressive (VAR) models. Tauchen and Hussey (1991) developed an extension of this method using quadrature formulas; but both of these methods fail to accurately approximate the dynamics of persistent processes (see Kopecky and Suen (2010)). Rouwenhorst (1995) developed a method which accurately approximates highly persistent processes. However, this method is limited to univariate first-order Gaussian autoregressive (AR) models. Gospodinov and Lkhagvasuren (2014) developed a method that builds on the Rouwenhorst method to better approximate persistent Gaussian VARs by matching low order conditional moments. Most recently, Farmer and Toda (2017) developed a method for approximating general nonlinear, non-Gaussian first-order Markov processes by matching conditional moments using maximum entropy.

A special case of the filtering algorithm developed in this paper, referred to as the "point-mass filter," was first proposed by Bucy (1969) and Bucy and Senne (1971). In the point-mass filter, the transition probabilities of the approximating process are chosen to be proportional to the conditional density of the state vector evaluated on a grid of evenly spaced points. In contrast, I show that any of the discretzation methods discussed in the previous paragraph can be used to approximate the state process and often lead to more accurate parameter estimates than the point-mass filter.

Furthermore, none of the previous papers on approximate grid-based filters consider the asymptotic properties of estimators resulting from these filtering approximations. The point-mass filter has exclusively been used to estimate the hidden states of

[2]Previous papers such as Aruoba, Cuba-Borda, and Schorfheide (2018) and Christiano, Eichenbaum, and Trabandt (2015) estimate linearized versions of New Keynesian models on pre-Great Recession data, and analyze the relative contribution of shocks during the ZLB. The stark differences in parameter estimates coming from the linearized and nonlinear models have significant implications for the study of the transmission of shocks at the ZLB. The ability to quickly and efficiently estimate highly nonlinear DSGE models is of critical importance for central banks and policy makers interested in studying the effects of monetary, fiscal, or macroprudential policies.

a known dynamical system. In this paper, I show that the discretization filter can also be used to estimate the parameters of the system when they are unknown. A comprehensive summary of filtering methods for state space models, including the point-mass filter, can be found in Chen (2003).

The theoretical results and proof techniques in this paper are most directly related to the work of Douc, Moulines, and Ryden (2004) and Douc, Moulines, Olsson, and Van Handel (2011). Douc, Moulines, and Ryden (2004) established the consistency and asymptotic normality of the maximum likelihood estimator in autoregressive models with a hidden Markov regime that has a compact support. Douc et al. (2011) extended the consistency result to a setting with unbounded support. These papers build on previous work from Baum and Petrie (1966), Leroux (1992), Bickel and Ritov (1996), Bickel, Ritov, and Ryden (1998), Bakry, Milhaud, and Vandekerkhove (1997), and Jensen and Petersen (1999), which establish asymptotic properties of the maximum likelihood estimator in several simpler state space models.

Section 7 of this paper contributes to the literature on likelihood-based estimation of nonlinear DSGE models by estimating a New Keynesian DSGE model with a ZLB with the discretization filter. A closely related paper by Gust, Herbst, López-Salido, and Smith (2017) estimates a medium scale New Keynesian DSGE model with a ZLB using a particle filter. Their model is estimated using highly optimized and parallelized Fortran code on the Federal Reserve Board's High Performance Computing cluster (with over 300 cores). Despite the application of this considerable computing power, it takes 10 days to produce 50,000 draws from an MCMC algorithm. My paper and methodology are focused on producing easily implementable and generalizable MATLAB code that enables economists to estimate interesting nonlinear DSGE models on a desktop computer.

Other related papers include Van Binsbergen, Fernández-Villaverde, Koijen, and Rubio-Ramírez (2012), who estimate a DSGE model with recursive preferences and stochastic volatility and Aruoba, Bocola, and Schorfheide (2017), who estimate a DSGE model with downward nominal wage and price rigidities. Closely related methodological papers discussing the properties and applications of the particle filter to the estimation of nonlinear DSGE models are Fernández-Villaverde and Rubio-Ramírez (2005), Fernández-Villaverde, Rubio-Ramírez, and Santos (2006), and Fernández-Villaverde and Rubio-Ramírez (2007).

3. The discretization filter

In this section, I introduce the notation used in the remainder of the paper and provide a brief overview of nonlinear state space models. I then explain how the state dynamics of any nonlinear state space model can be approximated by a discrete-state Markov chain. I show how this new state space system can be used to construct an approximation to the maximum likelihood estimator for the parameters and filtering distributions of the original model.

3.1 *The setting*

In what follows, I restrict attention to the analysis of Hidden Markov Models (HMMs). A HMM is a special type of nonlinear state space model where the observables in any

given time period are a function only of the state variables in that time period. However, the results can be generalized to the case when the observation equation additionally depends on some finite number of lags of the observables. Much of the exposition and notation follows Douc, Moulines, and Ryden (2004).

Let X_t denote the vector of hidden state variables of the state space system at time t. I assume that $\{X_t\}_{t=0}^{\infty}$ is a time-homogeneous, first-order,[3] stationary Markov chain and lies in a separable, compact set $\mathcal{X}$,[4] equipped with a metrizable topology and associated Borel σ-field $\mathcal{B}(\mathcal{X})$. Let $P_\theta(x, A)$, where $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$, be the transition kernel of the Markov chain. I further assume that for all $\theta \in \Theta$ and $x \in \mathcal{X}$, each conditional probability measure $P_\theta(x, \cdot)$ has a density $q_\theta(\cdot|x)$ with respect to a common finite dominating measure μ on $\mathcal{X}$.[5]

I assume that the observable sequence $\{Y_t\}_{t=1}^{\infty}$ takes values in a set $\mathcal{Y}$ that is separable and metrizable by a complete metric. I assume that for $t \geq 1$, Y_t is conditionally independent of $\{Y_s\}_{s=1}^{t-1}$ and $\{X_s\}_{s=1}^{t-1}$ given X_t. Note that this excludes models where the observation at time t depends on its own lagged values. This is purely for expositional simplicity and all of the results can be generalized to the case where Y_t depends on some fixed, finite number of lags of itself, $\{Y_{t-1}, \ldots, Y_{t-k}\}$, although this does complicate the construction of the transition matrices. I also assume that the observations conditional on any value of the state $X_t = x$, $x \in \mathcal{X}$, have a density $g_\theta(\cdot|x)$ with respect to a σ-finite measure ν on the Borel σ-field $\mathcal{B}(\mathcal{Y})$.

Define the joint process $\{Z_t\}_{t=0}^{\infty} \equiv \{(X_t, Y_t)\}_{t=0}^{\infty}$ on $\mathcal{Z} \equiv \mathcal{X} \times \mathcal{Y}$, which has transition kernel Π_θ given by

$$\Pi_\theta(z, A) = \int_A g_\theta\big(y'|x'\big) q_\theta\big(x'|x\big)\, dx'\, dy'$$

for any $z \equiv (x, y) \in \mathcal{Z}$ and $A \in \mathcal{B}(\mathcal{Z})$.

I am interested in conducting estimation and inference on the finite dimensional parameter $\theta \in \Theta$ by maximum likelihood. Θ is assumed to be a compact subset of $\mathbb{R}^p$. Denote the true parameter as θ^*.

A HMM is characterized by the following two equations:

$$X_t|X_{t-1} \sim q_\theta(X_t|X_{t-1}), \tag{3.1}$$

$$Y_t|X_t \sim g_\theta(Y_t|X_t). \tag{3.2}$$

Equation (3.1) is the state equation, and it characterizes the distribution of the latent state next period conditional on the current state. Equation (3.2) is the observation, or measurement equation, and it characterizes the distribution of the observables conditional on the current state.

[3]Assuming that X_t is a first-order Markov chain is not restrictive, because the state space can always be redefined to include additional lags of X_t as new state variables. For example, if X_t follows an AR(2) process, one can redefine the state vector to be $(X_t, X_{t-1})'$ and recover the first-order Markov assumption.

[4]Compactness of $\mathcal{X}$ simplifies much of the notation and proofs; however, many of the results can be generalized to the noncompact case using techniques developed in Douc et al. (2011)

[5]For two measures μ and ν, μ is said to *dominate* ν if for all A, $\mu(A) = 0$ implies $\nu(A) = 0$.

Let x_t and y_t denote particular realizations of the random variables X_t and Y_t. Given a sample $\{y_t\}_{t=1}^T$, the goal is to obtain estimates of the parameter vector θ and the unobserved states $\{x_t\}_{t=1}^T$, which I will denote by $\hat{\theta}_T$ and $\{\hat{x}_{t|t}\}_{t=1}^T$, respectively.[6] In order to do this, one must obtain an expression for the likelihood of the data:

$$L_T(\theta, x_0) \equiv p_\theta\big(\boldsymbol{Y}_1^T | X_0 = x_0\big), \tag{3.3}$$

where $\boldsymbol{Y}_1^T \equiv (Y_1, \ldots, Y_T)$, and X_0 refers to the initial condition of the state. For the remainder of the paper, the notation p_θ without explicit introduction will refer to a general density where the arguments and meaning will be clear from the context. Define the corresponding log likelihood as

$$\ell_T(\theta, x_0) \equiv \log p_\theta\big(\boldsymbol{Y}_1^T | X_0 = x_0\big). \tag{3.4}$$

In the subsequent section, I show how to approximate equation (3.1) by a discrete-valued Markov chain.

3.2 *Approximating the state dynamics*

The idea of discretization to alleviate computational problems in economics is not new. One of the first instances of this is Tauchen (1986). Tauchen's approximation, along with several more recent approximations proposed in the literature,[7] have been widely used to solve asset pricing and DSGE models. What is new in this paper is the application of the idea of discretization to the *estimation* of nonlinear, non-Gaussian state space models.

I construct a discrete-valued, first-order Markov process $\{X_{t,M}\}_{t=1}^\infty$, whose dynamics mimic those of the original continuous-valued process $\{X_t\}_{t=1}^\infty$. This allows me to summarize the dynamics of the unobserved state by a finite-dimensional transition matrix $P_{\theta,M}$.[8] Note that this is fundamentally different from forecasting the next period's state by taking a local approximation around the current estimate as is done in the extended Kalman filter. My approximation method is global yet does not rely on simulation techniques.

Define a discrete set of M points in $\mathcal{X}$, $\mathcal{X}_M \equiv \{x_{m,M}\}_{m=1}^M$, associated with sets $\{A_{m,M}\}_{m=1}^M$, which partition $\mathcal{X}$, and define a transition matrix $P_{\theta,M}$ such that the mm'-th element:

$$P_{\theta,M}\big(m, m'\big) = \mathbb{P}_\theta(X_{t,M} = x_{m',M} | X_{t-1,M} = x_{m,M}) \tag{3.5}$$

corresponds to the probability of transitioning from point $x_{m,M}$ to point $x_{m',M}$ between time $t-1$ and t. The matrix $P_{\theta,M}$ is assumed to be the same for all t, and thus $X_{t,M}$ follows a first-order, time homogeneous, M-state Markov chain.

[6] The notation $\hat{x}_{t|t}$ denotes the estimate of x_t conditional only on information through time t. Sometimes smoothed estimates of the unobserved state $\hat{x}_{t|T}$, incorporating all of the data, are of interest.

[7] See, for example, Tauchen and Hussey (1991), Rouwenhorst (1995), Adda and Cooper (2003), Flodén (2008), Tanaka and Toda (2013), Gospodinov and Lkhagvasuren (2014), and Farmer and Toda (2017).

[8] This is similar to the idea proposed in Tauchen and Hussey (1991). However, there the primary focus was on computing conditional expectations; here, it is approximating the dynamics of a state space model.

Note that each row of the matrix $P_{\theta,M}$ can be interpreted as a conditional probability distribution. Specifically, row m corresponds to the distribution of $X_{t,M}$ conditional on being at point $x_{m,M}$ at time $t-1$. It is critical that these conditional distributions be good approximations to the true conditional distributions $X_t|X_{t-1} = x_{m,M}$.

Define $s_{t,M}$ to be the state of the approximate system at time t. In particular, I will say that the system is in state $s_{t,M} = m$ and let $\zeta_{t,M} = e_m$ when $X_{t,M} = x_{m,M}$, where e_m is the mth column of the $(M \times M)$ identity matrix. The system outlined above is characterized by the equations:

$$\zeta_{t,M} = P'_{\theta,M}\zeta_{t-1,M} + \tilde{v}_{t,M}, \tag{3.6}$$

$$Y_t|X_{t,M} \sim g_\theta(Y_t|X_{t,M}), \tag{3.7}$$

where $\tilde{v}_{t,M} = \zeta_{t,M} - \mathbb{E}_\theta[\zeta_{t,M}|\zeta_{t-1,M}]$ and $P'_{\theta,M}$ is the transpose of the matrix $P_{\theta,M}$. Equations (3.6) and (3.7) are the state and observation equations of the new approximate model. The sequence $\{Y_t\}$ has the same distribution, conditional on the state $X_{t,M}$, as the sequence $\{Y_t\}$ generated by the original model. However, in the approximate model, the $X_{t,M}$ have been restricted to live on a discrete grid.

3.3 *Evaluating the likelihood*

In the previous section, I showed how to approximate any HMM by replacing the state equation, equation (3.1), with a discrete-state Markov chain, equation (3.6). In this section, I apply the results of Hamilton (1989) to construct an approximation to the likelihood function of the HMM. Hamilton (1989) shows that when the state dynamics of a HMM are characterized by a discrete-state Markov chain, simple prediction and updating equations exist that are analogous to the Kalman filter in the linear case. I use the notation developed in Hamilton (1994). I review these results here and show how they can be used to develop an approximation to the maximum likelihood estimator for θ.

Let $\hat{\zeta}_{t,M|t} = \mathbb{E}_\theta[\zeta_{t,M}|\boldsymbol{Y}_1^t]$ be the econometrician's best inference about the discretized state $\zeta_{t,M}$ conditional on time t information. Intuitively, $\hat{\zeta}_{t,M|t}$ is an $(M \times 1)$ vector of probabilities where each element represents the probability of being at a particular point in the state space at time t conditional on observations up to time t. The forecast of the approximate state today given the previous period's information is given by

$$\hat{\zeta}_{t,M|t-1} = \mathbb{E}_\theta\big[\zeta_{t,M}|\boldsymbol{Y}_1^{t-1}\big] = P'_{\theta,M}\hat{\zeta}_{t-1,M|t-1}. \tag{3.8}$$

Also define

$$\eta_{t,M} = \begin{bmatrix} g_\theta(Y_t|X_t = x_{1,M}) \\ \vdots \\ g_\theta(Y_t|X_t = x_{m,M}) \end{bmatrix}. \tag{3.9}$$

The mth element of $\eta_{t,M}$ is the likelihood of having observed Y_t conditional on being in state m at time t, that is, $s_{t,M} = m$.

Note that the marginal likelihood of Y_t given $\boldsymbol{Y}_1^{t-1}$ is then simply given by

$$p_{\theta,M}\big(Y_t|\boldsymbol{Y}_1^{t-1}\big) = \mathbf{1}'(\eta_{t,M} \odot \hat{\zeta}_{t,M|t-1}), \tag{3.10}$$

Algorithm 1: Discretization Filter.

1. *Approximate the State Dynamics*: Construct a discrete grid $\{x_{m,M}\}_{m=1}^{M}$ and its associated transition matrix $P_{\theta,M}$ using Algorithm 2 in Appendix B (Farmer (2021)) or any other method appropriate for the process X_t being considered.
2. *Initialization*: Set the initial distribution of the state $\hat{\zeta}_{0,m|0} = \pi_{\theta,M}^{X}$ or any arbitrary distribution. Set $t \rightsquigarrow 1$.
3. *Prediction*: Construct the forecast of the time t state $\hat{\zeta}_{t,M|t-1} = P'_{\theta,M}\hat{\zeta}_{t,M|t-1}$.
4. *Updating 1*: Evaluate the contemporaneous likelihood of having observed data y_t conditional on each possible value of the state, $\eta_{t,M}$, using equation (3.9). Compute and save the marginal likelihood of observation y_t given by equation (3.10).
5. *Updating 2*: Compute the time t filtered estimate of the state $\hat{\zeta}_{t,M|t}$ using (3.11). If $t < T$, set $t \rightsquigarrow t+1$ and go to step 3. Otherwise go to step 6.
6. *Likelihood*: Compute the approximate likelihood of the data, $\ell_{T,M}(\theta)$, using equation (3.12).

where $\odot$ is element by element multiplication of conformable matrices and $\mathbf{1}$ is an $(M \times 1)$ vector of ones. The updated inference about the state at time t is

$$\hat{\zeta}_{t,M|t} = \frac{\eta_{t,M} \odot \hat{\zeta}_{t,M|t-1}}{\mathbf{1}'(\eta_{t,M} \odot \hat{\zeta}_{t,M|t-1})} = \frac{\eta_{t,M} \odot \hat{\zeta}_{t,M|t-1}}{p_{\theta,M}\big(Y_t|\boldsymbol{Y}_1^{t-1}\big)}. \tag{3.11}$$

By iterating these equations from period 1 to the sample size T, one can obtain estimates of the filtering distributions $\{\hat{\zeta}_{t,M|t}\}_{t=1}^{T}$ and the parameters $\hat{\theta}_{T,M}$ by maximizing the log likelihood of the discretized system

$$\ell_{T,M}(\theta) = \sum_{t=1}^{T} \log p_{\theta,M}\big(Y_t|\boldsymbol{Y}_1^{t-1}\big). \tag{3.12}$$

Alternatively, given a prior distribution for the paramter vector θ, Bayesian methods can be used to sample from its posterior distribution.

Algorithm 1 summarizes the procedure for constructing the discrete approximation to the likelihood and the filtering distributions. This can then be embedded in either a classical or Bayesian procedure for performing likelihood-based estimation.

Note that the parameter estimates $\hat{\theta}_{T,M}$ and the log-likelihood function $\ell_{T,M}(\theta)$ are indexed by the number of discrete points M in addition to the sample size T to indicate that the estimates will depend on exactly how the space is discretized. I have omitted the explicit dependence of the likelihood function on the distribution of the initial state $x_{0,M}$. As part of the results in Section 5, I will show why this initial condition is irrelevant for the asymptotic properties of $\hat{\theta}_{T,M}$.

4. Recommendations for applied researchers

In this section, I provide recommendations for how to select the grid points of the approximate finite-state Markov chain and how to construct the transition matrix for the discretization filter.

4.1 *Choosing the number of grid points*

The asymptotic theory I developed in Section 5 shows that if the Farmer and Toda (2017) method with a trapezoidal quadrature rule is used to construct the transition matrix, the discretization error of the likelihood function is of the order $TM^{-2/d}$. While this is only a rate condition, I use it to recommend a rule of thumb choice for the number of points M used to construct the discretization. Setting this ratio equal to a constant and solving for M, one gets the rule of thumb

$$M = cT^{d/2} \tag{4.1}$$

where the constant c is a nuisance parameter. For example, if the dimension d of the state space is 1, the rule says to choose a number of points proportional to the square root of the sample size. If $d = 2$, then the rule recommends choosing the number of points equal to the sample size. I investigate the effect of choosing different values of c on the accuracy of the approximation in Section 6. Figure 1 plots the rule-of-thumb choice for M for state spaces of dimensions 1–4, for sample sizes up to $T = 100$ and $c = 1$.

The asymptotic analysis implies that M should be chosen to be as large as possible. However, for sufficiently large computational problems, it may not be possible to choose a large number for M. An applied researcher faces a tradeoff between computation time and the accuracy of the approximation, which I will elaborate on in Section 6. The rule

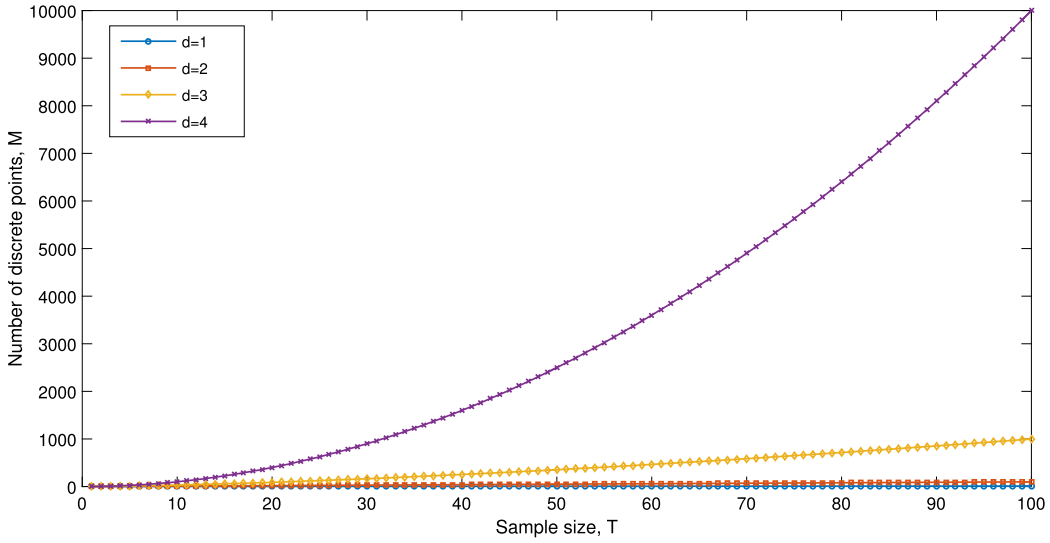

Figure 1. Rule of thumb choice for M.

of thumb proposed above can be thought of as a lower bound on the number of points to choose in order to retain validity of confidence intervals constructed for parameters using a normal approximation.

4.2 *Selecting the grid points*

When establishing my theoretical results, I assumed that the state space is compact. This is a convenient theoretical device that makes the proofs cleaner and more intuitive; but I conjecture that it is not necessary for my main results.[9] In general, practitioners specify state space models that take values in unbounded spaces. In this section, I address how to choose the support of the discretized probability measure when the state space is unbounded.

Consider the case where the number of discretization points, M, has been fixed and the goal is to choose the support of the discrete approximation, $\mathcal{X}_M$. In order for the discretized system to be a good approximation to the original model, the boundary points should be chosen to bracket the underlying state vector with high probability. This is analogous to picking boundary points from the tails of the ergodic distribution.

When the state follows a Gaussian VAR(1), a closed-form expression for the ergodic distribution is available. Gospodinov and Lkhagvasuren (2014) provided a method to discretize Gaussian VAR(1)s that is robust to high levels of persistence. They use mixtures of Rouwenhorst (1995) approximations to match conditional moments as closely as possible. For more general time series models, a closed-form expression for the ergodic distribution rarely exists.

Even when no expression for the unconditional distribution exists, it is often possible to compute the unconditional mean and standard deviation of the process. In this case, I recommend choosing a grid centered at the unconditional mean μ_x covering $\sqrt{M-1}$ unconditional standard deviations σ_x of the process on either side. That is, choose $\{x_{m,M}\}_{m=1}^{M}$ to be M evenly spaced points over the interval $[\mu_x - \sqrt{M-1}\sigma_x, \mu_x + \sqrt{M-1}\sigma_x]$.[10] In instances where discretization methods other than Rouwenhorst are used, I find that a 3 or 4 unconditional standard deviations typically suffice to provide an accurate approximation. In practice, selecting the width of the grid requires some experimentation on the part of the researcher.

If the computation of unconditional moments is infeasible, I propose simulating a path of the state and discarding a fixed fraction from the beginning as burn in. If the simulated sample and burn-in periods are sufficiently large, the remaining points can be treated as representative draws from the ergodic distribution. One can then estimate unconditional moments of the simulated process and use the method outlined above by replacing the population parameters μ_x and σ_x with their estimated counterparts. Alternatively, one can use empirical quantiles as the discretization points.

[9]The assumption of uniform ergodicity can be relaxed to geometric ergodicity, where the mixing rate of the Markov chain depends on the initial distribution. Under suitable restrictions on the initial distribution, consistency can still be established using the techniques in Douc et al. (2011).

[10]This is the way of constructing the grid employed in the Rouwenhorst (1995) and is needed to guarantee that the approximate process exactly matches first and second conditional and unconditional moments. It is also suggested for some applications in Farmer and Toda (2017).

Consider the case when $r = 1$, that is, the state vector is one-dimensional. Suppose one simulates S points from the state equation with S_{bi} used as burn in. Denote this simulated path as $\{x_s\}_{s=1}^{S}$. Then, to construct a grid that covers the state with approximately $1-\alpha$ probability, select

$$x_{m,M} = \hat{Q}_S\left(\frac{\alpha}{2} + \frac{m-1}{M}(1-\alpha)\right) \quad \text{for } m = 1, \ldots, M,$$

where $\hat{Q}_S : (0,1) \to \mathbb{R}$ is the empirical quantile function of the sample $\{x_s\}_{s=S_{bi}}^{S}$, defined as

$$\hat{Q}_S(p) = \left\{ \inf x \in \mathbb{R} : p \leq \frac{1}{S - S_{bi}} \sum_{s=S_{bi}}^{S} \mathbb{1}\{x_s \leq x\} \right\}.$$

Selecting the points in this way has the desirable property that roughly the same number of realizations of the state will fall between each pair of points.[11] By choosing α arbitrarily close to 1, it is possible to ensure that one has covered the ergodic set with any desired degree of confidence.[12] This method is also robust to skewness and fat tails in the stationary distribution.

While the simulation procedure outlined above is capable of handling very general models, it will introduce simulation error and increase the computational burden of the estimation. It is desirable to use prior knowledge of the particular model to help inform the choice of discretization whenever possible.[13]

4.3 *Constructing the transition matrix*

I recommend two ways of constructing the transition matrix for the discretization filter that are applicable to the widest range of economic models. However, there is no unique way to construct the transition matrix.[14]

First, I outline a way to extend the original method proposed by Tauchen (1986) to the nonlinear, non-Gaussian case. Create a partition of the state space $\{A_m\}_{m=1}^{M}$, where each A_m is associated with discretization point $x_{m,M}$ for all $m = 1, \ldots, M$ (this is equivalent to intervals in the one-dimensional case). Then define

$$P_{\theta,M}(m, m') = \int_{A_{m'}} q_\theta(x | X_{t-1} = x_{m,M}) \mu(dx). \tag{4.2}$$

[11] There is no unique way to define quantile functions in the multivariate case. However, one simple way to achieve the same goal is to take the univariate empirical quantiles covering $1 - \frac{\alpha}{d}$ probability for each dimension.

[12] Of course, a smaller α will require a larger number of data points for the same level of confidence in the approximation.

[13] Another possibility is to construct an ε-distinguishable set as proposed by Maliar and Maliar (2015), although this is subject to the same criticisms about introducing simulation.

[14] In addition to these two approaches, several others have been proposed in the literature: Tauchen and Hussey (1991), Rouwenhorst (1995), Adda and Cooper (2003), Flodén (2008), and Gospodinov and Lkhagvasuren (2014). However, all of these with the exception of Tauchen and Hussey (1991) only apply to linear autoregressive processes.

Intuitively, there are two layers of approximation in this expression. First, I am assuming that if X_{t-1} is in region A_m it is close to the point $x_{m,M}$ in the sense that the conditional distribution $q_\theta(X_t|X_{t-1})$ can be well approximated by $q_\theta(X_t|X_{t-1} = x_{m,M})$. Second, I am assuming that the probability of transitioning to region $A_{m'}$ from point $x_{m,M}$ is similar to the conditional density $q_\theta(X_t = x_{m',M}|X_{t-1} = x_{m,M})$ over the set $A_{m'}$.

A limitation of the application of the Tauchen (1986) approach is the difficulty it encounters in evaluating the integrals needed to construct the transition matrix. In general, the Tauchen (1986) method will only work well in practice when the A_m are hyperrectangles, and the transition density is easy to evaluate. Furthermore, there are no known results on the rate of weak convergence of the ergodic distribution of the approximate Markov chain to that of the underlying continuous process.

Second, I construct the transition matrix as in Farmer and Toda (2017). Their method finds the discrete distribution which is "closest" to a coarse approximation of the original continuous distribution in terms of Kullback–Leibler distance. This discrete distribution is chosen to match a set of conditional moments of the underlying continuous distribution.[15]

My Monte Carlo results in Section 6 demonstrate that when the primary aim is estimation of the parameters, very coarse discretizations are adequate. This is in line with my theoretical results which show that the estimates are consistent independently of the rate at which M grows. The discretization filter has the potential to scale to higher dimensional problems by exploiting sparse grid quadrature methods (e.g., Smolyak grids), quasi-Monte Carlo methods, or the more recently proposed ε-distinguishable set method in Maliar and Maliar (2015). I am pursuing this extension in ongoing research.

A final point related to computation is that constructing the transition matrix is easily parallelizable. Constructing each row of the transition matrix involves constructing a discrete approximation of a continuous distribution. This is independent across rows, and thus can be easily parallelized to reduce the computational burden for large M. Constructing the transition matrix is by far the most computationally expensive step of the discretization filter. While I do not employ parallelization in either my Monte Carlo examples or empirical estimation, parallelization will likely provide significant speed improvements over the fully iterative algorithm in larger scale applications.

5. Asymptotic properties of the maximum likelihood estimator

In this section, I establish strong consistency, asymptotic normality, and asymptotic efficiency of my proposed estimator. I consider joint asymptotics in both the sample size T and the number of discrete points M. I show that the accuracy of my approximation

[15] A special case of the discretization filter, known as the point mass filter, has been discussed at length in the computer science literature. The elements of the transition matrix are chosen to be proportional to the one-step-ahead density evaluated at the discretization points, that is, $P_{\theta,M}(m, m') \sim p(x_{m',M}|x_{m,M})$. However, since the primary aim in the computer science literature is to filter the states, the grid is chosen to be very fine. Tensor grid product approximations quickly become intractable in higher dimensions, and for this reason the point-mass filter is infrequently used. See Chen (2003) for a comprehensive survey of the properties and applications of filtering techniques.

is governed to first order by the proximity of the infinite history of filtering distributions of the approximate and true chains $X_{t,M}|\boldsymbol{Y}^t_{-\infty}$ and $X_t|\boldsymbol{Y}^t_{-\infty}$. The distance between these distributions is proportional to $h^*(M)$, where $h^*(M)$ is related to the approximation error between the approximate and true one-step-ahead conditional distributions of X_t. In what follows, double limits involving M and T should be interpreted as $T \to \infty$ and $M \to \infty$ as a function of T, $M(T)$. Strong consistency simply requires that $T \to \infty$ and $M \to \infty$. Asymptotic normality and asymptotic efficiency further require that $T \times h^*(M) \to 0$ as $M \to \infty$ and $T \to \infty$, that is, that $M \to \infty$ "fast enough."

A theoretical contribution of my paper is to establish a rate of convergence of the ergodic distribution of the approximate discrete chain to the true ergodic distribution. This result represents a new contribution to the literature on discrete approximations of Markov chains with continuous valued states. All intermediate results and proofs can be found in Appendix E in the replication file (Farmer (2021)).

5.1 *Preliminaries and assumptions*

Define the notation $\overline{\mathbb{P}}_\theta$, $\overline{\mathbb{E}}_\theta$, and $\overline{p}_\theta$ to denote probabilities, expectations, and densities evaluated under the assumption that the initial state X_0 is drawn from its ergodic distribution π^X_θ, or analogously $X_{0,M}$ from $\pi^X_{\theta,M}$ in the discrete case.

Before continuing, it is useful to define the extension of the transition kernel $P_{\theta,M}$ to $\mathcal{X}$. For $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$, let

$$P_{\theta,M}(x, A) \equiv \sum_{m=1}^{M} \sum_{m'=1}^{M} P_{\theta,M}\big(m, m'\big)\mathbb{1}\{x \in A_{m,M}\}\mathbb{1}\{x_{m',M} \in A\}.$$

Similarly, define the extension of the ergodic measure $\pi^X_{\theta,M}$ to $\mathcal{X}$. For $A \in \mathcal{B}(\mathcal{X})$, let

$$\pi^X_{\theta,M}(A) \equiv \sum_{m=1}^{M} \pi^X_{\theta,M}(m)\mathbb{1}\{x_{m,M} \in A\}.$$

Lastly, I define the limit as $M \to \infty$ of these objects in the natural way:

$$P_{\theta,\infty}(x, A) \equiv \lim_{M\to\infty} \sum_{m=1}^{M} \sum_{m'=1}^{M} P_{\theta,M}\big(m, m'\big)\mathbb{1}\{x \in A_{m,M}\}\mathbb{1}\{x_{m',M} \in A\}$$

and

$$\pi^X_{\theta,\infty}(A) \equiv \lim_{M\to\infty} \sum_{m=1}^{M} \pi^X_{\theta,M}(m)\mathbb{1}\{x_{m,M} \in A\}.$$

I will impose assumptions such that these limiting objects are well-defined. For the remainder of the section, I will use both the versions of $P_{\theta,M}$ and $\pi^X_{\theta,M}$, defined over $\mathcal{X}$ and $\mathcal{X}_M$, interchangeably and the meaning will be clear from the context.

I now list and discuss my basic assumptions. Assumptions that overlap with Douc, Moulines, and Ryden (2004) are labeled with an A, and assumptions that are new to

this paper are labeled with a B. Assumptions labeled A and B are paired by number, for example, (A1) and (B1).

(A1) (a) $0 < \sigma_- \equiv \inf_{\theta\in\Theta}\inf_{x,x'\in\mathcal{X}} q_\theta(x'|x)$ and $\sigma_+ \equiv \sup_{\theta\in\Theta}\sup_{x,x'\in\mathcal{X}} q_\theta(x'|x) < \infty$.

(b) For all $y' \in \mathcal{Y}$, $0 < \inf_{\theta\in\Theta}\int_{\mathcal{X}} g_\theta(y'|x)\,dx$ and $\sup_{\theta\in\Theta}\int_{\mathcal{X}} g_\theta(y'|x)\,dx < \infty$.

(B1) $Q_+^- \equiv \inf_{\theta\in\Theta}\inf_{M\in\mathbb{Z}^+}\inf_{m,m',m'',m'''} \frac{P_{\theta,M}(m,m')}{P_{\theta,M}(m'',m''')} > 0$.

Assumption (A1)(a) implies that there is a positive probability that the state variable can move from any part of the state space to any other part of the state space. This means that the state space $\mathcal{X}$ of the Markov chain $\{X_t\}$ is what is known as 1-small or petite. This further implies that for all $\theta \in \Theta$, $\{X_t\}$ has a unique invariant measure π_θ^X and is uniformly ergodic (see Meyn and Tweedie (1993) for a proof).

Assumption (B1) guarantees that the discrete process $\{X_{t,M}\}$ has a unique invariant distribution $\pi_{\theta,M}^X$ and is uniformly ergodic for every value $M < \infty$. Additionally it is needed so that the bound on the mixing rate of $X_{t,M}$ is independent of M and θ. This will be satisfied for any stochastic process satisfying (A1)(a) that is approximated using the methods reviewed in Section 4.3.[16] Note that while all elements of the transition matrix $P_{\theta,M}$ converge to 0 individually as $M \to \infty$, the limits of the ratios of these elements are still well-defined.

(A2) For all $\theta \in \Theta$, the transition kernel Π_θ is positive Harris recurrent and aperiodic with invariant distribution π_θ.

This assumption guarantees that the original joint Markov process $\{Z_t\}$ is itself uniformly ergodic. Assumption (A2) implies that for any initial measure λ,

$$\lim_{t\to\infty} \left\| \lambda \Pi_\theta^{(t)} - \pi_\theta \right\|_{\mathrm{TV}} = 0, \tag{5.1}$$

where $\|\cdot\|_{\mathrm{TV}}$ is the total variation norm, defined for any two probability measures μ_1 and μ_2 as

$$\|\mu_1 - \mu_2\|_{\mathrm{TV}} = \sup_A \left|\mu_1(A) - \mu_2(A)\right|$$

and $\Pi_\theta^{(t)}$ is the tth iterate of the transition kernel Π_θ. In words, for any initial measure of the joint process $\{Z_t\}$, the probability of being in any measurable set $A \in \mathcal{B}(\mathcal{Z})$ approaches the ergodic probability of being in that set uniformly over all measurable sets A as $t \to \infty$. This convergence is also independent of the initial measure λ. Developing a bound on this rate of convergence will be critical for the coming developments.

Lastly, assume that:

(A3) $b_+ \equiv \sup_{\theta\in\Theta}\sup_{y_1,x} g_\theta(y_1|x) < \infty$ and $\overline{\mathbb{E}}_{\theta^*}(|\log b_-(y_1)|) < \infty$, where $b_-(y_1) \equiv \inf_{\theta\in\Theta}\int_{\mathcal{X}} g_\theta(y_1|x)\mu(dx)$.

[16] An instance of where this would be violated is in the case of a stochastic process with bounded shocks. For example, an AR(1) process with bounded shocks cannot travel to any point in the state space from any other point in the state space. Therefore, some elements of the transition matrix $P_{\theta,M}$ would be exactly 0 for finite M, violating assumption (B1).

(B3) $\overline{\mathbb{E}}_{\theta^*}(|\log c_-(y_1)|) < \infty$, where $c_-(y_1) \equiv \inf_{\theta\in\Theta}\inf_{M\in\mathbb{Z}^+}\inf_{1\le m\le M}\sum_{m'=1}^{M} P_{\theta,M}(m, m')g_\theta(y_1|x_{m',M})$.

Assumptions (A3) and (B3) are additional boundedness conditions involving the observation density g_θ which will be necessary to establish the existence of certain limits. Additional assumptions are introduced and explained as needed.

5.2 *Consistency*

The proof of consistency can be broken down into two main parts. The first is to show that the approximation to the likelihood function implied by the discretization filter, properly normalized, converges to a well-defined asymptotic criterion function $\ell_M(\theta)$, for fixed M, as the sample size $T \to \infty$. It is important that this convergence be uniform with respect to the parameter $\theta \in \Theta$, the initial condition $x_0 \in \mathcal{X}_M$, and the number of discrete points $M \in \mathbb{Z}^+$. This step relies largely on the analysis in Douc, Moulines, and Ryden (2004), with the additional requirement that the conditions be strengthened so that the convergence is uniform with respect to the number of discrete points M used to construct the approximation. This will be a consequence of the uniform ergodicity of the filtering distributions $\{X_{t,M}|\boldsymbol{Y}_1^t\}_{M=1}^{\infty}$, which follows from the uniform ergodicity of the discrete Markov chains $\{X_{t,M}\}_{M=1}^{\infty}$. The proof of this first part, which relies on results for a fixed M, is detailed in Lemmas 1 through 3.

The second part, which is new to this paper, is to show that this approximate limiting criterion function $\ell_M(\theta)$, which is defined for any M, converges to the true limiting criterion function $\ell(\theta)$ as the number of points used in the approximation $M \to \infty$. I will show that this holds for any discretization method whose one-step-ahead conditional distributions $X_{t,M}|X_{t-1,M} = x$ converge in distribution to the one-step-ahead conditional distributions of the original continuous process $X_t|X_{t-1} = x$ as $M \to \infty$. The proof of this second part, which makes arguments for $M \to \infty$, is detailed in Propositions 1 and 2 and Lemmas 4 and 5. These results and their proofs can be found in Appendix A and Appendix D, respectively, available in the Online Supplementary Material and in the replication file (Farmer (2021)).

Together, these two pieces will imply that $T^{-1}\ell_{T,M}(\theta)$ converges uniformly to $\ell(\theta)$ as $T, M \to \infty$. Under some additional regularity conditions, this will imply that the estimator $\hat{\theta}_{T,M}$ converges to the true parameter θ^* almost surely as $T, M \to \infty$, given in Theorem 5.1.

THEOREM 5.1. *Assume* (A1)–(A5), (B1)–(B4), *and* (BT). *Then, for any sequence of initial points* $x_{0,M} \in \mathcal{X}_M$, $\hat{\theta}_{T,M,x_{0,M}} \to \theta^*$, $\overline{\mathbb{P}}_{\theta^*}$*-a.s. as* $T \to \infty$ *and* $M \to \infty$.

This is a powerful result. It states that the maximum likelihood estimator is not only consistent but strongly consistent. In addition, the estimator is strongly consistent *independently* of the rate at which the number of points M grows.

5.3 *Asymptotic normality*

Next, I turn to the asymptotic distribution of the maximum likelihood estimator. In order to establish asymptotic normality, I will need additional assumptions regarding the smoothness and boundedness of first and second derivatives of the likelihood function.

Let ∇_θ and ∇_θ^2 be the gradient and the Hessian operator with respect to the parameter θ, respectively. Assume there exists a positive real δ such that on $G \equiv \{\theta \in \Theta : |\theta - \theta^*| < \delta\}$, the following assumptions hold:

(A6) For all $x, x' \in \mathcal{X}$ and $y \in \mathcal{Y}$, the functions $\theta \mapsto q_\theta(x, x')$ and $\theta \mapsto g_\theta(y'|x')$ are twice continuously differentiable on G.

(A7) (a) $\sup_{\theta\in G} \sup_{x,x'} \|\nabla_\theta \log q_\theta(x, x')\| < \infty$ and $\sup_{\theta\in G} \sup_{x,x'} \|\nabla_\theta^2 \log q_\theta(x, x')\| < \infty$.

(b) $\overline{\mathbb{E}}_{\theta^*}[\sup_{\theta\in G} \sup_x \|\nabla_\theta \log g_\theta(Y_1|x)\|^2] < \infty$ and $\overline{\mathbb{E}}_{\theta^*}[\sup_{\theta\in G} \sup_x \|\nabla_\theta^2 \log g_\theta(Y_1|x)\|] < \infty$.

(A8) (a) For ν-almost all $y' \in \mathcal{Y}$, there exists a function $f_{y'} : \mathcal{X} \to \mathbb{R}^+ \in L^1(\mu)$ such that $\sup_{\theta\in G} g_\theta(y'|x) \leq f_{y'}(x)$.

(b) For μ-almost all $X \in \mathcal{X}$, there exist functions $f_x^1 : \mathcal{Y} \to \mathbb{R}^+$ and $f_x^2 : \mathcal{Y} \to \mathbb{R}^+$ in $L^1(\nu)$ such that $\|\nabla_\theta g_\theta(y'|x)\| \leq f_x^1(y')$ and $\|\nabla_\theta^2 g_\theta(y'|x)\| \leq f_x^2(y')$ for all $\theta \in G$.

Instead of restablishing asymptotic normality of my proposed estimator using the techniques in Douc, Moulines, and Ryden (2004), I use Theorem 7 from their paper. I reproduce the theorem here for completeness.

THEOREM 5.2 (Theorem 7 from Douc, Moulines, and Ryden (2004)). *Assume that $\tilde{\theta}_{T,x_0}$ is an estimator satisfying $\ell_T(\tilde{\theta}_{T,x_0}, x_0) \geq \sup_{\theta\in\Theta} \ell_T(\theta, x_0) - R_T$ and assumptions* (A1)–(A8) *hold. Then the following are true*:

(i) *If $R_T = o_p(T)$ (with $P = \overline{\mathbb{P}}_{\theta^*}$), then $\tilde{\theta}_{T,x_0}$ is consistent.*

(ii) *If $R_T = O_p(1)$, then $T^{1/2}(\tilde{\theta}_{T,x_0} - \theta^*) = O_p(1)$, that is, the sequence $\{\tilde{\theta}_{T,x_0}\}$ is $T^{1/2}$-consistent under $\overline{\mathbb{P}}_{\theta^*}$.*

(iii) *If $R_T = o_p(1)$, then $T^{1/2}(\tilde{\theta}_{T,x_0} - \theta^*) \to N(0, I(\theta^*)^{-1})$, $\overline{\mathbb{P}}_{\theta^*}$-weakly as $T \to \infty$.*

I derive an explicit expression for R_T as a function of M and T and provide conditions under which my proposed estimator satisfies condition (iii) of Theorem 2, which corresponds to asymptotic normality. Note that the bounds I have derived to establish consistency are not sufficient to establish asymptotic normality of my proposed estimator. I can only establish that condition (ii) of Theorem 3 is satisfied using the deterministic bounds applied thus far. To establish conditions under which (iii) is also satisfied, I use an Azuma–Hoeffding inequality derived in Douc et al. (2011). Using this new bound, I am able to state my second main result, asymptotic normality.

THEOREM 5.3. *Assume* (A1)–(A8), (B1)–(B4), (BT), *and that* $I(\theta^*)$ *is positive definite. Then for any sequence of initial points* $x_{0,M} \in \mathcal{X}_M$,

$$\sqrt{T}\big(\hat{\theta}_{T,M,x_{0,M}} - \theta^*\big) \to N\big(0, I\big(\theta^*\big)^{-1}\big)$$

$\overline{\mathbb{P}}_{\theta^*}$*-weakly as* $T \to \infty$, $M \to \infty$, *and* $T \times h^*(M) \to 0$.

Note that this result is actually stronger than just asymptotic normality. Theorem 5.3 establishes that my proposed estimator and the infeasible maximum likelihood estimator are asymptotically equivalent. That is, my estimator asymptotically achieves the Cramér–Rao lower bound.

5.4 *Relation to Bayesian estimation*

In practice, researchers often turn to Bayesian methods for the estimation of parameters. In this section, I briefly show how the results outlined in Section 4.2 lead to correct Bayesian inference as $M \to \infty$ for any fixed sample size T.

Let $p(\cdot) : \Theta \to \mathbb{R}$ denote a continuous and bounded probability density function on Θ. Given a sample $\boldsymbol{y}_1^T$ and a prior distribution over the parameters $p(\theta)$, the Bayesian is interested in computing draws from the posterior distribution $\theta|\boldsymbol{Y}_1^T = \boldsymbol{y}_1^T$ with probability density function $p(\theta|\boldsymbol{Y}_1^T = \boldsymbol{y}_1^T)$. This can be accomplished via various procedures such as importance sampling or Markov Chain Monte Carlo (MCMC).

If the discrete approximation to the likelihood function $p_{\theta,M}(\boldsymbol{Y}_1^T)$ is used in place of the true likelihood, standard Bayesian procedures will instead produce draws from the distribution $\theta_M|\boldsymbol{Y}_1^T = \boldsymbol{y}_1^T$ with associated probability density function

$$p_M\big(\theta|\boldsymbol{Y}_1^T = \boldsymbol{y}_1^T\big) \equiv \frac{p_{\theta,M}\big(\boldsymbol{Y}_1^T\big)p(\theta)}{\displaystyle\int_\Theta p_{\theta,M}\big(\boldsymbol{Y}_1^T\big)p(\theta)\,d\theta}.$$

PROPOSITION 1. *Assume* (A1)–(A3), (B1)–(B3), *and* (BT). *For any observed sample* $\boldsymbol{y}_1^T$ *and any continuous and bounded prior* $p(\cdot) : \Theta \to \mathbb{R}$,

$$\theta_M|\boldsymbol{Y}_1^T = \boldsymbol{y}_1^T \xrightarrow{d} \theta|\boldsymbol{Y}_1^T = \boldsymbol{y}_1^T$$

as $M \to \infty$.

The proof can be found in Appendix E, available in the replication file. This result guarantees that for sufficiently large M, samples drawn from the approximate posterior distribution $\theta_M|\boldsymbol{Y}_1^T = \boldsymbol{y}_1^T$ will be representative of the true posterior distribution $\theta|\boldsymbol{Y}_1^T = \boldsymbol{y}_1^T$.

This is an asymptotic result for $M \to \infty$, in contrast to particle filters, which provide unbiased approximations to the likelihood function for any number of particles. While particle filters theoretically provide an unbiased approximation to the likelihood function, the estimators can have large variances for small numbers of particles which can lead to poor performance of algorithms like Metropolis–Hastings (see Herbst and Schorfheide (2015) Chapter 9 for a detailed discussion).

6. Monte Carlo evidence

6.1 *Summary of results*

I next consider two simulation exercises to compare the performance of the discretization filter with existing alternatives: a stochastic volatility model and a linear state space model with increasing dimension. I summarize the key takeaways in the next several paragraphs and discuss the details of the stochastic volatility and increasing dimension models in Sections 6.2 and 6.3, respectively.

For both examples that I consider, I compare the speed and accuracy of the discretization filter with a simulation-based approximation, the bootstrap particle filter. For the stochastic volatility example, I also compare it to the auxiliary particle filter and a linearization-based approximation, the extended Kalman filter.

The first example I consider is a stochastic volatility model. Unlike the GDP measurement model, there is no closed-form recursive expression for evaluating the likelihood function. Thus, I use the root mean square error of parameter estimates from the different approximation methods to compare performance.

I find that the discretization filter achieves a similar degree of accuracy to both particle filters while being significantly faster. For a sample size of 100 observations, the discretization filter is 101 times faster than the bootstrap particle filter. For a sample size of 1000 observations, the discretization filter is 466 times faster than the bootstrap particle filter. The improvements in speed are even more dramatic relative to the auxiliary particle filter.

The extended Kalman filter is significantly faster than both the discretization filter and the particle filter. However, it fails to produce consistent estimates of the variance parameter, because it incorrectly imposes the assumption that the distribution of the innovations is Gaussian.

The second example I consider is a linear-Gaussian measurement error model. The advantage of considering a linear Gaussian state space model is that the Kalman Filter can be used to compute the true likelihood. This makes comparison across approximation methods of the accuracy of the likelihood approximation straightforward.

I find that the discretization filter is faster than the particle filter regardless of the desired degree of accuracy for models up to dimension 2. The discretization filter and bootstrap particle filter provide similar performance for the model of dimension 3. The particle filter provides a better tradeoff between accuracy and speed for models of dimension 4, which is driven by the use of tensor-product integration rules for the discretization filter.

It is also important to highlight the fact that unlike the particle filter, the discretization filter is deterministic, and thus does not inherit any of the additional approximation errors associated with simulation based methods. Even for the same data and parameter values, the particle filter produces a distribution of values for the likelihood function due to simulation variability. The discretization filter eliminates the error associated with simulation variability because, for a given approximation to the transition matrix, the filtering algorithm is determinsitic.

6.2 *Stochastic volatility*

In this section, I detail the performance of different estimation procedures on a stochastic volatility model.[17] The standard discrete time stochastic volatility model, as formulated in Taylor (1982), is given by

$$2X_t = \mu(1-\rho) + \rho X_{t-1} + v_t, \quad v_t \sim \text{i.i.d. } N\big(0, \sigma^2\big), \tag{6.1}$$

$$Y_t = e^{X_t/2} w_t, \quad w_t \sim \text{i.i.d. } N(0,1). \tag{6.2}$$

Note that the measurement equation can be equivalently rewritten as

$$\log\big(Y_t^2\big) = X_t + \log\big(w_t^2\big) \tag{6.3}$$

which leads to an additively separable state equation.[18] However, this simplification only applies to the most basic versions of the stochastic volatility model. I focus on results from the parameterization $\mu = -8.940$, $\rho = 0.9890$, and $\sigma = 0.1150$, which are empirical estimates of the parameters of the stochastic volatility model on daily returns data from the DAX in Hautsch and Ou (2008). The results are not sensitive to this parameterization.

I simulate data for $T = 100$, 500, and 1000 periods, and compute the likelihood of the model nine different ways: the DF using six different choices of the rule of thumb constant c, the boostrap PF with adaptive resampling using 1000 particles (BPF), an auxiliary PF with adaptive resampling using 1000 particles (APF), and the extended Kalman filter (EKF). The APF is implemented as in Herbst and Schorfheide (2015) Chapter 8. Each specification is simulated 1000 times and estimation is performed via maximum likelihood where optimization is done using MATLAB's fmincon initialized from the true value of θ. The random seed used to construct the particle filter approximation is fixed for a given sample in order to make the optimization well behaved.[19]

Figures 2, 3, and 4 display the sampling distributions of the maximum likelihood estimators. The rows of each figure correspond to a particular model parameter and the columns correspond to a particular method of approximating the likelihood. The distributions are standardized by subtracting off the true parameter values and scaling by an estimate of the Cholesky factor of the Fisher information matrix. The Fisher information matrix is estimated by computing the negative Hessian of the log-likelihood of a sample of 100,000 observations evaluated at the true $\boldsymbol{\theta}$ using a particle filter approximation with 50,000 particles. A standard normal distribution is overlayed in black for comparison

[17] I verify the assumptions outlined in Section 5 for a version of the stochastic volatility model with bounded shocks in Appendix C.

[18] This is the specification of the observation equation that I use in the EKF estimation. This can also be thought of as a misspecified Kalman filter where the measurement error is incorrectly assumed to be Gaussian.

[19] Note that traditional gradient based optimization methods are inapplicable to the PF without this trick because the likelihood function is simulated, which makes it nondifferentiable; see Flury and Shephard (2011) for a more detailed discussion.

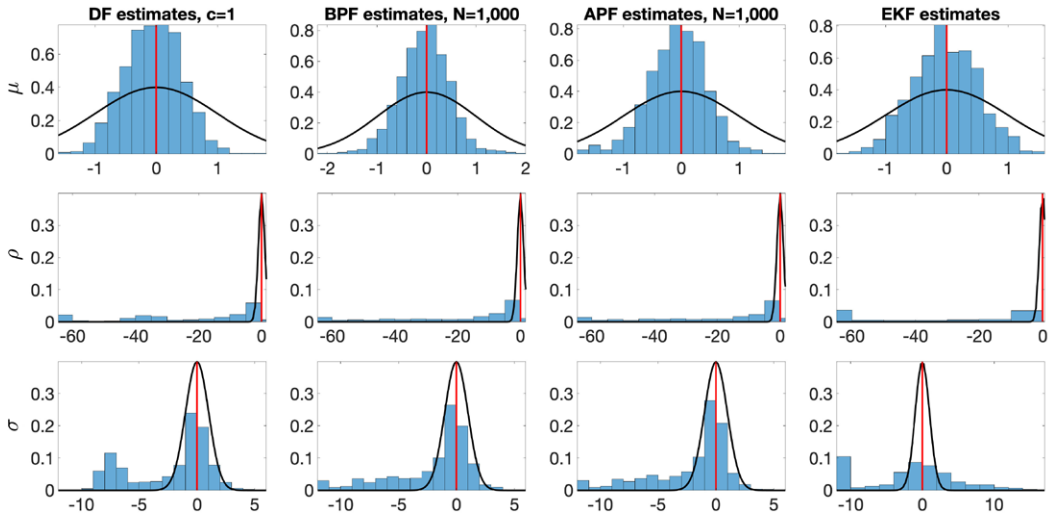

FIGURE 2. MLE sampling distributions for sample size $T = 100$.

with the asymptotic distribution. A vertical line is displayed at zero. All estimation using the discretization filter uses the Rouwenhorst (1995) discretization scheme.[20]

Note that for small sample sizes, $T = 100$, there is a considerable downward bias in the estimation of ρ and σ. That is, the optimization algorithm is picking values of ρ and σ extremely close to 0. This bias is most severe in the EKF estimates, especially for σ. However, this is not particularly surprising because the EKF is estimating a misspecified

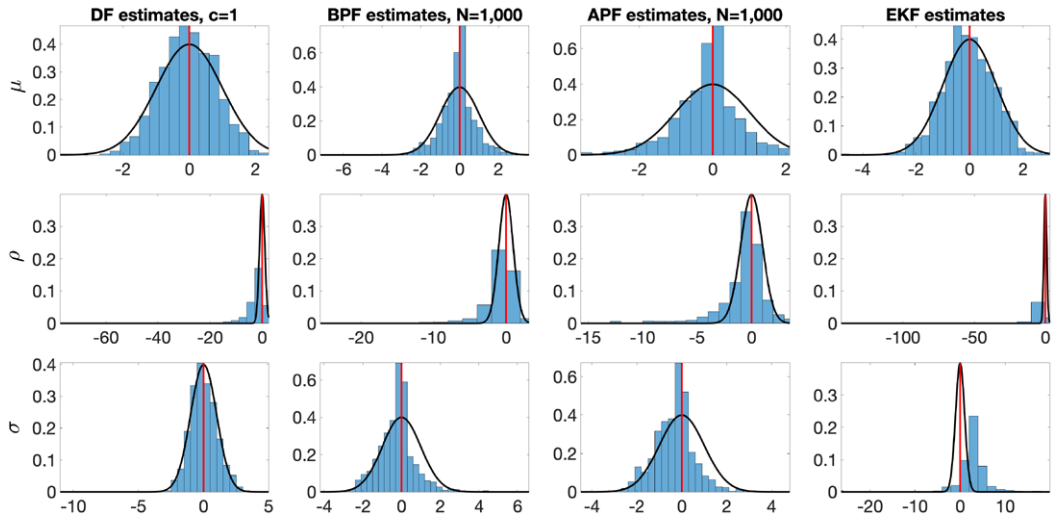

FIGURE 3. MLE sampling distributions for sample size $T = 500$.

[20]Estimation was also performed using the Farmer and Toda (2017) method, the Tauchen (1986) method, and the point-mass filter. The Rouwenhorst method performs the best although the relative gains of the discretization filter are similar across all discretization methods.

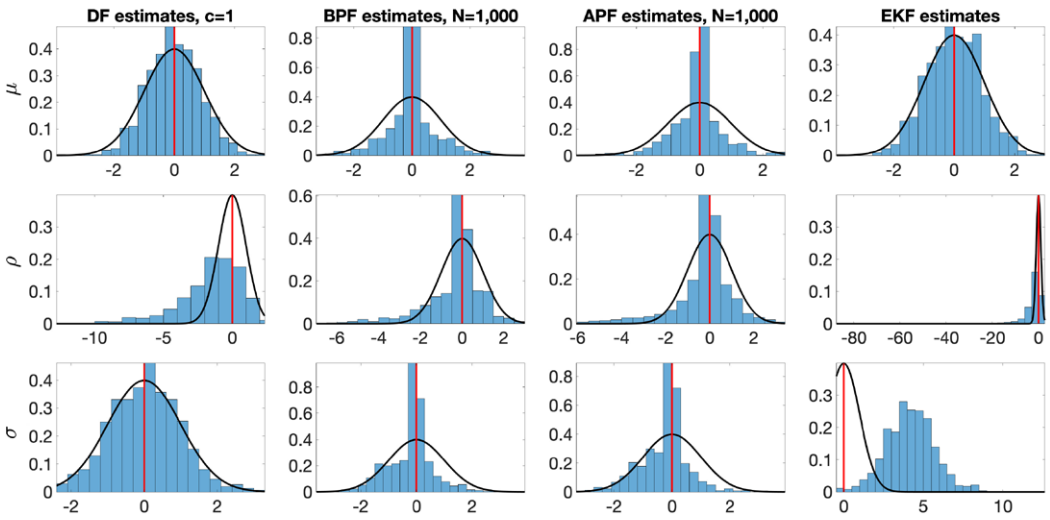

FIGURE 4. MLE sampling distributions for sample size $T = 1000$.

model, where it is treating the residual in the observation equation as a normal random variable, even though it has a $\log(\chi_1^2)$ distribution.

This bias vanishes for both the DF and the PFs in the larger sample simulations and the DF appears to produce tighter estimates of all 3 parameters, especially ρ. This is due, at least in part, to the fact that the accuracy of the Rouwenhorst approximation is independent of the persistence of the AR(1) process.

I also compute the root mean squared error (RMSE) and the bias of the parameter estimates, approximating the population expectation with an average across simulations. In particular, for the ith component of the parameter vector, I compute:

$$\text{RMSE}(\hat{\theta}_i) = \sqrt{\mathbb{E}\big[(\hat{\theta}_i - \theta_i)^2\big]}, \tag{6.4}$$

$$\text{Bias}(\hat{\theta}_i) = \mathbb{E}[\theta_i] - \theta_i \tag{6.5}$$

and report the results in Table 1.

First, consider the DF with $c = 5$ and its performance relative to the PF and the EKF. The DF and the two PFs are similar in terms of RMSE and bias for $T = 100$; however, the PFs generally slightly outperform the DF for $T = 500$ and $T = 1000$. There is not much of an appreciable difference between the BPF and APF in this application. The EKF is unambiguously the worst except for estimation of the mean parameter μ.

Next, I examine the performance of the DF for different values of the rule of thumb constant c. For $T = 100$ and to a lesser extent for $T = 500$, the RMSE and bias actually seem to increase for larger values of c. There are a couple of possible explanations for this phenomenon. The first is that the asymptotic analysis in Section 5 considers the case of a compact state space, whereas in this example as in most examples of economic interest, the state variable resides in an unbounded space. Thus, as the discretization is being constructed for larger values of c, the number of points is increasing, but so is

TABLE 1. Accuracy of parameter estimates.

		Discretization Filter						BPF	APF	EKF
ROT constant c		1/2	1	3	5	7	10	–	–	–
					Root Mean Squared Error					
	$T = 100$	0.497	0.494	0.503	0.496	0.496	0.496	0.539	0.536	0.565
μ	$T = 500$	0.395	0.396	0.389	0.389	0.389	0.389	0.421	0.408	0.441
	$T = 1000$	0.299	0.293	0.291	0.291	0.291	0.291	0.281	0.255	0.305
	$T = 100$	0.461	0.476	0.478	0.479	0.479	0.479	0.430	0.438	0.663
ρ	$T = 500$	0.063	0.070	0.070	0.070	0.070	0.070	0.032	0.024	0.203
	$T = 1000$	0.015	0.014	0.014	0.014	0.014	0.014	0.006	0.006	0.052
	$T = 100$	0.221	0.214	0.216	0.216	0.216	0.216	0.181	0.185	0.581
σ	$T = 500$	0.056	0.058	0.058	0.058	0.058	0.058	0.041	0.034	0.238
	$T = 1000$	0.028	0.028	0.027	0.027	0.027	0.027	0.019	0.020	0.127
					Bias					
	$T = 100$	−0.055	−0.056	−0.056	−0.057	−0.057	−0.057	−0.049	−0.075	−0.028
μ	$T = 500$	−0.034	−0.030	−0.030	−0.030	−0.030	−0.030	−0.045	−0.083	−0.026
	$T = 1000$	−0.006	−0.006	−0.005	−0.005	−0.006	−0.005	−0.003	−0.028	−0.005
	$T = 100$	−0.322	−0.346	−0.349	−0.351	−0.350	−0.350	−0.292	−0.300	−0.516
ρ	$T = 500$	−0.023	−0.024	−0.024	−0.024	−0.024	−0.024	−0.009	−0.008	−0.075
	$T = 1000$	−0.008	−0.008	−0.008	−0.008	−0.008	−0.008	−0.002	−0.002	−0.014
	$T = 100$	0.098	0.097	0.099	0.099	0.098	0.098	0.080	0.079	0.272
σ	$T = 500$	0.017	0.018	0.018	0.018	0.018	0.018	−0.000	−0.004	0.145
	$T = 1000$	0.007	0.007	0.007	0.007	0.007	0.007	−0.003	−0.004	0.108

the domain over which the approximation is constructed. This could potentially cause numerical issues for smaller sample sizes, because the discretization points cover large areas of the state space which are never visited in the sample.

A second possibility is that these larger numbers are actually more consistent with the RMSE and bias of the infeasible maximum likelihood estimator. In other words, the misspecification caused by small values of M is actually acting as a type of regularization which is outperforming the maximum likelihood estimator for small sample sizes. Note that this phenomenon is absent for larger sample sizes, and the estimates of the RMSE and bias appear stable across all values of c.

The first panel of Table 2 displays the average simulation times for all eight specifications. The differences in computational time relative to the PFs are stark. With $c = 1$, the EKF is about 4 times faster than the DF for small sample sizes and 13 times faster for large ones. However, this is at the cost of parameter estimates which are significantly less accurate for larger sample sizes. Furthermore, the EKF estimate of σ appears to be significantly biased, even asymptotically, due to the misspecification of the observation equation.

For estimates which are roughly the same accuracy for $T = 100$, the DF about 101 times faster than the BPF ($c = 1$). For $T = 1000$, the DF is slightly less accurate than the BPF (and APF) while still being about 466 times faster ($c = 1$). These results suggest that

TABLE 2. Computation times.

	Discretization Filter						BPF	APF
ROT constant c	1/2	1	3	5	7	10	–	–
	Computation Time of 1 Likelihood Evaluation relative to EKF							
$T = 100$	4.18	4.05	6.57	9.30	13.00	18.15	408.35	661.64
$T = 500$	6.44	8.19	22.850	44.61	80.52	119.77	2442.62	2883.08
$T = 1000$	8.67	12.58	47.71	128.33	189.18	310.03	5865.56	6180.33
	Percentage of Likelihood Evaluation Time Spent on Discretization (vs Filtering)							
$T = 100$	49.39	69.01	86.70	91.76	94.12	96.56	–	–
$T = 500$	35.10	51.08	75.17	83.51	86.47	91.66	–	–
$T = 1000$	28.74	44.36	68.05	76.28	83.92	93.29	–	–

the DF is significantly closer to the EKF than a PF in terms of computational burden, while delivering accurate parameter estimates. To give the reader a rough idea, all of the simulations for the DF and the EKF ran in a matter of minutes to hours whereas the most computationally burdensome PF specifications ($T = 1000$) took over a day to run operating in parallel on eight cores. These reductions in computation time make the estimation of many dynamic macroeconomic and financial models tractable. In the case of the Aruoba, Cuba-Borda, and Schorfheide (2018) New Keynesian model, the estimation takes a few days running MATLAB on a standard desktop computer, whereas estimation using a PF would likely take multiple weeks.

The second panel of Table 2 shows the average percentage of time (out of 100%) during one likelihood evaluation the discretization filter spends on constructing the grid and transition matrix relative to filtering the states and evaluating the likelihood. These numbers show that the primary computational burden of the discretization filter lies in constructing the grid points and especially the transition matrix, over 50% most of the time. This pattern is especially pronounced for smaller sample sizes which makes intuitive sense, but holds more weakly even for the largest sample size of $T = 1000$.

This analysis suggests that the computational burden lies primarily in the fixed cost of discretization even for large sample size estimation problems. Unlike with particle filters, the discretization filter provides fast and efficient filtering of the hidden states given a particular discretization. An interesting direction for refinement of the discretization filter is in finding ways to compute the transition matrix less frequently or perhaps only computing finitely many different transition matrices across parameter combinations at the outset of the estimation.

6.3 *Increasing the dimensionality*

In order to study how the performance of the discretization filter scales with the dimensionality of the state-space, I consider a linear state-space model where the true likelihood can be evaluated using the Kalman filter. I compare the performance, in terms of accuracy and computation time, of the discretization filter and bootstrap particle filter

relative to the Kalman filter in a d-dimensional system of AR(1) processes with measurement error:

$$X_{i,t} = \rho X_{i,t-1} + v_{i,t} \quad v_{i,t} \sim i.i.d.N\big(0, \sigma_u^2\big),$$
$$Y_{i,t} = X_{i,t} + w_{i,t}, \quad w_{i,t} \sim i.i.d.N\big(0, \sigma_o^2\big)$$

for $i = 1, \ldots, d$, where $\boldsymbol{\theta} = (\rho, \sigma_u, \sigma_o)'$ is assumed to be common across all processes.

I choose $\rho = 0.7$, $\sigma_u = 1$, and $\sigma_o = 0.1 \times \frac{\sigma_u}{\sqrt{1-\rho^2}} = 0.1 \times \frac{1}{\sqrt{1-0.49}}$ which is 10% of the unconditional standard deviation of the $X_{i,t}$ processes. I generate 1000 samples of 300 observations. For each sample, I evaluate the likelihood of the data at the true parameter vector $\boldsymbol{\theta}$ using the Kalman filter (KF), the discretization filter (DF) using a tensor-product combination of Rouwenhorst approximations, and the bootstrap particle filter (PF). I examine the following two statistics for assessing the quality of the likelihood approximation discussed in Herbst and Schorfheide (2015)

$$\hat{\Delta}_1 = \ln \hat{p}_\theta\big(\boldsymbol{Y}_1^T\big) - \ln p_\theta\big(\boldsymbol{Y}_1^T\big), \tag{6.6}$$
$$\hat{\Delta}_2 = \exp\big[\ln \hat{p}_\theta\big(\boldsymbol{Y}_1^T\big) - \ln p_\theta\big(\boldsymbol{Y}_1^T\big)\big] - 1, \tag{6.7}$$

where $\hat{p}_\theta(\boldsymbol{Y}_1^T)$ denotes the approximate likelihood computed with either the DF or the PF, and $p_\theta(\boldsymbol{Y}_1^T)$ denotes the true likelihood evaluated with the KF. Since the approximation to the likelihood provided by the PF is random, I use a 100 draws of the PF for every realization of the data. I consider several choices for the number of particles N used in the PF and for the proportionality constant used in the rule-of-thumb choice for the number of grid points M in the DF proposed in (4.1).

Table 3 presents the results of the simulation exercise for the accuracy of the likelihood approximations as measured by $\hat{\Delta}_1$ and $\hat{\Delta}_2$ and computation times. An important distinction between the PF and the DF is that the PF approximation to the likelihood is random. It depends on the particular path that is simulated for the particles. However, the DF approximation to the likelihood is deterministic and thus has no associated sampling uncertainty for a given draw of the data.

For the PF, the bias and standard deviation of the approximations for a particular realization of the data are computed as the average value and standard deviation of the likelihood discrepancies across the 100 draws of the particles respectively. Since the DF is deterministic, there is only one value of the bias per sample realization and the standard deviation is zero. The means of these statistics are then computed as the means across randomly generated samples.

To be more precise, index a draw of the data by s and a draw of the particles by g. Define $\hat{\Delta}_{i,s,g}^{\text{PF}}$ as the value of discrepancy measure $\hat{\Delta}_i$ computed by the PF for sample s and particle draw g. Similarly, define $\hat{\Delta}_{i,s}^{\text{DF}}$ as the value of discrepancy measure $\hat{\Delta}_i$ computed

TABLE 3. Increasing dimension Monte Carlo results.

Num particles N / Rule of thumb c	Bootstrap Particle Filter						Discretization Filter					
	100	500	1000	5000	10,000	50,000	1/2	1	3/2	2	3	4
						$d = 1$						
Mean Bias $\hat{\Delta}_1$	−49.74	−8.30	−3.93	−0.62	−0.22	0.085	−404.39	−89.18	−32.22	−10.57	0.48	2.63
Mean Std $\hat{\Delta}_1$	20.47	6.07	3.70	1.27	0.83	0.35	–	–	–	–	–	–
Mean Bias $\hat{\Delta}_2$	−0.95	0.12	0.24	0.23	0.23	0.23	−1	−1	−1	−0.84	18.90	21.43
Mean Time	0.05	0.07	0.10	0.35	0.65	3.65	0.00	0.00	0.00	0.01	0.01	0.01
Mean Relative Time	11.08	16.22	25.25	88.02	167.49	945.45	0.96	0.91	0.94	1.25	1.58	2.29
						$d = 2$						
Mean Bias $\hat{\Delta}_1$	−466.60	−93.11	−47.09	−9.65	−4.82	−0.78	−813.14	−178.42	−65.54	−21.65	0.71	5.23
Mean Std $\hat{\Delta}_1$	67.34	23.47	15.29	5.79	3.85	1.53	–	–	–	–	–	–
Mean Bias $\hat{\Delta}_2$	−1	−1	−0.99	0.32	0.48	0.52	−1	−1	−1	−0.99	131.28	508.35
Mean Time	0.05	0.08	0.11	0.42	0.83	4.64	0.07	0.09	0.10	0.13	0.20	0.31
Mean Relative Time	4.60	6.86	9.98	37.13	73.99	414.49	5.82	7.24	8.63	10.95	17.85	27.27
						$d = 3$						
Mean Bias $\hat{\Delta}_1$	−2034	−554.17	−309.82	−77.74	−42.28	−9.86	−1226	−275.37	−107.69	−42.80	−10.80	−4.90
Mean Std $\hat{\Delta}_1$	154.47	60.17	41.16	17.64	12.38	5.40	–	–	–	–	–	–
Mean Bias $\hat{\Delta}_2$	−1	−1	−1	−1	−0.98	0.80	−1	−1	−1	−1	−0.99	−0.97
Mean Time	0.06	0.09	0.13	0.50	1.01	5.30	0.11	0.75	1.58	2.99	7.64	17.27
Mean Relative Time	4.71	7.22	10.80	42.19	85.60	450.83	9.54	63.32	134.51	254.17	649.17	1468
						$d = 4$						
Mean Bias $\hat{\Delta}_1$	−5217	−1907	−1208	−391.38	−234.20	−67.38	−1635	−366.97	−143.19	−57.81	−14.23	−6.38
Mean Std $\hat{\Delta}_1$	276.63	121.43	87.02	41.32	30.14	14.60	–	–	–	–	–	–
Mean Bias $\hat{\Delta}_2$	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−0.99
Mean Time	0.06	0.09	0.13	0.52	1.05	5.83	1.04	10.08	36.46	103.95	472.82	1473
Mean Relative Time	1.89	3.13	4.73	18.52	37.37	207.67	30.81	344.14	1294	3687	16,582	52,535

by the DF for sample s. Then the PF statistics are computed as

$$\text{Mean Bias}\big(\hat{\Delta}_i^{\text{PF}}\big) = \frac{1}{S}\sum_{s=1}^{S}\left[\frac{1}{G}\sum_{g=1}^{G}\hat{\Delta}_{i,s,g}^{\text{PF}}\right], \tag{6.8}$$

$$\text{Mean Var}\big(\hat{\Delta}_i^{\text{PF}}\big) = \frac{1}{S}\sum_{s=1}^{S}\left[\hat{\Delta}_{i,s,g}^{\text{PF}} - \frac{1}{G}\sum_{g=1}^{G}\hat{\Delta}_{i,s,g}^{\text{PF}}\right]^2. \tag{6.9}$$

For the DF, the mean bias is given by

$$\text{Mean Bias}\big(\hat{\Delta}_i^{\text{DF}}\big) = \frac{1}{S}\sum_{s=1}^{S}\hat{\Delta}_{i,s}^{\text{DF}} \tag{6.10}$$

and Mean Var$(\hat{\Delta}_i^{\text{DF}}) = 0$ for the reason explained above.

Table 3 also reports the average absolute and relative evaluation times of the likelihood function across all specifications. The absolute times are reported in seconds. For the PF, these are computed as the average across samples and particle draws. For the DF, these are simply reported as averages across the samples. The relative times are computed as the time of one evaluation of the likelihood function relative to the time it takes for the KF.

For dimensions $d = 1$ and 2, the DF provides an improved tradeoff between computation time and accuracy for the log-likelihood discrepancy $\hat{\Delta}_1$, where the bias is smaller in absolute value. As examples, for $d = 1$ the DF with rule of thumb constant 3 provides similar accuracy to the PF with 5000 particles while being about 56 times faster; for $d = 2$ the DF with rule of thumb constant 3 provides similar accuracy to the PF with 50,000 particles while being about 23 times faster. The DF and PF deliver a similar tradeoff between speed and accuracy for $d = 3$, although the DF maintains the advantage of being a deterministic approximation. The PF becomes more computationally efficient for $d = 4$ where, for example, the DF with rule of thumb constant 2 and the PF with 50,000 particles provide a similar approximation quality with the DF being about 18 times slower. For dimensions 1 and 2 in particular, the particle filter provides a better approximation to the level of the likelihood itself, which is unsurprising given that it is an unbiased estimator of the likelihood function. This suggests that the PF may provide more accurate samples from a Bayesian posterior distribution when using MCMC methods in low dimensions. The comparison is unclear in higher dimensions as both methods provide poor approximations to the level of the likelihood function with mean biases of -1. A mean bias of -1 for $\hat{\Delta}_2$ signifies that the bias is too large to accurately quantify numerically because the exponential term is 0 to machine precision.

The main reason the DF requires so many points to maintain an accurate approximation is due to the tensor-product nature of the rule being used. The computational efficiency of the DF can be dramatically improved in applications of dimension $d = 3$ and above by employing nonproduct integration rules to construct the grid and transition matrix. In ongoing work, I develop an extension to the DF based on quasi-Monte Carlo integration rules.

7. Estimating a new Keynesian model with a ZLB

In this section, I solve and estimate a New Keynesian model with a ZLB on short-term interest rates. I closely follow the model used in Aruoba, Cuba-Borda, and Schorfheide (2018), a fairly standard New Keynesian model that uses Rotemberg adjustment costs to generate nominal rigidities. However, I differ from their approach in two significant ways.

First, I focus purely on the targeted inflation equilibrium. I do not consider multiple equilibria. Multiple equilibria are an important topic in their own right but the primary contribution of this paper is to provide a new estimation methodology. Furthermore, a conclusion of the Aruoba, Cuba-Borda, and Schorfheide (2018) paper is that there is weak to mixed evidence of a temporary switch to a deflationary regime in the U.S. data post 2008. One of their model specifications favors a full switch to the deflationary regime. However, the data on which they estimate their model are most consistent with a large discount factor shock in the targeted inflation regime.

Second, I do not use a kink in the monetary policy rule to enforce the zero lower bound. Instead, I use a smooth approximation to the kink. This is crucial because it allows me to solve the model with standard perturbation methods instead of using projection methods. Since the model needs to be solved thousands of times during the estimation, using projection methods would be impractical, even with the increased speed for the evaluation of the likelihood coming from the discretization filter.

The following subsections outline the model specification and present the data, solution method, and estimation results.

7.1 *Model description*

The model is a textbook New Keynesian model with a few small additions. For brevity, I omit most of the derivation of the model equilibrium conditions, which can be found in Aruoba, Cuba-Borda, and Schorfheide (2018) and the accompanying Appendix in the Online Supplementary Material. I outline the key difference in the monetary policy rule and present the final equilibrium conditions without derivation.

7.1.1 *Monetary policy* Instead of imposing that the nominal interest rate R_t must be 1 if the desired policy rate R_t^* falls below 1, I impose the smooth functional form:

$$R_t = g_{R_0}\big(R_t^*\big)e^{\sigma_R \varepsilon_{R,t}}, \tag{7.1}$$

where the function g is indexed by a parameter R_0 which determines the steepness of the function around 1. The function g is defined as

$$g_{R_0}(R) = \frac{(R-1)+\sqrt{(R-1)^2+4\times(R_0-1)}}{2}+1. \tag{7.2}$$

As $R_0 \to 1$ from the right, $g_{R_0}(R) \to \max\{1, R\}$. The function g asymptotes to 1 as $R_t^* \to -\infty$ and asymptotes to the 45 degree line as $R_t^* \to \infty$. However, unlike the max function, g has the advantage of being a smooth, differentiable function, so we can rely on standard perturbation methods for solving the model.

I assume that the central bank targets inflation and the output gap with a standard Taylor type rule of the form:

$$R_t^* = \left[r\pi_* \left(\frac{\pi_t}{\pi_*} \right)^{\psi_1} \left(\frac{Y_t}{Y_t^*} \right)^{\psi_2} \right]^{1-\rho_R} R_{t-1}^{\rho_R} \tag{7.3}$$

π_t is contemporaneous inflation and π_* is the inflation target. Y_t is aggregate output and Y_t^* is potential output. The parameter ρ_R governs the degree of interest rate smoothing.

7.1.2 *Equilibrium conditions* The equilibrium of the model can be characterized by five equations, along with the stochastic processes for the exogenous shocks. The control variables are consumption c_t, output y_t, and inflation π_t. The state variables can be split into endogenous states—lagged potential output y_{t-1}^* and the lagged interest rate R_{t-1}—and exogenous state variables—a discount factor shock d_t, a technology shock z_t, and a government spending (demand) shock g_t.

The model produces the following equilibrium conditions:

- Consumption Euler equation

$$c_t^{-\tau} = \beta R_t \mathbb{E}_t \left[\frac{d_{t+1}}{d_t} \frac{c_{t+1}^{-\tau}}{\gamma z_{t+1} \pi_{t+1}} \right]. \tag{7.4}$$

- Nonlinear Phillips curve

$$\begin{aligned} &\phi\beta \mathbb{E}_t \left[\frac{d_{t+1}}{d_t} c_{t+1}^{-\tau} y_{t+1} (\pi_{t+1} - \bar{\pi}) \pi_{t+1} \right] \\ &\quad = c_t^{-\tau} y_t \left\{ \frac{1}{\nu} \left(1 - \chi_H c_t^{\tau} y_t^{1/\eta} \right) + \phi (\pi_t - \bar{\pi}) \left[\left(1 - \frac{1}{2\nu} \right) \pi_t + \frac{\bar{\pi}}{2\nu} \right] - 1 \right\}. \end{aligned} \tag{7.5}$$

- Aggregate resource constraint

$$c_t = \left[1 - \frac{\phi}{2} (\pi_t - \bar{\pi})^2 \right] y_t. \tag{7.6}$$

- Monetary policy rule

$$R_t = g_{R_0} \left(\left[r\pi_* \left(\frac{\pi_t}{\pi_*} \right)^{\psi_1} \left(\frac{y_t}{y_{t-1}^*} z_t \right)^{\alpha\psi_2} \right]^{1-\rho_R} R_{t-1}^{\rho_R} \right) e^{\sigma_R \varepsilon_{R,t}}. \tag{7.7}$$

- Potential output

$$y_t^* = \alpha y_{t-1}^* + (1-\alpha) y_t - \alpha z_t. \tag{7.8}$$

I focus on targeted inflation equilibria of the model, where $\pi_t = \bar{\pi} = \pi_*$ in the steady state.

7.2 *Data*

I use the four time series considered in the 4-variable output gap specification of Aruoba, Cuba-Borda, and Schorfheide (2018) as the observables for the structural likelihood estimation. These are the log of real per-capita GDP relative to potential GDP (output gap),

the log consumption-output ratio, GDP deflator inflation, and the federal funds rate. The details of the data construction can be found in the Appendix in the Online Supplementary Material of their paper. The key methodological difference between their paper and mine is that I choose to directly compute potential GDP from the data and feed in the output gap directly. Aruoba, Cuba-Borda, and Schorfheide (2018) used the real per-capita GDP series in levels and explicitly account for the trend A_t as an additional state variable.

The data are quarterly and run from 1984Q1–2017Q4. This expands the original estimation sample of Aruoba, Cuba-Borda, and Schorfheide (2018) by 10 years, who estimate a linearized model using the subsample 1984Q1–2007Q4.

7.3 *Solution and estimation*

Since the model no longer has a kink at the ZLB, all of the equilibrium conditions are differentiable and standard perturbation methods can be used to solve for the policy functions. The model has three exogenous state variables: d_t, z_t, and g_t, and two endogenous state variables: y^*_{t-1} and R_{t-1}. Define the full state vector of the model as $x_t \equiv (y^*_{t-1}, R_{t-1}, d_t, z_t, g_t)'$.

Before proceeding, it will be useful to further reduce the set of equilibrium conditions outlined above. First, I substitute (7.6) and (7.7) into (7.4). Second, I substitute (7.6) into (7.5). This allows me to obtain expressions for the Euler equation and the Phillips curve only in terms of output, inflation, and the state variables. Output and inflation can be solved for using these two equations along with the equations which determine the evolution of the state variables.

For every value of the parameters that is considered, the policy functions for output and inflation are obtained using second-order perturbation around the targeted inflation steady state.[21] The remaining policy functions for c_t, y^*_t, and R_t are computed using the contemporaneous equilibrium conditions. This is done to ensure that the ZLB is respected in equilibrium and to reduce the computational burden of each perturbation step.

To evaluate the likelihood of the model, I use the discretization filter with five points for each exogenous state variable and seven points for both of the endogenous state variables. This results in a total of $5^3 \times 7^2 = 6125$ discretization points for the state vector.[22] The dynamics of the exogenous state variables are approximated using the Rouwenhorst method for each state variable individually. Approximating the dynamics of the endogenous state variables is more complicated and requires further explanation.

There are two key issues. First and most importantly, the values of the future endogenous state variables are predetermined by a continuous mapping, but must be forced to live on a grid to allow the use of the discretization filter. Second, the unconditional distribution of the endogenous states is unavailable in closed form, which makes choosing

[21] Only parameter groupings which produce a stable equilibrium with positive inflation are considered.

[22] This is clearly smaller than the rule-of-thumb $\lfloor T^{d/2} \rfloor = \lfloor 136^{5/2} \rfloor = 215{,}698$, and was chosen for computational feasibility reasons. Five state variables pushes the bounds of what is practical using a tensor grid approximation. In ongoing work, I explore the ability of sparse grids to greatly increase the number of state variables considered.

a range for the grid points problematic. To address these problems for a particular perturbation solution of the model, I construct a simulation-based approximation to the dynamics of the endogenous states.

I simulate 5500 draws of y_t^* and R_t using the first 500 as burn-in.[23] Using the remaining 5000 draws, I compute the unconditional mean and standard deviation of each of these two variables, for example, μ_R and $\sigma_{R,\text{unc}}$ for R_t. I set the grid for each variable to be seven equally spaced points covering $\sqrt{7-1}$ standard deviations on either side of its unconditional mean. So for R_t, this would be seven equally spaced points covering the interval $[\mu_R - \sqrt{6}\sigma_{R,\text{unc}}, \mu_R + \sqrt{6}\sigma_{R,\text{unc}}]$. All possible combinations of the elements of the two grids for the endogenous states and the three grids for the exogenous states implied by the Rouwenhorst approximation form the full set of 6125 discrete states, $(x_t^1, \ldots, x_t^{6125})'$, which will look like

$$\begin{bmatrix} x_t^1 \\ x_t^2 \\ \vdots \\ x_t^{6125} \end{bmatrix} = \begin{bmatrix} y_{t-1}^{*,1}, R_{t-1}^1, d_t^1, z_t^1, g_t^1 \\ y_{t-1}^{*,1}, R_{t-1}^1, d_t^1, z_t^1, g_t^2 \\ \vdots \\ y_{t-1}^{*,7}, R_{t-1}^7, d_t^5, z_t^5, g_t^5 \end{bmatrix}.$$

In order to compute the transition probabilities between endogenous grid points, I first compute the values of the two endogenous state variables y_t^* and R_t implied by their respective policy functions f_{y^*} and f_R for every discrete state x_t^i, $i = 1, \ldots, 6125$. No value will coincide exactly with any of the specified grid points for the endogenous states. There are three possibilities for the value of the policy functions for the endogenous states evaluated at each x_t^i:

1. The value is smaller than the smallest grid point. In this case, the probability of transitioning to the smallest grid point is 1.
2. The value is larger than the largest grid point. In this case, the probability of transitioning to the largest grid point is 1.
3. The value is in between two grid points. In this case, the probability of transitioning to the surrounding two grid points is defined to be the distance between the value and each of the two grid points, normalized so that they add to 1.

Note that the third situation results in a discretized policy function which is random. Another possibility is to map the value of the policy function to the nearest grid point. However, in practice, I find that the former provides a better approximation to the likelihood.

At this point, I have 3 (5×5) Rouwenhorst matrices for the exogenous states P_d, P_z, and P_g, and 2 (6125×7) transition matrices for the endogenous states P_{y^*} and P_R. The full transition matrix P can be computed as

$$P = \left(P_{y^*} \otimes \mathbb{1}'_{875}\right) \odot \left(\mathbb{1}'_7 \otimes P_R \otimes \mathbb{1}'_{125}\right) \odot (P_d \otimes P_z \otimes P_g \otimes I_{49}),$$

[23] Note that it is important to use the same random seed to simulate the draws across likelihood evaluations. This ensures that the likelihood function remains continuous with respect to the parameters.

where $\mathbb{1}_n$ is an $(n \times 1)$ vector of ones and I_n is the $(n \times n)$ identity matrix. The initial distribution for filtering, $\hat{\boldsymbol{\zeta}}_{0|0}$, is chosen to be a (6125×1) vector with every element equal to $\frac{1}{6125}$. For each observation, the measurement densities are evaluated at each of the 6125 discrete combinations of the states x_t^i to construct η_t.

Finally, I split the parameters of the model into two sets: those that are fixed a priori, ϑ, and those that are estimated, θ. The first set of parameters is

$$\vartheta \equiv (\bar{\pi}, \gamma, \beta, \bar{g}, \nu, \eta, \alpha, R_0)'$$

which includes the target inflation rate, the average technology growth rate, the discount factor, the average fraction of output dedicated to government spending, the elasticity of substitution between intermediate goods, the labor supply elasticity, the persistence of the output gap, and the curvature of the Taylor rule. The second set of parameters is

$$\theta \equiv (\tau, \kappa, \psi_1, \psi_2, \rho_R, \rho_z, \rho_d, \rho_g, \sigma_R, \sigma_z, \sigma_d, \sigma_g, \sigma_{\text{ME},gdp}, \sigma_{\text{ME},cy}, \sigma_{\text{ME,infl}}, \sigma_{\text{ME,fedfunds}})'$$

which includes the coefficient of relative risk aversion, the implied slope of the linearized Phillips curve, the Taylor Rule coefficients on inflation and the output gap, the degree of interest rate smoothing, the persistence and conditional standard deviation parameters of the exogenous shocks, and the measurement error standard deviations.

The model is estimated by MCMC using a random walk Metropolis–Hastings algorithm with 3 randomly generated parameter blocks at each iteration. To initialize the algorithm, I first estimate the posterior mode of the linearized model. This posterior mode is used as the starting point and the inverse Hessian computed at this posterior mode is used as the proposal variance–covariance matrix for the nonlinear MCMC algorithm. The scaling of the proposal variance–covariance matrix is calibrated to generate an acceptance rate of about 30% for the MCMC algorithm. I run the algorithm to produce 70,000 draws from the posterior distribution of θ and discard the first 10,000 as burn-in.

7.4 *Results*

I now present the results from the MCMC estimation outlined in the previous section.

Figure 5 displays the prior and estimated posterior distributions of all parameters in θ. The red lines are the prior densities and the blue histograms are the estimated posterior distributions re-scaled to be comparable to the priors. The priors are identical to those in Aruoba, Cuba-Borda, and Schorfheide (2018), except for those related to the measurement errors.

The main thing to notice is that all of the parameters appear to be well identified, with the only real concerns being the posterior distributions for the persistence of the technology shock and the standard deviation of the monetary policy shock, which exhibit some nonnormality and appear to be similar to the priors.

To facilitate the comparison with the results of Aruoba, Cuba-Borda, and Schorfheide (2018), in Table 4 I report my parameter estimates alongside theirs, and 95% credible intervals (2.5th and 97.5th percentiles of the posterior distribution). Recall that the key differences in the estimation are the sample, the specification of the monetary policy

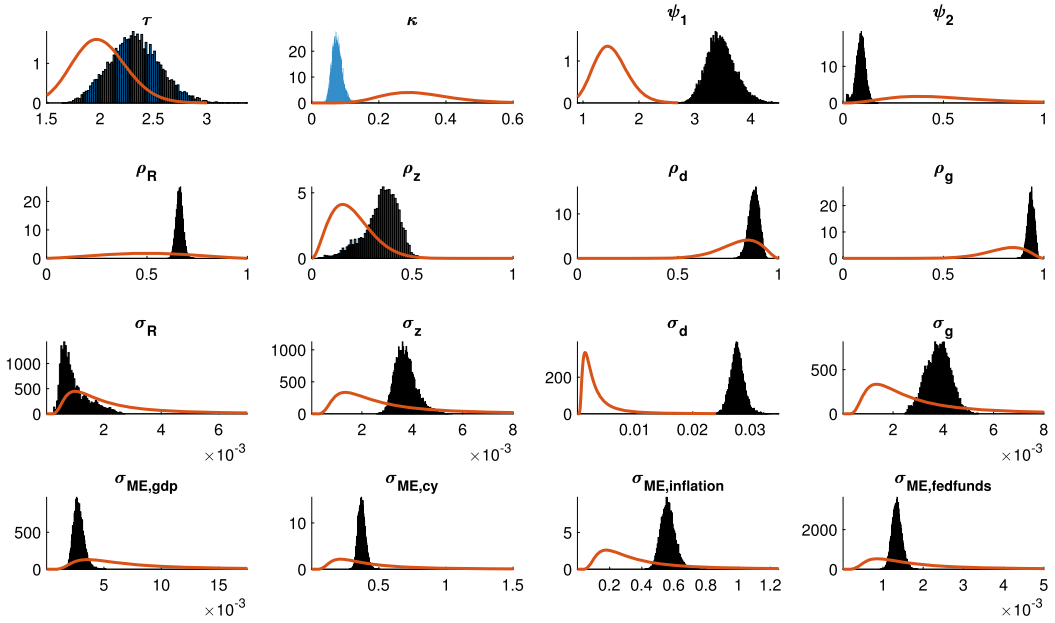

FIGURE 5. Prior and posterior distributions of model parameters.

rule, and the order of approximation. I use a second-order perturbation approximation rather than a linearized model, a smooth rather than a kinked Taylor rule, and I include ten additional years of data.

A couple of key differences are worth highlighting. First, the slope of the Phillips curve is estimated to be less than 1/3 of that in the linearized model, with a value of 0.076 compared to 0.27, which is well outside the 95% credible interval. This estimate implies that a 1% change in the output gap is associated with a 0.076% contemporaneous change in inflation rather than a 0.27% change. This corresponds to substantially higher

TABLE 4. Posterior estimates of model parameters.

Parameter	ACS Estimates (Mean)	Mean	Median	95% credible interval
τ	2.28	2.33	2.32	(1.88, 2.81)
κ	0.27	0.076	0.074	(0.049, 0.11)
ψ_1	2.55	3.47	3.45	(3.00, 4.03)
ψ_2	0.35	0.088	0.089	(0.030, 0.13)
ρ_R	0.81	0.66	0.66	(0.63, 0.70)
ρ_z	0.36	0.35	0.36	(0.15, 0.48)
ρ_d	0.94	0.88	0.88	(0.82, 0.92)
ρ_g	0.87	0.94	0.94	(0.90, 0.96)
$100\sigma_R$	0.14	0.096	0.081	(0.040, 0.22)
$100\sigma_z$	0.39	0.37	0.37	(0.31, 0.47)
$100\sigma_d$	2.47	2.77	2.77	(2.55, 3.02)
$100\sigma_g$	0.47	0.38	0.38	(0.28, 0.47)

price adjustment costs and thus larger real effects of unanticipated interest rate changes. My estimate falls in the range of DSGE model estimates surveyed in Schorfheide (2008), and is in line with the value of 0.052 found in Gust et al. (2017), who account for the ZLB.

Second, the Taylor rule parameters are significantly different. The coefficient on inflation is 36% larger, 3.47 compared to 2.55, while the coefficient on the output gap is 75% smaller, 0.088 compared to 0.35, implying that the monetary authority responds more aggressively to changes in inflation and less aggressively to changes in the output gap. Additionally, the degree of interest rate smoothing favored by the model is 19% lower, 0.66 compared to 0.81.

The remaining parameter estimates are all fairly similar to the linearized model, pre-ZLB estimates. Some mild differences are that demand shocks are estimated to be more persistent while discount factor shocks are estimated to be less persistent.

My results are also consistent with recent work by Lindé and Trabandt (2019), who show that the nonlinearity introduced by real price and wage-setting rigidities (in the form of a Kimball (1995) aggregation instead of the usual Dixit and Stiglitz (1977) aggregation) in conjunction with a ZLB on nominal interest rates is key to understanding the reduced comovement between output and inflation during the Great Recession.

8. Conclusion

Existing methods for estimating nonlinear dynamic models are either highly computationally costly or rely on local approximations which fail to capture adequately the nonlinear features of interest. In this paper, I have developed a new method, the discretization filter, for approximating the likelihood of nonlinear, non-Gaussian state space models. This approximation is simple to compute and can be used to estimate a model's parameters quickly and accurately using classical or Bayesian methods.

I showed that the maximum likelihood estimator implied by the discretization filter is strongly consistent, asymptotically normal, and asymptotically efficient. I demonstrated through simulations that the discretization filter is orders of magnitude faster than alternative nonlinear techniques for the same level of approximation error in low-dimensional settings and I provided practical guidelines for applied researchers. It is my hope that the method's simplicity will make the quantitative study of nonlinear models easier for and more accessible to applied researchers.

Lastly, I used the discretization filter to estimate a New Keynesian model with a zero lower bound on nominal interest rates. Accounting for the zero lower bound has important implications for key parameter estimates. Compared to estimates from linearized models, I find that the slope of the Phillips curve is more than 3 times smaller, the Taylor rule coefficients on inflation and the output gap are 36% larger and 75% smaller respectively, and the degree of interest rate smoothing is 19% lower. This suggests a strong decoupling of inflation from the output gap and larger real effects of unanticipated changes in interest rates in post Great Recession data.

Accurately identifying and estimating nonlinear mechanisms in dynamic economic models is critical for quantifying the effects of monetary, fiscal, and macroprudential

policies. Going forward, I hope that economists and policymakers working with nonlinear dynamic models will consider the discretization filter a valuable addition to their toolkit.

References

Adda, J. and R. W. Cooper (2003), *Dynamic Economics: Quantitative Methods and Applications*. MIT Press, Boston, MA. [46, 51]

Aruoba, S. B., L. Bocola, and F. Schorfheide (2017), "Assessing DSGE model nonlinearities." *Journal of Economic Dynamics and Control*, 83, 34–54. [44]

Aruoba, S. B., P. Cuba-Borda, and F. Schorfheide (2018), "Macroeconomic dynamics near the ZLB: A tale of two countries." *The Review of Economic Studies*, 85 (1), 87–118. [42, 43, 63, 67, 68, 69, 71]

Bakry, D., X. Milhaud, and P. Vandekerkhove (1997), "Statistique de Chaînes de Markov Cachées à Espace d'États Fini. Le Cas Non Stationnaire." *Comptes Rendus de l'Académie des Sciences—Series I—Mathematics*, 325 (2), 203–206. [44]

Baum, L. E. and T. Petrie (1966), "Statistical inference for probabilistic functions of finite state Markov chains." *Annals of Mathematical Statistics*, 37 (6), 1554–1563. [44]

Bickel, P. J. and Y. Ritov (1996), "Inference in hidden Markov models I: Local asymptotic normality in the stationary case." *Bernoulli*, 2 (3), 199–228. [44]

Bickel, P. J., Y. Ritov, and T. Ryden (1998), "Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models." *Annals of Statistics*, 26 (4), 1614–1635. [44]

Bucy, R. S. (1969), "Bayes theorem and digital realizations for non-linear filters." *Journal of the Astronautical Sciences*, 17, 80–94. [43]

Bucy, R. S. and K. D. Senne (1971), "Digital synthesis of non-linear filters." *Automatica*, 7 (3), 287–298. [43]

Chen, Z. (2003), "Bayesian filtering: From Kalman filters to particle filters, and beyond." *Statistics*, 182 (1), 1–69. [44, 52]

Christiano, L. J., M. S. Eichenbaum, and M. Trabandt (2015), "Understanding the great recession." *American Economic Journal: Macroeconomics*, 7 (1), 110–167. [43]

Dixit, A. K. and J. E. Stiglitz (1977), "Monopolistic competition and optimum product diversity." *American Economic Review*, 67 (3), 297–308. [73]

Douc, R., E. Moulines, J. Olsson, and R. Van Handel (2011), "Consistency of the maximum likelihood estimator for general hidden Markov models." *Annals of Statistics*, 39 (1), 474–513. [44, 45, 50, 56]

Douc, R., E. Moulines, and T. Ryden (2004), "Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime." *Annals of Statistics*, 32 (5), 2254–2304. [44, 45, 53, 55, 56]

Farmer, L. E., (2014), "Markov-chain approximation and estimation of nonlinear, non-Gaussian state space models." Available at SSRN: https://ssrn.com/abstract_id=2483346. [41]

Farmer, L. E. (2021), "Supplement to 'The discretization filter: A simple way to estimate nonlinear state space models'." *Quantitative Economics Supplemental Material*, 12, https://doi.org/10.3982/QE1353. [48, 53, 55]

Farmer, L. E. and A. A. Toda (2017), "Discretizing nonlinear, non-Gaussian Markov processes with exact conditional moments." *Quantitative Economics*, 8 (2), 651–683. [43, 46, 49, 50, 52, 60]

Fernández-Villaverde, J. and J. F. Rubio-Ramírez (2005), "Estimating dynamic equilibrium economies: Linear versus nonlinear likelihood." *Journal of Applied Econometrics*, 20 (7), 891–910. [44]

Fernández-Villaverde, J. and J. F. Rubio-Ramírez (2007), "Estimating macroeconomic models: A likelihood approach." *The Review of Economic Studies*, 74 (4), 1059–1087. [44]

Fernández-Villaverde, J., J. F. Rubio-Ramírez, and M. S. Santos (2006), "Convergence properties of the likelihood of computed dynamic models." *Econometrica*, 74 (1), 93–119. [44]

Flodén, M. (2008), "A note on the accuracy of Markov-chain approximations to highly persistent AR(1)-processes." *Economics Letters*, 99 (3), 516–520. [46, 51]

Flury, T. and N. Shephard (2011), "Bayesian inference based only on simulated likelihood: Particle filter analysis of dynamic economic models." *Econometric Theory*, 27 (5), 933–956. [59]

Gospodinov, N. and D. Lkhagvasuren (2014), "A moment-matching method for approximating vector autoregressive processes by finite-state Markov chains." *Journal of Applied Econometrics*, 29 (5), 843–859. [43, 46, 50, 51]

Gust, C., E. Herbst, D. López-Salido, and M. E. Smith (2017), "The empirical implications of the interest-rate lower bound." *American Economic Review*, 107 (7), 1971–2006. [44, 73]

Hamilton, J. D. (1989), "A new approach to the economic analysis of nonstationary time series and the business cycle." *Econometrica*, 57 (2), 357–384. [47]

Hamilton, J. D. (1994), *Time Series Analysis*. Princeton University Press, Princeton NJ. [47]

Hautsch, N. and Y. Ou (2008), "Stochastic volatility estimation using Markov chain simulation." In *Applied Quantitative Finance*, 249–274, Springer. Chapter 12. [59]

Herbst, E. P. and F. Schorfheide (2015), *Bayesian Estimation of DSGE Models*. Princeton University Press, Princeton NJ. [57, 59, 64]

Jensen, J. L. and N. V. Petersen (1999), "Asymptotic normality of the maximum likelihood estimator in state space models." *Annals of Statistics*, 27 (2), 514–535. [44]

Kimball, M. S. (1995), "The quantitative analytics of the basic neomonetarist model." *Journal of Money, Credit and Banking*, 27 (4), 1241–1277. [73]

Kopecky, K. A. and R. M. H. Suen (2010), "Finite state Markov-chain approximations to highly persistent processes." *Review of Economic Dynamics*, 13 (3), 701–714. [43]

Leroux, B. G. (1992), "Maximum-likelihood estimation for hidden Markov models." *Stochastic Processes and their Applications*, 40 (1), 127–143. [44]

Lindé, J. and M. Trabandt (2019), "Resolving the missing deflation puzzle." CEPR Discussion Paper No. DP13690. [73]

Maliar, L. and S. Maliar (2015), "Merging simulation and projection approaches to solve high-dimensional problems with an application to a new Keynesian model." *Quantitative Economics*, 6 (1), 1–47. [51, 52]

Meyn, S. P. and R. L. Tweedie (1993), *Markov Chains and Stochastic Stability*. Springer-Verlag, London. [54]

Rouwenhorst, K. G. (1995), "Asset pricing implications of equilibrium business cycle models." In *Frontiers of Business Cycle Research*, 294–330, Princeton University Press, Princeton, NJ. [43, 46, 50, 51, 60]

Schorfheide, F. (2008), "DSGE model-based estimation of the New Keynesian Phillips curve." *Economic Quarterly*, 94 (4), 397–433. [73]

Tanaka, K. and A. A. Toda (2013), "Discrete approximations of continuous distributions by maximum entropy." *Economics Letters*, 118 (3), 445–450. [46]

Tauchen, G. (1986), "Finite state Markov-chain approximations to univariate and vector autoregressions." *Economics Letters*, 20 (2), 177–181. [43, 46, 51, 52, 60]

Tauchen, G. and R. Hussey (1991), "Quadrature-based methods for obtaining approximate solutions to nonlinear asset pricing models." *Econometrica*, 59 (2), 371–396. [43, 46, 51]

Taylor, S. (1982), "Financial returns modelled by the product of two stochastic processes: A study of the daily sugar prices 1961–75." In *Time Series Analysis: Theory and Practice*, Vol. 1, 203–226, North-Holland, Amsterdam. [59]

Van Binsbergen, J. H., J. Fernández-Villaverde, R. S. Koijen, and J. Rubio-Ramírez (2012), "The term structure of interest rates in a DSGE model with recursive preferences." *Journal of Monetary Economics*, 59 (7), 634–648. [44]

Co-editor Tao Zha handled this manuscript.

Manuscript received 23 May, 2019; final version accepted 16 August, 2020; available online 15 September, 2020.