The Discretization Filter: A Simple Way to Estimate Nonlinear State Space Models*

Leland E. Farmer†
Department of Economics
University of California, San Diego

March 26, 2016

Abstract

I develop a new and simple method, the discretization filter, for approximating the likelihood of nonlinear, non-Gaussian state space models. I show that these models can be well approximated by a discrete-state Markov process and the likelihood evaluated using techniques developed in Hamilton (1989). I establish strong consistency and asymptotic normality of my proposed maximum likelihood estimator using results from the literature on uniformly ergodic Markov chains. Through Monte Carlo simulations, I demonstrate that the discretization filter outperforms popular existing methods both in terms of accuracy and computational efficiency for empirically relevant sample sizes. As a result, it can be used to study the quantitative implications of economic models for which existing methods are either intractable or inadequate. I use the discretization filter to estimate two popular nonlinear economic models, a term structure model with a zero lower bound, and the long run risks model of asset pricing. I show that small differences in parameter estimates implied by different methods translate into significant economic effects.

*I would like to especially thank the members of my thesis committee James Hamilton, Allan Timmermann, and Alexis Toda for their invaluable input. I am also grateful for helpful feedback and comments from Roy Allen, Brendan Beare, Roger E. A. Farmer, Jesús Fernández-Villaverde, Patrick Gagliardini, Émilien Gouin-Bonenfant, Michael Levy, Nelson Lind, Eric Renault, Juan Rubio-Ramirez, Andres Santos, Lawrence Schmidt, Frank Schorfheide, Neil Shephard, Harald Uhlig, Daniel Waggoner, Fan Dora Xia, and seminar participants at UCSD, the 2014 SoFiE Summer School for Financial Econometrics at Harvard University, and the 2015 California Econometrics Conference. All errors are my own.

†lefarmer@ucsd.edu

1 Introduction

Many economic models include important nonlinearities. Examples include the zero lower bound on interest rates (ZLB), stochastic volatility, Poisson jumps, non-convex adjustment costs, Markov-switching dynamics, and sovereign default. Models which include these features can often be represented as nonlinear state space models. These models are difficult to estimate because, in general, no closed form solution exists for the recursive updating of the filtering density. As a consequence, the evaluation of the likelihood requires the computation of an intractable $d \times T$-dimensional integral, where d is the dimension of the state space, and T is the sample size.

I propose a new method, the discretization filter, for estimating nonlinear, non-Gaussian state space models. The discretization filter works by constructing an approximation to a state space model where the state variables are assumed to be finite-valued. The resulting approximate system is a finite-state Markov-switching model of the kind studied in Hamilton (1989), for which simple closed form forecasting and updating equations exist, just like the Kalman filter for linear models.

I examine the asymptotic properties of the maximum likelihood estimator of the approximate model, as both the number of discrete points m and the sample size T go to infinity jointly. Building on the work of Douc et al. (2004)[1] and the literature on uniformly ergodic Markov chains, I establish two theoretical results.

First, the maximum likelihood estimator is strongly consistent, and this result does not depend on the rate at which the number of points m grows. Second, the maximum likelihood estimator is asymptotically normal at the usual rate $T^{1/2}$, as long as the number of discrete points m grows quickly enough as a function of the sample size. I show that discretizing the likelihood introduces a bias that is directly related to the discrepancy between the ergodic distributions of the approximate and original Markov chains. Under mild assumptions, this bias decreases to zero as the number of points grows.[2]

After establishing these theoretical results, I provide Monte Carlo evidence that the discretization filter outperforms simulation based methods like the particle filter while maintaining the computational tractability of sequential local approximation methods like the extended or unscented Kalman filters.

I then employ the discretization filter to estimate two nonlinear economic models: the ZLB term structure model proposed by Wu and Xia (2015), and the long run risks model of asset

[1] Other papers which examine the asymptotic properties of the likelihood in nonlinear state space models are Douc et al. (2011)

[2] However, the rate at which m needs to grow to ensure asymptotic normality depends on exactly how the state space is discretized. I show that if the discretization is constructed as in Farmer and Toda (2015), $m^{3/d}$ must grow faster than the sample size T. That is, for low dimensional state spaces, m need only grow linearly with or slower than T. However, as d gets large, m must grow increasingly quickly.

pricing first proposed in Bansal and Yaron (2004). I obtain significantly different estimates for the shadow rate of Wu and Xia (2015). In particular, I show that the shadow rate was as much as 2.2 percentage points lower in July 2012 than the estimate obtained by Wu and Xia (2015) when I use the discretization filter instead of the extended Kalman filter to estimate the model. My findings have important implications for policy makers who use the Wu-Xia shadow rate as an input to determining the effectiveness of unconventional monetary policy.

The paper is organized as follows. Section 3 introduces notation and provides a brief overview of nonlinear, non-Gaussian state space models. Section 4 explains how to approximate any nonlinear state space model with a finite-state Markov-switching model. Section 5 shows how to use the approximate model to construct a maximum likelihood estimator for the parameters of the original model. Section 6 presents the two main theoretical results regarding the asymptotic theory of my proposed estimator. Section 7 provides some practical implementation advice for applied researchers. Section 8 provides some suggestive Monte Carlo evidence in the case of a stochastic volatility model. Section 9 re-examines the Wu and Xia (2015) shadow rate term structure model. In section 10, I estimate the long run risks model. Section 11 concludes.

2 Literature Review

To be completed.

3 The Setting

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 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 et al. (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 homogeneous Markov chain and lies in a separable, compact set $\mathcal{X}$[3], equipped with a metrizable topology and associated Borel σ-field $\mathcal{B}(\mathcal{X})$. Let $Q_\theta(x, A)$, where $x \in \mathcal{X}$, $A \in \mathcal{B}(\mathcal{X})$, be the transition kernel of the Markov chain. I further assume that for all $\theta \in \Theta$ and $x \in \mathcal{X}$, each conditonal probability measure $Q_\theta(x, \cdot)$ has a density $q_\theta(\cdot\,|x)$ with

[3]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 developped in Douc et al. (2011)

respect to a common finite dominating measure μ on $\mathcal{X}$.[4]

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. This is the hidden Markov assumption. I also assume that the observations have a density $g_\theta(y|X_t)$ 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

$$\Pi_\theta f(x,y) = \int_{\mathcal{X}\times\mathcal{Y}} f(x',y')\, q_\theta(x'|x)\, g_\theta(y'|x')\, \mu(dx')\, \nu(dy')$$

for any bounded measurable function f on $\mathcal{Z}$.

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:

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

$$Y_t|X_t \sim g_\theta(Y_t|X_t) \tag{2}$$

Equation 1 is the state equation, and it characterizes the distribution of the state next period conditional on the current state. Equation 2 is the observation, or measurement equation, and it characterizes the distribution of the observables conditional on the current state.

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.[5] In order to do this, one must obtain an expression for the likelihood of the data:

$$L_T(\theta, x_0) = p_\theta\left(\mathbf{Y}_1^T | X_0 = x_0\right) \tag{3}$$

where $\mathbf{Y}_1^T \equiv (Y_1, \ldots, Y_T)$, and X_0 refers to the initial value of the state vector. In the subsequent section, I show how to approximate equation (1) by a finite-valued Markov chain.[6] In section 5, I show how to use this approximate model to obtain an expression for the likelihood. Finally in section 6, I present my main results on the asymptotic theory of this approximate maximum likelihood estimator.

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

[5] 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, might be of interest.

[6] I also provide overviews of two of the more popular methods for approximating the likelihood of nonlinear state space models, the extended Kalman filter and the bootstrap particle filter with adaptive resampling, in appendices **??** and **??** respectively.

4 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). He proposes a simple way of approximating any Gaussian VAR with a first-order, discrete-valued Markov chain, and shows that this approximation does a good job of matching unconditional and conditional moments for relatively coarse discretizations. His paper, along with several more recent suggestions, have been widely used in solving DSGE models where the ability to approximate the solutions to integral equations is of key importance.[7]

In this paper I apply this method to the estimation of nonlinear, non-Gaussian state space models. More specifically, I construct a finite-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 $Q_{\theta,m}$.[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 finite set of m points in $\mathcal{X}$, $\mathcal{X}_m \equiv \{x_{i,m}\}_{i=1}^{m}$, and define a transition matrix $Q_{\theta,m}$ such that the ij^{th} element:

$$Q_{\theta,m}(i,j) = \bar{\mathbb{P}}_\theta\left(X_{t,m} = x_{j,m} \,|X_{t-1,m} = x_{i,m}\right) \tag{4}$$

corresponds to the probability of transitioning from point $x_{i,m}$ to point $x_{j,m}$ between time $t-1$ and t. The matrix $Q_{\theta,m}$ is assumed to be the same for all t, and thus $X_{t,m}$ follows a first-order, time homogeneous, m-state Markov chain.

Note that each row of the matrix $Q_{\theta,m}$ can be interpreted as a conditional distribution. Specifically, row i correponds to the distribution of $X_{t,m}$ conditional on being at point $x_{i,m}$. It is critical that these conditional distributions be good approximations to the true conditional distributions $q_\theta\left(X_t \,|X_{t-1,m} = x_{i,m}\right)$.

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 $s_{t,m} = i$ and let $\zeta_{t,m} = e_i$ when $X_{t,m} = x_{i,m}$, where e_i is the i^{th} column of the $(m \times m)$ identity matrix. The system outlined above is characterized by the equations:

$$\zeta_{t,m} = Q_{\theta,m}^T \zeta_{t-1,m} + \tilde{v}_{t,m} \tag{5}$$

$$Y_t \,|X_{t,m} \sim g_\theta\left(Y_t \,|X_{t,m}\right) \tag{6}$$

[7] See e.g. Tauchen and Hussey (1991), Rouwenhorst (1995), Adda and Cooper (2003), Flodén (2008), Tanaka and Toda (2013), Gospodinov and Lkhagvasuren (2013), and Farmer and Toda (2015).

[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.

where $\tilde{v}_{t,m} = \zeta_{t,m} - \mathbb{E}\left[\zeta_{t,m} \left| \zeta_{t-1,m} \right.\right]$ and $Q_{\theta,m}^T$ is the transpose of the matrix $Q_{\theta,m}$. Equations (5) and (6) 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 finite grid.

5 Evaluating the Likelihood

In the previous section, I showed how to approximate any HMM by replacing the state equation, Equation (1), with a finite-state Markov chain, Equation (5). 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 finite-state Markov chain, simple prediction and updating equations exist that are analogous to the Kalman filter in the linear case. I review these results here and show how they can be used to develop a maximum likelihood estimator for θ.

Let $\hat{\zeta}_{t,m|t} = \mathbb{E}\left[\zeta_{t,m} \left| \mathbf{Y}_1^t \right.\right]$ be the econometrician's best inference about the discretized state $\zeta_{t,m}$ conditional on time t information. Inuitively, $\hat{\zeta}_{t,m|t}$ will be an $(m \times 1)$ vector of probabilities which represent the likelihood of being at a particular point in the state space. The forecast of the approximate state today given the previous period's information is given by:

$$\hat{\zeta}_{t,m|t-1} = \mathbb{E}\left[\zeta_{t,m} \left| \mathbf{Y}_1^{t-1} \right.\right] = Q_{\theta,m}^T \hat{\zeta}_{t-1,m|t-1} \tag{7}$$

Also define

$$\eta_{t,m} = \begin{bmatrix} g_\theta\left(Y_t \left| X_t = x_{1,m}\right.\right) \\ \vdots \\ g_\theta\left(y_t \left| X_t = x_{m,m}\right.\right) \end{bmatrix} \tag{8}$$

The i-th element of $\eta_{t,m}$ is the likelihood of having observed Y_t conditional on being in state i, i.e. $s_{t,m} = i$.

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

$$p_{\theta,m}\left(Y_t \left| \mathbf{Y}_1^{t-1} \right.\right) = 1'\left(\eta_{t,m} \odot \hat{\zeta}_{t,m|t-1}\right) \tag{9}$$

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

$$\hat{\zeta}_{t,m|t} = \frac{\eta_{t,m} \odot \hat{\zeta}_{t,m|t-1}}{1'\left(\eta_{t,m} \odot \hat{\zeta}_{t,m|t-1}\right)} = \frac{\eta_{t,m} \odot \hat{\zeta}_{t,m|t-1}}{p_{\theta,m}\left(Y_t \left| \mathbf{Y}_1^{t-1} \right.\right)} \tag{10}$$

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

$$\ell_{T,m}(\theta) = \frac{1}{T}\sum_{t=1}^{T} \log p_{\theta,m}\left(Y_t \left| \boldsymbol{Y}_1^{t-1}\right.\right) \tag{11}$$

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

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 6, I will show why this initial condition is irrelevant for the asymptotic properties of $\hat{\theta}_{T,m}$.

Section 6 proves the consistency and the asymptotic normality of the discretization filter. Those who are interested in applications of the discretization filter may wish to skip ahead to section 7.

6 Asymptotic Properties of the Maximum Likelihood Estimator

In this section I establish strong consistency and asymptotic normality 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 is governed to first order by the proximity of the ergodic distributions of the approximate and true chains. The distance between these distributions is governed by a number proportional to $m^{3/d}$ where d is the dimension of the state space and m is the number of points. Asymptotic normality requires that this number goes to infinity faster than the sample size T.

The assumptions and intermediate results leading to strong consistency follow the exposition and techniques in Douc et al. (2004). I require some additional assumptions that allow me to establish convergence uniformly over the number of discrete points used in the approximation. A key new 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 proofs can be found in Appendix A.

6.1 Preliminaries and Assumptions

I now list my basic assumptions. Assumptions labeled A are assumptions that overlap with Douc et al. (2004) and those labeled B are additional assumptions I make for my particular setting.

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

(B1) $0 < Q_+^- \equiv \sup_{m\in\mathbb{Z}^+} \frac{\inf_{\theta\in\Theta} \inf_{1\le i,j\le m} Q_{\theta,m}(i,j)}{\sup_{\theta\in\Theta} \sup_{1\le i,j\le m} Q_{\theta,m}(i,j)} < 1$.

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's 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 has a unique invariant distribution 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.

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

(B2) For all $\theta \in \Theta$, the transition kernel $Q_{\theta,\infty}$ is positive recurrent and aperiodic with invariant distribution $\pi_{\theta,\infty}$.

These assumptions guarantee that the original joint Markov process $\{Z_t\}$ and the limiting Markov chain with countably infinite state space $\{Z_{t,\infty}\}$ are themselves uniformly ergodic. Note that assumption (B2) is needed in addition to assumption (B1) to account for the limiting case of the chain. Strict positivity of every entry of $Q_{\theta,\infty}$ is not enough to guarantee that an ergodic distribution exists in the countably infinite case.

Assumption (A2) implies that for any initial measure λ,

$$\lim_{t\to\infty} \left\| \lambda \Pi_\theta^t - \pi_\theta \right\|_{TV} = 0 \tag{12}$$

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

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

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. This convergence is also independent of the initial measure λ. An analogous property holds for the process $\{Z_{t,\infty}\}$ by assumption (B2). Developing a bound on this rate of convergence will be critical for the coming developments.

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

(B3) $\bar{\mathbb{E}}_{\theta^*}(|\log c_-(y_1)|) < \infty$, where $c_-(y_1) \equiv \inf_{\theta\in\Theta} \inf_{m\in\mathbb{Z}^+} Q^-_{\theta,m} \sum_{i=1}^m g_\theta(y_1|x_{i,m})$, and $Q^-_{\theta,m} \equiv \inf_{i,j} Q_{\theta,m}(i,j)$

Assumptions (A3) and (B3) are additional regularity conditions. Additional assumptions will be introduced and explained as needed.

6.2 Consistency

Following Douc et al. (2004), I first establish that the conditional chain, $X_{t,m}$, given a history of observations $\mathbf{Y}_r^s$, is itself a uniformly ergodic (inhomogeneous) Markov chain with minorizing constant independent of the parameter $\theta \in \Theta$ and the number of discrete points $m \in \mathbb{Z}^+$. This is the analogous result to Lemma 1 in their paper.

Lemma 1. *Assume (A1) and (B1). Let $s, r \in \mathbb{Z}$, with $s \leq r$, $\theta \in \Theta$, and $m \in \mathbb{Z}^+$. Under $\bar{\mathbb{P}}_\theta$, conditionally on $\mathbf{Y}_r^s$, $\{X_{t,m}\}_{t\geq r}$ is an inhomogeneous Markov chain, and for all $t > r$ there exists a function $\mu_{t,m}(\boldsymbol{y}_t^s, A)$ such that:*

(i) for any $A \in \mathcal{B}(\mathcal{X}_m)$, $\boldsymbol{y}_t^s \mapsto \mu_{t,m}(\boldsymbol{y}_t^s, A)$ is a Borel function;

(ii) for any $\boldsymbol{y}_t^s$, $\mu_{t,m}(\boldsymbol{y}_t^s, \cdot)$ is a probability measure on $\mathcal{B}(\mathcal{X}_m)$. In addition, for all $\boldsymbol{y}_t^s$ it holds that $\mu_{t,m}(\boldsymbol{y}_t^s, \cdot) \ll \mu_{c,m}$ (where $\mu_{c,m}$ is counting measure on $\mathcal{X}_m$) and for all $\mathbf{Y}_r^s$,

$$\inf_{x\in\mathcal{X}_m} \bar{\mathbb{P}}_\theta(X_{t,m} \in A | X_{t-1,m} = x, \mathbf{Y}_r^s) \geq Q^-_+ \mu_{t,m}(\mathbf{Y}_t^s, A)$$

The major difference between this Lemma and the one established in Douc et al. (2004) is that for the following results, it will be crucial that the minorizing constant be the same for all m, in order to establish uniform convergence over $m \in \mathbb{Z}^+$ of the approximate likelihood function. Note that although the minorizing measure, $\mu_{t,m}(\mathbf{Y}_t^s, \cdot)$, *does* depend on both the number of points, m, and the observations the chain is conditioned on, $\mathbf{Y}_t^s$, it doesn't affect the mixing rate. The previous lemma leads to the following corollary, using standard results for uniformly minorized Markov chains.

Corollary 1. *Assume (A1) and (B1). Let $r, s \in \mathbb{Z}$ with $r \leq s$, $\theta \in \Theta$, and $m \in \mathbb{Z}^+$. Then for all $t \geq r$, all probability measures μ_1 and μ_2 on $\mathcal{B}(\mathcal{X}_m)$, and all $\mathbf{Y}_r^s$,*

$$\left\| \int_{\mathcal{X}_m} \bar{\mathbb{P}}_\theta \left(X_{t,m} \in \cdot \, | X_{r,m} = x, \mathbf{Y}_r^s \right) \mu_1 (dx) - \int_{\mathcal{X}_m} \bar{\mathbb{P}}_\theta \left(X_{t,m} \in \cdot \, | X_{r,m} = x, \mathbf{Y}_r^s \right) \mu_2 (dx) \right\|_{TV} \leq \rho^{t-r}$$

where $\rho \equiv 1 - Q_+^-$.

This corollary establishes that the Markov chain "uniformly forgets" its history at an exponential rate. That is, no matter where the chain is started, it converges to its ergodic distribution exponentially fast. This is analogous to ϕ-mixing, and the fact that the bound is deterministic will be important for establishing strong consistency.

The next step consists of showing that the approximate likelihood function $\ell_{T,m}(\theta, x_{0,m})$ with an arbitrary initial condition $x_{0,m}$ stays within a deterministic bound of $\ell_{T,m}(\theta)$ where $X_{0,m}$ is drawn from its ergodic distribution.

Lemma 2. *Assume (A1)-(A2) and (B1)-(B2). Then, for all $x_0 \in \mathcal{X}$ and $m \in \mathbb{Z}^+$,*

$$\sup_{\theta \in \Theta} |\ell_{T,m}(\theta, x_0) - \ell_{T,m}(\theta)| \leq 1/(1-\rho)^2, \quad \bar{\mathbb{P}}_{\theta^*}\text{-a.s.}$$

Next I show that $T^{-1}\ell_{T,m}(\theta)$ can be approximated by the sample mean of a $\bar{\mathbb{P}}_{\theta^*}$-stationary ergodic sequence of bounded random variables which has a well defined limit. To this end I first define the quantities:

$$\Delta_{t,r,x,m}(\theta) \equiv \log \bar{p}_{\theta,m}\left(Y_t \left| \mathbf{Y}_{-s}^{t-1}, X_{-s,m} = x \right. \right)$$
$$\Delta_{t,r,m}(\theta) \equiv \log \bar{p}_{\theta,m}\left(Y_t \left| \mathbf{Y}_{-s}^{t-1} \right. \right) = \int \log \bar{p}_{\theta,m}\left(Y_t \left| \mathbf{Y}_{-s}^{t-1}, X_{-s,m} = x \right. \right) \bar{\mathbb{P}}_\theta \left(dx_{-s,m} \left| \mathbf{Y}_{-s}^{t-1} \right. \right)$$

Consider the thought experiment of fixing the number of points m, but letting $T \to \infty$. Define the limiting object as

$$\ell_m(\theta) \equiv \bar{\mathbb{E}}_{\theta^*}\left[\Delta_{0,\infty,m}(\theta) \right]$$

I will show that such a limiting object is well-defined and that the sample analogue converges to this limit almost-surely. In particular, I will show that $\{\Delta_{t,r,m}\}_{r \geq 0}$ and $\{\Delta_{t,r,m,x}\}_{r \geq 0}$ converge uniformly w.r.t. $\theta \in \Theta$ $\bar{\mathbb{P}}_{\theta^*}$-a.s. by showing they are uniform Cauchy sequences.

Lemma 3. *Assume (A1)-(A3) and (B1)-(B3). Then for all $t \geq 1$, $r, r' \geq 0$, and $m \in \mathbb{Z}^+$,*

$\bar{\mathbb{P}}_{\theta^*}$-*a.s.*,

$$\sup_{\theta\in\Theta}\sup_{x,x'\in\mathcal{X}}\left|\Delta_{t,r,x,m}(\theta)-\Delta_{t,r',x',m}(\theta)\right|\leq\rho^{t+\min(r,r')-1}/(1-\rho), \tag{13}$$

$$\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left|\Delta_{t,r,x,m}(\theta)-\Delta_{t,r,m}(\theta)\right|\leq\rho^{t+r-1}/(1-\rho), \tag{14}$$

$$\sup_{\theta\in\Theta}\sup_{r\geq 0}\sup_{x\in\mathcal{X}}\left|\Delta_{t,r,x}(\theta)\right|\leq\max\left(\left|\log b_+\right|,\left|\log c_-(Y_t)\right|\right) \tag{15}$$

Lemma 3 shows that $\{\Delta_{t,r,m,x}\}_{r\geq 0}$ is a uniform Cauchy sequence w.r.t. $\theta\in\Theta$ and thus converges $\bar{\mathbb{P}}_{\theta^*}$-a.s. to a limit which does not depend on the initial value x. I label this limit $\Delta_{t,\infty,m}$ and intuitively this can be thought of as $\log\bar{p}_{\theta,m}\left(Y_t\left|\boldsymbol{Y}_{-\infty}^{t-1}\right.\right)$.

Corollary 2. *Assume (A1)-(A2) and (B1)-(B2). Then*

$$\sum_{t=1}^{T}\sup_{m\in\mathbb{Z}^+}\sup_{\theta\in\Theta}\left|\Delta_{t,0,m}(\theta)-\Delta_{t,\infty,m}\right|\leq 2/(1-\rho)^2,\quad\bar{\mathbb{P}}_{\theta^*}\text{-}a.s.$$

Corollary 2 shows that $T^{-1}\ell_{T,m}(\theta)$ can be approximated by the sample mean of a uniformly ergodic sequence. The next step is to prove that the ergodic distribution of the approximate discrete Markov chain converges weakly to that of the original continuous Markov chain. Proposition 1 establishes this convergence and provides a bound on the difference between the two distributions as a function of the number of points m.

Proposition 1. *Assume (A1)-(A3) and (B1)-(B3). Without loss of generality, assume that the state variable X_t takes values in the d_X-dimensional unit cube, $[0,1]^{d_X}$. Define J to be the set of d_X-dimensional boxes of the form*

$$\prod_{i=1}^{d_X}[0,a_i)=\left\{x\in\mathbb{R}^{d_X}:0\leq x_i<a_i\right\}$$

where $0\leq a_i<1$. Further assume that the approximation of the transition kernel satisfies

$$\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\sup_{A\in J}\left|Q_{\theta,m}(x,A)-Q_\theta(x,A)\right|=o(h(m))$$

Then it follows that

$$\sup_{\theta\in\Theta}\sup_{A\in J}\left|\pi_{\theta,m}^X(A)-\pi_\theta^X(A)\right|=o(h(m))$$

where the function $h(m)$ depends on the underlying formula used to construct the discretization. If the transition kernel is approximated as proposed in Farmer and Toda (2015), the rate of convergence is directly related to the underlying quadrature formula. For example, if the class

of Newton-Coates quadrature rules is used, the error will be at most of order

$$h(m) = \frac{1}{m^{(3-\delta)/d}}$$

for $\delta > 0$.

Note that even faster rates can be achieved through clever choice of the quadrature formula and the assumptions one is willing to make about the smoothness of the likelihood function. [9]

Combining proposition 1, corollary 2, and lemma 2 leads to the following pointwise convergence result

Corollary 3. *Assume (A1)-(A3) and (B1)-(B3). Then for all sequences of initial points* $\{x_{0,m}\}$ *and* $\theta \in \Theta$,

$$\lim_{m,T\to\infty} T^{-1}\ell_{T,m}(\theta, x_{0,m}) = \ell(\theta), \quad \bar{\mathbb{P}}_{\theta^*}\text{-a.s. and in } L^1\left(\bar{\mathbb{P}}_{\theta^*}\right)$$

The final step before I can state the strong consistency result involves showing that $\ell_m(\theta)$ is continuous w.r.t. θ for all $m \in \mathbb{Z}^+$. Note that by (15) and the dominated convergence theorem,

$$\ell_m(\theta) = \bar{\mathbb{E}}_{\theta^*}\left[\lim_{r\to\infty} \Delta_{0,r,m,x}(\theta)\right] = \lim_{r\to\infty} \bar{\mathbb{E}}_{\theta^*}\left[\Delta_{0,r,m,x}(\theta)\right]$$

It suffices to show that $\Delta_{0,r,m,x}(\theta)$ is continuous w.r.t θ, since $\{\Delta_{0,r,m,x}(\theta)\}_{r\geq 0}$ is a uniform Cauchy sequence $\bar{\mathbb{P}}_{\theta^*}$-a.s. which is uniformly bounded in $L^1\left(\bar{\mathbb{P}}_{\theta^*}\right)$.

The following additional assumptions are needed to establish continuity

(A4) For all $x, x' \in \mathcal{X}$ and all $y' \in \mathcal{Y}$, $\theta \mapsto q_\theta(x, x')$ and $\theta \mapsto g_\theta(y'|x)$ are continuous.

(B4) For all $m \in \mathbb{Z}^+$ and $i, j = 1, \ldots, m$, $\theta \mapsto Q_{\theta,m}(i, j)$ is continuous.

Lemma 4. *Assume (A1)-(A4), then*

$$\lim_{\delta\to 0} \bar{\mathbb{E}}_{\theta^*}\left[\sup_{m\in\mathbb{Z}^+} \sup_{|\theta'-\theta|\leq\delta} \left|\Delta_{t,\infty,m}(\theta') - \Delta_{t,\infty,m}(\theta)\right|\right] = 0.$$

Corollary 4. *Assume (A1)-(A3), then*

$$\lim_{m\to\infty} \sup_{\theta\in\Theta} |\ell_m(\theta) - \ell(\theta)| = 0$$

[9] There has been much research in the field of Quasi Monte-Carlo integration methods, which seek deterministic sequences to approximate high dimensional integrals which break the curse of dimensionality. These are referred to as low discrepancy sequences and their accuracy for numerical integration has been shown to depend only polynomially on the dimension d rather than exponentially. The use of these sequences to approximate the dynamics of high dimensional state processes is a promising area of study which I investigate in ongoing research. Further, there are no known convergence rates for the Tauchen or point mass filter approximations to the transition kernel and I leave this for future work.

Corollaries 2 and 4 allow me to establish almost-sure convergence of the average approximate sample likelihood.

Proposition 2. *Assume (A1)-(A4) and (B1)-(B4). Then*

$$\lim_{m,T\to\infty} \sup_{\theta\in\Theta} \sup_{x_{0,m}\in\mathcal{X}_m} \left|T^{-1}\ell_{T,m}\left(\theta, x_{0,m}\right) - \ell\left(\theta\right)\right| = 0, \quad \bar{\mathbb{P}}_{\theta^*}\text{-a.s.}$$

The last assumption needed to establish consistency is an identification assumption guaranteeing that θ^* is a unique maximizer of the likelihood function

(A5) $\theta = \theta^*$ if and only if

$$\bar{\mathbb{E}}_{\theta^*}\left[\log \frac{\bar{p}_{\theta^*}\left(\boldsymbol{Y}_1^t\right)}{\bar{p}_{\theta}\left(\boldsymbol{Y}_1^t\right)}\right] = 0 \qquad \text{for all } t \geq 1. \tag{16}$$

This is a high level assumption about the identification of the model. In general this is a difficult condition to verify because it relies on the ergodic distribution of the joint Markov chain $\{Z_t\}$. For a more thorough discussion on when this assumption is satisfied in the context of HMM, see Douc et al. (2011). Under the additional assumption (A5), I am ready to state my first main result, strong consistency of the maximum likelihood estimator

Theorem 1. *Assume (A1)-(A5) and (B1)-(B4). Then, for any sequence of initial points* $x_{0,m} \in \mathcal{X}_m$, $\hat{\theta}_{T,m,x_{0,m}} \to \theta^*$, $\bar{\mathbb{P}}_{\theta^*}$*-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, this consistency is *independent* of the rate at which the number of points m grows.

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 $\theta \mapsto q_\theta\left(x, x'\right)$ and $\theta \mapsto g_\theta\left(y' \left| x'\right.\right)$ are twice continuously differentiable on G.

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

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

(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 $\left\|\nabla_\theta^2 g_\theta(y'|x)\right\| \leq f_x^2(y')$ for all $\theta \in G$.

Instead of re-establishing asymptotic normality of my proposed estimator using the techniques in Douc et al. (2004), I use Theorem 7 from their paper. I reproduce the theorem here for completeness.

Theorem 2 (Theorem 7 from Douc et al. (2004))**.** *Assume that $\tilde{\theta}_{T,x_0}$ is an estimator satisfying $\ell_T\left(\tilde{\theta}_{T,x_0}, x_0\right) \geq \sup_{\theta \in \Theta} \ell_T(\theta, x_0) - R_T$ and assumptions (A1)-(A8) hold. Then the following are true:*

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

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

(iii) If $R_T = o_p(1)$, then $T^{1/2}\left(\tilde{\theta}_{T,x_0} - \theta^*\right) \to N\left(0, I(\theta^*)^{-1}\right)$, $\bar{\mathbb{P}}_{\theta^*}-$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 leverage 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 3. *Assume (A1)-(A8), (B1)-(B4), and that $I(\theta^*)$ is positive definite. Then for any sequence of initial points $x_{0,m} \in \mathcal{X}_m$,*

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

$\bar{\mathbb{P}}_{\theta^*}$-weakly as $T \to \infty$, $m \to \infty$, and $Th(m) \to 0$.

7 Recommendations for Applied Researchers

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

7.1 Constructing the Transition Matrix

There is no unique way to construct the transition matrix for the discretization filter. I propose two possible ways of doing so.[10]

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 $\{A_i\}_{i=1}^m$, where each A_i is associated with discretization point $x_{i,m}$ for all $i = 1, \ldots, m$ (think of intervals in the one dimensional case). Then define:

$$Q_{\theta,m}(i,j) = \int_{A_j} q_\theta(x \,|X_{t-1} = x_{i,m}) \, \mu(dx) \tag{17}$$

Intuitively, there are two layers of approximation in this expression. First, I am assuming that if X_{t-1} is in region A_i it is close to the point $x_{i,m}$ in the sense that the conditional distribution $q_\theta(X_t \,|X_{t-1})$ can be well approximated by $q_\theta(X_t \,|X_{t-1} = x_{i,m})$. Second, I am assuming that the probability of transitioning to region A_j from point $x_{i,m}$ is similar to the conditional density $q_\theta(X_t = x_{j,m} \,|X_{t-1} = x_{i,m})$ over the set A_j.

A limitation of this approach is the ability to evaluate the integrals needed to construct the transition matrix. In general, this method will only work well in practice when the A_i 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 the that of the underlying continuous process. Since this rate is critical to obtaining asymptotic normality, researchers should be cautious about standard errors when using this approach with a small number of points.

An alternative approach is construct the transition matrix as in Farmer and Toda (2015). They provide a general way of constructing finite-state Markov chain approximations to stochastic processes. Their method finds the discrete distribution which is "closest" to the original distribution from some prior distribution in terms of Kullback-Leibler distance, while matching a set of conditional moments of the underlying continuous distribution.

If the prior distribution is a valid quadrature formula for evaluating integrals with respsect to the original conditional density, the discrete approximation is guaranteed to converge weakly to the continuous distribution. Moreover, the rate of convergence is given by the rate of convergence of the selected quadrature formula.[11]

[10] 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 (2013). However, all of these with the exception of Tauchen and Hussey (1991) only apply to linear autoregressive processes.

[11] 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, i.e. $Q_{\theta,m}(i,j) \sim p(x_{i,m} \,|x_{j,m})$. 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

My Monte Carlo results in section 8 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 Judd et al. (2012).

7.2 Selecting the Grid Points

When establishing my theoretical results, I assumed that the state space is compact. 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, how many discretization points (m) to use, and how they should be spaced.

First, consider the case where the number of discretization points, m, has been fixed and the goal is to choose the support of the distribution. 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 its ergodic distribution.

When the state follows a Gaussian VAR(1), a closed form expression for the ergodic distribution is available. Gospodinov and Lkhagvasuren (2013) provide 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. I rely on this method in section 9 for my empirical application. However, for more general time series models, no such expression 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 covering some number of unconditional standard deviations of the process. $\sqrt{m-1}$ seems to work well in practice, which is the number employed in the Rouwenhorst (1995) approximation.

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 sample and burn in periods are sufficiently large, the remaining points can be treated as representative draws from the ergodic distribution. An appealing choice is to use empirical quantiles as the discretization points.

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

product approximations quickly become intractable in higher dimensions, and for this reason the point-mass filter is infrequently used. A comprehensive survey article on the properties and applications of filtering techniques is Chen (2003).

as $\{x_s\}_{s=1}^{S}$. Then, if one wants the state to lie in the grid with approximately $1-\alpha$ probability, select:

$$x_{i,m} = \hat{Q}_S\left(\tfrac{\alpha}{2} + \tfrac{i-1}{m}\left(1-\alpha\right)\right) \quad \forall i = 1, \ldots, m$$

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

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

Selecting the points in this way has the desirable property that one expects roughly the same number of realizations of the state to fall between each point.[12] 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.[13] 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.[14]

8 Monte Carlo Evidence

This section presents some Monte Carlo results comparing the performance of the discretization filter with the extended Kalman filter (EKF from now on) and bootstrap paticle filter (PF from now on). I consider the case of a basic stochastic volatility model.

8.1 Stochastic Volatility

The standard discrete time stochastic volatility model, as formulated in Taylor (1982), is given by

$$X_t = \mu + \rho\left(1-\mu\right)X_{t-1} + v_t \qquad v_t \sim \text{i.i.d. } N\left(0, \sigma^2\right) \tag{18}$$

$$Y_t = e^{X_t/2} w_t \qquad w_t \sim \text{i.i.d. } N\left(0,1\right) \tag{19}$$

[12] 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.

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

[14] Another possibility is to construct an ε-distinguishable set as proposed by Judd et al. (2012), although this is subject to the same criticisms about introducing simulation.

Note that the measurement equation can be equivalently rewritten as:

$$\log\left(Y_t^2\right) = X_t + \log\left(w_t^2\right) \tag{20}$$

which leads to an additively separable state equation.[15] 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). My results are not sensitive to this parameterization.

8.2 Parameter Estimation

I simulate data for T = 100, 500, and 1,000 periods, and compute the likelihood of the model five different ways. The first three ways are with my method, labeled as DF for "discretization filter," using 5, 9, and 13 discretization points respectively, the fourth is the EKF, and the fifth is the PF with adaptive resampling using 1,000 particles. Each specification is simulated 1,000 times and estimation is performed via maximum likelihood where optimization is done using MATLAB's genetic algorithm in the global optimization toolbox.[16]

Figures 1-3 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. A vertical line is displayed at the point of the true parameter value. All estimation using the DF uses the Rouwenhorst (1995) discretization scheme.[17]

Consider figures 1 through 3. 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 model, where it is treating the residual in the observation equation as a normal random variable, even though it has a $\log\left(\chi_1^2\right)$ distribution.

This bias vanishes for both the DF and PF 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.

[15] This is the specification of the observation equation 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.

[16] Note that traditional gradient based optimization methods are inapplicable to the PF because the likelihood function is simulated, which makes it non-differentiable. See Flury and Shephard (2011) for a more detailed discussion.

[17] Estimation was also performed using the Tauchen method and the point-mass filter. The Rouwenhorst method does the best though the relative gains of the DF are similar for all discretization methods.

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

$$\text{RMSE}\left(\hat{\theta}_i\right) = \sqrt{\mathbb{E}\left[\left(\hat{\theta}_i - \theta_i\right)^2\right]} \tag{21}$$

$$\text{MAE}\left(\hat{\theta}_i\right) = \mathbb{E}\left[\left|\hat{\theta}_i - \theta_i\right|\right] \tag{22}$$

$$\text{Bias}\left(\hat{\theta}_i\right) = \mathbb{E}\left[\theta_i\right] - \theta_i \tag{23}$$

and report the results in tables 1, 2, and 3. The results for the DF and PF are similar for $T = 100$, however the DF unambiguously outperforms both the PF and EKF for the larger sample sizes. It is also interesting to note that the EKF outperforms the PF for the largest sample size, which seems to be evidence of sample thinning, a well known problem in importance sampling methods.

Given computational constraints, I do not have results for the PF with different numbers of particles. However one would expect the particle filter to improve as the number of particles increases. Given that my method runs so much faster than this specification of the PF, it may be informative to compare it to the PF with a smaller number of particles, or to increase the number of discretization points used for my method. I leave these comparisons for future work.

8.3 Computational Complexity

This subsection presents some results on the computational complexity of the three methods. My code is written entirely in MATLAB and simulations were performed on an Intel Xeon CPU E31225, 3.10 GHz processor with 8 GB of RAM.

Table 4 displays the average simulation times for all 5 specifications. The differences in computational time are stark. With 5 discretization points, the EKF is 4 times faster than my method for small sample sizes and 12 times faster for large ones. However the PF is 32 times slower for small sample sizes and 293 times slower for large ones. These results suggest that my method is closer to the EKF in terms of computational burden than the PF. To give the reader a rough idea, all of the simulations for my method and the EKF ran in a matter of minutes whereas the most computationally burdensome PF specification (T = 1,000) took almost five days to run operating in parallel on four cores.

8.4 Filtered States

Another important dimension for comparison is the accuracy of the filtered states, $\left\{\hat{x}_{t|t}\right\}_{t=1}^{T}$. In this subsection, I provide results on the root mean square error (RMSE) and the mean absolute

error (MAE) of all the methods. For a given model specification and method, these are defined as:

$$\text{RMSE} = \left(\frac{1}{T} \sum_{t=1}^{T} \left(\hat{x}_{t|t} - x_t \right)^2 \right)^{1/2} \tag{24}$$

$$\text{MAE} = \frac{1}{T} \sum_{t=1}^{T} \left| \hat{x}_{t|t} - x_t \right| \tag{25}$$

I define the average RMSE (ARMSE) and average MAE (AMAE) to be the average of the RMSE and the MAE across simulations for a given method. Tables 5 and 6 display the ARMSE and AMAE of each method, where the filtering is done using the corresponding maximum likelihood estimates of the parameters for a given sample.

The DF and PF perform roughly the same for small samples and the DF seems to perform marginally better than the PF in the large samples. It appears as though the misspecification of the EKF is translating into poor estimates of the unobserved state.

9 A Term Structure Model with a Zero Lower Bound

In this section, I re-estimate the term structure model proposed in Wu and Xia (2015), and provide an updated estimate of their shadow rate series with data through January 2014. Using the discretization filter I am able to replicate most of their parameter estimates. While my filtered series and the Wu and Xia estimates match closely over most of the sample, they diverge after the onset of the zero lower bound in January 2009. My estimates indicate that the shadow rate was roughly 2.2 percentage points lower in July 2012 than the Wu and Xia estimates would indicate. I conjecture that the estimates differ because the DF provides a more accurate approximation than the EKF to nonlinearities in the state space when the zero lower bound is in effect. Furthermore, the EKF estimator is in general not consistent, while the DF estimator is.

I omit details of the derivation of their shadow rate model. What is key for my purposes is that under the presence of a zero lower bound on short term interest rates, they are able to derive an approximate nonlinear state space model characterizing movements of the yield curve:

$$X_t = \mu + \rho X_{t-1} + \Sigma v_t \qquad v_t \sim \text{i.i.d. } N(0, I_3) \tag{26}$$

$$Y_{n,n+1,t} = \underline{r} + \sigma_n^{\mathbb{Q}} g\left(\frac{a_n + b_n' X_t - \underline{r}}{\sigma_n^{\mathbb{Q}}} \right) + w_t \qquad w_t \sim \text{i.i.d. } N(0, \omega) \tag{27}$$

where $Y_{n,n+1,t}$ corresponds to the forward rate at time t for a loan starting at $t+n$ and maturing

at $t+n+1$, and X_t is a (3×1) vector of latent factors which explain movements in the yield curve. For a derivation of the expressions for a_n, b_n, and $\sigma_n^{\mathbb{Q}}$, I refer the reader to Wu and Xia (2015).

Using their data on the 3 and 6 month, 1, 2, 5, 7, and 10 year forward rates, one has 7 observation equations, one for each observed yield maturity. They further assume that each forward rate is observed with normally distributed measurement error with the same variance ω. θ is a (22×1) vector of structural parameters.[18]

Table 7 reports maximum likelihood estimates of the parameters from the model with QMLE robust standard errors[19], using the DF with 9 discretization points along each dimension (i.e. $9^3 = 729$ total discretization points). I use the Gospodinov and Lkhagvasuren (2013) method to discretize the VAR state dynamics, which generalizes the method of Rouwenhorst (1995) to VAR(1) systems. I also include the estimates from Wu and Xia (2015) for comparison.

Though the parameter estimates obtained using the DF are similar, they produce a drastically different shadow rate series in the zero lower bound period (from about January 2009 onward).[20]. These differences are illustrated in figure 4. I include 95% standard error bands for the shadow rate series estimated with the DF (where the randomness is coming from uncertainty about the state, not the parameters). This emphasizes the fact that the method used to estimate a nonlinear dynamic model can have important economic implications.

One appealing aspect of the discretization filter is that filtering directly outputs a full probability distribution over the unobserved state for each time period in the sample along with the transition matrix $Q_{\theta,m}$. It is then straightforward to compute any moments or forecasts of the unobserved state one wishes, and also use $Q_{\theta,m}$ to compute its stationary distribution. I demonstrate one of many potential uses of this information in figure 5, which plots the analogous picture to figure 10 in Wu and Xia (2015).

The blue dots and red squares, in figure 5, represent market expectations of when the economy will exit the zero lower bound, computed in two different ways. The blue dots are computed using the DF introduced in this paper, and the red squares are computed using the EKF, as explained in Wu and Xia (2015). The green vertical lines represent the dates of Fed announcements which explicitly mention a liftoff date.

The expected liftoff date is defined as the first period when the implied value of the shadow rate is greater than $\underline{r}$. Notice that the two methods lead to some significant differences in expected liftoff dates, and the DF indicates a generally more pessimistic market outlook over

[18] It has been pointed out that the results may be sensitive to the arbitrary choice of $\underline{r} = 0.25$, see Bauer and Rudebusch (2013). As a robustness check I also estimate $\underline{r}$ as a free parameter and find that it has little effect on subsequent analysis.

[19] See Hamilton (1994) for details.

[20] Note that once the parameter vector is estimated, I use the DF with 33 discretization points along each dimension to produce more smoothly varying filtered series. However, the qualitative difference remains even for coarser discretizations.

most of the period.

10 An Empirical Evaluation of the Long Run Risks Model

To be completed.

11 Conclusion

In this paper, I develop a new method, the discretization filter, for estimating nonlinear, non-Gaussian state space models. I establish that the maximum likelihood estimator associated with the discretization filter is strongly consistent and asymptotically normal.

Existing methods like the particle filter, on the one hand, are accurate, but too computationally intensive to be used for complex problems. On the other hand, the extended Kalman filter is fast, but incapable of capturing interesting nonlinearities. The discretization filter retains the best features of both. As a result, it can be used to study the quantitative implications of economic models for which existing methods are either intractable or inadequate.

I demonstrate that the filtering method used to estimate nonlinear models has sizeable effects on the accuracy of the estimated parameters and the accuracy of the filtered states. I show that these estimation differences translate into quantitatively significant economic differences using the Wu and Xia (2015) shadow rate model as an example. My estimation procedure leads one to conclude that the shadow rate was 2.2 percentage points lower in July 2012 than the estimates from their paper would indicate. My findings have important implications for policy makers who use the Wu-Xia shadow rate as an input to determining the effectiveness of unconventional monetary policy.

Each method has its advantages and disadvantages, but the discretization filter alleviates some of the more serious disadvantages of existing methods. I hope that economists working with nonlinear economic models will consider the discretization filter as a valuable addition to their toolkit.

References

Jerome Adda and Russell W. Cooper. *Dynamic Economics: Quantitative Methods and Applications*. MIT press, 2003.

Ravi Bansal and Amir Yaron. Risks for the long run: A potential resolution of asset pricing puzzles. *The Journal of Finance*, 59(4):1481–1509, 2004.

Michael D. Bauer and Glenn D. Rudebusch. Monetary Policy Expectations at the Zero Lower Bound. *Federal Reserve Bank of San Francisco Working Paper*, 18, 2013.

Zhe Chen. Bayesian Filtering: From Kalman Filters to Particle Filters, and Beyond. *Statistics*, 182(1):1–69, 2003.

Randal Douc, Eric Moulines, Tobias Ryden, et al. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *The Annals of Statistics*, 32(5):2254–2304, 2004.

Randal Douc, Eric Moulines, Jimmy Olsson, Ramon Van Handel, et al. Consistency of the maximum likelihood estimator for general hidden Markov models. *the Annals of Statistics*, 39(1):474–513, 2011.

Leland E Farmer and Alexis Akira Toda. Discretizing Stochastic Processes with Exact Conditional Moments. *Available at SSRN 2585859*, 2015.

Martin Flodén. A Note on the Accuracy of Markov-Chain Approximations to Highly Persistent AR(1)-Processes. *Economics Letters*, 99(3):516–520, 2008.

Thomas Flury and Neil Shephard. Bayesian Inference Based Only on Simulated Likelihood: Particle Filter Analysis of Dynamic Economic Models. *Econometric Theory*, 27(05):933–956, 2011.

Nikolay Gospodinov and Damba Lkhagvasuren. A Moment-Matching Method for Approximating Vector Autoregressive Processes by Finite-State Markov Chains. *Journal of Applied Econometrics*, 2013.

James D. Hamilton. A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica*, pages 357–384, 1989.

James D. Hamilton. *Time Series Analysis*, volume 2. Princeton University Press, Princeton NJ, 1994.

Nikolaus Hautsch and Yangguoyi Ou. Stochastic Volatility Estimation Using Markov Chain Simulation. In *Applied Quantitative Finance*, pages 249–274. Springer, 2008.

Kenneth L. Judd, Lilia Maliar, and Serguei Maliar. Merging Simulation and Projection Approaches to Solve High-Dimensional Problems. Technical report, National Bureau of Economic Research, 2012.

Sean P Meyn and Richard L Tweedie. *Markov Chains and Stochastic Stability*. Springer, 1993.

K. Geert Rouwenhorst. Asset Pricing Implications of Equilibrium Business Cycle Models. In *Frontiers of Business Cycle Research*, pages 294–330. Princeton University Press, 1995.

Ken'ichiro Tanaka and Alexis Akira Toda. Discretizing Distributions with Exact Moments: Error Estimate and Convergence Analysis. *SIAM Journal on Numerical Analysis*, 53(5): 2158–2177, 2015.

Ken'ichiro Tanaka and Alexis Akira Toda. Discrete approximations of continuous distributions by maximum entropy. *Economics Letters*, 118(3):445–450, 2013.

George Tauchen. Finite State Markov-Chain Approximations to Univariate and Vector Autoregressions. *Economics Letters*, 20(2):177–181, 1986.

George Tauchen and Robert Hussey. Quadrature-Based Methods for Obtaining Approximate Solutions to Nonlinear Asset Pricing Models. *Econometrica*, pages 371–396, 1991.

S.J. Taylor. Financial returns modelled by the product of two stochastic processes: a study of the daily sugar prices 1961-79. In *Time Series Analysis: Theory and Practice*, volume 1, pages 203–226. North-Holland, 1982.

Jing Cynthia Wu and Fan Dora Xia. Measuring the Macroeconomic Impact of Monetary Policy at the Zero Lower Bound. *Chicago Booth Research Paper*, (13-77), 2015.

A Proofs

Proof of Lemma 1. Note that by the Markov property, for $r < t \leq s$,

$$\bar{\mathbb{P}}_\theta \left(X_{t,m} \in A \left| \boldsymbol{X}_{r,m}^{t-1}, \boldsymbol{Y}_r^s \right. \right) = \bar{\mathbb{P}}_\theta \left(X_{t,m} \in A \left| X_{t-1,m}, \boldsymbol{Y}_{t-1}^s \right. \right)$$

Define the set of indices corresponding to the set A as $I_A \equiv \{ j \,|x_{j,m} \in A \}$. First consider the case where $t > s$.

$$\bar{\mathbb{P}}_\theta \left(X_{t,m} \in A \left| \boldsymbol{X}_{r,m}^{t-1}, \boldsymbol{Y}_r^s \right. \right) = \bar{\mathbb{P}}_\theta \left(X_{t,m} \in A \,| X_{t-1,m} = x \right) = \sum_{j \in I_A} Q_{\theta,m} \left(i, j \right)$$

Next consider the case where $t \leq s$,

$$\begin{aligned} &\bar{\mathbb{P}}_\theta \left(X_{t,m} \in A \,| X_{t-1,m} = x, \boldsymbol{Y}_r^s \right) \\ &= \sum_{j \in I_A} Q_{\theta,m} \left(i, j \right) \bar{p}_{\theta,m} \left(\boldsymbol{Y}_t^s \,| X_{t,m} = x_{j,m} \right) \times \left(\sum_{j=1}^{m} Q_{\theta,m} \left(i, j \right) \bar{p}_{\theta,m} \left(\boldsymbol{Y}_t^s \,| X_{t,m} = x_{j,m} \right) \right)^{-1} \end{aligned}$$

Define $Q_{\theta,m}^- \equiv \inf_{i,j} Q_{\theta,m}$ and $Q_{\theta,m}^+ \equiv \sup_{i,j} Q_{\theta,m}$. Since $Q_{\theta,m}^- \leq Q_{\theta,m} \left(i, j \right) \leq Q_{\theta,m}^+$ for all $i, j = 1, \ldots, m$, it follows that

$$\bar{\mathbb{P}}_\theta \left(X_{t,m} \in A \left| X_{t-1,m} = x, \boldsymbol{Y}_{t-1}^s \right. \right) \geq \frac{Q_{\theta,m}^-}{Q_{\theta,m}^+} \mu_{t,m} \left(\boldsymbol{Y}_t^s, A \right) \geq Q_+^- \mu_{t,m} \left(\boldsymbol{Y}_t^s, A \right)$$

where

$$\mu_{t,m} \left(\boldsymbol{Y}_t^s, A \right) \equiv \sum_{j \in I_A} \bar{p}_{\theta,m} \left(\boldsymbol{Y}_t^s \,| X_{t,m} = x_{j,m} \right) \times \left(\sum_{j=1}^{m} \bar{p}_{\theta,m} \left(\boldsymbol{Y}_t^s \,| X_{t,m} = x_{j,m} \right) \right)^{-1}$$

In the case $t > s$, it suffices to set $\mu_{t,m} \left(\boldsymbol{Y}_t^s, A \right) = \mu_{c,m} \left(A \right)$. □

Proof of Lemma 2. . Conditioning on a particular starting value $x_{0,m} \in \mathcal{X}_m$ is just a particular starting probability measure where probability 1 is assigned to that value. By Corollary 1, it follows that

$$\left\| \mathbb{P}_\theta \left(X_{t-1,m} \in \cdot \left| \boldsymbol{Y}_0^{t-1}, X_{0,m} = x_0 \right. \right) - \bar{\mathbb{P}}_\theta \left(X_{t-1,m} \in \cdot \left| \boldsymbol{Y}_0^{t-1} \right. \right) \right\|_{TV} \leq \rho^{t-1}$$

Thus, for $t \geq 1$,

$$
\begin{aligned}
&\left|p_{\theta,m}\left(Y_t \left| \boldsymbol{Y}_0^{t-1}, X_0 = x_0\right.\right) - \bar{p}_{\theta,m}\left(Y_t \left| \boldsymbol{Y}_0^{t-1}\right.\right)\right| \\
=&\left|\sum_{i=1}^{m}\sum_{j=1}^{m} Q_{\theta,m}\left(i,j\right) g_\theta\left(Y_t \left| x_{j,m}\right.\right) \times \left(\mathbb{P}_\theta\left(X_{t-1,m} = x_{i,m} \left| \boldsymbol{Y}_0^{t-1}, X_{0,m} = x_0\right.\right) - \bar{\mathbb{P}}_\theta\left(X_{t-1,m} = x_{i,m} \left| \boldsymbol{Y}_0^{t-1}\right.\right)\right)\right| \\
\leq&\ \rho^{t-1} \sup_i \sum_{j=1}^{m} Q_{\theta,m}\left(i,j\right) g_\theta\left(Y_t \left| x_{j,m}\right.\right) \\
\leq&\ \rho^{t-1} Q_{\theta,m}^{+} \sum_{j=1}^{m} g_\theta\left(Y_t \left| x_{j,m}\right.\right)
\end{aligned}
$$

In addition it holds that

$$
\begin{aligned}
p_{\theta,m}\left(Y_t \left| \boldsymbol{Y}_0^{t-1}, X_0 = x_0\right.\right) &= \sum_{i=1}^{m}\sum_{j=1}^{m} g_\theta\left(Y_t \left| x_{j,m}\right.\right) Q_{\theta,m}\left(i,j\right) \mathbb{P}_\theta\left(X_{t-1,m} = x_{i,m} \left| \boldsymbol{Y}_0^{t-1}, X_{0,m} = x_0\right.\right) \\
&\geq Q_{\theta,m}^{-} \sum_{i=1}^{m}\sum_{j=1}^{m} g_\theta\left(Y_t \left| x_{j,m}\right.\right) \mathbb{P}_\theta\left(X_{t-1,m} = x_{i,m} \left| \boldsymbol{Y}_0^{t-1}, X_{0,m} = x_0\right.\right) \\
&= Q_{\theta,m}^{-} \sum_{j=1}^{m} g_\theta\left(Y_t \left| x_{j,m}\right.\right)
\end{aligned}
$$

The same inequality also holds for $\bar{p}_{\theta,m}\left(Y_t \left| \boldsymbol{Y}_0^{t-1}\right.\right)$. It follows from the identity $\left|\log x - \log y\right| \leq \left|x - y\right| / \min\left(x, y\right)$ that

$$
\left|\log p_{\theta,m}\left(Y_t \left| \boldsymbol{Y}_t^{s}, X_{0,m} = x_0\right.\right) - \log \bar{p}_{\theta,m}\left(Y_t \left| \boldsymbol{Y}_t^{s}\right.\right)\right| \leq \frac{\rho^{t-1}}{1-\rho}
$$

By summing up the expression from $t = 1, \ldots, T$, we get

$$
\left|\ell_{T,m}\left(\theta, x_0\right) - \ell_{T,m}\left(\theta\right)\right| \leq \sum_{t=1}^{T} \frac{\rho^{t-1}}{1-\rho} = \frac{1-\rho^{T+1}}{\left(1-\rho\right)^2} \leq \frac{1}{\left(1-\rho\right)^2}
$$

Since this bound holds independently of θ and m, this concludes the proof. □

Proof of Lemma 3. Consider the first expression and let $r' \geq r$.

$$\begin{aligned}
&\bar{p}_{\theta,m}\left(Y_t \left| \boldsymbol{Y}_{-r}^{t-1}, X_{-r,m} = x\right.\right) - \bar{p}_{\theta,m}\left(Y_t \left| \boldsymbol{Y}_{-r'}^{t-1}, X_{-r',m} = x'\right.\right) \\
=&\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} g_\theta\left(Y_t \left| x_{k,m}\right.\right) Q_{\theta,m}\left(j,k\right) \bar{\mathbb{P}}\left(X_{t-1,m} = x_{j,m} \left| \boldsymbol{Y}_{-r}^{t-1}, X_{-r,m} = x_{i,m}\right.\right) \mathbb{1}\left\{x_{i,m} = x\right\} \\
-&\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} g_\theta\left(Y_t \left| x_{k,m}\right.\right) Q_{\theta,m}\left(j,k\right) \bar{\mathbb{P}}\left(X_{t-1,m} = x_{j,m} \left| \boldsymbol{Y}_{-r'}^{t-1}, X_{-r,m} = x_{i,m}\right.\right) \\
&\bar{\mathbb{P}}_\theta\left(X_{-r,m} = x_{i,m} \left| \boldsymbol{Y}_{-r'}^{t-1}, X_{-r',m} = x'\right.\right)
\end{aligned}$$

Thus, by Corollary 1

$$\begin{aligned}
&\left|\bar{p}_{\theta,m}\left(Y_t \left| \boldsymbol{Y}_{-r}^{t-1}, X_{-r,m} = x\right.\right) - \bar{p}_{\theta,m}\left(Y_t \left| \boldsymbol{Y}_{-r'}^{t-1}, X_{-r',m} = x'\right.\right)\right| \\
\leq\ & \rho^{t+r-1} Q_{\theta,m}^{+} \sum_{k=1}^{m} g_\theta\left(Y_t \left| x_{k,m}\right.\right)
\end{aligned}$$

Similary, I can obtain a lower bound on both expressions as in the proof of Lemma 2,

$$\begin{aligned}
&\bar{p}_{\theta,m}\left(Y_t \left| \boldsymbol{Y}_{-r}^{t-1}, X_{-r,m} = x\right.\right) \\
=&\sum_{i=1}^{m}\sum_{j=1}^{m} g_\theta\left(Y_t \left| x_{j,m}\right.\right) Q_{\theta,m}\left(i,j\right) \bar{\mathbb{P}}\left(X_{t-1,m} = x_{i,m} \left| \boldsymbol{Y}_{-r}^{t-1}, X_{-r,m} = x\right.\right) \\
\geq\ & Q_{\theta,m}^{-} \sum_{j=1}^{m} g_\theta\left(Y_t \left| x_{j,m}\right.\right)
\end{aligned}$$

Using the same inequality for logs applied in the proof of Lemma 2 we obtain the desired result. An analogous expression is obtained if $r' \leq r$. The second expression of the theorem follows from setting $r = r'$ and integrating with respect to $\bar{\mathbb{P}}_\theta\left(dx_{-r,m} \left| \boldsymbol{Y}_{-r}^{t-1}\right.\right)$.

Note that by Assumption (A3),

$$c_{-}\left(Y_t\right) \leq \bar{p}_{\theta,m}\left(Y_t \left| \boldsymbol{Y}_{-r}^{t-1}, X_{-r,m} = x\right.\right) \leq b_{+}$$

Taking logs concludes the proof.

□

Proof of Proposition 1. I proceed by providing a bound on the integration error

$$\left|\ell_m\left(\theta\right) - \ell\left(\theta\right)\right|$$

which is independent of $\theta \in \Theta$.

The Koksma-Hlawka inequality states that for any function f with bounded variation in

the sense of Hardy and Krause, then for any sequence $\{x_i\}_{i=1}^N$ with $x_i \in [0,1)^{d_X}$,

$$\left| \frac{1}{N} \sum_{i=1}^{N} f(x_i) - \int_{[0,1]^{d_X}} f(x)\, dx \right| \leq V(f)\, D_N^*(x_1, \ldots, x_N)$$

where an analogous version can be shown to hold for weighted integrals. D_N^* denotes the star-discrepancy and $V(\cdot)$ denotes the total variation.

Since $\pi_{\theta,m}^X$ can be thought of as an approximation of the true measure π_θ^X, using the Koksma-Hlawka inequality:

$$|\ell_m(\theta) - \ell(\theta)| \leq V(f)\, D_m^*\left(\pi_{\theta,m}^X\right)$$

where discrepancy in this instance is defined as

$$D_m^*\left(\pi_{\theta,m}^X\right) = \sup_{A \in J} \left| \pi_{\theta,m}^X(A) - \pi_\theta^X(A) \right|$$

where J is the set of d_X-dimensional boxes of the form

$$\prod_{i=1}^{d_X} [0, a_i) = \left\{ x \in \mathbb{R}^{d_X} : 0 \leq x_i < a_i \right\}$$

where $0 \leq a_i < 1$. Note that I originally defined $\pi_{\theta,m}^X$ using counting measure as the reference measure. To talk about the appropriate notion of convergence here I define the operation $\pi_{\theta,m}^X(A)$ where $A \in \mathcal{B}(\mathcal{X})$

$$\pi_{\theta,m}^X(A) = \sum_{i=1}^{m} \pi_{\theta,m}^X(i)\, \mathbb{1}\{x_{i,m} \in A\}$$

I similarly extend the definition of $Q_{\theta,m}$ to sets $A \in \mathcal{B}(\mathcal{X})$ and points $x \in \mathcal{X}$ as

$$Q_{\theta,m}(x, A) = \sum_{i=1}^{m} \sum_{j=1}^{m} Q_{\theta,m}(i,j)\, \mathbb{1}\{x \in A_{i,m}\}\, \mathbb{1}\{x_{j,m} \in A\}$$

Applying the triangle inequality twice to the expression of interest

$$
\begin{aligned}
&\sup_{A\in J}\left|\pi^X_{\theta,m}(A)-\pi^X_{\theta}(A)\right| = \sup_{x\in\mathcal{X}}\sup_{A\in J}\left|\pi^X_{\theta,m}(A)-\pi^X_{\theta}(A)\right| \\
=&\sup_{x\in\mathcal{X}}\sup_{A\in J}\left|\pi^X_{\theta,m}(A)-Q^T_{\theta,m}(x,A)+Q^T_{\theta,m}(x,A)-Q^T_{\theta}(x,A)+Q^T_{\theta}(x,A)-\pi^X_{\theta}(A)\right| \\
\leq&\sup_{x\in\mathcal{X}}\sup_{A\in J}\left|\pi^X_{\theta,m}(A)-Q^T_{\theta,m}(x,A)\right|+\sup_{x\in\mathcal{X}}\sup_{A\in J}\left|Q^T_{\theta,m}(x,A)-Q^T_{\theta}(x,A)\right|+\sup_{x\in\mathcal{X}}\sup_{A\in J}\left|\pi^X_{\theta}(A)-Q^T_{\theta}(x,A)\right| \\
\leq&\ \rho_1^T+\sup_{x\in\mathcal{X}}\sup_{A\in J}\left|Q^T_{\theta,m}(x,A)-Q^T_{\theta}(x,A)\right|+\rho_2^T \\
\leq&\ 2\rho_+^T+\sup_{x\in\mathcal{X}}\sup_{A\in J}\left|Q^T_{\theta,m}(x,A)-Q^T_{\theta}(x,A)\right|
\end{aligned}
$$

where ρ_1 and ρ_2 are the uniform minorizing constants (with respect to θ and m) for the Markov chains Π_θ and $\Pi_{\theta,m}$ respectively, and $\rho_+ = \max(\rho_1,\rho_2)$.

It remains to bound the second term. Through applications of the triangle inequality and the Cauchy-Schwarz inequality, we get that

$$
\begin{aligned}
&\sup_{x\in\mathcal{X}}\sup_{A\in J}\left|Q^T_{\theta,m}(x,A)-Q^T_{\theta}(x,A)\right| \\
=&\sup_{x\in\mathcal{X}}\sup_{A\in J}\left|Q_{\theta,m}Q^{T-1}_{\theta,m}(x,A)-Q_{\theta,m}Q^{T-1}_{\theta}(x,A)+Q_{\theta,m}Q^{T-1}_{\theta}(x,A)-Q_{\theta}Q^{T-1}_{\theta}(x,A)\right| \\
=&\sup_{x\in\mathcal{X}}\sup_{A\in J}\left|Q_{\theta,m}\left(Q^{T-1}_{\theta,m}(x,A)-Q^{T-1}_{\theta}(x,A)\right)-\left(Q_{\theta,m}-Q_{\theta}\right)Q^{T-1}_{\theta}(x,A)\right| \\
\leq&\ \sup_{x\in\mathcal{X}}\sup_{A\in J}\left|Q_{\theta,m}\left(Q^{T-1}_{\theta,m}(x,A)-Q^{T-1}_{\theta}(x,A)\right)\right|+\sup_{x\in\mathcal{X}}\sup_{A\in J}\left|\left(Q_{\theta,m}-Q_{\theta}\right)Q^{T-1}_{\theta}(x,A)\right| \\
\leq&\ \sup_{x\in\mathcal{X}}\sup_{A\in J}\left|Q_{\theta,m}(x,A)\right|\sup_{x\in\mathcal{X}}\sup_{A\in J}\left|Q^{T-1}_{\theta,m}(x,A)-Q^{T-1}_{\theta}(x,A)\right| \\
&+\sup_{x\in\mathcal{X}}\sup_{A\in J}\left|Q_{\theta,m}(x,A)-Q_{\theta}(x,A)\right|\sup_{x\in\mathcal{X}}\sup_{A\in J}\left|Q^{T-1}_{\theta}(x,A)\right| \\
\leq&\ \sup_{x\in\mathcal{X}}\sup_{A\in J}\left|Q^{T-1}_{\theta,m}(x,A)-Q^{T-1}_{\theta}(x,A)\right|+\sup_{x\in\mathcal{X}}\sup_{A\in J}\left|Q_{\theta,m}(x,A)-Q_{\theta}(x,A)\right|
\end{aligned}
$$

Applying this inequality recursively it is easy to see that

$$
\sup_{x\in\mathcal{X}}\sup_{A\in J}\left|Q^T_{\theta,m}(x,A)-Q^T_{\theta}(x,A)\right| \leq T\sup_{x\in\mathcal{X}}\sup_{A\in J}\left|Q_{\theta,m}(x,A)-Q_{\theta}(x,A)\right|
$$

As long as the Lebesgue measure of the discrete sets $\sup_i \lambda(A_{i,m}) \to 0$ as $m\to\infty$, the set of discrete points $\{x_{i,m}\}$ will become dense in $\mathcal{X}$. That is, for any $x\in\mathcal{X}$ and $\varepsilon>0$, $\exists m>0$ and $1\leq i\leq m$ s.t. $\|x-x_{i,m}\|<\varepsilon$. Thus the error in the expression above is bounded by the quality of approximation of the marginal distributions $Q_\theta(x,\cdot)$. Note that the specific rate achieved for the second term will depend on exactly how the Markov chain is discretized. As an example of the types of rates one might expect, consider the discretization method proposed in Farmer

and Toda (2015). In Tanaka and Toda (2015), it is shown that the discrete approximation will converge at a rate proportional to the underlying quadrature formula.

A simple example of quadrature formula is the trapezoidal rule, which is known to have integration error which is $O\left(\frac{1}{m^3}\right)$. The above expression can be thought of as a d_X-dimensional integral (d-dimensional state space), and thus if a simple product integration rule is used, we get

$$\sup_{A\in J}\left|\pi^X_{\theta,m}(A)-\pi^X_{\theta}(A)\right|\leq 2\rho_+^T+O\left(\frac{T}{m^{3/d}}\right)$$

This is equivalent to saying that $\exists M<\infty$ and $\exists c_m$ with $|c_m|<M$ for all m such that

$$\sup_{A\in J}\left|\pi^X_{\theta,m}(A)-\pi^X_{\theta}(A)\right|\leq 2\rho_+^{T_m}+c_m\frac{T_m}{m^{3/d}}$$

Let $\varepsilon>0$ and consider the sequence $T_m=m^{\varepsilon/d}$. Then

$$\begin{aligned}&2\rho_+^{T_m}+c_m\frac{T_m}{m^{3/d}}\\ =&2\rho_+^{m^{\varepsilon/d}}+c_m\frac{m^{\varepsilon/d}}{m^{3/d}}\\ =&2\rho_+^{m^{\varepsilon/d}}+c_m\frac{1}{m^{(3-\varepsilon)/d}}\end{aligned}$$

It is clear that the second term dominates asymptotically because it declines polynomially in m whereas the first term declines exponentially in m. This shows that

$$\sup_{A\in J}\left|\pi^X_{\theta,m}(A)-\pi^X_{\theta}(A)\right|=O\left(\frac{1}{m^{(3-\varepsilon)/d}}\right)$$

This implies that for any $\delta>\varepsilon$

$$\sup_{A\in J}\left|\pi^X_{\theta,m}(A)-\pi^X_{\theta}(A)\right|=o\left(\frac{1}{m^{(3-\delta)/d}}\right)$$

However since the choice of ε was arbitrary, we have that the above holds for any $\delta>0$.

For a general discretization method I will assume that a result of the form

$$\sup_{A\in J}\left|\pi^X_{\theta,m}(A)-\pi^X_{\theta}(A)\right|=o\left(h\left(m\right)\right)$$

can be established, which will again be independent of T for sensible discretization procedures. Note that the order of convergence of this integration error also does not depend on the value of the parameter $\theta\in\Theta$, which concludes the proof.

□

Proof of Corollary 4. Note that by Fubini's theorem:

$$\begin{aligned}\log \bar{p}_\theta\left(Y_0 \left| \boldsymbol{Y}_{-\infty}^{-1}\right.\right) &= \log \int_{\mathcal{Z}} \bar{p}_\theta\left(Y_0 \left| z_{-1}\right.\right) \pi_\theta\left(dz_{-1}\right) \\ &= \log \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} \bar{p}_\theta\left(Y_0 \left| z_{-1}\right.\right) g_\theta\left(y_{-1} \left| x_{-1}\right.\right) dy_{-1} \right) \pi_\theta^X\left(dx_{-1}\right)\end{aligned}$$

and

$$\begin{aligned}\log \bar{p}_{\theta,m}\left(Y_0 \left| \boldsymbol{Y}_{-\infty}^{-1}\right.\right) &= \log \int_{\mathcal{Z}_m} \bar{p}_{\theta,m}\left(Y_0 \left| z_{-1,m}\right.\right) \pi_{\theta,m}\left(dz_{-1,m}\right) \\ &= \log \sum_{i=1}^{m} \left(\int_{\mathcal{Y}} \bar{p}_{\theta,m}\left(Y_0 \left| y_{-1}, x_{-1} = x_{i,m}\right.\right) g_\theta\left(y_{-1} \left| x_{i,m}\right.\right) dy_{-1} \right) \pi_{\theta,m}^X\left(i\right)\end{aligned}$$

Thus, by the continuous mapping theorem, the dominated convergence theorem, and Proposition 1,

$$\lim_{m\to\infty} \sup_{\theta\in\Theta} \left|\ell_m\left(\theta\right) - \ell\left(\theta\right)\right| = 0, \quad \bar{\mathbb{P}}_{\theta^*}\text{-a.s.}$$

□

Proof of Lemma 4. To be completed. □

Proof of Proposition 2. Using the triangle inequality,

$$\begin{aligned}\sup_{\theta\in\Theta} \sup_{x_0\in\mathcal{X}} \left|T^{-1}\ell_{T,m}\left(\theta, x_0\right) - \ell\left(\theta\right)\right| &= \sup_{\theta\in\Theta} \sup_{x_0\in\mathcal{X}} \left|T^{-1}\ell_{T,m}\left(\theta, x_0\right) - \ell_m\left(\theta\right) + \ell_m\left(\theta\right) - \ell\left(\theta\right)\right| \\ &\leq \sup_{\theta\in\Theta} \sup_{x_0\in\mathcal{X}} \left|T^{-1}\ell_{T,m}\left(\theta, x_0\right) - \ell_m\left(\theta\right)\right| + \sup_{\theta\in\Theta} \left|\ell_m\left(\theta\right) - \ell\left(\theta\right)\right|\end{aligned}$$

The second term limits to 0 by Corollary 4. For the second term, note that by Lemma ?? it is sufficient to prove that

$$\limsup_{T\to\infty} \sup_{\theta\in\Theta} \sup_{m\in\mathbb{Z}^+} \left|T^{-1}\ell_{T,m}\left(\theta\right) - \ell_m\left(\theta\right)\right| = 0, \quad \bar{\mathbb{P}}_{\theta^*}\text{-a.s.}$$

Furthermore, since Θ is compact, this further reduces to proving that for all $\theta \in \Theta$,

$$\lim_{\delta\to 0} \limsup_{T\to\infty} \sup_{|\theta'-\theta|\leq\delta} \sup_{m\in\mathbb{Z}^+} \left|T^{-1}\ell_{T,m}\left(\theta'\right) - \ell_m\left(\theta\right)\right| = 0, \quad \bar{\mathbb{P}}_{\theta^*}\text{-a.s.}$$

This term can be further decomposed as

$$\begin{aligned}
&\limsup_{\delta\to 0}\limsup_{T\to\infty}\sup_{|\theta'-\theta|\leq\delta}\sup_{m\in\mathbb{Z}^+}\left|T^{-1}\ell_{T,m}\left(\theta'\right)-\ell_m\left(\theta\right)\right| \\
=&\limsup_{\delta\to 0}\limsup_{T\to\infty}\sup_{|\theta'-\theta|\leq\delta}\sup_{m\in\mathbb{Z}^+}\left|T^{-1}\ell_{T,m}\left(\theta'\right)-T^{-1}\ell_{T,m}\left(\theta\right)\right| \\
\leq& A+B+C
\end{aligned}$$

where

$$\begin{aligned}
A&=\limsup_{\delta\to 0}\limsup_{T\to\infty}\sup_{|\theta'-\theta|\leq\delta}\sup_{m\in\mathbb{Z}^+}T^{-1}\sum_{t=1}^{T}\left|\Delta_{t,0,m}\left(\theta'\right)-\Delta_{t,\infty,m}\left(\theta'\right)\right|, \\
B&=\limsup_{\delta\to 0}\limsup_{T\to\infty}\sup_{|\theta'-\theta|\leq\delta}\sup_{m\in\mathbb{Z}^+}T^{-1}\sum_{t=1}^{T}\left|\Delta_{t,\infty,m}\left(\theta'\right)-\Delta_{t,\infty,m}\left(\theta\right)\right|, \\
C&=\limsup_{T\to\infty}\sup_{m\in\mathbb{Z}^+}T^{-1}\sum_{t=1}^{T}\left|\Delta_{t,\infty,m}\left(\theta\right)-\Delta_{t,0,m}\left(\theta\right)\right|
\end{aligned}$$

Terms A and C are zero by Corollary 2, and by Lemma 4 and the ergodic theorem,

$$\begin{aligned}
B&\leq\limsup_{\delta\to 0}\limsup_{T\to\infty}T^{-1}\sum_{t=1}^{T}\sup_{|\theta'-\theta|\leq\delta}\sup_{m\in\mathbb{Z}^+}\left|\Delta_{t,\infty,m}\left(\theta'\right)-\Delta_{t,\infty,m}\left(\theta\right)\right| \\
&=\limsup_{\delta\to 0}\overline{\mathbb{E}}_{\theta^*}\left[\sup_{|\theta'-\theta|\leq\delta}\sup_{m\in\mathbb{Z}^+}\left|\Delta_{t,\infty,m}\left(\theta'\right)-\Delta_{t,\infty,m}\left(\theta\right)\right|\right] \\
&=0,\quad \overline{\mathbb{P}}_{\theta^*}\text{-a.s.}
\end{aligned}$$

□

Proof of Theorem 3. . In order to establish asymptotic normality of my proposed estimator, it is sufficient to show that $\ell_T\left(\hat{\theta}_{T,x_0},x_0\right)-\ell_T\left(\hat{\theta}_{T,m,x_0},x_0\right)=o_P\left(1\right)$ by Theorem 7 of Douc et al. (2004). Rewriting this term

$$\begin{aligned}
&\ell_T\left(\hat{\theta}_{T,x_0},x_0\right)-\ell_T\left(\hat{\theta}_{T,m,x_0},x_0\right) \\
=&\ell_T\left(\hat{\theta}_{T,x_0},x_0\right)-\ell_T\left(\hat{\theta}_{T,m,x_0},x_0\right)+\ell_{T,m}\left(\hat{\theta}_{T,x_0},x_0\right)-\ell_{T,m}\left(\hat{\theta}_{T,x_0},x_0\right) \\
\leq&\ell_T\left(\hat{\theta}_{T,x_0},x_0\right)-\ell_{T,m}\left(\hat{\theta}_{T,x_0},x_0\right)+\ell_{T,m}\left(\hat{\theta}_{T,m,x_0},x_0\right)-\ell_T\left(\hat{\theta}_{T,m,x_0},x_0\right)
\end{aligned}$$

Note that it is thus sufficient to show that for any $\theta\in\Theta$, $\mathbb{P}_\theta\left(\left|\ell_{T,m}\left(\theta,x_{0,m}\right)-\ell_T\left(\theta,x_0\right)\right|\geq\varepsilon\right)\to 0$ as $T\to\infty$, $m\to\infty$ at appropriate rates. It is possible to decompose this probability as fol-

lows:

$$\begin{aligned}
&\mathbb{P}_\theta\left(\left|\ell_{T,m}\left(\theta, x_{0,m}\right) - \ell_T\left(\theta, x_0\right)\right| \geq \varepsilon\right) \\
=&\mathbb{P}_\theta\left(\left|\ell_{T,m}\left(\theta, x_{0,m}\right) - \ell_T\left(\theta, x_0\right) + T\ell\left(\theta\right) - T\ell\left(\theta\right) + T\ell_m\left(\theta\right) - T\ell_m\left(\theta\right)\right| \geq \varepsilon\right) \\
\leq\ &\mathbb{P}_\theta\left(\left|\ell_T\left(\theta, x_0\right) - T\ell\left(\theta\right)\right| \geq \frac{\varepsilon}{3}\right) + \mathbb{P}_\theta\left(\left|\ell_{T,m}\left(\theta, x_{0,m}\right) - T\ell_m\left(\theta\right)\right| \geq \frac{\varepsilon}{3}\right) + \mathbb{P}_\theta\left(T\left|\ell_m\left(\theta\right) - \ell\left(\theta\right)\right| \geq \frac{\varepsilon}{3}\right)
\end{aligned}$$

Theorem 14 from Douc et al. (2011) states that for any V_θ-uniformly ergodic state process with transition kernel Q_θ, $f : \mathcal{Y}^{s+1}$ with $\|f\|_\infty < \infty$, there exists a constant $K < \infty$ such that

$$\mathbb{P}_\theta^\nu\left(\left|\sum_{t=1}^{T}\left\{f\left(\boldsymbol{Y}_t^{t+s}\right) - \bar{\mathbb{E}}_{\theta^*}\left[f\left(\boldsymbol{Y}_0^s\right)\right]\right\}\right| \geq \varepsilon\right) \leq K\nu\left(V\right)\exp\left[-\frac{1}{K}\left(\min\left(\frac{\varepsilon^2}{T}, \varepsilon\right)\right)\right]$$

for any initial probability measure ν and $\varepsilon > 0$. Both the original chain Q_θ and each discrete chain $Q_{\theta,m}$ are uniformly ergodic and thus V_θ-uniformly ergodic for $V_\theta = 1$.

Note that the first two terms are of the form considered in Theorem 14 from Douc et al. (2011). I explicity show the bound for the first term and the second term is analogous due to the uniform minorization of the sequence of discrete Markov chains for all $m \in \mathbb{Z}^+$ with the same minorizing constant

$$\begin{aligned}
&\mathbb{P}_\theta\left(\left|\ell_T\left(\theta, x_0\right) - T\ell\left(\theta\right)\right| \geq \frac{\varepsilon}{3}\right) \\
=&\mathbb{P}_\theta\left(\left|\sum_{t=1}^{T}\left\{\log p_\theta\left(Y_t\left|\boldsymbol{Y}_0^{t-1}, X_0 = x_0\right.\right) - \ell\left(\theta\right)\right\}\right| \geq \frac{\varepsilon}{3}\right) \\
\leq\ &K\exp\left[-\frac{1}{K}\left(\min\left(\frac{\varepsilon^2}{9T}, \frac{\varepsilon}{3}\right)\right)\right] = o_P\left(1\right) \quad \text{with } P = \bar{\mathbb{P}}_{\theta^*}
\end{aligned}$$

It follows from proposition 2 that

$$\left|\ell_m\left(\theta\right) - \ell\left(\theta\right)\right| = o\left(h\left(m\right)\right)$$

and thus

$$T\left|\ell_m\left(\theta\right) - \ell\left(\theta\right)\right| = o\left(Th\left(m\right)\right)$$

Returning to the original expression of interest

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\left|\ell_T\left(\hat{\theta}_{T,x_0}, x_0\right) - \ell_T\left(\hat{\theta}_{T,m,x_0}, x_0\right)\right| \geq \varepsilon\right) \\
\leq\ &\mathbb{P}_{\theta^*}\left(\left|\ell_T\left(\hat{\theta}_{T,x_0}, x_0\right) - \ell_{T,m}\left(\hat{\theta}_{T,x_0}, x_0\right) + \ell_{T,m}\left(\hat{\theta}_{T,m,x_0}, x_0\right) - \ell_T\left(\hat{\theta}_{T,m,x_0}, x_0\right)\right| \geq \varepsilon\right) \\
\leq\ &\mathbb{P}_{\theta^*}\left(\left|\ell_{T,m}\left(\hat{\theta}_{T,x_0}, x_{0,m}\right) - \ell_T\left(\hat{\theta}_{T,x_0}, x_0\right)\right| \geq \frac{\varepsilon}{2}\right) + \mathbb{P}_{\theta^*}\left(\left|\ell_{T,m}\left(\hat{\theta}_{T,m,x_0}, x_{0,m}\right) - \ell_T\left(\hat{\theta}_{T,m,x_0}, x_0\right)\right| \geq \frac{\varepsilon}{2}\right) \to 0
\end{aligned}$$

for $T \to \infty$, $m \to \infty$, $Th\,(m) \to 0$. This ensures that my proposed estimator satisfies condition (iii) of Theorem 7 from Douc et al. (2004). □

B Figures and Tables

		DF 5	DF 9	DF 13	PF 1000	EKF
$\boldsymbol{\mu}$	**T = 100**	0.4469	0.4539	0.4423	0.4877	0.5211
	T = 500	0.4256	0.4164	0.3795	0.5735	0.4446
	T = 1,000	0.3472	0.3492	0.3564	0.6139	0.3359
$\boldsymbol{\rho}$	**T = 100**	0.5882	0.5910	0.5981	0.5721	0.7269
	T = 500	0.1082	0.0987	0.1147	0.1079	0.3042
	T = 1,000	0.0189	0.0218	0.0191	0.0420	0.1260
$\boldsymbol{\sigma}$	**T = 100**	0.2345	0.2318	0.2344	0.2439	0.6214
	T = 500	0.0770	0.0667	0.0704	0.1114	0.2933
	T = 1,000	0.0433	0.0381	0.0398	0.0761	0.1629

Table 1: RMSE of parameter estimates

		DF 5	DF 9	DF 13	PF 1000	EKF
$\boldsymbol{\mu}$	**T = 100**	0.3575	0.3578	0.3539	0.3846	0.4126
	T = 500	0.3263	0.3148	0.2949	0.3841	0.3395
	T = 1,000	0.2653	0.2516	0.2404	0.3597	0.2607
$\boldsymbol{\rho}$	**T = 100**	0.4432	0.4466	0.4513	0.4418	0.6171
	T = 500	0.0373	0.0340	0.0376	0.0497	0.1371
	T = 1,000	0.0117	0.0115	0.0108	0.0193	0.0350
$\boldsymbol{\sigma}$	**T = 100**	0.1725	0.1702	0.1717	0.1781	0.4292
	T = 500	0.0500	0.0431	0.0431	0.0711	0.1948
	T = 1,000	0.0306	0.0263	0.0251	0.0453	0.1260

Table 2: MAE of parameter estimates

		DF 5	DF 9	DF 13	PF 1000	EKF
$\boldsymbol{\mu}$	**T = 100**	-0.0348	-0.0245	-0.0306	-0.0278	0.0046
	T = 500	-0.0124	-0.0169	-0.0070	0.0028	-0.0013
	T = 1,000	0.0137	-0.0002	0.0030	0.0163	0.0196
$\boldsymbol{\rho}$	**T = 100**	-0.4428	-0.4458	-0.4504	-0.4417	-0.6170
	T = 500	-0.0362	-0.0316	-0.0351	-0.0484	-0.1359
	T = 1,000	-0.0108	-0.0093	-0.0085	-0.0172	-0.0335
$\boldsymbol{\sigma}$	**T = 100**	0.1155	0.1099	0.1133	0.1338	0.3402
	T = 500	0.0390	0.0307	0.0305	0.0620	0.1860
	T = 1,000	0.0220	0.0153	0.0135	0.0354	0.1250

Table 3: Bias of parameter estimates

	DF 5	DF 9	DF 13	PF 1000	EKF
T = 100	0.62	0.72	0.83	19.65	0.17
T = 500	2.01	2.12	2.29	269.75	0.22
T = 1,000	3.71	3.92	4.21	1,088.52	0.30

Table 4: Average time of one maximum likelihood estimation (in seconds)

	DF 5	DF 9	DF 13	PF 1000	EKF
T = 100	0.3753	0.3753	0.3728	0.3742	0.4977
T = 500	0.3788	0.3759	0.3751	0.3830	0.4652
T = 1,000	0.3797	0.3769	0.3762	0.3828	0.4518

Table 5: ARMSE of filtered state estimates at $\hat{\theta}_{MLE}$

	DF 5	DF 9	DF 13	PF 1000	EKF
T = 100	0.3070	0.3073	0.3054	0.3065	0.3900
T = 500	0.3027	0.3002	0.3000	0.3062	0.3715
T = 1,000	0.3032	0.3008	0.3003	0.3054	0.3610

Table 6: AMAE of filtered state estimates at $\hat{\theta}_{MLE}$

	Discretization Filter			**Extended Kalman Filter**		
1200μ	-0.2251	-0.2061	0.0256	-0.3035	-0.2381	0.0253
	(0.0767)	(1.3693)	(0.0318)	(0.1885)	(0.1815)	(0.0160)
ρ	0.9648	0.0056	0.4541	0.9638	-0.0026	0.3445
	(0.0100)	(0.0212)	(1.5863)	(0.0199)	(0.0183)	(0.4821)
	-0.0234	0.9626	0.8170	-0.0226	0.9420	1.0152
	(0.1809)	(0.0762)	(5.2167)	(0.0202)	(0.0212)	(0.5111)
	0.0046	0.0035	0.7750	0.0033	0.0028	0.8869
	(0.0023)	(0.0067)	(0.1050)	(0.0018)	(0.0019)	(0.0385)
$eig\,(\rho)$	0.9765+0.006i	0.9765-0.006i	0.7513	0.9832	0.9642	0.8452
$\rho^{\mathbb{Q}}$	0.9983	0	0	0.9978	0	0
	(0.0026)			(0.0003)		
	0	0.9608	1	0	0.9502	1
		(0.0121)			(0.0012)	
	0	0	0.9608	0	0	0.9502
			(0.0121)			(0.0012)
$1200\delta_0$	13.2418			13.3750		
	(2.3324)			(1.0551)		
1200Σ	0.2511			0.4160		
	(0.3467)			(0.0390)		
	-0.0535	0.2541		-0.3999	0.2445	
	(0.3483)	(0.1978)		(0.0369)	(0.0233)	
	-0.0002	0.0026	0.0338	-0.0110	0.0033	0.0390
	(0.0026)	(0.0058)	(0.0095)	(0.0069)	(0.0034)	(0.0030)
$1200\sqrt{\omega}$	0.1638			0.0893		
	(0.0403)			(0.0027)		

Table 7: Maximum likelihood parameter estimates (QMLE standard errors in parantheses)

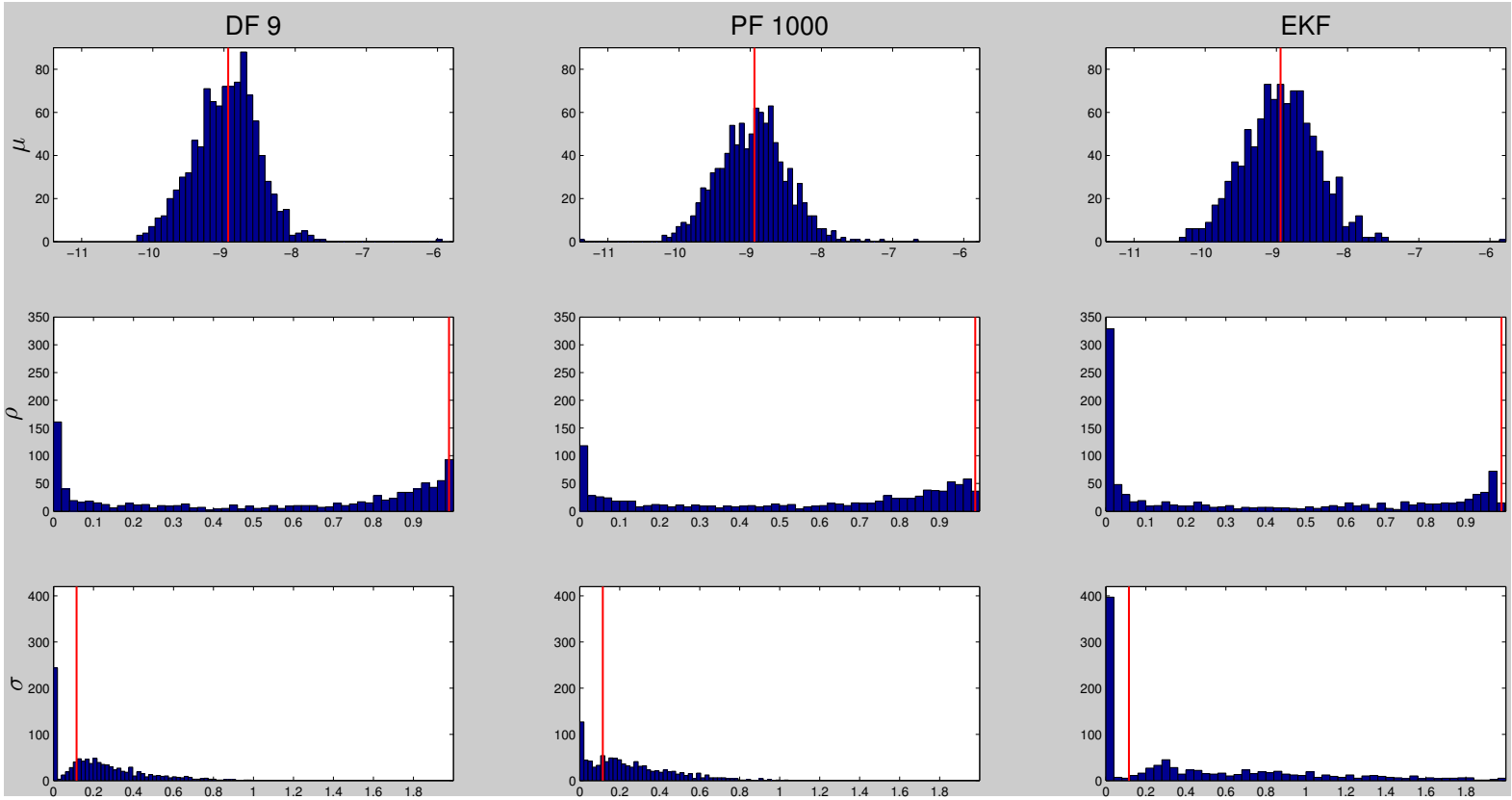


Figure 1: T = 100 MLE sampling distributions

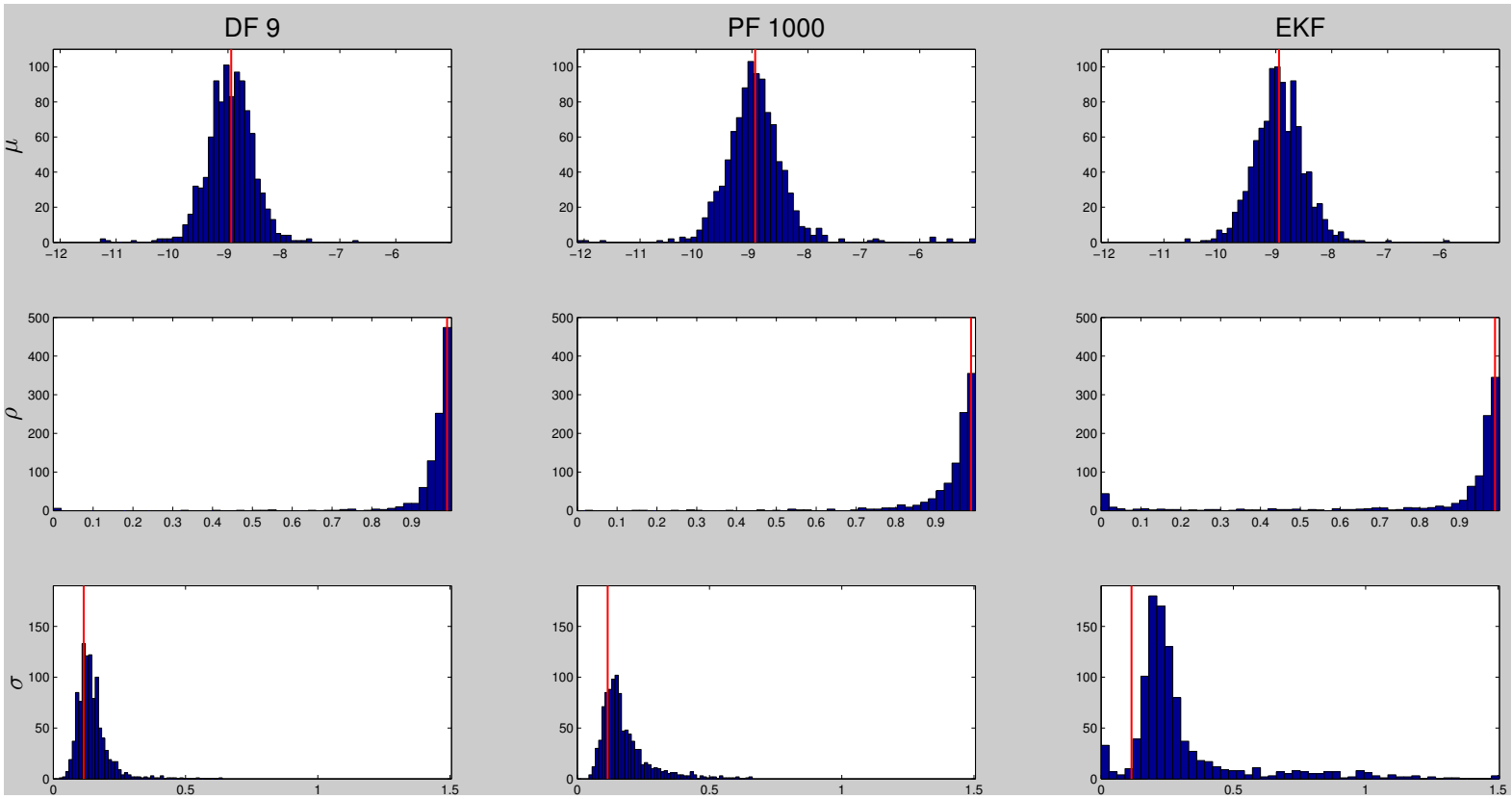


Figure 2: T = 500 MLE sampling distributions

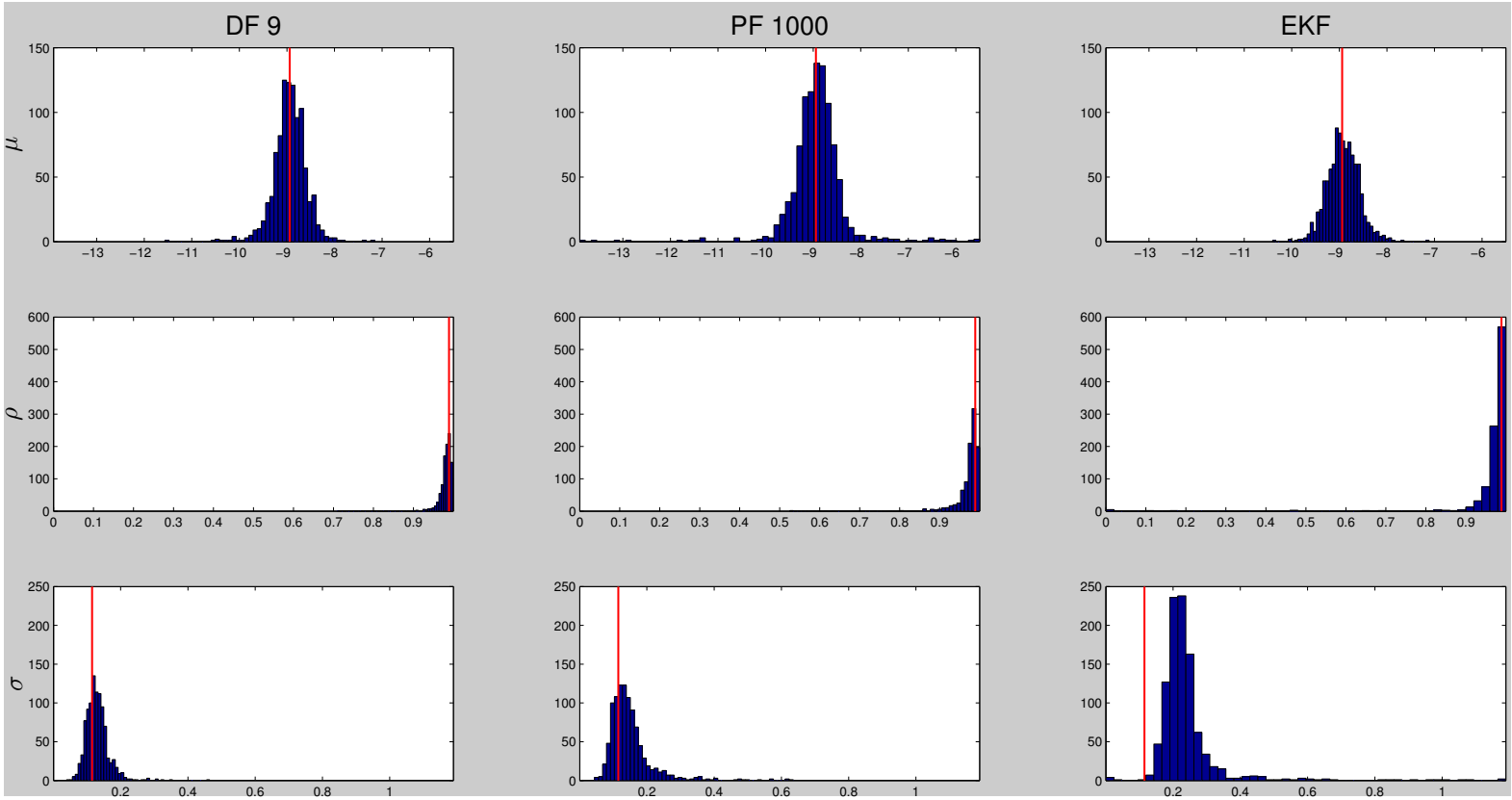


Figure 3: T = 1,000 MLE sampling distributions

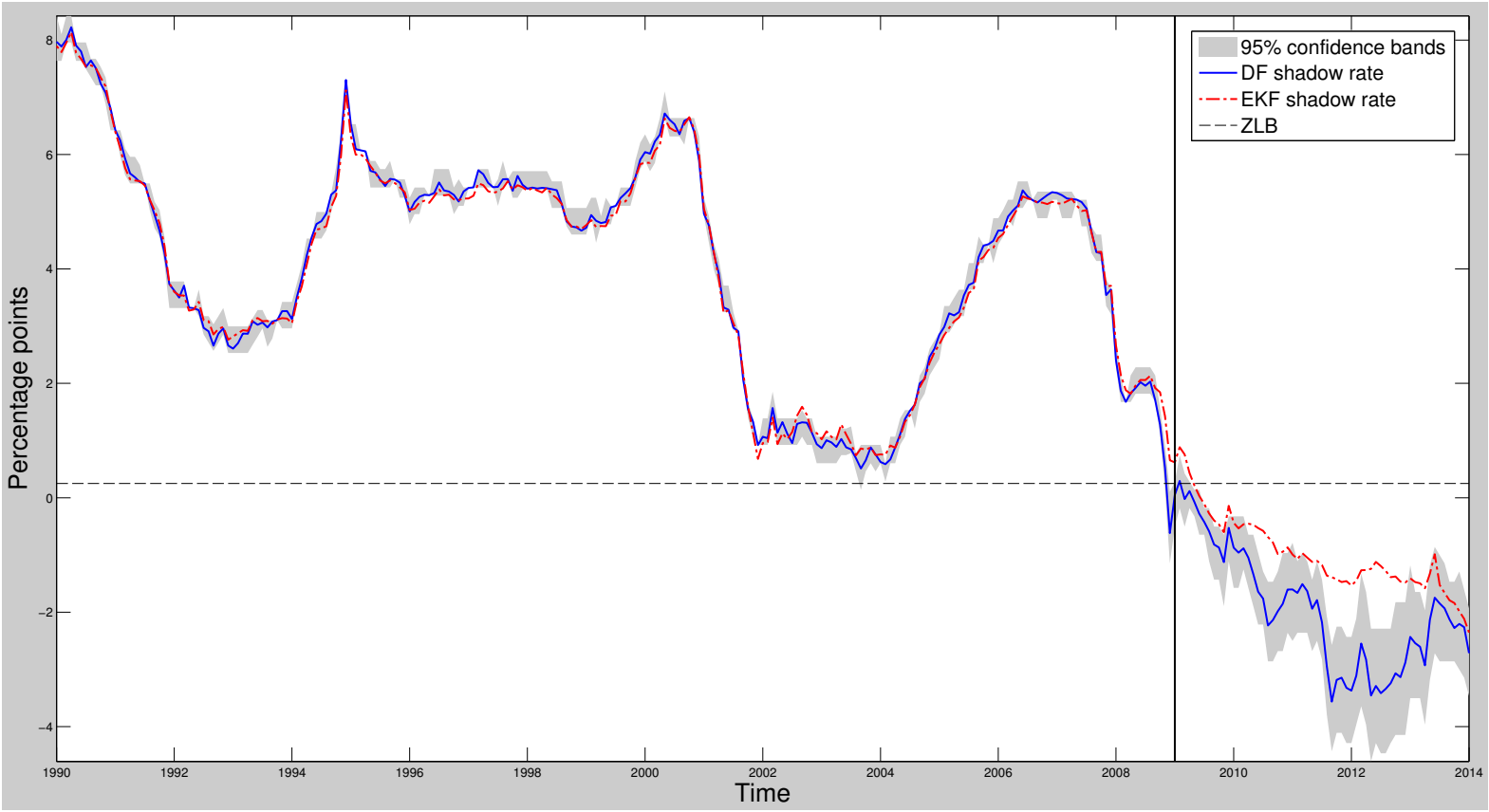


Figure 4: Shadow rates

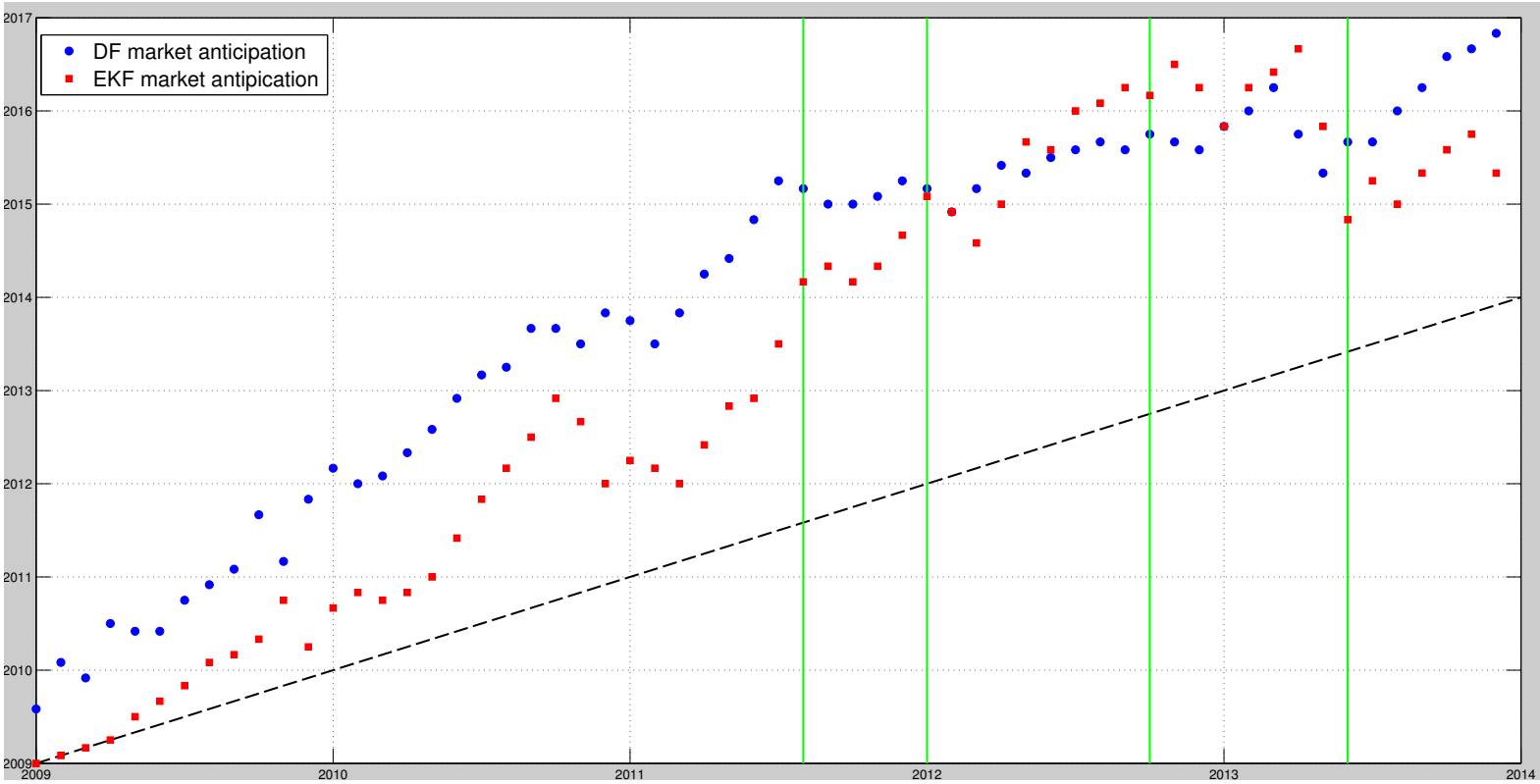


Figure 5: Expected liftoff dates