Finite-State Markov-Chain Approximations: A Hidden Markov Approach

Eva F. Janssens[†] and Sean McCrary[‡]

December 15, 2024

Abstract

This paper proposes a novel finite-state Markov chain approximation method for Markov processes with continuous support, providing both an optimal grid and transition probability matrix. The method can be used for multivariate processes, as well as processes with time-varying components. The method is based on minimizing the information loss between a Hidden Markov Model and the true data-generating process. We provide sufficient conditions under which this information loss can be made arbitrarily small if enough grid points are used. We compare our method to existing methods through the lens of an asset-pricing model, and a life-cycle consumption-savings model. We find our method can lead to more parsimonious discretizations and more accurate solutions, and the discretization matters for the welfare costs of risk, the marginal propensities to consume, and the amount of wealth inequality a life-cycle model can generate.

Keywords: Numerical methods, Kullback–Leibler divergence, misspecified model, earnings process

JEL classification codes: C63, C68, D15, E21

* Acknowledgements: Both authors are grateful for the invaluable comments from José-Víctor Ríos-Rull, Frank Kleibergen, Christian Stoltenberg, Robin Lumsdaine, as well as seminar and conference participants at the University of Zürich, University of Amsterdam, University of Michigan, University of Houston, University of Oxford, Federal Reserve Board, EEA/ESEM 2022, CEF 2022, CFE 2022, and the NBER SI 2022, especially Jesús Fernández-Villaverde, Frank Schorfheide, Borağan Aruoba, Mikkel Plagborg-Møller, Leland Farmer, Michael Wolf, Damian Kozbur, Florian Gunsilius, Elisabeth Proehl, and many others. Janssens is grateful to the Dutch Research Council for the NWO Research Talent Grant, project number 406.18.514 and to the Erasmus Trustfonds for the Professor Bruins Prize 2018, funding the research visit to University of Pennsylvania during which this paper was written, as well as to Frank Schorfheide for hosting this visit. We thank the Society of Computational Economics for the CEF 2022 Student Prize. Any errors are our own.

Contact information:
† Eva F. Janssens: University of Michigan, e-mail: evafj@umich.edu
‡ Sean McCrary: Ohio State University, e-mail: mccrary.65@osu.edu

1 Introduction

Numerical methods for solving nonlinear dynamic stochastic models often rely on finite-state Markov chain approximations of continuous stochastic processes. The stochastic process is an important input for these models, and its finite-state Markov chain approximation should closely resemble the continuous process. The modern econometrics literature has found that many driving processes exhibit nonlinear and non-Gaussian features, and discretizations that preserve these characteristics enable researchers to incorporate such processes into their models to evaluate their implications for economic quantities of interest. This paper proposes a novel full-information discretization method applicable to a wide variety of Markov processes.

Approximating a continuous stochastic process with a discrete Markov process, characterized by a grid of support points and a transition probability matrix, inherently involves selecting a misspecified approximating model. In line with the misspecified model literature, we propose a finite-state Markov chain approximation method that minimizes the information loss between the misspecified process and the true continuous process. Specifically, we assume that the misspecified process is a Hidden Markov Model (HMM), which embeds a discrete Markov chain into a continuous support process via a continuous measurement error, and enables the use of the Kullback-Leibler (KL) divergence between the two processes as a measure of information loss.

As a result, the implementation of our method is straightforward, as minimizing the KL divergence in this context is equivalent to quasi-maximum likelihood estimation. To do so, one simulates data from the true process and estimates the parameters of the HMM using the Expectation-Maximization algorithm (e.g., Rabiner, 1989). An attractive feature of this approach is that it yields both an optimal grid and transition probability matrix. Moreover, the method can be applied to multivariate processes, where the optimal grid mitigates the curse of dimensionality associated with tensor grids by accounting for dependencies between variables.

Our theoretical contribution is to prove that, under certain assumptions, as the number of unobserved states (and thus grid points) increases, the information loss between the misspecified HMM and the true continuous stochastic process becomes arbitrarily small. This relates to the literature on universal approximators, where we build on the results of Nguyen, Nguyen, Chamroukhi, and McLachlan (2020) regarding (Gaussian) mixtures and extend these to the non-i.i.d. setting of HMMs. Our proof provides insights into the properties of the process that determine how many grid points are needed to achieve a specific level of information loss.

For instance, more persistent processes require a larger number of grid points, which explains why finite-state Markov chain approximations of highly persistent processes tend to be less accurate.[1]

Different discretization methods are typically designed to accurately capture specific features of the true process. For example, moment-matching methods will attempt to match targeted moments, which might be appealing if those were the moments used to estimate the stochastic process. Ultimately, what matters is how the discretization affects the accuracy of model solutions and the endogenous economic quantities of interest. Therefore, we evaluate the performance of our method against existing approaches in two economic applications: an asset pricing model and a life-cycle consumption-saving model.

First, in the asset pricing model, we discretize dividend growth, which is assumed to follow an autoregressive process with stochastic volatility. This model has the appealing feature of admitting a closed-form solution, as shown by De Groot (2015). We use this solution as a benchmark to compare the performance of our method with existing approaches in the literature. Our findings show that our discretization reliably converges to the benchmark model solution as the number of grid points increases, and, because of the optimal placement of grid points, can be more parsimonious than existing methods.

Second, we evaluate the performance of our method within the framework of a life-cycle consumption-saving model. In this application, we focus on two earnings processes featuring life-cycle dependence: the process proposed by Guvenen, Karahan, Ozkan, and Song (2021), which incorporates non-employment shocks and earnings shocks with positive skewness, and the non-parametric process developed by Arellano, Blundell, and Bonhomme (2017).[2] In this application, we compare our method to binning methods based on Adda and Cooper (2003).[3]

HMM discretizations effectively reproduce the excess skewness and kurtosis in the processes of Guvenen et al. (2021) and Arellano et al. (2017) with fewer grid points than binning, resulting in more parsimonious model solutions. In contrast, binning with the same number of grid points underestimates these moments, leading to an underestimation of the welfare cost of risk and wealth inequality. Although life-cycle models often struggle to replicate the wealth distribution observed in the data (De Nardi and Fella, 2017), our accurate discretization of

[1] As discussed in/shown by Flodén (2008), Kopecky and Suen (2010) and Galindev and Lkhagvasuren (2010).

[2] These processes are regarded as being at the frontier of the earnings dynamics literature (Altonji, Hynsjö, and Vidangos, 2022).

[3] The life-cycle model does not admit a closed-form solution. Therefore, as a benchmark, we solve the model with a large number of gridpoints.

the Arellano et al. (2017) process generates a wealth Gini of 0.79 and top 1% wealth shares, closely aligned with U.S. data. We also compare the life-cycle implications of the two stochastic processes and observe significant differences. For instance, the Guvenen et al. (2021) process, relative to the Arellano et al. (2017) process, implies a higher welfare cost of risk (37% versus 19% of lifetime consumption), less wealth inequality (top 1% wealth share of 11.5% versus 35.7%), and a higher average marginal propensity to consume (21% versus 15%). Our findings on the importance of discretization methods for life-cycle model solutions also extend to mixture AR(1) processes, where HMM discretization achieves accurate solutions with few grid points.

The paper proceeds as follows. The next subsection discusses the related literature. Section 2 discusses our discretization method and theoretical contributions. Section 3 presents the asset pricing model with stochastic volatility. Section 4 discusses the life-cycle model applications. Section 5 concludes. The Supplementary Appendix provides the proof of our Main Theorem, details on the asset pricing model, and provides an additional application to the discretization of vector autoregressive processes.

Related literature. Several methods have been proposed for the discretization of stochastic processes. Most of these methods, including Tauchen (1986), Rouwenhorst (1995), Tauchen and Hussey (1991), Duan and Simonato (2001), Terry and Knotek II (2011), and Gospodinov and Lkhagvasuren (2014), are tailored to specific (linear) processes, such as AR(1) or VAR processes. For example, Fella, Gallipoli, and Pan (2019) adapt the methods of Rouwenhorst (1995), Tauchen and Hussey (1991), and Adda and Cooper (2003) to processes with a life-cycle component and examine their performance in settings where innovations are drawn from a mixture of normals. Similarly, Galindev and Lkhagvasuren (2010) modify Rouwenhorst (1995) to accommodate highly persistent, correlated AR(1) shocks, while Civale, Díez-Catalán, and Fazilet (2016) extend the Tauchen (1986) method to autoregressive processes with innovations from a normal mixture. Unlike these approaches, our method is applicable to a broader class of stochastic processes and simultaneously provides both an optimal grid and a transition probability matrix. In contrast, existing methods typically require a predefined grid as input and often assume equally spaced or equal-quantile grids.

Some discretization methods are applicable to a broader class of stochastic processes, such as binning methods described in Judd (1998) and Adda and Cooper (2003). However, binning methods only match one-step-ahead transitions between bins and require the grid spacing to be specified as an input. In contrast, our discretization method accounts for the full dynamics of the process and generates an optimal grid. Farmer and Toda (2017) propose a method to refine an initial discrete approximation using moment matching. By comparison, our method can be viewed as a full-information discretization approach, rather than a moment-matching method.

For multivariate processes, most existing methods rely on tensor grids, which suffer from the curse of dimensionality and are computationally inefficient. As noted by Gordon (2021), tensor grids are suboptimal because many of the grid points are rarely visited. To address this inefficiency, Gordon (2021) propose the use of pruning and sparse grids for VAR models. In contrast, our method produces optimal grids that limit the curse-of-dimensionality issue when the variables are correlated and is applicable to any type of process.

Our results contribute to the literature on misspecified models (Gourieroux, Monfort, and Trognon, 1984; White, 1982), and, more specifically, to the study of misspecified Hidden Markov Models (HMMs) (Douc and Moulines, 2012; Mevel and Finesso, 2004). The use of HMMs is prevalent in both economics and machine learning.[4] To the best of our knowledge, the application of HMMs as a finite-state Markov chain approximation method for continuous stochastic processes is novel, as is our theoretical result on the ability of HMMs to approximate such processes.[5] In the signal processing literature, Vidyasagar (2005), Finesso, Grassi, and Spreij (2010), and others address the problem of representing discrete state-space stationary processes as HMMs. However, their results do not extend to continuous stochastic processes.

2 Discretization using a Hidden Markov Model

Let $y_{it} \in \mathbb{R}^k$, $i = 1, ..., N$, $t = 1, ..., T$, denote a random variable for which the data generating process is a discrete-time continuous-support Markov process. Let $f(y_{it}|y_{it-1})$ denote its conditional probability density function, and denote the joint probability distribution of the full sequence $(y_{i1}, ..., y_{iT})$ by $f(\mathbf{y}_i)$. The index i denotes different sequences generated from the same stochastic process which will be useful when discussing the implementation of the algorithm for life-cycle processes with time-varying parameters. However, when we assume the data are generated by a stationary process, to ease notation, we suppress this index. The objective is to approximate the distribution of $\mathbf{y}$ by a misspecified model, with probability

[4] In the statistics and machine-learning literature, HMMs are often interpreted as a dimension reduction method for dependent data (McLachlan, Lee, and Rathnayake, 2019), with common applications including text processing. In econometrics, HMMs have been applied to detect structural breaks (Song, 2014) and model regime switches (beginning with Quandt (1958), Goldfeld and Quandt (1973), and Hamilton (1990)). Additionally, HMMs have been employed to approximate the dynamics of latent states in non-linear state-space models for estimation purposes, as in Kitagawa (1987), Langrock (2011), and Farmer (2021).

[5] This intuition is mentioned by Lehéricy (2021) but not formally proven.

distribution $p(\mathbf{y};\theta)$, by choosing parameter vector θ such that the relative entropy, also known as the information loss, between the approximating distribution and the true distribution is minimized. Minimizing information loss, which can be measured through the Kullback-Leibler (KL) divergence, is a common way to think about misspecified models and their consistency.

More precisely, let the relative entropy be defined as the logarithmic difference between the distributions $f(\mathbf{y})$ and $p(\mathbf{y};\theta)$, where the expectation is taken using the distribution $f(\mathbf{y})$, also known as the Kullback–Leibler (KL) divergence:

$$D^{KL}(f(\mathbf{y})||p(\mathbf{y};\theta)) = \int f(\mathbf{y}) \log \frac{f(\mathbf{y})}{p(\mathbf{y};\theta)} d\mathbf{y}, \tag{1}$$

Minimizing the KL divergence with respect to parameter vector θ requires taking the derivative of Equation (1) with respect to θ:

$$\begin{aligned} &\int \nabla_\theta \log p(\mathbf{y};\theta) f(\mathbf{y}) dy = 0 \\ \Leftrightarrow\ &\mathbb{E}_f \left[\nabla_\theta \log\left(p(\mathbf{y};\theta)\right)\right] = 0. \end{aligned}$$

Typically, $\mathbb{E}_f(\cdot)$ is hard to evaluate, and can be replaced by an estimate, by simulating data from $f(\mathbf{y})$, and evaluating $\nabla_\theta \log\left(p(\cdot;\theta)\right)$ in the simulated data. This is similar to a quasi-maximum likelihood approach, estimating a misspecified model using maximum likelihood estimation (Gourieroux et al., 1984; White, 1982).

2.1 Hidden Markov Model

As our approximating model, we propose using the following Hidden Markov Model. Denote the latent state by $x_{i,t}$, which lies in a finite discrete set $\{1, 2, \ldots, m\}$, evolving according to a first-order Markov process:[6]

$$y_{i,t}|x_{i,t} = \mu_t(x_{i,t}) + \text{diag}(\sigma_t)\varepsilon_{i,t}, \quad \varepsilon_{i,t} \sim N(0, I_k) \tag{2}$$

$$x_{i,t+1}|x_t \sim \Pi_{ij,t}, \tag{3}$$

[6] Assuming Gaussianity for $\varepsilon_{i,t}$ is convenient, because we will be using the EM algorithm to estimate θ, and, for Gaussian errors, the M step has a closed-form solution. In addition, the assumption of Gaussianity is used in our proof below.

where $i = 1, \dots, N$ and $t = 1, \dots, T$. The inclusion of a panel dimension allows for the estimation of parameters that vary with t, for example, over the life-cycle. Parameter vector θ in Equation (1) thus consists of:

(i) the parameters in transition probability matrix Π_t, denoted by $\Pi_{ij,t}$. In the case that there is no time dependence, that is, $\Pi_t = \Pi$ for all $t = 1, ..., T$, the number of parameters in Π is $m \times m$, of which $m \times (m-1)$ are linearly independent, given that each row sums to one.

(ii) the grid μ_t. When there is no time dependence, $\mu_t = \mu$ is an $m \times k$ matrix.

(iii) the variance of the error term σ_t^2. In the case that there is no time dependence, $\sigma_t^2 = \sigma^2$. If $y_{i,t} \in \mathbb{R}^k$ has $k > 1$, the variance is the diagonal matrix $\text{diag}(\sigma_{t,1}, ..., \sigma_{t,k})$.

These parameters $\theta = (\mu, \Pi, \sigma)$ result in a discretization of the process $f(\mathbf{y})$, where μ is the grid of the discretized process, and Π governs the transitions between the m states. By assuming that the approximating model is a Hidden Markov Model (HMM), this framework embeds a discrete Markov chain into a continuous support process via a continuous measurement error, which enables the use of the Kullback-Leibler (KL) divergence between the two processes as a measure of information loss. This framework is attractive because it allows us to build on existing approximation theory and computational algorithms. Our objective is to choose $\theta = (\mu, \Pi, \sigma)$ which minimizes the information loss between the true stochastic process and the approximating HMM:

$$\min_{(\Pi,\mu,\sigma)} D^{KL} \left(f(\mathbf{y}) || p \left(\mathbf{y}; \Pi, \mu, \sigma \right) \right). \tag{4}$$

The HMM framework also allows one to compute the information loss of existing discretization methods, by treating the transition probability matrix $\tilde{\Pi}$ and grid $\tilde{\mu}$ implied by such methods as fixed and defining the information loss as

$$\min_{\sigma} D^{KL} \left(f(\mathbf{y}) || p \left(\mathbf{y}; \tilde{\Pi}, \tilde{\mu}, \sigma \right) \right). \tag{5}$$

2.2 Properties of the KL divergence

Given our objective of minimizing the information loss between the true process and its approximation, two key questions arise. First, is there a consistent estimator of the Hidden Markov Model (HMM) parameters in this context? In the case of misspecified models, consistency is defined as the convergence of the estimator to the value that minimizes the Kullback-Leibler (KL) divergence. This property has been established for misspecified HMMs by Mevel and Finesso (2004) and later generalized by Douc and Moulines (2012). The second question concerns whether, with a sufficient number of hidden states (and thus grid points), the information loss between the true process and its approximation can be made arbitrarily small. Under a set of assumptions, we prove that the answer to this question is affirmative.

The Main Theorem builds on the results of Nguyen et al. (2020), who show that a mixture distribution with a sufficient number of components can approximate a large class of distribution functions arbitrarily well. We extend this result to the non-i.i.d. setting of continuous support Markov processes. That is, we show that a Hidden Markov Model in levels (as in Assumption (A5)) can approximate any stationary Markov process satisfying Assumptions (A1)-(A4) arbitrarily well as long as enough hidden states are used for the approximation.

For this result, we focus on the stationary case that omits time variation in the stochastic process and the parameters. Hence, in our notation, we suppress the panel dimension i in y and x, and the time dimension t in the parameters $\theta = (\Pi, \mu, \sigma)$. Denote the set of continuous functions with support on $\mathbb{R}^k$ by C. Define the set of locally compact functions (i.e., continuous functions that vanish at infinity) by

$$C_0 = \left\{ f \in C : \forall \epsilon > 0, \exists \text{ a compact } \mathbb{K} \in \mathbb{R}^k, \text{ such that } |f(x)| < \epsilon \text{ for all } x \notin \mathbb{K} \right\}.$$

We impose the following assumptions on the true process $f(\mathbf{y})$ and approximating model $p(\mathbf{y}, \theta)$:

(A1) $\mathbf{y} = \{y_t\}_{t=1}^{T}$ has a data generating process characterized by $f(\mathbf{y})$, $y_t \in \mathbb{R}^k$, that is first-order Markov and stationary, that is,

$$f(y_t|y_{t-1}, ..., y_1) = f(y_t|y_{t-1}),$$

and

$$f(y_{t+l}|y_{t+l-1}) = f(y_t|y_{t-1}) \quad \forall l \in \mathbb{N}.$$

(A2) $f(y_t|y_{t-1}) \in C_0$.

(A3) $\log f(y_t|y_{t-1})$ and $f(y_t|y_{t-1})$ are differentiable in y_{t-1}.

(A4) $\log f(y_t|y_{t-1})$ is locally Lipschitz continuous in y_{t-1}.

(A5) $p(\mathbf{y};\theta_m)$ is characterized by:

$$\begin{aligned} y_t|x_t &= \mu_m(x_t) + \text{diag}(\sigma_m)\varepsilon_t, \quad \varepsilon_t \sim N(0, I_k), \\ x_{t+1}|x_t &\sim \Pi_{ij,m} \end{aligned}$$

with parameters $\theta_m = (\mu_m, \Pi_m, \sigma_m)$, and $x_t \in \{1, ...m\}$ a latent state evolving according to a first-order Markov process with transition probability matrix Π_m. Denote the conditional distribution by $p(y_t|y_{t-1}, ..., y_1; \theta_m) \in C_0$.

The first-order Markov assumption (A1) is w.l.o.g., because any finite-order Markov process can be written as a multivariate first-order Markov process. Subscripts m are used to indicate the number of states of the HMM ("grid points"), also referred to as the complexity/size of the approximating model.

Main Theorem. *Under assumptions (A1)-(A5), given a sufficiently large number of grid points m, there exist a set of grid points μ_m, variance $\sigma_m \geq \tau > 0$ and transition probability matrix Π_m, collected in $\theta_m = (\mu_m, \Pi_m, \sigma_m)$ such that the KL divergence between $f(\mathbf{y})$ and $p(\mathbf{y};\theta)$, given by*

$$D^{KL}(f(\mathbf{y})||p(\mathbf{y};\theta)) = \int f(\mathbf{y}) \log \frac{f(\mathbf{y})}{p(\mathbf{y};\theta)} d\mathbf{y},$$

can be made arbitrarily small.

The full proof is given in Supplementary Appendix A.

Sketch of proof. First, we show the conditional distributions of a Hidden Markov Model, denoted by $p(y_t|y_{t-1}, ..., y_1; \theta_m)$, are Gaussian mixtures, whose mixture weights converge to a row of the transition probability matrix Π_m as m becomes large, such that $p(y_t|y_{t-1}, ..., y_1; \theta_m)$ converges to $p^0(y_t|y_{t-1}; \theta_m) := \sum_{j=1}^{m} \Pi_{lj}\phi_j(y_t)$, where l refers to the index of the closest grid point to y_{t-1} and $\phi_j(y_t)$ denotes the Gaussian probability density function evaluated at state j. The proof then applies the result of Nguyen et al. (2020) to m conditional distributions, that is, the true kernels $f(y_t|y_{t-1} = \mu_m(i))$ for $i = 1, ..., m$ are approximated by the m Gaussian mixtures $p^0(y_t|y_{t-1} = \mu_m(i); \theta_m)$. This results in an additional term in the KL divergence compared to the Nguyen et al. (2020) result, because in our setting, these m conditional distributions $f(y_t|y_{t-1} = \mu_m(i))$ are approximated by m Gaussian mixtures restricted to having the same location parameters μ_m. We show there are enough degrees of freedom for m different sets of convex mixture weights, because the transition probability matrix has $m \times m$ elements. The rest of the proof consists of three parts. First, we show that the additional term in the

KL divergence becomes arbitrarily small when m is large. Second, we show that when the KL divergences of these specific m conditional distributions are small, the KL divergences for all other potential realizations of $\{y_{t-k}\}_{k=1}^{t-1}$ are also small. Finally, we show that the KL divergence between $f(\mathbf{y})$ and $p(\mathbf{y};\theta)$ can be written as a function of the KL divergences between all conditional distributions, which concludes the proof.

While the theorem focuses on a stationary stochastic process, the core of the proof relies on approximating the conditional distributions $f(y_t \mid y_{t-1})$ using Gaussian mixtures. If these conditional distributions were time-varying, $f_t(y_t \mid y_{t-1})$, the corresponding transition matrix Π_t and grid μ_t would also vary with each time period t, a relaxation of (A1). In this case, the additional degrees of freedom would still allow the approximation to achieve arbitrary accuracy as the number of grid points increases provided (A2)-(A4) hold for $f_t(y_t|y_{t-1})$ and y_t is still first-order Markov.

Estimation. By the results of Mevel and Finesso (2004), given that the Maximum Likelihood Estimator (MLE) of a misspecified HMM is consistent, we know it minimizes the KL divergence for a given grid size m.[7] Therefore, we can use the Expectation-Maximization (EM) algorithm to obtain our grid points and transition probability matrix.

As a first step, simulate a panel of observations $y_{i,t}$, $i = 1, \ldots, N$ and $t = 1, \ldots, T$ from the continuous support process of interest. We follow the standard implementation of the EM algorithm (specifically, the Baum-Welch algorithm as it is for HMMs) as in Rabiner (1989) or Bilmes (1998). Let $\phi_{j,t}(y_{i,t})$ denote the Gaussian density function evaluated at $y_{i,t}$ conditional on $x_{i,t} = j$, i.e., with mean $\mu_t(j)$ and diagonal variance matrix Σ_t with elements $\sigma_{l,t}^2$, for $l = 1 \ldots, k$.

Given an initial (or updated) set of parameters $\theta = \left(\{\Pi_t, \mu_t, \sigma_t\}_{t=1}^T, \delta\right)$ for a time-varying process or $\theta = (\Pi, \mu, \sigma, \delta)$ for a stationary process, where δ denotes the initial state distribution, the E-step provides probabilities of states $x_{i,t}$ given observations $y_{i,t}$. These probabilities can be calculated recursively, enabling a numerically efficient implementation. Based on these probabilities, the parameter updates in the M-step have a closed-form solution. We now describe the E-step in detail. Let $y_i^t = (y_{i,1}, y_{i,2}, \ldots, y_{i,t})$, i.e., the observed values up to time t, let $\boldsymbol{y}_i = y_i^T$ be the full sequence of observations, and let $y_{i,t+1}^T = (y_{i,t+1}, y_{i,t+2}, \ldots, y_{i,T})$, i.e., the observed values from time $t+1$ to T. The forward probabilities $\boldsymbol{\alpha}_{i,t}(j)$ are given by

$$\boldsymbol{\alpha}_{i,t}(j) = p\left(y_i^t, x_{i,t} = j|\theta\right) \tag{6}$$

[7] This requires additional assumptions on the true stochastic process, including geometric ergodicity, and uniformly bounded moments of sufficiently high order which are satisfied in our applications.

and the backward probabilities $\boldsymbol{\beta}_{i,t}(s)$ are given by

$$\boldsymbol{\beta}_{i,t}(s) = p\left(y_{i,t+1}^{T} | x_{i,t} = s, \theta\right). \tag{7}$$

These are defined recursively as:

$$\boldsymbol{\alpha}_{i,1}(j) = \delta(j)\phi_j(y_{i,1}), \quad \boldsymbol{\alpha}_{i,t+1}(j) = \left(\sum_{s=1}^{m} \boldsymbol{\alpha}_{i,t}(s)\Pi_{sj}\right)\phi_j(y_{i,t+1}), \tag{8}$$

$$\boldsymbol{\beta}_{i,T}(s) = 1, \quad \boldsymbol{\beta}_{i,t}(s) = \sum_{j=1}^{m} \Pi_{sj}\phi_j(y_{i,t+1})\boldsymbol{\beta}_{i,t+1}(j). \tag{9}$$

Using these probabilities, we can define the probability of being in state j at time t, and observing the entire sequence $\boldsymbol{y}_i$ as

$$p(\boldsymbol{y}_i, x_{i,t} = j|\theta) = \boldsymbol{\alpha}_{i,t}(j)\boldsymbol{\beta}_{i,t}(j). \tag{10}$$

This leads to a posterior probability of being in state j conditional on observing $\boldsymbol{y}_i$, given by

$$\gamma_{i,t}(j) = p(x_{i,t} = j|\boldsymbol{y}_i, \theta) = \frac{p(\boldsymbol{y}_i, x_{i,t} = j|\theta)}{p(\boldsymbol{y}_i|\theta)} = \frac{\boldsymbol{\alpha}_{i,t}(j)\boldsymbol{\beta}_{i,t}(j)}{\sum_{s=1}^{m} \boldsymbol{\alpha}_{i,t}(s)\boldsymbol{\beta}_{i,t}(s)}. \tag{11}$$

We can also define the posterior probability of being in state s at time t and state j at time $t+1$ as

$$\begin{aligned}\xi_{i,t}(s,j) &= p(x_{i,t+1} = j, x_{i,t} = s|\boldsymbol{y}_i, \theta) \\ &= \frac{p(\boldsymbol{y}_i, x_{i,t+1} = j, x_{i,t} = s|\theta)}{p(\boldsymbol{y}_i|\theta)} \\ &= \frac{\boldsymbol{\alpha}_{i,t}(s)\Pi_{sj}\phi_j(y_{i,t+1})\boldsymbol{\beta}_{i,t+1}(j)}{\sum_{k=1}^{m} \boldsymbol{\alpha}_{i,t}(k)\boldsymbol{\beta}_{i,t}(k)}.\end{aligned}$$

When parameters vary with time t, the M-step is given by:

$$\mu_t^l(j) = \frac{\sum_{i=1}^N y_{i,t}^l \gamma_{i,t}(j)}{\sum_{i=1}^N \gamma_{i,t}(j)} \tag{12}$$

$$\sigma_{t,l}^2 = \frac{\sum_{i=1}^N \sum_{j=1}^m \left(y_{i,t}^l - \mu^l(j)\right)^2 \gamma_{i,t}(j)}{\sum_{i=1}^N \sum_{j=1}^m \gamma_{i,t}(j)} \tag{13}$$

$$\Pi_{t,sj} = \frac{\sum_{i=1}^N \xi_{i,t}(s,j)}{\sum_{i=1}^N \gamma_{i,t}(s)} \tag{14}$$

$$\delta(j) = \frac{\sum_{i=1}^N \gamma_{i,1}(j)}{\sum_{s=1}^m \sum_{i=1}^N \gamma_{i,1}(s)}, \tag{15}$$

for $t = 1, \ldots, T$ where $l = 1, ..., s$ denotes different elements of the vector $y_{i,t} = (y_{i,t}^1, ..., y_{i,t}^k)$. Alternatively, when parameters do not vary with t, the M step becomes:

$$\mu^l(j) = \frac{\sum_{i=1}^N \sum_{t=1}^T y_{i,t}^l \gamma_{i,t}(j)}{\sum_{i=1}^N \sum_{t=1}^T \gamma_{i,t}(j)} \tag{16}$$

$$\sigma_l^2 = \frac{\sum_{i=1}^N \sum_{t=1}^T \sum_{j=1}^m \left(y_{i,t}^l - \mu^l(j)\right)^2 \gamma_{i,t}(j)}{\sum_{i=1}^N \sum_{t=1}^T \sum_{j=1}^m \gamma_{i,t}(j)} \tag{17}$$

$$\Pi_{sj} = \frac{\sum_{i=1}^N \sum_{t=1}^T \xi_{i,t}(s,j)}{\sum_{i=1}^N \sum_{t=1}^T \gamma_{i,t}(s)}, \tag{18}$$

and the initial distribution δ is updated as the stationary distribution associated with Π (i.e., the left eigenvector associated with the unit eigenvalues of Π). The EM algorithm iterates between the E-step and the M-step until convergence.

Practical Guidelines. The EM algorithm provides local improvements to the likelihood given an initial guess of parameter values θ^0, and it is well known that this algorithm is sensitive to the initial guess. In practice, for univariate processes, we find that a good initial guess is an equal-spaced grid and transition probability matrix as implied by the binning method (Adda and Cooper, 2003) or the Tauchen (1986) method. As such, the HMM estimates are guaranteed to provide a local improvement over these standard methods. For multivariate processes, an initial grid point placement based on the ergodic distribution, and a transition matrix computed using nearest-neighbor binning works well. In our life-cycle applications in Section 4 where the stochastic processes vary by age, for every age t we compute an initial grid and transition probability matrix based on the binning method. Finally, for the choice of the panel dimension N and the time dimension T of the simulation, a good choice is to simulate data until the transition matrix and grid computed via binning (Adda and Cooper, 2003) stabilizes, which can be done recursively.[8]

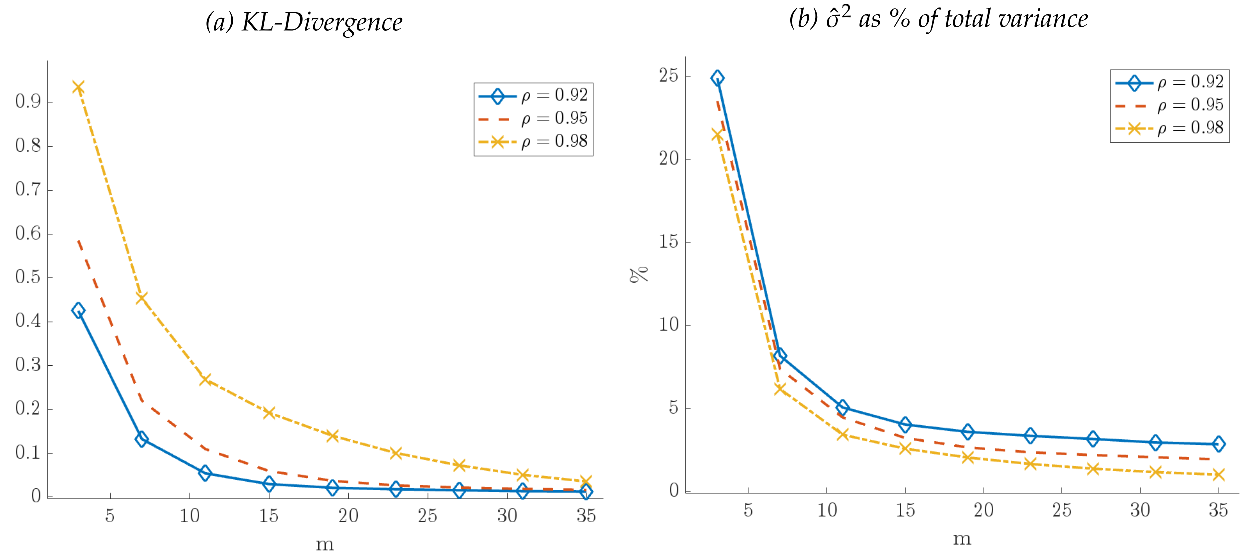
The EM algorithm estimates the parameters μ, Π, and σ. To incorporate the estimated HMM into typical economic models, i.e., a dynamic programming problem, there are two natural approaches. First, one could include the Hidden Markov Model as the state (x, ε) by discretizing ε such that both the latent state x and error ε are discrete.[9] Second, one can discard the error term, effectively treating μ as the grid and Π as the transition probability matrix of a discrete Markov chain. The advantage of the latter approach is the simplicity of implementation, but discarding ε may lead to a worse approximation of the model solution. The former approach is more computationally demanding, as the discretization of ε requires a tensor product with the grid μ. Given that as the number of grid points increases and the approximation improves, σ decreases and the contributions of ε becomes smaller, the following trade-off arises. One could either discretize ε onto an n point support and add it to the grid μ which has m points, increasing the state space to nm points, or, one could estimate the HMM with nm points and discard ε entirely. We illustrate this trade-off in Supplementary Appendix E for the example of an AR(1) process in the life-cycle model introduced in Section 4. We find, in this example, that it is more efficient to discard ε, even when n is as small as two. We acknowledge that this trade-off is likely process and model specific, but motivated by the computational simplicity of the second approach, we proceed by discarding ε in our applications and find this approach does well in these settings.

Number of grid points. When selecting the number of grid points m, one faces a trade-off between parsimoniousness for computational efficiency and accuracy of the approximation. In theory, the discretized process becomes arbitrarily accurate as the dimension of the grid goes to infinity. In practice, the grid must always have a finite dimension. One advantage of full-information discretization is that we can assess the fit of the approximating model with a finite number of grid points, as this fit is quantified by the KL divergence. We propose using a scree plot with the KL-divergence on the y-axis, and the number of grid points on the x-axis, as visualized in Figure 1 for three different parameterizations of an AR(1) process. This allows a practitioner to visualize the gain in approximation accuracy from adding an additional grid point.

[8] Codes to discretize the processes in this paper can be found here: https://github.com/SeanMcCrary/HMM_Discretization.

[9] Discretizing ε is straightforward. This is an i.i.d. Gaussian random variable, so one can use standard quadrature rules.

Figure 1: AR(1) Examples



(a) KL-Divergence

(b) $\hat{\sigma}^2$ as % of total variance

Notes: HMM as in in Equations (2)-(3) versus the true AR(1) process $y_t = \rho y_{t-1} + \varepsilon_t$, $\varepsilon_t \sim N(0, 0.01)$ for three values of ρ, where m is the number of grid points used for the discretization. Left panel reports the KL-Divergence between the HMM approximation and the true model. Right panel displays the variance of the HMM error terms $\hat{\sigma}^2$ relative to the unconditional variance of the true process.

Although the Main Theorem does not explicitly specify the rate of convergence (i.e., the number of grid points required to achieve a particular level of information loss), the proof offers insights into the properties of the true process that influence how many grid points are needed to obtain a given precision. Specifically, this depends on the local Lipschitz coefficient of $\log f(y_t|y_{t-1})$ and $f(y_t|y_{t-1})$. One property affecting the number of grid points is the persistence of the stochastic process. For an AR(1) process, it can be shown that these Lipschitz coefficients increase with the persistence of the process. Consequently, the more persistent a stochastic process, the more grid points are required to achieve the same level of information loss. Figure 1 demonstrates how the KL-divergence of our HMM approximation to an AR(1) process approaches zero as the number of grid points m increases, but at a slower rate when the AR(1) process exhibits higher persistence. These results provide insights into why discretizing highly persistent AR(1) processes presents challenges, as highlighted in Flodén (2008), Galindev and Lkhagvasuren (2010), and Kopecky and Suen (2010).

2.3 Imposing structure through restrictions

One can impose additional structure on the discretized process by estimating the process under a set of restrictions. For example, one might prefer a discretization that does match certain conditional or unconditional moments of the stochastic process, or reflects the symmetry in the underlying stochastic process. In our EM estimation procedure, this can be done by modifying the M step.

For symmetric processes, a symmetry restriction can be imposed on μ. In case of a process that is symmetric around zero and an odd number of grid points m, this means that:

$$\mu(\lceil m/2 \rceil) = 0, \text{ and } \mu(\lceil m/2 \rceil - r) = -\mu(\lceil m/2 \rceil + r), \quad \text{for } r = 1, ..., \lfloor m/2 \rfloor \tag{19}$$

Similarly, a process can also be symmetric in its dynamics, as reflected by the transition probability matrix. In that case, the restriction takes the form

$$\Pi_{i,j} = \Pi_{(m+1-i),(m+1-j)}. \tag{20}$$

For the specific restrictions in Equations (19)-(20), a closed-form solution is available for the M step.

3 Application I: Asset Pricing Model with Stochastic Volatility

In this section, we assess the performance of our method using an asset pricing model in which dividend growth exhibits stochastic volatility. Most models requiring the solution of a dynamic stochastic optimization problem with a continuous-support process lack a closed-form solution. However, as demonstrated by De Groot (2015), the asset pricing model we present below provides a closed-form solution for the price-dividend ratio, the risk-free interest rate, and the conditional expected return on equity. The existence of this analytical solution offers a benchmark for comparing results obtained using our discretization method against those produced by alternative approaches.

First, we present the analytically tractable asset pricing model of De Groot (2015). Next, we demonstrate how to discretize the AR(1)-SV process using ours and two other methods, and analyze their respective performance at capturing various moments of the stochastic process. Finally, we assess how the numerical solution corresponding to each method differs relative to the analytical benchmark solution.

3.1 Analytically tractable asset pricing model with AR(1)-SV dividend growth

We use the Lucas tree asset pricing model of De Groot (2015). A representative agent maximizes the expected discounted stream of utility:

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \frac{c_t^{1-\gamma}}{1-\gamma}$$
$$\text{s.t. } c_t + s_{t+1} p_t \leq (d_t + p_t) s_t,$$

where c_t is consumption, and s_t is an asset with price p_t and dividends d_t. Parameter $\beta \in (0,1)$ denotes the discount factor and γ is the coefficient of relative risk aversion.

The growth rate of dividends $y_t = \ln(d_t/d_{t-1})$ is assumed to follow an AR(1) process with stochastic volatility:[10]

$$y_t = \bar{y} + \rho(y_{t-1} - \bar{y}) + \sqrt{\eta}_t \varepsilon_t \tag{21}$$
$$\eta_t = \bar{\eta} + \rho_\eta(\eta_{t-1} - \bar{\eta}) + \omega \varepsilon_{\eta,t}. \tag{22}$$

with persistence in levels $\rho \in (-1,1)$, and ε_t is i.i.d. $N(0,1)$. The random variable η_t is the time-varying conditional variance of dividend growth. Parameter $\rho_\eta \in (-1,1)$ is the persistence of the stochastic volatility process, and $\varepsilon_{\eta,t}$ is also i.i.d. $N(0,1)$.

Market clearing, $s_t = 1$, implies that $c_t = d_t$. Defining the price-dividend ratio as $v_t := p_t/d_t$, the first-order condition of the representative agent's maximization problem is given by:

$$v_t = \mathbb{E}_t \beta \left(\frac{d_{t+1}}{d_t}\right)^{1-\gamma} (v_{t+1} + 1). \tag{23}$$

De Groot (2015) derives a closed-form solution for the price-dividend ratio v_t, the conditional expected return on equity, which is defined as:

$$\mathbb{E}_t R^e_{t+1} = \mathbb{E}_t \left(\frac{d_{t+1} + p_{t+1}}{p_t}\right), \tag{24}$$

[10] Note, that for this specification of the AR(1)-SV process, η_t can become negative, in which case $\sqrt{\eta_t}$ is imaginary. In the parametrization we use, taken from De Groot (2015), the probability of a negative value for η is very small, and in our long sample of simulations, it doesn't occur. The particular specification is necessary to obtain closed-form solutions in De Groot (2015).

and the risk-free rate defined as:

$$R_t^{rf} = \left[\beta \mathbb{E}_t \left(\frac{c_{t+1}}{c_t}\right)^{-\gamma}\right]^{-1}. \tag{25}$$

Details on the analytical solution of De Groot (2015) and the discretized solution are provided in Supplementary Appendix C.

Another object economists care about is the welfare cost of risk. In this application, we measure this using the certainty equivalent consumption. Define

$$V(d) = u(d) + \beta \mathbb{E}[V(d')|d],$$

where $V(d)$ is the value to the household of being in state d, where d is the level of aggregate dividends. $V(d)$ reflects the present discounted value of the risky dividend (i.e., consumption) stream. The certainty equivalent level of consumption is the certain (constant) level of consumption that the household is indifferent to receiving versus the stochastic consumption sequence. We denote that constant value by $x(d)$, which is defined implicitly as the solution to:

$$V(d) = \frac{u(x(d))}{1-\beta}.$$

We solve for $x(1)$ numerically by simulation using the true stochastic process for dividend growth and the discretized processes. Lower values of x indicate the household is willing to receive a lower constant consumption stream instead of facing uncertainty, so to the extent a discrete approximation fails to capture risk, it will overstate x relative to the true value.

Calibration. The stochastic volatility process we use is the one provided in De Groot (2015) which is an annualized parametrization of the process in Bansal and Yaron (2004). The parameters of the stochastic volatility process are $\rho_\eta = 0.855$, $\omega = 7.4000 \times 10^{-5}$, $\bar{\eta} = 0.0012$, $\rho = 0.868$, $\bar{y} = 0.0179$. These parameters, as well as risk aversion $\gamma = 1.5$ and the discount factor $\beta = 0.95$, are chosen such that the price-dividend ratio is finite and stable.[11]

3.2 Discretizing the AR(1)-SV process of De Groot (2015)

The process of Equations (21)-(22) is multivariate, which is why we discretize over both the levels y_t and variances η_t. We compare our discretization method with the method of Farmer

[11] De Groot (2015) provides parameter restrictions such that the price-dividend ratio is finite, see Supplementary Appendix C.

and Toda (2017) and the binning method of Adda and Cooper (2003).[12] Both methods use a tensor grid for multivariate processes, while our methods estimates the grid points optimally.

Figure 2 visualizes the KL-divergence of our discretization for different choices of grid size m relative to the true AR(1)-SV process. The figure also visualizes the KL divergences of the two other discretization methods, computed as in Equation (5). For the discretization methods that rely on tensor grids, we use a three-grid point discretization for η_t and vary the number of grid points for y_t from three to fifteen. The figure shows our method is more parsimonious; to capture the same amount of information as we do with 15 grid points, the Farmer and Toda method needs 33 grid points, and the binning method needs more than 45. This is due to both our method being a full-information method, as well as our method not relying on tensor grids but rather using an optimally chosen grid.

Figure 2 also visualizes three additional statistics to compare the performance of our method and the existing methods at capturing moments of y. Our method gives rise to the lowest Mean Squared Forecast Error (MSFE).[13] The MSFE of the other methods is 30-40% larger than ours when using few grid points, supporting that we give an agent a better process to make forecasts with. The Farmer and Toda (2017) method does well at the standard deviation, as this is a moment that method targets. Our method's performance at matching the unconditional kurtosis is comparable to Farmer and Toda (2017), while the binning method underestimates the kurtosis considerably.

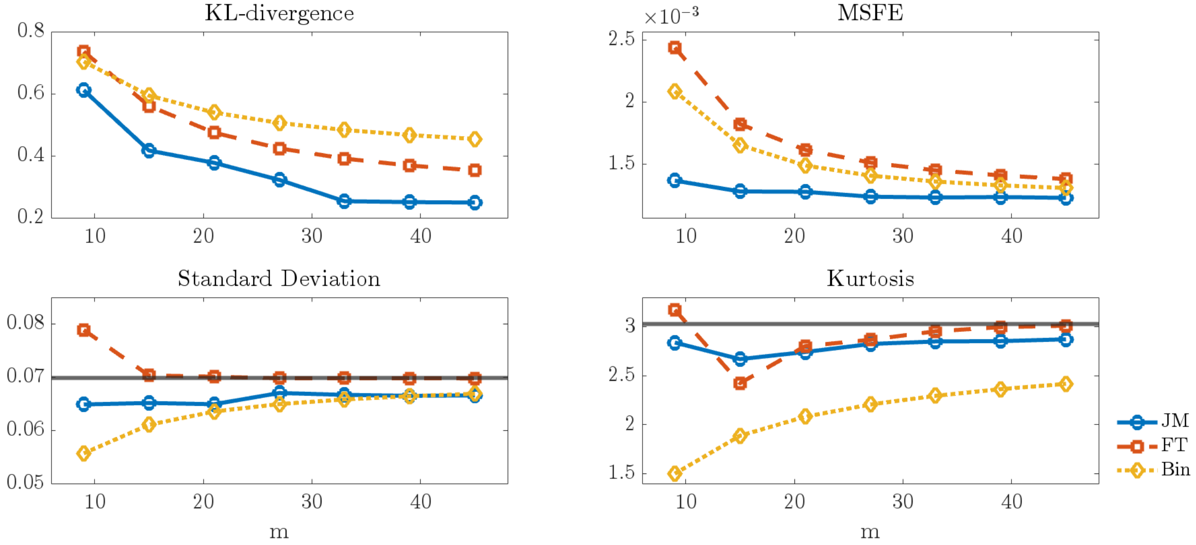
Figure 3 illustrates how our method optimally chooses the grid points for a multivariate process like the AR(1)-SV process.[14] The top two panels of Figure 3 show our optimal grids for $m = 9$ and $m = 15$ grid points, which can be compared to the bottom two panels showing standard tensor grids. As Figure 3 shows, our optimal grids take the shape of the ergodic distribution over (y_t, η_t). In the optimal grids, tensor-like structures emerge (i.e., grid points with similar levels of y and different η), but they are more likely to occur in high density regions of y. As the number of grid points increases, the number of tensor-like states increases as well, and they emerge further in the tails of y.

[12] We use the codes provided on the personal website of A.A. Toda, available at https://alexisakira.github.io/discretization/ for the implementation of the Farmer and Toda (2017) method. We adapt the Farmer and Toda (2017) method for the De Groot (2015) specification of the AR(1)-SV process, in particular, the method attempts to match the first two conditional moments in each grid point.

[13] The mean squared forecast error (MSFE) of the approximating model measures the one-step ahead forecasting error that the agent makes. For this statistic, we assume that an agent assigns the grid point closest to the current realization of y_t for forecasting y_{t+1}. Define MSFE $= \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2$, where $\hat{y}_t = \sum_j \Pi_{ij} \cdot \mu(x_t = j)$ and $i = \underset{i \in \{1,\ldots,m\}}{\operatorname{argmin}} |y_{t-1} - \mu(x_{t-1} = i)|$.

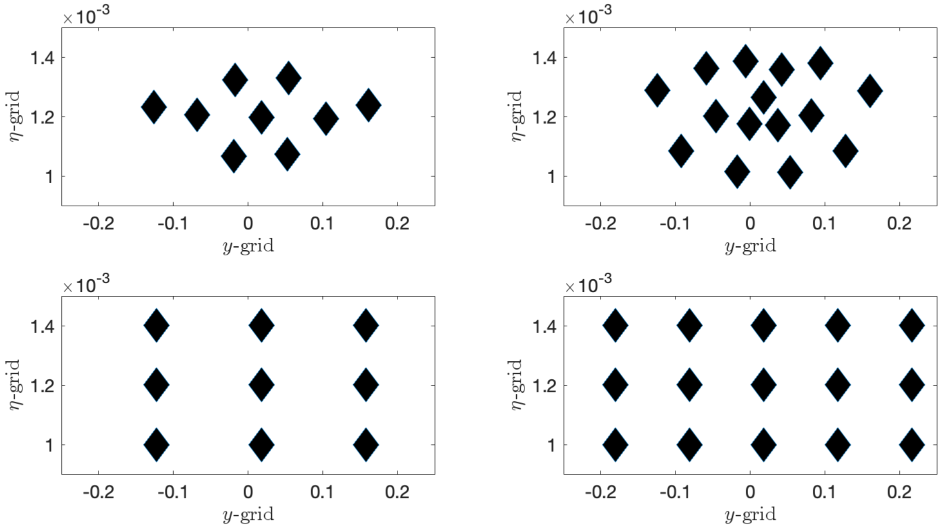
[14] In Supplementary Appendix C, we show the optimal grids for Vector Autoregressions, another multivariate process.

Figure 2: Statistical moments of three discretizations of the AR(1)-SV process



Notes: Visualization of KL-divergence, Mean Squared Forecast Error (MSFE), Standard Deviation and Kurtosis of three different discretization methods (ours (JM), Farmer and Toda (FT) and binning (Bin)). True process as in Equations (21)-(22), for different discretization methods and different grid sizes m. Note that the other methods rely on a tensor grid. For those methods, we fix the dimension of η at three, and vary the dimension of y, and m is the product of both dimensions.

Figure 3: Visualisation of the optimal grid for the AR(1)-SV process

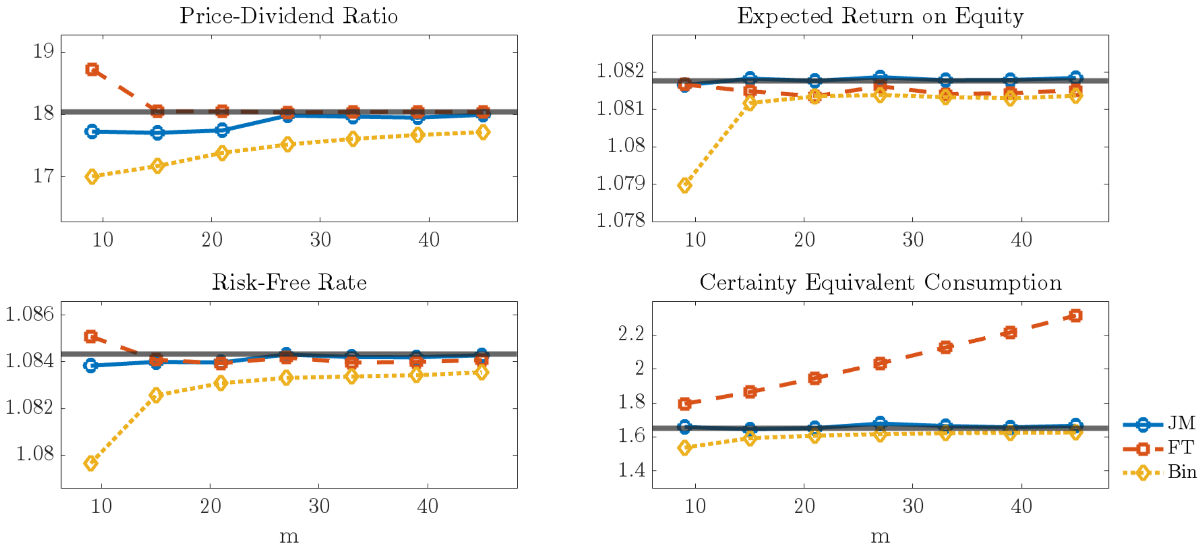


Notes: Optimal grids for $m = (9, 15)$ implied by our discretization method (top two panels) compared to a tensor grid (bottom two panels, as in, e.g., Farmer and Toda). True AR(1)-SV process as in Equation (21)-(22). The grids for the AR(1)-SV process are two-dimensional. The y-axis depicts the variance, while the positioning on the x-axis of the diamonds depicts the level of y.

3.3 Accuracy of the models solutions

To compare the relative performance of our method versus existing methods at solving the asset pricing model, we compute moments of the discrete solutions and the analytical benchmark. The results of this analysis are summarized in Figure 4. With as few as $m = 9$ grid points, our method captures most model moments well, especially compared to the other methods. The KL-divergence in Figure 2 suggests that, after $m = 33$ grid points, additional grid points do not noticeably improve the discretization. This is consistent with Figure 4, where the model implied price-dividend ratio, expected return on equity, risk-free rate, and certainty equivalent consumption are essentially equal to the analytical benchmark at $m = 33$.

Figure 4: Asset pricing model moments of three discretizations of the AR(1)-SV process



Notes: Visualization of the Price-Dividend Ratio, Expected Return on Equity, Risk Free Rate and Certainty Equivalent Consumption of the true benchmark (horizontal grey line) and three different discretization methods (ours (JM), Farmer and Toda (FT) and binning (Bin)) for different grid sizes. True process as in Equations (21)-(22), for different discretization methods and different grid sizes m. Note that the other methods rely on a tensor grid. For those methods, we fix the dimension of η at three, and vary the dimension of y, and m is the product of both dimensions.

The method of Farmer and Toda (2017) provides an accurate solution of the price-dividend ratio, and slightly underestimates the expected return on equity and the risk-free rate. Their method, however, noticeably overestimates the certainty equivalent consumption, and this gets worse as the number of grid points increases. This occurs because the method of Farmer and Toda (2017) aims to match the conditional mean and variance at each grid point. To achieve this, at the end points of the grid, it distorts the conditional kurtosis (a moment that is not matched). As these end points occur with low probability, the unconditional kurtosis displayed in Figure 2 is largely unaffected. However, this distortion of the tail risk has a large effect on the certainty equivalent consumption. The binning method reliably converges to the benchmark, but an accurate solution requires considerably more grid points than our method. Overall, we conclude our method provides an accurate model solution with fewer grid points than existing methods.

4 Application II: Life-cycle Model

In this section, we evaluate the quantitative implications of different discretization methods for consumption, wealth and welfare using an incomplete markets life-cycle model. While simple, this model forms the basis for most of the heterogeneous household quantitative macro literature. We expect that our results on the importance of accurate discretizations also hold in richer models. In addition, this application demonstrates how our discretization method can be applied to non-linear non-Gaussian processes with life-cycle dynamics where the grids and transition probability matrices are allowed to vary by age.

We first discuss the life-cycle model we will use in our analysis. Next, we discuss the different stochastic processes, our performance at discretizing these processes, and what the implications are for the model solutions, using ours and existing methods.

4.1 Model and calibration

We begin by discussing the model environment, followed by the household optimization problem, and the details of the calibration.

Environment. We consider a partial equilibrium life-cycle version of the canonical incomplete-markets model without aggregate uncertainty. Households live up to T periods, where the first $t < T_r$ are spent working, and the remaining periods are spent in retirement. Working households supply one unit of labor inelastically with pre-tax earnings $\mathfrak{e}_t$ that evolve stochastically as described in more detail below. Retired households receive pension b and survive with probability s_t each period. Asset markets are incomplete. Agents can borrow and save using an uncontingent bond, at risk-free interest rate r, up to an exogenous borrowing limit $\underline{a}$.

Household problem. At every age, agents choose consumption c and saving a' subject to the budget constraint which depends on the current state of assets a and earnings $\mathfrak{e}$. During their working life ($t < T_r$), households solve the following optimization problem:

$$V_t(a,\mathfrak{e}) = \max_{c,a'} \left\{ u(c) + \beta \mathbb{E}_t V_{t+1}\left(a',\mathfrak{e}'\right) \right\},$$

$$\begin{aligned} \text{s.t. } & c + a' = \tau(\mathfrak{e}) + (1+r)a \\ & a' \geq \underline{a}, \end{aligned}$$

where earnings satisfy

$$\mathfrak{e}_t = g_t y_t.$$

That is, earnings in levels $\mathfrak{e}_t$ are the product of a common deterministic age component g_t and an idiosyncratic stochastic component y_t that evolves according to a (possibly age-dependent) Markov transition matrix Π_t.

Retired households solve the following problem:

$$V_t(a) = \max_{c,a'} \left\{ u(c) + \beta s_t V_{t+1}\left(a'\right) \right\},$$

$$\begin{aligned} \text{s.t. } & c + a' = b + (1+r)a \\ & a' \geq \underline{a}. \end{aligned}$$

Calibration. Agents enter the model at age 25 and work until age $T_r = 65$ (60 for the ABB process), after which they can be retired up to 25 years. If agents reach age $T = T_r + 25$, they die with certainty. The exact year of death after retirement is stochastic, and the survival probabilities are taken from the Social Security Administration actuarial life table. Retirement benefit b is chosen to match the 45% replacement rate of average earnings, which is a good approximation of the system in the United States (Mitchell and Phillips, 2006).

Utility has CRRA form:

$$u(c) = c^{1-\gamma}/(1-\gamma).$$

The coefficient of relative risk aversion γ is set to 2. The risk free rate r is 4% and the borrowing limit $\underline{a}$ is 12% of average earnings, which De Nardi, Fella, and Paz-Pardo (2020) find is roughly the ratio of credit card limits to income in the Survey of Consumer Finances. The discount factor β is calibrated to match a wealth-to-income ratio of 3.1 for the working age population, and this will be re-calibrated for each process, and for each discretization method.

Following Benabou (2002), the labor income tax function has the form:

$$\tau(\mathfrak{e}) = (1-\chi)\tilde{\mathfrak{e}}^{1-\mu}.$$

where $\tilde{\mathfrak{e}} = \max\left\{\mathfrak{e}, \underline{\mathfrak{e}}\right\}$. We choose an earnings floor of $\underline{\mathfrak{e}}$ equal to 25% of average earnings. The parameters χ and μ govern the level and progressivity of the tax function. Following Krueger and Wu (2021), we set the progressivity parameter χ to 0.1327, and the level parameter μ to 0.1575. The specification for the deterministic component of earnings g_t is taken from Guvenen et al. (2021). The calibration is summarized in Table 1.

Table 1: Calibration of the life-cycle model parameters

Parameter	Description	Value	Motivation
γ	Risk aversion	2.0	De Nardi et al. (2020)
b	Retirement benefits	0.45	Mitchell and Phillips (2006)
r	Risk-free interest rate	0.04	De Nardi et al. (2020)
$\underline{a}$	Borrowing limit	-0.12	De Nardi et al. (2020)
μ	Income tax progressivity	0.1327	Krueger and Wu (2021)
χ	Income tax level	0.1575	Krueger and Wu (2021)
W/I	Wealth-to-income ratio	3.1	De Nardi et al. (2020)

Model statistics. When presenting the model solution, we report several statistics, such as correlations and standard deviations of consumption, asset holdings and earnings. In addition, we compute two other statistics. First, we compute the certainty equivalent value (CEV). This is the fraction of lifetime consumption an individual would be willing to give up to live in a world without risk.[15] The CEV is commonly used to evaluate the welfare cost of risk and policy experiments, so it is important to know its sensitivity to the discretization method. Second, we use the model solution to compute the Marginal Propensity to Consume out of transitory income shocks (MPC). We compute the MPC as the change in consumption divided by the (unexpected) increase in cash-on-hand. MPC's are a common object of interest when studying fiscal and monetary policy.

[15] Let c^1 be the sequence of consumption arising in an economy with risk and c^0 be the sequence of consumption without risk. The CEV is defined in terms of welfare W as $W\big((1-CEV)c^0\big) = W\big(c^1\big)$ (Krueger and Wu, 2021).

4.2 Discretizing Guvenen, Karahan, Ozkan and Song (2021)

Stochastic process. The first earnings process we consider is the process proposed by Guvenen et al. (2021). This earnings process is given by:[16]

$$\begin{aligned}
y_{it} &= (1-\nu_{it})e^{(z_{it}+\varepsilon_{it})} \\
z_{it} &= \rho z_{i,t-1} + \eta_{it} \\
z_{i,0} &\sim N(0, \sigma_{z_0}) \\
\eta_{it} &\sim \begin{cases} N(\mu_{\eta,1}, \sigma_{\eta,1}) & \text{with prob. } p_z \\ N(\mu_{\eta,2}, \sigma_{\eta,2}) & \text{with prob. } 1-p_z \end{cases} \\
\varepsilon_{it} &\sim \begin{cases} N(\mu_{\varepsilon,1}, \sigma_{\varepsilon,1}) & \text{with prob. } p_\varepsilon \\ N(\mu_{\varepsilon,2}, \sigma_{\varepsilon,2}) & \text{with prob. } 1-p_\varepsilon \end{cases} \\
v_{it} &\sim \begin{cases} 0 & \text{with prob. } 1-p_v(t, z_{it}), \\ \min\{1, \exp(\lambda)\} & \text{with prob. } p_v(t, z_{it}) \end{cases}
\end{aligned} \tag{26}$$

where p_v is given by

$$p_v(t, z_t) = \frac{e^{\xi_{it}}}{1+e^{\xi_{it}}}, \quad \text{where } \xi_{it} \equiv a + bt + cz_{it} + dz_{it}t.$$

Here y_{it} is the earnings level of individual i at time t, z_{it} is the persistent component of earnings, ε_{it} is the transitory component and v_{it} is a non-employment shock. The process is essentially a persistent-transitory earnings process, where the main features are: (i) the fat-tailed innovations to the persistent and transitory component, and (ii) the non-employment shocks ν_{it}.

Discretization. Our paper is one of the first to discretize the process in Guvenen et al. (2021).[17] We compare ourselves against a binning method. Simple quantile binning of earnings does not work for the Guvenen et al. (2021) process, because the presence of non-employment risk makes the process multivariate. We adapt the standard binning method by adding a fixed

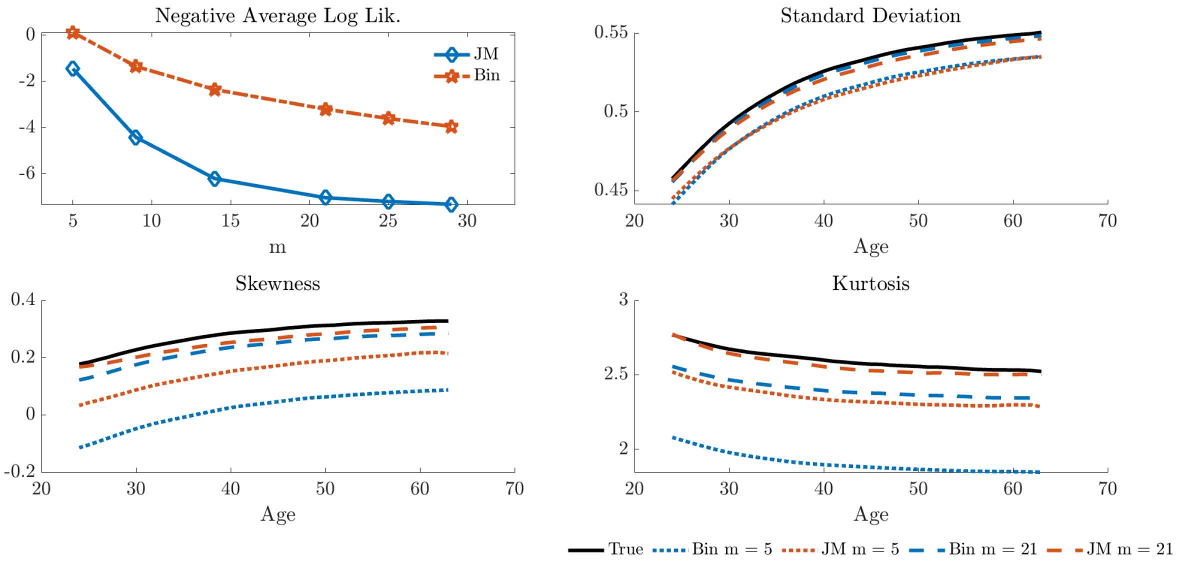
[16] We leave out the non-stochastic elements of the income-level, such as the fixed effect. Following Guvenen et al. (2021), we use the following parametrization: $\rho = 0.959$, $p_z = 0.407$, $\mu_{\eta,1} = -0.085$, $\mu_{\eta,2} = 0.085p_z/(1-p_z)$, $\sigma_{\eta,1} = 0.364$, $\sigma_{\eta,2} = 0.069$, $p_\varepsilon = 0.13$, $\mu_{\varepsilon,1} = 0.271$, $\mu_{\varepsilon,2} = -0.271p_\varepsilon/(1-p_\varepsilon)$, $\sigma_{\varepsilon,1} = 0.285$, $\sigma_{\varepsilon,2} = 0.037$, $\lambda = 0.0001$. We have $(a, b, c, d) = (-3.353, -0.859, -5.034, -2.895)$.

[17] Guvenen, Ozkan, and Madera (2024) solve a lifecycle model with a similar process using integration and interpolation over the Gaussian error terms instead of discretization.

number of zero earnings-states $y_{it} = 0$ and using the underlying dynamics of z_{it} to determine the transitions, effectively binning z_{it} conditional on non-employment. For positive values of earnings $y_{it} > 0$ we apply standard quantile binning (Adda and Cooper, 2003) to determine the grid points and transitions. The resulting grid and transition probability matrix vary over the lifecycle. Having multiple non-employment states generates heterogeneous job-finding probabilities, that is, non-employment states that differ in terms of their persistence.

For our discretization method, we use a multivariate discretization on $\log(y_{it} + 1)$ and z_{it} jointly. To make our method comparable to binning, we fix the number of non-employment states to be equal. We use the grids, transition matrices, and initial distribution over states from our binning method as the initial guess in our EM algorithm. As in binning, the grid and transition probabilities of our discretization are age-dependent and feature heterogeneous non-employment risk. We vary the number of grid points to evaluate the quality of the discretization in terms of statistical and model moments, and report its comparison with binning.

Figure 5: Statistical moments for two discretizations of Guvenen et al. (2021)



Notes: the first panel is the negative log-likelihood at different grid sizes m which is related to the KL-divergence. The remaining three panels report the standard deviation, skewness, and kurtosis of earnings by age for the true model and discrete approximations. JM refers to our method, and Bin refers to the binning method.

Figure 5 visualizes the negative log-likelihood (lower negative log-likelihood implies a smaller KL-divergence, as in Equation (5)), as well as the unconditional moments of the earnings levels of the Guvenen et al. (2021) process over the life-cycle, and the extent to which the discretized processes can replicate these moments. The first panel suggests that after $m = 21$ grid points, additional grid points only marginally improve the fit as measured through KL-divergence. This is consistent with the unconditional moments plots which show at $m = 21$ our discretization (red dashed line) coincides with the true model moments. Binning with $m = 21$ grid points is less accurate in terms of unconditional moments of earnings, suggesting it requires more grid points to achieve the same level of accuracy as our method.

4.3 Discretizing Arellano et al. (2017)

Stochastic process. Next, we consider the non-parametric earnings process in Arellano et al. (2017). Denote pre-tax labor earnings for individual i at age t as y_{it}, and decompose $\log y_{it}$ as follows:

$$\log y_{it} = \eta_{it} + \varepsilon_{it}$$

where η_{it} denotes the persistent component and ε_{it} denotes the transitory component. The transitory component is mean zero and is independent over time and from the persistent component. The persistent component η_{it} follows a general first-order Markov process, with its τth conditional quantile given $\eta_{i,t-1}$ by $Q_t(\eta_{i,t-1}, \tau)$ for each $\tau \in (0, 1)$, that is, without loss of generality:

$$\eta_{it} = Q_t(\eta_{i,t-1}, u_{it}), \quad (u_{it}|\eta_{i,t-1}, \eta_{i.t-2}, ...) \sim \text{Uniform}(0, 1)$$

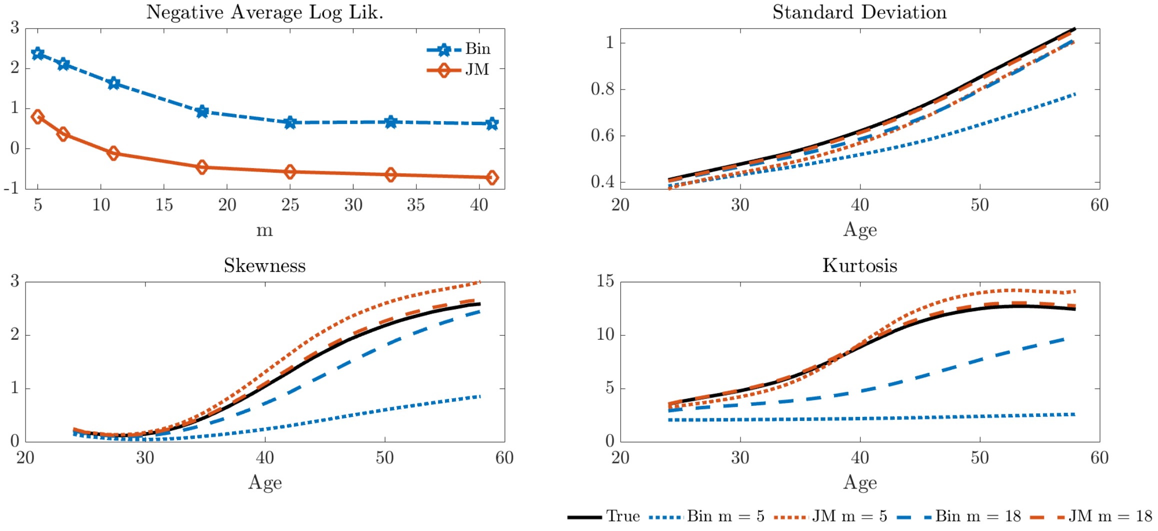
This model allows for nonlinear dynamics of earnings, and in particular, generates nonlinear persistence. Arellano et al. (2017) estimate this model non-parametrically, approximating Q using low-order products of Hermite polynomials.

Discretization. Our method only requires a simulated sample from the true stochastic process, and, therefore, can be applied to non-parametric processes like Arellano et al. (2017). We focus on the discretization of η_{it}, because the transitory component ε_{it} is i.i.d. The simulated values from the stochastic process are noisy, so we follow Arellano et al. (2017) in truncating the simulations at 3.3 age-dependent standard deviations around the mean.[18] We compare the performance of our discretization method with a textbook quantile binning method (Adda and Cooper, 2003).[19] This binning method also provides the initial guess for the grids, transition matrices, and initial distribution of our discretization routine.

[18] For the simulations from their earnings process, we use the publicly-available codes that accompany their publication.

[19] De Nardi et al. (2020) discretize the Arellano et al. (2017) process via binning to incorporate it into a lifecycle consumption saving model. In particular, their method adapts Adda and Cooper (2003) and uses simulation-

Figure 6: Statistical moments for two discretizations of Arellano et al. (2017)



Notes: the first panel is the negative log-likelihood at different grid sizes m which is related to the KL-divergence. The remaining three panels report the standard deviation, skewness, and kurtosis of earnings by age for the true model and discrete approximations. JM refers to our method, and Bin refers to the binning method.

Figure 6 visualizes the negative log-likelihood (lower negative log-likelihood implies a smaller KL-divergence, as in Equation (5)), as well as the unconditional moments of the earnings levels of the Arellano et al. (2017) process over the life-cycle, and the extent to which the discretized processes can replicate these moments. The first panel suggests that after $m = 18$ grid points, additional grid points only marginally improve the fit as measured through KL-divergence. This is consistent with the unconditional moments plots which show at $m = 18$ our discretization (red dashed line) coincides with the true model moments. Binning with $m = 18$ grid points is less accurate in terms of unconditional moments of earnings, in particular the kurtosis, suggesting it would take more grid points to achieve the same level of accuracy as our method.

4.4 Life-cycle model with the processes of Guvenen et al. (2021) and Arellano et al. (2017)

Next, we illustrate the importance of the choice of the discretization method for the earnings processes of Guvenen et al. (2021) and Arellano et al. (2017) through the lens of the life-cycle model. Figure 7 reports a selection of key models moments that are of interest to researchers using the canonical incomplete markets life-cycle model. These include moments on wealth

based binning, adding additional bins in the tails of the process. They discretize a re-estimated version of Arellano et al. (2017) that uses after-tax earnings, so our results are not directly comparable.

inequality, the welfare costs of risk, consumption volatility and the sensitivity of consumption to income fluctuations. We report these statistics for both our method and the binning methods for various choices of grid size m. The model does not allow for a closed-form solution, so to provide a benchmark, we solve the model numerically with a large age-dependent grid and transition matrix computed using the binning method.

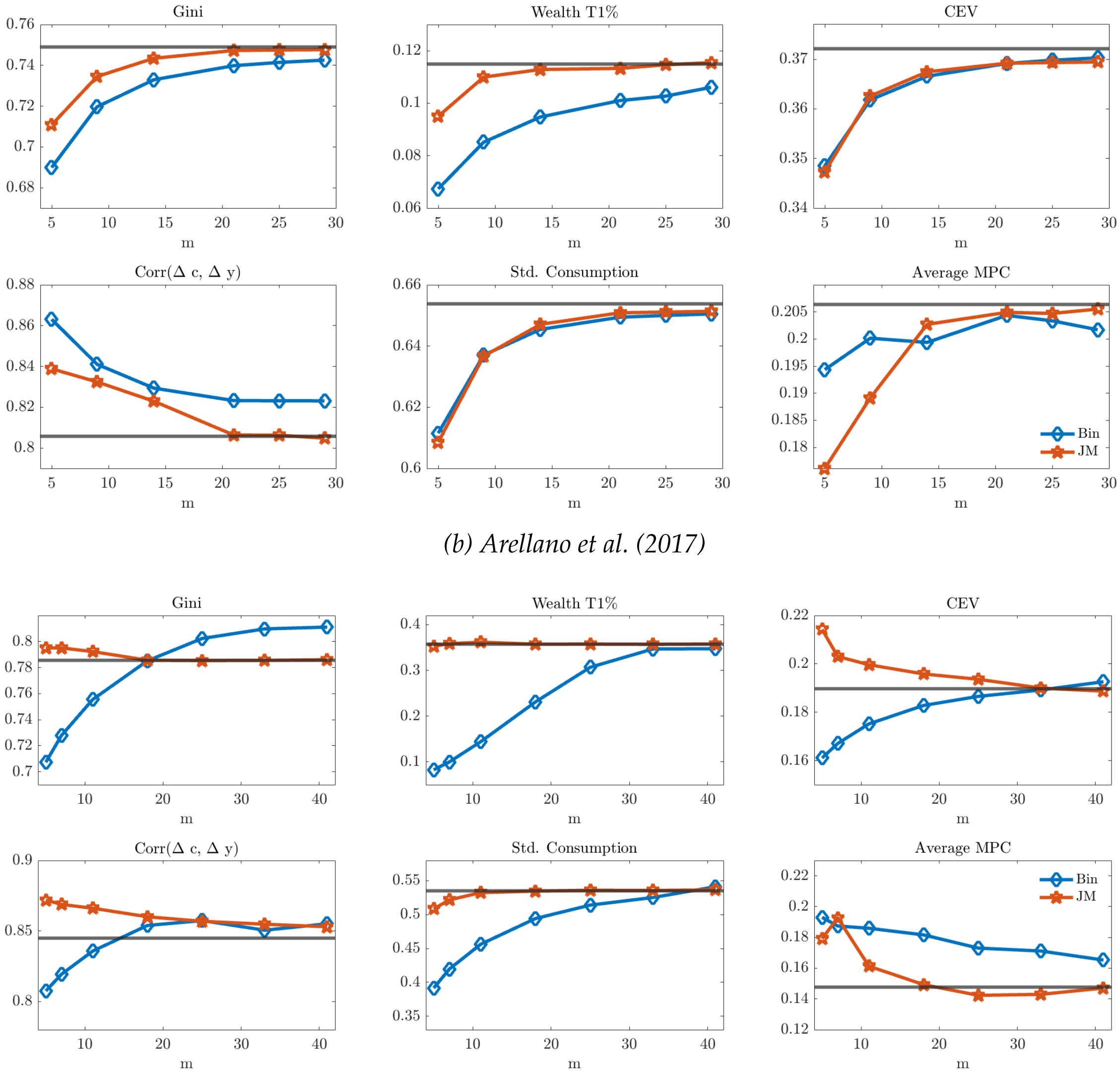
We draw three main lessons from this analysis. First, our method outperforms binning in terms of accuracy for a given grid size and converges faster to the benchmark moments. For the Guvenen et al. (2021) process, our method with $m = 21$ grid points provides an accurate approximation of the life-cycle model, which is consistent with the likelihood flattening at $m = 21$ in Figure 5. In contrast, binning with $m = 21$ grid points deviates considerably from the benchmark, and, in particular, it underestimates the amount of wealth inequality and overestimates the correlation between income and consumption changes. For the process of Arellano et al. (2017), our method is consistently closer to the benchmark than binning. Our method provides an accurate amount of wealth inequality with $m = 18$ grid points, but the correlation of consumption and income changes and welfare cost of risk take more grid points to converge to the benchmark.

Second, the processes of Arellano et al. (2017) and Guvenen et al. (2021) result in solutions to the lifecycle model that differ in economically significant ways. This is because of two main differences between Arellano et al. (2017) and Guvenen et al. (2021). Guvenen et al. (2021) model an earnings processes that features non-employment risk and a mixture distribution over innovations leading to highly volatile earnings conditional on working. The process of Arellano et al. (2017), on the other hand, does not model non-employment and features a longer right tail than Guvenen et al. (2021) and an increasing persistence of earnings over the life-cycle.

The economic consequences of the differing statistical descriptions of earnings are stark, in particular for the welfare cost of risk, wealth inequality, and marginal propensity to consume. The welfare cost of risk is 37.2% of lifetime consumption according to the Guvenen et al. (2021) process and 19.0% according to the Arellano et al. (2017) process, reflecting the downside risk of a non-employment spell as featured in Guvenen et al. (2021).

In terms of wealth inequality, for the Guvenen et al. (2021) process, the wealth share held by the top 1% wealthiest households is 11.5%, while, with the Arellano et al. (2017) process, it is 35.7%, driven by the larger positive skewness. Interestingly, the wealth Gini of 0.79 and the top 1% wealth share of 35.7 implied by the Arellano et al. (2017) process are close to the data counterparts of 0.78 and 33.5% from the 2007 Survey of Consumer Finances, and 0.77 and

Figure 7: Key model moments for two discretization methods



Notes: panels display Gini coefficient of wealth, wealth share of the top 1%, certainty equivalent variation of consumption, correlation between log consumption growth and log earnings growth, standard deviation of log consumption, and average marginal propensity to consume for various grid sizes m and two discrete approximation methods. The solid line is the benchmark model solution using binning with a large number of grid points ($m = 101$ for Guvenen et al. (2021) and 75 for Arellano et al. (2017)). JM refers to our method, and Bin refers to the binning method.

30.9% from the 2006 Panel Study of Income Dynamics (Krueger, Mitman, and Perri, 2016). This is in contrast to previous work, e.g., De Nardi and Fella (2017), that concludes that incomplete markets life-cycle models with non-Gaussian earnings processes generate too little top wealth inequality relative to the data.

We also find the marginal propensity to consume out of a transitory income shock differs substantially between the two processes, 20.6% with the Guvenen et al. (2021) process versus 14.8% with the Arellano et al. (2017) process. We conclude if one were using this class of models as a laboratory to study the positive or normative effects of fiscal policy, the results would differ substantially based on the choice of the earnings process, discretization method, and the number of grid points used.

As a third lesson, we conclude that selecting a discretization method based on an ad hoc moment accuracy criteria can lead to economically meaningful errors in the model solution. For example, if one were to use binning with $m = 21$ for the Guvenen et al. (2021) process based on the justification that if fits the life-cycle profile of the second moments of earnings (ignoring its poor fit to higher moments), one would understate wealth inequality, and overstate the correlation between consumption growth and income growth. Similarly, if one were to decide to use a binning method with $m = 18$ for the Arellano et al. (2017) process based on the same justification, one would understate top 1% wealth shares by more than 35% (23.1% versus 35.7%) and overstate the average marginal propensity to consume (18% versus 15%). This highlights the strength of a likelihood based approach to discretization, as which moments matter for the accuracy of the model solution is not obvious ex ante, and a likelihood approach considers the overall fit of the discretization to the true stochastic process without forcing the researcher to rely on moment selection.

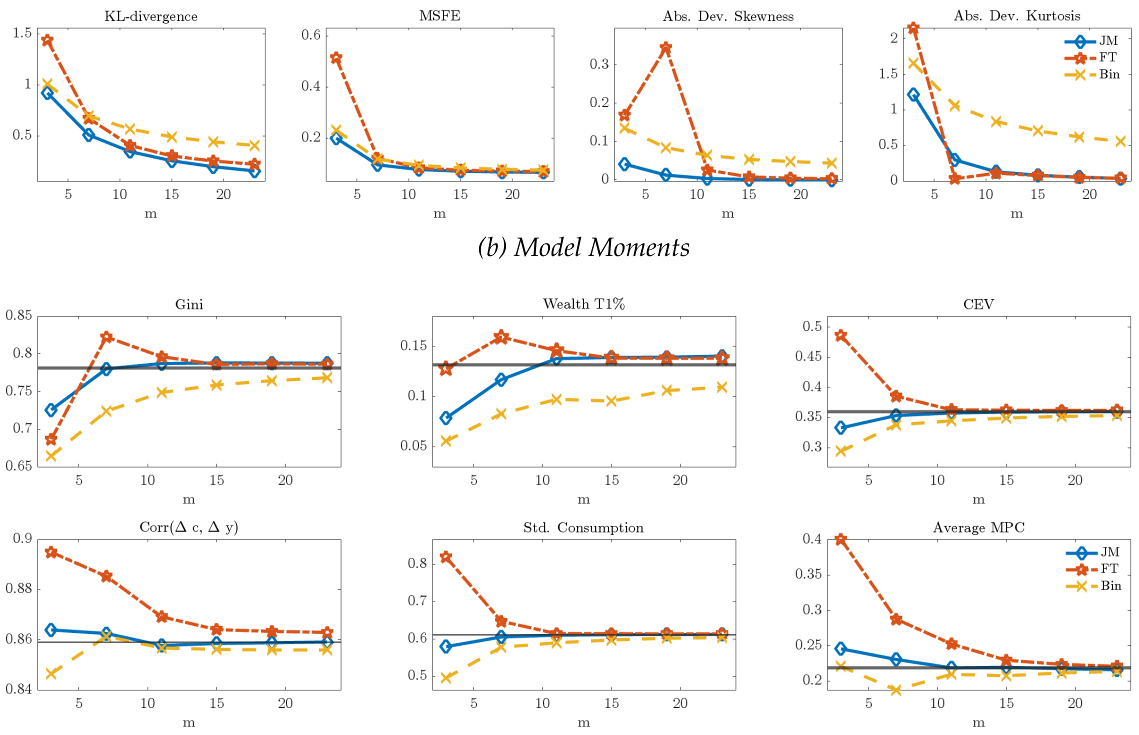
4.5 Canonical stochastic processes

To illustrate that discretization methods matter beyond the setting of highly non-linear processes like the ones presented above, this section considers the discretization of two simpler persistent-transitory earnings processes in the context of a life-cycle model. The first is the Guvenen et al. (2021) process without non-employment risk, effectively a permanent-transitory process where the innovations are drawn from normal mixtures, we refer to this as AR(1)-M. In the second simplified process, the innovations are drawn from a standard normal, we refer to this as AR(1). We parametrize the AR(1) process to have the same autocorrelation and variance as the AR(1)-M.[20] We compare our discretization method to the methods of Rouwenhorst (1995), Tauchen (1986), and Farmer and Toda (2017) for the AR(1) process, and to Farmer and

[20] For the AR(1) persistent-transitory process with Gaussian innovations, we use $\eta_t^i \sim N(0, \sigma_\eta^2)$ and $\varepsilon_t^i \sim N(0, \sigma_\varepsilon^2)$ where $\sigma_\eta^2 = p_\eta \sigma_{\eta,1}^2 + (1 - p_\eta)\sigma_{\eta,2}^2 + p_\eta \mu_{\eta,1}^2 + (1 - p_\eta)\mu_{\eta,2}^2$ and similar for σ_ε^2. For both processes, we discretize the persistent component and add an equal-quantile discretization of the transitory component to the model.

Toda (2017) and the binning method of Adda and Cooper (2003) for the AR(1)-M process. The results for the AR(1) are summarized in Figure D1 in Appendix D.[21]

Figure 8: Moments of the AR(1)-Mixture Process


(a) Statistical Moments

(b) Model Moments

Notes: Panel (a) plots the KL-divergence, the mean-squared forecast error (MSFE), skewness, and kurtosis of earnings for different grid sizes and approximation methods. Panel (b) displays the Gini coefficient of wealth, wealth share of the top 1%, certainty equivalent variation of consumption, correlation between log consumption growth and log earnings growth, standard deviation of log consumption, and average marginal propensity to consume for different grid sizes and approximation methods. The solid line is the benchmark model solution using binning with a large number of grid points ($m = 190$). JM refers to our method, Bin refers to the binning method, and FT refers to the Farmer and Toda (2017) method.

Focusing on the AR(1)-M results in Figure 8, we confirm the results from the previous subsection. Binning again underestimates wealth inequality and the welfare cost of risk. The moment matching method of Farmer and Toda (2017) may perform better or worse than the binning approach depending on the model moments of interest and the number of grid points. This happens because the Farmer and Toda (2017) method tries to match conditional moments

[21] For the AR(1) process, the methods of Farmer and Toda (2017) and Rouwenhorst (1995) are designed to match moments of the earnings process, and this helps them to match the standard deviation of log consumption and the certainty equivalent value of consumption with few grid points. Other model moments take longer to converge, and for those, the HMM discretizations perform similar. The method of Tauchen (1986) converges more slowly to the true model moments.

at each grid point, but this is not possible at the first and last grid point, hence the performance suffers when the total number of grid points is small. Our method performs better in terms of distance to the benchmark and converges more rapidly than the alternative discretization methods.

5 Conclusion

This paper proposes a novel finite-state Markov chain approximation method, based on minimizing the information loss between the true stochastic process and a Hidden Markov Model (HMM). Finite-state Markov chain approximations are inherently misspecified models, and minimizing the Kullback-Leibler (KL) divergence is a standard objective in the misspecified model literature. We demonstrate that this approach is consistent in our setting: under certain assumptions, and with a sufficient number of hidden states, the information loss between the approximating HMM and the true stochastic process can be made arbitrarily small. Our discretization method is applicable to a broad class of stochastic processes and provides both an optimally selected grid and a transition probability matrix. This optimal grid is especially powerful in the case of correlated multivariate processes, as it avoids the use of tensor grids.

We apply and compare our method in two applications. The first application is an asset-pricing model with stochastic volatility, which, as demonstrated by De Groot (2015), has a closed-form analytical solution. This analytical solution serves as a benchmark for evaluating the accuracy of solutions derived from different discretization methods. The second application examines the impact of the discretization method on solutions derived from a life-cycle model featuring various earnings processes, including the recently proposed non-linear and non-Gaussian processes of Arellano et al. (2017) and Guvenen et al. (2021). We find that the choice of discretization method significantly affects key economic outcomes, including the welfare cost of risk, the marginal propensity to consume, and wealth inequality measures. In both applications, our method provides accurate model solutions using fewer grid points than existing methods.

Discretized stochastic processes have many more applications than the ones we use to benchmark our method. The econometric literature has shown stochastic processes featuring non-linearities, excess skewness and kurtosis provide a better description of the data. Our method provides a tool for the use of richer statistical processes in structural economic models.

Supplementary Appendix to "Finite State Markov-Chain Approximations: A Hidden Markov Approach"

A Proof of Main Theorem

A.1 Preliminaries, notation and existing results

Denote the L^∞ norm as $||f||_{L^\infty} = \inf\{a \geq 0 : \lambda(\{x \in \mathbb{R}^k : |f(x)| > a\} = 0\} < \infty$. Denote the set of continuous functions with support on $\mathbb{R}^k$ by C. Define the set of locally compact functions (i.e., continuous functions that vanish at infinity) by

$$C_0 = \left\{ f \in C : \, \forall \epsilon > 0, \, \exists \text{ a compact } \mathbb{K} \in \mathbb{R}^k, \text{ such that } |f(x)| < \epsilon \text{ for all } x \notin \mathbb{K} \right\}.$$

We impose the following assumptions on the true process $f(\mathbf{y})$ and approximating model $p(\mathbf{y}, \theta)$:

(A1) $\mathbf{y} = \{y_t\}_{t=1}^T$ has a data generating process characterized by $f(\mathbf{y})$, $y_t \in \mathbb{R}^k$, that is first-order Markov and stationary, that is,

$$f(y_t|y_{t-1}, ..., y_1) = f(y_t|y_{t-1}),$$

and

$$f(y_{t+l}|y_{t+l-1}) = f(y_t|y_{t-1}) \quad \forall l \in \mathbb{N}.$$

(A2) $f(y_t|y_{t-1}) \in C_0$.

(A3) $\log f(y_t|y_{t-1})$ and $f(y_t|y_{t-1})$ are differentiable in y_{t-1}.

(A4) $\log f(y_t|y_{t-1})$ is locally Lipschitz continuous in y_{t-1}.

(A5) $p(\mathbf{y}; \theta_m)$ is characterized by:

$$\begin{aligned} &y_t|x_t = \mu_m(x_t) + \text{diag}(\sigma_m)\varepsilon_t, \quad \varepsilon_t \sim N(0, I_k), \\ &x_{t+1}|x_t \sim \Pi_{ij,m} \end{aligned}$$

with parameters $\theta_m = (\mu_m, \Pi_m, \sigma_m)$, and $x_t \in \{1, ...m\}$ a latent state evolving according to a first-order Markov process with transition probability matrix Π_m. Denote the conditional distribution by $p(y_t|y_{t-1}, ..., y_1; \theta_m) \in C_0$.

We denote the class of basic densities that we use in our approximation class as

$$\mathcal{M}_m^\phi = \left\{ f_m : f_m(x) = \sum_{i=1}^m \alpha_m(i) \cdot \phi\left(x, \mu_m(i), \sigma_m(i)\right), \quad \mu_m(i) \in \mathbb{R}^k,\ \sigma_m(i) \in \mathbb{R}_+^k,\ \alpha_m \in \Delta^{m-1},\ i \in [m] \right\}$$

where $\Delta^{m-1} = \{\alpha \in \mathbb{R}^m : \sum_{i=1}^m \alpha_i = 1,\ \alpha_m(i) \geq 0\ \forall\ i \in [m]\}$, and ϕ is the Gaussian probability density function. Denote $\theta_m = (\mu_m, \sigma_m)$.

Theorem 1 (Theorem 5a in Nguyen et al., 2020). *Assume f is a pdf in C_0, then there exists a sequence $\{f_m^0\}$ ($f_m^0 \in \mathcal{M}_m^\phi$) such that*

$$\lim_{m \to \infty} ||f - f_m^0||_{L^\infty} = 0.$$

Corollary 1. *By the Dominated Convergence Theorem, Theorem 1 implies* $\lim_{m\to\infty} D^{KL}(f||f_m^0) = 0$*, since $f_m^0 > 0$ whenever $f > 0$.*

A.2 Bound on the KL divergence of a Gaussian Mixture in a Given Grid

In Lemma 1, we show the KL divergence between a Gaussian mixture and a function f converges to 0 as $m \longrightarrow \infty$, even when the Gaussian mixture takes a choice of grid points $\tilde{\mu}_m$ and standard deviation $\tilde{\sigma}_m$ that may not be same as μ_m^0 and σ_m^0 of Theorem 1.

Lemma 1. *Let f_0^m and f as defined in Theorem 1, with f_m^0 characterized by $(\alpha_m^0, \mu_m^0, \sigma_m^0)$. Let $\tilde{f}_m$ be a mixture sequence with the same weights α_m^0 as f_m^0, but different means $\tilde{\mu}_m(i) \in \mathbb{R}^k$ and variances $\tilde{\sigma}_m \geq \tau > 0$. If $\tilde{\mu}_m$ forms a dense subset of $\mathbb{R}^k$ as $m \longrightarrow \infty$, and $\tilde{\sigma}_m$ goes to zero at the same rate as σ_m^0, then*

$$\lim_{m \to \infty} D^{KL}(f||\tilde{f}_m) = 0.$$

Proof: Given that we are comparing two Gaussian mixtures with the same mixture weights, from Do (2003), we obtain the following upper bound:

$$D^{KL}\left(f_m^0||\tilde{f}_m\right) \leq \sum_{i=1}^m \alpha_i D^{KL}\left(f_m^{0,i}||\tilde{f}_m^i\right),$$

where we denote the ith component of the mixture distribution with superscript i. By properties of the Gaussian distribution, we have:

$$D^{KL}\left(f_m^{0,i}||\tilde{f}_m^i\right) \leq \frac{1}{2}\left\{(\mu_m^0(i)-\tilde{\mu}_m(i))'\tilde{\Sigma}_m^{-1}(\mu_m^0(i)-\tilde{\mu}_m(i))+\text{tr}(\tilde{\Sigma}_m^{-1}\Sigma_m^0)-k+\ln\frac{|\tilde{\Sigma}_m|}{|\Sigma_m^0|}\right\}.$$

where Σ_m is a diagonal with $((\sigma_m^i)^2)$, $i=1,...k$ on the diagonal. Using that $\sum \alpha_i = 1$:

$$D^{KL}\left(f_m^0||\tilde{f}_m\right) \leq \frac{1}{2}\max_i\left\{(\mu_m^0(i)-\tilde{\mu}_m(i))'\tilde{\Sigma}_m^{-1}(\mu_m^0(i)-\tilde{\mu}_m(i))+\text{tr}(\tilde{\Sigma}_m^{-1}\Sigma_m^0)-k+\ln\frac{|\tilde{\Sigma}_m|}{|\Sigma_m^0|}\right\}.$$

Under the assumptions that $\tilde{\mu}_m$ forms a dense subset of $\mathbb{R}^k$ as $m \to \infty$, and $\tilde{\sigma}_m$ goes to zero at the same rate as σ_m^0, this object converges to 0 as $m \to \infty$. We can therefore write $f_m^0 - \kappa(m) \leq \tilde{f}_m \leq f_m^0 + \kappa(m)$, with $\kappa(m) \to 0$ as $m \to \infty$ by Pinsker's Inequality. Hence, we can write:

$$D^{KL}(f||f_m^0+\kappa(m)) \leq D^{KL}(f||\tilde{f}_m) \leq D^{KL}(f||f_m^0-\kappa(m))$$

We have

$$\begin{aligned}
D^{KL}(f||f_m^0 \pm \kappa(m)) &= \int f \ln\left(\frac{f}{f_m^0 \pm \kappa(m)}\right) dy \\
&= \int f \ln\left(\frac{f}{f_m^0}\cdot\frac{f_m^0}{f_m^0 \pm \kappa(m)}\right) dy \\
&= D^{KL}(f||f_m^0) + \mathbb{E}_f\left[\ln\left(\frac{f_m^0}{f_m^0 \pm \kappa(m)}\right)\right]
\end{aligned}$$

where $\mathbb{E}_f\left[\ln\left(\frac{f_m^0}{f_m^0\pm\kappa(m)}\right)\right]$ goes to zero because $\kappa(m) \to 0$ by the Dominated Convergence Theorem, and $D^{KL}(f||f_m^0)$ goes to zero by Theorem 1. □

A.3 Properties of the HMM

Lemma 2. *Let $p(\mathbf{y};\theta)$ as in Assumption (A5) and let there be a sequence $\{\mu_m(i)\}$ and $\{\sigma_m\}$ such that: (i) define $l = \text{argmax}_i\{\phi_i(y_{t-1}) : i \in [m]\}$, $\exists\ \eta(\sigma_m)$ s.t. $\phi_i(y_{t-1}) < \eta(\sigma_m)/(m-1)$ in $i \neq l$ where $\eta(\sigma_m) \downarrow 0$ as $\sigma_m \to 0$, and (ii) $K(\sigma_m) = \sum \phi_i(y_{t-1})$ satisfies $K(\sigma_m) > \varepsilon > 0$ as $m \to \infty$. For $h \geq 1$, $\log p(y_t|y_{t-1},...,y_{t-h},...,y_1;\theta)$ is Lipschitz continuous in y_{t-h}, and for $h \geq 2$, the Lipschitz constant goes to zero as m grows large.*

Proof. First of all, $\log p(y_t|y_{t-1},...,y_1;\theta)$ is everywhere differentiable in y_{t-1}. Therefore, to show Lipschitz continuity, we have to show its derivative is bounded.

$$\begin{aligned} p(y_t|y_{t-1},y_{t-2},...,y_1;\theta) &= \sum_{j=1}^{m}\left[p(y_t|x_t=j)\sum_{i=1}^{m}(P(x_t=j|x_{t-1}=i)P(x_{t-1}=i|y_{t-1},y_{t-2},...,y_1;\theta))\right] \\ &= \sum_{j=1}^{m}\left[\phi_j(y_t)\sum_{i=1}^{m}\left(\Pi_{ij}P(x_{t-1}=i|y_{t-1},y_{t-2},...,y_1)\right)\right] \end{aligned}$$

and $p(y_1)=\sum_{j=1}^{m}\left[\phi_j(y_1)\delta_{1i}\right]$. Here

$$P(x_t=i|y_t,y_{t-1},...,y_1)=\frac{\phi_i(y_t)\sum_{j=1}^{m}\Pi_{ji}P(x_{t-1}=j|y_{t-1},...,y_1)}{\sum_{i=1}^{m}\phi_i(y_t)\sum_{j=1}^{m}\Pi_{ji}P(x_{t-1}=j|y_{t-1},...,y_1)}:=\frac{A_{it}}{B_t}$$

where $P(x_1=i|y_1)=\delta_{1i}\phi_i(y_1)/\sum_{i=1}^{m}\delta_{1i}\phi_i(y_1)$.

We need to evaluate $\partial\log p(y_t|y_{t-1},y_{t-2},...,y_1)/\partial y_{t-h}$. For $h\geq 1$, we have:

$$\frac{\partial\log p(y_t|y_{t-1},y_{t-2},...,y_1;\theta)}{\partial y_{t-h}}=\frac{1}{p(y_t|y_{t-1},y_{t-2},...,y_1;\theta)}\frac{\partial p(y_t|y_{t-1},y_{t-2},...,y_1;\theta)}{\partial y_{t-h}}, \tag{A.1}$$

where

$$\frac{\partial p(y_t|y_{t-1},y_{t-2},...,y_1;\theta)}{\partial y_{t-h}}=\sum_{i=1}^{m}\phi_i(y_t)\sum_{j=1}^{m}\Pi_{ji}\frac{\partial P(x_{t-1}=j|y_{t-1},...,y_1)}{\partial y_{t-h}} \tag{A.2}$$

with, for $h=1$:

$$\frac{\partial P(x_{t-1}=i|y_{t-1},...,y_1)}{\partial y_{t-1}}= \frac{B_{t-1}\phi_i'(y_{t-1})\sum_{j=1}^{m}\Pi_{ji}P(x_{t-2}=j|y_{t-2},...,y_1)-A_{it-1}\sum_{l=1}^{m}\phi_l'(y_{t-1})\sum_{j=1}^{m}\Pi_{jl}P(x_{t-2}=j|y_{t-1},...,y_1)}{B_{t-1}^2} \tag{A.3}$$

The expression in Equation (A.1) is bounded and therefore Lipschitz in y_{t-1}. First of all, $\frac{1}{p(y_t|y_{t-1},y_{t-2},...,y_1)}$ is bounded from below and finite. A_{it} and B_t are finite, and $\phi'(\cdot)$ is bounded because the Gaussian distribution itself is Lipschitz continuous, so boundedness of the expressions follows.

For $h \geq 2$:

$$\frac{\partial P(x_{t-1}=i|y_{t-1},\dots,y_1)}{\partial y_{t-h}} = \\ \frac{B_{t-1}\phi_i(y_{t-1})\sum_{j=1}^m \Pi_{ji}\frac{\partial P(x_{t-2}=j|y_{t-2},\dots,y_1)}{\partial y_{t-h}} - A_{it-1}\sum_{l=1}^m \phi_l(y_{t-1})\sum_{j=1}^m \Pi_{jl}\frac{\partial P(x_{t-2}=j|y_{t-2},\dots,y_1)}{\partial y_{t-h}}}{B_{t-1}^2} \tag{A.4}$$

and, as this is recursive, we need the expression for $\partial P(x_1=i|y_1)/\partial y_1$, which is given by:

$$\frac{\partial P(x_1=i|y_1)}{\partial y_1} = \frac{\delta_{1i}\phi_i'(y_1)\sum_{j=1}^m \delta_{1j}\phi_j(y_1) - \delta_{1i}\phi_i(y_1)\sum_{j=1}^m \delta_{1j}\phi_j'(y_1)}{\left(\sum_{j=1}^m \delta_{1j}\phi_j(y_1)\right)^2}$$

Define $C_{it-1} := \phi_i(y_{t-1})\sum_{j=1}^m \Pi_{ji}\frac{\partial P(x_{t-2}=j|y_{t-2},\dots,y_1)}{\partial y_{t-h}}$ and $D_t = \sum_i C_{it-1}$. We rewrite Equation (A.4) as $(B_{t-1}C_{it-1} - A_{it-1}D_{t-1})/B_{t-1}^2$.

By our assumptions, there are two cases. If we are in the case that i and y_{t-1} are such that $\phi_i(y_{t-1}) < \eta(\sigma_m)/(m-1)$, denote $\hat{C}_i = \frac{\eta(\sigma_m)}{m-1}\sum_{j=1}^m \Pi_{ji}\frac{\partial P(x_{t-2}=j|y_{t-2},\dots,y_1)}{\partial y_{t-h}}$ and $\hat{A}_i = \frac{\eta(\sigma_m)}{m-1}\sum_{j=1}^m \Pi_{ji}P(x_{t-1}=i|y_{t-1},\dots,y_1) < \frac{\eta(\sigma_m)}{m-1}$. We have $B_{t-1}C_{it-1} < B_{t-1}\hat{C}_i$ and $A_{it-1}D_{t-1} < \hat{A}_i D_{t-1}$. Both $\hat{A}_i$ and $\hat{C}_i$ are decreasing in m, so in this case Equation (A.4) converges to 0 as $m \to \infty$. On the other hand, if i is such that $\phi_i(y_{t-1}) > \eta(\sigma_m)/(m-1)$, we have $B_{t-1}C_{i,t-1} - A_{i,t-1}D_{t-1}= (B_{t-1} - A_{i,t-1} + A_{i,t-1})C_{i,t-1} - A_{i,t-1}(D_{t-1} - C_{i,t-1} + C_{i,t-1}) = (B_{t-1} - A_{i,t-1})C_{i,t-1} - A_{i,t-1}(D_{t-1} - C_{i,t-1})$, with $B_{t-1} - A_{i,t-1} < \eta(\sigma_m)\sum_{k\neq i}\sum_{j=1}^m \Pi_{jk}P(x_{t-1}=i|y_{t-1},\dots,y_1)$ and $D_{t-1} - C_{i,t-1} < \eta(\sigma_m)\sum_{k\neq i}\sum_{j=1}^m \Pi_{jk}\frac{\partial P(x_{t-2}=j|y_{t-2},\dots,y_1)}{\partial y_{t-h}}$. Both terms in the numerator are decreasing towards zero in m. Note that B_{t-1} is bounded from below. Thus, in both cases, the derivative in Equation (A.4) decreases in m, so the Lipschitz coefficient of $\log p(y_t|y_{t-1},\dots,y_1;\theta_m)$ to y_{t-h}, $h \geq 2$ goes to zero as $m \to \infty$. □

Remark. This result is related to Le Gland and Mevel (2000) who show that Hidden Markov Models have exponential forgetting, which in this context means that $\partial p(y_t|y_{t-1},\dots,y_1;\theta)/\partial y_{t-h}$ declines in h at an exponential rate. However, for our result, exponential forgetting is not sufficient, because we need the Lipschitz constant not only to decline if the history is longer ago, but the Lipschitz constant also needs to become smaller as m grows larger, which is what we showed with Lemma 6. Intuitively, this result says that as the number of states grows large enough, and the filter becomes better, our HMM becomes approximately first-order Markov.

A.4 The KL divergence is a function of all conditional KL divergences

Lemma 3. *Under assumption (A1) and (A5), if $D^{KL}(f(y_t|y_{t-1})||p(y_t|y_{t-1},...,y_1;\theta_m))$ is bounded and can be made arbitrarily small for any sequences $\{y_k\}_{k=1}^{t-1}$ for all t, then $D^{KL}(f(\mathbf{y})||p(\mathbf{y};\theta_m))$ is also bounded and can be made arbitrarily small by picking m large.*

Proof. The first-order Markov assumption on the true DGP of $\mathbf{y}$ implies $f(y_t|y_0,y_1,...,y_{t-1}) = f(y_t|y_{t-1})$, such that we can write

$$f(\mathbf{y}) = f(y_1)\prod_{t=2}^{T} f(y_t|y_{t-1})$$

where $f(y_1)$ denotes some initial distribution.

Hidden Markov Models do not satisfy the Markov property for $\mathbf{y}$. We have

$$p(\mathbf{y};\theta) = p(y_1;\theta)\prod_{t=2}^{T} p(y_t|y_{t-1},y_{t-2},...,y_1;\theta)$$

with $p(y_1;\theta)$ again the initial distribution.

The KL divergence for T observations is given by

$$\int f(\mathbf{y})\log\left(\frac{f(\mathbf{y})}{p(\mathbf{y};\theta)}\right)d\mathbf{y} =$$
$$\int\int\cdots\int f(y_1)\prod_{t=2}^{T} f(y_t|y_{t-1})\log\left(\frac{f(y_1)\prod_{t=2}^{T} f(y_t|y_{t-1})}{p(y_1|\theta)p(y_2|y_1;\theta)\cdots p(y_T|y_{T-1},...,y_1;\theta)}\right)dy_T dy_{T-1}...dy_1$$

Straightforward algebra shows the KL divergence can be written as:

$$\int f(\mathbf{y})\log\left(\frac{f(\mathbf{y})}{p(\mathbf{y};\theta)}\right)d\mathbf{y} =$$
$$D^{KL}(f(y_1)||p(y_1|\theta))) + \sum_{t=2}^{T}\int f(\mathbf{y}_{1:t-1})D^{KL}(f(y_t|y_{t-1})||p(y_t|y_{t-1},...,y_1;\theta))d\mathbf{y}_{1:t-1}$$

Note that $f(y_{1:t-1})$ integrates to 1 and D^{KL} is non-negative. This implies if $D^{KL}(f(y_t|y_{t-1})||p(y_t|y_{t-1},y_{t-2},...,y_1;\theta)) \rightarrow 0$ for all $y_t,...,y_1$, and all $t > 1$, then $D^{KL}(p(\mathbf{y};\theta)||f(\mathbf{y})) \rightarrow 0$. □

A.5 Proof of Main Theorem

Main Theorem. *Under assumptions (A1)-(A5), given a sufficiently large number of grid points m, there exist a set of grid points μ_m, variance $\sigma_m \geq \tau > 0$ and transition probability matrix Π_m, collected in $\theta_m = (\mu_m, \Pi_m, \sigma_m)$ such that the KL divergence between $f(\mathbf{y})$ and $p(\mathbf{y};\theta)$, given by*

$$D^{KL}(f(\mathbf{y})||p(\mathbf{y};\theta)) = \int f(\mathbf{y}) \log \frac{f(\mathbf{y})}{p(\mathbf{y};\theta)} d\mathbf{y},$$

can be made arbitrarily small.

Proof. Let $f(y_t|y_{t-1} = \mu_m(i))$ denote the true conditional distribution and $p^0(y_t|y_{t-1} = \mu_m(i);\theta_m))$ the approximating mixture as in Lemma 1. By Lemma 1, $\lim_{m\to\infty} D^{KL}((f(y_t|y_{t-1} = \mu_m(i))||p^0(y_t|y_{t-1} = \mu_m(i);\theta_m)) = 0$. Let $p_m^0 = p^0(y_t|y_{t-1} = \mu_m(i);\theta_m)$, $p_m = p(y_t|y_{t-1} = \mu_m(i), ..., y_1;\theta_m)$, $f = f(y_t|y_{t-1} = \mu_m(i))$ and define $\kappa(m) = |p_m^0 - p_m|$. By Lemma 2, $\kappa(m) \longrightarrow 0$ as $m \longrightarrow \infty$. Note $p_m^0 - \kappa(m) \leq p_m \leq p_m^0 + \kappa(m)$. This implies

$$\int f \ln\left(\frac{f}{p_m^0 + \kappa(m)}\right) dy \leq \int f \ln\left(\frac{f}{p_m}\right) dy \leq \int f \ln\left(\frac{f}{p_m^0 - \kappa(m)}\right) dy,$$

or, equivalently,

$$D^{KL}(f||p_m^0 + \kappa(m)) \leq D^{KL}(f||p_m) \leq D^{KL}(f||p_m^0 - \kappa(m)).$$

Moreover,

$$D^{KL}(f||p_m^0 \pm \kappa(m)) = D^{KL}(f||p_m^0) + \mathbb{E}_f\left[\ln\left(\frac{p_m^0}{p_m^0 \pm \kappa(m)}\right)\right].$$

Note both $D^{KL}(f||p_m^0)$ and $\mathbb{E}_f\left[\ln\left(\frac{p_m^0}{p_m^0 \pm \kappa(m)}\right)\right]$ go to zero as $m \longrightarrow \infty$.

Hence, $D^{KL}(f(y_t|y_{t-1} = \mu_m(i)), p(y_t|y_{t-1} = \mu_m(i), ..., y_1;\theta_m))$ approaches zero when m becomes large.

Next, we show that when the KL-divergence of the distribution conditional on y_{t-1} being one of the m gridpoints, i.e., in $y_{t-1} = \mu_m(i)$, becomes arbitrarily small as m becomes large, then the KL divergence of distributions conditional on any $y_{t-1}, y_{t-2}, ..., y_1$ also becomes small.

By Assumptions (A3)-(A4), and Lemma 2, we have

$$\begin{aligned}
&D^{KL}(f(y_t|y_{t-1}=y)||p(y_t|y_{t-1}=y,y_{t-2},...,y_1;\theta_m)) \leq \\
&D^{KL}(f(y_t|y_{t-1}=\mu_m(i))||p(y_t|y_{t-1}=\mu_m(i),y_{t-2},...,y_1;\theta_m)) + ... \\
&O(K_p|y-\mu_m(i)|, K_f|y-\mu_m(i)|, K_{\log f}|y-\mu_m(i)|)
\end{aligned}$$

Here K_p denotes the Lipschitz coefficient of $p(y_t|y_{t-1},...,y_1;\theta_m)$ in y_{t-1}, K_f denotes the Lipschitz coefficient of $f(y_t|y_{t-1})$ in y_{t-1}, and $K_{\log f}$ the Lipschitz coefficient for $\log f(y_t|y_{t-1})$ in y_{t-1}. Note, the relevant $\mu_m(i)$ to consider is the one closest to y. $O(K|y-\mu_m(i)|, K_f|y-\mu_m(i)|, K_{\log f}|y-\mu_m(i)|)$ denotes some function increasing in the terms in between brackets. These three terms converge to zero as the grid points become dense (as assumed in Lemma 1), because the distance to the nearest grid point $|y-\mu_m(i)|$ goes to zero, hence $O(\cdot)$ will also converge to zero.

By Lemma 2, if $D^{KL}(f(y_t|y_{t-1})||p(y_t|y_{t-1},\{y_{t-k}\}_{k=2}^{t-1};\theta_m))$ can be made arbitrarily small for m large enough, then $D^{KL}(f(y_t|y_{t-1})||p(y_t|y_{t-1},\{\tilde{y}_{t-k}\}_{k=2}^{t-1};\theta_m)$ is arbitrarily small for any other sequence $\{\tilde{y}_{t-k}\}_{k=1}^{t-1}$, because $\log p(y_t|y_{t-1},\{y_{t-k}\}_{k=2}^{t-1};\theta_m))$ is Lipschitz continuous in $\{y_{t-k}\}_{k=2}^{t-1}$ with a coefficient that goes to zero as m becomes large. This implies the KL divergence for all $D^{KL}(f(y_t|y_{t-1})||p(y_t|y_{t-1},\{\tilde{y}_{t-k}\}_{k=2}^{t-1};\theta_m)$ goes to zero when m becomes large, for all $t \geq 2$.

For the initial distribution, the parameters δ_{1i} function as mixture weights, where $p(y_1) = \sum_{j=1}^{m} \phi_j(y_1)\delta_{1i}$ is also a mixture of Gaussians. Applying Lemma 4 shows this KL divergence is also bounded and can be made arbitrarily small.

Applying Lemma 3 to the conditional KL divergences concludes the proof. □

B Discretization of a VAR process

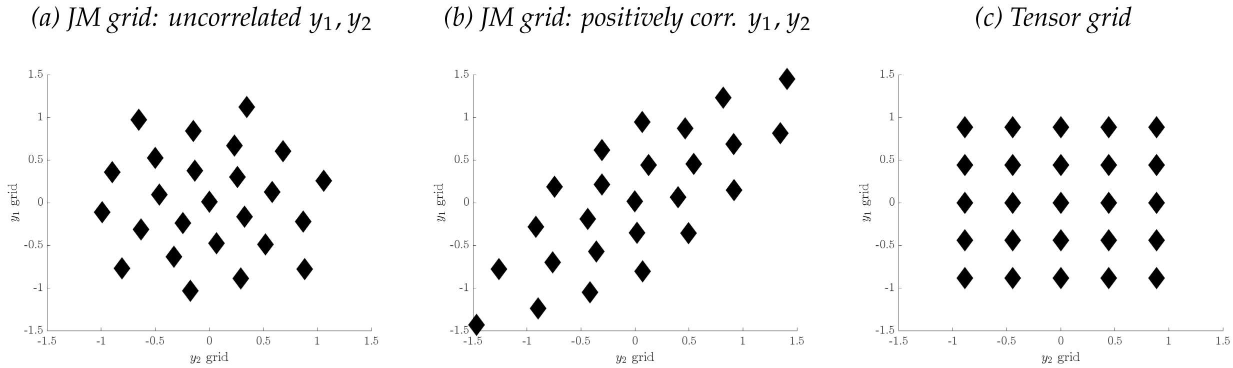
In this Appendix, we demonstrate the performance of our method for discretizing a bivariate VAR model of the form

$$y_{1,t} = \beta_{11} y_{1,t-1} + \beta_{12} y_{1,t-1} + \varepsilon_{1,t} \tag{B.1}$$

$$y_{2,t} = \beta_{21} y_{1,t-1} + \beta_{22} y_{2,t-1} + \varepsilon_{2,t}, \tag{B.2}$$

where $\varepsilon_t \sim N(0, \Sigma)$.

Figure B1: Visualisation of optimal grid for two different parametrizations of the data generating process in Equation (B.1)*, $m = 25$.*



(a) JM grid: uncorrelated y_1, y_2 *(b) JM grid: positively corr. y_1, y_2* *(c) Tensor grid*

For both parametrizations, $\Sigma = \text{diag}(0.1)$. Panel (a)/(c): $\beta_{11} = 0.7$, $\beta_{12} = 0$, $\beta_{21} = 0$, $\beta_{22} = 0.7$. Panel (b)/(c): $\beta_{11} = 0.7$ $\beta_{12} = 0.2$, $\beta_{21} = 0.2$, $\beta_{22} = 0.7$. JM stands for Janssens-McCrary.

We consider two different parametrizations but keep the grid size fixed to $m = 25$ to show how our discretization method optimally selects the grid. The optimal grids are visualized in Figure B1. As can be seen, as opposed to a tensor grid, our optimal grid incorporates the structure of the process into the grid. For example, in a VAR model where both variables are positively correlated ($\beta_{12} = \beta_{21} > 0$), if y_1 is large, y_2 is also likely large. Figure B1b shows how this is reflected in our optimal grid, while a standard tensor grid as in Figure B1c does not reflect this dependence.

C Asset Pricing Model with Stochastic Volatility

C.1 A closed-form solution

From De Groot (2015), we obtain closed-form expressions for the asset pricing model with stochastic volatility presented in Equations (21)-(22). The solution for the price-dividend ratio is given by:

$$v_t = \sum_{i=1}^{\infty} \beta^i \exp(B_i y_t + C_i \bar{\eta} + D_i(\eta_t - \bar{\eta}) + H_i),$$

where

$$\begin{aligned}
B_i &= \left(\frac{1-\gamma}{1-\rho}\right) \rho(1-\rho^i) \\
C_i &= \frac{1}{2}\left(\frac{1-\gamma}{1-\rho}\right)^2 \left(i - 2\rho\frac{1-\rho^i}{1-\rho} + \rho^2 \frac{1-\rho^{2i}}{1-\rho^2}\right) \\
D_i &= \frac{\rho_\eta}{2}\left(\frac{1-\gamma}{1-\rho}\right)^2 \left(\phi_1 + \phi_2 \rho_\eta \rho_\eta^{i-1} + \phi_3 \rho^{i-1} + \phi_4 \rho^{2(i-1)}\right) \\
H_i &= F_i \omega^2
\end{aligned}$$

where

$$\begin{aligned}
F_i = \frac{1}{8}\left(\frac{1-\gamma}{1-\rho}\right)^4 \Bigg(& i\phi_1^2 + \phi_2^2 \frac{1-\rho_\eta^{2i}}{1-\rho_\eta^2} + \phi_3^2 \frac{1-\rho^{2i}}{1-\rho^2} + \phi_4^2 \frac{1-\rho^{4i}}{1-\rho^4} \ldots \\
& \ldots + 2\phi_1\phi_2 \frac{1-\rho_\eta^i}{1-\rho_\eta} + 2\phi_1\phi_3 \frac{1-\rho^i}{1-\rho} + 2\phi_1\phi_4 \frac{1-\rho^{2i}}{1-\rho^2} + 2\phi_2\phi_3 \frac{1-(\rho_\eta\rho)^i}{1-\rho_\eta\rho} \ldots \\
& \ldots + 2\phi_2\phi_4 \frac{1-(\rho_\eta\rho^2)^i}{1-\rho_\eta\rho^2} + 2\phi_3\phi_4 \frac{1-\rho^{3i}}{1-\rho^3}\Bigg)
\end{aligned}$$

and

$$\begin{aligned}
\phi_1 &= \frac{1}{1-\rho_\eta}, \quad \phi_2 = \frac{-\rho_\eta(\rho_\eta + \rho)(1-\rho)^2}{(\rho^2 - \rho_\eta)(\rho - \rho_\eta)(1-\rho_\eta)}, \\
\phi_3 &= \frac{-2\rho^2}{\rho - \rho_\eta}, \quad \phi_4 = \frac{\rho^4}{\rho^2 - \rho_\eta}.
\end{aligned}$$

The conditional expected return on equity is defined as

$$\mathbb{E}_t R^e_{t+1} = \mathbb{E}_t \left(\frac{d_{t+1} + p_{t+1}}{p_t} \right) = \frac{\mathbb{E}_t \exp(y_{t+1}) + \mathbb{E}_t v_{t+1} \exp(y_{t+1})}{v_t}$$

The solution to this expression gives that

$$\mathbb{E}_t \exp(y_{t+1}) = \exp\left(\rho y_t + \frac{1}{2}\bar{\eta} + \frac{\rho_\eta}{2}(\eta_t - \bar{\eta}) + \frac{1}{8}\omega^2\right)$$

and

$$\mathbb{E}_t v_{t+1} \exp(y_{t+1}) = \sum_{i=1}^{\infty} \beta^i \exp\Big((B_i + 1)\rho y_t + (C_i + \frac{1}{2}(B_i + 1)^2)\bar{\eta} + \frac{1}{2}(B_i + 1)^2 \rho_\eta (\eta_t - \bar{\eta}) + \ldots$$
$$(F_i + \frac{1}{2}(\frac{1}{2}(B_i + 1)^2 + D_i)^2)\omega^2\Big).$$

The risk-free rate has the following solution:

$$R^{rf}_t = \beta^{-1} \exp\left(\gamma\bar{y} + \gamma\rho(y_t - \bar{y}) - \frac{\gamma^2}{2}\bar{\eta} - \frac{\gamma^2\rho_\eta}{2}(\eta_t - \bar{\eta}) - \frac{\gamma^4}{8}\omega^2\right)$$

As shown by De Groot (2015), there is a parameter restriction that guarantees a finite price-dividend ratio:

$$\beta \exp\left(\frac{1}{2}\left(\frac{1-\gamma}{1-\rho}\right)^2 \bar{\eta} + \frac{(1-\gamma)^4}{8(1-\rho)^4(1-\rho_\eta)^2}\omega^2\right) < 1.$$

We chose our parametrization of β and γ such that this condition is satisfied.

C.2 A discretized solution

Instead of solving the model using the continuous-support process in Equations (21)-(22), one can discretize the stochastic process and obtain approximate solutions for the price-dividend ratio, the conditional expected return on equity, and other objects of interest. If y_t follows a discrete-state-space first-order Markov process with states $y_s,\ s \in \{1, \ldots, m\}$ and transition probability matrix Π with elements $\Pi_{ss'} = P(y_{t+1} = y_{s'} | y_t = y_s)$, then we can rewrite Equation

(23) as

$$v(y_s) = \beta \sum_{s'=1}^{m} \exp((1-\gamma)y_{s'})(v(y_{s'})+1)\Pi_{ss'}$$

which solves to

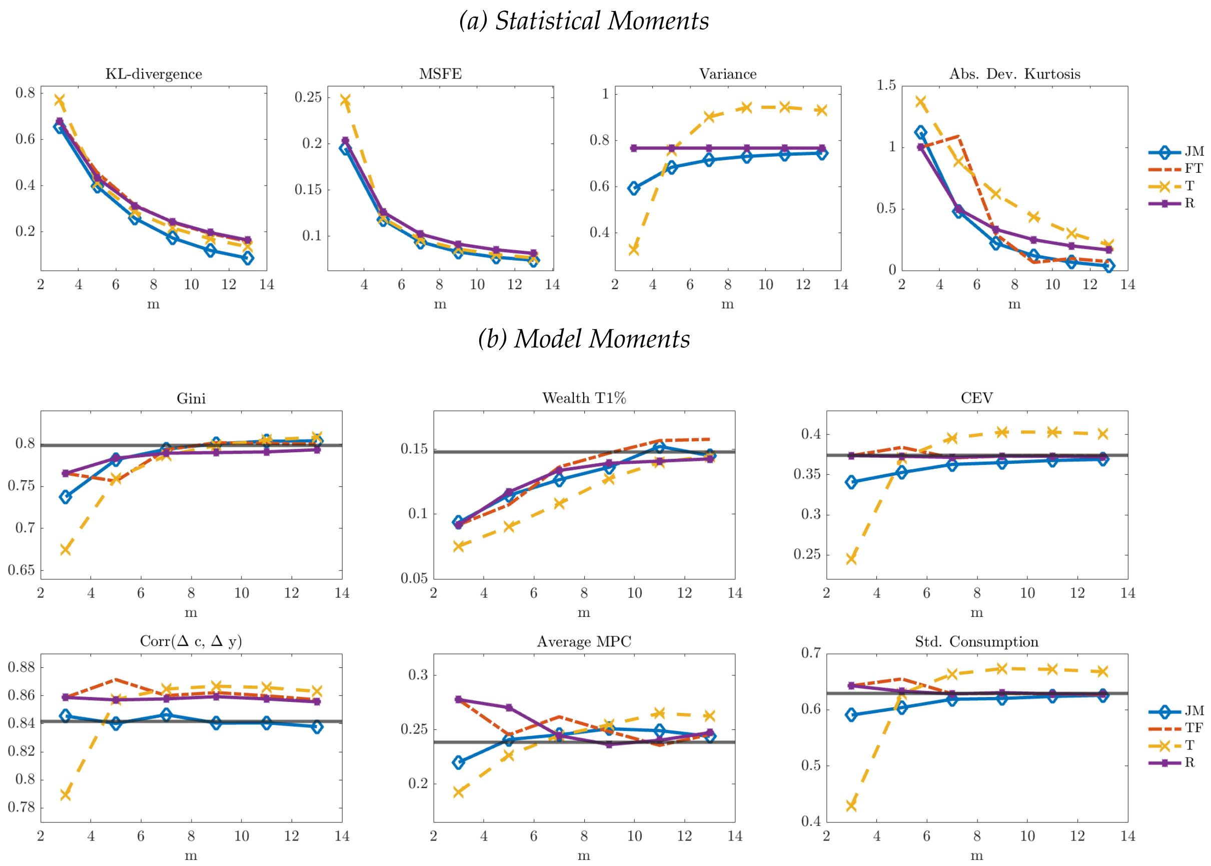
$$v = \left(I_m - \beta\Pi\mathrm{diag}(\exp(1-\gamma)y)\right)^{-1}\beta\Pi\exp((1-\sigma)y), \tag{C.1}$$

where m denotes the number of discrete states of y_t, y is an $s \times 1$ vector with all the levels y_t attains, and v is an $s \times 1$ vector with all discrete realizations of the price-dividend ratio in each discrete realization of y. Similarly, for the vector of conditional expected returns on equity at each value of the grid y_s, denoted $R^e(y_s)$, we have

$$R^e(y_s) = \Big(\sum_{s'} \Pi_{ss'}\exp(y_s)(1+v(y_{s'}))\Big)/v(y_s). \tag{C.2}$$

D AR(1) Process In a Life-Cycle Model

Figure D1: Moments of the AR(1) Process



(a) Statistical Moments

(b) Model Moments

Notes: Panel (a) plots the KL-divergence, the mean-squared forecast error (MSFE), variance, and kurtosis of earnings for different grid sizes and approximation methods. Panel (b) displays the Gini coefficient of wealth, wealth share of the top 1%, certainty equivalent variation of consumption, correlation between log consumption growth and log earnings growth, average marginal propensity to consume, and standard deviation of log consumption for different grid sizes and approximation methods. The solid line is the benchmark model solution using binning with a large number of grid points ($m = 100$). JM refers to our method, FT refers to the Farmer and Toda (2017) method, T refers to Tauchen (1986), and R to Rouwenhorst (1995).

E Discretize or Discard Approximation Error ε: Illustration for AR(1)

In this section, we evaluate whether it is preferable to ignore the approximation error ε in the hidden Markov model, treating the EM algorithm estimates as a discrete Markov chain with grid μ and transition probability matrix Π, or to explicitly include ε as part of the stochastic process with variance σ. As discussed in Section 2.2, incorporating ε increases computational complexity, as it requires approximating ε through integration or discretization. While ε is part of the process and might be theoretically relevant, its influence diminishes as the grid size m increases, reducing its standard deviation σ. This suggests that increasing m and discarding ε may be more efficient than discretizing ε and expanding the state space.

To evaluate this trade-off, we use the life-cycle model from our second application, where earnings follow the AR(1) process described in Section 4.5. Figure E1 compares key moments of the model under two approaches: (i) ignoring ε (blue solid line with diamond markers) in which case the model is identical to the one laid out in Section 4 and (ii) discretizing ε with two grid points $\varepsilon \in \{-\sigma_\varepsilon, \sigma_\varepsilon\}$ with equal weights $\omega = 1/2$ where σ_ε is the estimated variance of the noise term in the HMM (red dash-dotted line) in which case problem of a working age household becomes:

$$V_t(a, x, \varepsilon) = \max_{c, a'} \left\{ u(c) + \beta \sum_{x'} \sum_{\varepsilon'} V_{t+1}\left(a', x', \varepsilon'\right) \Pi_{x'|x} \omega \right\},$$

$$\text{s.t. } c + a' = \tau(\mathfrak{e}) + (1 + r)a$$

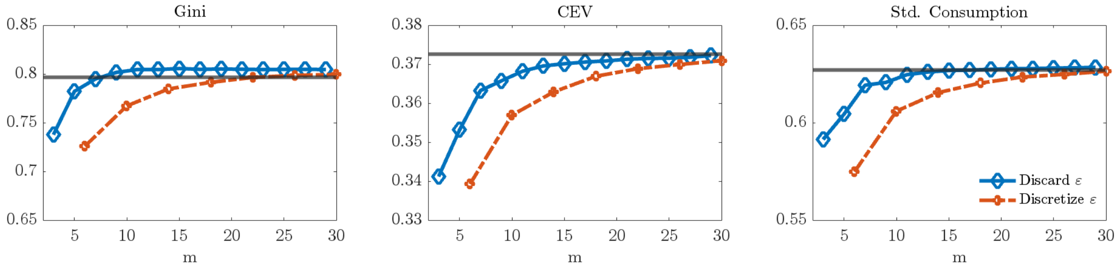
$$a' \geq \underline{a},$$

where earnings satisfy

$$\mathfrak{e}_t = g_t \cdot (\mu(x) + \varepsilon).$$

Here $\Pi_{x'|x}$ is the transition probability between latent states and $\mu(x)$ is the value of the grid at that state. Discretizing ε doubles the state-space size by forming the tensor product of μ and the discretized ε. The figure demonstrates that ignoring ε and using a finer grid (m) produces model moments closer to the benchmark than incorporating ε, after accounting for the larger state space ($2m$) in the latter approach. This highlights the trade-off between accuracy and computational cost, suggesting that, in this setting, it is more effective to discard ε.

Figure E1: Selected Life-Cycle Model Moments of the AR(1) Process With or Without Approximation Error ε



Notes: Panel visualizes the Gini coefficient of wealth, certainty equivalent variation of consumption, and standard deviation of log consumption for different grid sizes and approximation methods. The solid line is the benchmark model solution using binning with a large number of grid points ($m = 100$).

References

Adda, J., and Cooper, R. W. (2003). *Dynamic Economics: Quantitative Methods and Applications*. MIT press.

Altonji, J. G., Hynsjö, D. M., and Vidangos, I. (2022, May). *"Individual Earnings and Family Income: Dynamics and Distribution"* (Working Paper No. 30095). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w30095 doi: 10.3386/w30095

Arellano, M., Blundell, R., and Bonhomme, S. (2017). "Earnings and Consumption Dynamics: a Nonlinear Panel Data Framework". *Econometrica*, *85*(3), 693–734.

Bansal, R., and Yaron, A. (2004). "Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles". *Journal of Finance*, *59*(4), 1481–1509.

Benabou, R. (2002). "Tax and Education Policy in a Heterogenous-Agent Economy: What Levels of Redistribution Maximize Growth and Efficiency?". *Econometrica*, *70*(2), 481-517.

Bilmes, J. A. (1998). "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models". *International computer science institute*, *4*(510), 126.

Civale, S., Díez-Catalán, L., and Fazilet, F. (2016). "Discretizing a Process with Non-Zero Skewness and High Kurtosis". *Available at SSRN 2636485.*

De Groot, O. (2015). "Solving Asset Pricing Models with Stochastic Volatility". *Journal of Economic Dynamics and Control*, *52*, 308–321.

De Nardi, M., and Fella, G. (2017). "Saving and Wealth Inequality". *Review of Economic Dynamics*, *26*, 280–300.

De Nardi, M., Fella, G., and Paz-Pardo, G. (2020). "Nonlinear Household Earnings Dynamics, Self-Insurance, and Welfare". *Journal of the European Economic Association*, *18*(2), 890–926.

Do, M. N. (2003). "Fast Approximation of Kullback-Leibler Distance for Dependence Trees and Hidden Markov Models". *IEEE signal processing letters*, *10*(4), 115–118.

Douc, R., and Moulines, E. (2012). "Asymptotic Properties of the Maximum Likelihood Estimation in Misspecified Hidden Markov Models". *The Annals of Statistics*, *40*(5), 2697–2732.

Duan, J.-C., and Simonato, J.-G. (2001). "American Option Pricing under GARCH by a Markov Chain Approximation". *Journal of Economic Dynamics and Control*, *25*(11), 1689–1718.

Farmer, L. E. (2021). "The Discretization Filter: A Simple Way to Estimate Nonlinear State Space Models". *Quantitative Economics*, *12*(1), 41–76.

Farmer, L. E., and Toda, A. A. (2017). "Discretizing Nonlinear, Non-Gaussian Markov Processes with Exact Conditional Moments". *Quantitative Economics*, *8*(2), 651–683.

Fella, G., Gallipoli, G., and Pan, J. (2019). "Markov-Chain Approximations for Life-Cycle Models". *Review of Economic Dynamics*, *34*, 183–201.

Finesso, L., Grassi, A., and Spreij, P. (2010). "Approximation of Stationary Processes by Hidden Markov Models". *Mathematics of Control, Signals, and Systems*, *22*(1), 1–22.

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

Galindev, R., and Lkhagvasuren, D. (2010). "Discretization of Highly Persistent Correlated AR (1) Shocks". *Journal of Economic Dynamics and Control*, *34*(7), 1260–1276.

Goldfeld, S. M., and Quandt, R. E. (1973). "A Markov Model for Switching Regressions". *Journal of Econometrics*, *1*(1), 3–15.

Gordon, G. (2021). "Efficient VAR Discretization". *Economics Letters*, *204*, 109872.

Gospodinov, N., and Lkhagvasuren, D. (2014). "A Moment-Matching Method for Approximating Vector Autoregressive Processes by Finite-State Markov Chains". *Journal of Applied Econometrics*, *29*(5), 843–859.

Gourieroux, C., Monfort, A., and Trognon, A. (1984). "Pseudo Maximum Likelihood Methods: Theory". *Econometrica*, 681–700.

Guvenen, F., Karahan, F., Ozkan, S., and Song, J. (2021). "What do Data on Millions of US Workers Reveal About Lifecycle Earnings Dynamics?". *Econometrica*, *89*(5), 2303–2339.

Guvenen, F., Ozkan, S., and Madera, R. (2024). "Consumption Dynamics and Welfare Under Non-Gaussian Earnings Risk". *Journal of Economic Dynamics and Control*, *forthcoming*.

Hamilton, J. D. (1990). "Analysis of Time Series Subject to Changes in Regime". *Journal of Econometrics*, *45*(1-2), 39–70.

Judd, K. (1998). *Numerical Methods in Economics*. MIT Press.

Kitagawa, G. (1987). "Non-Gaussian State-Space Modeling of Nonstationary Time Series". *Journal of the American Statistical Association*, *82*(400), 1032–1041.

Kopecky, K. A., and Suen, R. M. (2010). "Finite State Markov-Chain Approximations to Highly Persistent Processes". *Review of Economic Dynamics*, *13*(3), 701–714.

Krueger, D., Mitman, K., and Perri, F. (2016). "Macroeconomics and Household Heterogeneity". In *Handbook of Macroeconomics* (Vol. 2, pp. 843–921). Elsevier.

Krueger, D., and Wu, C. (2021). "Concumption Insurance against Wage Risk: Family Labor Supply and Optimal Progressive Income Taxation". *American Economics Journal: Macroeconomics*, *13*(1), 79-113.

Langrock, R. (2011). “Some Applications of Nonlinear and Non-Gaussian State-Space Modelling by Means of Hidden Markov Models”. *Journal of Applied Statistics*, *38*(12), 2955–2970.

Lehéricy, L. (2021). “Nonasymptotic Control of the MLE for Misspecified Nonparametric Hidden Markov Models”. *Electronic Journal of Statistics*, *15*(2), 4916–4965.

McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). “Finite Mixture Models”. *Annual Review of Statistics and Its Applications*, *6*, 355-378.

Mevel, L., and Finesso, L. (2004). “Asymptotical Statistics of Misspecified Hidden Markov Models”. *IEEE Transactions on Automatic Control*, *49*(7), 1123–1132.

Mitchell, O. S., and Phillips, J. W. (2006). “Social Security Replacement Rates for Alternative Earnings Benchmarks”. *Benefits Quarterly*, *4*, 37-47.

Nguyen, T. T., Nguyen, H. D., Chamroukhi, F., and McLachlan, G. J. (2020). “Approximation by Finite Mixtures of Continuous Density Functions that Vanish at Infinity”. *Cogent Mathematics & Statistics*, *7*(1), 1750861.

Quandt, R. E. (1958). “The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes”. *Journal of the American Statistical Association*, *53*(284), 873–880.

Rabiner, L. R. (1989). “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”. *Proceedings of the IEEE*, *77*(2), 257–286.

Rouwenhorst, K. G. (1995). “Asset Pricing Implications of Equilibrium Business Cycle Models”. In T. F. Cooley (Ed.), *Frontiers of Business Cycle Research* (pp. 294–330). Princeton University Press.

Song, Y. (2014). “Modelling Regime Switching and Structural Breaks with an Infinite Hidden Markov Model”. *Journal of Applied Econometrics*, *29*(5), 825–842.

Tauchen, G. (1986). “Finite State Markov-chain Approximations to Univariate and Vector Autoregressions”. *Economics Letters*, *20*(2), 177–181.

Tauchen, G., and Hussey, R. (1991). “Quadrature-Based Methods for Obtaining Approximate Solutions to Nonlinear Asset Pricing Models”. *Econometrica*, 371–396.

Terry, S. J., and Knotek II, E. S. (2011). “Markov-chain Approximations of Vector Autoregressions: Application of General Multivariate-Normal Integration Techniques”. *Economics Letters*, *110*(1), 4–6.

Vidyasagar, M. (2005). “The Realization Problem for Hidden Markov Models: The Complete Realization Problem”. In *Proceedings of the 44th IEEE Conference on Decision and Control* (pp. 6632–6637).

White, H. (1982). “Maximum Likelihood Estimation of Misspecified Models”. *Econometrica*, 1–25.