Finance and Economics Discussion Series

Federal Reserve Board, Washington, D.C.
ISSN 1936-2854 (Print)
ISSN 2767-3898 (Online)

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

Janssens, Eva F. and McCrary, Sean

2023-040

Please cite this paper as:
Janssens, Eva F., and McCrary, Sean (2023). "Finite-State Markov-Chain Approximations: A Hidden Markov Approach," Finance and Economics Discussion Series 2023-040. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/FEDS.2023.040.

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

Eva F. Janssens† and Sean McCrary‡

May 19, 2023

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 non-stationary processes such as those with a life-cycle component. 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 leads 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

* Disclaimer: The views expressed in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System.

* 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, 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: Economist at the Board of Governors of the Federal Reserve System, e-mail: eva.f.janssens@frb.gov
‡ Sean McCrary: PhD student at the University of Pennsylvania, e-mail: smccrary@sas.upenn.edu

1 Introduction

Numerical methods to solve 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 therefore resemble the continuous process as closely as possible. This paper proposes a novel full-information method that can be used for the discretization of continuous Markov processes. We show that our method results in more accurate solutions to an asset-pricing model, and a better characterization of earnings risk in a life-cycle consumption-saving model with non-linear non-Gaussian earnings processes.

Approximating a continuous stochastic process by a discrete Markov process, characterized by a grid of support points and a transition probability matrix, inherently comes down to picking a misspecified approximating model. Borrowing from the misspecified model literature, we therefore propose a finite-state Markov chain approximation method that minimizes the information loss between the misspecified process and the true continuous process. We assume that the misspecified process is a Hidden Markov Model (HMM), that is, each realization is equal to the sum of a state-dependent level (a grid point) and an error term. This state is unobserved, and the evolution of the unobserved state is governed by a discrete first-order Markov process (with a transition probability matrix). This effectively embeds a discrete Markov chain into a continuous support process via a continuous measurement error. This allows us to use the standard Kullback-Leibler (KL) divergence between the two processes as our measure of information loss.

Consequently, the practical implementation of our method is simple, because in this setting, minimizing the KL divergence is essentially quasi-maximum likelihood estimation, fitting a HMM on data simulated from the continuous support process.[1] What is attractive about our approach is that it results in both an optimal grid and transition probability matrix, and can be applied to multivariate processes, in which case the optimal grid helps limit the curse-of-dimensionality issue posed by tensor grids by accounting for the dependency between variables.

Our theoretical contribution is to prove that, under some assumptions, as the number of unobserved states (and thus grid points) becomes large, the information loss between the

[1] As shown by Mevel and Finesso (2004) and later Douc and Moulines (2012), the maximum likelihood estimator of a misspecified HMM is consistent, in the sense that it minimizes the KL divergence between the model and the true process.

misspecified HMM and the true continuous stochastic process becomes arbitrarily small. This relates our paper to the literature on universal approximators, where we build on the result by Zeevi and Meir (1997) on (Gaussian) mixtures, and extend this to the non-i.i.d. setting of HMM's.[2] Our proof provides insights into what properties of the process determine how many grid points are needed to obtain a certain information loss. For example, more persistent processes require a larger number of grid points, which is why finite-state Markov chain approximations of highly persistent processes tend to be less accurate.[3]

We evaluate the performance of our method in two economic applications: an asset pricing-model and a life-cycle consumption-saving model. First, in our asset pricing model, we discretize dividend growth, which is assumed to follow an autoregressive (AR(1)) process with stochastic volatility, parametrized as in Bansal and Yaron (2004). As shown by De Groot (2015), this model has a closed-form solution. We use this solution as a benchmark to compare the performance of our method against the standards in the literature, and find our discretization captures higher-order moments of the true continuous process better, is more parsimonious, and results in more accurate model solutions. For example, we analyze the accuracy of these discretization methods for estimates of the certainty equivalent level of consumption (CE) and find that our method deviates 0.8-1.9% from the closed-form solution of De Groot (2015), while the comparison methods have deviations ranging from 4 to 12%. These results highlight the strength of a full-information approach, because for a non-linear object such as the CE, all information of the stochastic process matters.

Second, we analyze the performance of our method through the lens of a life-cycle consumption-saving model. In this application, we focus on two processes featuring life-cycle dependence; the process proposed in Guvenen, Karahan, Ozkan, and Song (2021) that features non-employment shocks, and innovations with positive skewness, and the non-parametric process in Arellano, Blundell, and Bonhomme (2017). These processes are considered to be at the frontier of the earnings dynamics literature (Altonji, Hynsjö, and Vidangos, 2022). Our discretizations better capture the excess skewness and kurtosis of the Guvenen et al. (2021) and Arellano et al. (2017) processes than commonly-used binning-based discretization methods.[4] For the Guvenen et al. (2021) process specifically, the binning method fails to capture the long-run dynamics of non-employment.

[2]The universal approximator property has also been shown to hold for Neural Networks, but to our knowledge also only in an i.i.d. setting, see, e.g., the seminal work by Hornik, Stinchcombe, and White (1989).

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

[4]These binning methods are adapted from the textbook treatment of Adda and Cooper (2003).

In the life-cycle model, we find that the discretization method matters for various economic quantities of interest, including the welfare cost of risk, wealth inequality measures, and marginal propensities to consume (MPC). By failing to fully capture the excess kurtosis and skewness of the processes over the life-cycle, binning-based methods underestimate the welfare cost of risk and the amount of precautionary saving in the economy. For the Guvenen et al. (2021) process, the binning-based method underestimates the welfare cost of risk by 23 percentage points relative to our method, and for the Arellano et al. (2017) process, the difference is 3 percentage points. Discretizations also matter for the amount of wealth inequality a life-cycle model can generate. While it is known that life-cycle models struggle to match the wealth distribution in the data (De Nardi and Fella, 2017), we show that more accurate discretizations of the earnings process can generate more wealth inequality. Our discretization of the Arellano et al. (2017) process generates a Wealth Gini of 0.76, close to that of the United States (0.77-0.78), while binning results in a value that is 0.06 lower. Similarly, our discretization results in top 1% wealth shares close to those in the data, while the binning-based estimates underestimate this moment.

Our results on the importance of discretization methods for life-cycle model solutions extends to simpler processes. For a Gaussian AR(1) and mixture AR(1), we show the life-cycle model solutions differ significantly between discretization methods, although the differences do become smaller when a sufficiently large number of grid points is used. This is an important insight, given the low number of grid points the literature tends to use for these processes. Our solution, on the other hand, changes little when adding more grid points, because our method is more parsimonious and captures more information of the true process than the other discretization methods. For these processes, the sensitivity of the marginal propensities to consume over the life-cycle stands out. Other discretization methods can underestimate the MPC's for younger age groups by as much as 20 percentage points when using a low number of grid points.

Finally, we compare the life-cycle implications across different stochastic processes. To our knowledge, this paper is the first to discretize the Guvenen et al. (2021) process and evaluate its implications in an incomplete markets model. Furthermore, representing the Arellano et al. (2017) and Guvenen et al. (2021) processes as discrete Markov chains allows for a consistent comparison between the two processes. We find the largest source of risk in the Guvenen et al. (2021) process comes from the probability of non-employment, which is a highly persistent state with rising persistence over the life-cycle. In contrast, most risk in the Arellano et al. (2017) process comes from the highest earnings state, which features a considerable probability of earnings loss next period, especially at younger ages, creating a strong precautionary savings

motive among high earners in the model. These differences between earnings processes result in different dynamics in our life-cycle model. The risk of non-employment in the Guvenen et al. (2021) process generates a life-cycle profile for MPC's that is flatter than that of the Arellano et al. (2017) process, and in a higher welfare cost of risk (0.69 instead of 0.19 in the model with Arellano et al. (2017)). The Arellano et al. (2017) process features more earnings inequality and consequently provides a better fit to the wealth distribution than the Guvenen et al. (2021) process.

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. Appendix A provides the proof of our Main Theorem. Appendix B provides details on the estimation of the HMM. Appendix C provides an additional application to the discretization of vector autoregressive processes.

Related literature. Several methods have been proposed to discretize stochastic processes. Most of these, such as Tauchen (1986), Rouwenhorst (1995), Tauchen and Hussey (1991), Duan and Simonato (2001), Terry and Knotek II (2011), and Gospodinov and Lkhagvasuren (2014), are designed for specific (linear) processes, such as AR(1) or VAR processes. 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 analyze how it performs under settings where the innovations are drawn from a mixture of normals. Galindev and Lkhagvasuren (2010) adapt Rouwenhorst (1995) to a setting with highly-persistent correlated AR(1) shocks. Civale, Díez-Catalán, and Fazilet (2016) adapt the Tauchen (1986) method to accommodate autoregressive processes with innovations drawn from a normal mixture. Unlike these methods, our method is applicable to any process, and provides both an optimal grid and transition probability matrix, while these methods typically take a grid as input, and/or assume equal-distant or equal-quantile grids.

Some discretization methods are applicable to a larger class of stochastic processes. Binning methods as in Judd (1998) and Adda and Cooper (2003), that discretize via a partition of the quantile space, are applicable to any stochastic process. However, binning methods only match one-step ahead transitions between bins and take the grid spacing as an input, while our discretization method looks at the full dynamics and provides an optimal grid. Farmer and Toda (2017) propose a method to refine discrete approximations by moment matching. Their method takes as inputs a grid, an initial transition probability matrix, and a set of moments to match, where the goal is to match these moments exactly – if possible – with a

transition matrix that is close, measured through relative entropy, to the initial approximation. Our method, in contrast, can be seen as a full-information discretization method, rather than moment-matching, that does not rely on prior information (i.e., an initial discretization) to obtain identification.

For multivariate processes, most existing methods rely on tensor grids, which leads to a curse of dimensionality and is computationally unattractive. As stated by Gordon (2021), tensor grids are inefficient, because many of the grid points will rarely be visited. Gordon (2021) proposes the use of pruning and sparse grids for VAR models. Our method results in optimal grids that limit the curse-of-dimensionality issue when the variables are correlated, and is applicable to any type of process.

Our results relate to the literature on misspecified models (Gourieroux, Monfort, and Trognon, 1984; White, 1982), and, specifically, misspecified Hidden Markov Models (Douc and Moulines, 2012; Mevel and Finesso, 2004). The use of HMM's is prevalent in economics and machine learning[5], but, to our knowledge, the application of HMM's as a finite-state Markov chain approximation method for continuous stochastic processes is novel, as is our theoretical result on the ability of HMM's to approximate such processes.[6] In the signal processing literature, Vidyasagar (2005), Finesso, Grassi, and Spreij (2010), and others, consider the problem of representing discrete state-space stationary processes as HMM's, but 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. Denote its probability distribution by $f(\mathbf{y})$. The objective is to approximate the distribution of $\mathbf{y}$ by a misspecified model, with probability 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

[5]The interpretation of a HMM as a dimension reduction method for dependent data is common in the statistics and machine-learning literature (McLachlan, Lee, and Rathnayake, 2019), where a common application of HMM's is text processing. Applications of HMM's in econometrics include the detection of structural breaks (Song, 2014) and modeling of regime switches (starting with Quandt (1958), Goldfeld and Quandt (1973), and Hamilton (1990)). HMM's have also been used to approximate the dynamics of the latent state in non-linear state space models for the purpose of estimation, as in Kitagawa (1987), Langrock (2011), and Farmer (2021).

[6]This is an intuition Lehéricy (2021) refers to but does not prove.

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:[7]

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

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

The transition matrix Π_t has stationary distribution $\boldsymbol{\delta}_t = (\delta_{1,t}, \delta_{2,t}, \ldots, \delta_{m,t})$. 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

[7] 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.

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. The intuition behind this HMM is that it provides a time-varying (soft) clustering of the continuous variable y into m discrete states x that each correspond to a grid point $\mu(x)$.

We consider time series settings where $N = 1$, as well as panels with $N \geq 2$. The inclusion of a panel dimension allows for the estimation of parameters that vary with t, for example, over the life-cycle.

2.2 Properties of the KL divergence

Given our objective of minimizing the information loss between the true and approximating process, two questions arise. First, whether there is a consistent estimator of the Hidden Markov model parameters in this setting. In the case of misspecified models, consistency is defined as whether the estimator converges to the value that minimizes the KL divergence. This has been shown to be true for misspecified Hidden Markov Models by Mevel and Finesso (2004), and later in a more general setting by Douc and Moulines (2012). The second question is whether, with a sufficient number of hidden states (and thus grid points), the information loss between the true and approximating process can be made arbitrarily small. We prove, under a set of assumptions, that the answer to the second question is positive.

The Main Theorem builds on the results of Zeevi and Meir (1997), 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.

As in Zeevi and Meir (1997), denote

$$\mathscr{F}_{c,\eta} = \{f \in \mathscr{F}_c | f \geq \eta > 0, \forall y \in \mathscr{Y}\}, \text{ with } \mathscr{F}_c = \left\{f | f \in C\mathscr{Y}, f \geq 0, \int f = 1\right\}$$

where $\mathscr{F}_c$ is the class of continuous density functions with compact support $\mathscr{Y} \subset \mathbb{R}^k$ fixed and given. $\mathscr{F}_{c,\eta} \subset \mathscr{F}_c$ is bounded below over $\mathscr{Y}$ by some positive constant, denoted by η.

We impose the following assumptions on the true process $f(\mathbf{y})$ and approximating model $p(\mathbf{y}, \theta)$:[8]

(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

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

(A2) $f(y_t|y_{t-1}) \in \mathscr{F}_{c,\eta}$.

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

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

(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 $p(y_t|y_{t-1}, ..., y_1; \theta_m) \in \mathscr{F}_{c,\eta}$.

The first-order Markov assumption (A1) is w.l.o.g., because any (finite) higher-order Markov process can be written as a (multivariate) first-order Markov process. Compared to the set-up of Section 2.1, Assumption (A5) omits time-variation in the parameters Π, μ and σ. 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 $\mu_m \in \mathscr{Y}$, variance $\sigma_m \geq \tau > 0$ and transition probability matrix Π_m,*

[8] Assumption (A4) is satisfied for some well-known processes. For example, straightforward algebra shows that for an AR(1) process, $f(y_t|y_{t-1}) = N(\rho y_{t-1}, \sigma^2)$ is Lipschitz, and $\log f(y_t|y_{t-1})$ is locally Lipschitz.

collected in $\theta_m = (\mu_m, \Pi_m, \sigma_m)$ *such that the KL divergence between* $f(\mathbf{y})$ *and* $p(\mathbf{y}; \theta)$ *on the compact subset* $y \in \mathcal{Y} \subset \mathbb{R}^k$*, given by*

$$D_{\mathcal{Y}}^{KL}(f(\mathbf{y})||p(\mathbf{y};\theta)) = \int_{\mathcal{Y}} f(\mathbf{y}) \log \frac{f(\mathbf{y})}{p(\mathbf{y};\theta)} d\mathbf{y},$$

can be made arbitrarily small.

The full proof is given in Appendix A.

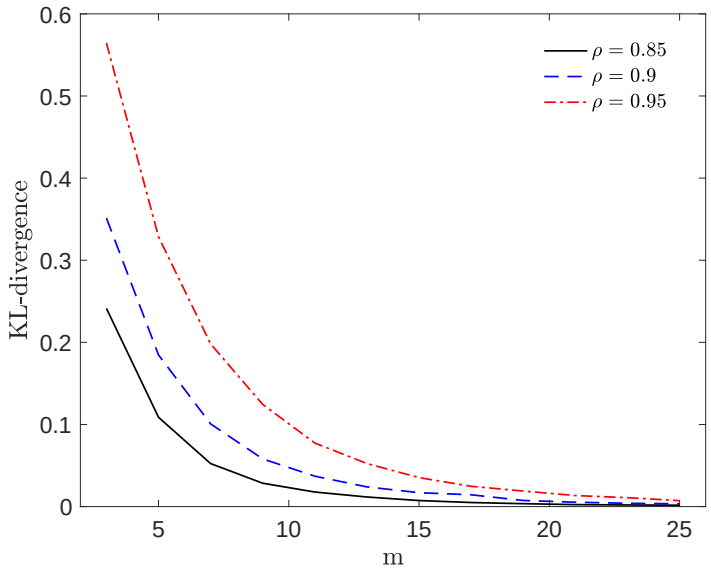
Sketch of proof. The first step of the proof consists of showing that the conditional distributions of our Hidden Markov Model, denoted by $p(y_t|y_{t-1}, ..., y_1; \theta_m)$, are Gaussian mixtures, whose mixture weights converge to a row of the transition probability matrix Π_m as m becomes large and the filter $p(x_t|y_t, ..., y_1; \theta_m)$ becomes better, 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 $\mu_m(l)$ denotes the closest grid point to y_{t-1}. The proof then applies the result of Zeevi and Meir (1997) to m conditional distributions at the same time, that is, to $f(y_t|y_{t-1} = \mu_m(i))$ and $p^0(y_t|y_{t-1} = \mu_m(i); \theta_m)$, conditioning on y_{t-1} being equal to one of m grid points $\{\mu_m(i)\}_{i=1}^m$. This results in an additional term in the KL divergence compared to the Zeevi and Meir (1997) result, because in our setting, these m conditional distributions $f(y_t|y_{t-1} = \mu_m(i))$ are approximated by m Gaussian mixtures that all have the same location parameters μ_m, as opposed to being freely chosen. However, we do have enough degrees of freedom for m different sets of convex mixture weights, because the transition probability matrix has $m \times m$ elements. This is summarized in Lemma 4 and 5 in the Appendix.

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 $\{y_{t-k}\}_{k=1}^{t-1}$ within the compact set $\mathcal{Y}$ are also small. This follows because: (i) the true process is assumed stationary and Markovian, (ii) the local Lipschitz assumption ensures that the KL divergences on the compact set are well behaved and bounded, and (iii) $p(y_t|y_{t-1}, ..., y_1; \theta_m)$ is Lipschitz in $y_{t-1}, ..., y_1$ and as m increases, the filter $p(x_t|y_t, ..., y_1; \theta_m)$ becomes better, and $p(y_t|y_{t-1}, ..., y_1; \theta_m)$ becomes approximately forgetting beyond $t-1$. 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.

Estimation. The results of Mevel and Finesso (2004) allow us to extend our theorem by the following insight. 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. Therefore, we can use the Expectation-Maximization (EM) algorithm to obtain our grid points and transition probability matrix.[9] We describe the estimation procedure in Appendix Section B.

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.

Figure 1: KL-divergence of the approximating model (in Equations (2)*) versus the true process, where the true process is an AR(1) process* $y_t = \rho y_{t-1} + \varepsilon_t$, $\varepsilon_t \sim N(0,1)$ *for three values of* ρ. m *is the number of grid points used for the discretization.*



Although the Main Theorem does not state the rate of convergence (that is, the number of grid points needed to achieve a particular information loss), the proof does provide insights on

[9]This requires additional assumptions on the true stochastic process, including geometric ergodicity, and uniformly bounded moments of sufficiently high order.

what properties of the true process matter for how many grid points are needed to obtain a particular precision. This will, among other things, depend on the local Lipschitz coefficient of $\log f(y_t|y_{t-1})$ and $f(y_t|y_{t-1})$, as well as the size of the compact set $\mathcal{Y}$. One property that affects the number of grid points is the persistence of the stochastic process. For an AR(1) process, one can show these Lipschitz coefficients as well as the unconditional variance are increasing in persistence. Consequently, the more persistent a stochastic process, the more grid point are needed to achieve the same information loss. Figure 1 shows how the KL-divergence of our HMM approximating an AR(1) process approaches zero when increasing the number of grid points m, but does so more slowly when the AR(1) process is more persistent. As such, our results shed some light on why discretizing highly persistent AR(1) processes poses a challenge, as discussed 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:

$$\mu(\lceil m/2 \rceil) = 0, \text{ and } \mu(\lceil m/2 \rceil - r) = -\mu(\lceil m/2 \rceil + r), \quad \text{for } r = 1, ..., \lfloor m/2 \rfloor \tag{4}$$

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

$$\Pi_{i,j} = \Pi_{(m+1-i),(m+1-j)}. \tag{5}$$

For the specific restrictions in Equations (4)-(5), a closed-form solution is available for the M step. In other cases, one may want to introduce restrictions through penalty terms rather than hard restrictions. For example, one may want the discretized process to target certain moments. Denote a certain set of moment functions of the discretized process by $\mathcal{M}(p(\mathbf{y};\theta))$ and the moments of the continuous process by $\mathcal{M}(f(\mathbf{y}))$. In that case, instead of maximizing

the log-likelihood of the simulated data $\mathbf{y}_{\text{sim}}$, maximize:

$$\log(\mathcal{L}(\theta|\mathbf{y},\mathbf{x})) - \lambda\mathcal{D}\left(\mathcal{M}(f(\mathbf{y})), \mathcal{M}(p(\mathbf{y};\theta))\right) \tag{6}$$

where $\lambda \in \mathbb{R}^+$ is a scalar parameter and $\mathcal{D}(\cdot,\cdot)$ a distance measure of choice. λ is chosen by the researcher. A higher λ should be chosen if the researcher considers it more important that the discretization matches the moments $\mathcal{M}$. When using this penalty term, the M step is no longer analytically tractable and numerical optimization is necessary.

3 Application I: Asset Pricing Model with Stochastic Volatility

In this section, we evaluate the performance of our method in an asset pricing model where dividend growth features stochastic volatility. Most models that involve solving a dynamic stochastic optimization problem with a continuous-support process do not have a closed-form solution. As shown by De Groot (2015), however, the model we present below does have a closed-form solution for the price-dividend ratio and the conditional expected return on equity. The existence of an analytical solution gives us a benchmark with which to compare a model solved with ours and other discretization methods.

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]

$$y_t = \bar{y} + \rho(y_{t-1} - \bar{y}) + \sqrt{\eta}_t \varepsilon_t \tag{7}$$

$$\eta_t = \bar{\eta} + \rho_\eta(\eta_{t-1} - \bar{\eta}) + \omega \varepsilon_{\eta,t}. \tag{8}$$

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:

$$v_t = \mathbb{E}_t \beta \left(\frac{d_{t+1}}{d_t} \right)^{1-\gamma} (v_{t+1} + 1). \tag{9}$$

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

$$\mathbb{E}_t R^e_{t+1} = \mathbb{E}_t \left(\frac{d_{t+1} + p_{t+1}}{p_t} \right). \tag{10}$$

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

The reason why we are interested in the performance of capturing $\mathbb{E}_t R^e_{t+1}$ in addition to v_t is because of its non-linear dependence on v_t, which is also approximated. The approximation errors will compound in a non-trivial way, and we are interested in how accurate the discretization methods are when these errors accumulate.

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

[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 Bansal and Yaron (2004), the probability of a negative value for η is very small, and in our long sample of simulations, it doesn't occur.

stream. One could ask what the certainty equivalent level of consumption is that would make the household indifferent between the risky consumption stream and a certain (constant) level of consumption. We denote that constant value by $x(d)$, which is 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 a higher willingness to pay, so to the extent the discretizations fail to capture risk, they will overstate x relative to the true value.

Calibration. The parametrization used for the results in the tables below are based on the estimates of the stochastic volatility process in Bansal and Yaron (2004), annualized as in De Groot (2015), that is, $\gamma = 1.5$, $\rho_\eta = 0.855$, $\omega = 7.4000 \times 10^{-5}$, $\bar{\eta} = 0.0012$, $\beta = 0.95$, $\rho = 0.868$, $\bar{y} = 0.0179$. We choose risk aversion γ and the discount factor β 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 (7)-(8) 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 and Toda and the binning method of Adda and Cooper (2003), both using their standard configurations.[12] Both methods use a tensor grid for multivariate processes.

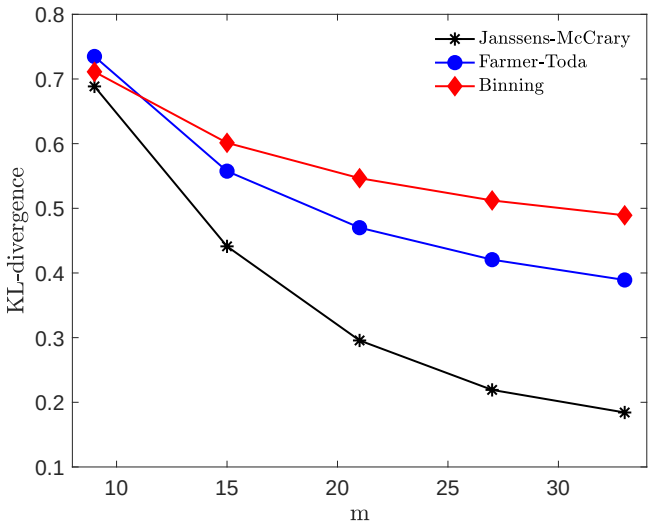
Figure 2 visualizes the KL divergence of our discretization for different choices of grid size m relative to the true AR(1)-SV process. The figure also visualizes the KL divergences of the two other discretization methods. The likelihoods for these methods are computed by interpreting the transition probability matrix and grid of the different discretization methods as parameters Π and μ in our HMM framework, re-estimating the variance of the approximation error. 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 eleven. 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 27 grid points, and the binning method needs

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

[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 method. We adapt the Farmer and Toda method for this specification of an AR(1)-SV, set to match the first two conditional moments in each grid point.

more than 33. 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: KL-divergence of the approximating model likelihood versus the likelihood of the true process for the AR(1)-SV process in Equations (7)-(8)*, for different discretization methods and different grid sizes* m*.*



Notes: We only visualize a selected number of grid points, because the other methods rely on a tensor grid, and cannot be computed for all choices of m. 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.

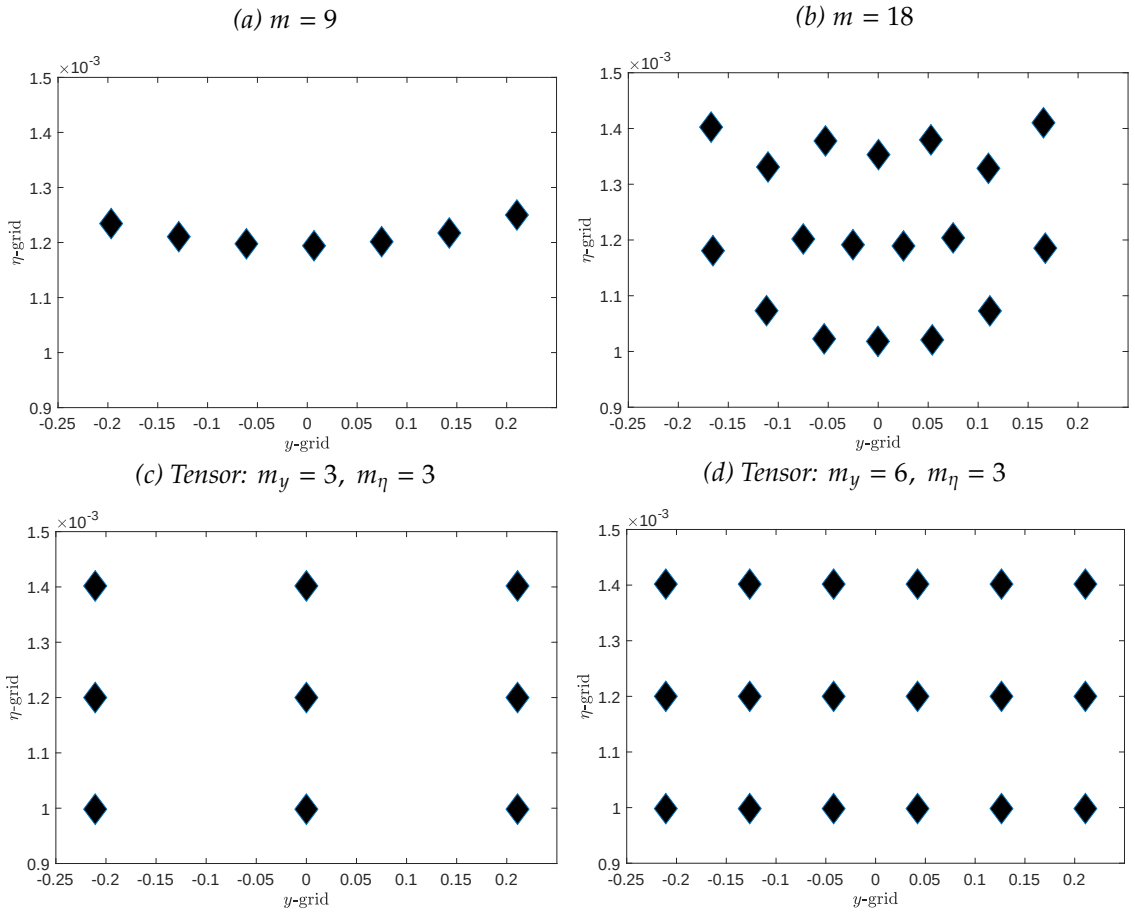
Figure 3 illustrates how our method optimally chooses the grid points for a multivariate process like the AR(1)-SV process.[13] Figure 3(a) shows the optimal grid for $m = 9$ grid points, and shows how our method assigns the grid points in the tails to have higher variances than in the center. This is consistent with the intuition behind an AR(1)-SV process, as it is more likely a high value of y_t is accompanied by a high realization of the variance η_t. As m becomes larger, our optimal grid adds what we call 'double' or 'triple' states. These are grid points with similar levels for y, but different values for the variance η. These grid points will have different dynamics to next period's state despite having the same level of y, as will be reflected by differences in the rows of the transition probability matrix for these states.

Table 1 computes several statistics to compare the performance of our method and the existing methods at capturing moments of y. As can be seen, the Farmer and Toda (2017) method does well at the mean and variance, as these are the moments they target, while we do well at higher order moments such as the skewness and kurtosis. The Mean Squared Forecast Error (MSFE) of the other methods is 30-40% larger than ours, supporting that we give an agent a better process to make forecasts with.[14]

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

[14] 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

Figure 3: Visualisation of the optimal grid for grid sizes $m = (9, 18)$ compared to a tensor grid, for the AR(1)-SV process as in Equation (7)-(8).



Notes: 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 ($\hat{M}$) and the analytical benchmark (M). To assess the accuracy of the different solutions, we compute the following summary statistic:

$$\log_{10}(|\hat{M}/M - 1|).$$

Lower values of $\log_{10}(|\hat{M}/M - 1|)$ indicate the moments of the discrete model are closer to those of the benchmark. The results of this analysis are summarized in Table 2. Overall,

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 $i = \underset{i\in\{1,\dots,m\}}{\operatorname{argmin}} |y_{t-1} - \mu(x_{t-1} = i)|$.

Table 1: Comparison for an AR(1) process with stochastic volatility as in Equation (7)-(8) *(based on a simulation of* $T = 200,000$*) parametrized as in Bansal and Yaron (2004).*

Method	Janssens-McCrary	Farmer-Toda	Binning
m = 15 ($m_y = 5$, $m_\eta = 3$ for Farmer-Toda and binning)			
Dev. uncond. mean y	0.018	**0.018**	0.018
% dev. uncond. variance y	−15.4	**0.525**	−23.7
Abs. dev. uncond. skewness y	**<0.001**	0.044	0.023
% dev. uncond. kurtosis y	**−9.42**	-19.2	−37.5
% dev. autocorrelation y	3.35	**−0.053**	−5.66
% abs. dev. cond. mean y	0.003	**0.002**	0.004
% abs. dev. cond. variance y	33.3	26.3	**26.0**
Abs. dev. cond. skewness y	**0.543**	1.16	0.580
% abs. dev. cond. kurtosis y	51.3	81.1	**18.5**
MSFE y	**0.0013**	0.0018	0.0017

our method always performs best at the mean, and, almost always, at the variance of both statistics. The differences in accuracy between discretization methods for the mean expected return of equity are larger than the differences in accuracy for the mean price-dividend ratio, because of the accumulation of approximation errors through a non-linear transformation.

Table 2: Accuracy of asset pricing model solutions for the price-dividend ratio v_t *and the conditional expected return on equity* $\mathbb{E}R^e_{t+1}$.

	M	$\log_{10}(\lvert\hat{M}/M - 1\rvert)$		
		Janssens-McCrary	Farmer-Toda	Binning
		$m = 9$	$m = 3$ x 3	$m = 3$ x 3
Mean v_t	18.10	**−1.67**	−1.51	−1.13
Variance v_t	9.61	**−1.33**	−0.29	−0.07
Mean $\mathbb{E}_t(R^e_{t+1})$	1.08	**−3.10**	−2.37	−2.67
Variance $\mathbb{E}_t(R^e_{t+1})$	0.01	**−0.64**	−0.49	−0.28
		$m = 15$	$m = 5$ x 3	$m = 5$ x 3
Mean v_t	18.10	**−2.77**	−2.23	−1.29
Variance v_t	9.61	−0.65	**−2.33**	−0.19
Mean $\mathbb{E}_t(R^e_{t+1})$	1.08	**−4.51**	−2.44	−2.73
Variance $\mathbb{E}_t(R^e_{t+1})$	0.01	**−0.64**	−0.49	−0.28

Notes: Comparison of the mean and variance of a simulated time-series from the discretized model solutions (denoted by $\hat{M}$) and the analytical closed-form model solution (denoted by M), where the relative accuracy of the solution for moment M is measured by $\log_{10}(\lvert\hat{M}/M - 1\rvert)$. The lower (more negative) this value is, the closer this moment of the simulated time series of the discrete model solution is to the moment of time series from the exact model solution. Lowest values are marked in bold.

Table 3: Accuracy of asset pricing model solutions for the certainty equivalent of consumption (CE): true value of CE compared to those following from three different methods. The lower the percentage deviation, the closer the solution of the discretized model is to the truth. Different grid sizes are presented.

	Janssens-McCrary	Farmer-Toda	Binning
CE (true)	% dev	% dev	% dev
	$m = 9$	$m = 3$ x 3	$m = 3$ x 3
1.65	**0.76%**	8.28%	5.41%
	$m = 15$	$m = 5$ x 3	$m = 5$ x 3
1.65	**1.93%**	12.22%	3.95%

Notes: Lowest values are marked in bold. Average is taken over 50 simulations of the CE.

In Table 3, we analyze the accuracy of the different methods when computing the CE for two different grid sizes. As follows from Table 3, our method produces the most accurate estimates of the CE, with deviations in percentage points 0.8-2% from the truth. The other two methods are at best 4% away from the truth, and at worst 12%, underestimating the amount of consumption the household is willing to give up to remove risk.

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 agent 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, using our adapted algorithm laid out in Appendix Section B.2.

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:

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

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

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

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. The specification for the deterministic component of earnings g_t is taken from Guvenen et al. (2021).

Retired households solve the following problem:

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

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

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

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:

$$\tau(y) = (1-\chi)y^{1-\mu}. \tag{11}$$

The parameters χ and μ govern the level and progressivity of the tax function. Following Krueger and Wu (2021), we set the progressivity parameter to 0.1327, and the level parameter to 0.1575. The calibration is summarized in Table 4.

Table 4: 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 over the life cycle. In addition, we compute three 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 policy experiments, so it is important to know its sensitivity to the discretization method.

[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 $W\big((1-CEV)c^0\big) = W\big(c^1\big)$ (Wu and Krueger, 2021).

Second, we report the partial insurance to persistent income shocks coefficient as in Blundell, Pistaferri, and Preston (2008):

$$\psi^{P}_{\text{BPP}} = 1 - \frac{\text{cov}(\Delta c_{it}, y_{i,t+1} - y_{i,t-2})}{\text{cov}(\Delta y_{it}, y_{i,t+1} - y_{i,t-2})}.$$

This statistic measures the extent to which consumption responds to unpredictable persistent changes in income. This statistic is used to validate the predictions of life-cycle models versus data in practice. Third, 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. We study MPC's over the life-cycle and across the wealth distribution. MPC's are a common object of interest when studying fiscal policy.

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]

$$\begin{aligned}
y^i_t &= (1 - v^i_t) e^{(z^i_t + \varepsilon^i_t)} \\
z^i_t &= \rho z^i_{t-1} + \eta^i_t \\
z^i_0 &\sim N(0, \sigma_{z_0}) \\
\eta^i_t &\sim \begin{cases} N(\mu_{\eta,1}, \sigma_{\eta,1}) & \text{with prob. } p_z \\ N(\mu_{\eta,2}, \sigma_{\eta,2}) & \text{with prob. } 1 - p_z \end{cases} \\
\varepsilon^i_t &\sim \begin{cases} N(\mu_{\varepsilon,1}, \sigma_{\varepsilon,1}) & \text{with prob. } p_\varepsilon \\ N(\mu_{\varepsilon,2}, \sigma_{\varepsilon,2}) & \text{with prob. } 1 - p_\varepsilon \end{cases} \\
v^i_t &\sim \begin{cases} 0 & \text{with prob. } 1 - p_v(t, z^i_t), \\ \min\{1, \exp(\lambda)\} & \text{with prob. } p_v(t, z^i_t) \end{cases}
\end{aligned} \tag{12}$$

[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: $\rho = 0.959$, $p_z = 0.407$, $\mu_{\eta,1} = -0.085$, $\mu_{\eta,2} = 0.085 p_z/(1 - p_z)$, $\sigma_{\eta,1} = 0.364$, $\sigma_{\eta,2} = 0.069$, $p_\varepsilon = 0.13$, $\mu_{\varepsilon,1} = 0.271$, $\mu_{\varepsilon,2} = -0.271 p_\varepsilon/(1 - p_\varepsilon)$, $\sigma_{\varepsilon,1} = 0.285$, $\sigma_{\varepsilon,2} = 0.037$, $\lambda = 0.0001$. We have $(a, b, c, d) = (-3.353, -0.859, -5.034, -2.895)$.

where p_v is given by

$$p_v(t, z_t) = \frac{e^{\xi_t^i}}{1 + e^{\xi_t^i}}, \quad \text{where } \xi_t^i \equiv a + bt + cz_t^i + dz_t^i t.$$

Here y_t^i is the earnings level of individual i at time t, z_t^i is the persistent component of earnings, ε_t^i is the transitory component and v_t^i 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_t.

Discretization. For our discretization, we use a multivariate discretization on $\log(y_t^i + 1)$ and z_t^i jointly. We allow the grid and transition probabilities of our discretization to be age-dependent. We use twelve grid points, because, as one can see below in the moment analysis, twelve grid points captures the main features of the process well.

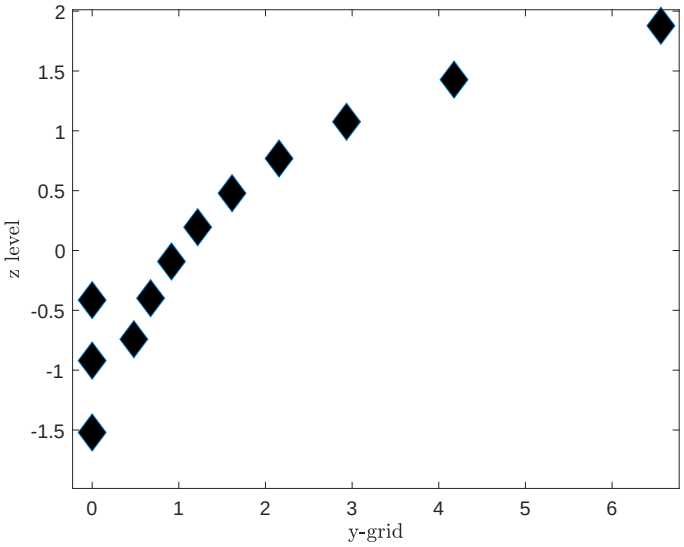
The resulting optimal age-dependent grids are visualized in Figure 4. Figure 4b shows that the grid points have a positive trend in age, capturing the increase in earnings dispersion over the life-cycle. Figure 4a shows that the discretization method generates a grid with multiple non-employment states. Having multiple states with an earnings level of zero generates heterogeneous job-finding probabilities, that is, non-employment states that differ in terms of their persistence. This is visualized in Figure E1 in the Appendix, depicting the age-dependent transition probability matrix. The first three rows represent the zero-earnings states, and by looking at the diagonal, we can see that these states indeed differ in terms of their persistence, and that this persistence changes over the life-cycle. Furthermore, Figure E1 shows how in the Guvenen et al. (2021) process, non-employment becomes highly persistent towards the end of working life.[17]

To the best of our knowledge, our paper is the first to discretize the process in Guvenen et al. (2021). We compare ourselves against a binning method. Standard binning methods, however, do not work for the Guvenen et al. (2021) process, because of the large number of zeros generated by the process. We adapt Adda and Cooper (2003) by adding a zero earnings-state and then use standard binning on the observations $y_{it} > 0$. We allow both the grid and transition probability matrix to vary by age.

[17]It should be noted that Guvenen et al. (2021) do not differentiate between unemployment and non-employment, which explains why these transition probabilities out of the zero-earnings states are different from those we know from the unemployment duration literature.

Figure 4: Visualisations of the optimal grid of the discretization of the stochastic process in Guvenen et al. (2021) with $m = 12$. Note that panel (b) only shows ten lines, because there are three grid points at zero.

(a) Optimal grid at age $t = 25$



(b) Optimal grid over the life-cycle

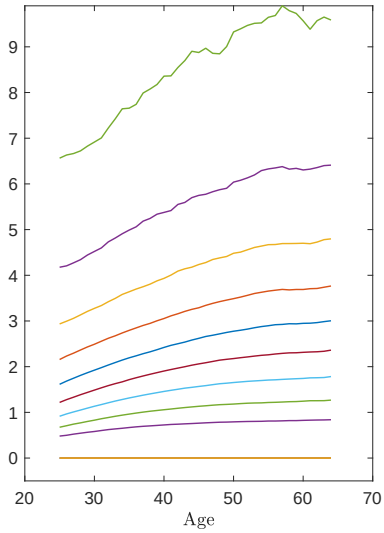


Figure 5 visualizes the unconditional moments of the earnings levels and arc-changes in earnings of the Guvenen et al. (2021) process over the life-cycle, and the extent to which the discretized processes can replicate these moments.[18] As these figures show, our discretization method captures the unconditional moments of the earnings levels well, and does so better than the binning method. The binning method performs similar to our method at the moments on arc-changes. In Figure 5c, the non-employment dynamics over the life-cycle are visualized for the two different discretizations. Our discretization is able to capture the life-cycle profile of the two- and three-period ahead conditional non-employment probabilities better than the binning method. The binning method by construction performs well at the one-period-ahead persistence of non-employment, but fails to capture the longer-run non-employment dynamics.

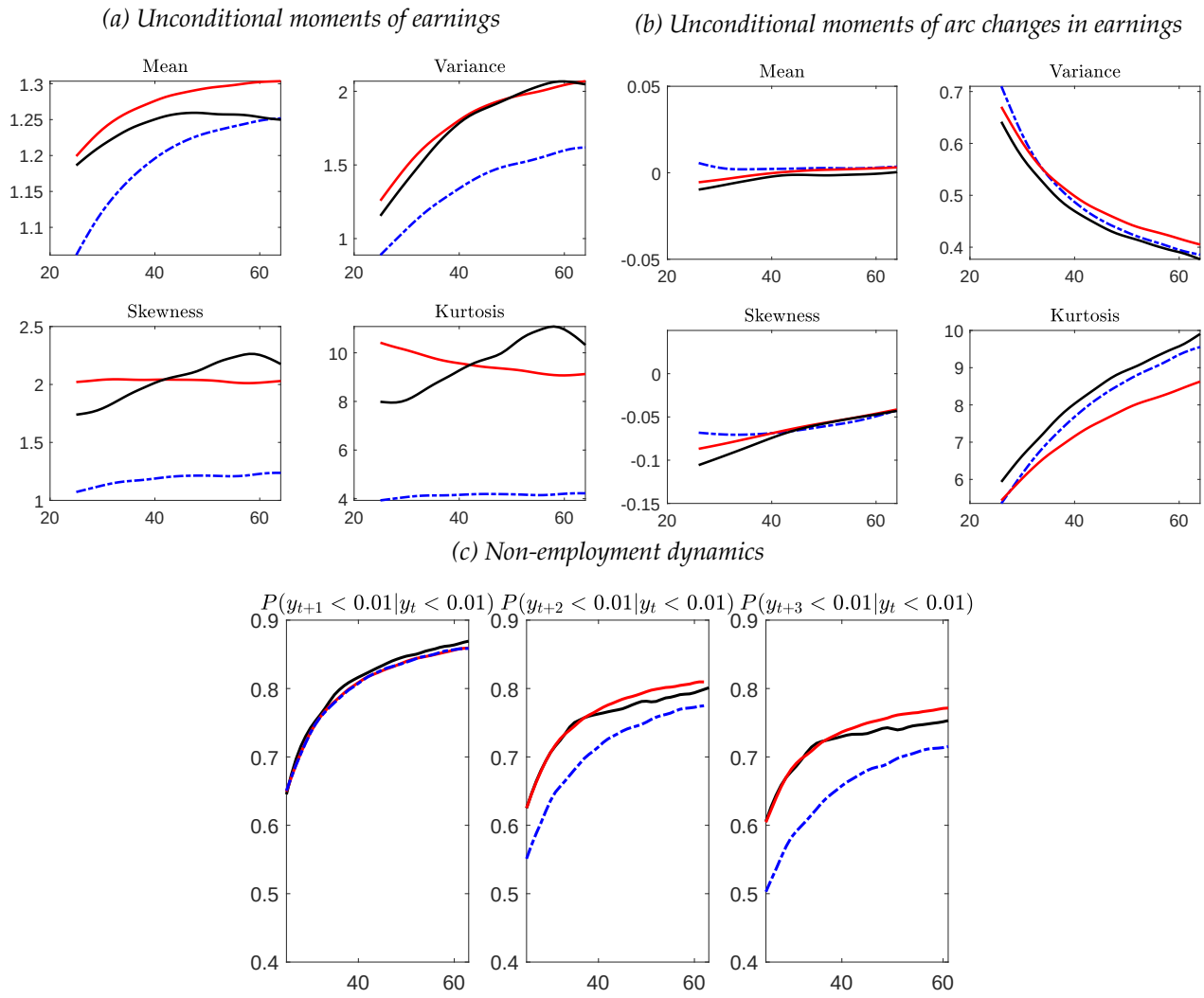
4.3 Discretizing Arellano et al. (2017)

Stochastic process. Next, we consider the nonparametric earnings process in Arellano et al. (2017). As in Arellano et al. (2017), let y_{it} be pre-tax labor earnings. Decompose $\log y_{it}$ as follows:

$$\log y_{it} = \eta_{it} + \varepsilon_{it}, \quad i = 1, ..., N, \quad t = 1, ..., T,$$

[18] Arc-changes are an important statistic in the Guvenen et al. (2021) paper.

Figure 5: Age-dependent moments, for two different discretizations of the stochastic process by Guvenen et al. (2021). $m = 12$ for both methods.



The solid red line represents the Guvenen et al. (2021) process, the solid black line is our discretization method, and the blue dash-dot line is the binning method. Arc changes are defined as $\frac{y_{i,t+1}-y_{i,t}}{(y_{i,t+1}+y_{i,t})/2}$.

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:

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

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 and limiting time-dependence to age-dependence, that is, $Q_t(\eta_{i,t-1}, \tau) = Q(\eta_{i,t-1}, \text{age}_{it}, \tau)$.

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 four age-dependent standard deviations around the mean.[19]

We allow the grids and transition probability matrices to vary by age, as visualized in Figure E2 in the Appendix. The grids are more dispersed than those of the Guvenen et al. (2021) process. In addition, while for the Guvenen et al. (2021) process most age-dependence in the transition probabilities is at the low-earnings states, for the Arellano et al. (2017) process this is at the high earnings states. For example, the highest earnings state becomes more persistent from age 35 onwards.

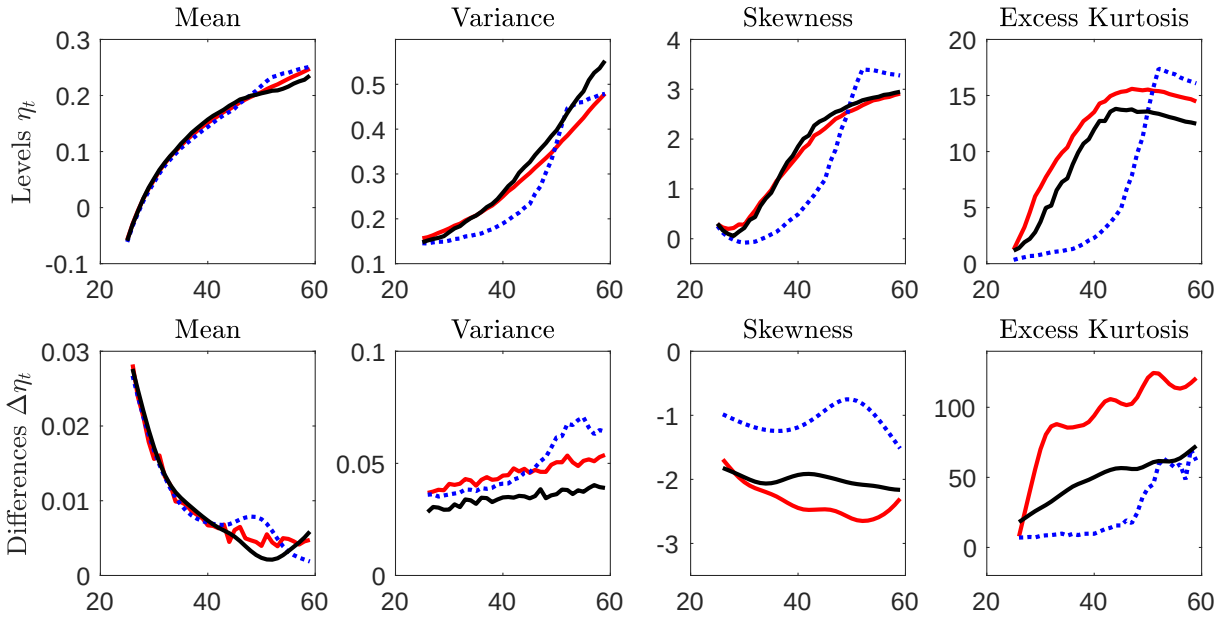
We compare the performance of our discretization method with the method De Nardi et al. (2020) propose to discretize the Arellano et al. (2017) process.[20] In particular, their method adapts Adda and Cooper (2003) and uses simulation-based binning, adding additional bins in the tails of the process. Their discretization for η_{it} uses 18 grid points, and we follow them in this choice. For details we refer to their paper. In what follows below, we refer to this adaptation of binning as "tail-binning".

Figure 6 visualizes the moments of the persistent component η_t and first-differences $\Delta\eta_t$ for the Arellano et al. (2017) process, our discretization and the tail-binning discretization. Our discretization method does a good job at capturing the first four unconditional moments of the levels of η_t. The tail-binning method misses the gradual increase in skewness and kurtosis over the life cycle, and instead catches up by rapidly increasing around age 45-50. Our method does better at capturing the skewness and excess kurtosis of the first-differences of η_t, but does still miss out on some of the excess kurtosis the process exhibits.

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

[20]Note that their discretization originally was applied to a re-estimated version of Arellano et al. (2017) that uses after-tax earnings, so our results are not directly comparable.

Figure 6: Moments of η_t and $\Delta\eta_t$ for the process of Arellano et al. (2017). The red line is data simulated from the Arellano et al. (2017) process, the black line follows from our discretization method, and the blue dotted line is based on the tail-binning method.



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

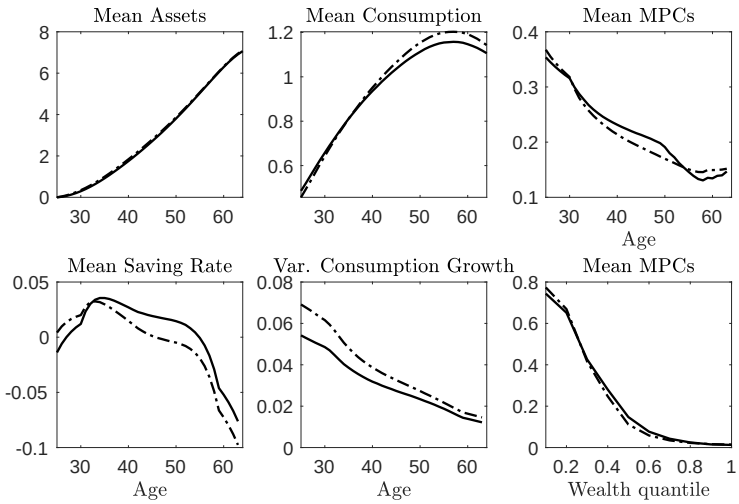
Next, we illustrate the importance of the choice of the discretization method for the earnings process through the lens of the life-cycle model. We use the discretizations of the persistent component of the earnings processes as presented above, and separately add a three-grid-point equal-quantile discretization of the transitory component to the model. Figure 7 plots how assets and consumption develop over the life cycle of an individual using the two different discretization methods. We observe that for both the Guvenen et al. (2021) and Arellano et al. (2017) process, the discretization method has economically meaningful implications for the development of mean consumption over the life-cycle, for the mean MPC's, and for the mean saving rate. In addition, the variance of consumption growth over the life-cycle is also sensitive to the discretization method used; discretizing the Guvenen et al. (2021) process using the binning method results in a larger variance of consumption growth over the life-cycle than our discretization method does. For the Arellano et al. (2017) process, what stands out is the difference in MPC's over the life-cycle between the two methods; our method generates higher MPC's for younger individuals (around 0.7) than the tail-binning method (around 0.6).

Table 5 summarizes several key statistics of the life-cycle model, and how these vary for the two different processes and their different discretizations. For the Guvenen et al. (2021) process,

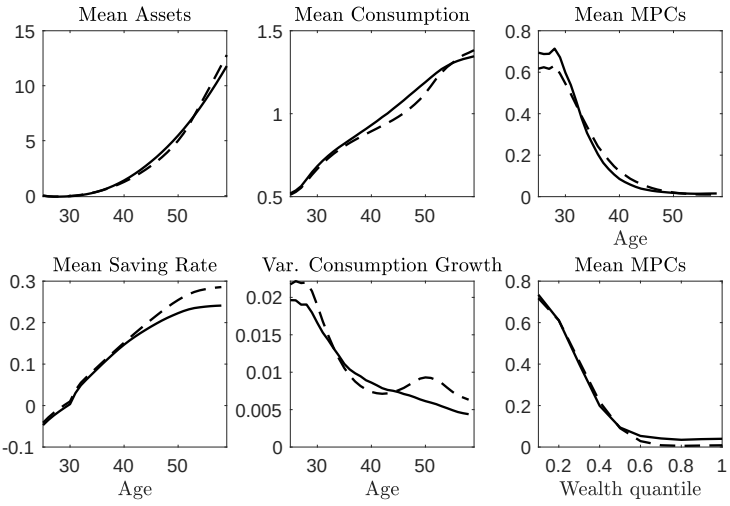
the most notable difference between the discretizations is in the welfare cost of risk (CEV). For our discretization, the CEV is 0.69, while the binning-method based solution implies a CEV that is considerably lower (0.46). We believe this is mainly driven by the binning method discretization understating the amount of longer-term non-employment risk. For the Arellano et al. (2017) process, the CEV estimates also differ between discretization methods; our method implies a CEV of 0.19, and tail-binning results in a CEV estimate of 0.16. Given that the CEV measures the welfare gain in applications with policy experiments, the sensitivity of CEV's to the choice of discretization method is an important finding.

Figure 7: Simulations from the life-cycle model for two different discretizations of the earnings process of Guvenen et al. (2021) and Arellano et al. (2017).

(a) Guvenen et al. (2021) for $m = 12$



(b) Arellano et al. (2017) for $m = 18$



Notes: Solid line represents our discretization method, and the dashed lines are the binning methods.

Table 5: Summary statistics computed from simulations from the life-cycle model for two different discretizations of the earnings processes of Guvenen et al. (2021) and Arellano et al. (2017).

	Model + Guvenen		Model + ABB	
Method	Janssens-McCrary	Binning	Janssens-McCrary	Tail-Binning
St.dev.($\log c_{it}$)	0.77	0.74	0.46	0.41
St.dev.($\Delta \log c_{it}$)	0.17	0.19	0.10	0.11
Corr($\log c_{it}$, $\log y_{it}$)	0.91	0.90	0.95	0.93
Corr($\Delta \log c_{it}$, $\Delta \log y_{it}$)	0.75	0.82	0.78	0.77
CEV	0.69	0.46	0.19	0.16
ψ^P_{BPP}	0.51	0.47	0.66	0.67
Mean MPC	0.22	0.23	0.22	0.21
Discount factor β	0.94	0.94	0.97	0.97

Table 6 summarizes several wealth inequality measures as found in the data (obtained from Krueger, Mitman, and Perri (2016)) and compares them to those computed from the life-cycle model solutions. We find that the discretization method matters for the amount of wealth inequality a life-cycle model can generate. Using binning to discretize both earnings processes results in less wealth inequality than when using our method. Most likely this is because binning misses out on the skewness and excess kurtosis present in the process. The differences between methods are largest for the Arellano et al. (2017) process. When using our discretization method for the Arellano et al. (2017) process, the model matches the wealth distribution of the data fairly well. For example, our discretization results in a wealth Gini index of 0.76, close to the 0.77-0.78 in the data (tail-binning: 0.7), and a top 1% wealth share of 34.5% (tail binning: 27.6), which is actually larger than the 30.9-33.5% reported by Krueger et al. (2016)). The ability of our model solution to match these aspects of the wealth distribution – without targeting it – is notable, given that the literature has documented that simple life-cycle models like this one typically struggle matching the right tail of the empirical wealth distribution (De Nardi and Fella, 2017).

Comparing the earnings process of Guvenen et al. (2021) with Arellano et al. (2017) in the context of a life-cycle model, we find that while both the mean MPC and the mean MPC's over the wealth distribution are comparable between processes, the Guvenen et al. (2021) process implies a flatter MPC profile over the life-cycle than the Arellano et al. (2017) process. This is because the presence of the non-employment shock in the Guvenen et al. (2021) process creates a strong precautionary savings motive for younger generations, resulting in lower MPC's for young ages than in the Arellano et al. (2017) process (0.35 instead of 0.75 for the 25 year olds). This large downside risk in the Guvenen et al. (2021) also result in a higher CEV (0.69) than

Table 6: Wealth inequality measures. Data from Krueger et al. (2016).

	Data		Model + Guvenen		Model + ABB	
% Share held by:	PSID, 06	SCF, 07	Janssens-McCrary	Binning	Janssens-McCrary	Tail-binning
Q1	-0.9	-0.2	-0.7	-0.6	-0.4	-0.3
Q2	0.8	1.2	0.9	1.4	1.5	2.3
Q3	4.4	4.6	6.4	7.6	7.1	9.5
Q4	13.0	11.9	19.6	21.1	15.7	19.2
Q5	82.7	82.5	73.0	70.6	76.0	69.3
T1%	30.9	33.5	10.9	8.9	34.5	27.6
Gini	0.77	0.78	0.73	0.69	0.76	0.70

in the Arellano et al. (2017) process (0.19). While the Guvenen et al. (2021) process by and large focuses on downside earnings risk, the Arellano et al. (2017) process features a longer right-tail, resulting in more wealth inequality (as in Table 6) than the Guvenen et al. (2021) process.

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. Both processes characterize the persistent component as an AR(1) process. The first process uses Gaussian innovations both to the persistent and transitory component (referred to as AR(1) below). The second process has Gaussian mixture innovations for both the persistent and transitory component (henceforth referred to as AR(1)-M). Specifically, for the second process, we use a simplified version of the Guvenen et al. (2021) process in Equation (12), disregarding the non-employment shock, with the same parameters. We parametrize the AR(1) process such that it has the same autocorrelation and variance as the AR(1)-M. In Equations, the AR(1)-M

process is given by:[21]

$$
\begin{aligned}
y_t^i &= e^{(z_t^i + \varepsilon_t^i)} \\
z_t^i &= \rho z_{t-1}^i + \eta_t^i \\
\eta_t^i &\sim \begin{cases} N(\mu_{\eta,1}, \sigma_{\eta,1}) & \text{with prob. } p_z \\ N(\mu_{\eta,2}, \sigma_{\eta,2}) & \text{with prob. } 1 - p_z \end{cases} \\
\varepsilon_t^i &\sim \begin{cases} N(\mu_{\varepsilon,1}, \sigma_{\varepsilon,1}) & \text{with prob. } p_\varepsilon \\ N(\mu_{\varepsilon,2}, \sigma_{\varepsilon,2}) & \text{with prob. } 1 - p_\varepsilon. \end{cases}
\end{aligned} \tag{13}
$$

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 only discretize the persistent component and separately add a three-grid-point equal-quantile discretization of the transitory component to the model.

Comparison with other methods. 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 Toda (2017) and the binning method of Judd (1998)/Adda and Cooper (2003) for the AR(1)-M process.[22]

Figure 8 presents the information loss of each discretization relative to the true process. To compute the information loss, we interpret the transition probability matrix and grid of the different discretization methods as parameters Π and μ in our HMM framework, and then re-estimate the variance of the approximation error. This results in a likelihood for each discretization. We compute this statistic for different grid sizes m.

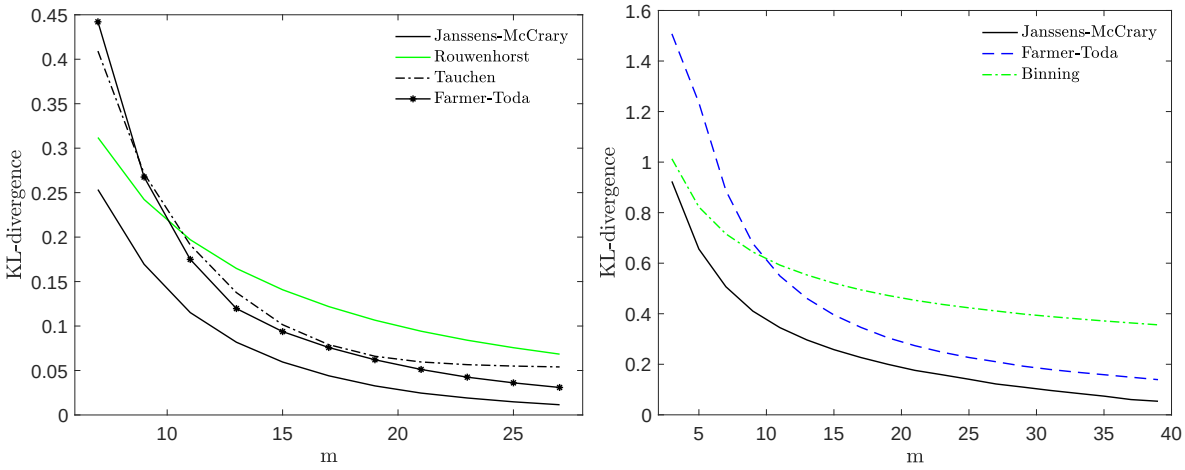
Given that our method minimizes information loss, it is no surprise that our method results in the lowest losses. Figure 8 shows the Farmer and Toda method is, for larger grids, closest to ours in terms of information loss, and the differences in information loss between ours and the alternative methods are large. For the AR(1) process, we achieve the same information loss as the Farmer and Toda method with 27 grid points using only 19. We achieve the same

[21] Parameters: $\rho = 0.959$, $p_z = 0.407$, $\mu_{\eta,1} = -0.085$, $\mu_{\eta,2} = 0.085 p_z/(1-p_z)$, $\sigma_{\eta,1} = 0.364$, $\sigma_{\eta,2} = 0.069$, $p_\varepsilon = 0.13$, $\mu_{\varepsilon,1} = 0.271$, $\mu_{\varepsilon,2} = -0.271 p_\varepsilon/(1-p_\varepsilon)$, $\sigma_{\varepsilon,1} = 0.285$ and $\sigma_{\varepsilon,2} = 0.037$.

[22] All implementations of these methods are standard, except for the grid width we use in the Farmer and Toda (2017) method. For the AR(1) process, we use a grid width equal to $\max\{3, 1.2\log(m-1)\}$ times the standard deviation of the process. This grid width is based on the proposal of Flodén (2008), and we find that this choice works better in this setting than the width of $\sqrt{m-1}$ that Farmer and Toda (2017) propose. For the AR(1)-M process, we use $\max\{4, 1.2\log(m-1)\}$. The Farmer-Toda method is set to match the first four conditional moments at each grid point.

Figure 8: KL divergence of the approximating model likelihood versus the likelihood of the true process for an AR(1) and AR(1)-M process, for different discretization methods and different grid sizes m.

(a) AR(1) *(b) AR(1)-M*



information loss as the Tauchen method with 27 using only 15 grid points, and the same loss as the Rouwenhorst method at 27 using only 13 grid points. For the AR(1)-M, we achieve the same information loss as the Farmer and Toda method at 39 grid points using only 25 grid points, and the same information loss as the binning method at 39 using only 11 grid points. Because the Rouwenhorst method is dominated by the Farmer and Toda method and Tauchen method for larger grids, we drop this method in the analysis that follows below.

Table 7 summarizes some other statistics of the discretized processes, being the unconditional and conditional moments of the distribution. The Farmer-Toda method is based on moment-matching, which is why they tend to perform well at most moments. However, their method is implemented such that when it cannot match a moment in one of the grid points, that specific moment restriction gets dropped for that grid point. This is why there are cases in which it doesn't match all moments even when targeting them, and ours or the Binning/Tauchen method may perform better at matching those moments. One statistic where we consistently outperform the other methods is the Mean Squared Forecast Error (MSFE), that is, if an agent would use the discretized process to make forecasts about the true process, what are the forecast errors the agent makes.

Implications in a life-cycle model. Next, we evaluate how the choice of the discretization method affects the solutions of the life-cycle model, where we focus on a selected number of statistics given in Table 8 and Figure 9. Our main conclusion from Table 8 is that for these two stochastic processes, the choice of the discretization method can matter for the model solution, and particularly when using a low number of grid points. For example, with an

Table 7: Summary statistics on unconditional and conditional moments for different discretizations of the AR(1) and AR(1)-M stochastic processes.

	AR(1)						AR(1)-M					
	$m = 7$			$m = 17$			$m = 17$			$m = 31$		
	JM	FT	T	JM	FT	T	JM	FT	Bin	JM	FT	Bin
Abs. dev. uncond. mean	**<0.01**	<0.01	<0.01	**<0.01**	<0.01	0.01	0.01	<0.01	**<0.01**	0.01	**<0.01**	0.01
% dev. uncond. var.	**5.33**	18.4	56.6	0.09	**<0.01**	11.7	2.99	**0.70**	7.3	2.51	**0.70**	4.04
% dev. uncond. skew.	0.03	**<0.01**	0.04	0.02	**0.01**	0.03	**7.88**	26.3	20.2	**0.81**	8.80	15.3
% dev. uncond. kurt.	**10.8**	61.8	11.3	3.62	**1.45**	7.83	**3.06**	6.36	18.8	1.34	**1.03**	12.7
% dev. autocor.	0.75	**0.07**	1.65	1.28	**0.03**	0.09	0.55	**<0.01**	0.69	0.26	**0.01**	0.29
Ave. abs. dev. cond. mean	1.07	**<0.01**	1.49	1.19	**<0.01**	0.25	0.01	**<0.01**	<0.01	0.01	**<0.01**	0.01
Ave. % dev. cond. var.	14.1	79.6	**4.17**	24.7	**<0.01**	16.1	19.1	**8.74**	19.1	26.1	**8.74**	13.0
Ave. % dev. cond. skew.	**1.12**	3.47	1.45	0.74	1.38	**0.14**	1.01	**0.60**	1.11	**1.13**	1.40	1.20
Ave. % dev. cond. kurt	336	1628	**265**	349	907	**2.55**	122	132	**84.0**	167	485	**84.0**
MSFE	**0.09**	0.12	0.12	**0.07**	0.07	0.07	**0.07**	0.08	0.08	**0.06**	0.07	0.07

Notes: For the skewness moments of the AR(1) process, these are the absolute deviations rather than the % deviation. For the conditional moments, the average is computed across grid points.

AR(1) process, the standard deviation of log consumption, the welfare cost of risk (CEV), the wealth Gini index, and the Wealth Share of the top 20% vary in economically significant ways across the different discretization methods when using a grid size of $m = 7$. When increasing the grid size to $m = 17$, the differences between the methods are smaller. In general, the solutions that follow from our discretization method change little when adding grid points. For the AR(1)-M method, we look at a larger m, because it requires more grid points to get to the same information loss (as shown in Figure 8b). At $m = 17$, the solutions differ for, particularly, the Gini Index, the mean MPC's and Top 1% Wealth Share, but the solutions between discretizations are more similar at $m = 31$. We think the sensitivity of model solutions at low m is an important insight, because it is common in the literature to use discretizations of AR(1) processes with few grid points.

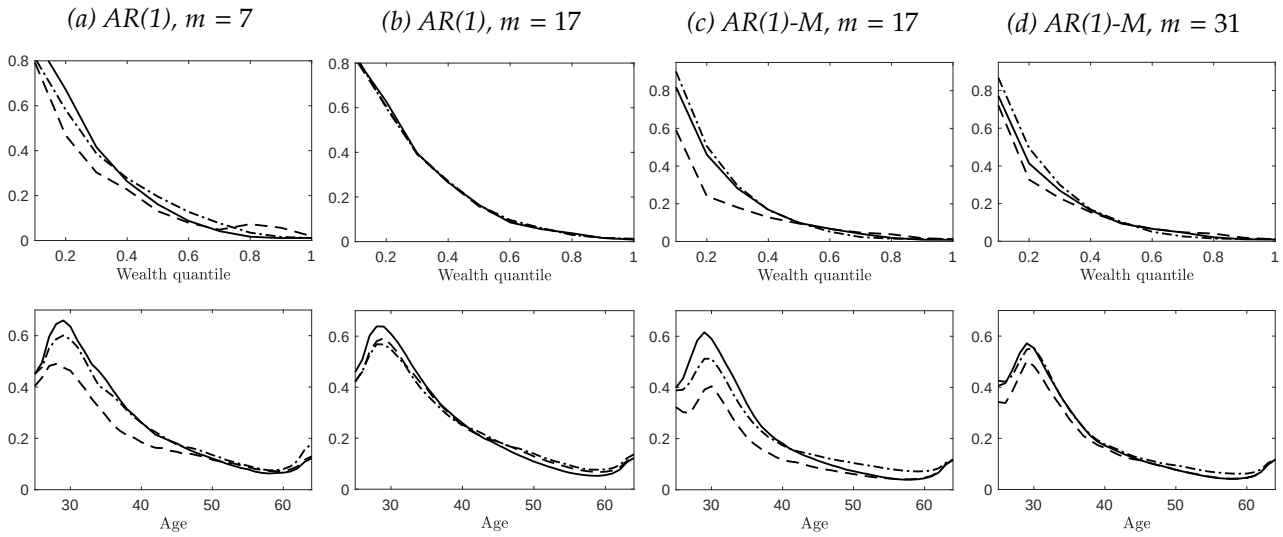
Figure 9 visualizes the mean MPC's across the life-cycle and the wealth distribution for the different AR(1) and AR(1)-M discretizations. We again see that the choice of the discretization method matters, mostly for low m. Interestingly, for the AR(1)-M, the aggregate statistics on consumption and earnings in Table 8 are similar for both $m = 17$ and $m = 31$, and appear insensitive to the choice of discretization, however, we see the life-cycle profile of MPC's and the MPC's across the wealth distribution differ significantly for $m = 17$ and even with $m = 31$ the choice of discretization matters. For some age groups, the mean MPC's can vary as much as 20% between methods. Compared with the other discretization methods, the MPC's that follow from our method change less when adding more grid points, in line with our method

Table 8: Summary statistics computed from simulations from the life-cycle model solved for an AR(1) and AR(1)-M earnings process discretized using different methods.

	Model + AR(1)						Model + AR(1)-M					
	$m = 7$			$m = 17$			$m = 17$			$m = 31$		
	JM	FT	T	JM	FT	T	JM	FT	Bin	JM	FT	Bin
St.dev($\log c_{it}$)	0.70	0.79	0.91	0.72	0.72	0.76	0.71	0.71	0.68	0.71	0.71	0.70
St.dev($\Delta \log c_{it}$)	0.14	0.15	0.15	0.14	0.15	0.16	0.14	0.15	0.15	0.14	0.15	0.15
Corr($\log c_{it}$, $\log y_{it}$)	0.96	0.95	0.97	0.97	0.96	0.96	0.96	0.96	0.95	0.96	0.96	0.96
Corr($\Delta \log c_{it}$, $\Delta \log y_{it}$)	0.78	0.82	0.82	0.79	0.80	0.82	0.89	0.89	0.89	0.89	0.89	0.89
CEV	0.37	0.42	0.46	0.38	0.38	0.39	0.40	0.40	0.38	0.40	0.40	0.39
ψ^P_{BPP}	0.51	0.55	0.46	0.52	0.51	0.52	0.51	0.51	0.52	0.51	0.51	0.52
Mean MPC	0.26	0.21	0.26	0.25	0.25	0.25	0.21	0.15	0.20	0.20	0.18	0.21
Gini index	0.78	0.83	0.81	0.79	0.78	0.78	0.75	0.75	0.72	0.75	0.75	0.73
Q5 Wealth Share	0.80	0.87	0.85	0.82	0.80	0.81	0.76	0.76	0.73	0.76	0.76	0.75
T1% Wealth Share	0.12	0.19	0.16	0.15	0.14	0.14	0.12	0.13	0.09	0.12	0.12	0.10
Discount factor β	0.95	0.93	0.93	0.95	0.94	0.94	0.95	0.94	0.94	0.95	0.94	0.94

Notes: JM stands for Janssens-McCrary, FT stands for the Farmer-Toda method, Bin refers to the binning method, and T stands for the Tauchen method.

Figure 9: Mean Marginal Propensities to Consume for three different discretizations of AR(1) and AR(1)-M processes in a life-cycle model, computed across the wealth distribution and by age.



Notes: The solid black line is our discretization method, and the dashed line is the Farmer-Toda method. For the AR(1), the dash-dot line is the Tauchen method, for AR(1)-M, it is the binning method.

being more parsimonious by capturing a larger fraction of the information using fewer grid points.

Comparison between an AR(1), AR(1)-M and the Guvenen et al. (2021) process. Finally, we use this model to ask what the differences are between an AR(1) and AR(1)-M process in a life-cycle context, that is, what do excess skewness and kurtosis imply for a life-cycle model. Consider Table 8 and the largest choice of m. Most notably, the Gaussian mixture leads to a larger correlation between consumption and income changes. In addition, we see an decrease of the mean MPC, and a decrease in wealth inequality. The intuition behind these results is that the income distribution in the economy with the mixture distribution is less unequal, because the mixture process is skewed towards lower incomes. This results in a less unequal wealth distribution. In addition, the increased left-tail risk increases the correlation between consumption and income changes, and it lowers MPC's because of a stronger precautionary savings motive. Comparing this with the Guvenen et al. (2021) process that features non-employment shocks in addition to Gaussian mixture innovations, we see that non-employment shocks substantially increase the CEV compared to the AR(1)-M process (from 0.40 to 0.69), lowers the wealth Gini index from 0.75 to 0.73, and increases mean MPC's from 0.20 to 0.22 because about 2% more people live hand-to-mouth.

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. A finite-state Markov chain approximation is inherently a misspecified model, and the objective of minimizing the KL divergence is standard in the misspecified model literature. We show that this is a consistent approach in our setting in the sense that under some assumptions, using enough hidden states, the information loss between the approximating Hidden Markov Model and the true stochastic process can be made arbitrarily small. Our discretization method is applicable to a large class of stochastic processes and provides both an optimally selected grid and 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 shown by De Groot (2015), has a closed-form analytical solution. This analytical solution is our benchmark when comparing the solutions based on different discretization methods, and we find our method results in numerical solutions closer to this benchmark. The second application evaluates the effect of the choice of the discretization method on the solutions that follow from a life-cycle model with a variety of different earnings processes, including Guvenen et al. (2021) and Arellano et al. (2017). We

find that the discretization method matters for, among other things, the welfare cost of risk, the marginal propensity to consume, and wealth inequality measures.

Discretized stochastic processes have many more applications than the ones we use to benchmark our method. The econometric literature has shown stochastic processes featuring nonlinearities, 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.

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.

Blundell, R., Pistaferri, L., and Preston, I. (2008). "Consumption Inequality and Partial Insurance". *American Economic Review*, *98*(5), 1887–1921.

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.

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

Hornik, K., Stinchcombe, M., and White, H. (1989). "Multilayer Feedforward Networks are Universal Approximators". *Neural Networks*, *2*(5), 359–366.

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.

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.

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.

Wu, C., and Krueger, D. (2021). "Consumption Insurance Against Wage Risk: Family Labor Supply and Optimal Progressive Income Taxation". *American Economic Journal: Macroeconomics*, *13*(1), 79–113.

Zeevi, A. J., and Meir, R. (1997). "Density Estimation through Convex Combinations of Densities: Approximation and Estimation Bounds". *Neural Networks*, *10*(1), 99–109.

A Proof of Main Theorem

A.1 Preliminaries, notation and existing results

As in Zeevi and Meir (1997), denote

$$\mathscr{F}_{c,\eta} = \{f \in \mathscr{F}_c | f \geq \eta > 0, \forall y \in \mathscr{Y}\}$$

where

$$\mathscr{F}_c = \left\{f | f \in C\mathscr{Y}, f \geq 0, \int f = 1\right\}$$

is the class of continuous density functions with compact support $\mathscr{Y} \subset \mathbb{R}^k$ fixed and given. $\mathscr{F}_{c,\eta} \subset \mathscr{F}_c$ is bounded below over $\mathscr{Y}$ by some positive constant, denoted by η.

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) $f(y_t | y_{t-1}) \in \mathscr{F}_{c,\eta}$.

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

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

(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 $p(y_t|y_{t-1}, ..., y_1; \theta_m) \in \mathscr{F}_{c,\eta}$.

Denote the L_p distance between two functions by

$$d_p(f, g) := \left(\int |f(x) - g(x)|^p dx \right)^{1/p}$$

and the l_p distance between two vectors as $d_p(x, x') = (|x_1 - x'_1|^p + ... + |x_d - x'_d|^p)^{1/p}$, for $p \geq 1$.

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

$$\Phi_{\eta,\tau} = \left\{ \phi_\sigma \in \Phi_\eta | \phi_\sigma = \sigma^{-d} \phi \left(\frac{\cdot - \mu}{\sigma} \right), \mu \in \mathscr{Y}, \sigma \in \mathbb{R} \text{ s.t. } \sigma \geq \tau > 0 \right\}$$

with $\Phi_\eta = \{\phi \in \Phi | \phi \geq \eta > 0, \forall y \in \mathscr{Y}\}$ and $\Phi = \{\phi | \phi \in C(\mathbb{R}^k), \phi > 0, \int \phi = 1\}$ the class of continuous densities. Note $\Phi_{\eta,\tau} \subset \Phi_\eta \subset \Phi$. The approximation class is given by

$$\mathscr{G}_n = \left\{ f_m^\theta | f_m^\theta(\cdot) = \sum_{i=1}^n \alpha_i \phi_\sigma(\cdot; \theta_i), \phi_\sigma \in \Phi_{\eta,\tau}, \alpha_i > 0, \sum_{i=1}^m \alpha_i = 1 \right\}$$

and we write $\theta = (\mu, \sigma)$, where $\mu = [\mu(1), ..., \mu(m)]$. That is, we consider a mixture distribution where all functions have the same scale parameter but a different location. Unlike Zeevi and Meir (1997), we will strictly refer to ϕ as the Gaussian probability density function, which falls into the class of functions they consider. For multivariate distributions, we have $\mu(i) = (\mu^1(i), ...\mu^k(i))$, and $\sigma = (\sigma_1, ..., \sigma_k)$, such that ϕ_σ is the product of k independent Gaussian pdf's, also known as a product kernel.

Define γ such that $\eta = \frac{1}{\gamma^2}$.

Lemma 1 (Eq. 14 in Zeevi and Meir, 1997). *For g, f s.t. $g, f \geq \frac{1}{\gamma^2} > 0$,*

$$D^{KL}(f||g) \leq \gamma^2 d_2^2(f, g).$$

That is, for densities f and g that are bounded below by $\frac{1}{\gamma^2}$, the KL divergence is bounded from above by the squared L_2 norm between f and g multiplied by γ^2.

Lemma 2 (Petersen, 1983 as in Zeevi and Meir, 1997). *Let* $1 \leq p < \infty$ *and let* $\phi \in L_1(\mathbb{R}^k)$, $\int \phi = 1$. *Letting* $\phi_\sigma(x) = \sigma^{-k}\phi(x/\sigma)$, *then for any* $f \in L_p(\mathbb{R}^k)$, *we have* $\phi_\sigma * f \to f$ *in* $L_p(\mathbb{R}^k)$ *as* $\sigma \to 0$ *where*

$$(\phi_\sigma * f)(x) := \int \phi_\sigma(x-y)f(y)dy.$$

Here, $L_1(\mathbb{R}^k)$ and $L_p(\mathbb{R}^k)$ denote the space of measurable functions for which $||f||_1 < \infty$ and $||f||_p < \infty$, respectively. If we define $\bar{f} := f * \phi_\sigma$, Lemma 2 implies $\forall \varepsilon > 0$ and $f \in \mathcal{F}_{c,\eta}$, there exists an $\bar{f}$ such that

$$d_2^2(f, \bar{f}) \leq \varepsilon. \tag{A.1}$$

Corollary 1 (Zeevi and Meir, 1997). *Function* $\bar{f}$ *belongs to the closure of the convex hull of* $\Phi_{\eta,\tau}$.

Lemma 3 (Barron, 1993 as in Zeevi and Meir, 1997). *If* $\bar{f}$ *is in the closure of the convex hull of a set* G *in Hilbert Space, with* $||g||_2 \leq b$ $\forall g \in G$, *then* $\forall m \geq 1$ *and* $\forall c > (b^2 - ||\bar{f}||_2^2)$, $\exists$ *a function* f_m^0 *in the convex hull of* m *points in* G *s.t.*

$$d_2^2(\bar{f}, f_m^0) \leq \frac{c}{m}.$$

Corollary 2 (Zeevi and Meir, 1997). For any $f \in \mathcal{F}_{c,\eta}$ and $\varepsilon > 0$, there exists a convex combination f_m^0 in $\mathcal{G}_m$ s.t.

$$d_2^2(f, f_m^0) \leq \varepsilon + \frac{c}{m}.$$

Note that Corollary 2 follows directly from the triangle inequality and Equation A.1 and Lemma 3. One of the implications of Corollary 2 is that the Gaussian mixture model is a universal approximator in the L_2 norm.

Combining Corollary 2 with Lemma 1, we have:

$$D_{\mathcal{Y}}^{KL}(f||f_m^0)) \leq \gamma^2\varepsilon + \gamma^2\frac{c}{m} \tag{A.2}$$

Note that although the KL divergence is not a strict metric and does not generally satisfy the triangle inequality, the d_2^2 distance function does. We can use the relationship between the d_2^2 metric and the KL divergence laid out by Lemma 1 in Lemma 4 below.

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

In Lemma 4, we provide an upper bound on the L_2 norm and KL divergence between a Gaussian mixture and a function f, when the Gaussian mixture takes a choice of grid points $\tilde{\mu}_m$ and variance and $\tilde{\sigma}_m$ that may not be same as μ_m^0 and σ_m^0 of Corollary 2.

Lemma 4. *Let ϕ_σ denote the Gaussian distribution function (or product of k independent Gaussian distribution functions), f_0^m and f are as defined in Corollary 2, and $\tilde{f}_m$ is the same function as f_m^0, characterized by $(\alpha_m^0, \mu_m^0, \sigma_m^0)$ except it is evaluated in a different μ and variance σ, with elements denoted by $\tilde{\mu}_m(i) \in \mathcal{Y} \subset \mathbb{R}^k$, and $\tilde{\sigma}_m \geq \tau > 0$ but with same mixture weights α_m^0. Then*

$$d_2^2(f, \tilde{f}_m) \leq \varepsilon + \frac{c}{m} + \frac{1}{4}\left(\max_i\{(\mu_m^0(i) - \tilde{\mu}_m(i))'(\tilde{\Sigma}_m^{-1})(\mu_m^0(i) - \tilde{\mu}_m(i))\} + tr(\tilde{\Sigma}_m^{-1}\Sigma_m^0) - k + \ln\frac{|\tilde{\Sigma}_m|}{|\Sigma_m^0|}\right)$$

and

$$D_{\mathcal{Y}}^{KL}(f||\tilde{f}_m) \leq \gamma^2\left(\varepsilon + \frac{c}{m} + \frac{1}{4}\left(\max_i\{(\mu_m^0(i) - \tilde{\mu}_m(i))'(\tilde{\Sigma}_m^{-1})(\mu_m^0(i) - \tilde{\mu}_m(i))\} + tr(\tilde{\Sigma}_m^{-1}\Sigma_m^0) - k + \ln\frac{|\tilde{\Sigma}_m|}{|\Sigma_m^0|}\right)\right)$$

with γ, ε and c given in Lemma 1, 2 and 3, respectively, and $\Sigma = diag(\sigma_1, ..., \sigma_k)$.

Proof. We use the L_2 - L_1 norm inequality: $d_2(f_m^0, \tilde{f}_m) \leq d_1(f_m^0, \tilde{f}_m)$ and Pinsker's inequality: $d_1(f_m^0, \tilde{f}_m) \leq \sqrt{\frac{1}{2}D^{KL}\left(f_m^0||\tilde{f}_m\right)}$. 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:

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

Using that $\sum \alpha_i = 1$:

$$\begin{aligned} d_2^2(f_m^0, \tilde{f}_m) &\leq \frac{1}{4}\max_i D^{KL}(f_m^{0,i}||\tilde{f}_m^i) \\ &\leq \frac{1}{4}\left(\max_i\{(\mu_m^0(i) - \tilde{\mu}_m(i))'(\tilde{\Sigma}_m^{-1})(\mu_m^0(i) - \tilde{\mu}_m(i))\} + \text{tr}(\tilde{\Sigma}_m^{-1}\Sigma_m^0) - k + \ln\frac{|\tilde{\Sigma}_m|}{|\Sigma_m^0|}\right) \end{aligned}$$

Combining this with $d_2^2(f, f_m^0) \leq \varepsilon + \frac{c}{m}$ from Corollary 2, and using the triangle inequality for the L_2 norm and Lemma 1, we conclude. □

As long as σ_m^0 and $\tilde{\sigma}_m$ go to zero at the same rate, and the distance between μ_m^o and $\tilde{\mu}_m$ goes to zero, the expressions in Lemma 4 will converge to those in Corollary 2 and Equation (A.2).

A.3 m Gaussian Mixtures

Lemma 5 extends Lemma 4, applying Lemma 4 to m conditional distributions at the same time.

Lemma 5. *Let $\tilde{\mu}_m(i) \in \mathscr{Y} \subset \mathbb{R}^k$, $i = 1, ..., m$ and $\tilde{\sigma}_m \geq \tau$ be given. Let $f^i \in \mathscr{F}_{c,\eta}$, for $i = 1, ..., m$ denote m distributions and let $f_m^{0,i}$, $i = 1, ..., m$ as in Corollary 2. Let $\tilde{f}_m^i$, $i = 1, ...m$ be the same as $f_m^{0,i}$, but all with the same location parameters $\tilde{\mu}_m$ and scale $\tilde{\sigma}_m \geq \tau$, but their own mixture weights $\alpha_m^{0,i}$. For every $\varepsilon > 0$, there exists $m > 0$ and $m \times m$ matrix with mixture weights $A_m = [\alpha_m^{0,1}, ..., \alpha_m^{0,m}]$ such that for all $j = 1, ..., m$*

$$D_{\mathscr{Y}}^{KL}(f^j || \tilde{f}_m^j)) \leq \gamma^2 \left(\varepsilon^{\max} + \frac{c^{\max}}{m} + ... \right.$$
$$\left. \frac{1}{4} \max_l \left\{ \max_i \{(\mu_m^{0,l}(i) - \tilde{\mu}_m(i))' \tilde{\Sigma}_m)^{-1} (\mu_m^{0,l}(i) - \tilde{\mu}_m(i))\} + tr(\tilde{\Sigma}_m^{-1} \Sigma_m^{0,l}) - k + \ln \frac{|\tilde{\Sigma}_m|}{|\Sigma_m^{0,l}|} \right\} \right) \tag{A.3}$$

and

$$d_2^2(f^j || \tilde{f}_m^j)) \leq \varepsilon^{\max} + \frac{c^{\max}}{m} + ...$$
$$\frac{1}{4} \max_l \left\{ \max_i \{(\mu_m^{0,l}(i) - \tilde{\mu}_m(i))' \tilde{\Sigma}_m)^{-1} (\mu_m^{0,l}(i) - \tilde{\mu}_m(i))\} + tr(\tilde{\Sigma}_m^{-1} \Sigma_m^{0,l}) - k + \ln \frac{|\tilde{\Sigma}_m|}{|\Sigma_m^{0,l}|} \right\} \tag{A.4}$$

where $\varepsilon^{\max} = \max_l \varepsilon^l$ and $c^{\max} = \max_l c^l$, where c^l, $\mu_m^{0,l}$, $\sigma_m^{0,l}$ and $\alpha_m^{0,l}$ are as in Corollary 2:

$$d_2^2(f^l, f_m^{0,l})) \leq \varepsilon^l + \frac{c^l}{m} \tag{A.5}$$

for each $l = 1, ..., m$.

Proof. Equation (A.3) follows from applying Corollary 2 to *each* conditional distribution f^i, $i = 1, ..., m$, such that Equation (A.5) holds for each of these distributions, except with a

different ε^i and c^i, and at different grids and variances (denoted $\mu_m^{0,i}$ and $\sigma_m^{0,i}$ respectively) for each of the $i = 1, ..., m$ conditional distributions. This results in m sets of m mixture weights α_m^0.

In addition to showing Equation (A.5) holds, where we use a different grid $\mu_m^{0,l}$ and variance $\sigma_m^{0,l}$ to fit each conditional distribution $l = 1, ..., m$, we need to argue this result also goes through when evaluating the KL divergence of these distributions all in the same grid $\tilde{\mu}_m$ and variance $\tilde{\sigma}_m$. For this, we use Lemma 4. This gives us

$$
\begin{aligned}
&D_{\mathcal{Y}}^{KL}(f^i || \tilde{f}_m^i)) \\
&\leq \gamma^2 \left(\varepsilon^l + \frac{c^l}{m} + \max_i \{(\mu_m^{0,l}(i) - \tilde{\mu}_m(i))' \tilde{\Sigma}_m)^{-1} (\mu_m^{0,l}(i) - \tilde{\mu}_m(i))\} + \text{tr}(\tilde{\Sigma}_m^{-1} \Sigma_m^{0,l}) - k + \ln \frac{|\tilde{\Sigma}_m|}{|\Sigma_m^{0,l}|} \right)
\end{aligned}
$$

for each l, and similar for the L_2 norm, such that we can take the maximum over all these $l = 1, ..., m$, to get the expression in Equation (A.3) which then holds for each conditional that is evaluated in one of the $j = 1, ..., m$ grid points. □

A.4 Properties of the HMM

Lemma 6. *If $p(\mathbf{y}; \theta)$ as in Assumption (A5) and $\{\mu_m(i)\}_{i=1}^m$ and $\sigma_m \geq \tau > 0$ are such that $\exists\ l \in \{1, ..., m\}$ s.t. $\phi_i(y_{t-1}) < \eta(\sigma_m)/(m-1)$ if $i \neq l$, for each $y_{t-1} \in \mathcal{Y}$, and $\sum_{i=1}^m \phi_i(y_{t-1}) = K(\sigma_m)$, where $\eta(\sigma_m) \to 0$ as $\sigma_m \to 0$ and K non-increasing in σ_m, then 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:

$$\frac{\partial \log p(y_t|y_{t-1}, y_{t-2}, ..., y_1; \theta)}{\partial y_{t-h}} = \frac{1}{p(y_t|y_{t-1}, y_{t-2}, ..., y_1; \theta)} \frac{\partial p(y_t|y_{t-1}, y_{t-2}, ..., y_1; \theta)}{\partial y_{t-h}}, \tag{A.6}$$

where

$$\frac{\partial p(y_t|y_{t-1}, y_{t-2}, ..., y_1; \theta)}{\partial y_{t-h}} = \sum_{i=1}^{m} \phi_i(y_t) \sum_{j=1}^{m} \Pi_{ji} \frac{\partial P(x_{t-1} = j|y_{t-1}, ..., y_1)}{\partial y_{t-h}} \tag{A.7}$$

with, for $h = 1$:

$$\frac{\partial P(x_{t-1} = i|y_{t-1}, ..., y_1)}{\partial y_{t-1}} = \frac{B_{t-1}\phi_i'(y_{t-1}) \sum_{j=1}^{m} \Pi_{ji} P(x_{t-2} = j|y_{t-2}, ..., y_1) - A_{it-1} \sum_{l=1}^{m} \phi_l'(y_{t-1}) \sum_{j=1}^{m} \Pi_{jl} P(x_{t-2} = j|y_{t-1}, ..., y_1)}{B_{t-1}^2} \tag{A.8}$$

The expression in Equation (A.6) 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 $h \geq 2$:

$$\frac{\partial P(x_{t-1} = i|y_{t-1}, ..., y_1)}{\partial y_{t-h}} = \frac{B_{t-1}\phi_i(y_{t-1}) \sum_{j=1}^{m} \Pi_{ji} \frac{\partial P(x_{t-2} = j|y_{t-2}, ..., y_1)}{\partial y_{t-h}} - A_{it-1} \sum_{l=1}^{m} \phi_l(y_{t-1}) \sum_{j=1}^{m} \Pi_{jl} \frac{\partial P(x_{t-2} = j|y_{t-2}, ..., y_1)}{\partial y_{t-h}}}{B_{t-1}^2} \tag{A.9}$$

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 $C_{it-1} := \phi_i(y_{t-1}) \sum_{j=1}^{m} \Pi_{ji} \frac{\partial P(x_{t-2}=j|y_{t-2},...,y_1)}{\partial y_{t-h}}$ and $D_t = \sum_i C_{it-1}$. Denote $C_{\text{low},i} = \frac{\eta(\sigma_m)}{m} \sum_{j=1}^{m} \Pi_{ji} \frac{\partial P(x_{t-2}=j|y_{t-2},...,y_1)}{\partial y_{t-h}}$ and $A_{\text{low},i} = \frac{\eta(\sigma_m)}{m-1} \sum_{j=1}^{m} \Pi_{ji} P(x_{t-1} = i|y_{t-1}, ..., y_1) < \frac{\eta(\sigma_m)}{m-1}$. We rewrite Equation (A.9) 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 $\phi_i(y_{t-1}) < \eta(\sigma_m)/(m-1)$, we have $B_{t-1}C_{it-1} < B_{t-1}C_{\text{low},i}$ and $A_{it-1}D_{t-1} < A_{\text{low},i}D_{t-1}$. Both $A_{\text{low},i}$ and $C_{\text{low},i}$ are decreasing in m, so in this case Equation (A.9) is decreasing in m. On the other hand, if i is such that $\phi_i(y_{t-1}) > (K - \eta(\sigma_m))$, 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 $B_{t-1} - A_{i,t-1} < \eta(\sigma_m) \sum_{k\neq i} \sum_{j=1}^{m} \Pi_{jk} P(x_{t-1} = i|y_{t-1}, ..., y_1)$ and $D_{t-1} - C_{i,t-1} < \eta(\sigma_m) \sum_{k\neq i} \sum_{j=1}^{m} \Pi_{jk} \frac{\partial P(x_{t-2}=j|y_{t-2},...,y_1)}{\partial y_{t-h}}$. Both terms in the numerator are decreasing towards zero in m. Note that B_{t-1} is bounded by $K(\sigma_m)$. Thus, in both cases, the derivative in Equation (A.9) decreases in m, so the Lipschitz coefficient of $\log p(y_t|y_{t-1}, ..., y_1; \theta_m)$ to y_{t-h}, $h \geq 2$ is decreasing in m. □

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 $\partial p(y_t|y_{t-1}, ..., y_1; \theta)/\partial y_{t-h}$ 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.

Corollary 3. *Under the assumptions of Lemma 6, the Hellinger distance between the conditional distribution $p(y_t|y_{t-1}, ..., y_1; \theta)$ and the Gaussian mixture $p^0(y_t|y_{t-1}; \theta) := \sum_{j=1}^{m} \phi_j(y_t)\Pi_{lj}$ with mixture weights $\{\Pi_{lj}\}_{j=1}^{m}$, with l as in Lemma 6, approaches zero as m becomes large.*

Note that $p(y_t|y_{t-1}, ..., y_1; \theta)$ is a Gaussian mixture with convex mixture weights $\sum_{i=1}^{m} \left(\Pi_{ij} P(x_{t-1} = i|y_{t-1}, y_{t-2}, ..., y_1)\right)$, $j = 1, ..., m$. From Lemma 6, the l_2-norm between the mixture weights $\sum_{i=1}^{m} \left(\Pi_{ij} P(x_{t-1} = i|y_{t-1}, y_{t-2}, ..., y_1)\right)$ and Π_{lj} goes to zero when m large. From this, it follows that $d_2(p(y_t|y_{t-1}, ..., y_1; \theta), p^0(y_t|y_{t-1}; \theta))$ is decreasing in m.

A.5 The KL divergence is a function of all conditional KL divergences

Lemma 7. *Under assumption (A1) and (A5), if $D_{\mathscr{Y}}^{KL}(f(y_t|y_{t-1})||p(y_t|y_{t-1}, ..., y_1; \theta_m))$ is bounded and can be made arbitrarily small for any sequences $\{y_k\}_{k=1}^{t-1}$ for all t, then $D_{\mathscr{Y}}^{KL}(f(\mathbf{y})||p(\mathbf{y}; \theta_m))$ 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:

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

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

A.6 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 $\mu_m \in \mathscr{Y}$, 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 $p(\mathbf{y};\theta)$ on the compact*

subset $y \in \mathcal{Y} \subset \mathbb{R}^k$, given by

$$D_{\mathcal{Y}}^{KL}(f(\mathbf{y})||p(\mathbf{y};\theta)) = \int_{\mathcal{Y}} f(\mathbf{y}) \log \frac{f(\mathbf{y})}{p(\mathbf{y};\theta)} d\mathbf{y},$$

can be made arbitrarily small.

Proof. By Corollary 3, as m becomes large, $d_2(p(y_t|y_{t-1}, ..., y_1; \theta_m), p^0(y_t|y_{t-1}; \theta))$ goes to zero. Next, we apply Lemma 5 to the m conditional distribution functions $f(y_t|y_{t-1} = \mu_m(i))$ and $p^0(y_t|y_{t-1} = \mu_m(i); \theta_m)$ for $i = 1, ...m$. By Lemma 5, the L_2 norm between these m conditional distributions is bounded and becomes arbitrarily small as m becomes large. This holds for a grid μ_m on $\mathcal{Y}$ for which the grid points get closer together as m grows larger, and $\sigma_m \geq \tau > 0$ approaches zero. By the triangle inequality,

$$\begin{aligned}
&d_2(f(y_t|y_{t-1} = \mu_m(i)), p(y_t|y_{t-1} = \mu_m(i), ..., y_1; \theta_m)) \leq \\
&d_2(p^0(y_t|y_{t-1} = \mu_m(i); \theta_m), p(y_t|y_{t-1} = \mu_m(i), ..., y_1; \theta_m)) + ... \\
&d_2(p^0(y_t|y_{t-1} = \mu_m(i); \theta_m), f(y_t|y_{t-1} = \mu_m(i)))
\end{aligned}$$

Together with Lemma 1, this implies $D_{\mathcal{Y}}^{KL}(f(y_t|y_{t-1} = \mu_m(i)), p(y_t|y_{t-1} = \mu_m(i), ..., y_1; \theta_m))$ 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 $y_{t-1}, y_{t-2}, ..., y_1$ in the compact set $\mathcal{Y}$ also becomes small.

By Assumptions (A3)-(A4), and Lemma 6, 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 here that 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. As can be seen, these three terms converge to zero as the grid points are closer together, because then the maximum distance $|y - \mu_m(i)|$ also goes to zero, so $O(\cdot)$ will also converge to zero.

By Lemma 6, if $D_{\mathscr{Y}}^{KL}(f(y_t|y_{t-1})||p(y_t|y_{t-1}, \{y_{t-k}\}_{k=2}^{t-1}; \theta_m))$ can be made arbitrarily small for m large enough, then $D_{\mathscr{Y}}^{KL}(f(y_t|y_{t-1})||p(y_t|y_{t-1}, \{\tilde{y}_{t-k}\}_{k=2}^{t-1}; \theta_m)$ 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 $D_{\mathscr{Y}}^{KL}(f(y_t|y_{t-1})||p(y_t|y_{t-1}, \{\tilde{y}_{t-k}\}_{k=2}^{t-1}; \theta_m)$ 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 7 to the conditional KL divergences concludes the proof. □

B Estimation procedures

B.1 Estimation of HMM's using the EM algorithm

We first discuss the general procedure we use for the estimation of the HMM. We omit the panel data dimension and assume all parameters are constant. Let $\phi_j(y_t) = P(y_t|x_t = j)$ denote the density of y_t conditional on x_t being in state j. That is,

$$\phi_j(y_t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y_t - \mu(j))^2}, \tag{B.1}$$

if $k = 1$, or $\det(2\pi\Sigma)^{-\frac{1}{2}} e^{-\frac{1}{2}(y_t - \mu(j))'(\Sigma)^{-1}(y_t - \mu_j)}$ for $k > 1$, where $\Sigma_t = \text{diag}(\sigma^2)$. It will be useful to think of the following matrix form for the observation densities:

$$\mathbf{\Phi}(y_t) = \begin{pmatrix} \phi_1(y_t) & & 0 \\ & \ddots & \\ 0 & & \phi_m(y_t) \end{pmatrix}, \tag{B.2}$$

that is, $\mathbf{\Phi}$ is an $m \times m$ diagonal matrix with the observation densities as diagonal elements.

Denote bold variables $\boldsymbol{y} = \{y_1, y_2, \ldots, y_T\}$ and $\boldsymbol{x} = \{x_1, x_2, \ldots, x_T\}$ as realizations of this random process. The complete data likelihood of the model in Equation (2) is given by

$$\mathcal{L}(\theta|\boldsymbol{y}, \boldsymbol{x}) = p(\boldsymbol{y}, \boldsymbol{x}|\theta) = p(\boldsymbol{y}|\boldsymbol{x}, \theta)p(\boldsymbol{x}|\theta), \tag{B.3}$$

and the maximum likelihood estimator is given by

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\theta|\boldsymbol{y}, \boldsymbol{x}). \tag{B.4}$$

If the latent states $\boldsymbol{x}$ were observed, the log-likelihood would be straightforward to maximize. This is because the log-likelihood is given by

$$\log\left(\mathcal{L}(\theta|\boldsymbol{y}, \boldsymbol{x})\right) = \log\left(p(\boldsymbol{y}|\boldsymbol{x}, \theta)\right) + \log\left(p(\boldsymbol{x}|\theta)\right), \tag{B.5}$$

and, conditional on $\boldsymbol{x}$, the parameters Π do not influence $\boldsymbol{y}$ and, similarly, the parameters $(\boldsymbol{\mu}, \sigma)$ do not matter for $\boldsymbol{x}$. Together, this implies the log-likelihood is given by

$$\log\left(\mathcal{L}(\theta|\boldsymbol{y}, \boldsymbol{x})\right) = \log\left(p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\mu}, \sigma)\right) + \log\left(p(\boldsymbol{x}|\Pi)\right). \tag{B.6}$$

That is, the parameters governing the observation equation and state transition equation could be solved for separately, given $\boldsymbol{x}$. Intuitively, if the states $\boldsymbol{x}$ are observed, one could estimate Π using only data on transitions from $\boldsymbol{x}$, estimate $\mu(j)$ by averaging the y_t that are observed when x_t is in state j, and then estimate Σ using the sample variance of the observations $\boldsymbol{y}$ demeaned by the estimates of $\boldsymbol{\mu}$.

In practice, the latent states $\boldsymbol{x}$ are unobservable, but we can use the EM algorithm to maximize the likelihood. The EM algorithm iterates between updating the posterior distribution over the latent states $p_{\boldsymbol{x}} = p(\boldsymbol{x}|\boldsymbol{y}, \theta)$ taking the parameters and observations $(\boldsymbol{y}, \theta)$ as fixed in the E step, and updating the parameters $\theta^{(i)} \to \theta^{(i+1)}$ taking the latent states and observations $(p_{\boldsymbol{x}}, \boldsymbol{y})$ as fixed in the M step.

We now describe the E-step. Let $y^t = (y_1, y_2, \ldots, y_t)$, i.e., the observed values up to time t. Similarly, let $y_{t+1}^T = (y_{t+1}, y_{t+2}, \ldots, y_T)$, i.e., the observed values from time $t+1$ to T. The forward probabilities $\boldsymbol{\alpha}_t(j)$ are given by

$$\boldsymbol{\alpha}_t(j) = p\left(y^t, x_t = j|\theta\right) \tag{B.7}$$

and the backward probabilities $\boldsymbol{\beta}_t(k)$ are given by

$$\boldsymbol{\beta}_t(k) = p\left(y_{t+1}^T|x_t = k, \theta\right). \tag{B.8}$$

These are defined recursively as:

$$\begin{aligned} \boldsymbol{\alpha}_1(j) &= \delta_{1,j}\phi_j(y_1) & \boldsymbol{\beta}_T(k) &= 1 \\ \boldsymbol{\alpha}_{t+1}(j) &= \left(\sum_{k=1}^{m} \boldsymbol{\alpha}_t(k)\Pi_{kj}\right)\phi_j(y_{t+1}), & \boldsymbol{\beta}_t(k) &= \sum_{j=1}^{m} \Pi_{kj}\phi_j(y_{t+1})\boldsymbol{\beta}_{t+1}(j), \end{aligned} \tag{B.9}$$

or in matrix form

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\Pi\boldsymbol{\Phi}(y_t) \quad \text{and} \quad \boldsymbol{\beta}_t' = \Pi\boldsymbol{\Phi}(y_{t+1})\boldsymbol{\beta}_{t+1}'. \tag{B.10}$$

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

$$p(\boldsymbol{y}, x_t = k|\theta) = \boldsymbol{\alpha}_t(k)\boldsymbol{\beta}_t(k). \tag{B.11}$$

This leads to a posterior probability of being in state k, given by

$$\gamma_t(k) = p(x_t = k|\boldsymbol{y}, \theta) = \frac{p(\boldsymbol{y}, x_t = k|\theta)}{p(\boldsymbol{y}|\theta)} = \frac{p(\boldsymbol{y}, x_t = k|\theta)}{\sum_{j=1}^{m} p(\boldsymbol{y}, x_t = j|\theta)} = \frac{\boldsymbol{\alpha}_t(k)\boldsymbol{\beta}_t(k)}{\sum_{j=1}^{m} \boldsymbol{\alpha}_t(j)\boldsymbol{\beta}_t(j)}. \tag{B.12}$$

We can also define the posterior transition probability between state i at time t and state j at time $t+1$ as

$$\begin{aligned} \xi_t(k, j) &= p(x_{t+1} = j, x_t = k|\boldsymbol{y}, \theta) \\ &\propto \boldsymbol{\beta}_{t+1}(j)\phi_j(y_{t+1})\Pi_{kj}\boldsymbol{\alpha}_t(k), \end{aligned} \tag{B.13}$$

where the last line follows from the definition of $\gamma_t(k)$ from above.

At last, the M step is given by

$$\mu_l(j) = \frac{\sum_{t=1}^{T} y_{l,t}\, p(x_t = j|\boldsymbol{y}, \theta)}{\sum_{t=1}^{T} p(x_t = j|\boldsymbol{y}, \theta)} = \frac{\sum_{t=1}^{T} y_{l,t}\gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{B.14}$$

$$(\sigma_l)^2 = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{m}(y_{l,t} - \mu_l(j))^2\gamma_t(j) \tag{B.15}$$

$$\Pi_{qj} = \frac{\sum_{t=2}^{T} p(x_t = j, x_{t-1} = q|\boldsymbol{y}, \theta)}{\sum_{t=2}^{T} p(x_{t-1} = q|\boldsymbol{y}, \theta)} = \frac{\sum_{t=2}^{T} \xi_t(q, j)}{\sum_{t=2}^{T} \gamma_t(q)}, \tag{B.16}$$

for $l = 1, ..., k$ subscripts denoting different elements of the vector $y_t = (y_{1t}, ..., y_{kt})$.

Given the updated transition matrix Π_t we can update the stationary probabilities as

$$\boldsymbol{\delta} = \mathbf{1}' \left(I_m - \Pi + U \right)^{-1} . \tag{B.17}$$

Here U is an $m \times m$ matrix of ones.

Note that this setting can be adapted to allow for the discretization of age-dependent earnings processes, with age-dependent transition probabilities and grid placement. In this case, the asymptotics depend on N and a different transition probability matrix and grid are estimated for every age group. However, in practice this implies the estimation of many parameters, which is why, for the estimation of models with rich life-cycle dynamics, we will use an iterative adaption of this algorithm. This algorithm is described in Appendix Section B.2. The idea behind this algorithm is that it only uses data on two time periods at a time, but passes the estimates for the filtered states on to the next time period.

B.2 A multi-step EM algorithm for HMM

In this subsection, we outline the multi-step EM algorithm we use for the estimation of the HMM in case of life-cycle dynamics, where the transition matrix Π_t and grid μ_t are allowed to vary over the life-cycle. The large number of parameters to be estimated here requires N to be large, and the EM algorithm has to converge for many parameters. A multi-step algorithm provides more stability.

Assume a panel of $y_{it} \in \mathbb{R}^k$, $t = 1, ..., T$ and $i = 1, ..., N$. Assume a given grid size m. Initialization:

- Estimate a Gaussian Mixture Model on y_{i1}, $i = 1, ..., N$. This gives a grid for the first time period and iteration, μ_1^1, stationary probabilities δ_1^1 and the filtered probabilities α_1^1. Set iteration $j = 1$.

We have a forward and backward step. For the forward step, set $t = 1$ and:

- Estimate the HMM of Section B for (y_{it}, y_{it+1}), $i = 1, ..., N$, restricting the grid of time period t to μ_t^j, the stationary probabilities of time period t to δ_t^j, the forward probabilities to α_t^j (except for $t = 1$, in which case they follow from Equation (B.9)). For $j > 1$, also restrict the backward probabilities for $t + 1$ to those obtained from the backward step, β_{t+1}^{j-1}, else set to 1. Estimate and store the grid μ_{t+1}^j, the transition probability matrix Π_t^j,

stationary probabilities δ^j_{t+1}, and forward probabilities α^j_{t+1}. Set $t = t + 1$ and repeat up until and including $t = T - 1$.

For the backward step, set $t = T$ and:

- Estimate the HMM of Section B for (y_{it-1}, y_{it}), $i = 1, ..., N$, restricting the grid of time period t to μ^j_t, the stationary probabilities of time period t to δ^j_t, the forward probabilities to α^j_t, the backward probabilities to β^j_t (for $t < T$). When $t = T$, all of these (except the backward probabilities) come from the last time period of the forward step. Estimate and store the grid μ^j_{t-1}, the transition probability matrix Π^j_{t-1}, stationary probabilities δ^j_{t-1}, and backward probabilities β^j_{t-1}. Set $t = t - 1$ and repeat up until and including $t = 2$.

Once can iterate multiple times between the forward and backward step until they stabilize. In that case, update $j = j + 1$.

C Discretization of a VAR process

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

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

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

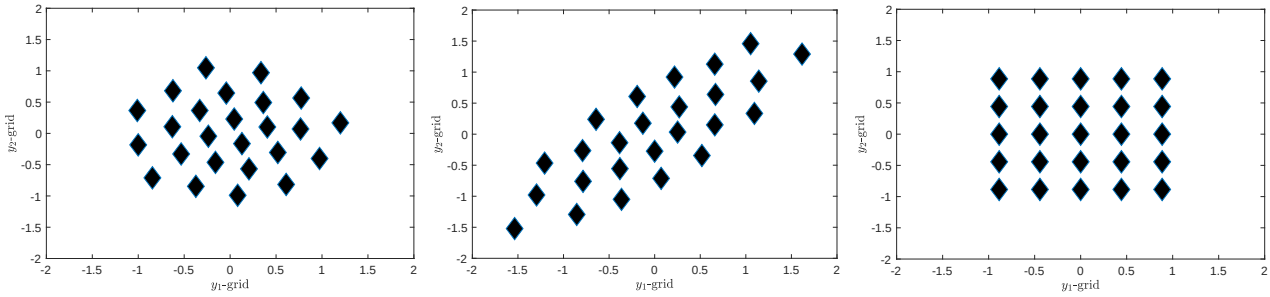
where $\varepsilon_t \sim N(0, \Sigma)$.

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 C1. 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 C1b shows how this is reflected in our optimal grid, while a standard tensor grid as in Figure C1c does not reflect this dependence.

Table C1 summarizes the performance of our discretization compared to the discretization of Farmer and Toda (2017) for two different parametrizations of the VAR model in Equation (C.1). In their discretization, Farmer and Toda (2017) target the first four conditional moments.

Figure C1: Visualisation of optimal grid for two different parametrizations of the data generating process in Equation (C.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.

As we can see, they outperform our discretization method in the first two conditional and unconditional moments, but for higher order moments, our method tends to be closer to the true process. Our method also has a smaller mean squared forecast error.

Table C1: Comparison for VAR model in Equation (C.1) *for* $m = 25$ ($m_{y_1} = 5,\ m_{y_2} = 5$ *for Farmer-Toda).*

Method	Janssens-McCrary	Farmer-Toda
Parametrization 1		
Abs. dev. uncond. mean y	0.103	$<$ **0.001**
% dev. uncond. variance y	-0.042	**0.083**
% dev. autocorrelation y	**0.111**	-0.264
Abs. dev. uncond. skewness y	**0.009**	-0.046
% dev. uncond. kurtosis y	**-0.049**	0.097
Abs. dev. correlation(y_1, y_2)	0.059	**0.007**
Abs. dev. cond. mean y	0.041	$<$ **0.001**
% abs. dev. cond. variance y	37.8	**23.3**
% abs. dev. cond. skewness y	**0.256**	0.466
% abs. dev. cond. kurtosis y	**12.7**	49.3
MSFE y	**0.109**	0.133
Parametrization 2		
Abs. dev. uncond. mean y	0.144	$<$ **0.001**
% dev. uncond. variance y	-0.122	**0.034**
% dev. autocorrelation y	**0.074**	-0.110
Abs. dev. uncond. skewness y	**-0.046**	0.278
% dev. uncond. kurtosis y	-0.030	**-0.015**
Abs. dev. correlation(y_1, y_2)	0.093	**-0.007**
Abs. dev. cond. mean y	0.030	$<$ **0.001**
% abs. dev. cond. variance y	38.3	**9.86**
% abs. dev. cond. skewness y	**0.337**	0.609
% abs. dev. cond. kurtosis y	**43.4**	81.7
MSFE y	**0.120**	0.211

Notes: Parametrization 1: $\beta_{11} = 0.7\ \beta_{12} = 0.2,\ \beta_{21} = 0.0,\ \beta_{22} = 0.7$. Parametrization 2: $\beta_{11} = 0.7\ \beta_{12} = 0.2,\ \beta_{21} = 0.2,\ \beta_{22} = 0.7$. The statistics average over y_1 and y_2.

D Asset Pricing Model with Stochastic Volatility

D.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 (7)-(8). 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

$$B_i = \left(\frac{1-\gamma}{1-\rho}\right)\rho(1-\rho^i)$$

$$C_i = \frac{1}{2}\left(\frac{1-\gamma}{1-\rho}\right)^2\left(i - 2\rho\frac{1-\rho^i}{1-\rho} + \rho^2\frac{1-\rho^{2i}}{1-\rho^2}\right)$$

$$D_i = \frac{\rho_\eta}{2}\left(\frac{1-\gamma}{1-\rho}\right)^2\left(\phi_1 + \phi_2\rho_\eta\rho_\eta^{i-1} + \phi_3\rho^{i-1} + \phi_4\rho^{2(i-1)}\right)$$

$$H_i = F_i\omega^2$$

where

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

and

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

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

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

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.

D.2 A discretized solution

Instead of solving the model using the continuous-support process in Equations (7)-(8), 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 (9) as

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

which solves to

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

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

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

E Age-dependent transition probabilities and grids

Figure E1: Visualisation of the age-dependent transition probabilities for a discretization of the stochastic process in Guvenen et al. (2021), with $m = 12$ grid points. The order of the matrix corresponds with a sorted (low-to-high) earnings grid, where the three lowest states are zero-earnings states. Age on the x-axis of all figures.

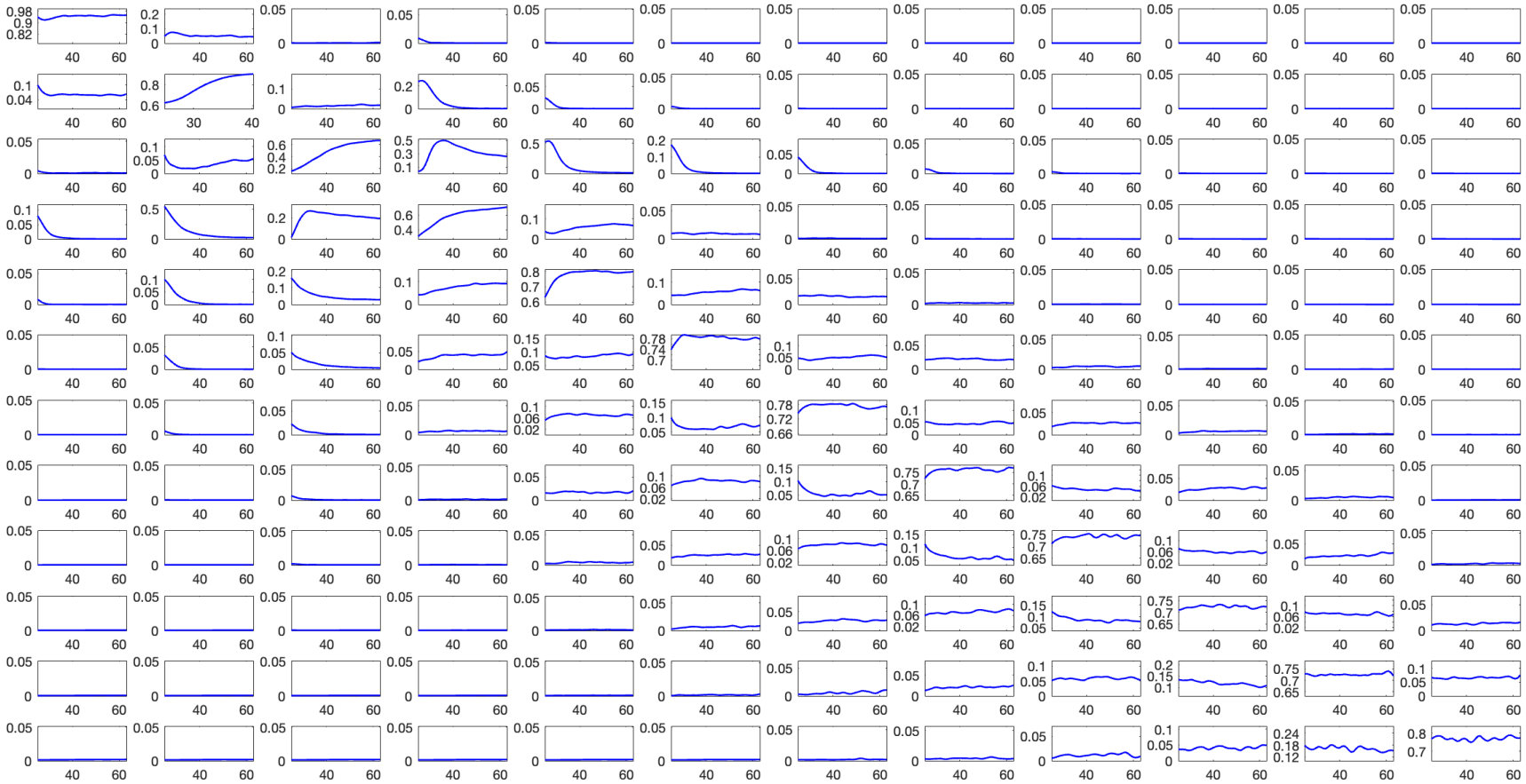
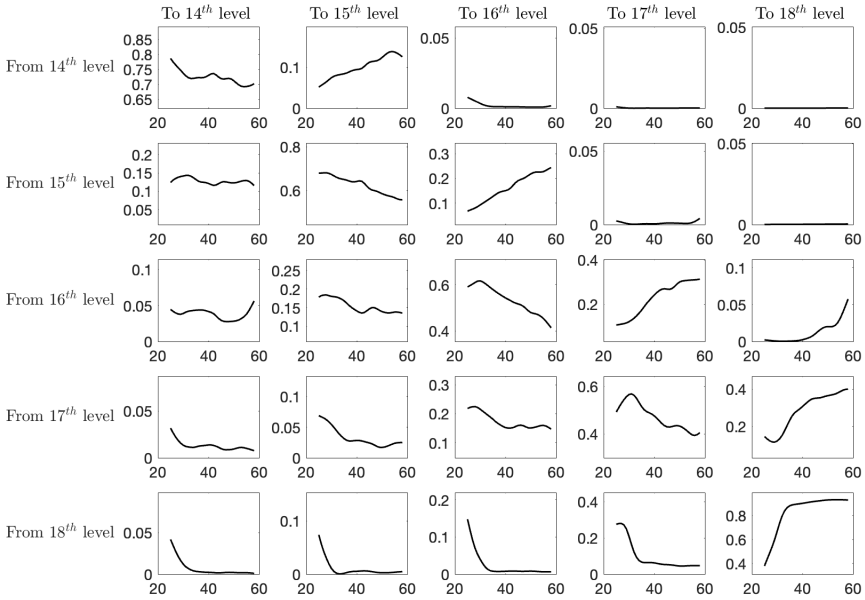


Figure E2: Visualisation of (selected) age-dependent transition probabilities and the age-dependent grid of the $m = 18$-discretization of the stochastic process in Arellano et al. (2017). Age on the x-axis of all figures.

(a) Transition probabilities of the five top earnings states



(b) Age-dependent grid (in logs)