UNIVERSITY OF CALIFORNIA, SAN DIEGO

Discrete Methods for the Estimation of Nonlinear Economic Models

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy

in

Economics

by

Leland Edward Farmer

Committee in charge:

Professor James D. Hamilton, Co-Chair
Professor Allan Timmermann, Co-Chair
Professor Dimitris N. Politis
Professor Alexis Akira Toda
Professor Rossen Valkanov

2017

The Dissertation of Leland Edward Farmer is approved and is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Co-Chair

University of California, San Diego

2017

DEDICATION

This dissertation is dedicated to my parents, Roger E. A. and Roxanne Farmer, for their constant love and support. You have always been there for me through the highest highs and lowest lows and without you, none of this would be possible.

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Table of Contents v

List of Figures viii

List of Tables x

Acknowledgements xii

Vita xiii

Abstract of the Dissertation xiv

Chapter 1 The Discretization Filter: A Simple Way to Estimate Nonlinear State Space Models 1
1.1 Introduction 1
1.2 Related Literature 4
1.3 The Discretization Filter 5
1.3.1 The Setting 6
1.3.2 Approximating the State Dynamics 8
1.3.3 Evaluating the Likelihood 10
1.4 Asymptotic Properties of the Maximum Likelihood Estimator 12
1.4.1 Preliminaries and Assumptions 13
1.4.2 Consistency 16
1.4.3 Asymptotic Normality 24
1.5 Recommendations for Applied Researchers 26
1.5.1 Choosing the Number of Grid Points 26
1.5.2 Selecting the Grid Points 27
1.5.3 Constructing the Transition Matrix 30
1.6 Monte Carlo Evidence 32
1.6.1 Measuring GDP: A Linear State Space Example 32
1.6.2 Stochastic Volatility 36
1.7 Variable Rare Disasters 43
1.7.1 Model Setup 44
1.7.2 Estimation 47
1.7.3 Implications of the Filtered State Estimates 51
1.7.4 Model Comparison 56
1.8 A Term Structure Model with a Zero Lower Bound 57
1.9 Conclusion 60

1.10 Acknowledgements 61

Chapter 2 Discretizing Nonlinear, Non-Gaussian Markov Processes with Exact Conditional Moments 62
2.1 Introduction 62
2.2 Maximum Entropy Method for Discretizing Markov Processes 66
2.2.1 Discretizing Probability Distributions 67
2.2.2 Discretizing General Markov Processes 71
2.3 Discretizing VAR(1)s and Stochastic Volatility Models 76
2.3.1 VAR(1) 76
2.3.2 AR(1) with Stochastic Volatility 81
2.3.3 Relation to the Existing Literature 82
2.4 Solution Accuracy of Asset Pricing Models 85
2.4.1 Model and Numerical Solution 85
2.4.2 Calibration 89
2.4.3 Solution Accuracy 90
2.5 Solution Accuracy of a Rare Disasters Model 97
2.5.1 Model 98
2.5.2 Solution Accuracy 99
2.6 Conclusion 105
2.7 Acknowledgements 106

Chapter 3 Pockets of Predictability 107
3.1 Introduction 107
3.2 Prediction Models and Estimation Methodology 113
3.2.1 Return Prediction Model with Constant Coefficients 114
3.2.2 Nonparametric Identification of Pockets 116
3.3 Empirical Results 121
3.3.1 Data 121
3.3.2 Anatomy of Pockets 126
3.3.3 Evaluating the Statistical Significance of the Results 128
3.3.4 Separating Spurious from Non-Spurious Pockets 132
3.4 Learning About Cash Flows 134
3.4.1 A Predictive Systems Model with Regime Switching 135
3.4.2 Filtering the State Variables and Evaluating the Likelihood 137
3.4.3 Asset Prices and Returns 140
3.4.4 Calibration of Model Parameters 141
3.4.5 Simulation Results 143
3.4.6 Learning Effects and Pockets 146
3.5 Economic Sources of Local Return Predictability 147
3.5.1 Pockets and Variation in the Business Cycle 148
3.5.2 Pockets and Variation in Sentiment 149
3.5.3 Out-of-Sample Return Predictability 149

3.6 Conclusion 152
3.7 Figures and Tables 153
3.8 Acknowledgements 167

Appendix 168
1.A Proofs for Chapter 1 168
1.B Discretizing Nonlinear, Non-Gaussian Markov Processes with Exact Conditional Moments 191
2.A Proofs for Chapter 2 194
2.B Accuracy of Discretization 200
2.B.1 VAR(1) 201
2.B.2 AR(1) with Stochastic Volatility 205
2.C Solving Asset Pricing Models 207
2.C.1 Analytical Solution with AR(1)/VAR(1) shocks 207
2.C.2 Discretizing the Rare Disasters Model 210
2.C.3 Solving the Rare Disasters Model 211
2.D Asset Pricing with Gaussian AR(1) Shocks 212
3.A Proof of Proposition 1 (Chapter 3) 218
3.B Details of Nonparametric Estimation 221

Bibliography 223

LIST OF FIGURES

Figure 1.1. Rule of Thumb Choice for M 28

Figure 1.2. MLE Sampling Distributions for Sample Size $T = 100$, Stochastic Volatility Model 38

Figure 1.3. MLE Sampling Distributions for Sample Size $T = 500$, Stochastic Volatility Model 39

Figure 1.4. MLE Sampling Distributions for Sample Size $T = 1,000$, Stochastic Volatility Model 39

Figure 1.5. Estimates of Disaster Probability, Rare Disasters Model 52

Figure 1.6. Estimates of Recovery Rate, Rare Disasters Model 54

Figure 1.7. Estimated Shadow Rates, Shadow Rate Term Structure Model . . . 58

Figure 2.1. $\log_{10}$ Relative Errors of Price-Dividend Ratio Approximations, VAR(1) Model 92

Figure 2.2. Densities Fitted to OLS Residuals, AR(1) Model 94

Figure 2.3. $\log_{10}$ Relative Errors of Price-Dividend Ratio Approximations, AR(1) Model with Gaussian Mixture Shocks 96

Figure 2.4. Ergodic Distribution of Resilience, Rare Disasters Model 100

Figure 2.5. $\log_{10}$ Relative Errors of Price-Dividend Ratio Approximations, Rare Disasters Model 101

Figure 3.1. Local Return Predictability from the Dividend Yield (Zero Coefficient Benchmark) 162

Figure 3.2. Local Return Predictability from the T-bill Rate (Zero Coefficient Benchmark) 163

Figure 3.3. Local Return Predictability from the Term Spread (Zero Coefficient Benchmark) 164

Figure 3.4. Local Return Predictability from the Corporate Spread (Zero Coefficient Benchmark) 165

Figure 3.5. Local Return Predictability from the Realized Variance (Zero Coefficient Benchmark) 166

Figure 2.D.1. $\log_{10}$ Relative Errors of Price-Dividend Ratio Approximations fo Different Numbers of Points, Gaussian AR(1) Model 214

Figure 2.D.2. $\log_{10}$ Relative Errors of Price-Dividend Ratio Approximations for Different Levels of Persistence, Gaussian AR(1) Model 215

Figure 2.D.3. $\log_{10}$ Relative Errors of Price-Dividend Ratio Approximations for a Highly Persistent Process, Gaussian AR(1) Model 216

Figure 2.D.4. $\log_{10}$ Relative Errors of Price-Dividend Ratio Approximations for a Highly Persistent Process, Alternate Parameterization, Gaussian AR(1) Model 219

LIST OF TABLES

Table 1.1. Likelihood Discrepancies, GDP Measurement Model 35

Table 1.2. Computation Time of 1 Likelihood Evaluation (in seconds), GDP Measurement Model 36

Table 1.3. Accuracy of Parameter Estimates, Stochastic Volatility Model 41

Table 1.4. Computation Time of 1 Likelihood Evaluation (in seconds), Stochastic Volatility Model 42

Table 1.5. Accuracy of Filtered State Estimates, Stochastic Volatility Model . 43

Table 1.6. Calibrated Parameters, Rare Disasters Model 47

Table 1.7. Estimated Parameters, Rare Disasters Model 49

Table 1.8. Parameter Values Implied by Estimation, Rare Disasters Model . . . 50

Table 1.9. Stock Market Moments 55

Table 1.10. Maximum Likelihood Parameter Estimates (QMLE Standard Errors in Parantheses), Shadow Rate Term Structure Model 59

Table 2.1. Mean and Maximum $\log_{10}$ Relative Errors of Price-Dividend Ratio Approximations, VAR(1) Model 91

Table 2.2. Parameter Values, AR(1) Model with Gaussian Mixture Shocks . . . 94

Table 2.3. Mean and Maximum $\log_{10}$ Relative Errors of Price-Dividend Ratio Approximations, AR(1) Model with Gaussian Mixture Shocks 97

Table 2.4. Mean and Maximum $\log_{10}$ Relative Errors of Price-Dividend Ratio Approximations, Rare Disasters Model 102

Table 2.5. Financial Moments, Rare Disasters Model 104

Table 3.1. Full Sample Regression Statistics 153

Table 3.2. Pocket Summary Statistics 154

Table 3.3. Statistical Significance Tests for Pocket Diagnostics (Zero Coefficient Benchmark) 155

Table 3.4. Statistical Significance Tests for Pocket Diagnostics (Constant Coefficient Benchmark) 156

Table 3.5. Integral $\bar{R}^2$ Measure and p-values for Individual Pockets 157

Table 3.6. Calibrated Parameters, Predictive Systems Model 158

Table 3.7. Average Integral $\bar{R}^2$, Predictive Systems Model 159

Table 3.8. Simulations from Predictive Systems Learning Model (Zero Coefficient Benchmark) 159

Table 3.9. Panel Regressions of Pocket Diagnostics on Belief Discrepancies, Predictive Systems Model 160

Table 3.10. Regressions of Pocket Diagnostics on Economic Indicators 161

Table 3.11. Out-of-Sample Measures of Forecasting Performance 161

Table 2.B.1. $\log_{10}$ Relative Bias, VAR(1) Model 203

Table 2.B.2. Computation Time for Discretizing the VAR(1) Process (in seconds) 205

Table 2.B.3. $\log_{10}$ Relative Bias, Stochastic Volatility Model 206

Table 2.D.1. Mispricing in Dollars when Investing \$1 Million, Gaussian AR(1) Model 217

ACKNOWLEDGEMENTS

I would like to thank Professors James D. Hamilton and Allan Timmermann for their continuing support, motivation, and guidance as the co-chairs of my committee. Their feedback and advice has proved invaluable.

Chapter 1, in full, is currently being prepared for submission for publication of the material. Farmer, Leland E. The dissertation author was the sole author of this paper.

Chapter 2, in full, is a reprint of the material that has been accepted for publication at Quantitative Economics. Farmer, Leland E.; Toda, Alexis Akira. The dissertation author was a primary author of this paper.

Chapter 3, in full, is currently being prepared for submission for publication of the material. Farmer, Leland E.; Schmidt, Lawrence D.W.; Timmermann, Allan. The dissertation author was a primary author of this paper.

VITA

2011	B.S. Mathematical and Computational Science, Stanford University
2013–2016	Teaching Assistant, Rady School of Management University of California, San Diego
2011–2017	Teaching Assistant, Department of Economics University of California, San Diego
2017	Ph.D. Economics, University of California, San Diego

PUBLICATIONS

"Discretizing Nonlinear, Non-Gaussian Markov Processes with Exact Conditional Moments." Quantitative Economics, forthcoming.

ABSTRACT OF THE DISSERTATION

Discrete Methods for the Estimation of Nonlinear Economic Models

by

Leland Edward Farmer

Doctor of Philosophy in Economics

University of California, San Diego, 2017

Professor James D. Hamilton, Co-Chair
Professor Allan Timmermann, Co-Chair

Economists increasingly use nonlinear methods to confront their theories with data. The switch from linear to nonlinear methods is driven, in part, by increased computing power, but also by a desire to understand economic phenomena that cannot easily be captured by linear models. My research is informed by questions at the intersection of macroeconomics and finance that cannot be addressed with standard methods.

Existing methods for estimating nonlinear dynamic models are either too computationally complex to be of practical use, or rely on local approximations which fail

to adequately capture the nonlinear features of interest. My research develops a new methodology for accurately estimating nonlinear dynamic models which is computationally simple and easy to apply. In my dissertation, I apply this methodology to study a model of interest rate dynamics near the zero lower bound, an asset pricing model of rare disasters, and a model of learning about cash flows in the presence of structural change.

Chapter 1

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

1.1 Introduction

Economists increasingly use nonlinear methods to confront their theories with data. The switch from linear to nonlinear methods is driven, in part, by increased computing power, but also by a desire to understand economic phenomena that cannot easily be captured by linear models. Examples include models which incorporate the zero lower bound on interest rates (ZLB), stochastic volatility, time-varying risk premia, Poisson jumps, credit constraints, borrowing constraints, non-convex adjustment costs, Markov-switching dynamics, and default.

Existing methods for estimating nonlinear dynamic models are either too computationally complex to be of practical use, or rely on local approximations which fail to adequately capture the nonlinear features of interest. In this paper, I develop a new method, the discretization filter, for approximating the likelihood of nonlinear, non-Gaussian state space models.

The major difficulty that arises when studying nonlinear state space models is that the likelihood cannot be evaluated recursively as it can in linear models with the Kalman

filter. The discretization filter solves this problem by constructing a discrete-valued Markov chain that approximates the dynamics of the state variables. The dynamics of the system are summarized by a transition matrix as opposed to an infinite dimensional transition kernel.

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

I apply results from the statistics literature on uniformly ergodic Markov chains to establish that the implied maximum likelihood estimator is strongly consistent, asymptotically normal, and asymptotically efficient. I demonstrate through simulations that the discretization filter is orders of magnitude faster than alternative nonlinear techniques for the same level of approximation error and I provide practical guidelines for applied researchers. It is my hope that the methods simplicity will make the quantitative study of nonlinear models easier for and more accessible to applied researchers.

I apply my approach to estimate two models at the intersection of macroeconomics and finance. The first is the Gabaix (2012) asset pricing model of variable rare disasters. The second is the Wu and Xia (2016) shadow rate term structure model. Both models are inherently nonlinear and neither can be consistently estimated with linear methods.

Gabaix (2012) develops a model of asset pricing which posits that the time-varying probability and severity of rare disasters explain why risk premia are large, volatile and time-varying. I provide the first quantitative estimates of the Gabaix model using data on equities and government bonds to identify the parameters and construct a measure of disaster risk for the U.S. economy. There have been several proposed

explanations for phenomena such as the equity premium puzzle, the excess volatility puzzle, and the riskfree rate puzzle. Most existing research on this topic calibrates a model and evaluates its ability to match a few select moments of the data. In contrast, the discretization filter allows researchers to formally estimate a series of models and evaluate their relative abilities to explain the data using model comparison statistics, thus facilitating model selection.

By using a likelihood-based method for estimation, I am able to construct estimates of the hidden states relating to real and nominal risk, which allow me to study additional implications of the model not captured by calibration or moment-matching procedures. In particular, I use these estimates to construct time series for the probability of a disaster, the conditional volatility of inflation, and the expected jump in inflation in the event of a disaster for the U.S. economy. I show that the model fails to identify the Great Recession as a disaster episode, assigning less than a 5% probability to a disaster having occurred between December of 2007 and June of 2009. This is because the model requires a positive jump in inflation in the event of disaster to match an upward sloping nominal yield curve. The model is unable to match the fact that the U.S. experienced low inflation and even deflation during the Great Recession in conjunction with an upward sloping nominal yield curve. This suggests that it is important to consider heterogeneity in the nature of disasters to capture the patterns of the U.S. data.

Wu and Xia (2016) develop a tractable approximation to a shadow rate term structure model. Their model provides a description of yield curve dynamics when the economy is near the zero lower bound on interest rates and provides a way of summarizing the effects of unconventional monetary policy. I show that when the model is estimated using the discretization filter, the estimates of the shadow rate are substantially lower over the zero lower bound period than those provided in their paper. This has important implications for policy makers who use this series as an input to their decision making

process. It implies, for example, that their estimates understate the effectiveness of unconventional monetary policy.

The paper is organized as follows. Section 1.2 reviews related literature. Section 1.3 explains the discretization filter. Section 1.4 establishes the strong consistency, asymptotic normality, and asymptotic efficiency of the approximate maximum likelihood estimator implied by the discretization filter. Section 1.5 provides practical implementation advice for applied researchers. Section 1.6 provides Monte Carlo comparisons with existing methods in the case of a linear measurement error model and a stochastic volatility model. In section 1.7, I estimate the Gabaix (2012) model of variable rare disasters and illustrate a couple of its shortcomings in explaining U.S. asset pricing data. Section 1.8 re-examines the Wu and Xia (2016) shadow rate term structure model and constructs an updated version of their shadow rate series. Section 1.9 concludes.

1.2 Related Literature

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

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

Toda (2016) develop a method for approximating general nonlinear, non-Gaussian first order Markov processes by matching conditional moments using maximum entropy.

A special case of the filtering algorithm proposed in this paper was first considered in Bucy (1969) and Bucy and Senne (1971), now referred to as the "point-mass filter." However, these papers and subsequent refinements only consider one specific method of discretizing the state process. Furthermore, none of these papers consider the asymptotic properties of estimators resulting from these filtering approximations. A comprehensive summary of filtering methods for state space models, including the point-mass filter, can be found in Chen (2003).

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

1.3 The Discretization Filter

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

distributions of the original model.

1.3.1 The Setting

In what follows I restrict attention to the analysis of Hidden Markov Models (HMMs). A HMM is a special type of nonlinear state space model where the observables in any given time period are a function only of the state variables in that time period. However, the results can be generalized to the case when the observation equation additionally depends on some finite number of lags of the observables. Much of the exposition and notation follows Douc, Moulines, and Ryden (2004).

Let X_t denote the vector of hidden state variables of the state space system at time t. I assume that $\{X_t\}_{t=0}^{\infty}$ is a time-homogeneous, first-order[1], stationary Markov chain and lies in a separable, compact set $\mathscr{X}$,[2] equipped with a metrizable topology and associated Borel σ-field $\mathscr{B}(\mathscr{X})$. Let $P_\theta(x,A)$, where $x \in \mathscr{X}$ and $A \in \mathscr{B}(\mathscr{X})$, be the transition kernel of the Markov chain. I further assume that for all $\theta \in \Theta$ and $x \in \mathscr{X}$, each conditonal probability measure $P_\theta(x,\cdot)$ has a density $q_\theta(\cdot|x)$ with respect to a common finite dominating measure μ on $\mathscr{X}$.[3]

I assume that the observable sequence $\{Y_t\}_{t=1}^{\infty}$ takes values in a set $\mathscr{Y}$ that is separable and metrizable by a complete metric. I assume that for $t \geq 1$, Y_t is conditionally independent of $\{Y_s\}_{s=1}^{t-1}$ and $\{X_s\}_{s=1}^{t-1}$ given X_t. Note that this excludes models where the observation at time t depends on its own lagged values. This is purely for expositional simplicity and all of the results can be generalized to the case where Y_t depends on some fixed, finite number of lags of itself, $\{Y_{t-1},\ldots,Y_{t-k}\}$, although this does complicate the

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

[2] Compactness of $\mathscr{X}$ simplifies much of the notation and proofs, however many of the results can be generalized to the noncompact case using techniques developped in Douc, Moulines, Olsson, and Van Handel (2011)

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

construction of the transition matrices. I also assume that the observations conditional on any value of the state $X_t = x$, $x \in \mathscr{X}$, have a density $g_\theta(\cdot|x)$ with respect to a σ-finite measure ν on the Borel σ-field $\mathscr{B}(\mathscr{Y})$.

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

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

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

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

A HMM is characterized by the following two equations:

$$X_t | X_{t-1} \sim q_\theta(X_t | X_{t-1}) \tag{1.1}$$

$$Y_t | X_t \sim g_\theta(Y_t | X_t) \tag{1.2}$$

Equation (1.1) is the state equation, and it characterizes the distribution of the latent state next period conditional on the current state. Equation (1.2) is the observation, or measurement equation, and it characterizes the distribution of the observables conditional on the current state.

Let x_t and y_t denote particular realizations of the random variables X_t and Y_t. Given a sample $\{y_t\}_{t=1}^T$, the goal is to obtain estimates of the parameter vector θ and the unobserved states $\{x_t\}_{t=1}^T$, which I will denote by $\hat{\theta}_T$ and $\{\hat{x}_{t|t}\}_{t=1}^T$ respectively.[4] In

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

order to do this, one must obtain an expression for the likelihood of the data:

$$L_T(\theta, x_0) \equiv p_\theta\left(\boldsymbol{Y}_1^T | X_0 = x_0\right) \tag{1.3}$$

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

$$\ell_T(\theta, x_0) \equiv \log p_\theta\left(\boldsymbol{Y}_1^T | X_0 = x_0\right) \tag{1.4}$$

In the subsequent section, I show how to approximate equation (1.1) by a discrete-valued Markov chain.

1.3.2 Approximating the State Dynamics

The idea of discretization to alleviate computational problems in economics is not new. One of the first instances of this is Tauchen (1986). He proposes a simple way of approximating any Gaussian VAR(1) with a first-order, discrete-valued Markov chain. He then shows that this approximation does a good job of matching unconditional and conditional moments for relatively coarse discretizations. Tauchen's approximation, along with several more recent approximations proposed in the literature,[5] have been widely used to solve asset pricing and DSGE models where the ability to approximate the solutions to integral equations is of key importance.

In this paper I apply this idea of discretization to the estimation of nonlinear, non-Gaussian state space models. More specifically, I construct a discrete-valued, first-order

[5] See e.g. Tauchen and Hussey (1991), Rouwenhorst (1995), Adda and Cooper (2003), Flodén (2008), Tanaka and Toda (2013), Gospodinov and Lkhagvasuren (2014), and Farmer and Toda (2016).

Markov process $\{X_{t,M}\}_{t=1}^{\infty}$, whose dynamics mimic those of the original continuous-valued process $\{X_t\}_{t=1}^{\infty}$. This allows me to summarize the dynamics of the unobserved state by a finite-dimensional transition matrix $P_{\theta,M}$.[6] Note that this is fundamentally different from forecasting the next period's state by taking a local approximation around the current estimate as is done in the extended Kalman filter. My approximation method is global yet does not rely on simulation techniques.

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

$$P_{\theta,M}\left(m,m'\right) = \mathbb{P}_\theta\left(X_{t,M} = x_{m',M} \left| X_{t-1,M} = x_{m,M}\right.\right) \tag{1.5}$$

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

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

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

$$\zeta_{t,M} = P'_{\theta,M}\zeta_{t-1,M} + \tilde{v}_{t,M} \tag{1.6}$$

$$Y_t \left| X_{t,M} \right. \sim g_\theta\left(Y_t \left| X_{t,M}\right.\right) \tag{1.7}$$

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

where $\tilde{v}_{t,M} = \zeta_{t,M} - \mathbb{E}_\theta \left[\zeta_{t,M} \middle| \zeta_{t-1,M} \right]$ and $P'_{\theta,M}$ is the transpose of the matrix $P_{\theta,M}$. Equations (1.6) and (1.7) are the state and observation equations of the new approximate model. The sequence $\{Y_t\}$ has the same distribution, conditional on the state $X_{t,M}$, as the sequence $\{Y_t\}$ generated by the original model. However, in the approximate model, the $X_{t,M}$ have been restricted to live on a discrete grid.

1.3.3 Evaluating the Likelihood

In the previous section, I showed how to approximate any HMM by replacing the state equation, equation (1.1), with a discrete-state Markov chain, equation (1.6). In this section, I apply the results of Hamilton (1989) to construct an approximation to the likelihood function of the HMM. Hamilton (1989) shows that when the state dynamics of a HMM are characterized by a discrete-state Markov chain, simple prediction and updating equations exist that are analogous to the Kalman filter in the linear case. I use the notation developed in Hamilton (1994). I review these results here and show how they can be used to develop an approximation to the maximum likelihood estimator for θ.

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

$$\hat{\xi}_{t,M|t-1} = \mathbb{E}_\theta \left[\zeta_{t,M} \middle| \boldsymbol{Y}_1^{t-1} \right] = P'_{\theta,M} \hat{\xi}_{t-1,M|t-1} \tag{1.8}$$

Also define

$$\eta_{t,M} = \begin{bmatrix} g_\theta\left(Y_t \left| X_t = x_{1,M}\right.\right) \\ \vdots \\ g_\theta\left(Y_t \left| X_t = x_{m,M}\right.\right) \end{bmatrix} \tag{1.9}$$

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

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

$$p_{\theta,M}\left(Y_t \left| \boldsymbol{Y}_1^{t-1}\right.\right) = 1'\left(\eta_{t,M} \odot \hat{\xi}_{t,M|t-1}\right) \tag{1.10}$$

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

$$\hat{\xi}_{t,M|t} = \frac{\eta_{t,M} \odot \hat{\xi}_{t,M|t-1}}{1'\left(\eta_{t,M} \odot \hat{\xi}_{t,M|t-1}\right)} = \frac{\eta_{t,M} \odot \hat{\xi}_{t,M|t-1}}{p_{\theta,M}\left(Y_t \left| \boldsymbol{Y}_1^{t-1}\right.\right)} \tag{1.11}$$

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

$$\ell_{T,M}(\theta) = \sum_{t=1}^{T} \log p_{\theta,M}\left(Y_t \left| \boldsymbol{Y}_1^{t-1}\right.\right) \tag{1.12}$$

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

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

Note that the parameter estimates $\hat{\theta}_{T,M}$ and the log-likelihood function $\ell_{T,M}(\theta)$

Algorithm 1: Discretization Filter

1. **Approximate the State Dynamics:** Construct a discrete grid $\{x_{m,M}\}_{m=1}^{M}$ and its associated transition matrix $P_{\theta,M}$ using algorithm 2 in appendix 1.B or any other method appropriate for the process X_t being considered.
2. **Initialization:** Set the initial distribution of the state $\hat{\zeta}_{0,m|0} = \pi_{\theta,M}^{X}$ or any arbitrary distribution. Set $t \rightsquigarrow 1$.
3. **Prediction:** Construct the forecast of the time t state $\hat{\zeta}_{t,M|t-1} = P'_{\theta,M}\hat{\zeta}_{t,M|t-1}$.
4. **Updating 1:** Evaluate the contemporaneous likelihood of having observed data y_t conditional on each possible value of the state, $\eta_{t,M}$, using equation (1.9). Compute and save the marginal likelihood of observation y_t given by equation (1.10).
5. **Updating 2:** Compute the time t filtered estimate of the state $\hat{\zeta}_{t,M|t}$ using (1.11). If $t < T$, set $t \rightsquigarrow t+1$ and go to step 3. Otherwise go to step 6.
6. **Likelihood:** Compute the approximate likelihood of the data, $\ell_{T,M}(\theta)$, using equation (1.12).

are indexed by the number of discrete points M in addition to the sample size T to indicate that the estimates will depend on exactly how the space is discretized. I have omitted the explicit dependence of the likelihood function on the distribution of the initial state $x_{0,M}$. As part of the results in section 1.4, I will show why this initial condition is irrelevant for the asymptotic properties of $\hat{\theta}_{T,M}$.

Section 1.4 establishes the strong consistency, asymptotic normality, and asymptotic efficiency of the discretization filter approximation to the maximum likelihood estimator. Those who are interested in applications of the discretization filter may wish to skip ahead to section 1.5.

1.4 Asymptotic Properties of the Maximum Likelihood Estimator

In this section I establish strong consistency, asymptotic normality, and asymptotic efficiency of my proposed estimator. I consider joint asymptotics in both the sample size

T and the number of discrete points M. I show that the accuracy of my approximation is governed to first order by the proximity of the infinite history filtering distributions of the approximate and true chains $X_{t,M}\left|\boldsymbol{Y}^t_{-\infty}\right.$ and $X_t\left|\boldsymbol{Y}^t_{-\infty}\right.$. The distance between these distributions is proportional to $h^*(M)$, where $h^*(M)$ is related to the approximation error between the approximate and true one-step-ahead conditional distributions of X_t. Strong consistency simply requires that $T\to\infty$ and $M\to\infty$. Asymptotic normality and asymptotic efficiency further require that $T\times h^*(M)\to 0$ as $M\to\infty$ and $T\to\infty$, i.e. that $M\to\infty$ "fast enough."

A key new theoretical contribution of my paper is to establish a rate of convergence of the ergodic distribution of the approximate discrete chain to the true ergodic distribution. This result represents a new contribution to the literature on discrete approximations of Markov chains with continuous valued states. All proofs can be found in Appendix 1.A.

1.4.1 Preliminaries and Assumptions

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

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

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

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

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

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

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

and

$$\pi^{X}_{\theta,\infty}(A) \equiv \lim_{M\to\infty} \sum_{m=1}^{M} \pi^{X}_{\theta,M}(m) \mathbb{1}\left\{x_{m,M} \in A\right\}$$

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

I now list and discuss my basic assumptions. Assumptions that overlap with Douc, Moulines, and Ryden (2004) are labeled with an A, and assumptions that are new to this paper are labeled with a B. Assumptions labeled A and B are paired by number, e.g. (A1) and (B1). Each B assumption can be thought of as an analog to the A assumption for the sequence of discrete approximations $X_{t,M}$.

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

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

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

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

Assumption (B1) guarantees that the discrete process $\{X_{t,M}\}$ has a unique invariant distribution $\pi^{X}_{\theta,M}$ and is uniformly ergodic for every value $M < \infty$. Additionally it is

needed so that the bound on the mixing rate of $X_{t,M}$ is independent of M and θ. This will be satisfied for any stochastic process satisfying (A1)(a) that is approximated using the methods reviewed in section 1.3.2. Note that while all elements of the transition matrix $P_{\theta,M}$ converge to 0 individually as $M \to \infty$, the limits of the ratios of these elements are still well defined.

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

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

These assumptions guarantee that the original joint Markov process $\{Z_t\}$ and the limiting approximating Markov chain $\{Z_{t,\infty}\}$ are themselves uniformly ergodic. Note that assumption (B2) is needed in addition to assumption (B1) to account for the limiting case of the chain.

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

$$\lim_{t\to\infty} \left\| \lambda \Pi_\theta^{(t)} - \pi_\theta \right\|_{TV} = 0 \tag{1.13}$$

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

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

and $\Pi_\theta^{(t)}$ is the t-th iterate of the transition kernel Π_θ. In words, for any initial measure of the joint process $\{Z_t\}$, the probability of being in any measurable set $A \in \mathscr{B}(\mathscr{Z})$ approaches the ergodic probability of being in that set uniformly over all measurable sets A as $t \to \infty$. This convergence is also independent of the initial measure λ. An analogous

property holds for the process $\{Z_{t,\infty}\}$ by assumption (B2). Developing a bound on this rate of convergence will be critical for the coming developments.

Lastly, assume that

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

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

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

1.4.2 Consistency

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

The second part, which is new to this paper, is to show that this approximate limiting criterion function $\ell_M(\theta)$, which is defined for any M, converges to the true

limiting criterion function $\ell(\theta)$ as the number of points used in the approximation $M \to \infty$. I will show that this holds for any discretization method whose one-step-ahead conditional distributions $X_{t,M} \left| X_{t-1,M} = x \right.$ converge in distribution to the one-step-ahead conditional distributions of the original continuous process $X_t \left| X_{t-1} = x \right.$ as $M \to \infty$.

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

Following Douc, Moulines, and Ryden (2004), I first establish that the distribution of $X_{t,M}$ given a history of observations $\boldsymbol{Y}_r^s$ is itself a uniformly ergodic (inhomogeneous) Markov chain with minorizing constant independent of the parameter $\theta \in \Theta$ and the number of discrete points $M \in \mathbb{Z}^+$. This is the analogous result to Lemma 1 in their paper. Note that a Markov chain with transition kernel P_θ is said to satisfy a uniform minorization condition if there exist a probability measure μ_Q, a positive integer n, and $\varepsilon > 0$ such that

$$P_\theta^{(n)}(x, A) \geq \varepsilon \mu_Q(A)$$

for all $x \in \mathscr{X}$ and $A \in \mathscr{B}(\mathscr{X})$, where $P_\theta^{(n)}$ is the n-step ahead transition kernel of the Markov chain.

Define $Q_M^- \equiv \inf_{m,m'} P_{\theta,M}(m, m')$, $Q_M^+ \equiv \sup_{m,m'} P_{\theta,M}(m, m')$, and $Q_+^- \equiv \frac{Q_M^-}{Q_M^+}$ for $M \in \mathbb{Z}^+$. I now state the first lemma

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

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

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

all $\boldsymbol{Y}_r^s$,

$$\inf_{x \in \mathscr{X}_M} \bar{\mathbb{P}}_\theta \left(X_{t,M} \in A \middle| X_{t-1,M} = x, \boldsymbol{Y}_r^s \right) \geq Q_+^- \mu_{t,M} \left(\boldsymbol{Y}_t^s, A \right)$$

The major difference between this Lemma and the one established in Douc, Moulines, and Ryden (2004) is that for the following results, it will be crucial that the minorizing constant be the same for all M, in order to establish uniform convergence over $M \in \mathbb{Z}^+$ of the approximate likelihood function. Note that although the minorizing measure, $\mu_{t,M}(\boldsymbol{Y}_t^s, \cdot)$, *does* depend on both the number of points, M, and the observations the chain is conditioned on, $\boldsymbol{Y}_t^s$, it doesn't affect the mixing rate. The previous lemma leads to the following corollary, using standard results for uniformly minorized Markov chains (see e.g. Lindvall (1992) Sections III.9-11).

. Assume (A1) and (B1). Let $r, s \in \mathbb{Z}$ with $r \leq s$, $\theta \in \Theta$, and $M \in \mathbb{Z}^+$. Then for all $t \geq r$, all probability measures μ_1 and μ_2 on $\mathscr{B}(\mathscr{X}_M)$, and all $\boldsymbol{Y}_r^s$,

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

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

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

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

. Assume (A1)-(A2) and (B1)-(B2). Then, for all $x_{0,M} \in \mathscr{X}_M$ and $M \in \mathbb{Z}^+$,

$$\sup_{\theta \in \Theta} \left| \ell_{T,M}(\theta, x_{0,M}) - \ell_{T,M}(\theta) \right| \leq 1/(1-\rho)^2, \quad \bar{\mathbb{P}}_{\theta^*}\text{-a.s.}$$

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

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

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

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

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

. Assume (A1)-(A3) and (B1)-(B3). Then for all $t \geq 1$, $r, r' \geq 0$, and $M \in \mathbb{Z}^+$, $\bar{\mathbb{P}}_{\theta^*}$-a.s.,

$$\sup_{\theta \in \Theta} \sup_{x, x' \in \mathscr{X}_M} \left| \Delta_{t,r,M,x}(\theta) - \Delta_{t,r',M,x'}(\theta) \right| \leq \rho^{t+\min(r,r')-1}/(1-\rho), \tag{1.14}$$

$$\sup_{\theta \in \Theta} \sup_{x \in \mathscr{X}_M} \left| \Delta_{t,r,M,x}(\theta) - \Delta_{t,r,M}(\theta) \right| \leq \rho^{t+r-1}/(1-\rho), \tag{1.15}$$

$$\sup_{\theta \in \Theta} \sup_{r \geq 0} \sup_{x \in \mathscr{X}_M} \left| \Delta_{t,r,M,x}(\theta) \right| \leq \max\left(\left|\log b_+\right|, \left|\log c_-(Y_t)\right|\right) \tag{1.16}$$

Equation (1.14) of Lemma 3 shows that $\{\Delta_{t,r,M,x}\}_{r\geq 0}$ is a uniform Cauchy sequence w.r.t. $\theta \in \Theta$ and thus converges $\bar{\mathbb{P}}_{\theta^*}$-a.s. to a limit which does not depend on the initial value x. I label this limit $\Delta_{t,\infty,M}$ and intuitively this can be thought of as $\log \bar{p}_{\theta,M}\left(Y_t \left| \boldsymbol{Y}_{-\infty}^{t-1}\right.\right)$, the marginal likelihood of an observation Y_t given an infinite history of data.

Equation (1.16) of Lemma 3 shows that $\{\Delta_{t,r,M,x}(\theta)\}_{r\geq 0}$ is uniformly bounded in $L^1\left(\bar{\mathbb{P}}_{\theta^*}\right)$ and thus its limit $\Delta_{t,\infty,M}(\theta)$ is also in $L^1\left(\bar{\mathbb{P}}_{\theta^*}\right)$. Furthermore, note that $\{\Delta_{t,\infty,M}(\theta)\}$ is a $\bar{\mathbb{P}}_{\theta^*}$-stationary ergodic process.

By setting $r = 0$ and letting $r' \to \infty$ in equation (1.14), it follows that

$$\sup_{\theta\in\Theta}\left|\Delta_{t,0,M,x}(\theta) - \Delta_{t,\infty,M}(\theta)\right| \leq \rho^{t-1}/(1-\rho)$$

Furthermore, setting $r = 0$ in equation (1.15) implies that

$$\sup_{\theta\in\Theta}\left|\Delta_{t,0,M,x}(\theta) - \Delta_{t,0,M}(\theta)\right| \leq \rho^{t-1}/(1-\rho)$$

By combining these two inequalities, applying the triangle inequality, and summing from 1 to T, I obtain Corollary 2.

. Assume (A1)-(A2) and (B1)-(B2). Then

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

Corollary 2 shows that $T^{-1}\ell_{T,M}(\theta)$ can be approximated by the sample mean of a stationary ergodic sequence, uniformly w.r.t. θ. Since $\Delta_{0,\infty,M} \in L^1\left(\bar{\mathbb{P}}_{\theta^*}\right)$, the ergodic theorem implies that $T^{-1}\ell_{T,M}(\theta) \to \ell_M(\theta)$ $\bar{\mathbb{P}}_{\theta^*}$-a.s. and in $L^1\left(\bar{\mathbb{P}}_{\theta^*}\right)$ as $T \to \infty$. Note that this convergence is uniform over $M \in \mathbb{Z}^+$. This will be important when I start considering joint asymptotics in T and M.

Define $\ell(\theta) \equiv \bar{\mathbb{E}}_{\theta^*}\left[\log \bar{p}_\theta\left(Y_0 \middle| \boldsymbol{Y}_{-\infty}^0\right)\right]$. The next step towards establishing consistency is to show that $\ell_M(\theta) \to \ell(\theta)$ as $M \to \infty$. The difference in these two quantities is related to the difference in the approximate and true filtering distributions for infinite histories of observations, $X_{t,M}\left|\boldsymbol{Y}_{-\infty}^t\right.$ and $X_t\left|\boldsymbol{Y}_{-\infty}^t\right.$.

I first prove that the ergodic distribution of the approximate discrete Markov chain converges weakly to that of the original continuous Markov chain, i.e. that $X_{t,M} \xrightarrow{d} X_t$ as $M \to \infty$. Proposition 1 establishes this convergence and provides a bound on the difference between the two distributions as a function of the number of points M.

Define $\mathscr{A}$ as the collection of all continuity sets of X_t. I make one further assumption regarding the approximation quality of the sequence of transition kernels $\left\{P_{\theta,M}\right\}$.

(BT) For all $A \in \mathscr{A}$, the sequence of approximations $P_{\theta,M}$ satisfy

$$\sup_{\theta\in\Theta}\sup_{x\in\mathscr{X}}\left|P_{\theta,M}(x,A) - P_\theta(x,A)\right| = O(h(M)) \tag{1.17}$$

where $h(M)$ satisfies $\lim_{M\to\infty} h(M) = 0$.

This assumption allows the practitioner to use *all* of the discretization methods outlined in 1.3.2 to construct $P_{\theta,M}$. I have chosen to illustrate the case where the Farmer and Toda (2016) method with trapezoidal quadrature rule is used. In this case, assumption (BT) is satisfied with $h(M) = M^{-2/d}$, where d is the dimension of the state space $\mathscr{X}$.[7]

. Assume (A1)-(A3), (B1)-(B3), and (BT). Then it follows that for any $A \in \mathscr{A}$,

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

where $h^*(M)$ satisfies $\lim_{M\to\infty} h^*(M) = 0$. If the transition kernel is approximated as

[7]For a discussion of error convergence properties see Tanaka and Toda (2015).

proposed in Farmer and Toda (2016) with a trapezoidal quadrature rule,

$$h^*(M) = M^{-(2-\delta)/d}$$

for any $\delta > 0$.

Note that even faster rates can be achieved through clever choice of the quadrature formula and the assumptions one is willing to make about the smoothness of the likelihood function.[8] By combining Proposition 1 with uniform ergodicity of $X_{t,M}$ and X_t, it can be shown that this approximation error directly translates to probabilities computed under the filtering distributions $X_{t,M}|\boldsymbol{Y}_r^t$ and $X_t|\boldsymbol{Y}_r^t$.

. Assume (A1)-(A3), (B1)-(B3), and (BT). Then

$$\sup_{\theta\in\Theta} |\ell_M(\theta) - \ell(\theta)| = o(h^*(M))$$

Combining Corollary 2, Lemma 2, and Lemma 4 leads to the following pointwise convergence result

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

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

The final step before I can state the strong consistency result involves showing

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

that $\ell_M(\theta)$ is continuous w.r.t. θ for all $M \in \mathbb{Z}^+$. This will allow me to strengthen Corollary 3 from pointwise convergence to uniform convergence in θ. Note that by (1.16) and the dominated convergence theorem,

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

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

The following additional assumptions are needed to establish continuity

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

(B4) For all $M \in \mathbb{Z}^+$, $x \in \mathscr{X}_M$, and $A \in \mathscr{B}(\mathscr{X}_M)$, $\theta \mapsto P_{\theta,M}(x,A)$ is continuous.

. Assume (A1)-(A4), (B1)-(B4), and (BT), then

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

A direct consequence of Lemma 5 is that the convergence established in Corollary 3 can be strengthened to uniform convergence in $\theta \in \Theta$.

. Assume (A1)-(A4), (B1)-(B4), and (BT). Then

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

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

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

$$\bar{\mathbb{E}}_{\theta^*}\left[\log \frac{\bar{p}_{\theta^*}\left(\boldsymbol{Y}_1^t\right)}{\bar{p}_\theta\left(\boldsymbol{Y}_1^t\right)}\right] = 0 \qquad \text{for all } t \geq 1. \tag{1.18}$$

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

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

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

1.4.3 Asymptotic Normality

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Note that this result is actually stronger than just asymptotic normality. Theorem 3 establishes that my proposed estimator and the infeasible maximum likelihood estimator are asymptotically equivalent. That is, my estimator asymptotically achieves the Cramér-Rao lower bound.

1.5 Recommendations for Applied Researchers

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

1.5.1 Choosing the Number of Grid Points

The asymptotic theory I developed in section 1.4 shows that if the Farmer and Toda (2016) method with a trapezoidal quadrature rule is used to construct the transition matrix, the discretization error of the likelihood function is of the order $TM^{-2/d}$. While this is only a rate condition, I use it to recommend a rule of thumb choice for the number of points M used to construct the discretization. Setting this ratio equal to a constant and solving for M, one gets the rule of thumb

$$M = cT^{d/2} \tag{1.19}$$

where the constant c is a nuisance parameter. For example, if the dimension d of the state space is 1, the rule says to choose a number of points proportional to the cube root of the sample size. If $d = 2$, then the rule recommends choosing the number of points equal to the sample size. I investigate the effect of choosing different values of c on the accuracy of the approximation in section 1.6.

Figure 1.1 plots the rule-of-thumb choice for M for state spaces of dimensions 1-4, for sample sizes up to $T = 100$ and $c = 1$.

The asymptotic analysis implies that M should be chosen to be as large as possible. However, for sufficiently large computational problems, it may not be possible to choose a large number for M. An applied researcher faces a tradeoff between computation time and the accuracy of the approximation, which I will elaborate on in section 1.6. This rule of thumb can be thought of as a lower bound on the number of points to choose in order to retain validity of confidence intervals constructed for parameters using a normal approximation.

1.5.2 Selecting the Grid Points

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

Consider the case where the number of discretization points, M, has been fixed

[9]The assumption of uniform ergodicity can be relaxed to geometric ergodicity, where the mixing rate of the Markov chain depends on the initial distribution. Under suitable restrictions on the initial distribution, consistency can still be established using the techniques in Douc, Moulines, Olsson, and Van Handel (2011). Asymptotic normality of the maximum likelihood estimator under geometric ergodicity appears to still be an open problem.

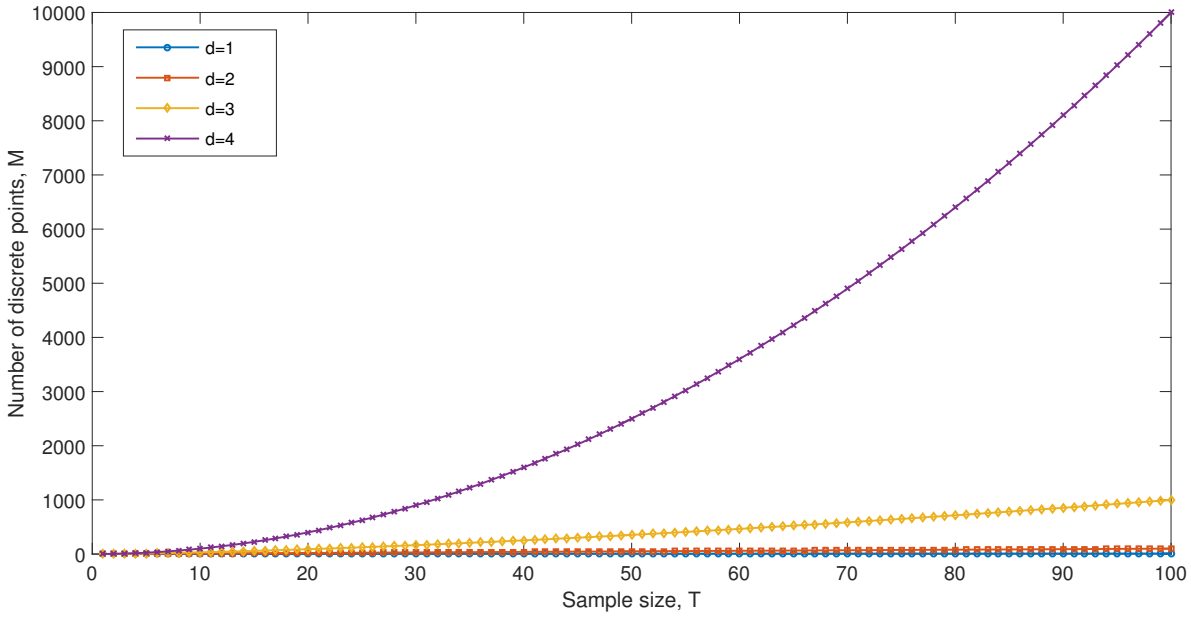


Figure 1.1. Rule of Thumb Choice for M

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

When the state follows a Gaussian VAR(1), a closed form expression for the ergodic distribution is available. Gospodinov and Lkhagvasuren (2014) provide a method to discretize Gaussian VAR(1)s that is robust to high levels of persistence. They use mixtures of Rouwenhorst (1995) approximations to match conditional moments as closely as possible. I rely on this method in section 1.8 for my empirical application. However, for more general time series models, no such expression exists.

Even when no expression for the unconditional distribution exists, it is often possible to compute the unconditional mean and standard deviation of the process. In this case, I recommend choosing a grid centered at the unconditional mean μ_x covering

$\sqrt{M-1}$ unconditional standard deviations σ_x of the process on either side. That is, choose $\{x_{m,M}\}_{m=1}^{M}$ to be M evenly spaced points over the interval $\left[\mu_x - \sqrt{M-1}\sigma_x, \mu_x + \sqrt{M-1}\sigma_x\right]$. [10]

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

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

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

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

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

Selecting the points in this way has the desirable property that roughly the same number of realizations of the state will fall between each pair of points.[11] By choosing α arbitrarily close to 1, it is possible to ensure that one has covered the ergodic set with any

[10] This is the way of constructing the grid employed in the Rouwenhorst (1995) approximation and suggested in Farmer and Toda (2016).

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

desired degree of confidence.[12] This method is also robust to skewness and fat tails in the stationary distribution.

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

1.5.3 Constructing the Transition Matrix

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

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

$$P_{\theta,M}\left(m,m'\right) = \int_{A_{m'}} q_\theta\left(x\,|X_{t-1} = x_{m,M}\right)\mu\left(dx\right) \tag{1.20}$$

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

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

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

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

to the conditional density $q_\theta\left(X_t = x_{m',M} \,|X_{t-1} = x_{m,M}\right)$ over the set $A_{m'}$.

A limitation of this approach is the ability to evaluate the integrals needed to construct the transition matrix. In general, this method will only work well in practice when the A_m are hyperrectangles, and the transition density is easy to evaluate. Furthermore, there are no known results on the rate of weak convergence of the ergodic distribution of the approximate Markov chain to the that of the underlying continuous process. Since this rate is critical to obtaining asymptotic normality, researchers should be cautious about standard errors when using this approach with a small number of points.

Second, I construct the transition matrix as in Farmer and Toda (2016). They provide a general way of constructing finite-state Markov chain approximations to stochastic processes. Their method finds the discrete distribution which is "closest" to the original distribution from some prior distribution in terms of Kullback-Leibler distance, while matching a set of conditional moments of the underlying continuous distribution.

If the prior distribution is a valid quadrature formula for evaluating integrals with respsect to the original conditional density, the discrete approximation is guaranteed to converge weakly to the continuous distribution. Moreover, the rate of convergence is given by the rate of convergence of the selected quadrature formula.[15]

My Monte Carlo results in section 1.6 demonstrate that when the primary aim is estimation of the parameters, very coarse discretizations are adequate. This is in line with my theoretical results which show that the estimates are consistent independently of the rate at which M grows. The discretization filter has the potential to scale to higher dimensional problems by exploiting sparse grid quadrature methods (e.g. Smolyak grids),

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

quasi-Monte Carlo methods, or the more recently proposed ε-distinguishable set method in Maliar and Maliar (2015). I leave the investigation of this extension for future research.

1.6 Monte Carlo Evidence

In this section, I consider two simulation exercises, a linear measurement error model and a stochastic volatility model, to compare the performance of the discretization filter with exisiting alternatives.

1.6.1 Measuring GDP: A Linear State Space Example

I first consider a simple linear Gaussian state space model to illustrate the performance of the discretization filter in a case where the exact evaluation of the likelihood is possible using the Kalman Filter.

Aruoba, Diebold, Nalewaik, Schorfheide, and Song (2016) propose extracting a common component of the two widely available measures of GDP using a simple measurement error model in order to provide a more accurate estimate of "true" GDP. Let $\Delta\text{GDP}_{E,t}$ and $\Delta\text{GDP}_{I,t}$ denote the expenditure and income-side estimates of GDP growth respectively, and let ΔGDP_t denote true GDP growth, which is assumed to be unobserved. Consider the following state space model

$$\begin{bmatrix} \Delta\text{GDP}_{E,t} \\ \Delta\text{GDP}_{I,t} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Delta\text{GDP}_t + \begin{bmatrix} \varepsilon_{E,t} \\ \varepsilon_{I,t} \end{bmatrix}$$

$$\Delta\text{GDP}_t = \mu(1-\rho) + \rho\Delta\text{GDP}_{t-1} + \varepsilon_{G,t}$$

where $\left(\varepsilon_{G,t}, \varepsilon_{E,t}, \varepsilon_{I,t}\right)' \sim i.i.d.N(\mathbf{0}, \Sigma)$, with

$$\Sigma = \begin{bmatrix} \sigma^2_{G,G} & 0 & 0 \\ 0 & \sigma^2_{E,E} & \sigma_{E,I} \\ 0 & \sigma_{E,I} & \sigma^2_{I,I} \end{bmatrix}$$

In their paper, Aruoba, Diebold, Nalewaik, Schorfheide, and Song (2016) also consider a more sophisticated specification of the model which allows for correlation between the measurement and state equation errors. The discretization filter can allow for this at the cost of introducing time-varying transition matrices but I omit the details for expositional simplicity. I focus on the restricted model outlined above.

I take the parameters estimated in the paper and simulate 500 samples of length $T = 204$, which is the amount of data used for estimation. For each sample, I evaluate the likelihood of the data using the Kalman filter (KF), the discretization filter (DF), and the bootstrap particle filter (PF). I examine the following two statistics for assessing the quality of the likelihood approximation discussed in Herbst and Schorfheide (2015)

$$\hat{\Delta}_1 = \ln \hat{p}_\theta \left(\boldsymbol{Y}_1^T \right) - \ln p_\theta \left(\boldsymbol{Y}_1^T \right) \tag{1.21}$$

$$\hat{\Delta}_2 = \exp \left[\ln \hat{p}_\theta \left(\boldsymbol{Y}_1^T \right) - \ln p_\theta \left(\boldsymbol{Y}_1^T \right) \right] - 1 \tag{1.22}$$

where $\hat{p}_\theta \left(\boldsymbol{Y}_1^T \right)$ denotes the approximate likelihood computed with either the DF or the PF, and $p_\theta \left(\boldsymbol{Y}_1^T \right)$ denotes the true likelihood evaluated with the KF. Since the approximation to the likelihood provided by the PF is random, I use a 100 draws of the PF for every realization of the data. I consider several choices for the number of particles N used in the PF and for the proportionality constant used in the rule-of-thumb choice for the number of grid points M in the DF proposed in (1.19).

Table 1.1 presents the results of the simulation exercise for the accuracy of the likelihood approximations as measured by $\hat{\Delta}_1$ and $\hat{\Delta}_2$. An important distinction between

the PF and the DF is that the PF approximation to the likelihood is random. It depends on the particular path that is simulated for the particles. However, the DF approximation to the likelihood is deterministic and thus has no associated sampling uncertainty for a given draw of the data.

For the PF, the bias and standard deviation of the approximations for a particular realization of the data are computed as the average value and standard deviation of the likelihood discrepancies across the 100 draws of the particles respectively. Since the DF is deterministic, there is only one value of the bias per sample realization and the standard deviation is zero. The RMSE is given by the familiar Bias^2 + Var formula. The means of these statistics are then computed as the means across randomly generated samples.

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

$$\text{Mean Bias}\left(\hat{\Delta}_i^{PF}\right) = \frac{1}{S}\sum_{s=1}^{S}\left[\frac{1}{G}\sum_{g=1}^{G}\hat{\Delta}_{i,s,g}^{PF}\right] \tag{1.23}$$

$$\text{Mean Var}\left(\hat{\Delta}_i^{PF}\right) = \frac{1}{S}\sum_{s=1}^{S}\left[\hat{\Delta}_{i,s,g}^{PF} - \frac{1}{G}\sum_{g=1}^{G}\hat{\Delta}_{i,s,g}^{PF}\right]^2 \tag{1.24}$$

$$\text{Mean RMSE}\left(\hat{\Delta}_i^{PF}\right) = \frac{1}{S}\sum_{s=1}^{S}\left[\frac{1}{G}\sum_{g=1}^{G}\left(\hat{\Delta}_{i,s,g}^{PF}\right)^2\right]^{1/2} \tag{1.25}$$

For the DF, the mean bias and RMSE are given by

$$\text{Mean Bias}\left(\hat{\Delta}_i^{DF}\right) = \frac{1}{S}\sum_{s=1}^{S}\hat{\Delta}_{i,s}^{DF} \tag{1.26}$$

$$\text{Mean RMSE}\left(\hat{\Delta}_i^{DF}\right) = \frac{1}{S}\sum_{s=1}^{S}\left[\left(\hat{\Delta}_{i,s}^{DF}\right)^2\right]^{1/2} \tag{1.27}$$

Table 1.1. Likelihood Discrepancies, GDP Measurement Model

	Bootstrap Particle Filter					
Number of particles N	100	500	1,000	5,000	10,000	50,000
Mean Bias $\hat{\Delta}_1$	-2.484	-0.505	-0.257	-0.054	-0.029	-0.006
Mean StdD $\hat{\Delta}_1$	2.292	0.991	0.698	0.310	0.221	0.100
Mean RMSE $\hat{\Delta}_1$	3.394	1.121	0.750	0.318	0.224	0.101
Mean Bias $\hat{\Delta}_2$	-0.014	-0.003	-0.001	-0.001	-0.002	0.000
Mean StdD $\hat{\Delta}_2$	3.889	1.164	0.771	0.318	0.225	0.100
Mean RMSE $\hat{\Delta}_2$	3.933	1.171	0.775	0.320	0.226	0.101
	Discretization Filter					
Rule of thumb constant c	1/2	1	3	5	7	10
Mean Bias $\hat{\Delta}_1$	-0.405	-0.040	0.001	0.002	0.001	0.001
Mean StdD $\hat{\Delta}_1$	-	-	-	-	-	-
Mean RMSE $\hat{\Delta}_1$	1.287	0.383	0.114	0.070	0.053	0.042
Mean Bias $\hat{\Delta}_2$	0.085	0.029	0.007	0.004	0.003	0.002
Mean StdD $\hat{\Delta}_2$	-	-	-	-	-	-
Mean RMSE $\hat{\Delta}_2$	1.119	0.391	0.113	0.069	0.051	0.039

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

Table 1.2 reports the average absolute and relative evaluation times of the likelihood function across all specifications. The absolute times are reported in seconds. For the PF, these are computed as the average across samples and particle draws. For the DF and the KF, these are simply reported as averages across the samples. The relative times are computed as the time of one evaluation of the likelihood function relative to the time it takes for the KF.

Considered together, tables 1.1 and 1.2 provide a better understanding of the tradeoff between accuracy and computational complexity that both the DF and PF exhibit. As an example, note that the evaluation of the likelihood using 100 particles for the PF and a rule of thumb constant of 7 for the DF take roughly the same amount of time, about 0.02 seconds. However, the DF is 2 orders of magnitude more accurate in terms of RMSE. Similarly, consider the PF with 50,000 particles and the DF with a rule of thumb constant of 3. These are roughly the same in terms of RMSE, but the DF evaluation

Table 1.2. Computation Time of 1 Likelihood Evaluation (in seconds), GDP Measurement Model

	Kalman Filter					
Mean Time	0.007					
	Bootstrap Particle Filter					
Number of particles N	100	500	1,000	5,000	10,000	50,000
Mean Time	0.020	0.039	0.055	0.194	0.351	1.845
Mean Relative Time	3.13	6.13	8.66	30.30	54.84	288.63
	Discretization Filter					
Rule of thumb constant c	1/2	1	3	5	7	10
Mean Time	0.009	0.009	0.011	0.014	0.020	0.032
Mean Relative Time	1.38	1.37	1.69	2.16	3.10	4.99

of the likelihood is about 170 times faster. Examining the other elements of the tables leads to a similar conclusion: the DF offers a much better tradeoff between accuracy and computation time than the PF.

1.6.2 Stochastic Volatility

Next, I compare the performance of different estimation procedures on a stochastic volatility model. The standard discrete time stochastic volatility model, as formulated in Taylor (1982), is given by

$$X_t = \mu(1-\rho) + \rho X_{t-1} + v_t \qquad v_t \sim \text{i.i.d. } N\left(0, \sigma^2\right) \tag{1.28}$$

$$Y_t = e^{X_t/2} w_t \qquad w_t \sim \text{i.i.d. } N(0,1) \tag{1.29}$$

Note that the measurement equation can be equivalently rewritten as:

$$\log\left(Y_t^2\right) = X_t + \log\left(w_t^2\right) \tag{1.30}$$

which leads to an additively separable state equation.[16] However, this simplification

[16]This is the specification of the observation equation I use in the EKF estimation. This can also be

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

I simulate data for T = 100, 500, and 1,000 periods, and compute the likelihood of the model eight different ways: the DF using six different choices of the rule of thumb constant c, the boostrap PF with adaptive resampling using 1,000 particles, and the extended Kalman filter (EKF). Each specification is simulated 1,000 times and estimation is performed via maximum likelihood where optimization is done using MATLAB's genetic algorithm in the global optimization toolbox. The random seed used to construct the particle filter approximation is fixed for a given sample in order to make the optimization better behaved.[17]

Figures 1.2, 1.3, and 1.4 display the sampling distributions of the maximum likelihood estimators. The rows of each figure correspond to a particular model parameter and the columns correspond to a particular method of approximating the likelihood. A vertical line is displayed at the point of the true parameter value. All estimation using the discretization filter uses the Rouwenhorst (1995) discretization scheme.[18]

Note that for small sample sizes, $T = 100$, there is a considerable downward bias in the estimation of ρ and σ. That is, the optimization algorithm is picking values of ρ and σ extremely close to 0. This bias is most severe in the EKF estimates, especially

thought of as a misspecified Kalman filter where the measurement error is incorrectly assumed to be Gaussian.

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

[18]Estimation was also performed using the Farmer and Toda (2016) method, the Tauchen (1986) method, and the point-mass filter. The Rouwenhorst method performs the best although the relative gains of the discretization filter are similar across all discretization methods.

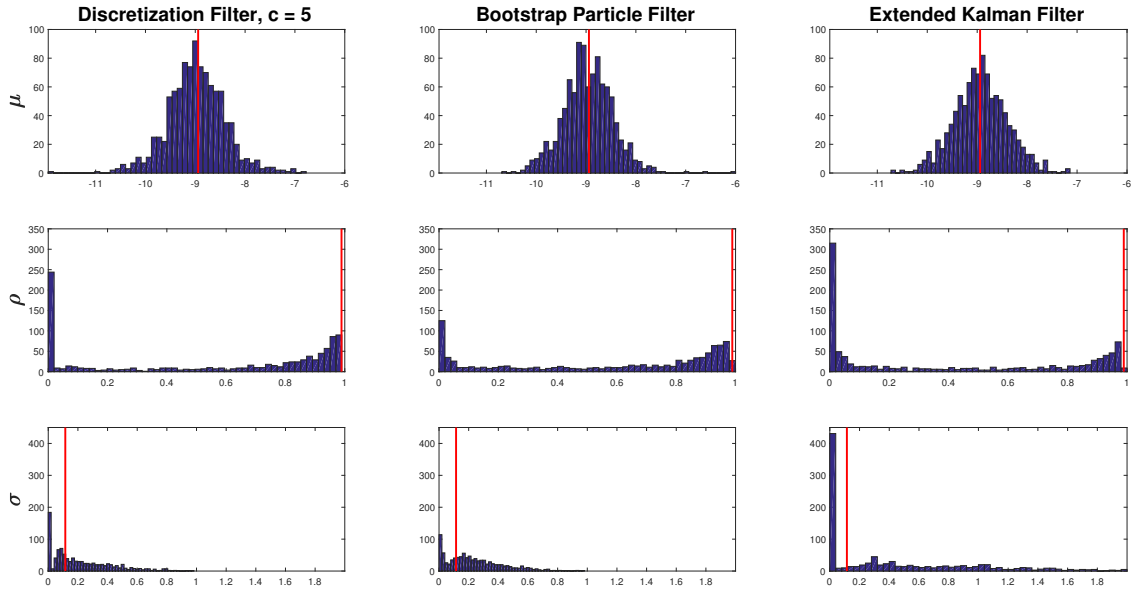


Figure 1.2. MLE Sampling Distributions for Sample Size $T = 100$, Stochastic Volatility Model

for σ. However, this is not particularly surprising because the EKF is estimating a misspecified model, where it is treating the residual in the observation equation as a normal random variable, even though it has a $\log\left(\chi_1^2\right)$ distribution.

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

I also compute the root mean squared error (RMSE) and the bias of the parameter estimates, approximating the population expectation with an average across simulations.

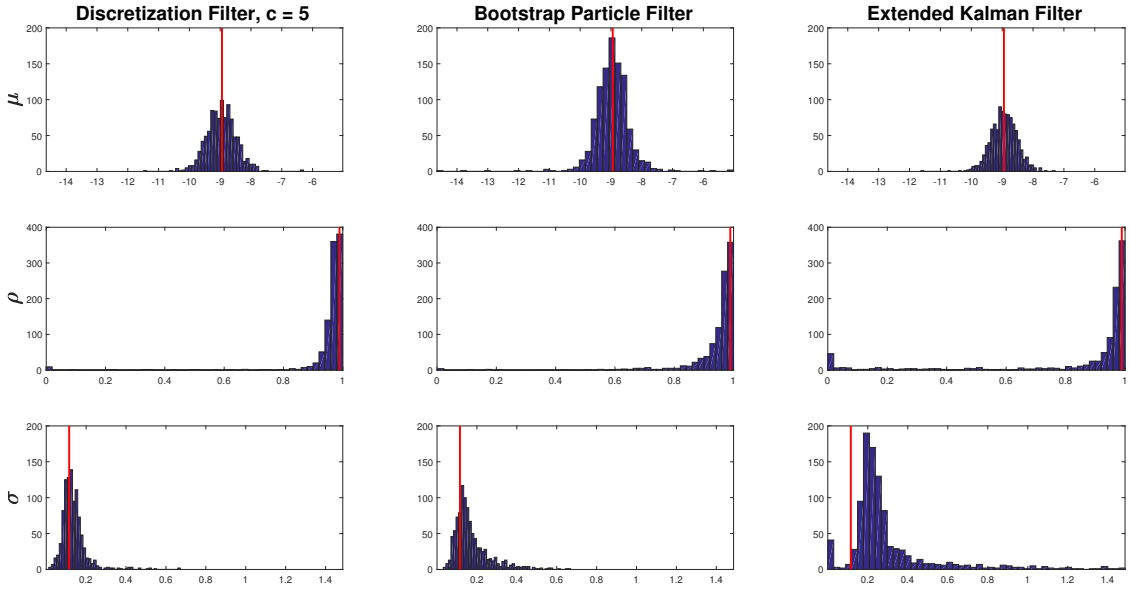


Figure 1.3. MLE Sampling Distributions for Sample Size $T = 500$, Stochastic Volatility Model

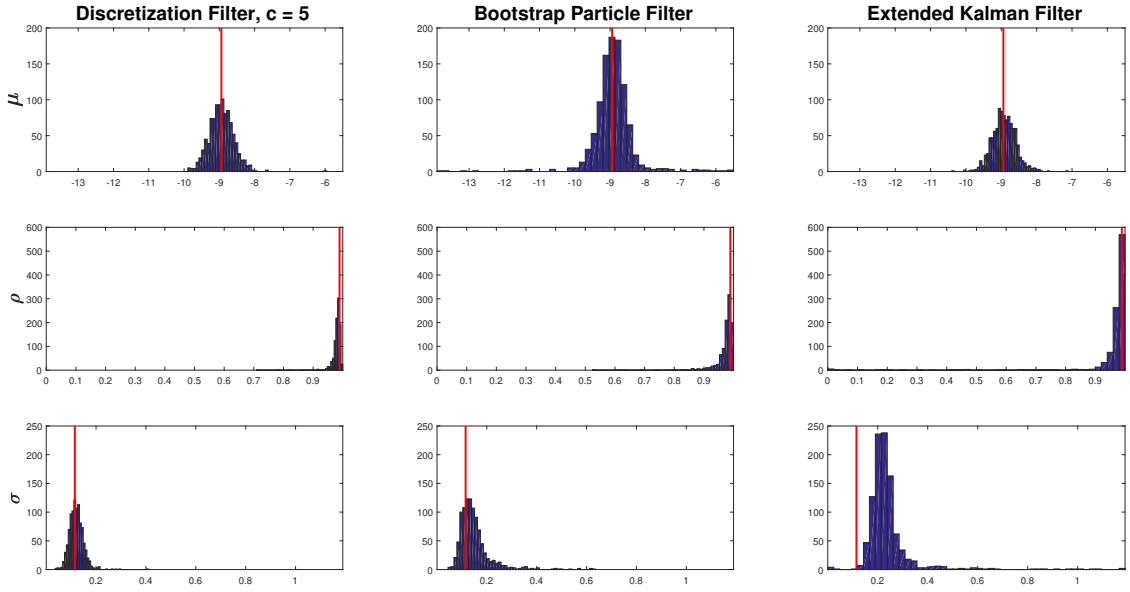


Figure 1.4. MLE Sampling Distributions for Sample Size $T = 1,000$, Stochastic Volatility Model

In particular, for the i-th component of the parameter vector, I compute:

$$\text{RMSE}\left(\hat{\theta}_i\right) = \sqrt{\mathbb{E}\left[\left(\hat{\theta}_i - \theta_i\right)^2\right]} \tag{1.31}$$

$$\text{Bias}\left(\hat{\theta}_i\right) = \mathbb{E}\left[\theta_i\right] - \theta_i \tag{1.32}$$

and report the results in table 1.3.

First consider the DF with $c = 5$ and its performance relative to the PF and the EKF. The DF and the PF are similar in terms of RMSE and bias for $T = 100$, however the DF generally outperforms the PF for the larger sample sizes. The EKF is unambiguously the worst except for estimation of the mean parameter μ. It is also interesting to note that the performance of the PF for estimating μ actually deteriorates for larger sample sizes, which seems to be evidence of sample thinning, a well known problem with importance sampling methods.

Next I examine the performance of the DF for different values of the rule of thumb constant c. For $T = 100$ and to a lesser extent for $T = 500$, the RMSE and bias actually seem to increase for larger values of c. There are a couple of possible explanations for this phenomenon. The first is that the asymptotic analysis in section 1.4 considers the case of a compact state space, whereas in this example as in most examples of economic interest, the state variable resides in an unbounded space. Thus, as the discretization is being constructed for larger values of c, the number of points is increasing, but so is the domain over which the approximation is constructed. This could potentially cause numerical issues for smaller sample sizes, because the discretization points cover large areas of the state space which are never visited in the sample.

A second possibility is that these larger numbers are actually more consistent with the RMSE and bias of the infeasible maximum likelihood estimator. In other words, the misspecification caused by small values of M is actually acting as a type of regularization

Table 1.3. Accuracy of Parameter Estimates, Stochastic Volatility Model

Root Mean Squared Error

		Discretization Filter						PF	EKF
ROT constant c		1/2	1	3	5	7	10	-	-
μ	**T = 100**	0.511	0.538	0.611	0.618	0.623	0.669	0.488	0.521
	T = 500	0.450	0.475	0.511	0.516	0.486	0.508	0.574	0.445
	T = 1,000	0.343	0.364	0.370	0.381	0.339	0.391	0.614	0.336
ρ	**T = 100**	0.598	0.584	0.603	0.617	0.630	0.637	0.572	0.727
	T = 500	0.108	0.080	0.101	0.103	0.121	0.134	0.108	0.304
	T = 1,000	0.014	0.014	0.015	0.015	0.014	0.015	0.042	0.126
σ	**T = 100**	0.228	0.225	0.236	0.238	0.246	0.251	0.244	0.621
	T = 500	0.061	0.057	0.061	0.064	0.072	0.072	0.111	0.293
	T = 1,000	0.027	0.027	0.027	0.027	0.027	0.027	0.076	0.163

Bias

		Discretization Filter						PF	EKF
ROT constant c		1/2	1	3	5	7	10	-	-
μ	**T = 100**	-0.036	-0.039	-0.035	-0.029	-0.027	-0.031	-0.028	0.005
	T = 500	-0.031	-0.027	-0.017	-0.015	-0.012	-0.018	0.003	-0.001
	T = 1,000	-0.008	0.015	0.007	0.015	0.015	0.026	0.016	0.020
ρ	**T = 100**	-0.441	-0.427	-0.455	-0.475	-0.493	-0.504	-0.442	-0.617
	T = 500	-0.030	-0.030	-0.036	-0.035	-0.038	-0.041	-0.048	-0.136
	T = 1,000	-0.008	-0.009	-0.009	-0.009	-0.009	-0.009	-0.017	-0.034
σ	**T = 100**	0.111	0.105	0.108	0.106	0.111	0.113	0.134	0.340
	T = 500	0.020	0.019	0.022	0.021	0.022	0.023	0.062	0.186
	T = 1,000	0.006	0.006	0.006	0.006	0.006	0.006	0.035	0.125

which is outperforming the maximum likelihood estimator for small sample sizes. Note that this phenomenon is absent for larger sample sizes, and the estimates of the RMSE and bias appear stable across all values of c.

Table 1.4 displays the average simulation times for all eight specifications. The differences in computational time are stark. With $c = 1$, the EKF is 32 times faster than the DF for small sample sizes and 78 times faster for large ones. However, this is at the cost of parameter estimates which are significantly less accurate for larger sample sizes. Furthermore, the EKF estimate of σ appears to be significantly biased, even asymptotically, due to the misspecification of the observation equation.

Table 1.4. Computation Time of 1 Likelihood Evaluation (in seconds), Stochastic Volatility Model

	Discretization Filter						**PF**	**EKF**
ROT constant c	1/2	1	3	5	7	10	-	-
T = 100	0.001	0.001	0.002	0.003	0.005	0.009	0.010	0.000
T = 500	0.002	0.003	0.008	0.017	0.034	0.083	0.120	0.000
T = 1,000	0.005	0.006	0.018	0.046	0.101	0.273	0.401	0.000

For estimates which are roughly the same accuracy for $T = 100$, the DF is an order of magnitude faster than the PF. For $T = 1,000$, the DF is between twice and three times as accurate as the particle filter while being 2 orders of magnitude faster. These results suggest that the DF is somewhere in between the EKF and the PF in terms of computational burden, while delivering accurate parameter estimates. To give the reader a rough idea, all of the simulations for the DF and the EKF ran in a matter of minutes to hours whereas the most computationally burdensome PF specification ($T = 1,000$) took almost five days to run operating in parallel on four cores. These reductions in computation time make the estimation of many dynamic macroeconomic and financial models feasible. In the case of the Gabaix (2012) rare disasters model, the estimation takes several hours running MATLAB on a standard desktop computer, whereas estimation using a PF would likely take several weeks.

Another important dimension for comparison is the accuracy of the filtered states, $\left\{\hat{x}_{t|t}\right\}_{t=1}^{T}$. I provide results on the root mean square error (RMSE) and the mean absolute error (MAE) of all the methods. For a given model specification and method, these are defined as:

$$\text{RMSE} = \left(\frac{1}{T}\sum_{t=1}^{T}\left(\hat{x}_{t|t} - x_t\right)^2\right)^{1/2} \tag{1.33}$$

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}\left|\hat{x}_{t|t} - x_t\right| \tag{1.34}$$

I define the average RMSE (ARMSE) and average MAE (AMAE) to be the average of the RMSE and the MAE across simulations for a given method. Table 1.5 displays the ARMSE and AMAE of each method, where the filtering is done using the corresponding maximum likelihood estimates of the parameters for a given sample.

Table 1.5. Accuracy of Filtered State Estimates, Stochastic Volatility Model

Average Root Mean Squared Error

	Discretization Filter						**BPF**	**EKF**
ROT constant c	1/2	1	3	5	7	10	-	-
T = 100	0.362	0.360	0.362	0.365	0.368	0.370	0.374	0.498
T = 500	0.378	0.379	0.385	0.388	0.390	0.391	0.383	0.465
T = 1,000	0.379	0.381	0.385	0.386	0.386	0.387	0.383	0.452

Average Absolute Mean Error

	Discretization Filter						**BPF**	**EKF**
ROT constant c	1/2	1	3	5	7	10	-	-
T = 100	0.297	0.294	0.295	0.297	0.299	0.300	0.307	0.390
T = 500	0.302	0.302	0.304	0.305	0.306	0.306	0.306	0.372
T = 1,000	0.302	0.303	0.304	0.304	0.304	0.304	0.305	0.361

The DF and PF perform roughly the same for all sample sizes. However, keep in mind that this is for dramatically different estimation times for the parameters as discussed above. The misspecification of the measurement error distribution using the EKF translates into poor estimates of the unobserved state.

1.7 Variable Rare Disasters

In this section, I provide the first estimates of the Gabaix (2012) model of variable rare disasters. I show how likelihood-based estimation can be used as a model diagnosis tool. In particular, I find that (i) the estimated model fails to identify the Great Recession as a disaster episode, and (ii) the model cannot capture the change in the dynamics of the price-dividend ratio starting in the 1990s. To explain (i), the model requires a positive

expected jump in inflation in the event of a disaster in order to generate an upward sloping nominal yield curve. However, during the Great Recssion, we observed a strongly upward sloping nominal yield curve in conjunction with close to zero inflation and even deflation. For (ii), the model specifies a process which is close to an AR(1) which governs the dynamics of the price-dividend ratio, while the price-dividend ratio starting the 1990s appears to exhibit a structural break both in its mean and its dynamics.

1.7.1 Model Setup

The model is an endowment economy where a representative agent has lifetime expected utility over consumption given by:

$$\mathbb{E}_0\left[\sum_{t=0}^{\infty} e^{-\rho t}\frac{C_t^{1-\gamma}}{1-\gamma}\right]$$

$\gamma > 0$ is the coefficient of relative risk aversion, and $\rho > 0$ is rate of time preference. Each period she receives consumption endowment C_t. For expositional purposes, I only present the version of the model with CRRA utility here, however for estimation purposes I consider the full Epstein-Zin version of the model which allows risk aversion and the IES to be independently estimated.

The endowment stream is hit by large but infrequent disasters. The dynamics of consumption are given by:

$$\Delta c_{t+1} = g_C + w_{t+1} b_{t+1} \tag{1.35}$$

where g_C is the normal-time growth rate of the economy, $B_{t+1}e^{g_C}$ is the growth rate if a disaster occurs ($b_{t+1} := \log B_{t+1}$), and w_{t+1} is an indicator for whether a disaster occurs at time $t+1$, which happens with probability p_t.

Consider a stock i which is a claim to a stream of dividend payments $(D_t)_{t\geq 0}$.

The growth rate of its dividends is assumed to follow

$$\Delta d_{t+1} = g_D + \varepsilon^D_{t+1} + w_{t+1} f_{t+1} \tag{1.36}$$

where g_D is the growth rate of dividends in normal times, ε^D_{t+1} is a mean zero shock that is independent of the disaster event, and F_{t+1} ($f_{t+1} := \log F_{t+1}$) is the recovery rate of the dividend. That is, in the event of a disaster, there can be "partial default." If $F_{t+1} = 0$, the asset is completely destroyed, and if $F_{t+1} = 1$ there is no loss relative to normal times.

In contrast with some of the other more recent papers on variable rare disasters such as Wachter (2013) and Gourio (2012), the probability of a disaster is fixed in the baseline model. It is the severity of a disaster which is time-varying. The combination of variations in the disaster probability and the severity are captured by a variable called "resilience." Define resilience H_t of the asset as

$$H_t \equiv p_t \mathbb{E}^D_t \left[B^{-\gamma}_{t+1} F_{t+1} - 1 \right] \tag{1.37}$$

Assets with high resilience are safer than assets with low resilience because they pay out more in disaster states, and thus will command lower risk premia.

As in Gabaix (2012), I split resilience into a constant part H_* and a variable part $\hat{H}_t$ with mean zero. The dynamics of $\hat{H}_t$ are assumed to follow a linearity-generating process (Gabaix 2009)

$$\hat{H}_{t+1} = \frac{1 + H_*}{1 + H_t} e^{-\phi_H} \hat{H}_t + \varepsilon^H_{t+1} \tag{1.38}$$

Linearity generating processes behave like first-order autoregressive processes close to their steady state but display nonlinear dynamics as they reach more extreme values.

Define $\delta \equiv \rho + \gamma g_C$, $h_* \equiv \log(1 + H_*)$, and $\delta_i \equiv \delta - g_D - h_*$. It can be shown

that the price-dividend ratio of the asset is given by

$$\frac{P_t}{D_t} = \frac{1}{1-e^{-\delta_i}}\left(1+\frac{e^{-\delta_i-h_*}\hat{H}_t}{1-e^{-\delta_i-\phi_H}}\right) \tag{1.39}$$

The unconditional equity premium for the asset is given by

$$r_t^e = \delta - H_t - p_t\mathbb{E}_t\left[1-F_{t+1}\right] - r_f \tag{1.40}$$

where r_f, the risk-free rate, is given by

$$r_f = \delta - p_t\mathbb{E}_t\left[B_{t+1}^{-\gamma}-1\right] \tag{1.41}$$

Turning to the nominal side of the economy, inflation $I_t = I_* + \hat{I}_t$ is assumed to vary exogenously and its non-constant component $\hat{I}_t$ also follows a linearity-generating process. In addition, inflation jumps by an amount $J_t = J_* + \hat{J}_t$ in the event of a disaster. J_* is the baseline jump in inflation in the event of a disaster, and $\hat{J}_t$ is a mean-reverting deviation in this jump size from its baseline. Their dynamics are jointly given by

$$\hat{I}_{t+1} = \frac{1-I_*}{1-I_t}\left(e^{-\phi_I}\hat{I}_t + w_{t+1}J_t\right) + \varepsilon_{t+1}^I \tag{1.42}$$

$$\hat{J}_{t+1} = \frac{1-I_*}{1-I_t}e^{-\phi_J}\hat{J}_t + \varepsilon_{t+1}^J \tag{1.43}$$

where ε_{t+1}^I and ε_{t+1}^J are mean zero shocks which are uncorrelated with disasters, but may be correlated with each other. This allows me to define the variable π_t, the variable part of the bond premium, as

$$\pi_t \equiv \frac{p_t\mathbb{E}_t\left[B_{t+1}^{-\gamma}F_{\$,t+1}\right]}{1+H_\$}\hat{J}_t$$

π_t is what controls deviations of the slope of the nominal yield curve from its typical

value, while inflation controls the level relative to the real yield.

Define $\Psi \equiv e^{-\delta}(1+H_{\$})(1-I_*)$, $\tilde{\rho}_I \equiv \frac{e^{-\phi_I}+\kappa}{1-\kappa}$, and $\tilde{\rho}_J \equiv \frac{e^{-\phi_J}}{1-\kappa}$. The price of a nominal zero-coupon bond of maturity T at time t is given by

$$Z_{\$t}(T) = (\Psi(1-\kappa))^T \times \left\{ 1 - \frac{1}{1-\kappa}\frac{1-\tilde{\rho}_I^T}{1-\tilde{\rho}_I}\left(\frac{\hat{I}_t}{1-I_*} - \kappa\right) - \frac{1}{(1-\kappa)^2}\frac{\frac{1-\tilde{\rho}_I^T}{1-\tilde{\rho}_I} - \frac{1-\tilde{\rho}_J^T}{1-\tilde{\rho}_J}}{\tilde{\rho}_I - \tilde{\rho}_J}\frac{\pi_t}{1-I_*} \right\}$$

The corresponding yield is

$$y_t(T) = -\frac{\ln Z_{\$t}(T)}{T}. \tag{1.44}$$

I now turn to the details of the estimation.

1.7.2 Estimation

I fix a subset of parameters related to the cash flow dynamics, the severity of disasters, and inflation in Table 1.6.

Table 1.6. Calibrated Parameters, Rare Disasters Model

Parameters	**Values**
Growth rate of consumption and dividends	$g = g_C = g_D = 2.5\%$
Volatility of dividend growth	$\sigma_D = 11\%$
Recovery rate of C after a disaster	$\bar{B} = 0.66$
Stock's recovery rate: typical value	$F_{i*} = \bar{B} = 0.66$
Inflation: typical value	$I^* = 3.8\%$

The means of consumption and dividend growth, the volatility of dividend growth, the recovery rate of consumption after a disaster, and the typical value of the stock's recovery rate are fixed to the values used in Gabaix (2012). I set the typical value of inflation to be 3.8%, which is the sample average of CPI inflation in my sample.

For my baseline estimation results, I use monthly data on the price-dividend ratio of the CRSP value-weighted portfolio, nominal yields on U.S. Treasury securities, and CPI inflation from June 1961 to December 2015. This is the longest sample for which all variables are available. The data on nominal yields are constructed as in Gürkaynak, Sack, and Wright (2007) and I use maturities of 3 and 6 months, 1, 2, 5, 7, and 10 years. Inflation is constructed as the 12-month change in log CPI.

The model has three state variables: resilience $\hat{H}_t$, inflation $\hat{I}_t$, and jumps in inflation $\hat{J}_t$. The mapping from resilience to the price-dividend ratio is given by (1.39) and the mapping from inflation and jumps in inflation to nominal yields is given by (1.44). I assume that the price-dividend ratio, nominal yields of all maturities, and inflation itself are observed with error with measurement errors given by $\varepsilon_{PD_{obs}} \sim N(0, \sigma_{PD_{obs}})$, $\varepsilon_{y_{obs}} \sim N(0, \sigma_{y_{obs}})$, and $\varepsilon_{I_{obs}} \sim N(0, \sigma_{I_{obs}})$ respectively. The measurement errors are assumed to be independent of each other and all other quantities in the model.

I estimate the vector of 13 parameters (10 structural and 3 measurement error variances)

$$\theta \equiv (\rho, \gamma, \psi, p, \phi_H, \sigma_I, \phi_I, J_*, \sigma_J, \phi_J, \sigma_{PD_{obs}}, \sigma_{y_{obs}}, \sigma_{I_{obs}})'$$

by maximum likelihood using the Farmer and Toda (2016) method with an 11 point grid for each state variable. This results in a total of $11^3 = 1,331$ discrete points. Table 1.7 shows the estimated parameters, with quasi maximum likelihood robust standard errors in parentheses, along with the calibrated values used in Gabaix (2012).

First, consider the values of the preference parameters ρ, γ, ψ. The estimated value of the rate of time preference ρ, 3.07%, is significantly lower than its calibrated value of 6.57%. In terms of annual discount factors, this translates into the difference between 0.970 and 0.936. Next, the estimated coefficient of relative risk aversion γ, 2.8, is significantly lower than its calibrated value of 4. This is heartening because

Table 1.7. Estimated Parameters, Rare Disasters Model

Parameters	**Estimated Values**	**Gabaix Calibration**
Time preference, ρ	3.07% (1.59%)	6.57%
Risk aversion, γ	2.812 (0.487)	4
Intertemporal elasticity of substitution, ψ	0.257 (0.101)	0.25
Probability of a disaster, p	4.81% (0.72%)	3.63%
Resilience:		
volatility, σ_F	7.0%	10%
speed of mean reversion, ϕ_H	12.48% (14.41%)	13%
Inflation:		
conditional volatility, σ_I	0.61% (3.08%)	1.5%
speed of mean reversion, ϕ_I	15.21% (5.14%)	18%
Jump in inflation:		
typical value, J_*	2.56% (0.53%)	2.1%
conditional volatility, σ_J	6.83% (6.71%)	15%
speed of mean reversion, ϕ_J	82.13% (95.33%)	92%
Volatility of measurement errors:		
price-dividend ratio, $\sigma_{PD_{obs}}$	2.62 (5.76)	-
nominal yields, $\sigma_{y_{obs}}$	0.43% (0.54%)	-
inflation, $\sigma_{I_{obs}}$	2.66% (7.14%)	-

traditionally asset pricing models require what are often considered unreasonably high values of risk aversion in order to match financial data. This number is more in line with typical macroeconomic calibrations of DSGE models. Lastly, the IES ψ is estimated to be 0.26, and importantly, is significantly less than 1. This is consistent with the empirical micro evidence, but at odds with values that are typically chosen in asset pricing models.

Second, the probability of a disaster is estimated to be 4.81% annually, compared to the calibrated value of 3.63% which comes from Barro and Ursúa (2008). Given that there are a very few observations of consumption disasters in the data, it seems reasonable to think that this probability may be higher than existing empirical estimates that rely on macro data.

Lastly, the estimates of the parameters governing the dynamics of inflation and jumps in inflation differ between the estimated and calibrated models. The estimated

model favors more peristent and less volatile processes for both of these quantities.

Table 1.8. Parameter Values Implied by Estimation, Rare Disasters Model

Parameters	**Estimated Values**	**Gabaix Calibration**
Ramsey discount rate, δ	12.8%	16.6%
Risk-adjusted probability of disaster, $p\mathbb{E}\left[B_{t+1}^{-\gamma}\right]$	15.5%	19.2%
Stocks:		
effective discount rate, δ_i	1.7%	5%
Stock resilience:		
typical value, H_*	8.6%	9.0%
volatility, σ_H	1.1%	1.9%
Stocks, equity premium:		
conditional on no disasters	5.3%	6.5%
unconditional	3.6%	5.3%
Real short-term rate	2.1%	1.0%
Resilience of one nominal dollar, $H_\$$	10.7%	16.0%
5-year nominal slope $y_t(5) - y_t(1)$:		
mean	0.55%	0.57%
volatility	0.81%	0.92%
Long-run − short-run yield:		
typical value, κ	3.5%	2.6%
Inflation:		
I_{**}	7.3%	6.3%
ψ_I	8.2%	13%
ψ_J	78.6%	90%
Bond risk premium:		
volatility, σ_π	0.95%	2.9%

I next consider some additional quantities implied by the model evaluated using both the estimated parameter values and the calibrated ones. The results are presented in table 1.8. A key quantity of interest is the risk-adjusted probability of a disaster, given by $p\mathbb{E}\left[B_{t+1}^{-\gamma}\right]$. This is the quantity that allows the model to match high average risk premia. The estimated model implies a value of 15.5%, less than 4 percentage points lower than the calibrated value of 19.2%. Given that $B_{t+1} = \bar{B}$ is fixed across both specifications, the differences in this quantity are coming from differences in the probability of a disaster and the coefficient of relative risk aversion. The higher probability of a disaster and

lower value of risk aversion estimated by maximum likelihood allow the model to remain broadly consistent with a wide variety of asset pricing facts.

In particular, the model still achieves an unconditional equity premium of 3.6%, roughly half of what it is in the data, while the calibration produces 5.3%. The estimated real short-term rate is a bit high at 2.1% relative to the calibration which targets 1%, although this is consistent with the historical average of data back to 1891 is considered. The estimated model matchces the average level and volatility of the 5-year slope of the nominal yield curve produced by the calibration.

1.7.3 Implications of the Filtered State Estimates

I now focus on two implications of the model which come from the ability to examine the filtered and smoothed states implied by the estimated parameters. First, from the processes for inflation and jumps in inflation, I can back out the implied probability of a disaster having occured in any given period. Note that in the event of a disaster, the conditional mean of inflation at time $t+1$ given information up to time t is

$$\frac{1-I_*}{1-I_t}\left(e^{-\phi_I}\hat{I}_t + J_t\right)$$

whereas in the event of no disaster, this conditional mean is

$$\frac{1-I_*}{1-I_t}e^{-\phi_I}\hat{I}_t$$

By running the discretization filter at the estimated parameter vector, I can obtain filtered and smoothed estimates of the time series $\left\{\hat{I}_t\right\}_{t=1}^{T}$ and $\{J_t\}_{t=1}^{T}$. Since the innovation to inflation each period, ε_{t+1}^{I}, has a normal distribution with standard deviation σ_I, I can compute the likelihood of having observed the value of $\hat{I}_{t+1}$ implied by these estimates in the event of a disaster and in the event of no disaster. Figure 1.5 plots the probability

of a disaster having occurred in each period of the sample by applying this procedure using both the filtered and smoothed estimates of the states. The results using the filtered states are in blue and the results using the smoothed states are in red.

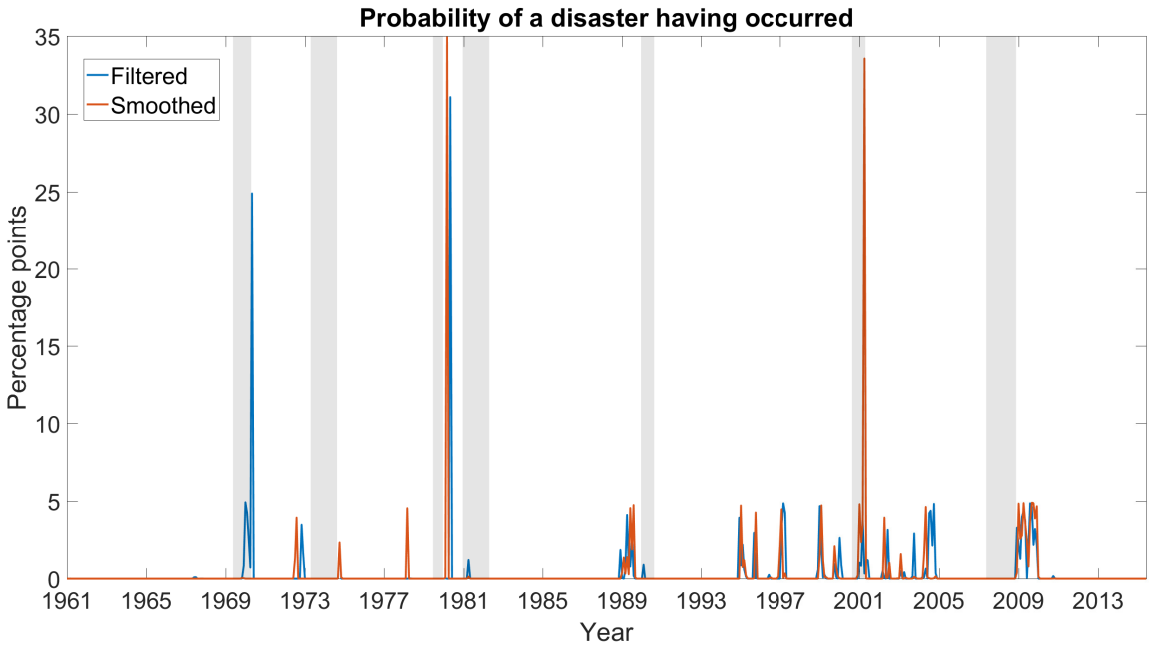


Figure 1.5. Estimates of Disaster Probability, Rare Disasters Model

First, note that the filtered and smoothed estimates together identify three potential disaster episodes over the sample. I will refer to a "potential disaster episode" as a period where the estimated model assigns more than a 20% chance to a disaster having occured. The filtered estimates identify the early 1970s and the late 1970s as potential disaster episodes. The smoothed estimates identify the same period in the late 1970s and a then a period in the early 2000s coinciding with the dot com bubble as potential disaster episodes.

While the two series do not fully agree on which periods are potential disaster episodes, they both come to the same conclusion regarding the Great Recession. At no point during the Great Recession does either series assign more than a 5% chance

of a disaster having occurred. While this may seem surprising at first, upon further investigation it makes a lot of sense.

During the Great Recession, the U.S. experienced low inflation relative to the rest of the sample, and even a period of deflation. However, at the same time, the nominal yield curve was upward sloping. The Gabaix model achieves an upward sloping nominal yield curve through an expected positive jump in inflation. Since the model is being fit to both inflation data and data on nominal yields, it is trying to reconcile a period of expected low inflation / deflation with a period of upward sloping nominal yield curves but ends up splitting the difference. This suggests a shortcoming of the Gabaix framework, which is that rare disasters are typically coupled with expected increases in inflation. However, in the U.S. and many other developed countries, financial crises are typically coupled with deflation and upward sloping nominal yield curves.

Next, I examine the model's implications for the recovery rate of stocks, F_t. Recall that F_t is the fraction of its value that a stock retains in the event of a disaster. The recovery rate is an affine function of the state variable resilience $\hat{H}_t$, for which I construct filtered and smoothed estimates and plot in figure 1.6. As above, the results using the filtered states are in blue and the results using the smoothed states are in red. The black line is the long run average of the recovery rate, which is calibrated to be 66% as in Gabaix (2012).

What immediately jumps out is that before the late 1990s, the recovery rate is estimated to be about 20% on average, persistently low relative to its long run average of 66%. This shoots up to almost 100% during the dot com bubble and crashes back down to its long run average in the mid 2000s. It again experiences a sharp decline during the Great Recession and bounces back close to its long run average at the end of the sample. This is counterintuitive because it suggests that the model considers the Great Moderation to be particularly risky relative to the rest of the sample. On average, investors were

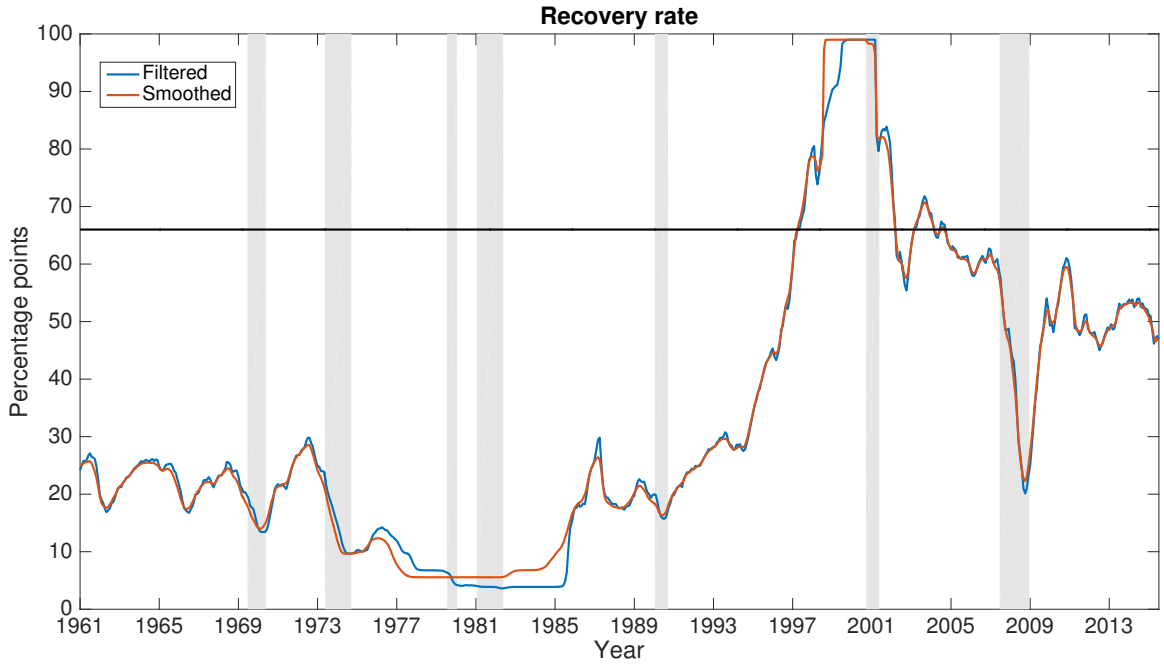


Figure 1.6. Estimates of Recovery Rate, Rare Disasters Model

expecting to lose about 80% of the value of their assets in the event of a disaster whereas the model implies that they should typically expect to lose 34%.

This result begins to make a lot more sense when one examines the connection between the recovery rate and the price-dividend ratio. In the Gabaix model, the price-dividend ratio of a stock is an affine function of its recovery rate. Unsurprisingly, movements in the recovery rate are closely linked to movements in the price-divdend ratio, with the only differences being attributed to measurement error.

The differences in the implied moments of the price-dividend ratio and stock returns are presented in table 1.9. The first thing that stands out from looking at this table is that the level of the price-dividend ratio implied by the estimated model parameters is about three times larger than the value implied by the calibrated model. This has a lot to do with the sample used in the estimation. The historical average of the price-dividend ratio targeted by Gabaix, 23, is computed using data that ends in 1997. However, the data

I use for estimation goes all the way up to 2015, which includes the dot-com bubble and subsequent Great Recession.

Table 1.9. Stock Market Moments

	Data (Campbell Sample 1891-1997)	**Data (Estimation Sample 1961-2015)**	**Estimated Model**	**Calibrated Model**
Mean P/D	23	39.2	57.6	18.2
Std. dev. $\ln P/D$	0.33	0.40	0.21	0.30
Std. dev. of stock returns	0.18	0.15	0.11	0.15

The price-dividend ratio reaches a maximum value of 92 in the early 2000s and has an average value of 39 over my sample, almost twice the value targeted by Gabaix. The estimation chooses values of the structural parameters that allow the model to achieve these high values of the price-dividend ratio. The estimated model also understates the volatility of the price-dividend ratio. Unsurprisingly, given the lower volatility and higher mean of the price-dividend ratio, this results in a lower volatility of stock returns than the calibrated model, 11% vs. 15%, for the same values of the cash flow parameters.

Given the pronounced change in the both the level of the price-dividend ratio and its dynamics (sharper decreases and increases) after 1997, the fit of the Gabaix model may be greatly improved by allowing for switches in the parameters governing the long-run average, persistence, and volatility of the recovery rate. This would help produce more sensible economic estimates of the recovery rate.

The estimation of both the probability of a disaster having occurred and the recovery rate is an exercise which can only be conducted using likelihood-based estimation procedures. This highlights an advantage of likelihood-based methods over calibration and other moment-matching based methods: the ability to construct estimates of the hidden state variables. By constructing estimates of the hidden state variables, one is able to consider the model's implications for dynamics in addition to moments. The calibrated

version of the Gabaix model does an excellent job of matching several moments of asset pricing data related to equities, bonds, and options. However, the estimation shows that the model also exhibits a couple of important shortcomings regarding the coupling of rare disasters with positive expected inflation and economically counterintuitive implications for the recovery rate of equity.

1.7.4 Model Comparison

Finally, I formally test the null hypothesis that the estimated model provides a better fit to the data than the calibrated model. To do this, I fix the parameters as calibrated in Gabaix (2012) and estimate the measurement errors using the same data as the full estimation outlined previously. Denote the resulting parameter vector as θ_0. I test the null hypothesis that $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta = \hat{\theta}$ using a likelihood ratio test. The likelihood ratio statistic is given by

$$LR = 2\left[\ell_{T,M}\left(\hat{\theta}\right) - \ell_{T,M}\left(\theta_0\right)\right] = 7,471$$

This is compared to a $\chi^2(10)$ because the unrestricted model has 10 extra parameters that are freely estimated. The 99% critical value for the test is 23.21 and thus the null hypothesis is overwhelmingly rejected in favor of the alternative. Again, this is unsurprising given the calibrated model's inability to match the extreme values of the price-dividend ratio observed in the 2000s. The only way the calibrated model can rationalize these observations is by choosing unreasonably large values of the measurement error variance for the price-dividend ratio.

1.8 A Term Structure Model with a Zero Lower Bound

In this section, I re-estimate the term structure model proposed in Wu and Xia 2016, and provide an updated estimate of their shadow rate series with data through January 2014. Using the discretization filter I am able to replicate most of their parameter estimates. While my filtered series and the Wu and Xia estimates match closely over most of the sample, they diverge after the onset of the zero lower bound in January 2009. My estimates indicate that the shadow rate was roughly 2.2 percentage points lower in July 2012 than the Wu and Xia estimates would indicate. I conjecture that the estimates differ because the DF provides a more accurate approximation than the EKF to nonlinearities in the state space when the zero lower bound is in effect. Furthermore, the EKF estimator is in general not consistent, while the DF estimator is.

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

$$X_t = \mu + \rho X_{t-1} + \Sigma v_t \qquad v_t \sim \text{i.i.d. } N(0, I_3) \tag{1.45}$$

$$Y_{n,n+1,t} = \underline{r} + \sigma_n^{\mathbb{Q}} g\left(\frac{a_n + b_n' X_t - \underline{r}}{\sigma_n^{\mathbb{Q}}}\right) + w_t \qquad w_t \sim \text{i.i.d. } N(0, \omega) \tag{1.46}$$

where $Y_{n,n+1,t}$ corresponds to the one-period forward rate at time t for a loan starting at $t+n$ and maturing at $t+n+1$, and X_t is a (3×1) vector of latent factors which explain movements in the yield curve. For a derivation of the expressions for a_n, b_n, and $\sigma_n^{\mathbb{Q}}$, I refer the reader to Wu and Xia 2016.

Using their data on the 3 and 6 month, 1, 2, 5, 7, and 10 year forward rates, one has 7 observation equations, one for each observed yield maturity. They further assume

that each forward rate is observed with normally distributed measurement error with the same variance ω. θ is a (22×1) vector of structural parameters.[19]

Table 1.10 reports maximum likelihood estimates of the parameters from the model with QMLE robust standard errors[20], using the DF with 9 discretization points along each dimension (i.e. $9^3 = 729$ total discretization points). I use the Gospodinov and Lkhagvasuren (2014) method to discretize the VAR state dynamics, which generalizes the method of Rouwenhorst (1995) to VAR(1) systems. I also include the estimates from Wu and Xia (2016) for comparison.

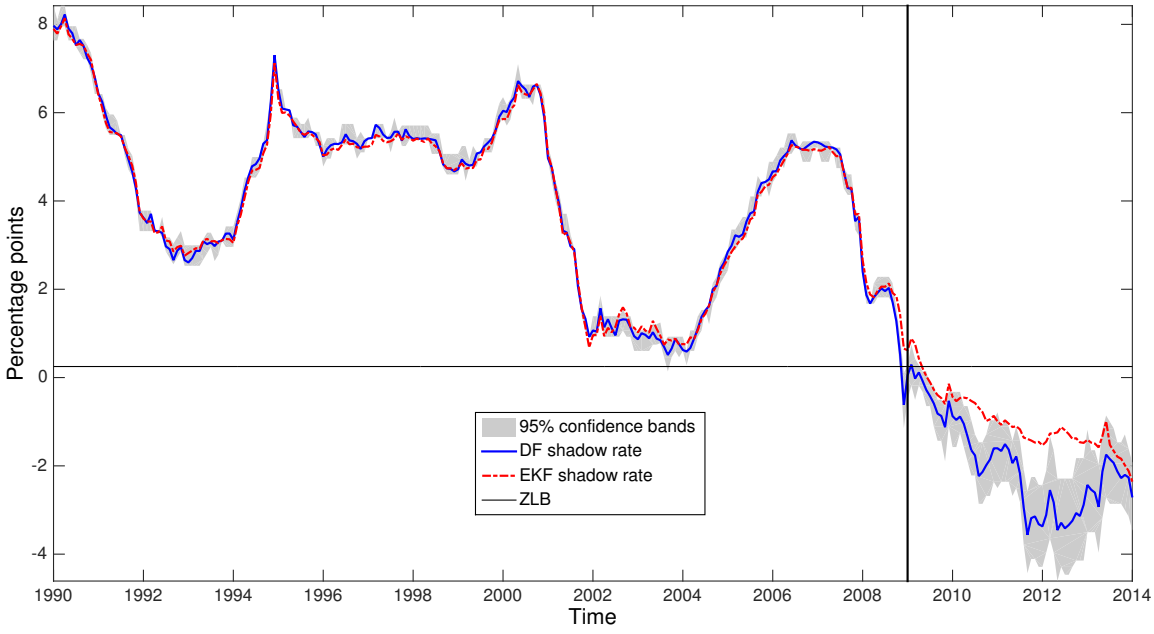


Figure 1.7. Estimated Shadow Rates, Shadow Rate Term Structure Model

Though the parameter estimates obtained using the DF are similar, they produce a drastically different shadow rate series in the zero lower bound period (from about January

[19] It has been pointed out that the results may be sensitive to the arbitrary choice of $\underline{r} = 0.25$, see Bauer and Rudebusch (2016). As a robustness check I also estimate $\underline{r}$ as a free parameter and find that it has little effect on subsequent analysis.

[20] See Hamilton (1994) for details.

Table 1.10. Maximum Likelihood Parameter Estimates (QMLE Standard Errors in Parantheses), Shadow Rate Term Structure Model

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

2009 onward).[21] These differences are illustrated in figure 1.7. I include 95% standard error bands for the shadow rate series estimated with the DF (where the randomness is coming from uncertainty about the state, not the parameters). This emphasizes the fact that the method used to estimate a nonlinear dynamic model can have important economic implications.

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

1.9 Conclusion

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

I apply results from the statistics literature on uniformly ergodic Markov chains to establish that the maximum likelihood estimator implied by the discretization filter is strongly consistent, asymptotically normal, and asymptotically efficient. I demonstrate through simulations that the discretization filter is orders of magnitude faster than alternative nonlinear techniques for the same level of approximation error and I provide practical guidelines for applied researchers.

I demonstrate that the filtering method used to estimate nonlinear models has sizeable effects on the accuracy of the estimated parameters and the accuracy of the filtered states. I show that these estimation differences translate into quantitatively significant economic differences using the Wu and Xia (2016) shadow rate model as an example. My findings have important implications for policy makers who use the Wu and Xia shadow rate as an input to determining the effectiveness of unconventional monetary policy. My estimation procedure leads one to conclude that the shadow rate was 2.2 percentage points lower in July 2012 than the estimates from their paper would indicate.

Additionally, I provide the first estimates of structural parameters in the Gabaix (2012) model of variable rare disasters. I show that the estimated model fails to identify the Great Recession as a disaster episode. This is due to the model's need to have a

positive expected jump in inflation in the event of a disaster in order to capture an upward sloping nominal yield curve. Furthermore, I show that model fails to capture the sharp change in dynamics exhibited by the price-dividend ratio starting in the 1990s.

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

1.10 Acknowledgements

Chapter 1, in full, is currently being prepared for submission for publication of the material. Farmer, Leland E. The dissertation author was the sole author of this paper.

Chapter 2

Discretizing Nonlinear, Non-Gaussian Markov Processes with Exact Conditional Moments

2.1 Introduction

Many nonlinear dynamic economic models such as dynamic stochastic general equilibrium (DSGE) models, asset pricing models, or optimal portfolio problems imply a set of integral equations that do not admit explicit solutions. Finite-state Markov chain approximations of stochastic processes are a useful way of reducing computational complexity when solving and estimating such models because integration is replaced by summation.[1] However, existing methods only work on a limited case by case basis, and apply mostly to linear Gaussian autoregressive processes.

In this paper, we provide a new method for accurately discretizing general nonlinear, non-Gaussian Markov processes. The dynamics of any Markov process are characterized by its transition kernel, which summarizes the conditional distribution of

[1] Examples include heterogeneous-agent incomplete markets models (Aiyagari 1994; Heaton and Lucas 1996), optimal taxation (Aiyagari 1995; Dávila, Hong, Krusell, and Ríos-Rull 2012), portfolio problems (Haliassos and Michaelides 2003; Judd, Kubler, and Schmedders 2011), asset pricing (Zhang 2005; Guvenen 2009), DSGE models (Aruoba, Fernández-Villaverde, and Rubio-Ramírez 2006; Caldara, Fernández-Villaverde, Rubio-Ramírez, and Yao 2012), estimating dynamic games (Aguirregabiria and Mira 2007), inflation dynamics and monetary policy (Vavra 2014), among many others.

the subsequent state for all possible current states. We construct a discrete approximation to the underlying Markov process by approximating a finite set of its conditional distributions.[2] Given a set of discrete points in the state space, we construct a transition matrix, where each row corresponds to a discrete probability measure which mimics the dynamics of the continuous process in that particular state. This is accomplished by starting from a coarse approximation of the underlying process and modifying the transition probabilities so as to exactly match a set of conditional moments, such as the mean and variance. Because there are typically more grid points than there are conditional moments of interest, there are infinitely many candidates for the approximate conditional distribution. To deal with this underdetermined system, we obtain the discrete approximation by minimizing the relative entropy (Kullback-Leibler information) of the conditional distribution from an initial approximation, subject to the given moment constraints. Although this primal problem is a high dimensional constrained optimization problem, its dual is a computationally tractable, *low dimensional unconstrained* optimization problem. We provide recommendations for how to choose the initial approximation and the moments to match.

The two ingredients of our method—matching conditional moments to approximate a Markov process and using the maximum entropy principle to match moments—have already been proposed separately in the literature. Our main contribution is that we combine these two ingredients and show that this idea can be used to discretize a wide variety of nonlinear, non-Gaussian Markov processes, for which there is currently no systematic way of discretizing. Furthermore, we provide sufficient conditions for the existence of a discretization with exact moments and study economic applications to which existing methods do not apply.

[2]For the remainder of the paper, "discrete" should be understood to refer to the state space of the Markov process. Time is always discrete.

The closest papers to ours are Tanaka and Toda 2013; Tanaka and Toda 2015 and Gospodinov and Lkhagvasuren 2014. Tanaka and Toda 2013 construct discrete approximations of continuous probability distributions (as opposed to stochastic processes) by modifying an initial discretization so as to exactly match low order moments using the maximum entropy principle. While they briefly discuss how to apply their method to discretize vector autoregressive processes (VARs), because they need a closed-form expression for the ergodic distribution—which is not available in most situations—their method cannot be directly used for discretizing general Markov processes. Tanaka and Toda 2015 prove that their approximation method weakly converges to the true distribution as the number of grid points tends to infinity. They also show that the integration error diminishes by a factor proportional to the error when the integrand is approximated using the functions defining the moments of interest as basis functions. Therefore, the approximation quality of the Tanaka-Toda method depends on two factors, (i) the quality of the initial discretization, and (ii) how well the moment defining functions approximate the integrand.

Gospodinov and Lkhagvasuren 2014 (henceforth GL) propose a discretization method of VARs that targets the first and second conditional moments. According to their numerical results, the GL method seems to be the most accurate finite-state Markov chain approximation for VARs currently available in the literature. As in GL, we target the conditional moments in order to discretize VARs. However, our method improves upon theirs in three important ways.

First, unlike the GL method, our approach is not limited to the approximation of VARs. It applies to *any* Markov process for which we can compute conditional moments and thus has a much wider range of applicability. For instance, we can discretize stochastic processes with interesting nonlinear and non-Gaussian conditional dynamics. Additionally, we do not require a parametric specification of the Markov process to

use our approach. Given sufficient data, we can estimate the conditional moments and transition kernel nonparametrically, and use these to construct our discrete approximation.

Second, GL adjust the transition probabilities to match moments directly, whereas we solve the dual problem, which is a low dimensional unconstrained convex minimization problem. The gradient and Hessian of the objective function can be computed in closed form, which allows us to use a standard Newton-type algorithm to find the minimum. Consequently, our method is computationally tractable even when the number of grid points is large. This is an important property, particularly for the case of high dimensional processes.

Finally, for general VARs (which may even feature stochastic volatility), under certain regularity conditions we prove that our method matches all k-step ahead conditional mean, variance, and covariance as well as the unconditional ones. This property has been known only for the Rouwenhorst 1995 method for discretizing univariate AR(1) processes. We further discuss the relation of our method to the existing literature in Section 2.3.3.

In order to illustrate the general applicability of our method, we solve for the price-dividend ratio in Lucas-tree asset pricing models, under different assumptions about the stochastic processes driving consumption and dividend growth, including more standard AR(1) and VAR(1) processes with Gaussian shocks, an AR(1) model with non-Gaussian shocks, and the variable rare disasters model of Gabaix 2012, whose underlying stochastic process is highly nonlinear and non-Gaussian. In each case, we show that our method produces more accurate solutions than all existing discretization methods,[3] often by

[3]Several papers such as Aruoba, Fernández-Villaverde, and Rubio-Ramírez 2006 and Caldara, Fernández-Villaverde, Rubio-Ramírez, and Yao 2012 compare the accuracy of various solution techniques (log-linearization, value function iteration, perturbation, projection, etc.), *given the discretization method*. To the best of our knowledge, Kopecky and Suen 2010 is the only paper that compares the solution accuracy across various discretization methods, fixing the solution technique. However, they consider only Gaussian AR(1) processes.

several orders of magnitude, requiring only minor modifications between specifications and trivial computing time. We also show that solving general asset pricing models (*e.g.*, with recursive utility and complicated dynamics) using discretization and projection (Judd 1992) is actually equivalent to solving a discrete-state model (which is a matter of inverting a matrix) and interpolating. Therefore our method provides a simple but systematic way for solving asset pricing models.

We emphasize that our method has many potential applications beyond the asset pricing models considered here. For example, our method can be used to facilitate the estimation of nonlinear state space models. In parallel work, Farmer 2017 shows that by discretizing the dynamics of the state variables, one can construct an approximate state space model with closed-form expressions for the likelihood and filtering recursions, as in Hamilton 1989. The parameters of the model can then be estimated using standard likelihood or Bayesian techniques. This procedure offers an alternative to computationally expensive, simulation-based methods like the particle filter, and simple but often inaccurate linearization approaches like the extended Kalman filter. Our paper provides a computationally tractable method for discretizing general nonlinear Markov processes governing the state dynamics.

2.2 Maximum Entropy Method for Discretizing Markov Processes

In this section we review the maximum entropy method for discretizing probability distributions proposed by Tanaka and Toda 2013; Tanaka and Toda 2015 and apply it to discretize general Markov processes.

2.2.1 Discretizing Probability Distributions

Description of Method

Suppose that we are given a continuous probability density function $f : \mathbb{R}^K \to \mathbb{R}$, which we want to discretize. Let X be a random vector with density f, and $g : \mathbb{R}^K \to \mathbb{R}$ be any bounded continuous function. The first step is to pick a quadrature formula

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^K} g(x)f(x)\,\mathrm{d}x \approx \sum_{n=1}^{N} w_n g(x_n) f(x_n), \tag{2.1}$$

where N is the number of integration points, $\{x_n\}_{n=1}^N$, and $w_n > 0$ is the weight on the integration point x_n.[4] Let $D_N = \{x_n \mid n = 1,\ldots,N\}$ be the set of grid points. For example, if we let

$$D_N = \{(m_1 h,\ldots,m_K h) \mid m_1,\ldots,m_K = 0,\pm 1,\ldots,\pm M\},$$

which consists of $N = (2M+1)^K$ lattice points with grid size h, setting the weight $w_n = h^K$ in quadrature formula (2.1) gives the trapezoidal formula.

For now, we do not take a stance on the choice of the initial quadrature formula, but take it as given. Given the quadrature formula (2.1), a coarse but valid discrete approximation of the density f would be to assign probability q_n to the point x_n proportional to $w_n f(x_n)$, so

$$q_n = \frac{w_n f(x_n)}{\sum_{n=1}^{N} w_n f(x_n)}. \tag{2.2}$$

However, this is not necessarily a good approximation because the moments of the discrete distribution $\{q_n\}$ do not generally match those of f.

Tanaka and Toda 2013 propose exactly matching a finite set of moments by

[4] Since the grid points $\{x_n\}$ and weights $\{w_n\}$ may depend on the number of grid points N, a more precise notation might be $x_{n,N}$ and $w_{n,N}$. Since there is no risk of confusion, we keep the simpler notation x_n and w_n.

updating the probabilities $\{q_n\}$ in a particular way. Let $T : \mathbb{R}^K \to \mathbb{R}^L$ be a function that defines the moments that we wish to match and let $\overline{T} = \int_{\mathbb{R}^K} T(x) f(x)\,\mathrm{d}x$ be the vector of exact moments. For example, if we want to match the first and second moments in the one dimensional case ($K = 1$), then $T(x) = (x, x^2)'$. Tanaka and Toda 2013 update the probabilities $\{q_n\}$ by solving the optimization problem

$$\begin{aligned} \underset{\{p_n\}}{\text{minimize}} \quad & \sum_{n=1}^{N} p_n \log \frac{p_n}{q_n} \\ \text{subject to} \quad & \sum_{n=1}^{N} p_n T(x_n) = \overline{T},\ \sum_{n=1}^{N} p_n = 1,\ p_n \geq 0. \end{aligned} \tag{P}$$

The objective function in the primal problem (P) is the Kullback and Leibler 1951 information of $\{p_n\}$ relative to $\{q_n\}$, which is also known as the relative entropy. This method matches the given moments exactly while keeping the probabilities $\{p_n\}$ as close to the initial approximation $\{q_n\}$ as possible in the sense of the Kullback-Leibler information.[5] Note that since (P) is a convex minimization problem, the solution (if one exists) is unique.

The optimization problem (P) is a constrained minimization problem with a large number (N) of unknowns ($\{p_n\}$) with $L+1$ equality constraints and N inequality constraints, which is in general computationally intensive to solve. However, it is well-known that entropy-like minimization problems are computationally tractable by using duality theory (Borwein and Lewis 1991). Tanaka and Toda 2013 convert the primal

[5]The Kullback-Leibler information is not the only possible loss function. One may also use other criteria such as the L^2 norm or other generalized entropies. However, the Kullback-Leibler information has the unmatched feature that (i) the domain of the dual function is the entire space, so the dual problem becomes unconstrained, and (ii) the constraint $p_n \geq 0$ never binds, so the dual problem becomes low dimensional. See Borwein and Lewis 1991 for more details on duality in entropy-like minimization problems and Owen 2001, Tsao 2004, Kitamura 2007, and Tsao and Wu 2013 for discussions on the computational aspects of empirical likelihood methods, which is mathematically related.

problem (P) to the dual problem

$$\max_{\lambda \in \mathbb{R}^L} \left[\lambda' \overline{T} - \log \left(\sum_{n=1}^{N} q_n \mathrm{e}^{\lambda' T(x_n)} \right) \right], \tag{D}$$

which is a *low dimensional* (L unknowns) *unconstrained* concave maximization problem and hence computationally tractable. The following theorem shows how the solutions to the two problems (P) and (D) are related. Below, the symbols "int" and "co" denote the interior and the convex hull of sets.

. 1. The primal problem (P) has a solution if and only if $\overline{T} \in \operatorname{co} T(D_N)$. If a solution exists, it is unique.

2. The dual problem (D) has a solution if and only if $\overline{T} \in \operatorname{int} \operatorname{co} T(D_N)$. If a solution exists, it is unique.

3. If the dual problem (D) has a (unique) solution λ_N, then the (unique) solution to the primal problem (P) is given by

$$p_n = \frac{q_n \mathrm{e}^{\lambda_N' T(x_n)}}{\sum_{n=1}^{N} q_n \mathrm{e}^{\lambda_N' T(x_n)}} = \frac{q_n \mathrm{e}^{\lambda_N' (T(x_n) - \overline{T})}}{\sum_{n=1}^{N} q_n \mathrm{e}^{\lambda_N' (T(x_n) - \overline{T})}}. \tag{2.3}$$

Practical Implementation

Theorem 4 provides a practical way to implement the Tanaka-Toda method. After choosing the initial discretization $Q = \{ q_n \}$ and the moment defining function T, one can numerically solve the unconstrained optimization problem (D). To this end, we can instead solve

$$\min_{\lambda \in \mathbb{R}^L} \sum_{n=1}^{N} q_n \mathrm{e}^{\lambda' (T(x_n) - \overline{T})} \tag{D$'$}$$

because the objective function in (D$'$) is a monotonic transformation (-1 times the exponential) of that in (D). Since (D$'$) is an unconstrained convex minimization problem

with a (relatively) small number (L) of unknowns (λ), solving it is computationally simple. Letting $J_N(\lambda)$ be the objective function in (D′), its gradient and Hessian can be analytically computed as

$$\nabla J_N(\lambda) = \sum_{n=1}^{N} q_n \mathrm{e}^{\lambda'(T(x_n)-\overline{T})}(T(x_n)-\overline{T}), \tag{2.4a}$$

$$\nabla^2 J_N(\lambda) = \sum_{n=1}^{N} q_n \mathrm{e}^{\lambda'(T(x_n)-\overline{T})}(T(x_n)-\overline{T})(T(x_n)-\overline{T})', \tag{2.4b}$$

respectively. In practice, we can quickly solve (D′) numerically using optimization routines by supplying the analytical gradient and Hessian.[6]

If a solution to (D′) exists, it is unique, and we can compute the updated discretization $P = \{ p_n \}$ by (2.3). If a solution does not exist, it means that the regularity condition $\overline{T} \in \operatorname{int}\operatorname{co} T(D_N)$ does not hold and we cannot match moments. Then one needs to select a smaller set of moments. Numerically checking whether moments are matched is straightforward: by (2.3), (D′), and (2.4a), the error is

$$\sum_{n=1}^{N} p_n T(x_n) - \overline{T} = \frac{\sum_{n=1}^{N} q_n \mathrm{e}^{\lambda_N'(T(x_n)-\overline{T})}(T(x_n)-\overline{T})}{\sum_{n=1}^{N} q_n \mathrm{e}^{\lambda_N'(T(x_n)-\overline{T})}} = \frac{\nabla J_N(\lambda_N)}{J_N(\lambda_N)}. \tag{2.5}$$

Error Estimate and Convergence

Tanaka and Toda 2015 prove that whenever the quadrature approximation (2.1) converges to the true value as the number of grid points N tends to infinity, the discrete distribution $\{ p_n \}$ in (2.3) also weakly converges to the true distribution f and improves the integration error as follows. Let g be the integrand in (2.1) and consider approximating

[6] Since the dual problem (D) is a concave maximization problem, one may also solve it directly. However, according to our experience, solving (D′) is numerically more stable. This is because the objective function in (D) is close to linear when $\|\lambda\|$ is large, so the Hessian is close to singular and not well-behaved. On the other hand, since the objective function in (D′) is the sum of exponential functions, it is well-behaved.

g using $T = (T_1, \ldots, T_L)$ as basis functions:

$$g(x) \approx \widehat{g}_T(x) = \sum_{l=1}^{L} b_l T_l(x),$$

where $\{ b_l \}_{l=1}^{L}$ are coefficients. Let $r_{g,T} = \frac{g - \widehat{g}_T}{\|g - \widehat{g}_T\|_\infty}$ be the normalized remainder term, where $\|\cdot\|_\infty$ denotes the supremum norm. Letting

$$E_{g,N}^{(Q)} = \left| \int_{\mathbb{R}^K} g(x) f(x) \, \mathrm{d}x - \sum_{n=1}^{N} q_n g(x_n) \right|$$

be the integration error under the initial discretization $Q = \{ q_n \}$ and $E_{g,N}^{(P)}$ be the error under $P = \{ p_n \}$, Tanaka and Toda 2015 prove the error estimate

$$E_{g,N}^{(P)} \leq \|g - \widehat{g}_T\|_\infty \left(E_{r_{g,T},N}^{(Q)} + \frac{2}{\sqrt{C}} E_{T,N}^{(Q)} \right), \tag{2.6}$$

where C is a constant explicitly given in the paper. Equation (2.6) says that the integration error improves by the factor $\|g - \widehat{g}_T\|_\infty$, which is the approximation error of the integrand g by the basis functions $\{ T_l \}_{l=1}^{L}$ that define the targeted moments. It is clear from (2.6) that the approximation quality of the Tanaka-Toda method depends on two factors, (i) the quality of the initial discretization (how small $E_{g,N}^{(Q)}$ is), and (ii) how well the moment defining functions approximate the integrand (how small $\|g - \widehat{g}_T\|_\infty$ is).

2.2.2 Discretizing General Markov Processes

Next we show how to extend the Tanaka-Toda method to the case of time-homogeneous Markov processes.

Description of method

Consider the time-homogeneous first-order Markov process

$$P(x_t \leq x' | x_{t-1} = x) = F(x', x),$$

where x_t is the vector of state variables and $F(\cdot, x)$ is a cumulative distribution function (CDF) that determines the distribution of $x_t = x'$ given $x_{t-1} = x$. The dynamics of any Markov process are completely characterized by its Markov transition kernel. In the case of a discrete state space, this transition kernel is simply a matrix of transition probabilities, where each row corresponds to a conditional distribution. We can discretize the continuous process x by applying the Tanaka-Toda method to each conditional distribution separately.

More concretely, suppose that we have a set of grid points $D_N = \{x_n\}_{n=1}^N$ and an initial coarse approximation $Q = (q_{nn'})$, which is an $N \times N$ probability transition matrix. Suppose we want to match some conditional moments of x, represented by the moment defining function $T(x)$. The exact conditional moments when the current state is $x_{t-1} = x_n$ are

$$\overline{T}_n = \mathbb{E}[T(x_t) \mid x_n] = \int T(x) \, \mathrm{d}F(x, x_n),$$

where the integral is over x, fixing x_n. (If these moments do not have explicit expressions, we can use highly accurate quadrature formulas to compute them.) By Theorem 4, we can match these moments exactly by solving the optimization problem

$$\begin{aligned} \underset{\{p_{nn'}\}_{n'=1}^N}{\text{minimize}} \quad & \sum_{n'=1}^N p_{nn'} \log \frac{p_{nn'}}{q_{nn'}} \\ \text{subject to} \quad & \sum_{n'=1}^N p_{nn'} T(x_{n'}) = \overline{T}_n, \ \sum_{n'=1}^N p_{nn'} = 1, \ p_{nn'} \geq 0 \end{aligned} \tag{P_n}$$

for each $n = 1, 2, \ldots, N$, or equivalently the dual problem

$$\min_{\lambda \in \mathbb{R}^L} \sum_{n'=1}^{N} q_{nn'} e^{\lambda'(T(x_{n'}) - \overline{T}_n)}. \tag{D$'_n$}$$

(D$'_n$) has a unique solution if and only if the regularity condition

$$\overline{T}_n \in \operatorname{int} \operatorname{co} T(D_N) \tag{2.7}$$

holds. We summarize our procedure in Algorithm 2 below.

Algorithm 2: Discretization of Markov Processes

1. Select a discrete set of points $D_N = \{x_n\}_{n=1}^N$ and an initial approximation $Q = (q_{nn'})$.
2. Select a moment defining function $T(x)$ and corresponding exact conditional moments $\{\overline{T}_n\}_{n=1}^N$. If necessary, approximate the exact conditional moments with a highly accurate numerical integral.
3. For each $n = 1, \ldots, N$, solve minimization problem (D$'_n$) for λ_n. Check whether moments are matched using formula (2.5), and if not, select a smaller set of moments. Compute the conditional probabilities corresponding to row n of $P = (p_{nn'})$ using (2.3).

The resulting discretization of the process is given by the transition probability matrix $P = (p_{nn'})$. Since the dual problem (D$'_n$) is an unconstrained convex minimization problem with a typically small number of variables, standard Newton type algorithms can be applied. Furthermore, since the probabilities (2.3) are strictly positive by construction, the transition probability matrix $P = (p_{nn'})$ is a strictly positive matrix, so the resulting Markov chain is stationary and ergodic.

The Regularity Condition

How stringent is the regularity condition (2.7)? Note that $\operatorname{co} T(D_N)$ is the convex hull of the image of the grid D_N under the moment defining function T, so any element

of $\operatorname{co} T(D_N)$ has the form $\sum_n \alpha_n T(x_n)$, where $\alpha_n \geq 0$, $\sum_n \alpha_n = 1$, and $x_n \in D_N$. Also, by definition $\overline{T}_n = \mathbb{E}[T(x_t) \mid x_{t-1} = x_n]$, which is a weighted average of $T(x)$'s. Therefore in practice it is not hard to meet the regularity condition $\overline{T}_n \in \operatorname{int} \operatorname{co} T(D_N)$. The only case difficulty arises is when x_n is close to the boundary of (the convex hull of) D_N and the stochastic process is highly persistent. Then $\overline{T}_n$ also tends to be close to the boundary of $\operatorname{co} T(D_N)$, and it may happen to be outside the set, violating (2.7). But since the boundary of a convex set has measure zero, for the vast majority of the grid points we are able to match moments exactly. A practical solution to the potential failure of the regularity condition is thus to match moments whenever we can by solving the minimization problem (D'_n), and if a solution fails to exist (which can be checked by computing the error (2.5)), we can match only a subset of the moments $T = (T_1, \ldots, T_L)$.

How to Choose the Grid

In order to implement our method in practice, we need to overcome two issues: (i) the choice of the grid, and (ii) the choice of the targeted moments.

According to the convergence analysis in Tanaka and Toda 2015, the grid D_N should be chosen as the integration points of the quadrature formula (2.1), which is used to obtain the initial coarse approximation in (2.2). For simplicity we often choose the trapezoidal formula and therefore even-spaced grids. Alternatively, we can place points using the Gaussian quadrature nodes as in Tauchen and Hussey 1991, or, for that matter, any quadrature formula with positive weights such as Simpson's rule, low-degree Newton-Cotes type formulas, or the Clenshaw-Curtis quadrature (see Davis and Rabinowitz 1984 for quadrature formulas); or quantiles as in Adda and Cooper 2003.

Although tensor grids work well in low dimensional problems, in higher dimensions they are not computationally tractable because the number of grid points increases

exponentially with the dimension.[7] In such cases, one needs to use sparse grids (Krueger and Kubler 2004; Heiss and Winschel 2008) or select the grid points to delimit sets that the process visits with high probability (Maliar and Maliar 2015).

In practice, we find that the even-spaced grid (trapezoidal formula) works very well and is robust across a wide range of different specifications. However, if there is some special structure to the conditional distribution, such as normality, a Gaussian quadrature approximation can result in better solution accuracy for dynamic models.

How to Choose the Moments to Match

Our method approximates a continuous Markov process by a discrete transition matrix. A good approximation is one for which the integral of any bounded continuous function using the discrete measure is close to the integral using the original continuous measure. The quality of this approximation depends on how accurately the integrand can be approximated by the moment defining functions (see $\|g - \widehat{g}_T\|_\infty$ in (2.6)).

In the case of a single probability distribution, we can choose a grid over a set with high probability and therefore match as many moments as we wish, up to 1 fewer than the number of grid points. In the case of stochastic processes, the situation is more restrictive. As an illustration, consider the AR(1) process

$$x_t = \rho x_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0,1),$$

with ρ close to 1.

Let $D_N = \{\bar{x}_1, \ldots, \bar{x}_N\}$ be the grid, with $\bar{x}_1 < \cdots < \bar{x}_N$. When $x_{t-1} = \bar{x}_N$, the conditional distribution of x_t is $N(\rho\bar{x}_N, 1)$. But when ρ is close to 1, this (true) distribution

[7] Note that with our method, having a large number of grid points is not an issue for solving the dual problem (D'_n). The number of unknowns is equal to the number of targeted moments, which is fixed. The issue with tensor grids is that the *number of dual problems* we need to solve grows exponentially with the dimension.

has nearly 1/2 of its probability mass on the interval $(\bar{x}_N, \infty)$, which lies outside the grid. Since there is such a discrepancy between the location of the grid points and the probability mass, we do not have the flexibility to match many moments, because the regularity condition $\overline{T}_n \in \text{int}\,\text{co}\,T(D_N)$ may fail to hold near the boundary. In the examples below, we consider matching up to 4 conditional moments whenever we can.

2.3 Discretizing VAR(1)s and Stochastic Volatility Models

Applied researchers often specify vector autoregressive processes (VARs) to describe the underlying shocks in their models. In this section we explain how our method can be used to discretize general VARs and stochastic volatility models, and prove some theoretical properties.

2.3.1 VAR(1)

Suppose we want to discretize a VAR(1) process

$$x_t = (I - B)\mu + Bx_{t-1} + \eta_t, \quad \eta_t \sim N(0, \Psi), \tag{2.8}$$

where all vectors are in $\mathbb{R}^K$, μ is the unconditional mean of x_t, Ψ is the conditional variance matrix, and B is a $K \times K$ matrix with all eigenvalues smaller than 1 in absolute value in order to guarantee stationarity. Using the Cholesky decomposition, without loss of generality, we can rewrite (2.8) as

$$y_t = Ay_{t-1} + \varepsilon_t, \tag{2.9}$$

where $y_t = C^{-1}(x_t - \mu)$, $A = C^{-1}BC$, $\varepsilon_t = C^{-1}\eta_t \sim N(0,D)$, C is lower triangular, D is diagonal (typically $D = I$), and $\Psi = CDC'$.[8] Once we have a discretization for y_t, we have one for $x_t = \mu + Cy_t$.

Description of Method

First we introduce some additional notation. Let $y_t = (y_{1t}, \ldots, y_{Kt})$ and assume that the discrete approximation of y_{kt} takes N_k values denoted by $D_{k,N_k} = \{\bar{y}_{kn}\}_{n=1}^{N_k}$. In total, there are $J = N_1 \times \cdots \times N_K$ states.[9] Let $j = 1, \ldots, J$ be an index of the state, corresponding to a particular combination of points $(\bar{y}_{1n}(j), \ldots, \bar{y}_{Kn}(j))$. Let $p_{kn}(j)$ be the probability that $y_{kt} = \bar{y}_{kn}$ conditional on being in state j. Define the conditional mean and variance of y_{kt} given state j as $\mu_k(j)$ and $\sigma_k(j)^2$, respectively. We outline the procedure in Algorithm 3. (Although we describe it for the case of two conditional moments, the case with higher order moments is similar.)

In order to determine $\{p_{kn}(j)\}$ using Algorithm 3, we need an initial coarse approximation $\{q_{kn}(j)\}$. The simplest way is to take the grid points $\{\bar{y}_{kn}\}_{n=1}^{N_k}$ to be evenly spaced and assign $q_{kn}(j)$ to be proportional to the conditional density of y_{kt} given state j, which corresponds to choosing the trapezoidal rule for the initial quadrature formula. Alternatively, we can use the nodes and weights of the Gauss-Hermite quadra-

[8] Clearly there are infinitely many such decompositions. Experience tells that the quality of discretization is best when each compoment of the y_t process in (2.9) has the same unconditional variance. We can do as follows to construct such a decomposition. First, take $\tilde{C}$ such that $\Psi = \tilde{C}\tilde{C}'$, so $D = I$. Define $\tilde{y}_t = \tilde{C}^{-1}(x_t - \mu)$, $\tilde{A} = \tilde{C}^{-1}B\tilde{C}$, and $\tilde{\varepsilon}_t = \tilde{C}^{-1}\eta_t \sim N(0,I)$. Let $\tilde{\Sigma}$ be the unconditional variance of the $\tilde{y}$ process. Let $y_t = U'\tilde{y}_t$ for some orthogonal matrix U, and define $A = U'\tilde{A}U$, $\varepsilon_t = U'\tilde{\varepsilon}_t$, and $C = \tilde{C}U'$. Then $\mathrm{Var}[\varepsilon_t] = U'IU = I$. The unconditional variance of the y process is then $\Sigma = U'\tilde{\Sigma}U$. Since $\mathrm{tr}\,\Sigma = \mathrm{tr}\,\tilde{\Sigma}$, the diagonal elements of Σ become equal if $\Sigma_{kk} = (U'\tilde{\Sigma}U)_{kk} = \frac{1}{K}\mathrm{tr}\,\tilde{\Sigma}$. We can make this equation (approximately) true by solving the optimization problem

$$\begin{aligned} &\text{minimize} && \sum_{k=1}^{K}\left((U'\tilde{\Sigma}U)_{kk} - \frac{1}{K}\mathrm{tr}\,\tilde{\Sigma}\right)^2 \\ &\text{subject to} && U'U = I. \end{aligned}$$

With this choice of U, the unconditional variances of the compoments of $\{y_t\}$ are close to each other, and in fact equal if the objective function takes the value zero.

[9] In practice, we take $N_1 = N_2 = \cdots = N_K = N$, so $J = N^K$.

Algorithm 3: Discretization of VAR(1) Processes

1 For each component of $y_t = (y_{1t}, \ldots, y_{Kt})$, select a discrete set of points $D_{k,N_k} = \{\bar{y}_{kn}\}_{n=1}^{N_k}$.
2 For $j = 1, \ldots, J$,
3 For $k = 1, \ldots, K$ (note that we can treat each component k separately because the variance-covariance matrix D is diagonal),
4 Define the moment defining function and exact moments by

$$T_{kj}(x) = \begin{bmatrix} x \\ (x - \mu_k(j))^2 \end{bmatrix} \quad \text{and} \quad \overline{T}_{kj} = \begin{bmatrix} \mu_k(j) \\ \sigma_k(j)^2 \end{bmatrix}.$$

5 Select an initial approximation $\{q_{kn}(j)\}_{n=1}^{N_k}$, where $q_{kn}(j)$ is the probability of moving to point $\bar{y}_{kn}$ conditional on being in state j.
6 Solve minimization problem (D'_n) for λ_{kj} and compute the conditional probabilities $\{p_{kn}(j)\}_{n=1}^{N_k}$ using (2.3).
7 Compute the conditional probabilities $\{p_{jj'}\}_{j'=1}^{J}$ by multiplying together the conditional probabilities $p_{kn}(j)$ that make up transitions to elements of state j'.
8 Collect the conditional probabilities $\{p_{jj'}\}_{j'=1}^{J}$ into a matrix $P = (p_{jj'})$.

ture as in Tauchen and Hussey 1991,[10] or take the grid points $\{\bar{y}_{kn}\}_{n=1}^{N_k}$ as quantiles of the unconditional distribution and assign probabilities according to the cumulative distribution function, as in Adda and Cooper 2003.[11] Which grid/quadrature formula is best is a practical problem and we explore this issue in subsequent sections.

This method can be generalized to VAR(p) processes, although the dimension of the state space would grow exponentially in p unless we use a sparse grid.

[10] Following the original paper by Tauchen and Hussey 1991, we always use the conditional variance matrix D to construct the Gauss-Hermite quadrature. This is the most logical way since dynamic economic models involve conditional expectations (*e.g.*, Euler equations), which are integrals that use the conditional distributions.

[11] The specific procedure is as follows. Let the stationary distribution of y_{kt} be $N(0, \sigma_k^2)$. Since there are N_k discrete points for y_{kt}, we divide the real line $\mathbb{R}$ into N_k intervals using the n-th N_k-quantile ($n = 1, \ldots, N_k - 1$), which we denote by $I_{k1}, \ldots, I_{kN}$. The discrete points are then the median of each interval, so $\bar{y}_{kn} = F^{-1}((2n-1)/2N_k)$ ($n = 1, 2, \ldots, N_k$), where F is the CDF of $N(0, \sigma_k^2)$. When the $t-1$ state is j, since the conditional distribution of y_{kt} is $N(\mu_k(j), \sigma_k^2(j))$, we assign initial probability $q_{kn}(j) = P(I_{kn})$ to the point $\bar{y}_{kn}$ under the conditional distribution $N(\mu_k(j), \sigma_k^2(j))$.

Theoretical Properties of the Discretization

If a solution to the dual problem (D'_n) exists, by construction our method generates a finite-state Markov chain approximation of the VAR with exact 1-step ahead conditional moments. But how about k-step ahead conditional moments and unconditional moments? The following theorem provides an answer.

. Consider the VAR(1) process in (2.9), with grid D_N. Suppose that the regularity condition $\overline{T}_n \in \operatorname{int}\operatorname{co} T(D_N)$ holds, and hence our method matches the conditional mean and variance. Then the method also matches any k-step ahead conditional mean and variance, as well as the unconditional mean and all autocovariances (hence spectrum).

This result holds even for a certain class of stochastic volatility models (Theorem 2.A.1). According to its proof, there is nothing specific to the choice of the grid, the normality of the process, or the diagonalization. Therefore the result holds for any non-Gaussian linear process.

So far, we have assumed that the regularity condition (2.7) holds, so that a discrete approximation with exact conditional moments using our method exists. As we see in the numerical examples below, such a discretization exists most of the time, but not always. Therefore it is important to provide easily verifiable conditions that guarantee existence. For general VARs, the following proposition shows that it is always possible to match conditional means.

Proposition 1. *Consider the VAR(1) process in* (2.9) *with coefficient matrix* $A = (a_{kk'})$. *Let* $|A| = (|a_{kk'}|)$ *be the matrix obtained by taking the absolute value of each element of A. If the spectral radius of* $|A|$ *is less than 1 (*i.e.*, all eigenvalues are less than 1 in absolute value), then there exists a tensor grid such that we can match all conditional means.*

How about the conditional mean and variance? Since addressing this issue for general VAR processes is challenging, we restrict our analysis to the case of an AR(1) process. The following proposition shows that a solution exists if the grid is symmetric, sufficiently fine, and the grid points span more than one unconditional standard deviation around 0.

Proposition 2. *Consider the AR(1) process*

$$x_t = \rho x_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim (0,1),$$

where $0 \leq \rho < 1$. *Suppose that (i) the grid* $\{\bar{x}_n\}_{n=1}^N$ *is symmetric and spans more than one unconditional standard deviation around 0, so* $\max_n |\bar{x}_n| > 1/\sqrt{1-\rho^2}$, *and (ii) either the maximum distance between two neighboring grid points is less than 2, or for each positive grid point* $\bar{x}_n > 0$ *there exists a grid point* $\bar{x}_{n'}$ *such that*

$$\rho \bar{x}_n - \frac{1}{(1-\rho)\bar{x}_n} < \bar{x}_{n'} \leq \rho \bar{x}_n. \tag{2.10}$$

Then (D'_n) *has a unique solution for all n.*

When the grid $\{\bar{x}_n\}$ is even-spaced, we can obtain a simple sufficient condition for existence.

Corollary 3. *Let the grid points* $\{\bar{x}_n\}_{n=1}^N$ *be symmetric and even-spaced,* $\sigma = \frac{1}{\sqrt{1-\rho^2}}$ *be the unconditional standard deviation, and* $M = \max_n \bar{x}_n$. *Suppose that either*

1. $\rho \leq 1 - \frac{2}{N-1}$ *and* $\sigma < M \leq \sqrt{2}\sigma\sqrt{N-1}$, *or*
2. $\rho > 1 - \frac{2}{N-1}$ *and* $\sigma < M \leq \sigma\sqrt{N-1}$.

Then (D'_n) *has a unique solution for all n.*

Interestingly, Kopecky and Suen 2010 show that the Rouwenhorst 1995 method matches the first and second conditional moments when the grid span is $M = \sigma\sqrt{N-1}$, the upper bound in Corollary 3 for the case $\rho > 1 - \frac{2}{N-1}$. Choosing a grid span of order $\sqrt{N}$ can also be theoretically justified. In that case, the grid spacing is of order $N/\sqrt{N} = 1/\sqrt{N}$. Since the grid gets finer while the grid span tends to infinity, the trapezoidal formula converges to the true integral. Therefore the approximation error can be made arbitrarily small by increasing N. For general VARs, we do not have theoretical results for the existence of a discretization that matches second moments. However, we recommend using a grid span $M = \sigma\sqrt{N-1}$ in each dimension, where σ is the square root of the smallest eigenvalue of the unconditional variance of the VAR.

Theorem 5, Proposition 2, and Corollary 3 are significant. Note that among all existing methods, the Rouwenhorst 1995 method for discretizing Gaussian AR(1) processes is the only one known to match the first and second conditional moments exactly.[12]

2.3.2 AR(1) with Stochastic Volatility

Consider an AR(1) process with stochastic volatility of the form

$$y_t = \lambda y_{t-1} + u_t, \qquad u_t \sim N(0, \mathrm{e}^{x_t}), \tag{2.11a}$$

$$x_t = (1-\rho)\mu + \rho x_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma^2), \tag{2.11b}$$

where x_t is the unobserved log variance process and y_t is the observable, *e.g.*, stock returns. We assume that y_t is mean zero without loss of generality.

Since the log variance process x_t evolves independently of the level y_t as an AR(1) process, we can discretize it using Algorithm 3. For y_t, note that the unconditional

[12] Kopecky and Suen 2010 prove that the 1-step ahead conditional moments are exact. By Theorem 5, all k-step ahead conditional moments are also exact.

variance is given by

$$\sigma_y^2 = \mathbb{E}[y_t^2] = \frac{\mathbb{E}[e^{x_t}]}{1-\lambda^2}.$$

Since the unconditional distribution of x_t is $N\left(\mu, \frac{\sigma^2}{1-\rho^2}\right)$, we have

$$\mathbb{E}[e^{x_t}] = \exp\left(\mu + \frac{\sigma^2}{2(1-\rho^2)}\right)$$

using the properties of lognormal random variables. We can then construct an even-spaced grid for y_t spanning some number of unconditional standard deviations around 0.

With some more algebra, we can show that

$$y_t|x_{t-1}, y_{t-1} \sim N\left(\lambda y_{t-1}, \exp\left((1-\rho)\mu + \rho x_{t-1} + \sigma^2/2\right)\right).$$

We discretize these conditional distributions for each (x_{t-1}, y_{t-1}) pair using our method and combine them with the discretization obtained for $x_t|x_{t-1}$ above, to come up with a joint transition matrix for the state (x_t, y_t).

2.3.3 Relation to the Existing Literature

In this section we discuss the existing literature in detail.

The standard method for approximating an AR(1) process is that of Tauchen 1986, which divides the state space into evenly spaced intervals, with the grid chosen as the midpoints of those intervals. Tauchen constructs each approximate conditional distribution by matching the probabilities of transitioning from a particular point to each interval. The Tauchen method is intuitive, simple, and reasonably accurate when the number of grid points is large enough. It is easily generalized and widely used for the approximation of VAR processes. Variants of the Tauchen method have been developed

in the literature by using Gauss-Hermite quadrature (Tauchen and Hussey 1991), placing grid points using quantiles instead of even-spaced intervals (Adda and Cooper 2003), and using multivariate normal integration techniques (Terry and Knotek 2011). Rouwenhorst 1995 proposes an alternative discretization method of a Gaussian AR(1) process that matches the unconditional first and second moments exactly. His idea is to approximate a normal distribution by binomial distributions.

VARs are highly persistent in typical macroeconomic applications. It has been recognized that the Tauchen and Tauchen-Hussey methods often fail to give accurate approximations to such processes (Zhang 2005; Flodén 2008),[13] which has spurred a renewed research interest in accurately discretizing autoregressive processes. Kopecky and Suen 2010 prove that for a certain choice of the grid, the Rouwenhorst method actually matches the autocorrelation and the *conditional* mean and variance. This means that the Rouwenhorst method is suitable for discretizing highly persistent Gaussian AR(1) processes, for which earlier methods failed. Applying it to typical macroeconomic models such as stochastic growth and income fluctuation models, they show that the relative error in the solution accuracy is less than 1% with the Rouwenhorst method, compared with 10–20% with earlier methods.

Galindev and Lkhagvasuren 2010 generalize the Rouwenhorst method to the multivariate case by transforming a VAR into a set of cross-correlated AR(1) processes. However, their method works only when the AR(1) processes are equally persistent (a knife-edge case), for otherwise the state space is not finite.

Gospodinov and Lkhagvasuren 2014 propose an alternative discretization method of VARs by first discretizing independent AR(1) processes using the Rouwenhorst method and then targeting the first and second conditional moments to mimic the conditional

[13] In the original paper, Tauchen 1986 himself admits that "[e]xperimentation showed that the quality of the approximation remains good except when λ [the persistence parameter] is very close to unity."

distributions of the actual VAR process. Solving a stochastic growth model with a highly persistent bivariate VAR, they find that the relative error in the solution accuracy is about 1–3% with their method, compared with 10–30% with the Tauchen method.

Since our method matches conditional moments, it is similar in spirit to Rouwenhorst 1995 (AR(1)) and Gospodinov and Lkhagvasuren 2014 (VAR(1)), though our method is not limited to VARs. Here we contrast our method to these two in more details. According to Proposition 3 in Kopecky and Suen 2010, the ergodic distribution of the resulting Markov chain of the Rouwenhorst method is a standardized binomial distribution with parameter $N-1$ and $s=1/2$, so by the central limit theorem it converges to $N(0,1)$ as $N\to\infty$. This argument suggests that the Rouwenhorst method is designed to discretize a Gaussian AR(1). It immediately follows that neither our method (for AR(1)) nor the Rouwenhorst method is a special case of the other: our method is not limited to Gaussian AR(1) processes (Proposition 2 and Corollary 3 do not assume normality), and generally has a different grid.

With regard to VARs, both the Gospodinov and Lkhagvasuren 2014 (GL) method and ours target the first and second conditional moments. The GL method uses the Rouwenhorst method to obtain a preliminary discretization and then targets the moments. As GL acknowledge in their paper, the GL method has fewer free variables than the number of targeted moments, and hence it is generally impossible to match all moments. While we do not have a proof that our method matches all first and second conditional moments (Proposition 1 shows that it is possible to match conditional means), according to our experience it seems that for most applications we can indeed match all first two conditional moments when we use the even-spaced grid. Again neither of the two methods is a special case of the other.

We do not claim that our method is always preferable, although we emphasize that our method is not limited to the discretization of linear Gaussian processes. Whether

our method is superior or not can only be answered by studying the accuracy in specific problems. The Online Appendix compares the accuracy of discretization and shows that our method outperforms existing ones by several orders of magnitude. However, discretization is not an end in itself. A more important question is whether different discretization methods lead to substantial differences in the solution accuracy of dynamic economic models, and whether these differences matter economically. We provide answers to these questions in the next sections.

2.4 Solution Accuracy of Asset Pricing Models

Whenever one proposes a new numerical method for solving dynamic models, it must be evaluated by two criteria: (i) Does the new method improve the solution accuracy of well-known, standard dynamic economic models? (ii) Can the new method be applied to solve more complicated models for which existing methods are not readily available? In order for a new method to be useful, it must meet at least one (preferably both) of these two criteria.

This section addresses these questions by solving simple asset pricing models with or without Gaussian shocks. We use the closed-form solutions obtained by Burnside 1998 for Gaussian shocks and Tsionas 2003 for non-Gaussian shocks as comparison benchmarks.[14]

2.4.1 Model and Numerical Solution

Consider a representative agent with additive CRRA utility function

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \frac{C_t^{1-\gamma}}{1-\gamma},$$

[14] Collard and Juillard 2001 and Schmitt-Grohé and Uribe 2004 also use this model in order to evaluate the solution accuracy of the perturbation method.

where C_t is consumption, $\beta > 0$ is the discount factor, and $\gamma > 0$ is the coefficient of relative risk aversion. The agent is endowed with aggregate consumption $\{C_t\}_{t=0}^{\infty}$, and can trade assets in zero net supply. Let D_t be the dividend to an asset and P_t be its price. When log consumption and dividend growth

$$x_t = (\log(C_t/C_{t-1}), \log(D_t/D_{t-1}))$$

follow a VAR(1) process with i.i.d. shocks, it is possible to obtain a closed-form solution for the price-dividend ratio $V_t = P_t/D_t$, which depends only on x_t. See the Online Appendix for details.

We obtain numerical solutions as follows. By the Euler equation, we have

$$P_t = \mathbb{E}_t[\beta(C_{t+1}/C_t)^{-\gamma}(P_{t+1} + D_{t+1})]. \tag{2.12}$$

Dividing (2.12) by D_t, we obtain

$$V_t = \beta \mathbb{E}_t[\exp(\alpha' x_{t+1})(V_{t+1} + 1)], \tag{2.13}$$

where $\alpha = (-\gamma, 1)'$. Suppose that the process for consumption and dividend growth is discretized. Let $s = 1, \ldots, S$ be the states, x_s be the vector of log consumption/dividend growth in state s, and $P = (\pi_{ss'})$ be the transition probability matrix. Then the discrete analog of (2.13) is

$$v_s = \beta \sum_{s'=1}^{S} \pi_{ss'} \mathrm{e}^{\alpha' x_{s'}} (v_{s'} + 1), \tag{2.14}$$

where v_s is the price-dividend ratio in state s. Let $v = (v_1, \ldots, v_S)'$ $(S \times 1)$ and $X = (x_1', \ldots, x_S')'$ $(S \times 2)$ be the matrices of those values. Then (2.14) is equivalent to the linear

equation

$$v = \beta P \operatorname{diag}(\mathrm{e}^{X\alpha})(v+1) \iff v = (I - \beta P \operatorname{diag}(\mathrm{e}^{X\alpha}))^{-1} \beta P \mathrm{e}^{X\alpha}. \tag{2.15}$$

This formula gives the price-dividend ratio only at the grid points, and one might be interested in computing the value at any point. In this case, we can use the projection method (Judd 1992). The idea of the projection method with Chebyshev collocation is to approximate the unknown policy function using Chebyshev polynomials as a basis.[15] Suppose we approximate $V(x)$ as

$$\widehat{V}(x;b) = \sum_{s=1}^{S} b_s \Psi_s(x),$$

where $\{\Psi_s\}_{s=1}^{S}$ is a set of basis functions (Chebyshev polynomials) and $b = \{b_s\}_{s=1}^{S}$ is the vector of coefficients to be determined. We can solve for b that sets the Euler equation (2.13) to exactly zero at each of the S grid points implied by each discretization method, which leads to an exactly identified system. The equation becomes

$$\widehat{V}(x_s;b) = \beta \sum_{s'=1}^{S} \pi_{ss'} \mathrm{e}^{\alpha' x_{s'}} \left(\widehat{V}(x_{s'};b) + 1 \right). \tag{2.16}$$

However, if we set $v_s = \widehat{V}(x_s;b)$, then (2.16) becomes identical to (2.14)! Therefore finding coefficients $\{b_s\}$ that solve (2.16) is equivalent to first solving the linear equation (2.14) (whose solution is given by (2.15)) and then finding an interpolating polynomial. We summarize the above discussion in the following proposition.

Proposition 4. *Solving an asset pricing model with a continuous state space using*

[15] Unlike standard Chebyshev collocation, we are constrained to solve for coefficients that set the Euler equation residuals equal to 0 at the discretization points rather than the zeroes of the Chebyshev polynomial. This in general means we are only guaranteed pointwise convergence of our approximation rather than uniform convergence.

discretization and projection is equivalent to solving a model with a discrete state space, which can be done by inverting a matrix as in (2.15)*. The continuous solution can be obtained by interpolating the discrete solution.*

Proposition 4 is quite powerful. Note that there is nothing specific to the preferences of the agent or the underlying stochastic process needed to apply the proposition. For example, suppose that the agent has a general recursive utility of the form

$$U_t = f(C_t, \mathscr{M}_t(U_{t+1})), \tag{2.17}$$

where U_t is the utility at time t, C_t is consumption, f is the aggregator, and $\mathscr{M}_t$ is the certainty equivalent of the continuation utility U_{t+1}.[16] Suppose that $f, \mathscr{M}$ are homogeneous of degree 1 (which is true for almost all applications) and the underlying stochastic process is discretized. Dividing (2.17) by C_t, we can solve for the S nonlinear equations in S unknowns

$$u_s = f(1, \mathscr{M}_s(\mathrm{e}^{x_{ss'}} u_{s'})), \tag{2.18}$$

where $x_{ss'}$ is log consumption growth from state s to s' and $u_s = (U_t/C_t)(s)$ is the utility-consumption ratio in state s. After solving for these values $\{u_s\}$, one can compute the pricing kernel and price any assets by inverting a matrix as in (2.15). In practice, solving (2.18) and inverting a matrix to compute asset prices take only a fraction of a second to carry out.[17]

[16] A typical example is $f(c,v) = ((1-\beta)c^{1-1/\psi} + \beta v^{1-1/\psi})^{\frac{1}{1-1/\psi}}$ (CES aggregator with elasticity of intertemporal substitution ψ) and $\mathscr{M}_t(X) = \mathbb{E}_t[X^{1-\gamma}]^{\frac{1}{1-\gamma}}$ (CRRA certainty equivalent with relative risk aversion γ) in which case we obtain the Epstein-Zin preference.

[17] The idea of using discretization to solve asset pricing models is not particularly new: see, for example, Mehra and Prescott 1985, Cecchetti, Lam, and Mark 1993, and Bonomo, Garcia, Meddahi, and Tédongap 2011, among others. The point is that there have been no systematic ways to accurately discretize the underlying stochastic process in the literature to make discretization a viable option.

2.4.2 Calibration

We calibrate the model at annual frequency. We select the preference parameters $\beta = 0.95$ and $\gamma = 2$, which are relatively standard in the macro literature. We consider three specifications for the law of motion of x_t: Gaussian AR(1), Gaussian VAR(1), and AR(1) with non-Gaussian shocks. We estimate the parameters of each of these models using data on real personal consumption expenditures per capita of nondurables from FRED, and 12-month moving sums of dividends paid on the S&P 500 obtained from the spreadsheet in Welch and Goyal 2008.[18] For the two univariate specifications, we assume that $C_t = D_t$, *i.e.*, $x_{1,t} = x_{2,t} = x_t$, and use the data on dividends to estimate the parameters.

The reason why we use dividend data instead of consumption data for the univariate models is as follows. Given the mean μ and persistence ρ of the AR(1) process, according to Tsionas 2003 the price-dividend ratio depends only on the moment generating function (MGF) $M(s)$ of the shock distribution in the range $\frac{1-\gamma}{1-\rho} \leq s \leq 1-\gamma$ (assuming $\gamma > 1$ and $\rho > 0$). But if two shock distributions have identical mean and variance, then the Taylor expansion of their MGF around $s = 0$ will coincide up to the second order term. Therefore, in order to make a difference for asset pricing, we either need to (i) move away from $s = 0$ by increasing γ, (ii) make the domain of the MGF larger by increasing ρ, or (iii) make the MGF more nonlinear by increasing the variance or skewness. Since dividend growth is more persistent, volatile, and skewed than consumption growth, using dividend growth will make the contrasts between methods more stark.

[18] http://www.hec.unil.ch/agoyal/

2.4.3 Solution Accuracy

After computing the numerical and closed-form solutions as described in the Online Appendix, we evaluate the accuracy by the $\log_{10}$ relative errors

$$\log_{10}\left|\widehat{V}(x)/V(x)-1\right|,$$

where $V(x)$ is the true price-dividend ratio at x and $\widehat{V}(x)$ is the approximate (numerical) solution corresponding to each method obtained by the interpolating polynomial as in Proposition 4. To compare the relative errors of each method, we first take the largest common support across all discretization methods so that the approximation is well defined, and then compute the relative errors on a fine grid (say 1,001 points in each dimension) on this support. All methods beginning with "ME" refer to the maximum entropy method developed in this paper with different choices of the underlying grid and quadrature formula. For example, "ME-Even" refers to the maximum entropy method using an even-spaced grid.

Gausian AR(1)

Modeling the dynamics of dividend growth by a Gaussian AR(1) is straightforward and we relegate the details to the Online Appendix.

Gaussian VAR(1)

We next consider specifying the joint dynamics of dividend growth and consumption growth as a Gaussian VAR(1)

$$x_t = (I-B)\mu + Bx_{t-1} + \eta_t, \quad \eta_t \sim N(0,\Psi)$$

where μ is a 2×1 vector of unconditional means, B is a 2×2 matrix with eigenvalues less than 1 in absolute value, η is a 2×1 vector of shocks, and Ψ is a 2×2 variance covariance matrix. The estimated parameters of the VAR(1) model are

$$\mu = \begin{bmatrix} 0.0128 \\ 0.0561 \end{bmatrix}, \quad B = \begin{bmatrix} 0.3237 & -0.0537 \\ 0.2862 & 0.3886 \end{bmatrix}, \quad \Psi = \begin{bmatrix} 0.000203 & 0.000293 \\ 0.000293 & 0.003558 \end{bmatrix}.$$

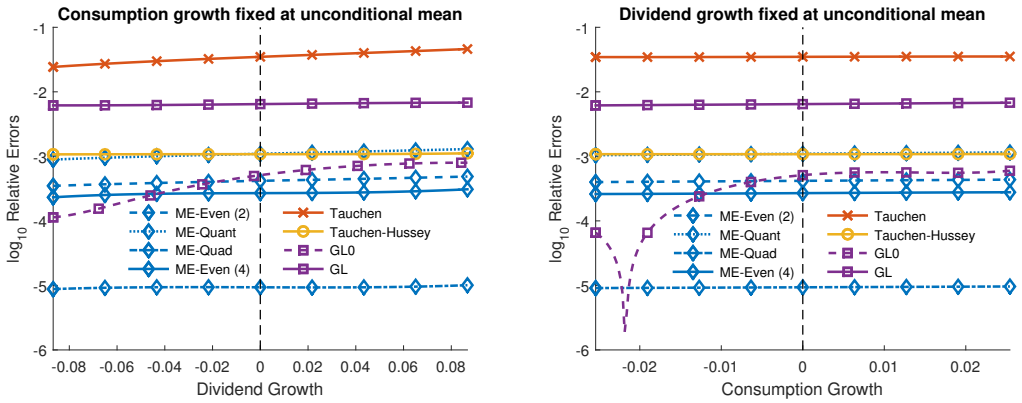
The eigenvalues of B are $0.3561 \pm 0.1196i$, with spectral radius $\rho(B) = 0.3757$, so the VAR is moderately persistent.

We consider eight different discretization methods. For our method, we consider the even-spaced grid with 2 or 4 moments (ME-Even (2,4)), the quantile grid (ME-Quant), and the Gauss-Hermite quadrature grid (ME-Quad). For existing methods, we consider those of Tauchen 1986(Tau), Tauchen and Hussey 1991 (TH), and Gospodinov and Lkhagvasuren 2014 with (GL) and without (GL0) moment matching. Figure 2.1 shows the graphs of $\log_{10}$ relative errors for the VAR(1) model. Table 2.1 shows the mean and maximum $\log_{10}$ relative errors over the entire grid.

Table 2.1. Mean and Maximum $\log_{10}$ Relative Errors of Price-Dividend Ratio Approximations, VAR(1) Model

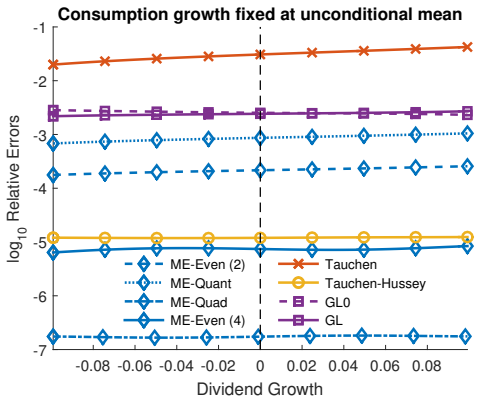
	ME methods				Existing methods			
N	Even (2)	Quant	Quad	Even (4)	Tau	TH	GL0	GL
Mean $\log_{10}$ *errors*								
5	-3.381	-2.963	-5.028	-3.570	-1.463	-2.964	-3.439	-2.191
7	-3.667	-3.066	-6.758	-5.134	-1.520	-4.920	-2.586	-2.618
9	-3.949	-3.146	-8.563	-6.739	-1.546	-6.900	-2.449	-3.106
Maximum $\log_{10}$ *errors*								
5	-3.292	-2.865	-4.975	-3.485	-1.327	-2.890	-2.365	-1.982
7	-3.566	-2.954	-6.717	-4.891	-1.360	-4.838	-2.125	-2.140
9	-3.838	-3.022	-8.451	-5.730	-1.370	-6.581	-2.212	-2.471

Mean and maximum $\log_{10}$ relative errors for the asset pricing model with VAR(1) consumption/dividend growth.

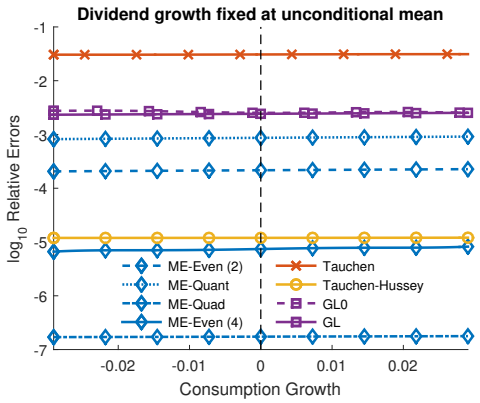


(a) $N = 5$, fixed consumption.

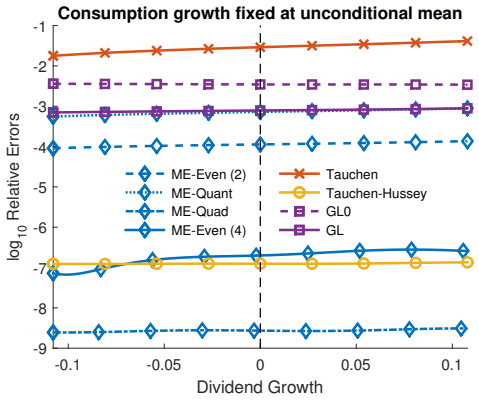
(b) $N = 5$, fixed dividend.



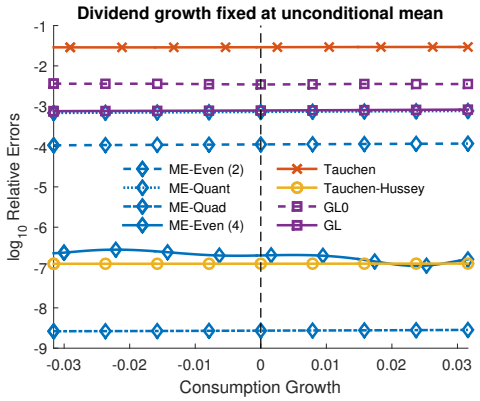
(c) $N = 7$, fixed consumption.



(d) $N = 7$, fixed dividend.



(e) $N = 9$, fixed consumption.



(f) $N = 9$, fixed dividend.

Figure 2.1. $\log_{10}$ Relative Errors of Price-Dividend Ratio Approximations, VAR(1) Model

Note: each row corresponds to a certain number of grid points ($N = 5, 7, 9$). The left panels show the accuracy along the dividend growth dimension, fixing consumption growth at its unconditional mean. The right panels fix dividend growth at its unconditional mean and vary consumption growth. The grids are demeaned so that the unconditional mean corresponds to 0 in the figures.

For all choices of N, the Gaussian quadrature based methods, ME-Quad and TH, perform the best, with ME-Quad being always about two orders of magnitude more accurate than TH. For even-spaced methods, the order of accuracy is always ME-Even (4) $>$ ME-Even (2) $>$ GL0, GL $>$ Tauchen, and ME-Even (4) is as accurate as Tauchen-Hussey. ME-Quant is not particularly accurate but its performance is similar to the GL methods. According to Table 2.1, the conclusions drawn from Figure 2.1 are robust.

AR(1) with Non-Gaussian Shocks

Researchers often assume normality of the conditional shock distributions for analytical and computational convenience. However, there is much evidence of non-normality in financial data. One might prefer to specify a parametric distribution with fatter tails, or refrain from parametric specifications altogether. For this reason, we consider an AR(1) with i.i.d., but non-Gaussian shocks:

$$x_t = (1-\rho)\mu + \rho x_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim F.$$

We model the shock distribution F by a Gaussian mixture, because it is flexible yet analytically tractable (all moments and moment generating function have closed-form expressions). Table 2.2 shows the parameter estimates.

Figure 2.2 plots the PDFs of ε_t fit to the dividend growth data under the assumptions of normal and Gaussian mixture shocks, as well as the nonparametric kernel density estimate. The Gaussian mixture with three components appears to capture the skewness and kurtosis lacking in the normal specification by placing more weight on large negative realizations of the shock as well as ones close to zero.

We consider six different discretizations for the log dividend growth process. The first two are the Rouwenhorst 1995 and the Tauchen and Hussey 1991 methods, which

Table 2.2. Parameter Values, AR(1) Model with Gaussian Mixture Shocks

Parameter	Symbol	Value
Mean dividend growth	μ	0.0559
Persistence of dividend growth	ρ	0.4049
Volatility of dividend growth	σ	0.0589
Proportion of mixture components	w_j	0.0304, 0.8489, 0.1207
Mean of mixture components	μ_j	-0.2282, -0.0027, 0.0766
S.D. of mixture components	σ_j	0.0513, 0.0316, 0.0454

Note: this table shows the parameter estimates of the AR(1) process with Gaussian mixture shocks $x_t = (1-\rho)\mu + \rho x_{t-1} + \varepsilon_t$, where $x_t = \log(D_t/D_{t-1})$ is log dividend growth and $\varepsilon_t \sim N(\mu_j, \sigma_j^2)$ with probability w_j, $j = 1,\ldots,J$. μ, ρ are estimated by OLS. $\sigma = \sqrt{\text{Var}[\varepsilon_t]}$ is computed from the squared sum of residuals. The Gaussian mixture parameters are estimated by maximum likelihood from the residuals, and the number of components $J = 3$ is chosen to minimize the Akaike Information Criterion (AIC).

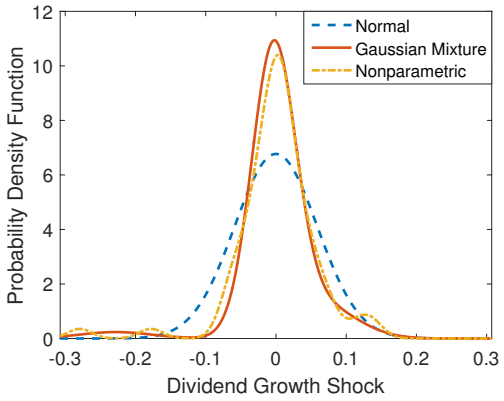


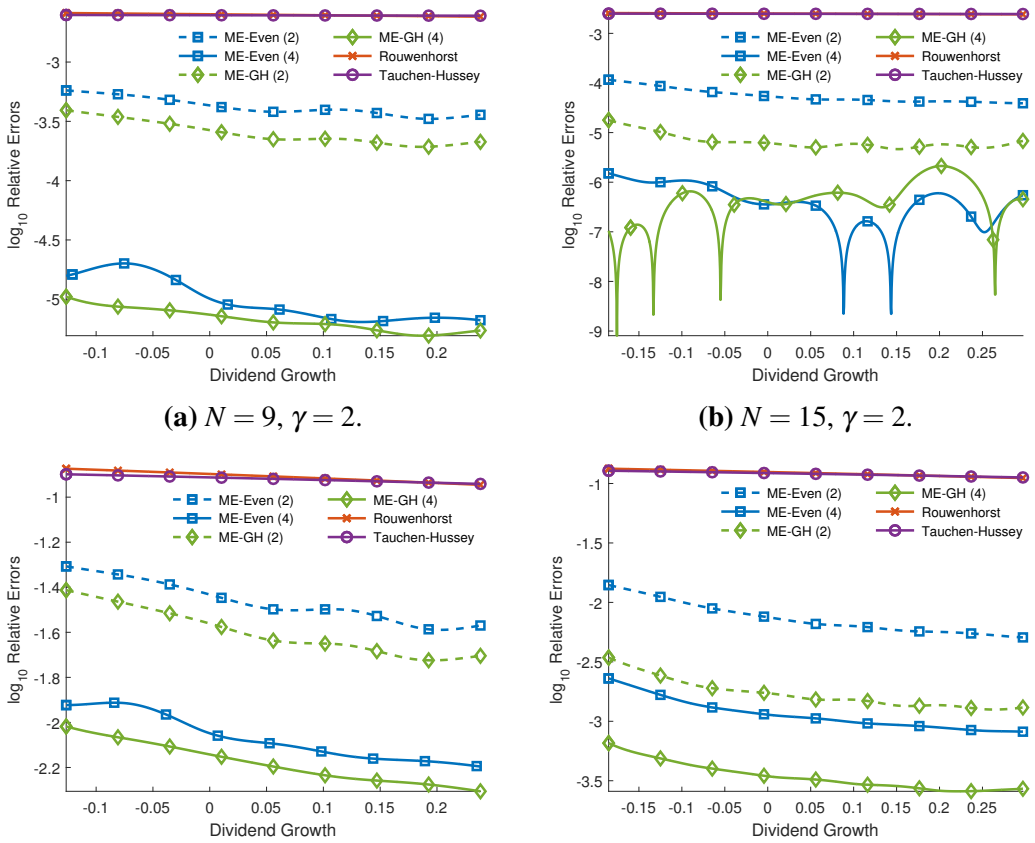
Figure 2.2. Densities fitted to AR(1) OLS residuals.

can be thought of as a case where the researcher incorrectly believes the conditional density to be Gaussian. The other four methods are the ME methods with even-spaced (ME-Even) or Gauss-Hermite quadrature grid (ME-GH), each with 2 or 4 moments matched. For ME-Even, we implement the discretization exactly as in Algorithm 3, except that we use the Gaussian mixture density instead of the normal density. We choose the grid spacing as the upper bound in Corollary 3. For ME-GH, we take the following approach. Suppose the true (Gaussian mixture) density at a given grid point is $f(x)$. Let $\phi(x)$ be the normal density with mean 0 and the same standard deviation as $f(x)$. Then the expectation of a function $g(x)$ is

$$\int g(x)f(x)\,\mathrm{d}x = \int g(x)\frac{f(x)}{\phi(x)}\phi(x)\,\mathrm{d}x \approx \sum_{n=1}^{N} w_n \frac{f(x_n)}{\phi(x_n)} g(x_n),$$

where $\{x_n\}$ and $\{w_n\}$ are nodes and weights for the Gauss-Hermite quadrature corresponding to $\phi(x)$. This argument suggests that we can use the Gauss-Hermite quadrature grid with weights $w'_n = w_n \frac{f(x_n)}{\phi(x_n)}$ in order to discretize $f(x)$. Figure 2.3 plots the $\log_{10}$ relative errors of the AR(1) model with Gaussian mixture shocks. Table 2.3 shows the mean and maximum $\log_{10}$ relative errors.

As we can see from the figure and the table, the order of accuracy is always ME-GH $\approx$ ME-Even $>$ Rouwenhorst $\approx$ Tauchen-Hussey, and matching 4 moments instead of 2 increases the solution accuracy by about 1 to 2 orders of magnitude. For low risk aversion ($\gamma = 2$), even the misspecified models (Rouwenhorst and Tauchen-Hussey) have relative errors less than 10^{-2} or 1%, so the choice of the discretization method does not matter so much. However, with higher risk aversion ($\gamma = 5$), the misspecified models are off by more than 10^{-1} (10%), while ME methods with 4 moments has errors less than 10^{-2} (1%) with 9 points and 10^{-3} (0.1%) with 15 points. Hence the choice of the discretization method makes an economically significant difference when risk aversion is



(a) $N = 9$, $\gamma = 2$. **(b)** $N = 15$, $\gamma = 2$.

(c) $N = 9$, $\gamma = 5$. **(d)** $N = 15$, $\gamma = 5$.

Figure 2.3. $\log_{10}$ Relative Errors of Price-Dividend Ratio Approximations, AR(1) Model with Gaussian Mixture Shocks

Note: the top panels show the accuracy for approximations to the benchmark model with risk aversion $\gamma = 2$ and different number of grid points $N = 9, 15$. The bottom panels show the results for an alternative specification in which the risk aversion is higher at $\gamma = 5$.

Table 2.3. Mean and Maximum $\log_{10}$ Relative Errors of Price-Dividend Ratio Approximations, AR(1) Model with Gaussian Mixture Shocks

		ME methods				Existing methods	
N	γ	Even (2)	Even (4)	GH (2)	GH (4)	R	TH
Mean $\log_{10}$ *errors*							
9	2	-3.381	-5.013	-3.602	-5.176	-2.602	-2.606
15	2	-4.264	-6.445	-5.189	-6.414	-2.604	-2.606
9	5	-1.466	-2.071	-1.602	-2.182	-0.909	-0.919
15	5	-2.137	-2.948	-2.774	-3.467	-0.913	-0.919
Maximum $\log_{10}$ *errors*							
9	2	-3.239	-4.698	-3.406	-4.978	-2.587	-2.603
15	2	-3.935	-5.821	-4.748	-5.673	-2.591	-2.602
9	5	-1.307	-1.913	-1.413	-2.018	-0.874	-0.900
15	5	-1.854	-2.639	-2.464	-3.184	-0.875	-0.892

Note: Even (L): even-spaced grid with L moments; GH (L): Gauss-Hermite quadrature grid with L moments; R: Rouwenhorst 1995 method; TH: Tauchen and Hussey 1991 method.

moderately high, which is often the case for many asset pricing models in the literature.

2.5 Solution Accuracy of a Rare Disasters Model

To illustrate the general applicability of our method, in this section we solve an asset pricing model with variable rare disasters (Gabaix 2012). There are several good reasons to consider this model. First, the dynamics of the underlying stochastic process are nonlinear and non-Gaussian, which makes our method more useful. Second, Gabaix's model admits closed-form solutions, which makes the accuracy comparison particularly simple. Finally, since rare disaster models have recently become quite popular in the literature (Rietz 1988; Barro 2006; Gourio 2012; Wachter 2013), providing a simple yet accurate solution algorithm seems to be useful, especially for the purpose of calibration and estimation.

2.5.1 Model

Gabaix 2012 considers a representative-agent asset pricing model in an endowment economy. The representative agent has CRRA preferences

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \mathrm{e}^{-\rho t} \frac{C_t^{1-\gamma}}{1-\gamma},$$

where $\rho > 0$ is the discount rate and $\gamma > 0$ is relative risk aversion. Disasters occur with probability p_t at time $t+1$. The consumption growth is given by

$$\frac{C_{t+1}}{C_t} = \mathrm{e}^{g_C} \times \begin{cases} 1, & \text{(no disaster)} \\ B_{t+1}, & \text{(disaster)} \end{cases}$$

where g_C is the growth rate in normal times and $B_{t+1} \in (0,1]$ is the consumption recovery rate after a disaster. Similarly, the dividend growth is

$$\frac{D_{t+1}}{D_t} = \mathrm{e}^{g_D} \times \begin{cases} 1, & \text{(no disaster)} \\ F_{t+1}, & \text{(disaster)} \end{cases}$$

where g_D is the growth rate in normal times and $F_{t+1} \in (0,1]$ is the dividend recovery rate after a disaster. Gabaix 2012 defines the following quantity, which he calls "resilience":

$$H_t = p_t \mathbb{E}_t^{\mathrm{D}} [B_{t+1}^{-\gamma} F_{t+1} - 1], \tag{2.19}$$

where $\mathbb{E}_t^{\mathrm{D}}$ denotes the expectation conditional on *d*isaster. Instead of specifying the dynamics of the fundamentals p_t, B_t, F_t individually, Gabaix directly specifies the dynamics

of $H_t = H_* + \widehat{H}_t$ as follows:

$$\widehat{H}_{t+1} = \frac{1+H_*}{1+H_t} \mathrm{e}^{-\phi_H} \widehat{H}_t + \varepsilon^H_{t+1}, \tag{2.20}$$

where H_* is a constant, $\phi_H > 0$ is the speed of mean reversion at $H_t = H_*$, and ε^H_{t+1} is an innovation. Since $1 + H_t$ appears in the denominator of the right-hand side, (2.20) is a highly nonlinear process. It turns out that the price-dividend ratio at time t depends only on $\widehat{H}_t$ independent of the distribution of ε^H_{t+1}, and Gabaix obtains a closed-form solution (see Eq. (13) in his paper).

2.5.2 Solution Accuracy

To compare numerical solutions obtained by our method to the exact solution, we need to discretize the process (2.20). Since the distribution of the innovation ε^H_{t+1} does not matter, and since Gabaix shows that the process $\{\widehat{H}_t\}$ must be bounded, we assume that the distribution of $\widehat{H}_{t+1}$ given $\widehat{H}_t$ is a beta distribution (properly rescaled) with mean and variance implied by (2.20). Once we specify the conditional distribution this way, it is straightforward to discretize the Markov process using our method. See the Online Appendix for the details on discretization and the computation of the numerical solution. Although there are no accepted standard ways for solving the rare disasters model, we also compare the solution accuracy of our method to the perturbation method proposed in Levintal 2014.[19]

For the parameter values, following Gabaix 2012 we set the discount rate $\rho = 0.0657$, relative risk aversion $\gamma = 4$, consumption and dividend growth rate $g_C = g_D = 0.025$, disaster probability $p = 0.0363$, consumption recovery rate $B = 0.66$, and the speed of mean reversion $\phi_H = 0.13$. The implied value for the constant H_* in (2.20)

[19] https://sites.google.com/site/orenlevintal/5th-order-perturbation

is 0.09. Figure 2.4 shows the ergodic distribution of the variable part of resilience $\widehat{H}$ computed from the discrete approximation with $N = 201$ points. The distribution is bimodal.

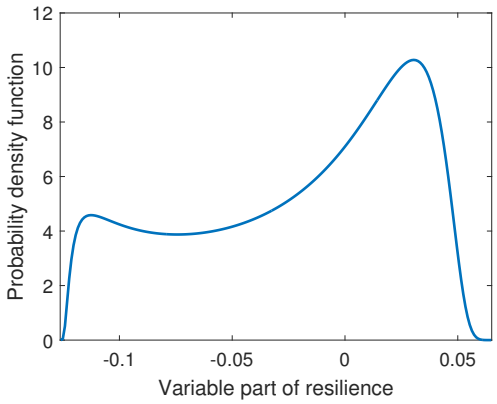
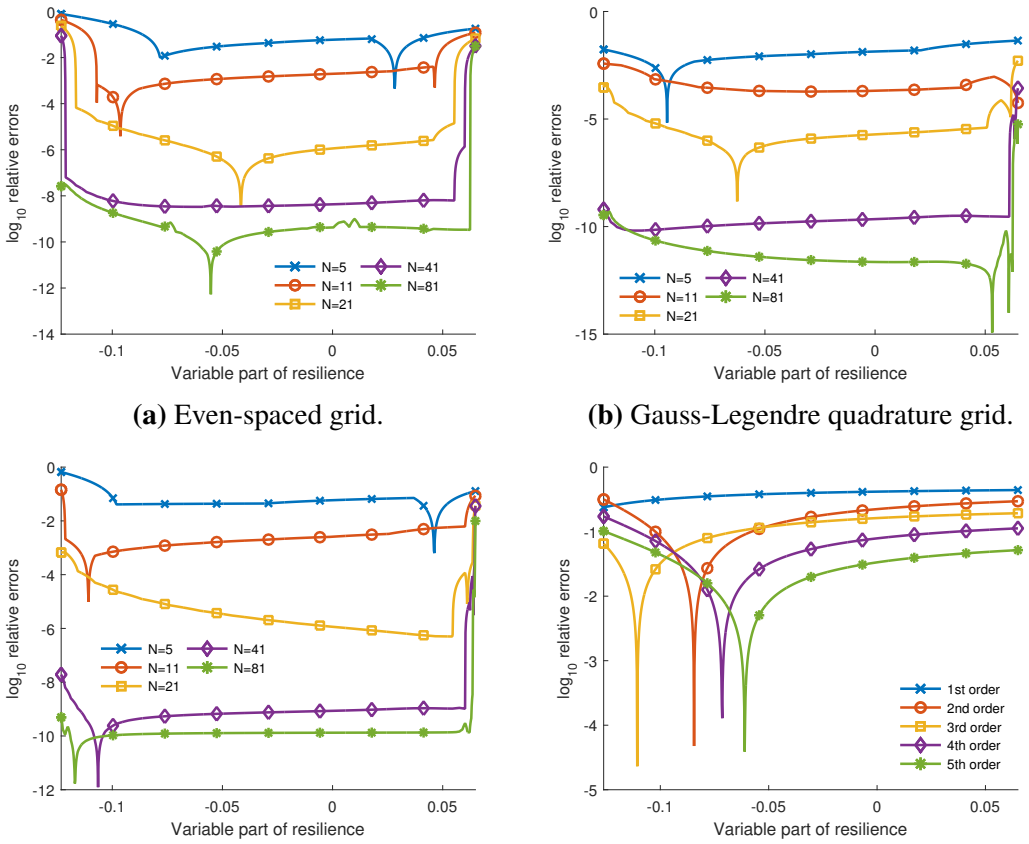


Figure 2.4. Ergodic distribution of the variable part of resilience $\widehat{H}$.

For our method, we consider the even-spaced grid, Gauss-Legendre quadrature grid, and the Clenshaw-Curtis quadrature grid, which are the most natural choices since the integration is over a bounded interval. The number of points are $N = 5, 11, 21, 41, 81$. For the perturbation method in Levintal 2014, we consider up to the fifth-order approximation (the maximum allowed). In order to apply the perturbation method, we need to supply the unconditional standard deviation of the innovation in resilience, ε^H_{t+1}. We compute this number using the ergodic distribution in Figure 2.4, which is 0.0174. We also simulated the true process (2.20) for a long time and verified that we obtain the same number up to four decimal places. Figure 2.5 shows the $\log_{10}$ relative errors of the price-dividend ratio. Table 2.4 shows the mean and maximum $\log_{10}$ relative errors over the entire grid.

Because the resilience process (2.20) is highly nonlinear, we need many grid points in order to obtain an accurate solution. Overall using the Gauss-Legendre quadrature grid (Figure 2.5b) is the most accurate, with relative errors about 10^{-3} with $N = 11$



(a) Even-spaced grid.

(b) Gauss-Legendre quadrature grid.

(c) Clenshaw-Curtis quadrature grid.

(d) Perturbation method.

Figure 2.5. $\log_{10}$ Relative Errors of Price-Dividend Ratio Approximations, Rare Disasters Model

Table 2.4. Mean and Maximum $\log_{10}$ Relative Errors of Price-Dividend Ratio Approximations, Rare Disasters Model

	ME methods			Perturbation	
N	Even	Gauss-Legendre	Clenshaw-Curtis	Order	
Mean $\log_{10}$ *errors*					
5	-1.187	-1.982	-1.218	1	-0.422
11	-2.582	-3.451	-2.676	2	-0.856
21	-5.383	-5.560	-5.354	3	-1.007
41	-8.007	-9.679	-9.040	4	-1.268
81	-9.228	-11.23	-9.873	5	-1.590
Maximum $\log_{10}$ *errors*					
5	-0.107	-1.353	-0.182	1	-0.356
11	-0.365	-2.422	-0.841	2	-0.501
21	-0.628	-2.291	-1.430	3	-0.715
41	-1.053	-3.567	-1.447	4	-0.765
81	-1.503	-5.245	-2.003	5	-0.992

points, 10^{-5} with $N = 21$ points, and 10^{-10} with $N = 41$ points. Hence for practical purposes 11 points are enough. Clenshaw-Curtis quadrature (Figure 2.5c) is similar to Gauss-Legendre, as documented in Trefethen 2008. The performance of the even-spaced grid (Figure 2.5a) is worse near the boundary points. This is because the conditional variance of the resilience process (2.20) approaches zero near the boundary, which makes it hard to match the conditional variance. Since there are many grid points near the boundary for Gauss-Legendre and Clenshaw-Curtis, a low variance is not a problem. The perturbation method (Figure 2.5d) is not so accurate, with about 10% error with 3rd-order approximation and 2.6% error with 5th-order. Even the 5-point Gauss-Legendre discretization is more accurate than the 5th-order perturbation in terms of both mean and maximum $\log_{10}$ errors.

Do these differences in solution accuracy economically matter? To address this question, we simulate the resilience process (2.20) for $T = 100,000$ periods and compute some financial moments from the true solution as well as the numerical solutions. Table

2.5 shows the results. As expected from Figure 2.5 and Table 2.4, the 11-point Gauss-Legendre discretization gives accurate results up to the third significant digit (0.1%). The perturbation method does not fare well: with the 1st-order approximation, the stock return is 4 percentage points higher than the true value; the 3rd order approximation is off by 10–20%, and the 5th-order approximation is off by about 10% for the standard deviation.

Based on the numerical results in the last two sections, we provide some recommendations to allow the reader to make an informed decision on what kind of computational strategy to adopt. The perturbation method is fast but it is inherently a local approximation. When the model is highly nonlinear and shocks are large, the solution accuracy can be poor. Discretization is easy to implement and seems to be accurate enough for most problems. For Gaussian VARs, our method (with even-spaced or quadrature grid) seems best. Numerical results in the appendix suggest that for univariate Gaussian AR(1) process, ME-Quad is most accurate for persistence less than 0.8, ME-Even is most accurate for persistence between 0.8 and 0.99, and the Rouwenhorst method is best for persistence 0.99 and beyond (because the Rouwenhorst method is error-free, *i.e.*, it does not involve any numerical optimization). However, for persistence beyond 0.99, it may be better to use the projection method. Pohl, Schmedders, and Wilms 2014 suggest that for solving the long run risk model (Bansal and Yaron 2004), which features very persistent processes, using the projection method makes an economically meaningful difference in the solution accuracy. For nonlinear or non-Gaussian processes, as in the rare disasters model, our discretization method would be the first choice since there may not be any readily available quadrature formulas to use along with the projection method.

Table 2.5. Financial Moments, Rare Disasters Model

	ME methods			Perturbation	
N	Even	Gauss-Legendre	Clenshaw-Curtis	Order	
Mean P/D				True = 16.7330	
5	17.5101	16.9876	17.8134	1	9.9614
11	16.8498	16.7268	16.6894	2	13.6059
21	16.7523	16.7330	16.7329	3	14.2745
41	16.7351	16.7330	16.7330	4	15.6998
81	16.7331	16.7330	16.7330	5	16.3267
Standard deviation of $\log(P/D)$				True = 0.3366	
5	0.2432	0.3467	0.2955	1	0.2640
11	0.3129	0.3371	0.3342	2	0.1859
21	0.3309	0.3366	0.3366	3	0.2718
41	0.3359	0.3366	0.3366	4	0.2717
81	0.3366	0.3366	0.3366	5	0.3020
Mean stock returns (%)				True = 6.9574	
5	6.2558	6.9003	6.3332	1	11.4419
11	6.7882	6.9627	6.9637	2	7.9205
21	6.9187	6.9575	6.9577	3	7.8651
41	6.9527	6.9574	6.9574	4	7.1212
81	6.9572	6.9574	6.9574	5	6.9676
Standard deviation of stock returns (%)				True = 11.8058	
5	10.2217	12.1749	11.3956	1	9.9833
11	11.5335	11.8175	11.7561	2	6.7575
21	11.7549	11.8062	11.8069	3	9.7367
41	11.8003	11.8058	11.8058	4	9.6381
81	11.8055	11.8058	11.8058	5	10.6445

Note: this table shows the financial moments from $T = 100,000$ simulations. "True" indicates the values from the exact solution. The numbers are slightly different from Table III of Gabaix 2012 because (i) we simulate at the annual frequency, while he simulates at the monthly frequency, and (ii) in Gabaix's calibration, the stock resilience volatility is $\sigma_H = 0.019$ while we have $\sigma_H = 0.0174$ because we specify beta distributions for the conditional dynamics.

2.6 Conclusion

In this paper, we provide a new method for discretizing a general class of stochastic processes by matching low order conditional moments. Our method is computationally tractable and allows researchers to approximate a wide variety of nonlinear non-Gaussian Markov processes. We demonstrate that our method produces discrete approximations which are often several orders of magnitude more accurate than existing methods for both linear and nonlinear stochastic processes. This is the case whether we consider the relative bias of unconditional moments implied by the discretization or the accuracy of solutions to asset pricing models.

Our maximum entropy procedure has a wide range of potential applications beyond asset pricing models. It is common in the quantitative macro literature to use an AR(1) specification for technology or income. We believe that researchers use AR(1) specifications because existing methods do not easily allow for more realistic assumptions. Recent work on the dynamics of the income distribution has shown that while income shocks have roughly constant variance, skewness and kurtosis display significant time-variation (Guvenen, Ozkan, and Song 2014). Our method can be used to solve a life cycle model with a realistic income process by matching the dynamics of these higher order moments. Our method can also be used for estimating nonlinear, non-Gaussian state space models (Farmer 2017). In this paper we considered only tensor grids since our applications involved only one or two state variables. An interesting and important future research topic is to explore the performance of our method in conjunction with sparse grids for solving dynamic models with many state variables.

2.7 Acknowledgements

Chapter 2, in full, is a reprint of the material that has been accepted for publication at Quantitative Economics. Farmer, Leland E.; Toda, Alexis Akira. The dissertation author was a primary author of this paper.

Chapter 3

Pockets of Predictability

3.1 Introduction

A large body of empirical evidence suggests that stock returns are predictable.[1] This evidence has mostly been established using parametric return prediction models–most commonly a linear, constant-coefficient specification–followed by inference on the coefficients capturing the effect of time-varying predictors. Such regressions pool information across long historical periods and are thus designed to capture how certain state variables capture return predictability "on average", across potentially very different economic states.

In this paper, we offer a fundamentally different view of both the nature and source of return predictability. We present new evidence that return predictability is far more concentrated or "local" in time and tends to fall in certain (contiguous) "pockets". For example, using the T-bill rate as a predictor variable, our approach identifies seven pockets whose duration lasts between 62 and 411 days. In total, twelve percent of the sample is spent inside pockets with return predictability. We contrast this with the temporal patterns in return predictability implied by return predictability models

[1] For early studies, see, e.g., Campbell (1987), Fama and French (1988), Fama and French (1989), Keim and Stambaugh (1986), and Pesaran and Timmermann (1995). Lettau and Ludvigson (2010) and Rapach and Zhou (2013) review the extensive literature on return predictability.

conventionally used in the vast academic literature on the subject.

Our analysis uses a new empirical approach to study how return predictability evolves over time. Specifically, we adopt a nonparametric strategy that is capable of identifying "local" patterns in return predictability. This approach offers advantages and novel insights compared with existing methods.[2] Unlike conventional methods that impose parametric restrictions on how return predictability evolves over time, we do not need to take a stand on the return generating process. Instead, our approach lets the data determine the number of "pockets", if any, with locally predictable stock returns.

Our evidence of pockets with local return predictability could, in principle, be due to the repeated use of a test for identifying pockets. For example, if we repeatedly apply our pockets test at a 5% significance level, we would expect on average to wrongly reject the null of no (local) return predictability five percent of the time. To see if the evidence on return predictability identified by our non-parametric approach is consistent with random variations generated under a null of no return predictability (constant expected returns), we simulate stock returns from popular models for return dynamics, including a model that allows for time-varying volatility. We find that standard models fail to match the patterns observed in returns, most notably the number of pockets, the proportion of the sample spent inside pockets, and the mean and maximum integral R^2 value inside pockets. These shortcomings continue to hold if we allow for time-varying return predictability, but assume a constant coefficient on the predictor variable as is common practice in the return predictability literature. We conclude from this evidence that the commonly-used, constant-coefficient, linear return predictability model fails to generate time-variation in expected returns that is consistent with the empirical evidence we observe.

[2]Studies such as Henkel, Martin, and Nardari (2011), Dangl and Halling (2012), and Johannes, Korteweg, and Polson (2014) propose models with time-varying coefficients. However, these studies introduce strong parametric assumptions about changes in the return generating model–either regime switching or a time-varying parameter model.

We next ask what type of asset pricing model is consistent with the presence of return predictability pockets? The dominant interpretation of return predictability is that it reflects a time-varying risk premium. Thus, it is consistent with changes in investors' marginal utility of consumption across different economic states as modeled by many asset pricing models.[3] We provide a new theoretical result which shows that, in fact, linear constant-coefficient return predictability models are consistent with a broad class of asset pricing specifications in common use, including models that allow for time-varying volatility and compound Poisson jumps. When simulating returns from such models we find, however, that this type of specification does not match the pockets of return predictability that we find.

Having shown that the empirical evidence is at odds with conventional constant-coefficient return prediction models, we next advance an alternative explanation of return predictability. Stock prices depend on expected cash flows that occur in the distant future and so are surrounded by considerable uncertainty. The high sensitivity of aggregate stock prices to even minor variations in beliefs about future cash flow growth rates means that cash flow learning effects could be an important source of return movements.[4] Consistent with this intuition, we show that a new type of cash flow learning dynamics can generate return predictability patterns that look like time-varying risk premia in a setting where, by construction, we know that the risk premium is constant.

Building on the predictive systems model of Pástor and Stambaugh (2009), we assume that the cash flow process can be decomposed into a highly persistent, unobserved component that tracks expected cash flows and a temporary shock that is not predictable.[5]

[3] See, for example, Balvers, Cosimano, and McDonald (1989), Bansal and Yaron (2004), Campbell and Cochrane (1999), and Cechetti, Lam and Nelson (1990).

[4] In a model with paradigm shifts, Hong, Stein, and Yu (2007) find that investors learning about the underlying model that generates dividends can give rise to predictable variation in returns and help to match volatility and skewness patterns in returns. In their analysis, agents switch between models that are under-dimensioned specifications relative to the true dividend generating process.

[5] A key difference to Pástor and Stambaugh (2009) is that we model the unobserved component in

While the expected cash flow process is unobserved, investors do, however, observe a state variable that is correlated with variation in the persistent component in expected cash flows and thus can be used to predict future cash flows.

Generalizing the predictive systems approach, we allow both the drift in the expected cash flow process as well as its correlation with the observed state (predictor) variable to undergo discrete changes that we capture through a regime switching process. For commonly used predictor variables such as the T-bill rate and the term spread, it is plausible to expect that the extent to which these variables are informative over future cash flows will vary over time and depend on the underlying monetary policy regime. Moreover, we consider a setting where the expected growth rate of cash flows is highly persistent, using parameter values similar to those adopted by Bansal and Yaron (2004) and Bansal, Kiku, and Yaron (2012) in the long-run risk literature.

We use our regime switching predictive systems model to compare two scenarios. In the first "no-learning" scenario agents observe the regime process underlying the cash flow process. In the second "learning" scenario investors do not observe the underlying regime and so have to recursively update their estimates of the state probabilities using information on returns and the predictor variable to track the state of the economy.

Next, we simulate asset prices under the no-learning and learning scenarios. By construction, the ex-ante risk premium is constant in these simulations. We find that the no-learning model cannot match the empirical evidence on return predictability pockets generated by our prediction models fitted to historical data. In contrast, the model with learning about cash flows is capable of generating pockets with similar duration and return predictability characteristics as those we observe for the actual returns data.

---

expected cash flows and use an asset pricing model to study its implications for prices and returns. Instead, Pastor and Stambaugh directly model the dynamics in expected returns and use economic arguments to constrain the sign of the correlation between innovations in the predictive system. As these constraints do not apply to the cash flow process, they are not imposed in our analysis.

The simulations from the predictive systems model for cash flows suggest that investors' learning about the underlying cash flow process can induce patterns that look, ex-post, like local return predictability even in a model in which ex-ante expected returns are constant. It follows that the existence of local return predictability pockets should not be viewed as a sign of market inefficiency. We confirm this point by comparing ex-post identifications of pockets – which use the full sample, including data after the end of the pocket – to ex-ante identified pockets, which only use data available in real time. This comparison allows us to determine whether a pocket could be detected in real time, with the potential for possible trading gains.[6] Empirically, we find that this is not the case. This suggests that exploiting local variation in return predictability in real time is not an easy task.

Some key differences between our analysis and earlier studies are worth highlighting. Our analysis uses daily stock market returns. This differs from existing studies of return predictability which generally use monthly, quarterly, or annual returns. Using daily stock market returns enables us to study the local nature of return predictability which can change rapidly over short periods of time. Sampling returns at monthly or longer horizons would likely lead us to miss such patterns in return predictability. [7]

Although the nature of return predictability identified by our approach is fundamentally new and we use a very different empirical strategy in our analysis, some of our findings are consistent with earlier studies in the return predictability literature which indicate that return predictability varies over time. For example, Paye and Timmermann (2006), Rapach and Wohar (2006), and Chen and Hong (2012) find evidence of model instability for stock market return prediction models. Similarly, Henkel, Martin, and Nardari (2011) use regime switching models to capture changes in stock return pre-

[6]This is equivalent to comparing results based on two-sided versus one-sided estimation windows.

[7]Since dividends are not available at the daily horizon, our estimates of the regime switching predictive systems model uses a daily proxy for the state of the economy.

dictability, while Dangl and Halling (2012) and Johannes, Korteweg, and Polson (2014) use time-varying parameter models to model predictability in stock returns.

There are also key differences between our findings of local return predictability pockets and the evidence presented in these earlier studies. For example, Henkel, Martin, and Nardari (2011), along with Rapach, Strauss, and Zhou (2010), argue that return predictability is closely linked to the economic cycle. We find that although there exists a link between economic recessions and the return predictability pockets, this link is weak and the stage of the economic cycle only explains a very small part of the time-variation in expected returns that we document.

Authors such as Schwert (2003), Green, Hand, and Soliman (2011) (2011), and McLean and Pontiff (2016) have also found evidence that return predictability patterns can be learned away over time. These papers show that the strength of the evidence of return predictability, either from time-series regressions or from the cross-section, weakens upon the publication of such evidence. A plausible mechanism is that investors' attempts at exploiting predictive patterns leads to their diminishing as new money flows in to undervalued assets or out of overvalued assets. The mechanism in our paper is quite different from these papers. We assume that investors learn about the cash flow process and the asset price is derived endogenously as a function of the expected discounted cash flows. We also document how long it takes for this cash flow learning mechanism to be completed to the point where no additional return predictability is detectable and the amount of return predictability that is present in the interim.[8]

Our findings contribute to several areas of the finance literature. Gaining a

[8]We distinguish between learning about a fixed number of parameters–which eventually (asymptotically) will reveal the true value of the parameters–and incomplete learning for which agents will never learn the true value. The latter situation arises in settings with a latent state whose dimensions increase with the time period. Learning about the underlying regime in a regime switching model is one example since the dimension of the state vector increases in the sample size and so the current state cannot be consistently estimated.

better understanding of both the patterns of return predictability and the source of such predictability has important implications for several areas in finance. The belief that returns are predictable has influenced key areas of finance such as asset allocation (e.g., Aït-Sahalia and Brandt (2001), Barberis (2000), Campbell and Viceira (1999), and Kandel and Stambaugh (1996)), performance evaluation of mutual funds (e.g., Ferson and Schadt (1996), Avramov and Wermers (2006), and Banegas, Gillen, Timmermann, and Wermers (2013)), and theoretical asset pricing models (e.g., Bansal and Yaron (2004)).

The rest of the paper proceeds as follows. Section 2 discusses conventional approaches to modeling return predictability, derives a new result that establishes the class of asset pricing models that lead to the conventional constant-coefficient return predictability model, and introduces our nonparametric methodology for identifying pockets with return predictability. Section 3 introduces our daily data and presents empirical evidence on return predictability pockets using a variety of predictor variables from the literature on return predictability. This section also uses simulations to address whether the pockets could be generated spuriously as a result of the repeated use of (correlated) tests for local return predictability. Section 4 introduces our Markov switching predictive systems model for cash flows and presents evidence on the extent to which incomplete learning about cash flows can generate return predictability pockets that are similar to those found in the data. Section 5 discusses possible alternative explanations and sources of return predictability pockets, while Section 6 concludes. Two appendices contain additional technical material.

3.2 Prediction Models and Estimation Methodology

In this section we first derive a result showing the type of asset pricing models that are consistent with the benchmark linear regression specification that is commonly used in empirical studies of return predictability. Next, we introduce the alternative

non-parametric regression methodology that we use to measure time-variation in return predictability.

3.2.1 Return Prediction Model with Constant Coefficients

We start by providing a set of very general conditions under which the constant coefficient specification holds almost exactly within a fairly general endowment economy. We parameterize cash flow risks in the economy, allowing for stochastic volatility and compound Poisson jumps. To this end, let z_t be an $L \times 1$ vector of state variables capturing the aggregate state of the economy. We assume that this evolves according to a process with the following properties:

. The aggregate state of the economy follows a stationary VAR process:

$$z_{t+1} = \mu + F z_t + \varepsilon_{t+1} \tag{3.1}$$

with y_0 given, where the $L \times L$ matrix F has all of its eigenvalues inside the unit circle. In addition,

1. The log of aggregate dividend growth, Δd_{t+1}, equals $S_d' z_{t+1}$ for some $L \times 1$ vector S_d

2. For any $\gamma \in \mathbb{R}^L$, the conditional Laplace transform of ε_{t+1} satisfies

$$\log E_t[\exp(\gamma' \varepsilon_{t+1})|z_t] = f(\gamma) + g(\gamma)' z_t, \tag{3.2}$$

where $f(\gamma)\colon \mathbb{R}^L \to \mathbb{R}$ and $g(\gamma)\colon \mathbb{R}^L \to \mathbb{R}^L$

Part 1 of Assumption 1 states that aggregate dividend growth can be captured by a linear combination of the elements of a finite-dimensional, stationary vector autoregressive process, z_t. Part 2 of Assumption 1 requires that the logarithm of the

moment-generating function of the innovation vector is affine in the state vector. This restriction is satisfied for a wide class of distributions used in the theoretical asset pricing literature, as the affine structure greatly facilitates analytical tractability.[9]

In addition to the restrictions on the cash flow process in Assumption 1, we also put restrictions on preferences. The main restriction is that the pricing kernel is an exponential affine function of the z_t vector that summarizes the aggregate state of the economy.

. The continuously compounded market return, r_{t+1}, satisfies the Euler equation

$$1 = E_t[\exp(\lambda_0 - \Lambda' z_{t+1} + r_{t+1})] \tag{3.3}$$

where λ_0 is a scalar and Λ is an $L \times 1$ vector

Assumption 2 requires that the pricing kernel has an affine solution, a property that is satisfied by a large class of models. For example, it holds approximately in a representative agent model where agents have Epstein-Zin (1989) preferences when aggregate consumption growth is also an affine function of the state vector.[10] Schmidt (2014) shows that this property also holds in an incomplete markets setting with state-dependent higher moments of uninsurable idiosyncratic shocks.

In order to solve for asset prices, we apply a standard Campbell and Shiller (1988) log-linearization of market returns, r_{t+1}, as a function of the log-dividend growth rate, Δd_{t+1}, and the log price-dividend ratios at time $t+1$ and t, pd_{t+1} and pd_t:

$$r_{t+1} \approx \kappa + \Delta d_{t+1} + \rho \cdot pd_{t+1} - pd_t. \tag{3.4}$$

[9]For example, the property holds for affine jump-diffusion models, e.g., Eraker and Shaliastovich (2008) and Drechsler and Yaron (2011). In these models, ε_{t+1} is the sum of Gaussian and jump components, where the variance-covariance matrix for Gaussian shocks and the arrival intensities for the jump shocks are affine functions of y_t.

[10]See e.g. Bansal and Yaron (2004), Hansen, Heaton, and Li (2007,2008), Eraker and Shaliastovich (2008) and Drechsler and Yaron (2011).

Here κ and $\rho < 1$ are linearization constants. Using this linearization and assumptions 1 and 2, we can show the following result:

. Suppose Assumptions 1-2 hold. Then the following hold

(i) The market price-dividend ratio is $pd_t = A_{0,m} + A_m' z_t$;

(ii) The risk-free return from t to $t+1$ is $rf_{t+1} = A_{0,f} + A_f' z_t$;

(iii) The expected excess return is $E_t[r_{t+1}] - rf_t = \beta_0 + \beta' z_t$,

where $A_{0,m}, A_{0,f}, \beta_0$ are scalars and $A_m, A_f, \beta \in \mathbb{R}^L$.

Proposition 1 shows that the price-dividend ratio, the risk-free rate, and the expected excess return are (approximate) affine functions of the aggregate state vector.

The result motivates why a large empirical literature summarized in Goyal and Welch (2008) and Rapach and Zhou (2013) studies predictability of U.S. stock returns by means of linear regression models with constant coefficients:

$$r_{t+1} = x_t' \beta + \varepsilon_{t+1}, \tag{3.5}$$

where r_{t+1} is the period-$t+1$ stock return, measured in excess of the risk free rate, x_t is a $(d \times 1)$ vector of covariates, and ε_t is an unobservable disturbance with $\mathbb{E}[\varepsilon_t | x_t] = 0$. Provided that $x_t \sqsubseteq z_t$, Proposition 1 justifies using the linear return prediction model in equation (3.5).

3.2.2 Nonparametric Identification of Pockets

The assumption in (3.5) of constant regression coefficients has been challenged in numerous studies such as Paye and Timmermann (2006), Rapach and Wohar (2006), Chen and Hong (2012), all of which find strong statistical evidence that this assumption is empirically rejected for U.S. stock returns using standard predictor variables.

Following insights from these studies, we generalize (3.5) to allow for time-varying return predictability of the form:

$$r_{t+1} = x_t' \beta_t + \varepsilon_{t+1}, \tag{3.6}$$

where the regression coefficients β_t are now subscripted with t to indicate that they are functions of time as a means of allowing for time-varying return predictability. We also allow for general forms of conditional heteroskedasticity $\sigma_t^2 \equiv \mathbb{E}\left[\varepsilon_t^2 | x_t\right] = \sigma^2(x_t)$. The commonly used constant coefficient model (3.5) is obtained as a special case of (3.6) when $\beta_t = \beta$ for all t.

To identify periods with return predictability, we follow the nonparametric estimation strategy developed in Robinson (1989) and Cai (2007). We want to use an approach that is valid regardless of whether the linear return prediction model in (3.5) is correctly specified. Using nonparametric methods for pocket identification offers the major advantage that we do not need to take a stand on the dynamics of local return predictability, e.g., whether such predictability is short-lived or long-lived and whether it disappears slowly or rapidly. Instead, our nonparametric methods allow us to characterize the anatomy of the pockets, e.g., the duration and frequency of pockets and the degree of return predictability inside the pockets. Such characteristics can provide important clues about the economic sources of return predictability.

The nonparametric approach views $\beta : [0,1] \to \mathbb{R}^d$ as a smooth function of time that can have at most finitely many discontinuities. The problem of estimating β_t for $t = 1, \ldots, T$ can then be thought of as estimating the function β at finitely many points $\beta_t = \beta\left(\frac{t}{T}\right)$.[11]

[11] Because time, t, is normalized by the number of observations T, α is a function whose domain is $[0,1]$ as opposed to $[0,T]$. This is useful because we need more and more local information to estimate α_t consistently as $T \to \infty$.

Appendix B provides more details about how we implement the nonparametric analysis. Specifically, we use a local constant model to compute the estimator of β_t as

$$\hat{\beta}_t = \underset{\beta_0 \in \mathbb{R}^d}{\arg\min} \sum_{s=1}^{T} K_{hT}(s-t)\left[r_{t+1} - x_s'\beta_0\right]^2. \tag{3.7}$$

The weights on the local observations get controlled through the kernel $K_{hT}(u) \equiv K(u/hT)/(hT)$, where h is the bandwidth. The estimator in (3.7) can be viewed as a series of weighted least squares regressions with Taylor expansions of α around each point t/T. The weighting of observations in (3.7) can be contrasted with the familiar rolling window estimator which uses a flat kernel that puts equal weights on observations in a certain neighborhood. For this estimator $K_{hT}(s-t) = 1$ if $t \in [t - \lfloor hT \rfloor, t + \lfloor hT \rfloor]$, otherwise $K_{hT}(s-t) = 0$. A weakness of this conventional approach is that it assigns the same weight to local observations, making it less suited for picking up time variation in α if the build-up and disappearance of such patterns is more gradual, as we would expect a priori.

To identify periods with return predictability ("pockets"), we need a decision rule for determining what constitutes significant return predictability. To this end we use a bootstrap procedure to compute standard errors for the local slope coefficients, β_t, and evaluate their statistical significance. For the estimator of a particular ordinate $\hat{\beta}_t$, the estimated asymptotic variance-covariance matrix is given by:

$$\hat{\Sigma}_{\beta,t} = \frac{\kappa_2}{hT}\left(\sum_{s=1}^{T} K_{hT}(s-t)\,\hat{e}_s^2\right)\left(\sum_{s=1}^{T} K_{hT}(s-t)\,x_s x_s'\right)^{-1}, \tag{3.8}$$

where $\kappa_2 \equiv \int_0^1 K^2(u)\,du$. The limiting distribution is normal and thus a valid 95%

pointwise confidence interval for the i_{th} element of $\hat{\beta}_t$, $\hat{\beta}_{i,t}$, is given by

$$\left[\hat{\beta}_{i,t} - z_{(1-\gamma)/2}\hat{\Sigma}^{1/2}_{\beta,t}(i,i), \hat{\beta}_{i,t} + z_{(1-\gamma)/2}\hat{\Sigma}^{1/2}_{\beta,t}(i,i)\right], \tag{3.9}$$

where $z_{(1-\gamma)/2}$ is the $(1-\gamma)/2$ quantile of the standard normal distribution.

To quantify the degree of local return predictability, we compute a measure of the local R^2 at time t, R^2_t:

$$R^2_t = 1 - \frac{\sum_{s=1}^{T} K_{hT}(s-t)\hat{e}^2_s}{\sum_{s=1}^{T} K_{hT}(s-t)y^2_s}, \tag{3.10}$$

where $\hat{e}_s = r_s - x'_{s-1}\hat{\beta}_{s-1}$ is the residual at time s obtained from the nonparametric regression. To identify local variations in the regression coefficients of our model (3.6), we use a two-sided Epanechnikov Kernel and an effective sample size of one year, i.e., six months of data before and six months after each observation. The Epanechnikov Kernel function has an inverted parabola shape and takes the form

$$K(u) = \frac{3}{4}\left(1-u^2\right)1\{|u| \leq 1\}. \tag{3.11}$$

Thus, for each day in the sample, we estimate nonparametrically the return prediction model in (3.6) after trimming the first and last six months of the data. At each point we test if the local slope coefficient is significantly different from zero (using a two-sided test), assigning a value of unity to the pocket indicator $\mathscr{I}_t = 1\{\left|\hat{\beta}_t / se(\hat{\beta}_t)\right| > c\}$, where c is a cutoff value that determines the size of the test.

The overlap in adjacent windows (kernel weighting schemes) for nearby dates t, t' yields a sequence of highly correlated test statistics. Moreover, repeating the test multiple times can be expected to generate false rejections and identify evidence of spurious return predictability. We address this concern in Section 3.3 by simulating from different data generating processes for returns and addressing to what extent different

models can match the characteristics of the pockets of predictability identified by our methodology. Pocket characteristics are measured in a variety of ways. At the most basic level, we want to know how many contiguous pockets our procedure detects. We refer to this as N_p. Second, it is of interest to ask how long the pockets last. To this end, let $\mathscr{I}_{jt} = 1$ for time-series observations inside the jth pocket, while $\mathscr{I}_{jt} = 0$ outside pockets. Letting t_{0j} and t_{1j} be the start and end date of the jth pocket, the duration of pocket j is defined as

$$Dur_j = \sum_{\tau=1}^{T} \mathscr{I}_{j\tau} = t_{1j} - t_{0j} + 1, \quad j = 1, ..., N_p. \tag{3.12}$$

Presumably, it is easier for investors to detect and exploit long-lived pockets as the power of any tests for the presence of pockets grows with the length of the pocket. We characterize the distribution of pocket durations by reporting the mean, minimum, and maximum durations and also report the fraction of observations inside a pocket, i.e., $\sum_{j=1}^{N_p} Dur_j/T$, where T is the total sample size.

Pocket durations do not shed light on the magnitude of the (local) predictability within a pocket. This matters a great deal because investors are more likely to identify local predictability if the R^2 is high. To get at this, we compute a measure of the average R^2 within each pocket as

$$\bar{R}_j^2 = \frac{\sum_{\tau=1}^{T} \mathscr{I}_{j\tau} R_\tau^2}{Dur_j}. \tag{3.13}$$

We report the average R^2 value across all pockets, $\bar{\bar{R}}^2 = \sum_{j=1}^{N_p} Dur_j \bar{R}_j^2 / \sum_{j=1}^{N_p} Dur_j$, as well as the maximum value of the average $\bar{R}^2$, computed across all pockets, $Max_{j=1,...,N_p} \left\{ \bar{R}_j^2 \right\}$.

The duration and R^2 values measure the length and magnitude of spells with predictability. However, they do not quantify the total amount of predictability which accounts for both the duration and the magnitude of the local predictability. We capture

this by means of the integral R^2 measure defined as

$$IR_j^2 = \sum_{\tau=t_{0j}}^{t_{1j}} R_\tau^2 = \sum_{\tau=1}^{T} \mathscr{I}_{j\tau} R_\tau^2. \tag{3.14}$$

Visually, this measure captures the area marked under a time-series plot of the local R_τ^2 values, summed across each of the pocket indicators. We report the mean, minimum and maximum values of IR_j computed across the pockets $j = 1,...,N_p$. Pockets are more detectable either when the degree of predictability within a pocket is very high, possibly for a brief period of time, or when a pocket lasts long (even with low average predictability), or both. By combining the duration of a pocket with the magnitude of the predictability inside this pocket, the integral R^2 measure provides both economic insights into how much predictability is present as well as the possibility that investors can detect this predictability.[12]

3.3 Empirical Results

This section introduces our daily data on stock returns and a set of predictor variables, presents evidence from applying the non-parametric approach to identifying local return predictability pockets, and finally, tests whether this evidence is consistent with simple models of stock market returns.

3.3.1 Data

Most studies on return predictability use monthly, quarterly, or annual returns data. However, since we are concerned with local return predictability which may be of a relatively short-lived nature, we use daily data on both stock returns and the predictor

[12]Note the analogy to the integral R^2 measure from the literature on breakpoint testing which finds that tests for breaks cannot easily distinguish between frequent, but small breaks to parameters versus rare, but large breaks that move the parameters by the same distance over a particular sample. See Elliott and Müller (2006).

variables. Data observed at the standard frequencies may miss episodes with return predictability at times when the slope coefficients (β_t) change relatively quickly.

In all return regressions, the dependent variable is the value-weighted CRSP US stock market return minus the one-day return on a short T-bill rate. These data are taken from Ken French's website.

Following studies such as Goyal and Welch (2008), Dangl and Halling (2012), Johannes, Korteweg, and Polson (2014), and Pettenuzzo, Timmermann, and Valkanov (2014), our main empirical analysis considers univariate prediction models and so only includes one time-varying predictor at a time, i.e., $r_{t+1} = x_t\beta_t + \varepsilon_{t+1}$. The univariate approach is particular well suited to our nonparametric analysis which benefits from keeping the dimensionality of the set of predictors low.

Specifically, we consider five variables that have been used in numerous studies on return predictability and are included in the list of predictors considered by Goyal and Welch (2008). First, we use the lagged dividend yield, defined as dividends over the most recent 12-month period divided by the current stock price. This predictor has been used in studies such as Keim and Stambaugh (1986), Campbell (1987), Campbell and Shiller (1988), Fama and French (1988), Fama and French (1989), and many others to predict stock returns. Second, we consider the yield on a 3-month Treasury bill. Campbell (1987) and Ang and Bekaert (2007) use this as a predictor of stock returns. Next, we use the term spread, defined as the difference in yields on a 10-year Treasury bond and a three month Treasury bill and the corporate default spread, defined as the yield differential between BAA and AAA rated corporate bonds.[13] Finally, we also consider a realized variance measure, defined as the realized variance over the previous 60 days. Following Merton's work on the ICAPM, the conditional volatility has been used as a predictor of

[13] See Keim and Stambaugh (1986) and Welch and Goyal (2008) for studies using these predictors.

stock returns.[14]

The data samples vary across these predictor variables and begin in 11/4/1926 for the dividend yield (22,778 observations), 1/4/1954 for the 3-month T-bill rate (14,852 obs.), 1/2/1962 (12,838 obs.) for the term spread, 1/2/1986 (6,808 obs.) for the corporate spread and 1/15/1927 (22,719 obs.) for the realized variance. The final sample date for all series is 12/31/2012.

On economic grounds, we would expect return predictability to be very weak at the daily horizon. Table 1 confirms that this is indeed the case. The table shows full-sample coefficient estimates obtained from the linear regression model in (3.5) along with t-statistics and R^2 values. Only the regressions that use the T-bill rate (t-statistic of -2.72) and the term spread (t-statistic of 2.4) generate statistically significant slope coefficients. As expected, there the average predictability is extremely low at the daily frequency with in-sample $\bar{R}^2$ values varying from -0.014% for the default spread to a maximum of 0.056% (i.e., 0.00056) for the regression that uses the T-bill rate as a predictor.

Campbell and Thompson (2008) suggest comparing the R^2 to the squared Sharpe ratio to get a measure of the economic value of return predictability. For our daily data, the Sharpe ratio is 0.0255 and so the squared Sharpe ratio is $S^2 = 0.00065$. Using equations (13) and (14) in Campbell and Thompson (2008), the in-sample R^2 value for the dividend yield regression translates into a gain of 0.42% in the return of a mean-variance investor with a coefficient of risk aversion of three or, equivalently, a 7% proportional increase in the investor's utility.[15] Even ignoring the fact that these are in-sample estimates and omit

[14]The daily predictor variables are highly persistent at the daily frequency, posing challenges for estimation and inference with daily data. We therefore detrended the predictors by subtracting a 6-month moving average. This is a common procedure for variables such as the nominal interest rate even at longer horizons such as monthly data, see, e.g., Ang and Bekaert (2007). However, we found that the results did not change very much due to this type of detrending and so decided to go with the simpler approach of using the raw data.

[15]These numbers are computed by comparing the expected return of an investor with access to the

any transaction costs (and trading limits) associated with exploiting the prediction signals, this shows that there would not have been great economic benefits to investors from exploiting daily return predictability from the dividend yield. Notably bigger values are seen for the regression based on the T-bill rate for which the R^2 value of 0.056 translates into an increase in the expected return of 4.7% per annum (assuming again a coefficient of risk aversion of three) or, equivalently, an 86% proportional increase in the investor's expected return. We emphasize again that these are not feasible gains and instead should be viewed as an upper bound on the economic value of the daily return predictability signals from the constant coefficient regression model.

Figures 1-5 provide graphical illustrations of the pockets identified by our non-parametric procedure. Each figure covers a different predictor variable. The top panel in each figure plots time series of non-parametric kernel estimates of the local slope coefficient ($\hat{\beta}_t$) from regressions of daily excess stock returns on the lagged predictors. Dashed lines surrounding the solid line represent plus or minus two standard error bands. The bottom panel in each figure plots the local R^2 measure against time. Shaded areas underneath the local R^2 curve represent the integral R^2 measure for periods identified as pockets of predictability. Using a methodology described below, areas colored in red represent pockets that have less than a 5% chance of being spurious, areas colored in orange represent pockets that have between a 5% and a 10% chance of being spurious, and areas colored in yellow represent pockets with more than a 10% chance of being spurious. We comment more on this below.

First consider the predictability plots for the dividend yield predictor, shown in Figure 1. The plots for this variable indicate 13 separate pockets with significant return predictability. The two longest pockets occur during the second world war and around the Korean War. Moreover, both the number and average duration of the pockets has

(in-sample) predictions relative to the return of the same investor who assumes a constant expected return.

come down over time with only five pockets appearing after 1970 and no pocket showing up in the last 25-years of our sample, i.e., after 1987. For all but two short-lived pockets, the coefficient on the dividend yield is positive inside the pocket. Inside pockets, the R^2 goes as high as 0.04 in the pocket in 1954. but mostly hovers substantially below this level at around 0.01.

For the T-bill rate predictor (Figure 2), we identify seven pockets, only one of which occurs after 1990. Unlike the plots for the dividend yield–and consistent with existing studies such as Ang and Bekaert (2008)–the local coefficient estimates for the T-bill rate are mostly negative, the only exception being the pocket in 1995. Notably, our nonparametric T-bill rate model identifies the episode in 1973-74 with negative excess returns. The local R^2-values exceed 0.02 during three of these episodes, but are very low during most of the remaining sample, including the period after 2000 which saw low and downward trending interest rates.

The plots for the term spread in Figure 3 identify three pockets–all with positive coefficients–in 1969, 1973-74, and in 1981-82. Interestingly, the last pocket coincides with the changes to the Federal Reserve's operating procedures during the monetarist experiment in 1979-1982 which led to significantly higher and more volatile interest rates. The local R^2 is notably higher during these three episodes, ranging between 0.015 and 0.025.

The corporate spread regression in Figure 4 only identifies two pockets–a fairly long-lived pocket in 1990 and a shorter one in 1995. In both cases, the local R^2 is around 0.01.

The plots for the realized variance, shown in Figure 5, identify eight pockets. Interestingly, whereas the estimated coefficients on this variable are negative during the four pockets identified in the first half of the sample up to 1968, they switch sign and become positive in the three longest pockets identified in the second half of the sample.

This behavior is consistent with the difficulty the finance literature has experienced in establishing a consistently positive risk-return trade-off.

3.3.2 Anatomy of Pockets

Having illustrated the presence of pockets with return predictability, we next move on to study the properties of such pockets in more detail. To this end, the first five columns in Table 2 show statistics on the number of pockets identified by our methodology. This includes their minimum and maximum lengths, the average pocket lengths, and the fraction of the total sample for which a pocket is identified. Results in Panel A use a 5% significance level to identify pockets, while results in Panel B use a 1% significance level.

The length of the pockets varies significantly, even for a particular forecasting model. For example, using a 5% significance level to identify pockets (Panel A), the model based on the dividend yield finds a pocket that lasts only 24 days (a little more than one month) while the longest pocket lasts 876 days, or more than three years. Similarly, if less extreme, variations in pocket length are observed for the other predictor variables. The average duration of periods with return predictability varies from 126 days (five months) for the corporate spread variable to 378 days (18 months) for the term spread.

As seen from the plots in figures 1-5, the number of pockets identified by our approach also varies substantially across predictors–from 13 for the dividend yield model to only 2 for the corporate spread. This reflects differences in both the sample length (the sample is much longer for the dividend yield than for the corporate spread), but also differences in the proportion of the sample spent inside pockets. If we use a 5% test size and repeat the test multiple times, by random chance we should expect to find pockets 5% of the time. In fact, for both the dividend yield and T-bill rate predictors, 11 percent of the sample is spent inside pockets–over twice the rate we would expect by chance.

We also find a higher-than-expected pocket frequency for the term spread (9%) and the realized variance (8.3%) predictors, but a lower-than-expected rate for the default spread (3.8%).

Comparing the periods spent inside pockets (columns 1-5) to periods spent outside pockets (columns 6-10), we find that the average duration of spells outside pockets is far greater than that spent inside pockets. This is, of course, a reflection of the fact that most of the time (at least 88% of the sample) is spent outside pockets, but the duration measures for the "out-of-pocket" episodes serve the purpose of showing that there are decade-long periods with no significant return predictability.

Panel B repeats the analysis in Panel A, now using a significance level of 1%. The advantage of using this more stringent level of significance is that it is likely to trigger fewer cases of "spurious" pockets due to the repeated use of the test statistic. Although the number of pockets, as well as their average and maximum length decline, we see continued strong evidence of pockets even for this more stringent threshold. For the T-bill rate, the term spread and the realized variance predictors, pockets occupy 4.3%, 4.9%, and 4.2% of the sample, respectively. This is four times higher than the frequency (1%) expected due to the repeated use of the pocket test statistic. For the dividend yield variable, 2.8% of the sample is spent inside pockets, whereas the model based on the default spread no longer identifies any pockets.

Panels C and D report sample statistics on the mean, standard deviation, skewness, kurtosis and persistence of returns inside the predictability pockets identified by our methodology (left columns) as well as outside the pockets (right columns). Focusing on the results based on the 5% cutoff (Panel C), return distributions of stock returns inside versus outside the pockets can differ by large amounts. For example, the daily mean return inside the pockets identified by the dividend yield predictor is 4.7 basis points (bps) per day which is nearly twice as high as outside the pockets (2.4 bps). Even larger

differences are observed for the pockets identified by the T-bill rate and the term spread predictors, for which we observe negative mean returns (-2.5 bps) in the pockets, but positive means (3.0 and 2.7bps, respectively) outside the pockets.

Returns inside the pockets also tend to be less volatile (with the exception of pockets identified by the term spread) with positive skews for four out of the five predictors (the exception being the realized variance predictor). The positive skews inside pockets contrast with the large negative skews observed outside pockets. Kurtosis is also markedly smaller inside the pockets than outside for four of five variables. This suggests that the pockets overall have lower risk than during normal periods.

We conclude from these results that return predictability varies significantly over time. Our nonparametric regression approach detects local pockets of return predictability and the return distribution appears to be quite different inside versus outside such pockets. Of course, we have not yet conducted any formal inference on these findings–a topic we turn to next.

3.3.3 Evaluating the Statistical Significance of the Results

Because we use a new approach for identifying local return predictability, it is worth further exploring its statistical properties. For example, we are interested in knowing to what extent our approach spuriously identifies pockets of return predictability. Since our approach repeatedly computes local (overlapping) test statistics, we are bound to find evidence of some pockets even in the absence of any genuine return predictability. The question is whether we find more pockets than we would expect by random chance, given a reasonable model for the daily return dynamics. Another issue is whether shorter pockets are more likely to be spurious than the longer ones and whether the degree of return predictability (as measured by the local R^2) inside pockets is consistent with standard models for return dynamics.

To address these questions, we consider two different models for return dynamics. Our simplest model assumes a random walk model with a drift for stock prices and so takes the form

$$r_{t+1} = \mu + \varepsilon_{t+1}. \tag{3.15}$$

To allow returns to follow a non-Gaussian distribution, we draw the zero-mean innovations, ε_{t+1}, by means of an i.i.d. bootstrap. This is clearly not a very good model for daily stock returns, but it serves as a benchmark that allows us to gauge the importance of adding more realistic features of return dynamics.

Specifically, we need to account for the pronounced time-varying volatility in daily returns. To this end, we estimate a GARCH(1,1) model which has been used extensively to characterize stock market volatility. In addition to allowing for volatility dynamics in returns, we allow for (constant) return predictability from a time-varying state variable, x_t, whose volatility is also time-varying, so that the model we simulate from takes the form

$$\begin{aligned}
r_{t+1} &= \gamma x_t + \varepsilon_{rt+1} \equiv \gamma x_t + \sqrt{h_{rt}}u_{rt+1}, \;\; u_{rt+1} \sim (0,1), \\
h_{rt+1} &= \omega + \alpha_1 \varepsilon_{rt}^2 + \beta_1 h_{rt}, \\
x_t &= \rho x_{t-1} + \varepsilon_{xt} \equiv \rho x_{t-1} + \sqrt{h_{xt}}u_{xt}, \quad u_{xt+1} \sim (0,1), \\
h_{xt+1} &= \omega_x + \alpha_x \varepsilon_{xt}^2 + \beta_x h_{xt},
\end{aligned} \tag{3.16}$$

where u_{rt+1} and u_{xt+1} are mutually independent. The specification in (3.16) is very flexible: We allow for time-varying volatility both in the return shocks and in the predictor variable and the shocks to returns and the predictor variable can be correlated. This constant-coefficient specification is of particular interest for two reasons. First, because it is the conventional return predictability model used in the empirical literature.

Second, it is implied by many asset pricing specifications as shown in Proposition 1. The GARCH(1,1) model allows for the possibility that local pockets of return predictability could arise due to periods with large variations in the predictor variable.

To simulate from the model in (3.16), we first estimate the parameters $\gamma, \omega, \alpha_1, \beta_1$, ρ, ω_x, α_x and β_x by fitting GARCH(1,1) models to daily values of excess returns and the predictors. Using these estimates, we next construct values of x_t as $\hat{\rho} x_{t-1} + \sqrt{\hat{h}_{xt}}\hat{u}_{xt}$, where $\hat{h}_{xt}$ is the fitted variance of x_t from a GARCH(1,1) model and $\hat{u}_{xt}$ is obtained by bootstrapping (with replacement) from the normalized residuals of the x process. Finally, we construct a series of conditional variances $\left\{\hat{h}_{rt+1}\right\}_{t=0}^{T-1}$ and obtain normalized residuals $\{\hat{u}_{rt+1}\}_{t=0}^{T-1}$, where $\hat{u}_{rt+1} = r_{t+1} - \hat{\gamma} x_t / \sqrt{\hat{h}_{rt+1}}$. Next, we construct 1,000 bootstrap samples by first drawing $T+1$ bootstrap residuals $\left\{u_{rt}^b\right\}_{t=0}^{T}$ at random from $\{\hat{u}_{rt+1}\}_{t=0}^{T-1}$ with replacement. We then construct a bootstrap sample of returns $\left\{r_{t+1}^b\right\}_{t=0}^{T-1}$ from (3.16), with $h_0^b = \hat{\omega}/(1-\hat{\alpha}_1 - \hat{\beta}_1)$.

For both specifications (3.15) and (3.16) we use the available sample to fit the model and estimate parameters such as the persistence of the predictor variable (x_t). Our simulations follow the empirical analysis and define pockets as periods where the estimated coefficient on the lagged predictor variable is found to be significant at the 95% level. For each bootstrap sample, we record the number of such pockets, along with the minimum, maximum and average values for the pocket duration (measured in days), the R^2 and the integral R^2, described in equations (3.12)-(3.14), along with the fraction of time spent inside pockets, measured as a proportion of the full sample.

Table 3 shows results for the actual data (first column) and the bootstrapped average, standard errors and p-values–the latter computed as the proportion of simulations that generate a statistic as large as or bigger than that found in the actual data. Columns two through four assume the simple return generating model in (3.15), while columns five through seven present results for the extended GARCH model in (3.16).

First consider the results for the model that uses the dividend yield as a predictor variable (Panel A). On average there are 6.6 pockets in the simulations as compared with 13 in the actual data and this difference is statistically significant: Only 1.1% of the simulations generate at least 13 pockets. The simulated data also fail to match the maximum pocket duration but can match both the minimum and average duration statistics. A similar pattern arises for the statistics based on the R^2 measure for which the simulated data match the minimum and average R^2 value within pockets but not its maximum value. The simulated data also fail to match both the average and maximum values of the integral R^2 and the fraction of the sample inside pockets which is 11% in the actual data and close to 5% in the simulations.

For many of the measures of local return predictability, similar patterns are found for the other predictor variables: whereas simulations based on the benchmark specifications in (3.15) and (3.16) can generate the same number of pockets as in the original sample and also match the shortest pocket duration and the minimum IR^2, they have a much harder time matching the mean or maximum IR^2 value or the maximum R^2. Results do vary for the remaining statistics, e.g., the maximum pocket length which is matched for the pockets identified by the T-bill rate, default spread and realized variance predictors but not for the term spread.

Interestingly, the simulated data can match all the test statistics for the pockets identified by the corporate spread variable (Panel D). This suggests that the pockets identified by this predictor are consistent with one of the simple return prediction models in (3.15) and (3.16).

Looking across the different benchmark specifications, it does not make a big difference to most of the results if the random walk with a constant expected return or the GARCH model with a constant slope coefficient is used in the simulations.

These simulations suggest that the shortest predictability pockets can be due to

"chance" as they are likely to occur in simulations with zero coefficients; whereas neither models with constant expected returns or a constant slope coefficient and time-varying volatility come close to matching the amount of predictability observed in the longer-lived pockets.

Table 4 repeats the analysis, but now defines the pockets relative to a constant coefficient benchmark. Hence the results in this table are testing if there is evidence of significant time variation in the slope coefficient of the predictors, again measured relative to a constant, non-zero baseline. The number of pockets should now be thought of as contiguous periods with evidence of significant time variation in the slope coefficient. Because the coefficients can move either up or down relative to the baseline specification, there is no mechanical relation between the number of pockets–or the time spent inside pockets–in Table 4 versus the number of pockets reported in Table 3. However, we do see that the two numbers are broadly similar with most pockets identified for the dividend yield (16 pockets) and fewest for the corporate spread (one pocket).

While the duration statistics can be matched by the simulations, the mean and maximum values of the integral R^2, the maximum R^2, as well as the fraction of days with a significant pocket indicator cannot, in most cases, be matched. This is evidence of significant time variation in the regression coefficients of the univariate return prediction models and evidence against the class of affine asset pricing models as shown in Proposition 1.

3.3.4 Separating Spurious from Non-Spurious Pockets

We previously discussed the concern that our local, non-parametric approach may detect spurious pockets due to the repeated use of tests based on overlapping data. This naturally raises the question whether we can tell if some of the pockets identified by our approach are more or less likely to be spurious. Our finding from Table 3 that the

simulations can match some properties of the pocket distribution but not others, suggests that we can discriminate between spurious and non-spurious pockets by looking at each individual pocket's integral R^2 value–a measure found to be particularly hard to match in the simulations–and computing the percentage of simulations with at least one pocket matching this value. This measure gives an odds ratio with small values indicating how difficult it is to match the total amount of predictability observed for each particular pocket.

Following this idea, for each of the pockets identified in Table 2, Table 5 reports the associated integral R^2 measure and the probability that this value is matched by at least one pocket in a simulation. First consider the 13 pockets identified by the return prediction models that use the dividend yield as a predictor (first column). Some of the pockets are highly unlikely to be due to chance–for example, the third and fifth pockets generate integral R^2 values around 13 and not a single of the simulations is able to match these high values. Other pockets, notably the second, seventh, tenth, and twelfth, are more likely to be spurious as their integral R^2 values are matched in at least five percent of the simulations.

All of the seven pockets identified by the T-bill rate generate integral R^2 values less than 5%–most substantially smaller even than 1%, and similar conclusions hold for the three pockets identified by the term spread and the realized variance (with two exceptions).

Using the analysis in Table 5, Figures 1-5 mark in red the integral R^2 of pockets with less than a 5% chance of being spurious, pockets colored in orange have between a 5% and a 10% chance of being spurious, while pockets colored in yellow have more than a 10% chance of being spurious. As expected, pockets that are more short-lived and have lower peaks in the R^2 are more likely to be deemed spurious.

These results suggest that, with the exception of some of the pockets identified

by the model that uses the dividend yield or the default spread as a predictor variable, the vast majority of pockets of predictability identified by our local kernel approach are not spurious and, so, represent periods where returns are genuinely predictable.

3.4 Learning About Cash Flows

This section explores whether the evidence of local pockets with return predictability is consistent with learning dynamics induced by an asset pricing model where expected returns are constant but the cash flow process is partially predictable. Specifically, we propose a new specification for cash flow dynamics that builds on, and generalizes, the predictive systems approach pioneered by Pástor and Stambaugh (2009). As in the predictive systems approach, we assume that cash flows consist of an unobserved expected growth component that is highly persistent and a temporary "unexpected growth" shock. This latent process is correlated with a set of observable state variables which, through their correlation with expected growth, gain predictive power over future cash flows. We argue that the correlation between expected cash flows and many of the conventional predictor variables used in the return predictability literature–notably nominal interest rates and interest rate spreads–is likely to vary over time and we model such variation through a regime switching process. A likely source of such shifts in correlations is changes in monetary policy and these can be discrete, as captured by the regime switching process. Thus, how informative an observed predictor variable is with regards to the underlying (unobserved) cash flow growth process is likely to vary over time. Sometimes the economic state variable has strong predictive power over future cash flows, and at other times it does not. Our approach captures such variation.

3.4.1 A Predictive Systems Model with Regime Switching

We develop a model for the dividend process that captures a small predictable component in daily cash flows. Specifically, let $\Delta d_{t+1} = \log(D_{t+1}/D_t)$ be the growth rate in (log-) dividends and assume that this can be decomposed into an expected cash flow component, μ_t, and a purely temporary shock, u_{t+1} :

$$\Delta d_{t+1} = \mu_t + u_{t+1}, \tag{3.17}$$

The mean of the expected cash flow process, μ_t, is affected by a state variable, s_t, which captures discrete shifts to the process. Moreover, μ_t can be persistent as captured by an autoregressive component. Finally, expected cash flows are affected by a transitory shock, w_{t+1} :

$$\mu_{t+1} = \mu_{\mu,s_{t+1}} + \rho_\mu \mu_t + w_{t+1}. \tag{3.18}$$

While investors do not observe the expected cash flow process, they observe a predictor variable, x_{t+1}, that is affected by the same state variable, s_{t+1}, and follows a similar dynamic process:

$$x_{t+1} = \mu_{x,s_{t+1}} + \rho_x x_t + v_{t+1}. \tag{3.19}$$

We assume that the innovations to the processes in (3.17) - (3.19) are normally distributed with mean zero $(u_{t+1} w_{t+1}, v_{t+1})' \sim N(0, \Sigma_{s_{t+1}})$, where $\Sigma_{s_{t+1}}$ is a state-dependent variance-covariance matrix:

$$\Sigma_{s_t} = \begin{bmatrix} \sigma_u^2 & \sigma_{uv} & \sigma_{uw} \\ \sigma_{uv} & \sigma_{v,s_t}^2 & \sigma_{vw,s_t} \\ \sigma_{uw} & \sigma_{vw,s_t} & \sigma_{w,s_t}^2 \end{bmatrix}. \tag{3.20}$$

Note that we constrain this covariance matrix to have a very particular form as only the variance of the expected cash flow (σ_{w,s_t}^2) and predictor variable (σ_{v,s_t}^2), in addition to

their cross-moment (σ_{vw,s_t}), are state dependent. In contrast, the variance of the purely temporary shocks to dividend growth (σ_u^2), or their correlation with the other shocks in the model (σ_{uv}, σ_{uw}), do not depend on the underlying state variable. We impose these constraints to ensure that the identified states capture changes to how informative the predictor variable, x_{t+1}, is over the expected cash flow process, μ_{t+1}.

We focus on the case with two states so that $s_t \in \{1,2\}$ and assume that s_t follows a first-order homogenous Markov chain with transitions

$$\pi_{ii} = prob(s_{t+1} = i | s_t = i), \quad i = 1,2. \tag{3.21}$$

We further assume that s_t is independent of μ_t and collect the state transitions in a 2×2 transition probability matrix Π_s.

Using results from Timmermann (2000), the unconditional mean and variance of the μ_t process are given by[16]

$$\mathbb{E}[\mu_t] = \overline{\pi}' \boldsymbol{\mu}_{\mu,s}, \tag{3.22}$$

$$Var(\mu_t) = \overline{\pi}' \left(\left(\boldsymbol{\mu}_{\mu,s} - \mu_\mu \boldsymbol{\iota} \right) \odot \left(\boldsymbol{\mu}_{\mu,s} - \mu_\mu \boldsymbol{\iota} \right) + \frac{\boldsymbol{\sigma}_{w,s}^2}{1-\rho_\mu^2} \right), \tag{3.23}$$

where

$$\boldsymbol{\mu}_{\mu,s} = \begin{bmatrix} \frac{\mu_{\mu,1}}{1-\rho_\mu} \\ \frac{\mu_{\mu,2}}{1-\rho_\mu} \end{bmatrix}, \quad \boldsymbol{\sigma}_{w,s} = \begin{bmatrix} \sigma_{w,1} \\ \sigma_{w,2} \end{bmatrix}, \quad \overline{\pi} = \begin{bmatrix} \mathbb{P}(s_t = 1) \\ \mathbb{P}(s_t = 2) \end{bmatrix}.$$

are the steady-state mean, volatility and ergodic state probabilities, respectively.

[16] This result uses that the μ_t process can equivalently be re-written as

$$\mu_{t+1} = \mu_{\mu,s_{t+1}} + \rho_\mu \mu_{\mu,s_t} + \rho_\mu \left(\mu_t - \mu_{\mu,s_t} \right) + w_{t+1}.$$

3.4.2 Filtering the State Variables and Evaluating the Likelihood

The predictive systems model in (3.17) - (3.19) is a nonlinear state space model as it contains a combination of linear and regime-switching dynamics, and thus standard Kalman filtering cannot be used to filter the states and evaluate the likelihood. To address this issue, we approximate the likelihood function using a discretization of μ_t as proposed in Farmer (2017). We briefly explain how this is done.

In the first step, we construct a discrete approximation to the stochastic process governing the dynamics of the state variables, μ_t and s_t. Because the shocks to the measurement and state equations are correlated, the distribution of the state in the next period conditional on the state in the current period depends on the specific values of the observables and so the transition matrix constructed to approximate the dynamics will be time-varying. We handle this issue as follows. Using properties of correlated normal random variables, we have

$$w_t \left| u_t, v_t, s_t \right. \sim N\left(\mu_{cond,t}, \sigma^2_{cond,t}\right),$$

where $\mu_{cond,t}$ and $\sigma_{cond,t}$ are given by

$$\mu_{cond,t} = \frac{\left(\sigma_{uw}\sigma^2_{v,s_t} - \sigma_{vw,s_t}\sigma_{uv}\right)u_t + \left(\sigma_{vw,s_t}\sigma^2_u - \sigma_{uw}\sigma_{uv}\right)v_t}{\sigma^2_u\sigma^2_{v,s_t} - \sigma^2_{uv}},$$
$$\sigma^2_{cond,t} = \sigma^2_{w,s_t} - \frac{\sigma^2_{uw}\sigma^2_{v,s_t} + \sigma_{vw,s_t}\sigma^2_u}{\sigma^2_u\sigma^2_{v,s_t} - \sigma^2_{uv}}.$$

Next, define a new random variable $\mu_{t,M}$ which takes M discrete values $\left(\mu^1, \ldots, \mu^M\right)$. For a given choice of M, define a grid from the set of M equally spaced points between $\mathbb{E}\left[\mu_t\right] - \sqrt{(M-1)\operatorname{Var}\left(\mu_t\right)}$ and $\mathbb{E}\left[\mu_t\right] + \sqrt{(M-1)\operatorname{Var}\left(\mu_t\right)}$. Let each point μ^m be associated with the interval $\left[\underline{\mu}^m, \overline{\mu}^m\right]$ where $\underline{\mu}^1 = -\infty$, $\overline{\mu}^M = \infty$, and $\overline{\mu}^m = \underline{\mu}^{m+1} = \frac{\mu^m + \mu^{m+1}}{2}$ for $m = 1, \ldots, M-1$. Construct the transition probabilities for

$\mu_{t,M}$ at time t as

$$\begin{aligned}
&\mathbb{P}\left(\mu_{t+1,M} = \mu^{m'} | \mu_{t,M} = \mu^{m}, u_{t+1}, v_{t+1}, s_{t+1}\right) \\
=&\Phi\left(\frac{\overline{\mu}^{m'} - \mu_{\mu,s_{t+1}} - \rho_{\mu}\mu^{m} - \mu_{cond,t+1}}{\sigma_{cond,t+1}}\right) \\
-&\Phi\left(\frac{\underline{\mu}^{m'} - \mu_{\mu,s_{t+1}} - \rho_{\mu}\mu^{m} - \mu_{cond,t+1}}{\sigma_{cond,t+1}}\right)
\end{aligned} \tag{3.24}$$

where Φ is the standard normal CDF.

With two variance regimes, we get a total of $2 \times M$ states in the discrete chain. The ordering convention for the states is

$$\psi = \begin{bmatrix} \mu_1 & s_1 \\ \mu_2 & s_1 \\ \vdots & \vdots \\ \mu_M & s_1 \\ \mu_1 & s_2 \\ \vdots & \vdots \\ \mu_M & s_2 \end{bmatrix} \tag{3.25}$$

The state probabilities, $\hat{\xi}_{t|t}$, can therefore be represented by a $(2 \times M) \times 1$ vector whose individual entries refer to the probability of being in each of the state pairs listed in the ψ matrix above. Because the innovations to the state and observation equations are correlated, the discrete Markov chain is non-homogeneous with transition matrix at time t given by Π_t.

The forecast of next period's state is given by

$$\hat{\xi}_{t+1|t} = \Pi_t \hat{\xi}_{t|t}, \tag{3.26}$$

while the state probabilities are updated recursively as follows

$$\hat{\xi}_{t+1|t+1} = \frac{\hat{\xi}_{t+1|t} \odot \eta_{t+1}}{1'\left(\hat{\xi}_{t+1|t} \odot \eta_{t+1}\right)}, \tag{3.27}$$

where 1 is a $(2 \times M) \times 1$ vector of ones and η_{t+1} denote the joint conditional densities

$$\eta_{t+1} = \begin{bmatrix} p\left(u_{t+1}, v_{t+1} \middle| \mu_t = \mu^1, s_{t+1} = 1\right) \\ \vdots \\ p\left(u_{t+1}, v_{t+1} \middle| \mu_t = \mu^M, s_{t+1} = 1\right) \\ p\left(u_{t+1}, v_{t+1} \middle| \mu_t = \mu^1, s_{t+1} = 2\right) \\ \vdots \\ p\left(u_{t+1}, v_{t+1} \middle| \mu_t = \mu^M, s_{t+1} = 2\right) \end{bmatrix}.$$

From the time series $\left\{\hat{\xi}_{t|t}\right\}_{t=1}^{T}$ we can construct filtered estimates of μ_t as

$$\hat{\mu}_{t|t} = (\psi \iota_1)' \hat{\xi}_{t|t} \tag{3.28}$$

where ι_1 is the first column of the (2×2) identity matrix.

Lastly, the log likelihood of the approximate predictive system model can be constructed as

$$\ell_{T,M}(\theta) = \frac{1}{T} \sum_{t=1}^{T} \log\left[1'\left(\hat{\xi}_{t+1|t} \odot \eta_{t+1}\right)\right]. \tag{3.29}$$

We next discuss how the parameters of the model are calibrated using daily data on dividend growth and the T-bill rate.

3.4.3 Asset Prices and Returns

We next develop a simple present value model for pricing stocks under the assumption that dividends follow the predictive systems model in (3.17) - (3.19). To this end, we use a simple log-linearized present value model. Following Campbell and Shiller (1988), the present value stock price can be written as

$$p_t = d_t + \frac{c}{1-\rho} + E_t\left[\sum_{j=0}^{\infty} \rho^j [\Delta d_{t+1+j} - r_{t+1+j}]\right], \tag{3.30}$$

where p_t and d_t denote the log of the stock price and dividends, respectively, Δd_{t+j} and r_{t+j} are the dividend growth rate and (log-) returns in period $t+j$, and c, ρ are (linearization) constants. Under the assumption that expected returns are constant, $E_t\left[r_{t+j}\right] = \bar{r}$ for all j, (3.30) simplifies to

$$p_t = d_t + \frac{c-\bar{r}}{1-\rho} + E_t\left[\sum_{j=0}^{\infty} \rho^j \Delta d_{t+1+j}\right]. \tag{3.31}$$

Thus, calculating the stock price in (3.31) only requires us to compute the expected future dividend growth, Δd_{t+1+j} for $j \geq 0$. Recall from (3.28) that the filtered state estimate of μ_t is given by $\hat{\mu}_{t|t} = (\psi \iota_1)' \hat{\xi}_{t|t}$. To compute an expression for the expected value of future dividend growth, we assume that the transition matrix remain as Π_t, which amounts to assuming that agents do not account for the effect of their future learning when projecting cash flows. Under this assumption, we have

$$\mathbb{E}_t\left[\Delta d_{t+1+j}\right] = (\psi \iota_1)' \left(\Pi_t^j \hat{\xi}_{t|t}\right). \tag{3.32}$$

Using (3.31) we can compute an expression for the log stock price

$$\begin{aligned} p_t &= d_t + \frac{c-\bar{r}}{1-\rho} + \sum_{j=0}^{\infty} \rho^j \mathbb{E}_t \left[\Delta d_{t+1+j}\right] \\ &= d_t + \frac{c-\bar{r}}{1-\rho} + \sum_{j=0}^{\infty} \rho^j \left(\psi \iota_1\right)' \left(\Pi_t^j \hat{\xi}_{t|t}\right) \\ &= d_t + \frac{c-\bar{r}}{1-\rho} + \left(\psi \iota_1\right)' \left(\left(I - \rho \Pi_t\right)^{-1} \hat{\xi}_{t|t}\right). \end{aligned} \tag{3.33}$$

From this we obtain an expression for returns:

$$\begin{aligned} r_{t+1} &= c + \rho\left(p_{t+1} - d_{t+1}\right) + d_{t+1} - p_t \\ &= c + \rho \frac{c-\bar{r}}{1-\rho} + \left(\psi \iota_1\right)' \left(\left(I - \rho \Pi_t\right)^{-1} \hat{\xi}_{t+1|t+1}\right) \\ &+ \Delta d_{t+1} - \frac{c-\bar{r}}{1-\rho} - \left(\psi \iota_1\right)' \left(\left(I - \rho \Pi_t\right)^{-1} \hat{\xi}_{t|t}\right) \\ &= \bar{r} + \Delta d_{t+1} + \left(\psi \iota_1\right)' \left(I - \rho \Pi_t\right)^{-1} \left[\hat{\xi}_{t+1|t+1} - \hat{\xi}_{t|t}\right] \end{aligned} \tag{3.34}$$

We use equations (3.33) and (3.34) along with the dividend process in (3.17) - (3.19) to simulate dividends, stock prices and stock returns.

3.4.4 Calibration of Model Parameters

In practice, the daily dividend process is not observed, and so we have to use a proxy for Δd_{t+1}.[17] To this end, we use the ADS index proposed by Aruoba, Diebold, and Scotti (2009). This is a daily business cycle index that is constructed using daily updates to "real" economic variables observed at different frequencies such as weekly payroll figures, monthly industrial production, and quarterly GDP growth. The ADS

[17] Although one can technically construct a daily cash flow composite for US dividends, the resulting series is extremely irregular with outliers on days with large dividend payouts and a strong seasonal components that varies from quarter to quarter, depending on the exact timing of individual firms' dividend payments.

index is updated daily by the Federal Reserve Bank of Philadelphia and closely tracks the business cycle. The daily ADS time series is highly persistent but is constructed to revert to a mean of zero. As observed by Rossi and Timmermann (2015), the ADS index is a good candidate for picking up a slow-moving component in variables such as consumption growth which, at least in endowment economies, are closely related to dividend growth.[18] We rescale the ADS index so that when it is simulated at a daily frequency, the time-aggregated mean and standard deviation at a yearly frequency are 5.44% and 5.71% respectively.

As our predictor variable, we use the yield on a 3-month Treasury bill which we saw earlier gives rise to several pockets with return predictability.

The values of the calibrated parameters are listed in Table 6. We choose the diagonal elements of the state transition matrix ($\pi_{11} = 0.996, \pi_{22} = 0.999$) so that the expected duration of the first and second regimes are 250 and 1,000 days, respectively. The mean of the observable predictor variable (the T-bill rate) in the first regime, $\mu_{x,1}$ is set so that its expectation in this state equals 0.040 with its expectation in the second state ($\mu_{x,2}$), at 0.053, is calibrated such that its unconditional mean across the two states matches the overall sample mean of the 3-month Treasury bill rate.

We choose a value of the autoregressive parameter for the expected growth process, $\rho_\mu = 0.9999$, that implies very high persistence in cash flows. Part of the reason for choosing such a high value is that we use daily data, and we would expect growth in daily cash flows to be highly persistent. Moreover, the level of persistence chosen in our simulations is comparable to that assumed in the literature on "long-run risk", see, e.g., Bansal and Yaron (2004). For example, the coefficient on the expected growth rate in Bansal and Yaron (2004) is 0.979 for monthly data which translates into roughly

[18] Rossi and Timmermann (2015) find that the correlation between the economic activity index, constructed using a similar methodology to that used for the ADS index, and growth in real personal, nondurable consumption is 15.4% and 39.7% at the quarterly and annual horizons, respectively.

0.999 at the daily frequency, assuming 21 days in a month. Similarly, Bansal, Kiku, and Yaron (2012) analyze a model where the persistence of the expected growth rate and of volatility is 0.9988 and 0.99995, respectively, at the daily frequency.

Therefore, we are effectively studying investors' learning about a long-run risk component in the cash flow process. The more persistent the underlying latent growth process, the slower investors' learning is expected to be, and so investors face a particularly difficult learning problem in our setting.

Given this choice for ρ_μ, the means of the expected growth process, μ, at $\frac{\mu_{\mu,1}}{1-\rho_\mu} = 3.30 \times 10^{-4}$ and $\frac{\mu_{\mu,2}}{1-\rho_\mu} = 1.87 \times 10^{-4}$, are again calibrated so that the unconditional mean matches the sample mean of the rescaled ADS series. The same is true for the standard deviation of the innovations to x. The standard deviation of the innovations to μ are chosen to be equal across regimes. That is, we impose $\sigma_{w,1} = \sigma_{w,2}$. We also impose that $\sigma_{uv} = 0$ across both regimes, and that $\sigma_{vw,1} = 0$.

3.4.5 Simulation Results

Using the calibrated parameters of the generalized predictive systems model, we generate simulated data and run the local, non-parametric time-varying coefficient regressions exactly as in the empirical specification to identify pockets and to see if the characteristics of such pockets match the characteristics of the pockets identified in the actual data. Since we are interested in matching the pocket evidence in the actual data (Table 2), for each predictor variable we generate a sample whose length matches that of the corresponding predictor variable listed in Table 1.

We show results for the two significance levels (5% and 1%) considered in our study. For the case with no learning, investors are assumed to know the state of the underlying Markov chain and μ_t is set equal to the value corresponding to the interval in which the true continuous value lies. For the case with learning, agents do not observe

the state, s_t, and instead have to form beliefs about the probability that they are in any particular (μ_t, s_t) regime. In both cases, we compare the simulated statistics to the results from the model that uses the T-bill rate as a predictor variable.

Before studying the ability of the simulated model to generate predictability pockets, first consider the overall area under the local R^2 curves displayed in Figures 1-5 as well as the areas above and below the zero line, labeled "positive" and "negative", respectively. If a returns model does not match the overall or average R^2, this suggests that it does not generate much predictability. Using the actual data, the results reported in Table 7 show that we find a local average R^2 of 0.39%. This value cannot be matched by the simulations with no learning which generate, on average, a local R^2 of -0.34%, whereas the models with learning easily match this measure (average local R^2 of 0.57%). The key reason for the no-learning model's failure to match the amount of predictability is that it generates large negative values of the local R^2 (-2.31% on average)–something that is not matched in the actual sample (-0.098%) which in turn closely lines up with the measure under the learning model (-0.11%).

Turning to the emergence of return predictability pockets, first consider the results based on the 5% significance level (Panel A in Table 8). The model with no learning generates an average of only 3.8 pockets as opposed to the seven pockets observed in the actual sample and only 4.6% of the sample is spent inside pockets compared to 11.7% in the actual data. The no-learning model does not get close to matching the values observed in the actual data of the mean or maximum integral $\bar{R}^2$ statistics or the maximum R^2 measure.

Considering next the model with learning dynamics, we see that this is capable of matching all sample statistics based either on the value of the integral R^2 measure or the length of the pockets. For example, the number of pockets is seven in the sample as compared to an average value of 7.62 in the simulations, and the simulations with

learning also match the the fraction of the sample spent inside pockets (13.6% versus 11.7%) quite closely. Only the mean and maximum values of the integral R^2 are not matched to the full extent in the simulations with learning effects–with p-values around 0.07. However, even here we find that the effect of learning dynamics is to move the simulated values much closer towards the values observed in the actual data sample.

These findings carry over to the results that use the 1% significance level to identify pockets (Panel B in Table 8). For example, whereas the model with no learning only generates 1.7 pockets on average, the model with learning generates an average of 3.6 pockets, a number that, while slightly below the four pockets observed in the data, is within sampling error of that number. The fraction of the sample spent inside pockets with predictability in the actual data (4.3%) is also matched more closely in the simulations with learning (5.9%) than in the simulations without learning (1.9%).

We find similar results in the simulations with a constant coefficient benchmark. Again the no-learning simulations have a hard time matching the integral $\bar{R}^2$ measures in addition to the maximum R^2. Moreover, this model cannot simultaneously match the number of pockets and the average R^2 with pockets (p-value of 0.017)–a task which the learning model finds much easier to accomplish (p-value of 0.336).

We conclude the following from these simulations of our predictive systems model with discrete changes in how informative the observed predictor is over the (unobserved) conditional mean process for the growth rate in cash flows. First, in the absence of learning, our model cannot match the local nature (pockets) of the temporal patterns we observe in return predictability. Second, a model that introduces learning about the underlying state process is capable of generating return predictability pockets with similar features as those observed in the actual data. Significantly, both the number of pockets and the average time spent in pockets is matched by this model. Third, since our simulations assumed a constant risk premium, the results suggest that learning about

cash flow dynamics could be an alternative explanation to the time-variation in return predictability that we document in the first part of the paper.

3.4.6 Learning Effects and Pockets

Next, we investigate the ability of an investor's misperception of expected growth to explain the rise and fall of pockets. Define the belief discrepancy measure

$$\tilde{\mu}_t \equiv \hat{\mu}_{t|t} - \mu_t \tag{3.35}$$

which is the difference between an agent's inference about expected growth ($\hat{\mu}_{t|t}$) and the true value of expected growth at time t, μ_t. We consider several different regression specifications of the following form

$$y_{it} = \alpha + x_{it}'\beta + \gamma \mathbb{1}\{edge_{it}\} + \mathbb{1}\{edge_{it}\} x_{it}'\delta + \varepsilon_{i,t} \tag{3.36}$$

Here the i subscript refers to the simulation number, from 1 to 1,000, and t refers to the time period within a simulated sample, from 1 to 13,300 (the sample size for the 3-month Treasury bill). The dummy variable $edge_{it}$ takes the value 1 for the first and last 126 (half of the kernel regression bandwidth) periods of the sample, and zero otherwise. The dependent variable y_{it} is chosen to be either an indicator for whether a pocket is identified at period t in sample i or the local R^2 measure from the kernel regression. The vector x_{it} contains various functions of $\tilde{\mu}_{it}$. We consider three specifications for x_{it}, namely (i) $x_{it} = \tilde{\mu}_t$, (ii) $x_{it} = \begin{bmatrix} \tilde{\mu}_t & \tilde{\mu}_t^2 \end{bmatrix}$, and (iii) $x_{it} = |\tilde{\mu}_t|$.

To aggregate the results across simulations, we report the coefficient estimate as the average coefficient estimate across simulations. The standard errors of the coefficient estimates are computed as the standard deviation of the estimates across simulations scaled by the square root of the number of simulations. These standard errors are then

used to compute p-values. We consider both choices for y_{it} and allowing for 5% and 1% significance thresholds for identifying pockets. Note that the choice of significance threshold does not affect the local R^2 results, only the pocket indicator variable results. The results are reported in Table 9.

The results show that the belief discrepancy measure, while not strongly correlated on its own with either the pocket indicator or the local R^2 measure, is strongly correlated with both the pocket indicator and the local R^2 when combined with the squared belief discrepancy. Similarly, the absolute value of the belief discrepancy measure is highly correlated on its own than either the pocket indicator or the local R^2. Moreover, the explanatory power of these discrepancy measures over variation in future stock returns can be quite high–ranging from about 6.5% for the regression of the pocket indicator on the simple belief discrepancy measure, $\tilde{\mu}_t$, to about 15% for the regression of the local R^2 on $\tilde{\mu}_t$ and $\tilde{\mu}_t^2$.

Interestingly, in unreported results we find that the switching indicator is not significantly correlated with either measure of return predictability–in part because regime switches are quite rare in our sample.

Together, these findings suggest that variation in investors' learning about the highly persistent growth rate of the cash flow process can create pockets of return predictability.

3.5 Economic Sources of Local Return Predictability

We argued earlier that the return predictability pockets detected by our analysis can be used as a diagnostic that helps identify the sources of return predictability. In this section we use this idea to explore whether the evidence of local return predictability is associated with business cycle movements and movements in variables known to track market sentiment. Moreover, we also study whether the pockets with return predictability

could have been detected in real time. This is a question with implications for whether investors could have exploited localized return predictability.

3.5.1 Pockets and Variation in the Business Cycle

Studies such as Henkel, Martin, and Nardari (2011) and Dangl and Halling (2012) find a strong relationship between return predictability in the stock market and economic recessions. To explore this relationship, we regress the pocket indicator generated by our univariate linear regressions, I_t^{pocket}, on a constant and the NBER recession indicator, $NBER_t$

$$I_t^{pocket} = \mu + \beta NBER_t + \varepsilon_t.$$

A positive coefficient β suggests that return predictability pockets are more likely to occur during economic recessions while a negative value of β suggests the opposite.

To see whether the extent of return predictability depends on the state of the economy, we also regress the local R^2 measure on the NBER indicator

$$R_t^2 = \mu + \beta NBER_t + \varepsilon_t.$$

Here a positive coefficient indicates that return predictability tends to be higher during recessions, while a negative coefficient would indicate the opposite

The results, reported in Panels A and B of Table 10, show that local predictability of stock returns is indeed related to the business cycle. However, two observations suggest that business cycle variation is not the main driver of local pockets of return predictability. First, the R^2-values of these regressions are very low, less than four percent for all predictors with exception of the term spread for which the R^2 is close to 17%.[19] Second, although the local return predictability identified by the models that use the

[19] We find similar results when we project the pocket indicator on an early recession indicator (the three months after the peak of the cycle) or a late recession indicator (three months before the trough).

T-bill rate, the term spread, and volatility rises during recessions, the opposite holds for the dividend yield for which return predictability tends to fall during recessions.

3.5.2 Pockets and Variation in Sentiment

Our second regression uses the sentiment indicators proposed by Baker and Wurgler (2006), Baker and Wurgler (2007) as a means to see whether return predictability is correlated with market sentiment. We first assign to each day within a given month the value of the Baker-Wurgler sentiment indicator, BW, of the same month. Then, analogously to the analysis of a business cycle component in return predictability, we estimate daily regressions

$$\begin{aligned} R_t^2 &= \mu + \beta BW_t + \varepsilon_t, \\ I_t^{pocket} &= \mu + \beta BW_t + \varepsilon_t, \end{aligned}$$

Panels C and D in Table 10 show evidence that large values of the BW index are associated with a greater degree of local return predictability, with the R^2 of the relation being particularly high (15-20%) for the forecasts generated by the T-bill rate and the realized variance.

3.5.3 Out-of-Sample Return Predictability

So far our methods for identifying return predictability have used two-sided kernels, i.e., windows consisting of data both before and after the point at which local return predictability is being tested. In real time, investors only have access to data prior to and including the point at which the forecast is being generated and so must use a one-sided window to estimate their model.

If return predictability is not driven by a time-varying risk premium, then we would not expect a one-sided prediction approach to be able to generate better return

forecasts than a simple model with a constant equity premium. To see if this holds, we estimate the same model as in the earlier Section but use a one-sided analog of the Epanechnikov Kernel in (3.11):

$$K(z) = \frac{3}{2}\left(1-z^2\right)1\{-1 < z < 0\}, \tag{3.37}$$

so that only past data are used to estimate the time varying relationship between y and x as indicated by $1\{-1 < z < 0\}$.[20]

We construct two forecasts of excess returns at time $t+1$. The first is the prevailing mean benchmark of Goyal and Welch (2008):

$$\bar{r}_{t+1|t} = \frac{1}{t}\sum_{s=1}^{t} r_s. \tag{3.38}$$

The second forecast is generated by the nonparametric model:

$$\hat{r}_{t+1|t}^{local} = \bar{r}_{t+1|t} + x_t'\hat{\beta}_t, \tag{3.39}$$

where $x_t'\hat{\beta}_t$ is rescaled by the unconditional standard deviation of r.

To see if local return predictability could have been exploited in real time, we test the null of equal predictive accuracy (equal squared forecast errors) for the prevailing mean model in (3.38) and the time varying mean model in (3.39). To this end, table 11 reports values for the test statistics proposed by Diebold and Mariano (1995) and Clark and West (2007). The Diebold and Mariano (1995) statistic is based on the difference in

[20]Note that the multiplicative factor becomes $\frac{3}{2}$ instead of $\frac{3}{4}$ in (3.11) so that the kernel function in (3.37) still integrates to one.

mean squared forecasts errors

$$\Delta MSE_{t+1} = \left(r_{t+1} - \bar{r}_{t+1|t}\right)^2 - \left(r_{t+1} - f_{t+1}^{local}\right)^2.$$

Clark and West (2007) propose a test statistic that accounts for the effect of estimation error by recentering the Diebold-Marino test. The test does so by using the mean of the adjusted MSE measure

$$\Delta MSE_{t+1}^{adj} = \left(r_{t+1} - \bar{r}_{t+1|t}\right)^2 - \left(\left(y_{t+1} - \hat{r}_{t+1|t}^{local}\right)^2 - \left(\bar{r}_{t+1|t} - \hat{r}_{t+1|t}^{local}\right)^2\right) \tag{3.40}$$

Positive values of these test statistics suggest that the time-varying mean model performs best, while negative values suggest that the prevailing mean model produces the most accurate one-sided forecasts. Using the Diebold-Mariano test, we find for all variables that the prevailing mean model (3.38) produces better out-of-sample forecasts than the model with a local time-varying mean, (3.39), i.e., the MSE of the prevailing mean model is lower than that of the model that allows for a time-varying mean, although none of the test statistics are significant at conventional levels. Turning to the Clark-West results, we continue to find no evidence that the time-varying mean model performs significantly better than the prevailing mean specification.

These results suggest little evidence that local return predictability could have been exploited in real time to produce more accurate return forecasts than a model that assumes a constant equity premium. In fact, the one-sided estimates of the regression coefficients in (3.6) are notably noisier than their two-sided equivalents. The stark difference between the one-sided and two-sided results can thus be explained by the latter's use of more information, and thus improved power, to identify local return predictability.[21]

[21] Lettau and Van Nieuwerburgh (2008) report a similar finding for a return predictability model with

3.6 Conclusion

We use a novel nonparametric methodology to establish evidence that while stock returns may not be predictable "all the time," as identified by full-sample constant coefficient regressions, there is strong evidence that returns are predictable "some of the time" using some of the most popular predictor variables considered in the finance literature on return predictability.

Predictability of stock market returns are particularly suited for studying learning effects due to the dependency of stock prices on cash flows expected to occur in the distant future and the considerable uncertainty surrounding such expectations. The high sensitivity of aggregate stock prices to even minor variations in beliefs about future cash flow growth rates means that cash flow learning effects are likely to be an important source of return movements.[22]

Our nonparametric methodology for identifying local return pockets can be used as a diagnostic for determining whether a particular asset pricing model matches the data. Specifically, by simulating from asset pricing models and applying our pocket methodology to the resulting data, we can compute whether a specific asset pricing model generates approximately the right number of pockets. This approach is particularly insightful for models with incomplete learning about some underlying (latent) state such as that proposed by David and Veronesi (2013). Comparing results with and without the incomplete learning mechanism in place, we can see how this mechanism affects the ability to generate return predictability pockets.

---

breaks to the dividend yield.

[22] In a model with paradigm shifts, Hong, Stein, and Yu (2007) find that investors learning about the underlying model that generates dividends can give rise to predictable variation in returns and help to match volatility and skewness patterns in returns. In their model, agents switch between models that are under-dimensioned specifications relative to the true dividend generating process.

3.7 Figures and Tables

Table 3.1. Full Sample Regression Statistics

Variables	Full sample beta	*t*-statistic	$\bar{R}^2$ in %	Start date	No. of obs.
dy	0.060	1.255	0.005	11/5/1926	22,778
tbl	-0.008	-2.724	0.056	1/4/1954	14,852
tsp	0.018	2.415	0.047	1/2/1962	12,838
cpsp	0.007	0.095	-0.014	1/2/1986	6,808
svar	0.000	0.344	-0.001	1/15/1927	22,719

This table reports full-sample beta estimates, t-statistics, and $\bar{R}^2$ values for univarate regressions of daily excess stock returns on the predictor variables listed in the rows. All series run through the end of 2012.

Table 3.2. Pocket Summary Statistics

	In-pocket					Out-of-pocket				
Variables	**Num pockets**	**Min length**	**Max length**	**Avg. length**	**Frac signif**	**Num pockets**	**Min length**	**Max length**	**Avg. length**	**Frac signif**
	Panel A: 5% pocket statistics									
dy	13	24	876	196	0.113	13	118	6,612	1,537	0.887
tbl	7	62	411	243	0.117	7	164	4,330	1842	0.883
tsp	3	301	487	378	0.090	4	790	7,481	2,863	0.910
cpsp	2	194	57	126	0.038	3	955	4,376	2,101	0.962
svar	8	52	337	232	0.083	9	717	3,414	2,290	0.917
	Panel B: 1% pocket statistics									
dy	4	6	345	159	0.028	5	25	13,560	4,378	0.972
tbl	4	63	224	157	0.043	4	830	9,718	3493	0.957
tsp	3	98	299	205	0.049	4	1,061	7,525	2,992	0.951
cpsp	0	0	0	0	0.000	1	6,555	6,555	6,555	1.000
svar	5	86	251	190	0.042	6	888	7,727	3,585	0.956
	Avg. R^2	**Mean**	**Std. dev**	**Skew**	**Kurtosis**	**Avg. R^2**	**Mean**	**Std. dev**	**Skew**	**Kurtosis**
	Panel C: 5% return statistics									
dy	0.015	0.047	0.820	0.175	27.792	0.004	0.024	1.113	-0.113	18.821
tbl	0.016	-0.025	0.859	0.121	5.931	0.003	0.030	0.976	-0.602	20.525
tsp	0.017	-0.025	1.038	0.294	4.177	0.003	0.027	0.993	-0.602	20.891
cpsp	0.010	0.130	0.783	0.368	3.873	0.003	0.024	1.183	-0.644	18.252
svar	0.021	0.059	0.812	-0.506	8.048	0.003	0.024	1.106	-0.084	19.541
	Panel D: 1% return statistics									
dy	0.026	0.131	0.606	-1.000	7.247	0.005	0.024	1.094	-0.092	19.372
tbl	0.021	-0.029	0.826	-0.259	3.922	0.004	0.026	0.968	-0.548	19.849
tsp	0.020	-0.007	1.156	0.446	3.979	0.004	0.024	0.988	-0.588	20.451
cpsp	-	-	-	-	-	0.003	0.028	1.170	-0.641	18.366
svar	0.026	0.043	0.684	-0.502	5.068	0.003	0.026	1.099	-0.096	19.362

This table reports summary statistics on the number of pockets with significant return predictability from the predictor variable listed in the left column, using a non-parametric kernel regression approach with significance levels of 5% (Panel A) or 1% (Panel B) to identify pockets. We show, for each predictor variable, the number of pockets identified, the minimum, maximum and average pocket length (all measured in days) as well as the fraction of day in the sample identified to have significant local return predictability. Left columns show summary statistics for periods inside pockets while, for comparison, right columns show summary statistics for periods outside pockets. Panels C and D report summary statistics for daily excess returns inside (left panels) and outside (right panels) pockets, including the average value of the local $\bar{R}^2$, the mean, standard deviation, skewness and kurtosis of daily returns. The sample periods vary across the predictor variables and begin in 11/5/1926 for the dividend yield (22,778 observations), 1/2/1954 for the 3-month T-bill rate (14,852 obs.), 1/2/1962 (12,838 obs.) for the term spread, 1/2/1986 (6,808 obs.) for the corporate spread and 1/15/1927 (22,719 obs.) for the realized variance.

Table 3.3. Statistical Significance Tests for Pocket Diagnostics (Zero Coefficient Benchmark)

		Random Walk			GARCH		
Stats	**Sample**	**Avg.**	**Std. err.**	**p-val**	**Avg.**	**Std. err.**	**p-val**
			dy				
Num pockets	13	6.690	2.379	0.011	6.615	2.456	0.011
Min length	24	49.298	41.452	0.679	52.027	46.429	0.701
Max length	876	319.575	104.638	0.000	328.7780	106.6839	0.0010
Avg. length	196	164.725	46.864	0.229	169.277	48.790	0.264
Min integral $\bar{R}^2$	0.289	0.269	0.521	0.261	0.229	0.506	0.237
Mean integral $\bar{R}^2$	2.972	0.802	1.028	0.042	0.713	0.964	0.043
Max integral $\bar{R}^2$	13.600	1.392	1.634	0.001	1.254	1.557	0.000
Frac signif	0.113	0.048	0.021	0.006	0.049	0.021	0.002
Avg. $\bar{R}^2$ within pockets	0.015	0.014	0.002	0.229	0.014	0.003	0.340
Max $\bar{R}^2$	0.040	0.021	0.006	0.011	0.024	0.008	0.036
			tbl				
Num pockets	7	4.576	2.117	0.180	4.433	1.907	0.142
Min length	62	66.275	58.737	0.432	64.829	57.369	0.411
Max length	411	281.470	113.442	0.112	283.066	111.941	0.112
Avg. length	243.29	160.739	59.713	0.082	161.880	60.896	0.083
Min integral $\bar{R}^2$	0.415	0.299	0.512	0.223	0.290	0.543	0.198
Mean integral $\bar{R}^2$	3.795	0.747	0.878	0.011	0.754	0.921	0.017
Max integral $\bar{R}^2$	7.225	1.287	1.520	0.010	1.312	1.502	0.006
Frac signif	0.117	0.049	0.026	0.013	0.049	0.026	0.011
Avg. $\bar{R}^2$ within pockets	0.016	0.013	0.004	0.153	0.014	0.003	0.214
Max $\bar{R}^2$	0.063	0.020	0.007	0.000	0.021	0.007	0.000
			tsp				
Num pockets	3	3.759	1.784	0.752	3.754	1.820	0.755
Min length	301	79.087	70.040	0.012	79.326	69.949	0.015
Max length	487	273.983	118.532	0.037	276.760	119.453	0.044
Avg. length	377.67	166.918	71.208	0.015	168.976	71.243	0.008
Min integral $\bar{R}^2$	4.543	0.261	0.545	0.002	0.283	0.683	0.004
Mean integral $\bar{R}^2$	6.604	0.647	0.943	0.002	0.674	1.028	0.005
Max integral $\bar{R}^2$	8.394	1.113	1.676	0.006	1.118	1.510	0.006
Frac signif	0.090	0.049	0.027	0.085	0.050	0.029	0.097
Avg. $\bar{R}^2$ within pockets	0.017	0.013	0.003	0.084	0.014	0.004	0.161
Max $\bar{R}^2$	0.024	0.019	0.006	0.157	0.021	0.008	0.252
			cpsp				
Num pockets	2	2.180	1.404	0.658	2.053	1.327	0.631
Min length	57	94.749	85.912	0.598	99.286	90.738	0.611
Max length	194	205.702	124.609	0.536	199.837	128.579	0.497
Avg. length	125.50	146.959	89.229	0.595	145.655	92.881	0.574
Min integral $\bar{R}^2$	0.670	0.511	0.775	0.217	0.539	0.906	0.216
Mean integral $\bar{R}^2$	1.269	0.807	1.027	0.179	0.842	1.147	0.175
Max integral $\bar{R}^2$	1.868	1.120	1.466	0.169	1.149	1.515	0.165
Frac signif	0.038	0.052	0.039	0.568	0.048	0.038	0.516
Avg. $\bar{R}^2$ within pockets	0.010	0.013	0.004	0.763	0.014	0.004	0.751
Max $\bar{R}^2$	0.012	0.017	0.007	0.745	0.017	0.008	0.749
			svar				
Num pockets	8	6.759	2.496	0.369	6.676	2.415	0.357
Min length	52	48.196	40.947	0.361	47.473	45.421	0.331
Max length	337	323.377	106.942	0.402	337.783	106.292	0.492
Avg. length	231.75	165.572	46.779	0.082	170.781	50.012	0.099
Min integral $\bar{R}^2$	0.613	0.236	0.492	0.101	0.219	0.424	0.112
Mean integral $\bar{R}^2$	4.844	0.783	0.942	0.007	0.770	0.912	0.007
Max integral $\bar{R}^2$	9.091	1.453	1.569	0.005	1.420	1.552	0.003
Frac signif	0.083	0.049	0.022	0.075	0.050	0.021	0.070
Avg. $\bar{R}^2$ within pockets	0.020	0.014	0.002	0.012	0.015	0.003	0.075
Max $\bar{R}^2$	0.034	0.021	0.006	0.035	0.025	0.009	0.148

This table reports the outcome of Monte Carlo simulations of daily excess returns using either a random walk model with constant mean and volatility (columns 2-4) or a model that allows for a time-varying expected return and time-varying volatility (columns 5-7). Using these respective models, each simulation draws a sample with the same length as the original sample for the respective predictor variables and computes the pocket measures listed in each row, including the number of pockets, the minimum, maximum and average length (in days) of the pockets, the minimum, mean and maximum integral $\bar{R}^2$, the fraction of the sample with a significant pocket indicator, the average and maximum values of the $\bar{R}^2$ inside pockets. The average values, standard errors and p-values for the pocket measures are computed using 1,000 simulations and are based on a zero coefficient benchmark.

Table 3.4. Statistical Significance Tests for Pocket Diagnostics (Constant Coefficient Benchmark)

		Random Walk			GARCH		
Stats	**Sample**	**Avg.**	**Std. err.**	**p-val**	**Avg.**	**Std. err.**	**p-val**
			dy				
Num pockets	16	6.553	2.513	0.000	6.525	2.377	0.001
Min length	4	51.110	45.346	0.954	51.421	43.967	0.965
Max length	414	313.919	109.073	0.138	325.440	113.093	0.172
Avg. length	124.56	162.211	49.435	0.778	168.459	51.107	0.818
Min integral $\bar{R}^2$	0.026	0.244	0.398	0.722	0.251	0.504	0.735
Mean integral $\bar{R}^2$	1.977	0.759	0.830	0.086	0.753	0.980	0.088
Max integral $\bar{R}^2$	11.434	1.356	1.438	0.000	1.352	1.611	0.003
Frac signif	0.089	0.046	0.021	0.026	0.048	0.022	0.047
Avg. $\bar{R}^2$ within pockets	0.016	0.014	0.003	0.133	0.014	0.003	0.202
Max $\bar{R}^2$	0.040	0.021	0.0067	0.017	0.023	0.007	0.023
			tbl				
Num pockets	6	4.311	2.115	0.280	4.395	2.047	0.271
Min length	39	64.905	55.002	0.598	68.659	60.534	0.644
Max length	352	263.357	114.100	0.193	275.110	118.703	0.212
Avg. length	240	152.970	58.250	0.057	159.153	61.224	0.089
Min integral $\bar{R}^2$	0.808	0.295	0.523	0.098	0.311	0.562	0.113
Mean integral $\bar{R}^2$	3.429	0.742	0.926	0.026	0.758	0.971	0.030
Max integral $\bar{R}^2$	5.952	1.300	1.572	0.027	1.273	1.517	0.024
Frac signif	0.099	0.045	0.026	0.033	0.047	0.027	0.038
Avg. $\bar{R}^2$ within pockets	0.014	0.013	0.003	0.348	0.013	0.003	0.364
Max $\bar{R}^2$	0.063	0.020	0.0068	0.000	0.0210	0.008	0.001
			tsp				
Num pockets	7	3.543	1.791	0.059	3.861	1.862	0.086
Min length	76	75.148	67.709	0.371	75.282	72.951	0.366
Max length	307	254.624	115.128	0.308	277.774	122.442	0.386
Avg. length	183.29	156.775	67.576	0.302	165.073	72.327	0.346
Min integral $\bar{R}^2$	0.612	0.263	0.499	0.148	0.248	0.532	0.134
Mean integral $\bar{R}^2$	2.324	0.598	0.797	0.046	0.630	0.863	0.053
Max integral $\bar{R}^2$	5.691	0.994	1.241	0.013	1.093	1.424	0.015
Frac signif	0.102	0.044	0.027	0.035	0.050	0.030	0.065
Avg. $\bar{R}^2$ within pockets	0.013	0.013	0.003	0.453	0.013	0.004	0.499
Max $\bar{R}^2$	0.024	0.018	0.006	0.137	0.020	0.007	0.235
			cpsp				
Num pockets	1	1.9470	1.331	0.871	2.003	1.365	0.882
Min length	170	93.710	91.559	0.162	91.665	86.793	0.150
Max length	170	186.210	128.122	0.547	193.966	135.338	0.545
Avg. length	170	137.326	94.621	0.332	139.239	92.373	0.334
Min integral $\bar{R}^2$	1.785	0.548	0.851	0.077	0.575	1.047	0.070
Mean integral $\bar{R}^2$	1.785	0.790	1.019	0.119	0.854	1.252	0.121
Max integral $\bar{R}^2$	1.785	1.055	1.363	0.179	1.149	1.608	0.194
Frac signif	0.026	0.044	0.036	0.631	0.047	0.039	0.639
Avg. $\bar{R}^2$ within pockets	0.011	0.013	0.004	0.623	0.013	0.005	0.600
Max $\bar{R}^2$	0.012	0.017	0.007	0.712	0.017	0.008	0.759
			svar				
Num pockets	10	6.532	2.456	0.123	6.511	2.398	0.108
Min length	28	49.182	48.203	0.613	46.621	41.746	0.597
Max length	354	316.317	106.171	0.3360	325.881	109.583	0.378
Avg. length	206.70	163.116	51.477	0.161	166.386	50.056	0.185
Min integral $\bar{R}^2$	0.238	0.225	0.428	0.282	0.248	0.648	0.287
Mean integral $\bar{R}^2$	4.084	0.729	0.828	0.007	0.792	1.021	0.017
Max integral $\bar{R}^2$	9.367	1.340	1.394	0.001	1.475	1.657	0.006
Frac signif	0.092	0.047	0.021	0.032	0.048	0.022	0.040
Avg. $\bar{R}^2$ within pockets	0.020	0.014	0.002	0.017	0.015	0.003	0.058
Max $\bar{R}^2$	0.034	0.022	0.007	0.053	0.025	0.009	0.126

This table reports the outcome of Monte Carlo simulations of daily excess returns using either a random walk model with constant mean and volatility (columns 2-4) or a model that allows for a time-varying expected return and time-varying volatility (columns 5-7). Using these respective models, each simulation draws a sample with the same length as the original sample for the respective predictor variables and computes the pocket measures listed in each row, including the number of pockets, the minimum, maximum and average length (in days) of the pockets, the minimum, mean and maximum integral $\bar{R}^2$, the fraction of the sample with a significant pocket indicator, the average and maximum values of the $\bar{R}^2$ inside pockets. The average values, standard errors and p-values for the pocket measures are computed using 1,000 simulations and are based on a constant coefficient benchmark, computed as the full-sample slope coefficient for each predictor.

Table 3.5. Integral $\bar{R}^2$ Measure and p-values for Individual Pockets

Pocket #	dy	tbl	tsp	cpsp	svar
1	0.897	2.465	4.543	1.868	1.686
	(0.247)	(0.073)	(0.018)	(0.106)	(0.134)
2	0.522	4.925	8.394	0.670	7.402
	(0.381)	(0.013)	(0.003)	(0.322)	(0.003)
3	13.600	7.225	6.874	-	9.091
	(0.000)	(0.002)	(0.006)		(0.000)
4	2.214	0.415	-	-	3.528
	(0.078)	(0.452)			(0.035)
5	12.555	6.280	-	-	6.892
	(0.000)	(0.005)			(0.004)
6	1.756	2.232	-	-	6.597
	(0.109)	(0.088)			(0.005)
7	0.435	3.023	-	-	0.613
	(0.427)	(0.050)			(0.375)
8	1.353	-	-	-	2.944
	(0.152)				(0.050)
9	2.020	-	-	-	-
	(0.085)				
10	0.400	-	-	-	-
	(0.447)				
11	0.672	-	-	-	-
	(0.318)				
12	0.289	-	-	-	-
	(0.525)				
13	1.924	-	-	-	-
	(0.093)				

This table reports the integral $\bar{R}^2$ measure for each of the pockets identified by our nonparametric kernel regression approach, assuming a 5% cutoff value to define pockets with p-values in brackets. To compute p-values We use the Monte Carlo simulations in Table 3 to compute the proportion of simulations that can generate integral $\bar{R}^2$ meaures as high as the value associated with a particular pocket.

Table 3.6. Calibrated Parameters, Predictive Systems Model

Parameter	Parameter value	Parameter description
π_{11}	0.996	Probability of staying in regime 1
π_{22}	0.999	Probability of staying in regime 2
ρ_x	0.99	Persistence of observed predictor variable x
$\frac{\mu_{x,1}}{1-\rho_x}$	0.040	Unconditional mean of observed predictor variable in regime 1
$\frac{\mu_{x,1}}{1-\rho_x}$	0.053	Unconditional mean of observed predictor variable in regime 2
ρ_μ	0.9999	Persistence of expected cash flows
$\frac{\mu_{\mu,1}}{1-\rho_\mu}$	3.30×10^{-4}	Unconditional mean of expected cash flows in regime 1
$\frac{\mu_{\mu,2}}{1-\rho_\mu}$	1.87×10^{-4}	Unconditional mean of expected cash flows in regime 2
$\frac{\sigma_{v,1}}{\sqrt{1-\rho_x^2}}$	0.015	Unconditional standard deviation of observed predictor variable in regime 1
$\frac{\sigma_{v,2}}{\sqrt{1-\rho_x^2}}$	0.032	Unconditional standard deviation of observed predictor variable in regime 2
$\frac{\sigma_w}{\sqrt{1-\rho_\mu^2}}$	1.5×10^{-5}	Unconditional standard deviation of expected cash flows
σ_u	1.5×10^{-4}	Standard deviation of realized cash flows
$\frac{\sigma_{vw,2}}{\sigma_{v,2}\sigma_{w,2}}$	-0.8	Correlation between innovations to observed predictor variable and expected cash flows in regime 2

This table reports the values and descriptions for the calibrated parameter values in the predictive systems model.

Table 3.7. Average Integral $\bar{R}^2$, Predictive Systems Model

		No learning			Learning		
	Sample	**Avg.**	**Std. err.**	**p-val**	**Avg.**	**Std. err.**	**p-val**
Positive	0.506	0.243	0.131	0.051	0.642	0.265	0.629
Negative	-0.098	-2.315	1.111	0.000	-0.113	0.086	0.581
Net	0.393	-0.336	0.277	0.003	0.566	0.234	0.775

This table reports the average integral $\bar{R}^2$ conditional on it being positive, negative, and over the whole sample. For the "positive" measure, the average of the local $\bar{R}^2$ is taken over all periods where it is positive and multiplied by 100. The analogous procedure is done for the "negative" measure. For the "net" measure, the average of the local $\bar{R}^2$ is taken over the whole sample and multiplied by 100. These statistics are computed for the actual data under "Sample," and the average, standard error, and one-sided p-values are computed for the predictive systems model simulations under both the no learning and learning specifications.

Table 3.8. Simulations from Predictive Systems Learning Model (Zero Coefficient Benchmark)

		No learning			Learning		
Stats	**Sample**	**Avg.**	**Std. err.**	**p-val**	**Avg.**	**Std. err.**	**p-val**
5% significance results							
Num pockets	7.000	3.863	2.495	0.151	7.622	3.148	0.614
Min pocket length	62.000	60.792	73.950	0.343	66.532	55.632	0.439
Avg. pocket length	243.290	143.107	84.589	0.093	221.703	74.298	0.345
Max pocket length	411.000	269.550	181.829	0.171	477.542	219.585	0.552
Min integral $\bar{R}^2$	0.415	-0.894	5.257	0.127	0.353	0.747	0.244
Mean integral $\bar{R}^2$	3.795	0.082	3.201	0.044	1.243	1.512	0.066
Max integral $\bar{R}^2$	7.225	1.046	4.119	0.044	2.507	3.050	0.068
Fraction significant	0.117	0.046	0.039	0.068	0.136	0.078	0.529
Max R^2	0.063	0.024	0.016	0.036	0.036	0.018	0.092
Num pockets & Avg. pocket length	-	-	-	0.013	-	-	0.238
1% significance results							
Num pockets	4.000	1.678	1.903	0.160	3.590	2.599	0.436
Min pocket length	63.000	61.298	83.350	0.349	96.863	91.881	0.571
Avg. pocket length	156.500	95.870	98.616	0.260	180.533	108.991	0.593
Max pocket length	224.000	145.299	166.509	0.272	304.616	206.597	0.635
Min integral $\bar{R}^2$	1.207	-0.219	3.287	0.070	0.521	1.016	0.124
Mean integral $\bar{R}^2$	3.233	0.150	3.516	0.056	1.042	1.487	0.074
Max integral $\bar{R}^2$	4.798	0.626	4.658	0.053	1.679	2.542	0.086
Fraction significant	0.043	0.019	0.027	0.148	0.059	0.055	0.500
Max R^2	0.063	0.024	0.016	0.036	0.036	0.018	0.092
Num pockets & Avg. pocket length	-	-	-	0.076	-	-	0.336

This table presents simulation results from the predictive systems model with regime switching in the cash flow growth rate. Investors observe a predictor variable that is correlated with the latent process driving the mean dividend growth rate, but whose correlation is also affected by the regime switching. In the scenario with no learning (columns 2-4), investors are assumed to observe the latent state variable while in the scenario with learning (columns 5-7), investors update their estimates of the mean dividend growth rate based on their probability estimates of the underlying state. The reported sample average, standard errors and p-values for the simulated data are based on 1,000 simulations of the same length as the sample for the T-bill rate and assume a mean dividend-price ratio of 0.038 which is the historical sample average. Pockets in both the actual and simulated data sample are computed around a zero coefficient benchmark.

Table 3.9. Panel Regressions of Pocket Diagnostics on Belief Discrepancies, Predictive Systems Model

Regressor	**Pocket Indicator**			**Local R^2**		
	5% significance results					
	(1)	(2)	(3)	(4)	(5)	(6)
$\hat{\mu}_{t\|t} - \mu_t$	449*	1.38×10^4***	-	6	235***	-
	(0.07)	(< 0.01)		(0.40)	(< 0.01)	
$\left(\hat{\mu}_{t\|t} - \mu_t\right)^2$	-	4.82×10^8***	-	-	1.36×10^7***	-
		(< 0.01)			(< 0.01)	
$\lvert\hat{\mu}_{t\|t} - \mu_t\rvert$	-	-	-2,052***	-	-	-56***
			(< 0.01)			(< 0.01)
$\bar{R}^2$	6.52%	10.30%	6.30%	8.88%	14.63%	8.38%
	1% significance results					
	(1)	(2)	(3)	(4)	(5)	(6)
$\hat{\mu}_{t\|t} - \mu_t$	-54	3,975**	-	6	235***	-
	(0.76)	(0.02)		(0.40)	(< 0.01)	
$\left(\hat{\mu}_{t\|t} - \mu_t\right)^2$	-	2.72×10^8***	-	-	1.36×10^7***	-
		(< 0.01)			(< 0.01)	
$\lvert\hat{\mu}_{t\|t} - \mu_t\rvert$	-	-	-1,229***	-	-	-56***
			(< 0.01)			(< 0.01)
$\bar{R}^2$	7.81%	11.50%	7.50%	8.88%	14.63%	8.38%

This table reports coefficient estimates and p-values (in brackets) from regressions of the local $\bar{R}^2$ measure for return predictability or the binary pocket indicator that is one inside pockets with return predictability and is zero otherwise on an intercept, a dummy for being within the first or last 126 (half of the kernel regression bandwidth) periods of the sample, functions of the difference between the true simulated expected cash flows and the agent's filtered beliefs about expected cash flows, and their interactions with the dummy. All specifications contain the intercept, the dummy, and its interactions with any other regressors that are included. The coefficient estimate is the average coefficient across simulations, and the p-values are computed by dividing the average by the standard deviation of the estimates across samples and multiplying by the square root of the number of simulations. The $\bar{R}^2$ is computed as the average $\bar{R}^2$ across simulations for each regression specification.

Table 3.10. Regressions of Pocket Diagnostics on Economic Indicators

Variables	Slope	$\bar{R}^2$ (in %)	Slope	$\bar{R}^2$ (in %)
	NBER Recession indicator			
	Panel A: Local $\bar{R}^2$		**Panel B: Pocket indicator**	
dy	0.0022 (0.00)	2.223	-0.0055 (0.30)	0.000
tbl	0.0030 (0.00)	3.830	0.1539 (0.00)	2.840
tsp	0.0061 (0.00)	16.390	0.3439 (0.00)	17.188
cpsp	0.0013 (0.00)	1.935	0.1182 (0.00)	3.648
svar	0.0015 (0.00)	0.835	0.0251 (0.00)	0.122
	BW index			
	Panel C: Local $\bar{R}^2$		**Panel D: Pocket indicator**	
dy	0.0010 (0.00)	6.980	-0.0035 (0.04)	0.022
tbl	0.0020 (0.00)	14.972	0.0467 (0.00)	2.165
tsp	0.0010 (0.00)	3.603	0.0454 (0.00)	2.297
cpsp	0.0005 (0.00)	0.648	-0.0821 (0.00)	4.445
svar	0.0026 (0.00)	19.623	0.1008 (0.00)	13.698

This table reports coefficient estimates and p-values (in brackets) along with the $\bar{R}^2$ value from regressions of the local $\bar{R}^2$ measure for return predictability (Panel A) or the binary pocket indicator that is one inside pockets with return predictability and is zero otherwise (Panel B) on an intercept and either the NBER recession indicator (Panels A and B) or the Baker-Wurgler sentiment index (Panels C and D). All regressions use daily data with the samples described in the caption to Table 1.

Table 3.11. Out-of-Sample Measures of Forecasting Performance

Variables	Clark-West statistic	Diebold-Mariano statistic
dy	-0.335	-1.658
tbl	0.514	-0.748
tsp	1.230	-1.137
cpsp	-1.519	-1.647
svar	-1.026	-1.399

This table reports the Clark and West (2010) and Diebold-Mariano (1995) test statistics for out-of-sample return predictability measured relative to a prevailing mean model that assumes constant expected excess returns. These test statistics approximately follow a normal distribution with postive values indicating more accurate out-of-sample return forecasts than the prevailing mean benchmark and negative values indicating the opposite.

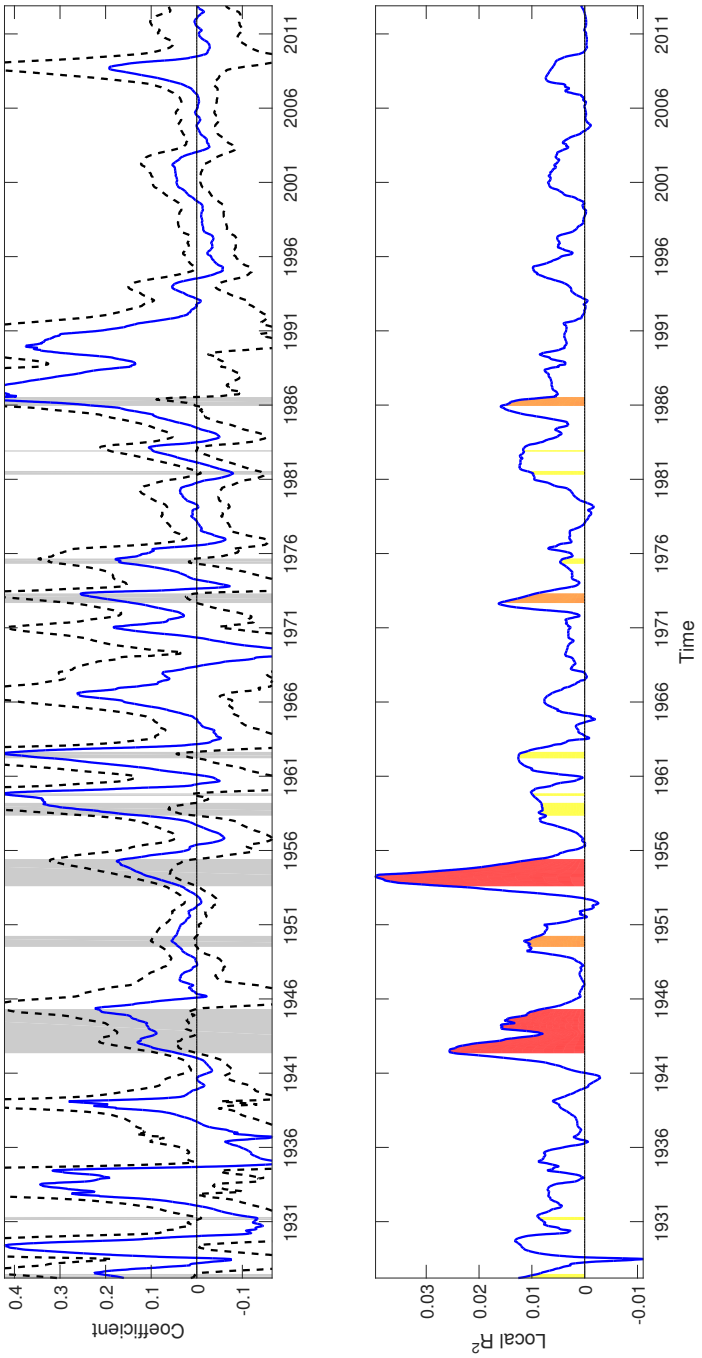


Figure 3.1. Local Return Predictability from the Dividend Yield (Zero Coefficient Benchmark)

Local return predictability from the dividend yield. The top panel in this figure plots non-parametric kernel estimates of the local slope coefficient from a regression of daily excess stock returns on the lagged dividend yield. Dashed lines represents plus or minus two standard error bands. The bottom panel plots the local $\bar{R}^2$ measure with shaded areas tracking periods identified as pockets of return predictability using a 5% critical value. The shaded areas represent the integrated $\bar{R}^2$ inside pockets with areas colored in red representing pockets that have less than a 5% chance of being spurious, areas colored in orange representing pockets that have between a 5% and a 10% chance of being spurious, and areas colored in yellow representing pockets that have more than 10% chance of being spurious.

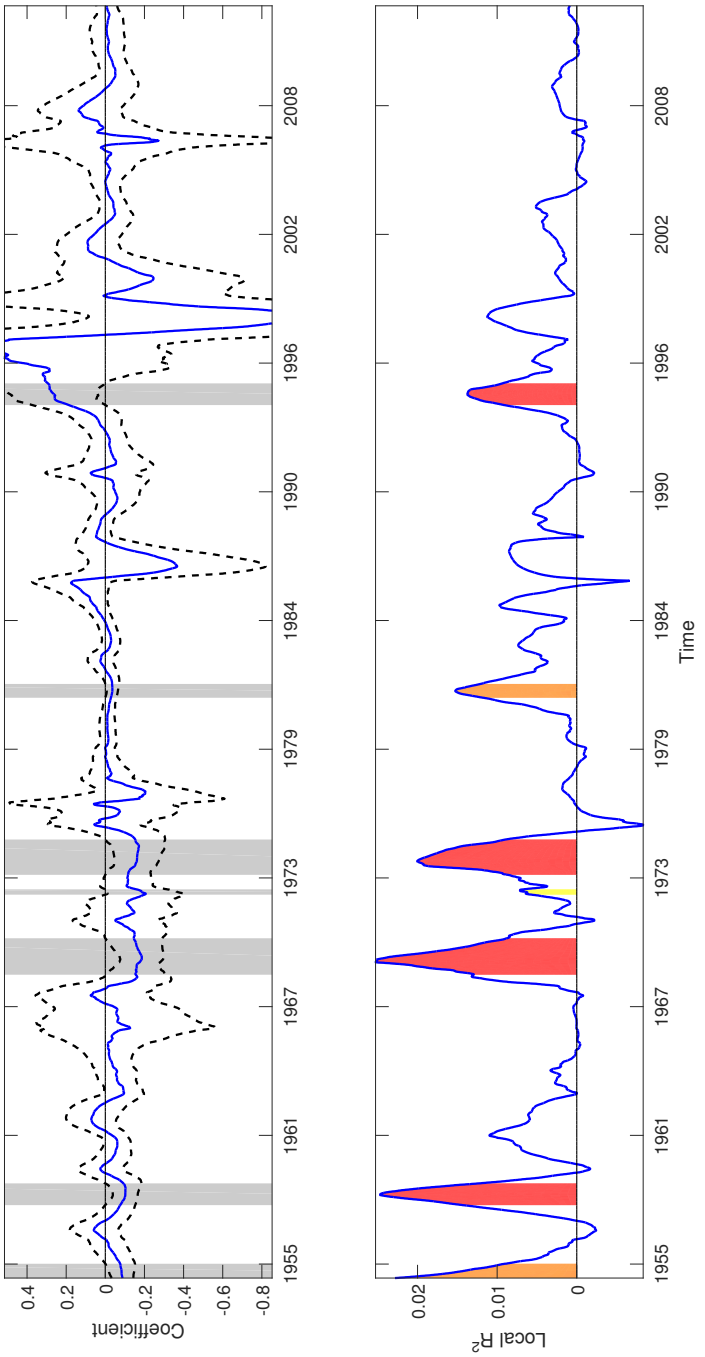


Figure 3.2. Local Return Predictability from the T-bill Rate (Zero Coefficient Benchmark)

Local return predictability from the T-bill rate. The top panel in this figure plots non-parametric kernel estimates of the local slope coefficient from a regression of daily excess stock returns on the lagged T-bill rate. Dashed lines represents plus or minus two standard error bands. The bottom panel plots the local $\bar{R}^2$ measure with shaded areas tracking periods identified as pockets of return predictability using a 5% critical value. The shaded areas represent the integrated $\bar{R}^2$ inside pockets with areas colored in red representing pockets that have less than a 5% chance of being spurious, areas colored in orange representing pockets that have between a 5% and a 10% chance of being spurious, and areas colored in yellow representing pockets that have more than 10% chance of being spurious.

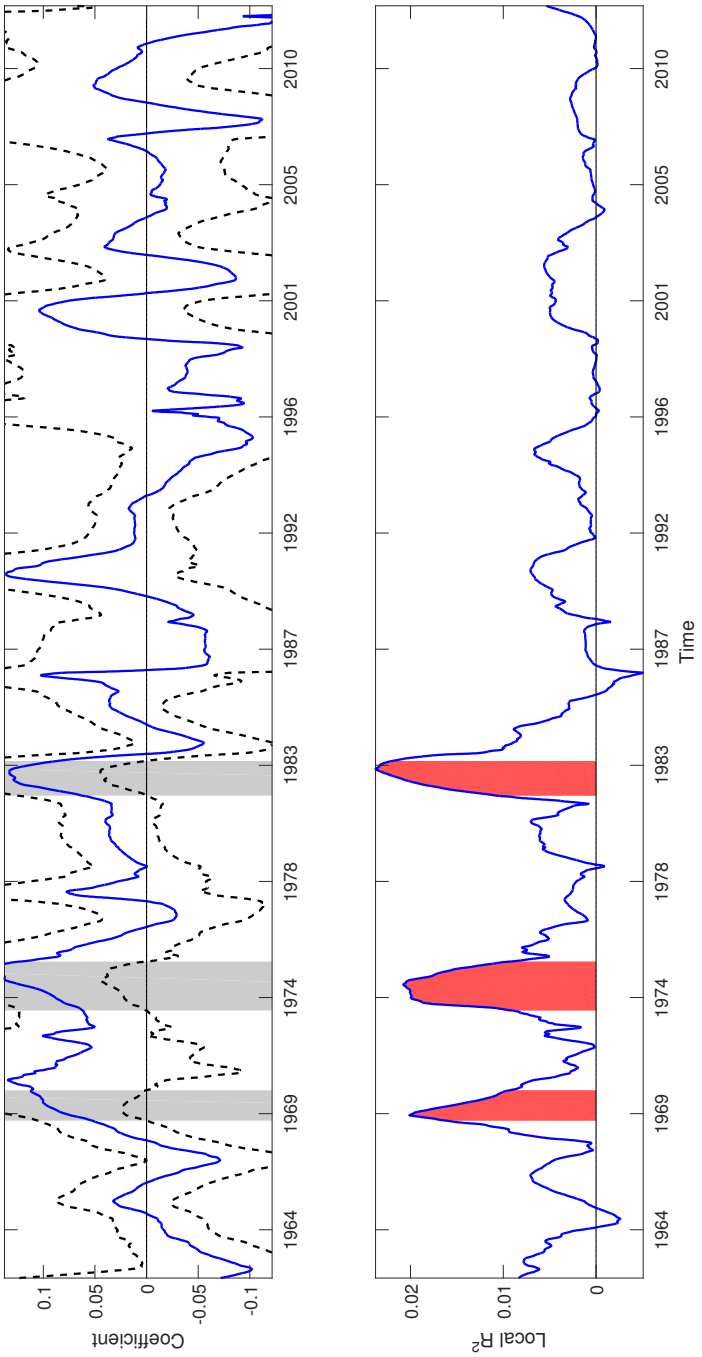


Figure 3.3. Local Return Predictability from the Term Spread (Zero Coefficient Benchmark)

Local return predictability from the term spread. The top panel in this figure plots non-parametric kernel estimates of the local slope coefficient from a regression of daily excess stock returns on the lagged term spread. Dashed lines represents plus or minus two standard error bands. The bottom panel plots the local $\bar{R}^2$ measure with shaded areas tracking periods identified as pockets of return predictability using a 5% critical value. The shaded areas represent the integrated $\bar{R}^2$ inside pockets with areas colored in red representing pockets that have less than a 5% chance of being spurious, areas colored in orange representing pockets that have between a 5% and a 10% chance of being spurious, and areas colored in yellow representing pockets that have more than 10% chance of being spurious.

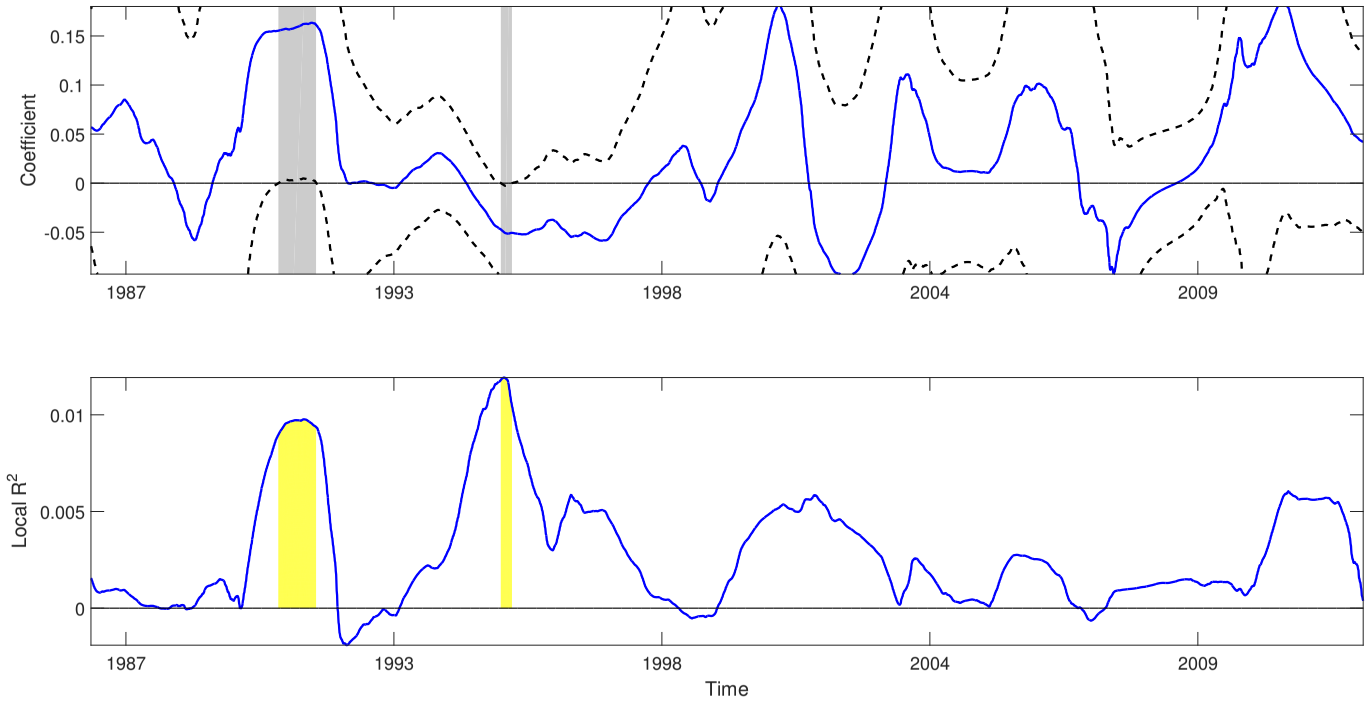


Figure 3.4. Local Return Predictability from the Corporate Spread (Zero Coefficient Benchmark)

Local return predictability from the corporate spread. The top panel in this figure plots non-parametric kernel estimates of the local slope coefficient from a regression of daily excess stock returns on the lagged dividend yield. Dashed lines represents plus or minus two standard error bands. The bottom panel plots the local$\bar{R}^2$ measure with shaded areas tracking periods identified as pockets of return predictability using a 5% critical value. The shaded areas represent the integrated $\bar{R}^2$ inside pockets with areas colored in red representing pockets that have less than a 5% chance of being spurious, areas colored in orange representing pockets that have between a 5% and a 10% chance of being spurious, and areas colored in yellow representing pockets that have more than 10% chance of being spurious.

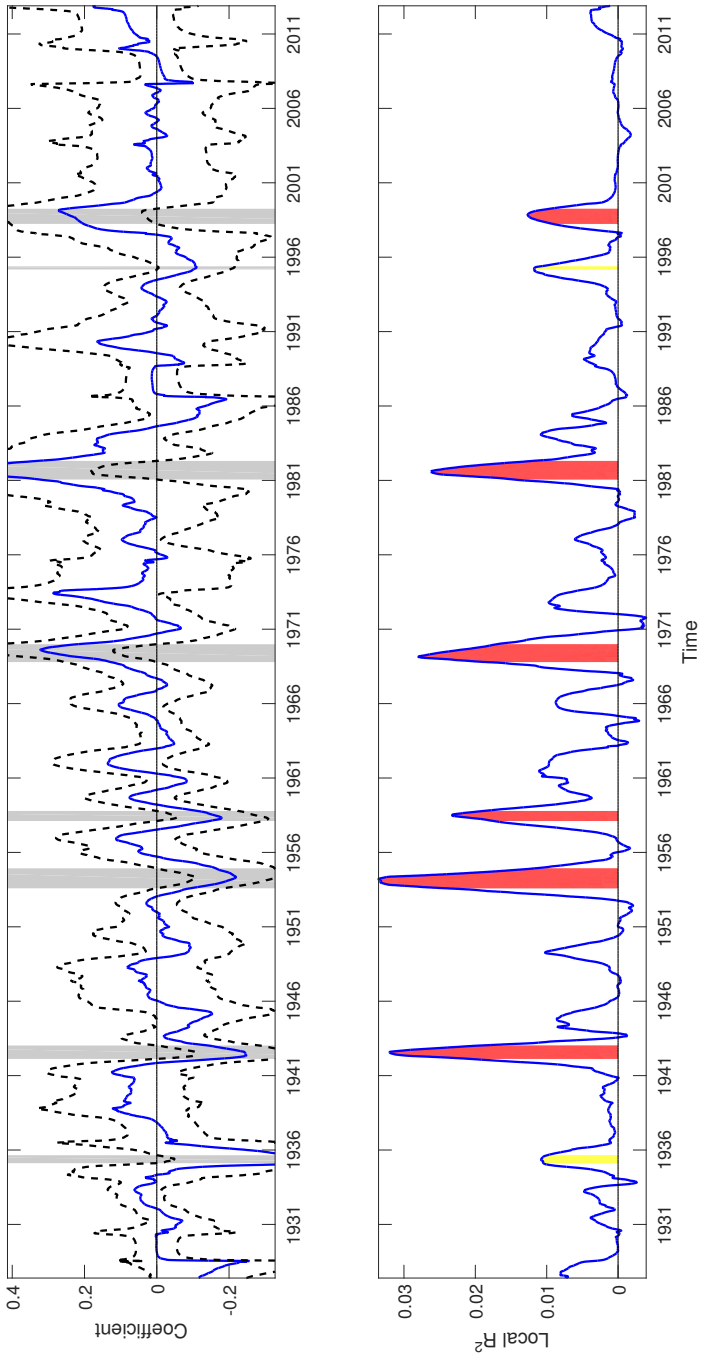


Figure 3.5. Local Return Predictability from the Realized Variance (Zero Coefficient Benchmark)

Local return predictability from the realized variance. The top panel in this figure plots non-parametric kernel estimates of the local slope coefficient from a regression of daily excess stock returns on the lagged realized variance. Dashed lines represents plus or minus two standard error bands. The bottom panel plots the local R^2 measure with shaded areas tracking periods identified as pockets of return predictability using a 5% critical value. The shaded areas represent the integrated R^2 inside pockets with areas colored in red representing pockets that have less than a 1% chance of being spurious, areas colored in orange representing pockets that have between a 5% and a 10% chance of being spurious, and areas colored in yellow representing pockets that have more than 10% chance of being spurious.

3.8 Acknowledgements

Chapter 3, in full, is currently being prepared for submission for publication of the material. Farmer, Leland E.; Schmidt, Lawrence D.W.; Timmermann, Allan. The dissertation author was a primary author of this paper.

Appendix

1.A Proofs for Chapter 1

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

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

Let $I_A \equiv \{ m \,|x_{m,M} \in A \}$ be the set of indices of the points in $\mathscr{X}_M$ contained in A. First consider the case where $t > s$.

$$\bar{\mathbb{P}}_\theta \left(X_{t,M} \in A \left| \boldsymbol{X}_{r,M}^{t-1}, \boldsymbol{Y}_r^s \right. \right) = \bar{\mathbb{P}}_\theta \left(X_{t,M} \in A \left| X_{t-1,M} = x \right. \right) = \sum_{m' \in I_A} P_{\theta,M} \left(m, m' \right)$$

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

$$\begin{aligned} & \bar{\mathbb{P}}_\theta \left(X_{t,M} \in A \left| X_{t-1,M} = x, \boldsymbol{Y}_r^s \right. \right) \\ = & \sum_{m' \in I_A} P_{\theta,M} \left(m, m' \right) \bar{p}_{\theta,M} \left(\boldsymbol{Y}_t^s \left| X_{t,M} = x_{m',M} \right. \right) \\ \times & \left(\sum_{m'=1}^{M} P_{\theta,M} \left(m, m' \right) \bar{p}_{\theta,M} \left(\boldsymbol{Y}_t^s \left| X_{t,M} = x_{m',M} \right. \right) \right)^{-1} \end{aligned}$$

By assumption (B1),

$$
\begin{aligned}
&\bar{\mathbb{P}}_\theta\left(X_{t,M} \in A \,\middle|\, X_{t-1,M} = x, \boldsymbol{Y}_{t-1}^s\right) \\
&\geq \sum_{m' \in I_A} Q_M^- \bar{p}_{\theta,M}\left(\boldsymbol{Y}_t^s \,\middle|\, X_{t,M} = x_{m',M}\right) \times \left(\sum_{m'=1}^{M} Q_M^+ \bar{p}_{\theta,M}\left(\boldsymbol{Y}_t^s \,\middle|\, X_{t,M} = x_{m',M}\right)\right)^{-1} \\
&= \frac{Q_M^-}{Q_M^+} \sum_{m' \in I_A} \bar{p}_{\theta,M}\left(\boldsymbol{Y}_t^s \,\middle|\, X_{t,M} = x_{m',M}\right) \times \left(\sum_{m'=1}^{M} \bar{p}_{\theta,M}\left(\boldsymbol{Y}_t^s \,\middle|\, X_{t,M} = x_{m',M}\right)\right)^{-1} \\
&\geq Q_+^- \mu_{t,M}\left(\boldsymbol{Y}_t^s, A\right)
\end{aligned}
$$

where

$$
\mu_{t,M}\left(\boldsymbol{Y}_t^s, A\right) \equiv \sum_{m' \in I_A} \bar{p}_{\theta,M}\left(\boldsymbol{Y}_t^s \,\middle|\, X_{t,M} = x_{m',M}\right) \times \left(\sum_{m'=1}^{M} \bar{p}_{\theta,M}\left(\boldsymbol{Y}_t^s \,\middle|\, X_{t,M} = x_{m',M}\right)\right)^{-1}
$$

In the case $t > s$, it suffices to set $\mu_{t,M}\left(\boldsymbol{Y}_t^s, A\right) = \frac{\mu_{c,M}(A)}{\mu_{c,M}(\mathscr{X}_M)}$, where $\mu_{c,M}$ is counting measure on $\mathscr{X}_M$. □

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

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

Thus, for $t \geq 1$, by Corollary 1 and assumption (A3),

$$\begin{aligned}
&\left| p_{\theta,M}\left(Y_t \left| \boldsymbol{Y}_0^{t-1}, X_0 = x_0\right.\right) - \bar{p}_{\theta,M}\left(Y_t \left| \boldsymbol{Y}_0^{t-1}\right.\right)\right| \\
&= \left| \sum_{m=1}^{M} \sum_{m'=1}^{M} P_{\theta,M}\left(m,m'\right) g_\theta\left(Y_t \left| x_{m',M}\right.\right) \right. \\
&\quad \left. \times \left(\mathbb{P}_\theta\left(X_{t-1,M} = x_{m,M} \left| \boldsymbol{Y}_0^{t-1}, x_{0,M} = x_0\right.\right) - \bar{\mathbb{P}}_\theta\left(X_{t-1,M} = x_{m,M} \left| \boldsymbol{Y}_0^{t-1}\right.\right)\right)\right| \\
&\leq \rho^{t-1} \sup_{1\leq m\leq M} \sum_{m'=1}^{M} P_{\theta,M}\left(m,m'\right) g_\theta\left(Y_t \left| x_{m',M}\right.\right)
\end{aligned}$$

In addition, by assumption (B3),

$$\begin{aligned}
&p_{\theta,M}\left(Y_t \left| \boldsymbol{Y}_0^{t-1}, X_0 = x_0\right.\right) \\
&= \sum_{m=1}^{M} \sum_{m'=1}^{M} g_\theta\left(Y_t \left| x_{m',M}\right.\right) P_{\theta,M}\left(m,m'\right) \mathbb{P}_\theta\left(X_{t-1,M} = x_{m,M} \left| \boldsymbol{Y}_0^{t-1}, x_{0,M} = x_0\right.\right) \\
&\geq \sum_{m=1}^{M} \left(\inf_{1\leq m\leq M} \sum_{m'=1}^{M} P_{\theta,M}\left(m,m'\right) g_\theta\left(Y_t \left| x_{m',M}\right.\right) \right) \\
&\times \mathbb{P}_\theta\left(X_{t-1,M} = x_{m,M} \left| \boldsymbol{Y}_0^{t-1}, x_{0,M} = x_0\right.\right) \\
&= \inf_{1\leq m\leq M} \sum_{m'=1}^{M} P_{\theta,M}\left(m,m'\right) g_\theta\left(Y_t \left| x_{m',M}\right.\right)
\end{aligned}$$

The same inequality also holds for $\bar{p}_{\theta,M}\left(Y_t \left| \boldsymbol{Y}_0^{t-1}\right.\right)$. It follows from the identity

$$\left|\log x - \log y\right| \leq \left|x - y\right| / \min\left(x, y\right)$$

that

$$\begin{aligned}
&\left|\log p_{\theta,M}\left(Y_t \left| \boldsymbol{Y}_t^s, x_{0,M} = x_0\right.\right) - \log \bar{p}_{\theta,M}\left(Y_t \left| \boldsymbol{Y}_t^s\right.\right)\right| \\
&\leq \rho^{t-1} \frac{\sup_{1\leq m\leq M} \sum_{m'=1}^{M} P_{\theta,M}(m,m')\, g_\theta\left(Y_t \left| x_{m',M}\right.\right)}{\inf_{1\leq m\leq M} \sum_{m'=1}^{M} P_{\theta,M}(m,m')\, g_\theta\left(Y_t \left| x_{m',M}\right.\right)} \\
&\leq \rho^{t-1} \frac{1}{Q_+^-} \\
&\leq \frac{\rho^{t-1}}{1-\rho}
\end{aligned}$$

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

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

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

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

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

Thus, by Corollary 1

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

Similary, I can obtain a lower bound on each term in the difference above as in the proof of Lemma 2,

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

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

Note that by Assumption (A3),

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

Taking logs leads to the third inequality and concludes the proof.

□

Proof of Proposition 1. I wish to show that for any $A \in \mathscr{A}$, where $\mathscr{A}$ is the collection of continuity sets of X_t, that

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

By the Portmenteau Lemma, this is equivalent to showing that $X_{t,M} \xrightarrow{d} X_t$ as $M \to \infty$. From assumption (BT), I know that for any $A \in \mathscr{A}$,

$$\sup_{\theta \in \Theta} \sup_{x \in \mathscr{X}} \left| P_{\theta,M}(x,A) - P_{\theta}(x,A) \right| = O\left(h(M)\right)$$

I will use this assumption and the fact that X_t and $X_{t,M}$ are uniformly ergodic to establish a bound on the difference in probability assigned to the set A by the approximate and true ergodic distributions.

By applying the triangle inequality twice, I can bound the expression of interest by the difference between the ergodic distribution of X_t and its T-step ahead transition kernel, the difference between $X_{t,M}$ and its T-step ahead transition kernel, and the difference between the two T-step ahead transition kernels

$$\begin{aligned}
&\sup_{\theta \in \Theta} \left| \pi^X_{\theta,M}(A) - \pi^X_{\theta}(A) \right| \\
=\ &\sup_{\theta \in \Theta} \sup_{x \in \mathscr{X}} \left| \pi^X_{\theta,M}(A) - \pi^X_{\theta}(A) \right| \\
=\ &\sup_{\theta \in \Theta} \sup_{x \in \mathscr{X}} \left| \pi^X_{\theta,M}(A) - P^{(T)}_{\theta,M}(x,A) + P^{(T)}_{\theta,M}(x,A) - P^{(T)}_{\theta}(x,A) + P^{(T)}_{\theta}(x,A) - \pi^X_{\theta}(A) \right| \\
\leq\ &\sup_{\theta \in \Theta} \sup_{x \in \mathscr{X}} \left| \pi^X_{\theta,M}(A) - P^{(T)}_{\theta,M}(x,A) \right| + \sup_{\theta \in \Theta} \sup_{x \in \mathscr{X}} \left| P^{(T)}_{\theta,M}(x,A) - P^{(T)}_{\theta}(x,A) \right| \\
&+ \sup_{\theta \in \Theta} \sup_{x \in \mathscr{X}} \left| \pi^X_{\theta}(A) - P^{(T)}_{\theta}(x,A) \right|
\end{aligned}$$

Let ρ_1 and ρ_2 denote the uniform minorizing constants of the Markov chains X_t and $X_{t,M}$

respectively, and define $\rho_+ \equiv \max(\rho_1, \rho_2)$. By the definition of uniform ergodicity, the first and third terms in the above expression can be bounded by their uniform minorizing constants to the power T

$$\begin{aligned}
&\sup_{\theta\in\Theta} \left|\pi^X_{\theta,M}(A) - \pi^X_\theta(A)\right| \\
&\leq \rho_1^T + \sup_{\theta\in\Theta}\sup_{x\in\mathscr{X}} \left|P^{(T)}_{\theta,M}(x,A) - P^{(T)}_\theta(x,A)\right| + \rho_2^T \\
&\leq 2\rho_+^T + \sup_{\theta\in\Theta}\sup_{x\in\mathscr{X}} \left|P^{(T)}_{\theta,M}(x,A) - P^{(T)}_\theta(x,A)\right|
\end{aligned}$$

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

$$\begin{aligned}
&\sup_{\theta\in\Theta}\sup_{x\in\mathscr{X}} \left|P^{(T)}_{\theta,M}(x,A) - P^{(T)}_\theta(x,A)\right| \\
&= \sup_{\theta\in\Theta}\sup_{x\in\mathscr{X}} \left|P_{\theta,M}P^{(T-1)}_{\theta,M}(x,A) - P_{\theta,M}P^{(T-1)}_\theta(x,A) + P_{\theta,M}P^{(T-1)}_\theta(x,A) - \right. \\
&\quad \left. P_\theta P^{(T-1)}_\theta(x,A)\right| \\
&= \sup_{\theta\in\Theta}\sup_{x\in\mathscr{X}} \left|P_{\theta,M}\left(P^{(T-1)}_{\theta,M}(x,A) - P^{(T-1)}_\theta(x,A)\right) - \left(P_{\theta,M} - P_\theta\right)P^{(T-1)}_\theta(x,A)\right| \\
&\leq \sup_{\theta\in\Theta}\sup_{x\in\mathscr{X}} \left|P_{\theta,M}\left(P^{(T-1)}_{\theta,M}(x,A) - P^{(T-1)}_\theta(x,A)\right)\right| + \sup_{\theta\in\Theta}\sup_{x\in\mathscr{X}} \left|\left(P_{\theta,M} - P_\theta\right)P^{(T-1)}_\theta(x,A)\right| \\
&\leq \sup_{\theta\in\Theta}\sup_{x\in\mathscr{X}} \left|P_{\theta,M}(x,A)\right| \sup_{\theta\in\Theta}\sup_{x\in\mathscr{X}} \left|P^{(T-1)}_{\theta,M}(x,A) - P^{(T-1)}_\theta(x,A)\right| \\
&\quad + \sup_{\theta\in\Theta}\sup_{x\in\mathscr{X}} \left|P_{\theta,M}(x,A) - P_\theta(x,A)\right| \sup_{\theta\in\Theta}\sup_{x\in\mathscr{X}} \left|P^{(T-1)}_\theta(x,A)\right| \\
&\leq \sup_{\theta\in\Theta}\sup_{x\in\mathscr{X}} \left|P^{(T-1)}_{\theta,M}(x,A) - P^{(T-1)}_\theta(x,A)\right| + \sup_{\theta\in\Theta}\sup_{x\in\mathscr{X}} \left|P_{\theta,M}(x,A) - P_\theta(x,A)\right|
\end{aligned}$$

Applying this inequality recursively one can show that

$$\sup_{\theta\in\Theta}\sup_{x\in\mathscr{X}} \left|P^{(T)}_{\theta,M}(x,A) - P^{(T)}_\theta(x,A)\right| \leq T \sup_{\theta\in\Theta}\sup_{x\in\mathscr{X}} \left|P_{\theta,M}(x,A) - P_\theta(x,A)\right|$$

As long as the Lebesgue measure of the discrete sets $\sup_m \lambda(A_{m,M}) \to 0$ as $M \to \infty$, the set of discrete points $\{x_{m,M}\}$ will become dense in $\mathscr{X}$. That is, for any $x \in \mathscr{X}$ and $\varepsilon > 0$, $\exists M > 0$ and $1 \leq m \leq M$ s.t. $\|x - x_{m,M}\| < \varepsilon$. Thus the error in the expression above is bounded by the quality of approximation of the marginal distributions $P_\theta(x, \cdot)$.

Combining this last inequality with the bounds derived above, the original expression of interest can be bounded by

$$\begin{aligned} &\sup_{\theta \in \Theta} \left| \pi^X_{\theta,M}(A) - \pi^X_\theta(A) \right| \\ &\leq 2\rho_+^T + T \times O(h(M)) \end{aligned}$$

Letting T be a function of M, T_M, this means that $\exists\ 0 < c < \infty$ and $\exists\ N < \infty$ such that for all $M \geq N$,

$$\sup_{\theta \in \Theta} \left| \pi^X_{\theta,M}(A) - \pi^X_\theta(A) \right| \leq 2\rho_+^{T_M} + cT_M h(M)$$

Thus in order to control the above expression, it must be the case that T_M is chosen such that

$$2\rho_+^{T_M} + cT_M h(M) \to 0$$

as $M \to \infty$. Note that since $0 < \rho_+ < 1$, the term $2\rho_+^{T_M}$ decays exponentially fast as a function of T_M. Thus T_M can be chosen to be any function of M such that $2\rho_+^{T_M} \to 0$ and $T_M \times h(M) \to 0$ as $N \to \infty$. This will determine $h^*(M)$.

I now focus on the specific case of using the Farmer and Toda (2016) method with a trapezoidal rule quadrature rule. The trapezoidal rule has integration error which is $O\left(M^{-2/d}\right)$. Thus

$$\sup_{\theta\in\Theta}\left|\pi^X_{\theta,M}(A)-\pi^X_\theta(A)\right|\leq 2\rho_+^{T_M}+T_M\times O\left(M^{-2/d}\right)$$

This is equivalent to saying that $\exists\ 0<c<\infty$ and $\exists\ N<\infty$ such that for all $M\geq N$,

$$\sup_{\theta\in\Theta}\left|\pi^X_{\theta,M}(A)-\pi^X_\theta(A)\right|\leq 2\rho_+^{T_M}+cT_MM^{-2/d}$$

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

$$\begin{aligned}&2\rho_+^{T_M}+cT_MM^{-2/d}\\=&2\rho_+^{M^{\varepsilon/d}}+cM^{\varepsilon/d}M^{-2/d}\\=&2\rho_+^{M^{\varepsilon/d}}+cM^{(\varepsilon-2)/d}\end{aligned}$$

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

$$\sup_{\theta\in\Theta}\left|\pi^X_{\theta,M}(A)-\pi^X_\theta(A)\right|=O\left(M^{(\varepsilon-2)/d}\right)$$

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

$$\sup_{\theta\in\Theta}\left|\pi^X_{\theta,M}(A)-\pi^X_\theta(A)\right|=o\left(M^{(\delta-2)/d}\right)$$

However since the choice of ε was arbitrary, we have that the above holds for any $\delta>0$. This shows that for the case of the Farmer and Toda (2016) method with trapezoidal quadrature rule, $h^*(M)=M^{(\delta-2)/d}$ for any $\delta>0$.

□

Proof of Lemma 4. My goal is to show that the discrete approximation to the filter-

ing distribution converges in distribution to the true filtering distribution as $M \to \infty$. Define $\boldsymbol{X}^0_{-r,M} \equiv (X_{0,M}, \ldots, X_{-r,M})$ and $\boldsymbol{X}^0_{-r} \equiv (X_0, \ldots, X_{-r})$. I will first show that $\boldsymbol{X}^0_{-r,M} \xrightarrow{d} \boldsymbol{X}^0_{-r}$ for $r \geq 0$ as $M \to \infty$. I will then show this implies that the joint distribution $\left(\boldsymbol{X}^0_{-r,M}, \boldsymbol{Y}^0_{-r}\right) \xrightarrow{d} \left(\boldsymbol{X}^0_{-r}, \boldsymbol{Y}^0_{-r}\right)$ as $M \to \infty$. This will imply my desired result, that $X_{0,M} \left| \boldsymbol{Y}^0_{-r} \right. \xrightarrow{d} X_0 \left| \boldsymbol{Y}^0_{-r} \right.$ as $M \to \infty$.

Let $f_r : \mathscr{X}^{r+1} \to \mathbb{R}$ be a bounded, continuous function. I will establish convergence in distribution by showing that the expectation of $f_r\left(\boldsymbol{X}^0_{-r,M}\right)$ converges to the expectation of $f_r\left(\boldsymbol{X}^0_{-r}\right)$ as $M \to \infty$ for any bounded, continuous f_r. Define the difference of these two expectations as

$$\Delta_E \equiv \left| \overline{\mathbb{E}}_\theta \left[f_r\left(\boldsymbol{X}^0_{-r,M}\right) \right] - \overline{\mathbb{E}}_\theta \left[f_r\left(\boldsymbol{X}^0_{-r}\right) \right] \right| \tag{.1}$$

Recall the definitions of the transition kernel and ergodic distribution of the discrete approximation extended to $\mathscr{X}$

$$P_{\theta,M}(x,A) \equiv \sum_{m=1}^{M} \sum_{m'=1}^{M} P_{\theta,M}\left(m,m'\right) \mathbb{1}\left\{x \in A_{m,M}\right\} \mathbb{1}\left\{x_{m',M} \in A\right\} \tag{.2}$$

$$\pi^X_{\theta,M}(A) \equiv \sum_{m=1}^{M} \pi^X_{\theta,M}(m) \mathbb{1}\left\{x_{m,M} \in A\right\} \tag{.3}$$

This extended transition kernel $P_{\theta,M}$ and probability measure $\pi^X_{\theta,M}$ admit densities with respect to the measure μ on $\mathscr{X}$ which I will label as $q_{\theta,M}(\cdot\,|x) : \mathscr{X} \to \mathbb{R}$ for $x \in \mathscr{X}$, and $p_{\theta,M} : \mathscr{X} \to \mathbb{R}$. This allows me to replace summation by integration and keep the notation consistent across the discrete and continuous random variables.

I next factor the joint distribution of the sequence of $r+1$ X's into the product of the marginal distribution of the initial X and the distribution of the remaining X's conditional on the initial one. This is a straightforward application of Bayes' Rule.

$$\begin{aligned}\Delta_E = &\left| \int_{\mathscr{X}^{r+1}} f_r(x_0,\ldots,x_{-r})\, p_{\theta,M}(x_0,\ldots,x_{-r})\, dx_0 \cdots dx_{-r} \right. \\ &\left. - \int_{\mathscr{X}^{r+1}} f_r(x_0,\ldots,x_{-r})\, p_{\theta}(x_0,\ldots,x_{-r}.)\, dx_0 \cdots dx_{-r} \right| \\ = &\left| \int_{\mathscr{X}^{r+1}} f_r(x_0,\ldots,x_{-r})\, p_{\theta,M}(x_0,\ldots,x_{-r+1}\,|x_{-r})\, p_{\theta,M}(x_{-r})\, dx_0 \cdots dx_{-r} \right. \\ &\left. - \int_{\mathscr{X}^{r+1}} f_r(x_0,\ldots,x_{-r})\, p_{\theta}(x_0,\ldots,x_{-r+1}\,|x_{-r})\, p_{\theta}(x_{-r})\, dx_0 \cdots dx_{-r} \right|\end{aligned}$$

Since both $X_{0,M}$ and X_0 are first order Markov processes, these distributions can be further factored into the product of the initial distribution with the sequence of r one-step-ahead conditional distributions.

$$\begin{aligned}\Delta_E = &\left| \int_{\mathscr{X}^{r+1}} f_r(x_0,\ldots,x_{-r})\, q_{\theta,M}(x_0\,|x_{-1}) \cdots q_{\theta,M}(x_{-r+1}\,|x_{-r})\, p_{\theta,M}(x_{-r})\, dx_0 \cdots dx_{-r} \right. \\ &\left. - \int_{\mathscr{X}^{r+1}} f_r(x_0,\ldots,x_{-r})\, q_{\theta}(x_0\,|x_{-1}) \cdots q_{\theta}(x_{-r+1}\,|x_{-r})\, p_{\theta}(x_{-r})\, dx_0 \cdots dx_{-r} \right|\end{aligned}$$

Before proceeding, it will be useful to define the operators associated with the transition kernels P_θ and $P_{\theta,M}$ and their r-step counterparts $P_\theta^{(r)}$ and $P_{\theta,M}^{(r)}$. For a function $f : \mathscr{X}^2 \to \mathbb{R}$, define

$$(P_\theta f)(x) \equiv \int_{\mathscr{X}} f(x',x)\, q_\theta(x'\,|x)\, dx' \tag{.4}$$

$$(P_{\theta,M} f)(x) \equiv \int_{\mathscr{X}} f(x',x)\, q_{\theta,M}(x'\,|x)\, dx' \tag{.5}$$

For $r > 1$, $0 \le n < r$ and $f_r : \mathscr{X}^{r+1} \to \mathbb{R}$, define $f_{r-n} : \mathscr{X}^{r-n+1} \to \mathbb{R}$ as

$$f_{r-n}(x_0,\ldots,x_{-r+n}) \equiv f_r(x_0,\ldots,x_{-r+n};x_{-r+n-1},\ldots,x_{-r}) \tag{.6}$$

where arguments after the semi-colon are held fixed. In other words, f_{r-n} can be thought of as the function f_r where the last n arguments are held fixed. This then allows me to define the i-step versions of P_θ and $P_{\theta,M}$. Define the 1-step versions as

$$\left(P_\theta^{(1)} f_1\right)(x) \equiv \left(P_\theta f_1\right)(x) = \int_{\mathscr{X}} f_r\left(x_0, x; x_{-2}, \ldots, x_{-r}\right) q_\theta\left(x_0 \,|x\right) dx_0$$
$$\left(P_{\theta,M}^{(1)} f_1\right)(x) \equiv \left(P_{\theta,M} f_1\right)(x) = \int_{\mathscr{X}} f_r\left(x_0, x; x_{-2}, \ldots, x_{-r}\right) q_{\theta,M}\left(x_0 \,|x\right) dx_0$$

For $i = 2, \ldots, r$, define

$$\left(P_\theta^{(i)} f_i\right)(x) \equiv \left(P_\theta\left(P_\theta^{(i-1)} f_{i-1}\right)\right)(x) = \int_{\mathscr{X}} \left(P_\theta^{(i-1)} f_{i-1}\right)(x_{-i+1}) q_\theta\left(x_{-i+1} \,|x\right) dx_{-i+1} \tag{.7}$$

$$\left(P_{\theta,M}^{(i)} f_i\right)(x) \equiv \left(P_{\theta,M}\left(P_{\theta,M}^{(i-1)} f_{i-1}\right)\right)(x) = \int_{\mathscr{X}} \left(P_{\theta,M}^{(i-1)} f_{i-1}\right)(x_{-i+1})$$
$$q_{\theta,M}\left(x_{-i+1} \,|x\right) dx_{-i+1} \tag{.8}$$

These are distinct are from what is referred to as the i-step ahead transition kernel and its associated operator. An i-step ahead transition kernel characterizes the probability of transitioning from a point in the space to any measurable set in that space i periods ahead. However, this operator implicitly characterizes the probability of moving from any point in the space to any sequence of i measurable sets. In other words, it computes probabilities over paths of the Markov chain. Note that these i-step ahead operators can

be equivalently written in terms of one-step-ahead conditional densities as

$$\left(P_{\theta}^{(i)} f_i\right)(x) = \int_{\mathscr{X}^i} f_r\left(x_0,\ldots,x_{-i+1},x;x_{-i-1},\ldots,x_{-r}\right) q_{\theta}\left(x_0\,|x_{-1}\right)\cdots q_{\theta}\left(x_{-i+1}\,|x\right) dx_0\cdots dx_{-i+1}$$

$$\left(P_{\theta,M}^{(i)} f_i\right)(x) = \int_{\mathscr{X}^i} f_r\left(x_0,\ldots,x_{-i+1},x;x_{-i-1},\ldots,x_{-r}\right) q_{\theta,M}\left(x_0\,|x_{-1}\right)\cdots q_{\theta,M}\left(x_{-i+1}\,|x\right) dx_0\cdots dx_{-i+1}$$

With this new notation in hand, Δ_E can equivalenthly be rewritten in terms of the r-step operators as

$$\Delta_E = \left|\int_{\mathscr{X}} \left(P_{\theta,M}^{(r)} f_r\right)(x) p_{\theta,M}(x)\,dx - \int_{\mathscr{X}} \left(P_{\theta}^{(r)} f_r\right)(x) p_{\theta}(x)\,dx\right| \tag{.9}$$

Next, I seek to establish that Δ_E can be bounded by the sum of two terms, one involving the difference in one step ahead transition kernels, the second involving the difference in $r-1$-step operators. By assumption a bound is known for the difference in integrals with respect to the one-step-ahead conditional distributions. Thus I can iteratively apply this logic to obtain a bound for Δ_E in terms of only the one-step-ahead approximation error.

Replace integration with respect to $p_{\theta,M}$ by integration with respect to p_{θ} in the first term of (.9), and add and subtract the result from equation (.9). Then apply the

triangle inequality to bound Δ_E by the sum of two new terms.

$$\begin{aligned}\Delta_E = &\left|\int_{\mathscr{X}}\left(P^{(r)}_{\theta,M}f_r\right)(x)p_{\theta,M}(x)\,dx-\int_{\mathscr{X}}\left(P^{(r)}_{\theta,M}f_r\right)(x)p_{\theta}(x)\,dx\right.\\ &\left.+\int_{\mathscr{X}}\left(P^{(r)}_{\theta,M}f_r\right)(x)p_{\theta}(x)\,dx-\int_{\mathscr{X}}\left(P^{(r)}_{\theta}f_r\right)(x)p_{\theta}(x)\,dx\right| \qquad (.10)\\ \leq &\left|\int_{\mathscr{X}}\left(P^{(r)}_{\theta,M}f_r\right)(x)p_{\theta,M}(x)\,dx-\int_{\mathscr{X}}\left(P^{(r)}_{\theta,M}f_r\right)(x)p_{\theta}(x)\,dx\right|\\ &+\left|\int_{\mathscr{X}}\left(P^{(r)}_{\theta,M}f_r\right)(x)p_{\theta}(x)\,dx-\int_{\mathscr{X}}\left(P^{(r)}_{\theta}f_r\right)(x)p_{\theta}(x)\,dx\right| \qquad (.11)\end{aligned}$$

Consider the first term on the right hand side of inequality (.11). It is simply the difference of integrals of $\left(P^{(r)}_{\theta,M}f_r\right)(x)$ with respect to $p_{\theta,M}$ and p_θ respectively. By Proposition 1, this difference is $o\left(h^*(M)\right)$.

$$\begin{aligned}&\left|\int_{\mathscr{X}}\left(P^{(r)}_{\theta,M}f_r\right)(x)p_{\theta,M}(x)\,dx-\int_{\mathscr{X}}\left(P^{(r)}_{\theta,M}f_r\right)(x)p_{\theta}(x)\,dx\right|\\ \leq &\sup_{|f|\leq 1}\left|\int_{\mathscr{X}}f(x)\,p_{\theta,M}(x)\,dx-\int_{\mathscr{X}}f(x)p_{\theta}(x)\,dx\right|\\ = &\,2\left\|\pi^X_{\theta,M}-\pi^X_{\theta}\right\|_{TV}\\ = &\,o\left(h^*(M)\right)\end{aligned}$$

Next consider the second term on the right hand side of inequality (.11). By definition, the r-step operator can be written as the composition of the one-step-ahead operator with the $r-1$-step ahead operator.

$$\begin{aligned}&\left|\int_{\mathscr{X}}\left(P^{(r)}_{\theta,M}f_r\right)(x)p_{\theta}(x)\,dx-\int_{\mathscr{X}}\left(P^{(r)}_{\theta}f_r\right)(x)p_{\theta}(x)\,dx\right|\\ = &\left|\int_{\mathscr{X}}\left(P_{\theta,M}\left(P^{(r-1)}_{\theta,M}f_{r-1}\right)\right)(x)p_{\theta}(x)\,dx-\int_{\mathscr{X}}\left(P_{\theta}\left(P^{(r-1)}_{\theta}f_{r-1}\right)\right)(x)p_{\theta}(x)\,dx\right|\end{aligned} \qquad (.12)$$

Take the first term of equation (.12), replace the first $P_{\theta,M}$ by P_θ, and add and subtract it to equation (.12). Then apply the triangle inequality again.

$$
\begin{aligned}
&\left|\int_{\mathscr{X}}\left(P_{\theta,M}\left(P_{\theta,M}^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx-\int_{\mathscr{X}}\left(P_\theta\left(P_\theta^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx\right| \\
=&\left|\int_{\mathscr{X}}\left(P_{\theta,M}\left(P_{\theta,M}^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx-\int_{\mathscr{X}}\left(P_\theta\left(P_{\theta,M}^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx\right. \\
&\left.+\int_{\mathscr{X}}\left(P_\theta\left(P_{\theta,M}^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx-\int_{\mathscr{X}}\left(P_\theta\left(P_\theta^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx\right| \qquad (.13) \\
\leq&\left|\int_{\mathscr{X}}\left(P_{\theta,M}\left(P_{\theta,M}^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx-\int_{\mathscr{X}}\left(P_\theta\left(P_{\theta,M}^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx\right| \\
&+\left|\int_{\mathscr{X}}\left(P_\theta\left(P_{\theta,M}^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx-\int_{\mathscr{X}}\left(P_\theta\left(P_\theta^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx\right| \\
&\qquad (.14)
\end{aligned}
$$

The first term of inequality (.14) depends only on the approximation error of the one-step-ahead distribution, and the second term depends on the approximation error of the $r-1$-step ahead distribution. Define the function $\phi:\mathscr{X}^2\to\mathbb{R}$

$$
\begin{aligned}
\phi(x_{-r+1},x_{-r})\equiv\int_{\mathscr{X}^{r-1}}&f_r(x_0,\ldots,x_{-r})\,q_{\theta,M}(x_0\,|x_{-1})\cdots q_{\theta,M}(x_{-r+2}\,|x_{-r+1}) \\
&dx_0\cdots dx_{-r+2} \qquad (.15)
\end{aligned}
$$

Consider the first term on the right hand side of inequality (.14) and substitute in the definitions of the r-step operators in terms of one-step-ahead conditional distributions. I will show that this term can be thought of as the difference in integrals of the function ϕ

with respect to the one-step-ahead conditional distribution and its discrete approximation.

$$
\begin{aligned}
&\left|\int_{\mathscr{X}}\left(P_{\theta,M}\left(P_{\theta,M}^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx-\int_{\mathscr{X}}\left(P_\theta\left(P_{\theta,M}^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx\right| \\
=&\left|\int_{\mathscr{X}^{r+1}}f_r(x_0,\ldots,x_{-r})\,q_{\theta,M}(x_0\,|x_{-1})\cdots q_{\theta,M}(x_{-r+1}\,|x_{-r})\,p_\theta(x_{-r})\,dx_0\cdots dx_{-r}\right. \\
&-\int_{\mathscr{X}^{r+1}}f_r(x_0,\ldots,x_{-r})\,q_{\theta,M}(x_0\,|x_{-1})\cdots q_{\theta,M}(x_{-r+2}\,|x_{-r+1})\,q_\theta(x_{-r+1}\,|x_{-r}) \\
&\left.p_\theta(x_{-r})\,dx_0\cdots dx_{-r}\right| \\
=&\left|\int_{\mathscr{X}^2}\phi(x_{-r+1},x_{-r})\,q_{\theta,M}(x_{-r+1}\,|x_{-r})\,p_\theta(x_{-r})\,dx_{-r+1}dx_{-r}\right. \\
&\left.-\int_{\mathscr{X}^2}\phi(x_{-r+1},x_{-r})\,q_\theta(x_{-r+1}\,|x_{-r})\,p_\theta(x_{-r})\,dx_{-r+1}dx_{-r}\right|
\end{aligned}
$$

The term on the right hand side of this last equality can be rewritten in terms of the one-step-ahead operators P_θ and $P_{\theta,M}$

$$
\begin{aligned}
&\left|\int_{\mathscr{X}^2}\phi(x_{-r+1},x_{-r})\,q_{\theta,M}(x_{-r+1}\,|x_{-r})\,p_\theta(x_{-r})\,dx_{-r+1}dx_{-r}\right. \\
&\left.-\int_{\mathscr{X}^2}\phi(x_{-r+1},x_{-r})\,q_\theta(x_{-r+1}\,|x_{-r})\,p_\theta(x_{-r})\,dx_{-r+1}dx_{-r}\right| \\
=&\left|\int_{\mathscr{X}}\left(P_{\theta,M}\phi\right)(x)p_\theta(x)dx-\int_{\mathscr{X}}\left(P_\theta\phi\right)(x)p_\theta(x)dx\right|
\end{aligned}
$$

By proposition 1 the error between integrals with respect to $q_{\theta,M}$ and q_θ is $o(h^*(M))$.

$$
\begin{aligned}
&\left|\int_{\mathscr{X}}\left(P_{\theta,M}\phi\right)(x)p_\theta(x)dx-\int_{\mathscr{X}}\left(P_\theta\phi\right)(x)p_\theta(x)dx\right| \\
\leq&\sup_{|f|\leq 1}\left|\int_{\mathscr{X}}\left(P_{\theta,M}f\right)(x)p_\theta(x)dx-\int_{\mathscr{X}}\left(P_\theta f\right)(x)p_\theta(x)dx\right| \\
=&\,2\left\|\pi_\theta P_{\theta,M}-\pi_\theta P_\theta\right\|_{TV} \\
=&\,o(h^*(M))
\end{aligned}
$$

This leaves one term to bound to establish convergence in distribution of $\boldsymbol{X}_{-r,M}^0$ to $\boldsymbol{X}_{-r}^0$

as $M \to \infty$. Consider the second term on the right hand side of inequality (.14). Similar to the above argument, it will be useful to define a new function $\varphi : \mathscr{X}^{r-1} \to \mathbb{R}$

$$\varphi(x_0,\ldots,x_{-r+2}) = \int_{\mathscr{X}^2} f_r(x_0,\ldots,x_{-r})\, q_\theta(x_{-r+1}|x_{-r})\, p_\theta(x_{-r})\, dx_{-r+1}dx_{-r}$$

By using Fubini's theorem, I will show that by switching the order of integration in the second term on the right hand side of inequality (.14), this term can be expressed as the $(r-1)$-step operators $P_\theta^{(r-1)}$ and $P_{\theta,M}^{(r-1)}$ applied to the same function φ. I take the supremum over the conditioning value for x_{-r+2} in order to break the dependence of the terms not captured by φ on x_{-r+1}.

$$\begin{aligned}
&\left|\int_{\mathscr{X}} \left(P_\theta\left(P_{\theta,M}^{(r-1)} f_{r-1}\right)\right)(x) p_\theta(x)\, dx - \int_{\mathscr{X}} \left(P_\theta\left(P_\theta^{(r-1)} f_{r-1}\right)\right)(x) p_\theta(x)\, dx\right| \\
&= \Bigg|\int_{\mathscr{X}^{r+1}} f_r(x_0,\ldots,x_{-r})\, q_{\theta,M}(x_0|x_{-1})\cdots q_{\theta,M}(x_{-r+2}|x_{-r+1})\, q_\theta(x_{-r+1}|x_{-r}) \\
&\quad p_\theta(x_{-r})\, dx_0\cdots dx_{-r} \\
&\quad - \int_{\mathscr{X}^{r+1}} f_r(x_0,\ldots,x_{-r})\, q_\theta(x_0|x_{-1})\cdots q_\theta(x_{-r+2}|x_{-r+1})\, q_\theta(x_{-r+1}|x_{-r}) \\
&\quad p_\theta(x_{-r})\, dx_0\cdots dx_{-r}\Bigg| \\
&\leq \sup_{x\in\mathscr{X}} \Bigg|\int_{\mathscr{X}^{r+1}} f_r(x_0,\ldots,x_{-r})\, q_{\theta,M}(x_0|x_{-1})\cdots q_{\theta,M}(x_{-r+2}|x)\, q_\theta(x_{-r+1}|x_{-r}) \\
&\quad p_\theta(x_{-r})\, dx_0\cdots dx_{-r} \\
&\quad - \int_{\mathscr{X}^{r+1}} f_r(x_0,\ldots,x_{-r})\, q_\theta(x_0|x_{-1})\cdots q_\theta(x_{-r+2}|x)\, q_\theta(x_{-r+1}|x_{-r}) \\
&\quad p_\theta(x_{-r})\, dx_0\cdots dx_{-r}\Bigg| \\
&= \sup_{x\in\mathscr{X}} \Bigg|\int_{\mathscr{X}^{r-1}} \varphi(x_0,\ldots,x_{-r+2})\, q_{\theta,M}(x_0|x_{-1})\cdots q_{\theta,M}(x_{-r+2}|x)\, dx_0\cdots dx_{-r+2} \\
&\quad - \int_{\mathscr{X}^{r-1}} \varphi(x_0,\ldots,x_{-r+2})\, q_{\theta,M}(x_0|x_{-1})\cdots q_\theta(x_{-r+2}|x)\, dx_0\cdots dx_{-r+2}\Bigg|
\end{aligned}$$

Note that the last term in the right hand side of the above equality can be thought of as the $(r-2)$-step operator applied to the function φ

$$\begin{aligned}
&\sup_{x\in\mathscr{X}}\left|\int_{\mathscr{X}^{r-1}}\varphi\left(x_0,\ldots,x_{-r+2}\right)q_{\theta,M}\left(x_0\left|x_{-1}\right.\right)\cdots q_{\theta,M}\left(x_{-r+2}\left|x\right.\right)dx_0\cdots dx_{-r+2}\right.\\
&\left.-\int_{\mathscr{X}^{r-1}}\varphi\left(x_0,\ldots,x_{-r+2}\right)q_{\theta,M}\left(x_0\left|x_{-1}\right.\right)\cdots q_{\theta}\left(x_{-r+2}\left|x\right.\right)dx_0\cdots dx_{-r+2}\right|\\
=&\sup_{x\in\mathscr{X}}\left|\left(P_{\theta,M}^{(r-2)}\varphi\right)(x)-\left(P_{\theta}^{(r-2)}\varphi\right)(x)\right|
\end{aligned}$$

By appyling the same logic to this component as the $(r-1)$-step ahead component, it can be shown that the second term on the right hand side of inequality (.11) will be $o\left(r\times h^*\left(M\right)\right)$. Combining that result with the bound on the first term on the right hand side of inequality (.11) and returning to the original expression of interest, it can be seen that

$$\Delta_E\leq o\left(h^*\left(M\right)\right)+o\left(r\times h^*\left(M\right)\right)=o\left(r\times h^*\left(M\right)\right)$$

For any fixed r, this difference converges to 0 because by assumption $h^*\left(M\right)\to 0$ as $M\to\infty$.

Next I seek to show that $\left(\boldsymbol{X}_{-r,M}^0,\boldsymbol{Y}_{-r}^0\right)\xrightarrow{d}\left(\boldsymbol{X}_{-r}^0,\boldsymbol{Y}_{-r}^0\right)$ as $M\to\infty$. The joint density can be written and then factored as:

$$\begin{aligned}
&p_\theta\left(X_0,\ldots,X_{-r},Y_0,\ldots,Y_{-r}\right)\\
=&\,p_\theta\left(Y_0,\ldots,Y_{-r}\left|X_0,\ldots,X_{-r}\right.\right)p_\theta\left(X_0,\ldots,X_{-r}\right)\\
=&\,g_\theta\left(Y_0\left|X_0\right.\right)\cdots g_\theta\left(Y_{-r}\left|X_{-r}\right.\right)p_\theta\left(X_0,\ldots,X_{-r}\right)
\end{aligned}$$

The same factorization can be done for the discrete approximations. Consider the expectation of an arbitrary bounded, continuous function $f:\mathscr{X}^{r+1}\times\mathscr{Y}^{r+1}\to\mathbb{R}$. In order to establish convergence in distribution it is sufficient to establish the expectation

of any bounded, continuous function of the sequence of approximations converges to the expectation of the function of the limit. The difference in the expectations of the function f is given by

$$\left|\overline{\mathbb{E}}_\theta\left[f\left(\boldsymbol{X}^0_{-r,M},\boldsymbol{Y}^0_{-r}\right)\right]-\overline{\mathbb{E}}_\theta\left[f\left(\boldsymbol{X}^0_{-r},\boldsymbol{Y}^0_{-r}\right)\right]\right|$$

Define the new function $f^*:\mathscr{X}^{r+1}\to\mathbb{R}$ as:

$$f^*\left(\boldsymbol{x}^0_{-r}\right)\equiv\int_{\mathscr{Y}^{r+1}} f\left(\boldsymbol{x}^0_{-r},\boldsymbol{y}^0_{-r}\right)g_\theta\left(y_0\left|x_0\right.\right)\cdots g_\theta\left(y_{-r}\left|x_{-r}\right.\right)dy_0\cdots dy_{-r}$$

Since $g_\theta\left(\cdot\left|x\right.\right)$ is a continuous and bounded function, so is their $(r+1)$-fold product and thus their product with f. Furthermore, since integration is a continuous operator over the space $\mathscr{Y}^{r+1}$, it follows from $\boldsymbol{X}^0_{-r,M}\xrightarrow{d}\boldsymbol{X}^0_{-r}$ that $\left(\boldsymbol{X}^0_{-r,M},\boldsymbol{Y}^0_{-r}\right)\xrightarrow{d}\left(\boldsymbol{X}^0_{-r},\boldsymbol{Y}^0_{-r}\right)$ as $M\to\infty$. This implies that the filtering distribution $X_{0,M}\left|\boldsymbol{Y}^0_{-r}\right.\xrightarrow{d}X_0\left|\boldsymbol{Y}^0_{-r}\right.$ as $M\to\infty$. Making an analogous argument to that in Proposition 1, r can be chosen as a function of M, r_M, so as to maintain the convergence in distribution as both r and M go to infinity. The sufficient condition is that $r_M\times h^*(M)\to 0$ as $M\to\infty$.

Consider the initial object of interest

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|\ell_M(\theta)-\ell(\theta)\right| \\
=\;&\sup_{\theta\in\Theta}\left|\overline{\mathbb{E}}_{\theta^*}\left[\log\bar{p}_{\theta,M}\left(Y_0\left|\boldsymbol{Y}^0_{-\infty}\right.\right)-\log\bar{p}_\theta\left(Y_0\left|\boldsymbol{Y}^0_{-\infty}\right.\right)\right]\right| \\
\leq\;&\sup_{\theta\in\Theta}\overline{\mathbb{E}}_{\theta^*}\left[\left|\log\bar{p}_{\theta,M}\left(Y_0\left|\boldsymbol{Y}^0_{-\infty}\right.\right)-\log\bar{p}_\theta\left(Y_0\left|\boldsymbol{Y}^0_{-\infty}\right.\right)\right|\right] \\
\leq\;&\sup_{\theta\in\Theta}\overline{\mathbb{E}}_{\theta^*}\left[\frac{\left|\bar{p}_{\theta,M}\left(Y_0\left|\boldsymbol{Y}^0_{-\infty}\right.\right)-\bar{p}_\theta\left(Y_0\left|\boldsymbol{Y}^0_{-\infty}\right.\right)\right|}{\min\left(\bar{p}_{\theta,M}\left(Y_0\left|\boldsymbol{Y}^0_{-\infty}\right.\right),\bar{p}_\theta\left(Y_0\left|\boldsymbol{Y}^0_{-\infty}\right.\right)\right)}\right]
\end{aligned}$$

This quantity converges to 0 as $M\to\infty$ due to the convergence in distribution of the filtering distributions for infinite histories. When the Farmer and Toda (2016) method

with a trapezoidal quadrature rule is used,

$$\sup_{\theta\in\Theta}\left|\ell_M(\theta)-\ell(\theta)\right| = o\left(h^*(M)\right) = o\left(M^{-(2-\delta)/d}\right)$$

for $\delta > 0$, by arguments analogous to those made in proposition 1.

□

Proof of Lemma 5. I first establish that for any fixed $x \in \mathscr{X}_M$, r, and M, $\Delta_{0,r,M,x}(\theta)$ is continuous w.r.t. θ. By definition

$$\bar{p}_{\theta,M}\left(Y_0 \left| \boldsymbol{Y}_{-r}^{-1}, X_{-r,M} = x\right.\right) = \frac{\bar{p}_{\theta,M}\left(\boldsymbol{Y}_{-r+1}^{0} \left| Y_{-r}, X_{-r,M} = x\right.\right)}{\bar{p}_{\theta,M}\left(\boldsymbol{Y}_{-r+1}^{-1} \left| Y_{-r}, X_{-r,M} = x\right.\right)}$$

Note that for $s \in \{-1, 0\}$, and assuming $x = x_{m_{-r},M}$ without loss of generality,

$$\begin{aligned}
&\bar{p}_{\theta,M}\left(\boldsymbol{Y}_{-r+1}^{s} \left| Y_{-r}, X_{-r,M} = x\right.\right) \\
&= \sum_{m_{-r},\ldots,m_s} \left[P_{\theta,M}(m_{-r}, m_{-r+1}) \mathbb{1}\left\{x_{m_{-r}} = x\right\} \prod_{i=-r+2}^{s} P_{\theta,M}(m_{i-1}, m_i) \right. \\
&\quad \left. \prod_{i=-r+1}^{s} g_\theta\left(Y_i \left| X_i = x_{m_i,M}\right.\right) \right]
\end{aligned}$$

Thus $\bar{p}_{\theta,M}\left(\boldsymbol{Y}_{-r+1}^{s} \left| Y_{-r}, X_{-r,M} = x\right.\right)$ is continuous w.r.t. θ by continuity of $P_{\theta,M}$ and g_θ. Therefore the sequence $\left\{\Delta_{0,r,M,x}\right\}$ is also continuous w.r.t. θ because it is the composition of continuous functions. Since $\left\{\Delta_{0,r,M,x}(\theta)\right\}$ converges uniformly w.r.t. $\theta \in \Theta$, $\bar{\mathbb{P}}_{\theta^*}$-a.s., $\Delta_{0,\infty,M}(\theta)$ is also continuous w.r.t. $\theta \in \Theta$, $\bar{\mathbb{P}}_{\theta^*}$-a.s. The proof follows by using Lemma 3 and the dominated convergence theorem. □

Proof of Proposition 2. Using the triangle inequality,

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

The second term limits to 0 by Lemma 4. For the second term, note that by Lemma 2 it is sufficient to prove that

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

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

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

This term can be further decomposed as

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

where

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

Terms A and C are zero by Corollary 2, and by Lemma 5 and the ergodic theorem,

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

□

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

$$\begin{aligned}
&\ell_T\left(\hat{\theta}_{T,x_0},x_0\right)-\ell_T\left(\hat{\theta}_{T,M,x_{0,M}},x_0\right) \\
&= \ell_T\left(\hat{\theta}_{T,x_0},x_0\right)-\ell_T\left(\hat{\theta}_{T,M,x_{0,M}},x_0\right)+\ell_{T,M}\left(\hat{\theta}_{T,x_0},x_{0,M}\right)-\ell_{T,M}\left(\hat{\theta}_{T,x_0},x_{0,M}\right) \\
&\le \ell_T\left(\hat{\theta}_{T,x_0},x_0\right)-\ell_{T,M}\left(\hat{\theta}_{T,x_0},x_{0,M}\right)+\ell_{T,M}\left(\hat{\theta}_{T,M,x_{0,M}},x_{0,M}\right)-\ell_T\left(\hat{\theta}_{T,M,x_{0,M}},x_0\right)
\end{aligned}$$

Note that it is thus sufficient to show that for any $\theta\in\Theta$,

$$\mathbb{P}_\theta\left(\left|\ell_{T,M}\left(\theta,x_{0,M}\right)-\ell_T\left(\theta,x_0\right)\right|\ge\varepsilon\right)\to 0$$

as $T \to \infty$ and $M \to \infty$ at appropriate rates. It is possible to decompose this probability as follows:

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

Theorem 14 from Douc, Moulines, Olsson, and Van Handel (2011) states that for any V_θ-uniformly ergodic state process with transition kernel P_θ, $f : \mathscr{Y}^{s+1}$ with $\|f\|_\infty < \infty$, there exists a constant $K < \infty$ such that

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

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

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

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

For the third term, it follows from Lemma 4 that

$$|\ell_M(\theta) - \ell(\theta)| = o(h^*(M))$$

and thus

$$T|\ell_M(\theta) - \ell(\theta)| = o(T \times h^*(M))$$

Returning to the original expression of interest

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

for $T \to \infty$, $M \to \infty$, and $T \times h^*(M) \to 0$. This ensures that my proposed estimator satisfies condition (iii) of Theorem 7 from Douc, Moulines, and Ryden (2004). □

1.B Discretizing Nonlinear, Non-Gaussian Markov Processes with Exact Conditional Moments

This appendix briefly summarizes the method for discretizting stochastic processes proposed in Farmer and Toda (2016).

Consider the time-homogeneous first-order Markov process

$$\mathbb{P}(X_t \leq x' | X_{t-1} = x) = F(x'|x),$$

where X_t is the random vector of state variables and $F(\cdot|x)$ is a cumulative distribution function (CDF) that determines the distribution of $X_t = x'$ given $X_{t-1} = x$. The dynamics

of any Markov process are completely characterized by its Markov transition kernel. In the case of a discrete state space, this transition kernel is simply a matrix of transition probabilities, where each row corresponds to a conditional distribution. One can discretize the continuous process X_t by applying the Tanaka and Toda (2013) method to each conditional distribution separately.

More concretely, suppose that one has a set of grid points $D_M = \{x_m\}_{m=1}^M$ and an initial coarse approximation $Q = (q_{mm'})$, which is an $M \times M$ probability transition matrix. Additionally, suppose one wants to match some conditional moments of X_t, represented by the moment defining function $T(x)$. The exact conditional moments when the current state is $X_{t-1} = x_m$ are

$$\overline{T}_m = \mathbb{E}\left[T(X_t)\,|X_{t-1} = x_m\right] = \int T(x) dF(x\,|x_m),$$

where the integral is over x, fixing $X_{t-1} = x_m$. (If these moments do not have explicit expressions, highly accurate quadrature formulas can be used to compute them.) By Theorem 2.1 in Farmer and Toda (2016), these moments can be matched exactly by solving the optimization problem

$$\begin{aligned} \min_{\{p_{mm'}\}_{m'=1}^M} \quad & \sum_{m'=1}^M p_{mm'} \log \frac{p_{mm'}}{q_{mm'}} \\ \text{subject to} \quad & \sum_{m'=1}^M p_{mm'} T(x_{m'}) = \overline{T}_m,\ \sum_{m'=1}^M p_{mm'} = 1,\ p_{mm'} \geq 0 \end{aligned} \tag{.16}$$

for each $m = 1, 2, \ldots, M$, or equivalently the dual problem

$$\min_{\lambda \in \mathbb{R}^L} \sum_{m'=1}^M q_{mm'} \mathrm{e}^{\lambda'(T(x_{m'}) - \overline{T}_m)}. \tag{.17}$$

(.17) has a unique solution if and only if the regularity condition

$$\overline{T}_m \in \text{int co } T(D_M) \tag{.18}$$

holds. Furthermore, if the dual problem has a unique solution λ_m, then the solution to the primal problem (.16) is given by

$$p_{mm'} = \frac{q_{mm'} e^{\lambda_m' \left(T\left(x_{m'}\right) - \overline{T}_m \right)}}{\sum_{m'=1}^{M} q_{mm'} e^{\lambda_m' \left(T\left(x_{m'}\right) - \overline{T}_m \right)}} \tag{.19}$$

Lastly, define the errors associated with the moment matching as:

$$\varepsilon_m \equiv \sum_{m'=1}^{M} p_{mm'} T\left(x_{m'}\right) - \overline{T}_m \tag{.20}$$

The procedure for constructing the finite-state Markov chain approximation to X_t is summarized in Algorithm 2 below.

Algorithm 4: Discretization of Markov Processes

1. Select a discrete set of points $D_M = \{x_m\}_{m=1}^{M}$ and an initial approximation $Q = (q_{mm'})$.
2. Select a moment defining function $T(x)$ and corresponding exact conditional moments $\left\{\overline{T}_m\right\}_{m=1}^{M}$. If necessary, approximate the exact conditional moments with highly accurate numerical integrals. Set $m \rightsquigarrow 1$ and define an error tolerance $\kappa > 0$.
3. Solve minimization problem (.17) and store the resulting solution λ_m.
4. Compute ε_m using (.20). If $\|\varepsilon_m\|_\infty < \kappa$, move to step 5. If not, select a smaller set of moments to match and return to step 3.
5. Compute the conditional probabilities corresponding to row m of $P = (p_{mm'})$ using (.19). Set $m \rightsquigarrow m+1$. If $m \leq M$, move to step 3, otherwise move to step 6.
6. Collect the computed conditional probability measures in the matrix $P = (p_{mm'})$.

The resulting finite-state Markov chain approximation to X_t takes values in the

set D_M and has associated transition matrix P. Since the dual problem (.17) is an unconstrained convex minimization problem with a typically small number of variables, standard Newton type algorithms can be applied. Furthermore, since the probabilities (.19) are strictly positive by construction, the transition probability matrix $P = (p_{mm'})$ is a strictly positive matrix, so the resulting Markov chain is stationary and uniformly ergodic by construction.

2.A Proofs for Chapter 2

Proof of Theorem 4. 1. The constraint set in (P) is nonempty if and only if $\overline{T} \in \mathrm{co}\, T(D_N)$. Since $\mathrm{co}\, T(D_N)$ is nonempty, compact, convex, and the objective function in (P) is strictly convex (a well-known property of the Kullback-Leibler information), the claim is trivial.

2. The "if" part is Theorem 2 of Tanaka and Toda 2013. To show the "only if" part, suppose that λ_N is a solution to (D). Since the objective function is differentiable, by taking the derivative we get

$$\overline{T} - \sum_{n=1}^{N} \frac{q_n \mathrm{e}^{\lambda_N' T(x_n)}}{\sum_{n=1}^{N} q_n \mathrm{e}^{\lambda_N' T(x_n)}} T(x_n) = 0.$$

Letting p_n as in (2.3), this equation shows $\overline{T} = \sum_{n=1}^{N} p_n T(x_n)$, $\sum_{n=1}^{N} p_n = 1$, and $p_n > 0$ for all n. Therefore $\overline{T} \in \mathrm{int}\, \mathrm{co}\, T(D_N)$.

3. This is Theorem 1 of Tanaka and Toda 2013. □

Proof of Theorem 5. Special case of the following theorem by setting $\Sigma_t = D$ (constant). □

Theorem 2.A.1 . Let $\{ y_t \}$ be a VAR with stochastic volatility

$$y_t = Ay_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim (0, \Sigma_{t-1}),$$

where all eigenvalues of A are less than 1 in absolute value and $\{ \Sigma_t \}$ is an exogenous, stationary, ergodic finite-state Markov chain. Let $z_t = (y_t, \Sigma_t)$. Suppose that $z_t^d = (y_t^d, \Sigma_t)$ is a stationary and ergodic Markov chain approximation of z_t such that the conditional mean and variance of y_t are exact, so

$$\begin{aligned} \mathbb{E}\left[y_t^d \;\middle|\; z_{t-1}^d \right] &= \mathbb{E}\left[y_t \;\middle|\; z_{t-1}^d \right] = Ay_{t-1}^d, \\ \mathrm{Var}\left[y_t^d \;\middle|\; z_{t-1}^d \right] &= \mathrm{Var}\left[y_t \;\middle|\; z_{t-1}^d \right] = \Sigma_{t-1}. \end{aligned}$$

Then the *unconditional* mean, variance, and all autocovariance (hence the spectrum) of $\{ y_t \}$ and $\{ y_t^d \}$ are identical, and so are all k-step ahead conditional mean and variance.

Proof of Theorem 2.A.1. By assumption, $\Sigma := \mathbb{E}[\Sigma_t]$ exists and $\mathbb{E}[y_t] = 0$.

Define the discretized error term $\varepsilon_t^d := y_t^d - Ay_{t-1}^d$. First we prove that the first two unconditional moments are exact. Since by assumption the conditional mean is exact, we have

$$\mathbb{E}\left[\varepsilon_t^d \;\middle|\; z_{t-1}^d \right] = \mathbb{E}\left[y_t^d \;\middle|\; z_{t-1}^d \right] - Ay_{t-1}^d = Ay_{t-1}^d - Ay_{t-1}^d = 0,$$

and hence $\mathbb{E}[\varepsilon_t^d] = 0$. Since by assumption $\{ y_t^d \}$ is stationary and the eigenvalues of A are less than 1 in absolute value, taking the unconditional expectation of both sides of $y_t^d = Ay_{t-1}^d + \varepsilon_t^d$, we get $\mathbb{E}[y_t^d] = 0$. Therefore the unconditional mean is exact. To

compute the variance, note that

$$\begin{aligned} y_t^d (y_t^d)' &= (Ay_{t-1}^d + \varepsilon_t^d)(Ay_{t-1}^d + \varepsilon_t^d)' \\ &= Ay_{t-1}^d (y_{t-1}^d)' A' + Ay_{t-1}^d (\varepsilon_t^d)' + \varepsilon_t^d (y_{t-1}^d)' A' + \varepsilon_t^d (\varepsilon_t^d)'. \end{aligned}$$

Since $\mathbb{E}\left[\varepsilon_t^d \mid z_{t-1}^d\right] = 0$ and the conditional variance is exact, taking the conditional expectation we obtain

$$\mathbb{E}\left[y_t^d (y_t^d)' \,\middle|\, z_{t-1}^d\right] = Ay_{t-1}^d (y_{t-1}^d)' A' + \Sigma_{t-1}.$$

Taking the unconditional expectation, using the law of iterated expectations, and noting that $\{y_t^d\}$ is stationary, we get

$$\begin{aligned} \operatorname{Var}[y_t^d] &= \mathbb{E}\left[\mathbb{E}\left[y_t^d (y_t^d)' \,\middle|\, z_{t-1}^d\right]\right] = A\mathbb{E}[y_{t-1}^d (y_{t-1}^d)']A' + \mathbb{E}[\Sigma_{t-1}] \\ &= A \operatorname{Var}[y_{t-1}^d] A' + \Sigma = A \operatorname{Var}[y_t^d] A' + \Sigma. \end{aligned}$$

But the variance matrix of the true process $\{y_t\}$ satisfies the same equation. Since the eigenvalues of A are less than 1 in absolute value, the solution is unique. Therefore $\operatorname{Var}[y_t^d] = \operatorname{Var}[y_t]$.

Let $\Gamma(k) = \mathbb{E}[y_{t+k} y_t']$ be the true k-th order autocovariance matrix and $\Gamma^d(k) = \mathbb{E}[y_{t+k}^d (y_t^d)']$ be that of the discretized process. Multiplying $(y_t^d)'$ from the right to both sides of $y_{t+k+1}^d = Ay_{t+k}^d + \varepsilon_{t+k+1}^d$ and taking expectations, we obtain $\Gamma^d(k+1) = A\Gamma^d(k)$. By iteration, we get $\Gamma^d(k) = A^k \Gamma^d(0)$. Similarly, $\Gamma(k) = A^k \Gamma(0)$. Since $\Gamma(0) = \operatorname{Var}[y_t] = \operatorname{Var}[y_t^d] = \Gamma^d(0)$, it follows that $\Gamma^d(k) = \Gamma(k)$ for all k. Therefore all autocovariances of $\{y_t\}$ are exact, and so is the spectrum.

To evaluate the k-step ahead conditional moments, note that

$$y^d_{t+k} = \varepsilon^d_{t+k} + \cdots + A^{k-1}\varepsilon^d_{t+1} + A^k y^d_t.$$

Since $\{ y^d_t \}$ is a Markov process, we have

$$\mathbb{E}\left[\varepsilon^d_{t+j} \,\middle|\, z^d_t\right] = \mathbb{E}\left[\mathbb{E}\left[\varepsilon^d_{t+j} \,\middle|\, y^d_{t+j-1}\right] \,\middle|\, z^d_t\right] = 0$$

for any $j \geq 1$. Therefore $\mathbb{E}\left[y^d_{t+k} \,\middle|\, z^d_t\right] = A^k y^d_t$, so the k-step ahead conditional mean is exact. The proof for the conditional variance is analogous. □

Remark. If the conditional variance of ε_t is unknown at $t-1$, say $\varepsilon_t \sim (0, \Sigma_t)$, then the same result holds by replacing Σ_{t-1} in the proof by $\mathbb{E}[\Sigma_t \mid \Sigma_{t-1}]$.

Proof of Proposition 1. Let $\rho(M)$ denote the spectral radius of the matrix M. Since $\rho(|A|) < 1$, there exists $\delta > 0$ such that $\alpha := \rho(\delta I + |A|) < 1$. By the Perron-Frobenius theorem, $\delta I + |A|$ has a strictly positive eigenvector $v = (v_1, \ldots, v_K) \gg 0$. Take a tensor grid D_N with convex hull $\operatorname{co} D_N = [-v_1, v_1] \times \cdots \times [-v_K, v_K]$. Let $\bar{y}_n$ be any grid point of D_N, and let $T(x) = x$ be the moment defining function for the conditional mean (therefore it is the identity map). Then $T(D_N) = D_N$, and

$$\overline{T}_n := \mathbb{E}[T(y_t) \mid y_{t-1} = \bar{y}_n] = \mathbb{E}[y_t \mid y_{t-1} = \bar{y}_n] = A\bar{y}_n.$$

Taking absolute values element-by-element, since $0 < \alpha < 1$ we get

$$\left|\overline{T}_n\right| \leq |A|\,|\bar{y}_n| \leq |A|\,v \leq (\delta I + |A|)v = \alpha v \ll v,$$

so $\overline{T}_n \in \operatorname{int} \operatorname{co} T(D_N)$. □

Proof of Proposition 2. Let $D = \{\bar{x}_n\}_{n=1}^N$ be the set of grid points and $M = \max_n |\bar{x}_n|$. Suppose $x_{t-1} = x$, where $x \in D$. By symmetry, without loss of generality we may assume $x \geq 0$. Then the conditional first and second (uncentered) moments of x_t are ρx and $(\rho x)^2 + 1$, respectively. The moment defining function is $T(x) = (x, x^2)$. By Theorem 4, it suffices to show that $(\rho x, (\rho x)^2 + 1) \in \operatorname{int} \operatorname{co} T(D)$.

Define the points $P = (M, M^2)$, $Q = (-M, M^2)$, $X = (x, x^2)$, and $X' = (\rho x, (\rho x)^2 + 1)$. If $x = M$, in order for $X' \in \operatorname{int} \operatorname{co} T(D)$ it is necessary that X' lies below the segment PQ, so we need

$$(\rho M)^2 + 1 < M^2 \iff M > \frac{1}{\sqrt{1-\rho^2}},$$

which is condition (i) in Proposition 2. Therefore X' lies below PQ. Now take any $x \in D$ and set $\mu = \rho x$. Take two grid points $a_1 < a_2 \in D$ such that $\mu \in [a_1, a_2]$. Let $A_1 = (a_1, a_1^2)$ and $A_2 = (a_2, a_2^2)$. If X' lies above the segment A_1A_2, then X' is in the interior of the quadrilateral A_1A_2PQ, which is a subset of $\operatorname{co} T(D)$. Therefore it suffices to show that X' lies above A_1A_2. The equation of the straight line A_1A_2 is

$$y = \frac{a_2^2 - a_1^2}{a_2 - a_1}(x - a_1) + a_1^2 = (a_1 + a_2)(x - a_1) + a_1^2.$$

Therefore X' lies above A_1A_2 if and only if

$$\mu^2 + 1 > (a_1 + a_2)(\mu - a_1) + a_1^2 \iff (\mu - a_1)(a_2 - \mu) < 1. \tag{.21}$$

First, consider the case in which the maximum distance between neighboring points is $d < 2$. Take a_1, a_2 as neighboring points. By the arithmetic mean-geometric mean inequality, we have

$$(\mu - a_1)(a_2 - \mu) \leq \left(\frac{(\mu - a_1) + (a_2 - \mu)}{2}\right)^2 = \left(\frac{a_2 - a_1}{2}\right)^2 \leq (d/2)^2 < 1,$$

so (.21) holds. Next, we show (2.10). Setting $a_2 = x$ and $\mu = \rho x$ in (.21) and solving the inequality, a sufficient condition for existence is

$$\rho x = \mu \geq a_1 > \rho x - \frac{1}{(1-\rho)x},$$

which is (2.10) by setting $x = \bar{x}_n$ and $a_1 = \bar{x}_{n'}$. □

Proof of Corollary 3. Since the grid $\{\bar{x}_n\}_{n=1}^N$ spans from $-M$ to M and is even-spaced, the grid size is $d = \frac{2M}{N-1}$. Suppose that $M > \sigma = 1/\sqrt{1-\rho^2}$, so condition (i) of Proposition 2 holds. Note that the grid has at least three points $0, \pm M$, so $N \geq 3$.

$\rho \leq 1 - \frac{2}{N-1}$. By Proposition 2, it suffices to show $d < 2 \iff M < N-1$. Since $M \leq \sqrt{2}\sigma\sqrt{N-1}$ by assumption, it suffices to show

$$\frac{\sqrt{2}\sqrt{N-1}}{\sqrt{1-\rho^2}} < N-1 \iff \rho^2 < 1 - \frac{2}{N-1}.$$

But this inequality is trivial because $\rho^2 < \rho \leq 1 - \frac{2}{N-1}$.

$\rho > 1 - \frac{2}{N-1}$. Let $-M = \bar{x}_1 < \cdots < \bar{x}_N = M$ be the grid points. By Proposition 2, it suffices to show that (2.10) holds for all n such that $\bar{x}_n > 0$, which means that the interval $\left(\rho\bar{x}_n - \frac{1}{(1-\rho)\bar{x}_n}, \rho\bar{x}_n\right)$ contains a grid point. Since the length of this interval is $d_n := \frac{1}{(1-\rho)\bar{x}_n}$, if $d < d_n$, then the interval contains a grid point. Furthermore, since $d_n = \frac{1}{(1-\rho)\bar{x}_n}$ is decreasing in $\bar{x}_n$, it follows that if $d < d_n$ for some n, then $d < d_{n'}$ for all $n' < n$ such that $\bar{x}_{n'} > 0$.

Consider the point $n = N-1$. Since $d = \frac{2M}{N-1}$, we have $\bar{x}_{N-1} = M - d = M\frac{N-3}{N-1}$. Hence $d_{N-1} = \frac{1}{M(1-\rho)}\frac{N-1}{N-3}$. Therefore

$$d < d_{N-1} \iff \frac{2M}{N-1} < \frac{1}{M(1-\rho)}\frac{N-1}{N-3} \iff M < \frac{N-1}{\sqrt{2(1-\rho)(N-3)}}.$$

Since $M \leq \sigma\sqrt{N-1}$ by assumption, to show $d < d_{N-1}$, it suffices to show

$$\frac{\sqrt{N-1}}{\sqrt{1-\rho^2}} < \frac{N-1}{\sqrt{2(1-\rho)(N-3)}} \iff 1+\rho > \frac{2(N-3)}{N-1} \iff \rho > 1 - \frac{4}{N-1},$$

which trivially holds because $\rho > 1 - \frac{2}{N-1}$.

Therefore it remains to show that the two inequalities in (2.10) also hold for $n = N$, the boundary point. Take $n' = N-1$. Since $\bar{x}_{N-1} = M\frac{N-3}{N-1}$, the right inequality holds because

$$\bar{x}_{N-1} \leq \rho \bar{x}_N \iff M\frac{N-3}{N-1} \leq \rho M \iff \rho \geq 1 - \frac{2}{N-1},$$

which is trivial. The left inequality is equivalent to

$$\begin{aligned} \rho\bar{x}_N - \frac{1}{(1-\rho)\bar{x}_N} < \bar{x}_{N-1} &\iff \rho M - \frac{1}{(1-\rho)M} < M\frac{N-3}{N-1} \\ &\iff M^2\left(\rho - \frac{N-3}{N-1}\right) < \frac{1}{1-\rho}. \end{aligned}$$

Since $M \leq \sigma\sqrt{N-1}$, it suffices to show

$$\frac{N-1}{1-\rho^2}\left(\rho - \frac{N-3}{N-1}\right) < \frac{1}{1-\rho} \iff (N-1)\rho - (N-3) < 1+\rho \iff \rho < 1,$$

which is trivial. □

2.B Accuracy of Discretization

The accuracy of discretization has traditionally been evaluated by simulating the resulting Markov chain (Tauchen 1986; Gospodinov and Lkhagvasuren 2014). However, we think that such simulations have limited value, for the following reason. According to

Theorem 5, for VARs the first two population moments—both k-step ahead conditional and unconditional—are exact whenever the 1-step ahead conditional moments are exact. Since the population moments will be identical for such discretizations, any difference in the simulation performance must be due to sampling error.

A better approach is to directly compare the population moments of interest of the true process with those of the discretized Markov chains. For example, suppose that $(x_t, y_t)_{t=0}^{\infty} \subset \mathbb{R}^K \times \mathbb{R}$ is generated by some covariance stationary process such that

$$y_t = \beta' x_t + \varepsilon_t,$$

where $\mathbb{E}[x_t \varepsilon_t] = 0$. Then the population OLS coefficient is

$$\beta = \mathbb{E}[x_t x_t']^{-1} \mathbb{E}[x_t y_t].$$

If $(x_t^d, y_t^d)_{t=0}^{\infty}$ is a discretized Markov chain, then we can define its OLS coefficient by

$$\beta^d = \mathbb{E}[x_t^d (x_t^d)']^{-1} \mathbb{E}[x_t^d y_t^d],$$

where the expectation is taken under the ergodic distribution of the Markov chain. Then the bias of the discretization is $\beta^d - \beta$. Here we used the OLS coefficient as an example, but it can be any quantity that is defined through the population moments.

2.B.1 VAR(1)

As a concrete example, following Gospodinov and Lkhagvasuren 2014, consider the two-dimensional VAR(1) process

$$x_t = B x_{t-1} + \eta_t,$$

where

$$x_t = \begin{bmatrix} z_t \\ g_t \end{bmatrix}, \quad \eta_t = \begin{bmatrix} e_{z,t} \\ e_{g,t} \end{bmatrix}, \quad B = \begin{bmatrix} 0.9809 & 0.0028 \\ 0.0410 & 0.9648 \end{bmatrix}$$

and the shocks $e_{z,t}, e_{g,t}$ are uncorrelated, i.i.d. over time, and have standard deviations 0.0087 and 0.0262, respectively. The implied unconditional variance-covariance matrix is

$$\begin{bmatrix} \sigma_z^2 & \sigma_{zg} \\ \sigma_{zg} & \sigma_g^2 \end{bmatrix} = \begin{bmatrix} 0.00235 & 0.00241 \\ 0.00241 & 0.01274 \end{bmatrix}$$

and the eigenvalues of the coefficient matrix B are $\zeta_1 = 0.9863$ and $\zeta_2 = 0.9594$.

To evaluate the accuracy of discretization, we compute the Markov chain counterpart θ^d of the parameter $\theta = \sigma_z^2, \sigma_g^2, \sigma_{zg}, 1-\zeta_1, 1-\zeta_2$ and calculate the $\log_{10}$ relative bias $\log_{10}\left|\theta^d/\theta - 1\right|$ for various number of nodes in each dimension, $N = 5, 9, 15, 21$. For our method, we consider the even-spaced, quantile, and Gauss-Hermite quadrature grids, which we label as "ME-Even," "ME-Quant," and "ME-Quad," respectively. As a comparison, we consider the existing methods of Tauchen 1986, Tauchen and Hussey 1991 (TH), and Gospodinov and Lkhagvasuren 2014 (GL).[1] The GL method has two versions, one that is the VAR generalization of the Rouwenhorst method (referred to as GL0) and another that fine-tunes this method by targeting the first and second conditional moments (referred to as GL). Table 2.B.1 shows the results.

We can make a few observations from Table 2.B.1. First, as is well-known, the accuracy of discretization for the Tauchen and Tauchen-Hussey methods are poor, with relative bias of order about 10^0. Consistent with Gospodinov and Lkhagvasuren 2014, the GL methods improve upon earlier methods by several orders of magnitude.

[1] For the Tauchen method, we need to specify the grid spacing. To give it the best chance, following Kopecky and Suen 2010 we set the grid size proportional to the unconditional standard deviation of the VAR, and choose the constant of proportionality in order to make the unconditional variance as close to the true VAR as possible.

Table 2.B.1. $\log_{10}$ Relative Bias, VAR(1) Model

		Existing Methods				ME Methods		
N	Param.	Tauchen	TH	GL0	GL	Even	Quant	Quad
5	σ_z^2	-0.106	-0.052	-1.061	-1.500	-3.062	-1.465	-0.138
	σ_g^2	-0.106	-0.087	-0.918	-1.331	-2.369	-0.772	-0.138
	σ_{zg}	-0.001	-0.006	-4.394	-1.015	-2.408	-0.811	-0.138
	$1-\zeta_1$	1.641	1.178	-1.100	-1.235	-7.932	-8.178	-7.604
	$1-\zeta_2$	1.158	0.657	-1.865	-1.949	-9.303	-8.554	-8.538
9	σ_z^2	-0.106	-0.098	-1.004	-2.342	-9.321	-8.126	-0.379
	σ_g^2	-0.106	-0.166	-0.859	-2.156	-8.918	-9.372	-0.372
	σ_{zg}	-0.001	-0.021	-1.024	-1.915	-9.337	-7.787	-0.373
	$1-\zeta_1$	1.639	0.950	-1.904	-2.171	-8.690	-7.694	-8.410
	$1-\zeta_2$	1.157	0.396	-2.487	-2.713	-9.271	-9.077	-8.292
15	σ_z^2	-0.106	-0.170	-1.093	-3.730	-8.712	-9.085	-1.454
	σ_g^2	-0.106	-0.285	-0.944	-3.545	-8.783	-9.086	-0.760
	σ_{zg}	-0.001	-0.059	-1.052	-3.357	-10.015	-9.082	-0.800
	$1-\zeta_1$	1.639	0.696	-3.188	-3.664	-8.424	-8.774	-8.846
	$1-\zeta_2$	1.156	0.093	-3.650	-4.106	-8.729	-9.627	-9.790
21	σ_z^2	-0.106	-0.244	-1.174	-4.369	-9.539	-9.171	-8.966
	σ_g^2	-0.106	-0.403	-1.025	-4.140	-9.694	-8.538	-11.359
	σ_{zg}	-0.001	-0.114	-1.129	-4.240	-10.124	-8.524	-8.672
	$1-\zeta_1$	1.638	0.494	-4.517	-5.195	-9.373	-9.202	-8.589
	$1-\zeta_2$	1.156	-0.157	-4.894	-5.563	-9.665	-9.226	-9.301

Note: N: number of discrete points in each dimension; TH: Tauchen and Hussey 1991 method; GL, GL0: Gospodinov and Lkhagvasuren 2014 methods with or without moment targeting; ME: maximum entropy methods. The ME methods target the first two conditional moments. For ME-Even, the grid for the $\{y_t\}$ process (2.8) spans $[-\sigma\sqrt{N-1}, \sigma\sqrt{N-1}]$ in each dimension, where σ^2 is the smallest eigenvalue of the unconditional variance of $\{y_t\}$.

Second, the relative bias of ME-Even and ME-Quant is substantially smaller (of order about 10^{-9}, except when $N = 5$), which makes our method about 4 to 6 orders of magnitude more accurate than the GL methods. The reason why the bias is not exactly zero—although it should theoretically be zero if the regularity condition (2.7) holds—is because our method involves the numerical minimization of the dual function in (D'_n), in which we set the error tolerance to 10^{-10}.[2] Therefore this result suggests that for

[2]This point also explains why the accuracy does not monotonically improve as N gets larger for

this particular example, ME-Even and ME-Quant match all first and second conditional moments of the VAR.

Third, our method with Gauss-Hermite quadrature grid (ME-Quad) is poor for $N = 5, 9, 15$, especially for the unconditional variance. This is because, by construction, the quadrature method uses the Gauss-Hermite quadrature nodes of the *conditional* variance. When the process is highly persistent (as in this case since the spectral radius is $\zeta_1 = 0.9863$, which is close to 1), the *unconditional* variance is much larger than the conditional variance. Since the grid is much smaller than typical values of the true process, the regularity condition (2.7) may be violated and a solution to the dual problem may not exist. Note that ME-Quad is still quite accurate for the parameters $\theta = 1 - \zeta_1, 1 - \zeta_2$. The reason is that since $1 - \zeta_1, 1 - \zeta_2$ depend only on the coefficient matrix B and not on the variance, if the discretization method is able to match all first conditional moments, then the coefficient matrix will be exact. But B in this example satisfies the assumption of Proposition 1, so we can match $1 - \zeta_1, 1 - \zeta_2$ exactly.

While Table 2.B.1 shows the high accuracy of discretization by ME methods, is it computationally efficient? Table 2.B.2 shows the computing time for discretizing the VAR(1) process using various methods and number of grid points in each dimension. The TH and GL0 methods, which require no optimization, are clearly very fast. All other methods involve solving optimization problems. According to the table, the ME methods are faster than the GL method, probably because we solve the unconstrained dual problem using the Newton algorithm by supplying the analytical gradient and Hessian.

ME-Even and ME-Quant: since the relative bias is essentially the error tolerance (which is constant), it need not be monotonic in N. In contrast, since the relative bias is not zero for existing methods and ME-Quad, the accuracy of these methods monotonically improves with larger N.

Table 2.B.2. Computation Time for Discretizing the VAR(1) Process (in seconds)

	Existing Methods				ME Methods		
N	Tauchen	TH	GL0	GL	Even	Quant	Quad
5	0.490	0.008	0.013	0.559	0.684	0.616	1.017
9	1.198	0.016	0.047	2.107	1.397	1.268	1.851
15	3.487	0.049	0.265	5.910	3.212	3.031	3.525
21	8.324	0.078	0.730	12.074	5.561	5.616	6.301

Note: the table shows the computing time in seconds for discretizing the VAR(1) process in this section using a Windows 10 laptop computer with 2.2GHz Intel Core i5 processor. The Tauchen method matches the unconditional variance. The codes for the ME methods are available on our website discussed in Appendix E. The GL methods use the codes supplied in the online appendix of Gospodinov and Lkhagvasuren 2014.

2.B.2 AR(1) with Stochastic Volatility

Next, we consider the accuracy of the stochastic volatility discretization in Section 2.3.2. As a comparison, we construct an alternative approximation which uses the Rouwenhorst method to discretize the x_t process and the Tauchen method to discretize the conditional distributions $y_t|x_{t-1}, y_{t-1}$. This is the most logical choice since x is just and AR(1) process (for which the Rouwenhorst method is accurate) and there is no obvious way to discretize the y process except by the Tauchen method. We choose the spacing of the y process to target the unconditional variance σ_y^2. As in the simple autoregressive case, when discretizing the log variance process (x_t), we use $\sqrt{N-1}$ standard deviations for the Rouwenhorst method and either the even-spaced grid, Gauss-Hermite quadrature grid, or the quantile grid for our method. A similar type of discretization is considered in Caldara, Fernández-Villaverde, Rubio-Ramírez, and Yao 2012, although they use Tauchen's method to discretize both the log variance and the level of the process.

Following Caldara, Fernández-Villaverde, Rubio-Ramírez, and Yao 2012, we set the parameter values to $\lambda = 0.95$, $\rho = 0.9$, $\sigma = 0.06$, and choose $\mu = -9.9426$ to make the conditional standard deviation of the y process equal to 0.007. As a robustness

check, we also vary λ, the persistence of technology shocks, between 0 and 0.99. We focus on characteristics of the time series of y_t (the OLS coefficient λ and the unconditional variance σ_y^2), because the component approximations of x_t are just the standard autoregressive processes we studied before. For each discretization procedure, we vary N (the number of log variance and technology points) between 9, 15, and 21. Table 2.B.3 shows the results.

Table 2.B.3. $\log_{10}$ Relative Bias, Stochastic Volatility Model

N	λ	TR		ME-Even		ME-Quant		ME-Quad	
Parameter		$1-\lambda$	σ_y^2	$1-\lambda$	σ_y^2	$1-\lambda$	σ_y^2	$1-\lambda$	σ_y^2
9	0	$-\infty$	-9.781	$-\infty$	-6.101	$-\infty$	-5.034	$-\infty$	-5.282
	0.5	-1.819	-9.352	-9.556	-6.102	-9.997	-5.034	-8.755	-5.281
	0.9	-0.982	-8.265	-9.458	-6.102	-9.790	-5.034	-8.857	-5.281
	0.95	-0.718	-9.666	-9.117	-6.102	-9.153	-5.034	-9.409	-5.281
	0.99	-1.381	-8.034	-8.390	-6.102	-8.091	-5.034	-8.455	-5.281
15	0	$-\infty$	-11.15	$-\infty$	-7.371	-14.33	-5.203	-14.70	-6.060
	0.5	-2.189	-8.943	-9.079	-7.367	-9.647	-5.203	-9.630	-6.060
	0.9	-1.337	-8.502	-9.376	-7.364	-9.845	-5.203	-9.269	-6.060
	0.95	-1.061	-8.334	-9.902	-7.363	-9.245	-5.203	-9.158	-6.060
	0.99	-0.540	-8.112	-8.652	-7.399	-7.777	-5.204	-8.059	-6.067
21	0	$-\infty$	-9.336	-14.78	-8.625	-15.96	-5.317	-15.66	-6.898
	0.5	-2.436	-9.821	-10.09	-8.668	-9.813	-5.317	-10.46	-6.900
	0.9	-1.575	-8.693	-9.663	-8.700	-9.556	-5.317	-9.725	-6.900
	0.95	-1.296	-9.755	-10.44	-8.645	-9.993	-5.317	-10.24	-6.899
	0.99	-0.705	-8.193	-9.537	-8.750	-7.823	-5.319	-8.974	-6.909

Since the state space of the volatility process is continuous, Theorem 2.A.1 does not apply, so the unconditional moments need not be exact. However, Table 2.B.3 shows that our method is highly accurate, with a relative bias on the order of 10^{-8} or less for $1-\lambda$ and 10^{-5} or less for σ_y^2. This is likely because the finite-state Markov chain approximation of the volatility process is so accurate that Theorem 2.A.1 "almost" applies. As expected, the Tauchen-Rouwenhorst (TR) method does extremely well for

the unconditional variance because it is designed to match by construction. However, it does very poorly compared to the ME methods for the persistence, and this gap widens as λ gets closer to 1.

2.C Solving Asset Pricing Models

2.C.1 Analytical Solution with AR(1)/VAR(1) shocks

Burnside 1998 iterates (2.13) forward and obtains a closed-form solution as follows. In order to be consistent with the notation in Section 2.3, let

$$x_t = (I-B)\mu + Bx_{t-1} + \eta_t,$$

where μ is the unconditional mean of $\{x_t\}$, and $\eta_t \sim N(0,\Psi)$. Let

$$\begin{aligned}
\tilde{\Psi} &= (I-B)^{-1}\Psi(I-B')^{-1}, \\
\Psi_n &= \sum_{k=1}^{n} B^k \tilde{\Psi}(B')^k, \\
C_n &= B(I-B^n)(I-B)^{-1}, \\
\Omega_n &= n\tilde{\Psi} - C_n\tilde{\Psi} - \tilde{\Psi}C_n' + \Psi_n.
\end{aligned}$$

Then we have

$$V(x) = \sum_{n=1}^{\infty} \beta^n \exp\left(n\alpha'\mu + \alpha' C_n(x-\mu) + \frac{1}{2}\alpha'\Omega_n\alpha\right). \tag{.22}$$

A similar formula can be derived even if the shock distribution is non-Gaussian. For example, for the AR(1) case (so $C_t = D_t$), Tsionas 2003 shows that the price-dividend

ratio is

$$V(x) = \sum_{n=1}^{\infty} \beta^n \exp(a_n + b_n(x-\mu)), \tag{.23}$$

where

$$\begin{aligned} b_n &= (1-\gamma)\rho\frac{1-\rho^n}{1-\rho}, \\ a_n &= (1-\gamma)\mu n + \sum_{k=1}^{n} \log M\left((1-\gamma)\frac{1-\rho^k}{1-\rho}\right), \end{aligned}$$

and $M(\cdot)$ is the moment generating function of ε_t.

In general, the infinite series (.22) or (.23) have to be approximated. Burnside 1999 notes that truncating the series (.22) may not be accurate when α is close to zero since each term would have order β^n, so for β close to 1 the truncation error is substantial. A better way is to use the exact terms up to some large number N, and for $n > N$ we can replace C_n, Ψ_n by their limits $C_\infty = B(I-B)^{-1}$, $\Psi_\infty = \sum_{k=1}^{\infty} B^k \tilde{\Psi}(B')^k$, and Ω_n by

$$n\tilde{\Psi} - C_\infty\tilde{\Psi} - \tilde{\Psi}C_\infty' + \Psi_\infty,$$

in which case the infinite sum can be calculated explicitly. The result is

$$\begin{aligned} V(x) \approx \sum_{n=1}^{N} \beta^n \exp\left(n\alpha'\mu + \alpha' C_n(x-\mu) + \frac{1}{2}\alpha'\Omega_n\alpha\right) \\ + \frac{r^{N+1}}{1-r}\exp\left(\alpha' C_\infty(x-\mu) + \frac{1}{2}\alpha'(\Psi_\infty - C_\infty\tilde{\Psi} - \tilde{\Psi}C_\infty')\alpha\right), \end{aligned} \tag{.24}$$

where $r = \beta\exp\left(\alpha'\mu + \frac{1}{2}\alpha'\tilde{\Psi}\alpha\right) < 1$. If $r \geq 1$, the price-dividend ratio is infinite. Proposition 5 shows that the approximation error of (.24) is $O((r\rho)^N)$, where ρ is the absolute value of the largest eigenvalue of B. On the other hand, if we simply truncate the series (.22) at N, the error would be $O(r^N)$, which is much larger.

Proposition 5. *Consider the asset pricing formula* (.23)*. Let $V_N(x)$ be the value of $V(x)$, where ρ^n is replaced by 0 for $n > N$. Let a_n, b_n be as in* (.23)*, $m_n = \log M((1-\gamma)(1-\rho^n)/(1-\rho))$, $S_n = \sum_{k=1}^{n} m_n$, $b = \lim b_n = \frac{1-\gamma}{1-\rho}\rho$, $m = \lim m_n = \log M(\frac{1-\gamma}{1-\rho})$, and assume $r = \beta((1-\gamma)\mu + m) < 1$. Then*

$$V_N(x) = \sum_{n=1}^{N} \beta^n \exp(a_n + b_n(x-\mu)) + \frac{r^{N+1}}{1-r} \exp(S_N - mN + b(x-\mu)).$$

Furthermore, the approximation error $|V(x) - V_N(x)|$ is of order $(r\rho)^N$.

Proof. Let a'_n be the value of a_n, where ρ^k is set to 0 for $k > N$. Since $a'_n = (1-\gamma)\mu n + S_N + m(n-N)$, we get

$$\begin{aligned} V_N(x) - \sum_{n=1}^{N} \beta^n \exp(a_n + b_n(x-\mu)) &= \sum_{n=N+1}^{\infty} \beta^n \exp(a'_n + b(x-\mu)) \\ &= \sum_{n=N+1}^{\infty} \beta^n \exp((1-\gamma)\mu n + S_N + m(n-N) + b(x-\mu)) \\ &= \sum_{n=N+1}^{\infty} r^n \exp(S_N - mN + b(x-\mu)) = \frac{r^{N+1}}{1-r} \exp(S_N - mN + b(x-\mu)). \end{aligned}$$

If we replace ρ^n by 0 for $n > N$, since $\log M(\cdot)$ is differentiable and the domain of M for the asset pricing formula is bounded (hence $\log M$ is Lipschitz continuous), $|m_n - m|$ and $|b_n - b|$ are both of the order ρ^n. Since a_n contains the sum of m_n's, we have $|a_n - a'_n| \approx \sum_{k=N+1}^{n} \rho^k = O(\rho^N)$. Since $|\rho| < 1$, letting $c_n = a_n + b_n(x-\mu)$ and $c'_n = a'_n + b(x-\mu)$, we have $|c_n - c'_n| < 1$ eventually, so by the mean value theorem $|\exp(c_n - c'_n) - 1| \leq$

e $|c_n - c'_n| = O(\rho^N)$. Therefore

$$\begin{aligned}|V(x) - V_N(x)| &\leq \sum_{n=N+1}^{\infty} \beta^n \left|\exp(a_n + b_n(x-\mu)) - \exp(a'_n + b(x-\mu))\right| \\ &= \sum_{n=N+1}^{\infty} \beta^n \exp(a'_n + b(x-\mu)) \left|\exp(c_n - c'_n) - 1\right| \\ &\approx \sum_{n=N+1}^{\infty} r^n \rho^N = O((r\rho)^N). \qquad \square\end{aligned}$$

2.C.2 Discretizing the Rare Disasters Model

In this appendix we provide the details of the discretization of the resilience process (2.20). The discussion is partly based on footnote 9 in Gabaix 2012 and his online appendix. First, in order for (2.20) to be stable, we need

$$\frac{1+H_*}{1+H_t}\mathrm{e}^{-\phi_H} \leq 1 \iff \widehat{H}_t \geq (1+H_*)(\mathrm{e}^{-\phi_H} - 1). \tag{.25}$$

Since in Gabaix 2012 $p_t = p$ and $B_{t+1} = B$ are constant, and by definition $0 \leq F_{t+1} \leq 1$, from (2.19) we obtain

$$-p \leq H_* + \widehat{H}_t \leq p(B^{-\gamma} - 1). \tag{.26}$$

We can take $H_* = p(B^{1-\gamma} - 1)$ because Gabaix assumes that the average dividend recovery rate is the same as consumption. The inequalities (.25) and (.26) define bounds for $\widehat{H}_t$, which we denote by $[\widehat{H}_{\min}, \widehat{H}_{\max}]$. In order for the process to remain within this bound, Gabaix assumes that the conditional variance of ε^H_{t+1} shrinks to 0 as we approach the boundary. Namely, he assumes

$$\sigma^2(\widehat{H}) = 2K(1 - \widehat{H}/\widehat{H}_{\min})^2(1 - \widehat{H}/\widehat{H}_{\max})^2,$$

where $K = 0.2\phi_H \left|\widehat{H}_{\min}\widehat{H}_{\max}\right|$. See Eq. (59) in the online appendix of Gabaix 2012. We use the exact same functional form.

We define the grid of discretization to be $[\widehat{H}_{\min}+\varepsilon, \widehat{H}_{\max}-\varepsilon]$, where $\varepsilon > 0$ is a small number which we set to be $\varepsilon = 10^{-3} \times (\widehat{H}_{\max} - \widehat{H}_{\min})$. The reason for shrinking the interval slightly is because otherwise the conditional variance becomes exactly zero at the boundary points, which is impossible for a discrete Markov chain. Once we have defined the end points of the grid this way, we put grid points and discretize the beta distribution at each point by matching the conditional moments using our method. We consider the even-spaced grid (trapezoidal formula), Clenshaw-Curtis quadrature (Clenshaw and Curtis 1960; Trefethen 2008), and Gauss-Legendre quadrature, which are the most natural choices since the integration is over a bounded interval.

2.C.3 Solving the Rare Disasters Model

In this appendix we explain how to numerically solve the variable rare disaster model using discretization. We follow the notation in Gabaix 2012.

The stochastic discount factor between time t and $t+1$ is

$$M_{t+1} = \mathrm{e}^{-\rho}(C_{t+1}/C_t)^{-\gamma} = \mathrm{e}^{-\delta} \times \begin{cases} 1, & \text{(no disaster)} \\ B_{t+1}^{-\gamma}, & \text{(disaster)} \end{cases}$$

where $\delta = \rho + \gamma g_C$. Letting P_t be the cum-dividend price of the stock and $V_t = P_t/D_t$ be the price-dividend ratio, it follows from the Euler equation that

$$\begin{aligned} & P_t = D_t + \mathbb{E}_t[M_{t+1}P_{t+1}] \\ \Longrightarrow\ & V_t = 1 + \mathbb{E}_t\left[M_{t+1}\frac{D_{t+1}}{D_t}V_{t+1}\right] \\ & \quad = 1 + \mathrm{e}^{-\delta+g_D}\left((1-p_t)\mathbb{E}_t^{\mathrm{ND}}[V_{t+1}] + p_t\mathbb{E}_t^{\mathrm{D}}[B_{t+1}^{-\gamma}F_{t+1}V_{t+1}]\right), \end{aligned}$$

where p_t is the disaster probability and $\mathbb{E}_t^{\mathrm{ND}}, \mathbb{E}_t^{\mathrm{D}}$ denote the expectation conditional on no disaster or disaster. By the structure of the model, V_{t+1} depends only on the resilience (2.19), which evolves independently from disasters. Therefore $\mathbb{E}_t^{\mathrm{ND}}[V_{t+1}] = \mathbb{E}_t^{\mathrm{D}}[V_{t+1}] = \mathbb{E}_t[V_{t+1}]$. Using the definition of resilience, it follows that

$$V_t = 1 + \mathrm{e}^{-\delta + g_D}(1 + H_t)\mathbb{E}_t[V_{t+1}].$$

To solve for the price-dividend ration using discretization, suppose the state space of resilience H_t is discretized, and let $s = 1, \ldots, S$ be the states. Since the disaster probability is constant, it follows that

$$v_s = 1 + \mathrm{e}^{-\delta + g_D}(1 + h_s) \sum_{s'=1}^{S} \pi_{ss'} v_{s'},$$

where v_s is the price-dividend ratio in state s, h_s is the resilience in state s, and $\pi_{ss'}$ is the transition probability from state s to s'. Letting $v = (v_1, \ldots, v_S)$ and $h = (h_1, \ldots, h_S)$ be the vectors of those values, and $P = (\pi_{ss'})$ be the transition probability matrix, it follows that

$$v = 1 + \mathrm{e}^{-\delta + g_D} \operatorname{diag}(1 + h) P v \iff v = (I - \mathrm{e}^{-\delta + g_D} \operatorname{diag}(1 + h) P)^{-1} 1.$$

The continuous solution is obtained by interpolating these values over the entire grid (see Proposition 4).

2.D Asset Pricing with Gaussian AR(1) Shocks

In this appendix we solve the simple asset pricing model with Gaussian AR(1) shocks

$$x_t = (1 - \rho)\mu + \rho x_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2),$$

where x_t is log dividend growth. Using postwar data, the OLS estimates are $\mu = 0.0559$, $\rho = 0.405$, and $\sigma = 0.0589$. Preference parameters are risk aversion $\gamma = 2$ and discount factor $\beta = 0.95$. In order to avoid cherry-picking, we consider all major existing methods, Tauchen 1986,[3] Tauchen and Hussey 1991, and Rouwenhorst 1995. For the ME methods, we consider ME-Even, ME-Quant, ME-Quad (all with two moments) as well as ME-Even with 4 moments.[4] We consider two robustness checks, (i) changing the number of grid points N, and (ii) changing the persistence of dividend growth ρ.[5] The number of grid points is always $N = 9$ unless otherwise stated.

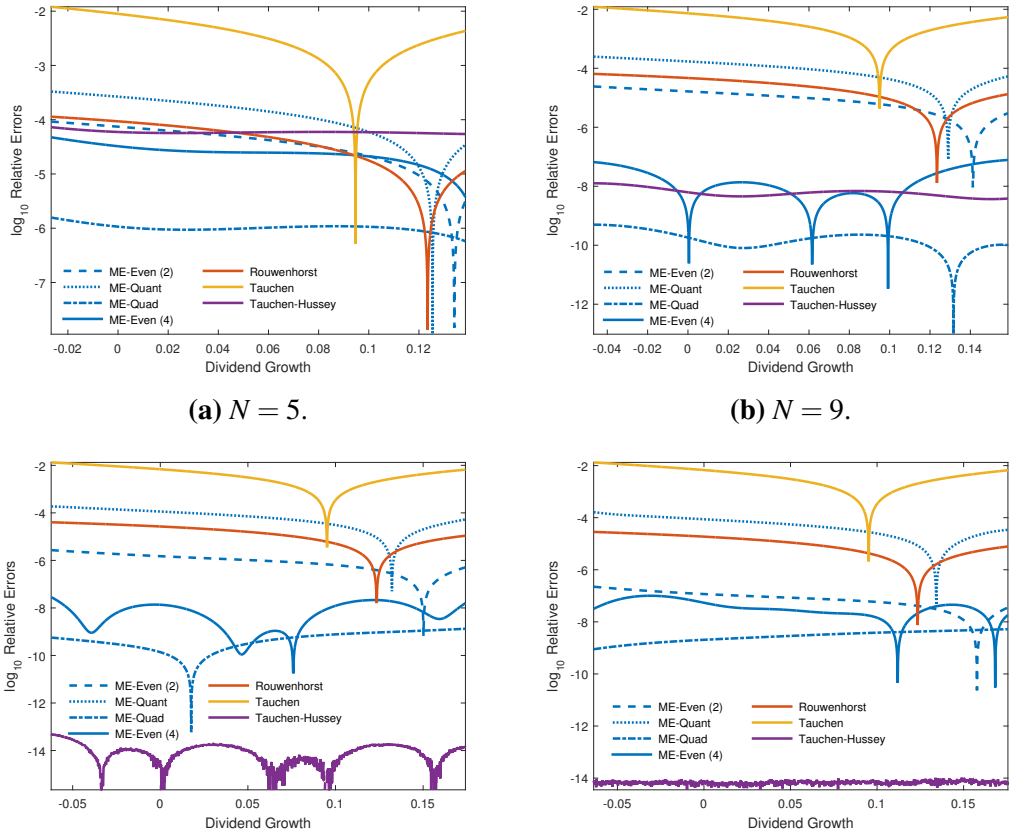
Figure 2.D.1 shows the $\log_{10}$ relative errors of the price-dividend ratio with various discretization methods and number of points N. We can make a few observations. First, as we increase N, all methods become more accurate, as expected. This is especially true for Tauchen-Hussey, whose performance is sensitive to N. Second, for methods other than Tauchen-Hussey, the order of the performance is generally ME-Quad $>$ ME-Even (4) $>$ ME-Even (2) $>$ Rouwenhorst $>$ ME-Quant $>$ Tauchen. ME-Quad and ME-Even (4 moments) give a solution accuracy of order 10^{-4} to 10^{-9}. Third, the performance of ME-Quad does not improve beyond $N = 9$. This is because since ME methods involve a numerical optimization, in which we set the error tolerance to 10^{-10}, the theoretical lower bound for the $\log_{10}$ errors is about -10.

Figure 2.D.2 shows the $\log_{10}$ relative errors when we increase the persistence ρ, fixing the number of points at $N = 9$. Not surprisingly, the performance worsens for all

[3]For the Tauchen method, we need to specify the grid spacing. To give it the best chance, following Kopecky and Suen 2010 we choose the grid spacing in order to match the unconditional variance exactly. We also experimented with $\sqrt{N-1}$ standard deviations (as in ME-Even and Rouwenhorst) or $1.2 \log N$ (as in Flodén 2008) but the performance was worse.

[4]As discussed below, ME-Quant is uniformly dominated by other ME methods, so there is no point in considering ME-Quant with 4 moments. The results for ME-Quad with 4 moments are similar to 2 moments. We also considered matching 6 moments, but the performance is similar to 4 moments.

[5]Collard and Juillard 2001 perform robustness checks across other parameters such as the discount factor, risk aversion, and volatility. They find that the solution accuracy is most susceptible to turning up the persistence.



(a) $N = 5$. **(b)** $N = 9$.

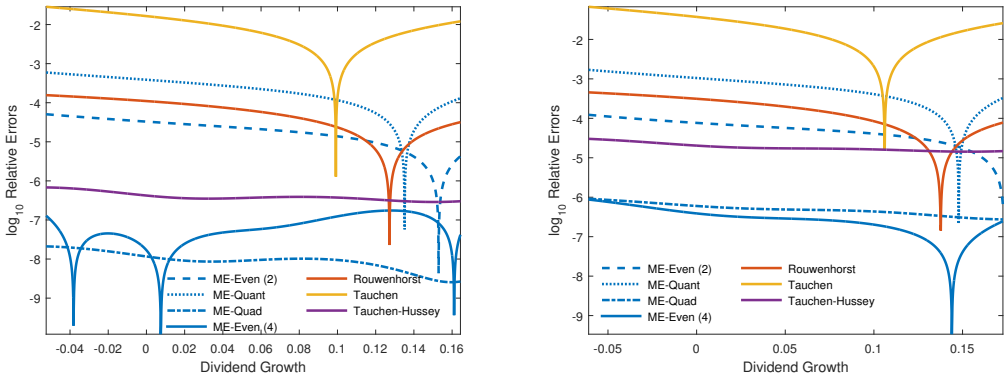
(c) $N = 15$. **(d)** $N = 21$.

Figure 2.D.1. $\log_{10}$ Relative Errors of Price-Dividend Ratio Approximations fo Different Numbers of Points, Gaussian AR(1) Model

$\log_{10}$ relative errors of price-dividend ratio with various discretization methods and number of points for the Gaussian AR(1) model. ME-Even (L) shows the result with L moments.

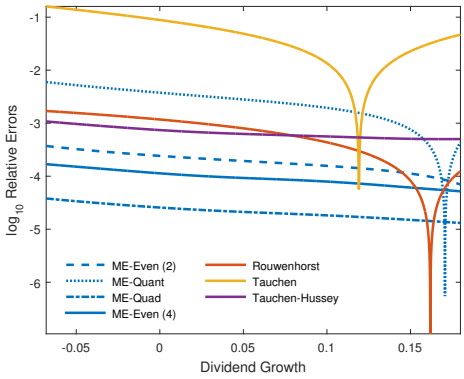
methods as we make the dividend process more persistent. However, the performance of the Tauchen-Hussey method deteriorates quickly, as is well-known. ME-Quad, which uses the same Gauss-Hermite quadrature grid as Tauchen-Hussey, also gets poorer, but it is still the best performer along with ME-Even (4 moments). The performance of the Rouwenhorst method is robust, although it is uniformly dominated by ME-Even (2 or 4 moments) and ME-Quad.

It is well-known that existing methods except Rouwenhorst are poor when the

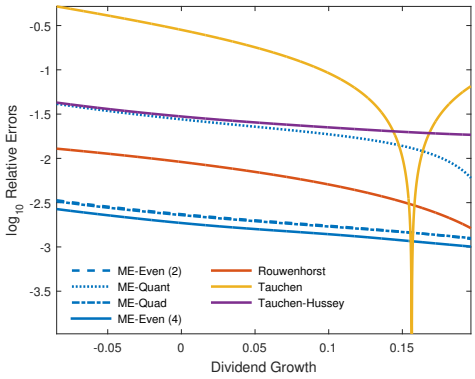


(a) $\rho = 0.5$.

(b) $\rho = 0.6$.



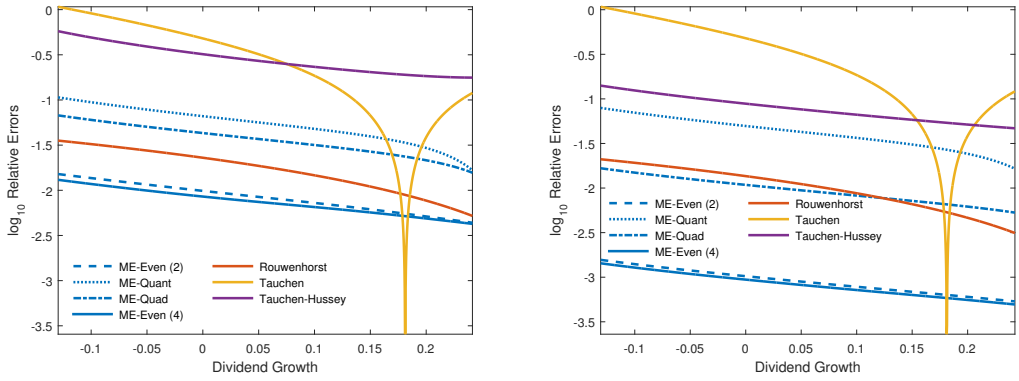
(c) $\rho = 0.7$.



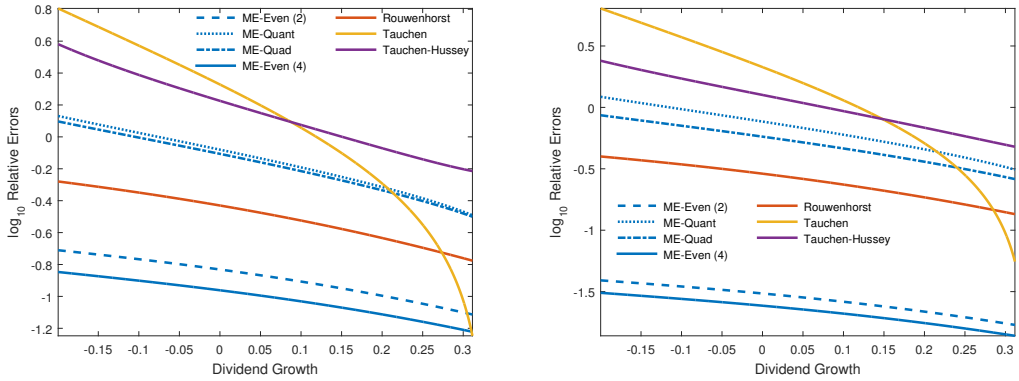
(d) $\rho = 0.8$.

Figure 2.D.2. $\log_{10}$ Relative Errors of Price-Dividend Ratio Approximations for Different Levels of Persistence, Gaussian AR(1) Model

process is persistent (Flodén 2008; Kopecky and Suen 2010). However, since the price-dividend ratio is infinite (*i.e.*, the series (.22) diverges) beyond $\rho = 0.8$ with the baseline specification $\gamma = 2$ and $\beta = 0.95$, the performance of the ME methods when persistence is high is still unanswered. In order to see what happens when the AR(1) process is very persistent, we set $(\rho, \gamma) = (0.9, 1.5), (0.95, 1.3)$, for which the price-dividend ratio is finite. Figure 2.D.3 shows the results. With $\rho = 0.9$, Tauchen-Hussey is one of the worst performers. ME-Quad also deteriorates, and is slightly worse (better) than Rouwenhorst with $N = 9$ ($N = 15$) grid points. The best performers are ME-Even, with comparable performance with 2 or 4 moments.



(a) $\rho = 0.9$, $\gamma = 1.5$, $N = 9$. **(b)** $\rho = 0.9$, $\gamma = 1.5$, $N = 15$.



(c) $\rho = 0.95$, $\gamma = 1.3$, $N = 15$. **(d)** $\rho = 0.95$, $\gamma = 1.3$, $N = 21$.

Figure 2.D.3. $\log_{10}$ Relative Errors of Price-Dividend Ratio Approximations for a Highly Persistent Process, Gaussian AR(1) Model

$\log_{10}$ relative errors of price-dividend ratio with various discretization methods for the highly persistent Gaussian AR(1) model with $(\rho, \gamma) = (0.9, 1.5), (0.95, 1.3)$.

To get a better idea of the solution accuracy, consider an investor purchasing \$1 Million worth of the asset. If the investor uses each discretization method to compute the fair price of the asset, what is the mistake in dollar amounts? Table 2.D.1 shows the mispricing using the average $\log_{10}$ relative errors. With the baseline specification ($N = 9$, $\rho = 0.405$), the mispricing for \$1M investment is only 1 cent with ME-Even (4 moments). With ME-Quad and Tauchen-Hussey, the pricing error is virtually zero. Even with the Rouwenhorst method, the mispricing is only \$18, so it does not make a

material difference across methods except the Tauchen method, which is off by more than \$3,000. However, the choice of the discretization method matters as we increase the persistence of the dividend process. With $\rho = 0.8$, the Tauchen method is off by 12%, Tauchen-Hussey by 2.6%, Rouwenhorst by 0.6%, as opposed to 0.16% with ME-Even (4 moments). The result is even more stark with $\rho = 0.9, 0.95$.

Table 2.D.1. Mispricing in dollars when investing \$1 Million.

		ME methods				Existing methods		
N	ρ	Even (2)	Quant	Quad	Even (4)	R	Tauchen	TH
Changing number of grid points ($\gamma = 2$)								
5	0.405	31.6	103	10.1	23.3	33.8	3,161	58.9
9		7.27	71.1	0	0.011	18.1	3,136	0.006
15		0.767	51.7	0	0.005	11.2	3,380	0
21		0.065	39.8	0	0.03	7.89	3,363	0
Changing persistence ($\gamma = 2$)								
9	0.5	16.1	172	0.009	0.051	43.6	7.2K	0.393
	0.6	46.6	507	0.491	0.235	127	17K	18.3
	0.7	185	2.1K	21.4	92.3	501	39K	652
	0.8	2.0K	21K	2.0K	1.6K	6.1K	120K	26K
Highly persistent case ($\gamma = 1.5$)								
9	0.9	8.3K	53K	36K	7.4K	17K	218K	280K
15		0.89K	41K	9.3K	0.82K	9.9K	218K	77K
Highly persistent case ($\gamma = 1.3$)								
15	0.95	13K	70K	67K	9.8K	32K	1.3M	1.4M
21		2.7K	65K	50K	2.2K	25K	1.3M	1.1M

Note: Even (L): ME-Even method with L moments; R: Rouwenhorst 1995 method; TH: Tauchen and Hussey 1991 method. K, M denote thousands and millions of dollars.

In summary, we find that for discretizing a Gaussian AR(1) process, (i) Tauchen-Hussey is best if there are many points ($N \geq 15$) and the process is not so persistent ($\rho \leq 0.4$), (ii) ME-Quad is best if the process is moderately persistent ($0.4 \leq \rho \leq 0.8$), with ME-Even (4 moments) comparable, (iii) ME-Even and Rouwenhorst perform well over all choices of grid points N and persistence ρ (especially $\rho > 0.8$), with solution

accuracy ME-Even (4) > ME-Even (2) > Rouwenhorst, and (iv) ME-Quant is poor.

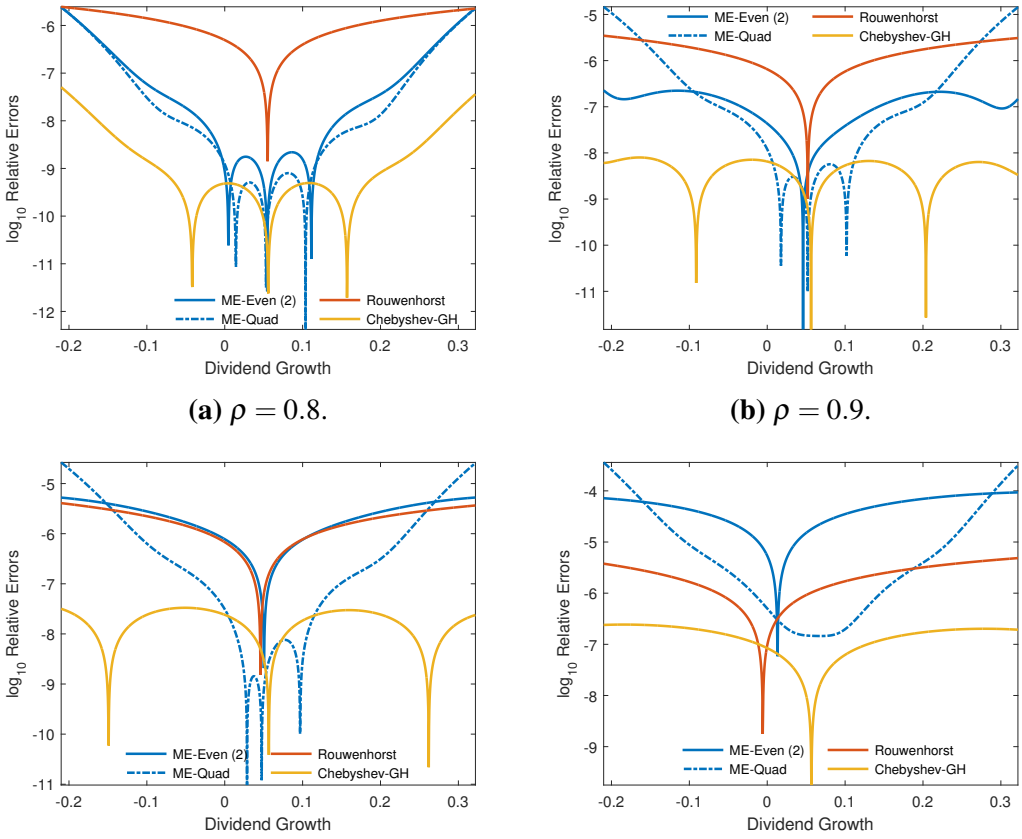
Finally, one may be interested in how the discretization solution fares against conventional methods such as projection, and how the performance of discretization deteriorates as the persistence increases. To address this issue, we fix the preference parameters at $\beta = 0.2$ and $\gamma = 1.3$, number of points $N = 9$, and consider the autocorrelation $\rho = 0.8, 0.9, 0.95, 0.99$. (It is necessary to reduce the discount factor β to an unrealistically small number so that the analytical solution exists even for high persistence.) For this exercise, we only consider ME-Even (2), ME-Quad, Rouwenhorst, and the projection method. For the projection method, we make the Euler equation errors zero at the Chebyshev collocation points, and the conditional expectation is computed using a highly accurate Gauss-Hermite quadrature (see Pohl, Schmedders, and Wilms 2014 for details). Figure 2.D.4 shows the results.

Unsurprisingly, the projection method is extremely accurate, since a highly accurate Gauss-Hermite quadrature nodes are chosen for each Chebyshev collocation point. The performance of discretization methods deteriorates as we increase the persistence. The maximum entropy methods are more accurate for persistence less than 0.95, but beyond that the Rouwenhorst method becomes more accurate. This is probably because the Rouwenhorst method does not involve any numerical optimization.

3.A Proof of Proposition 1 (Chapter 3)

Proof. (Part i)

We guess and verify that the price-dividend ratio is $pd_t = A_{0,m} + A_m' z_t$. By Assumption 1, $\Delta d_t = S_d' z_t$. Using $r_{t+1} \approx k + \rho(p_{t+1} - d_{t+1}) + \Delta d_{t+1} + d_t - p_t$ and plugging



(a) $\rho = 0.8$. **(b)** $\rho = 0.9$.

(c) $\rho = 0.95$. **(d)** $\rho = 0.99$.

Figure 2.D.4. $\log_{10}$ Relative Errors of Price-Dividend Ratio Approximations for a Highly Persistent Process, Alternate Parameterization, Gaussian AR(1) Model

$\log_{10}$ relative errors of price-dividend ratio with discretization and projection methods for the highly persistent Gaussian AR(1) model with $\beta = 0.2$, $\gamma = 1.3$, and $N = 9$. "Chebyshev-GH" refers to the projection method with Chebyshev collocation and Gauss-Hermite quadrature.

the log-linearized return into the Euler equation, we have

$$\begin{aligned} 1 &= \exp[\lambda_0 + \kappa + (\rho - 1)A_{0,m} - A'_m z_t] \times E_t \left[\exp\left\{[-\Lambda' + S'_d + \rho A'_m] z_{t+1}\right\}\right] \\ 0 &= \lambda_0 + \kappa + (\rho - 1)A_{0,m} - A'_m z_t + [-\Lambda' + S'_d + \rho A'_m](\mu + F z_t) \\ &+ f(-\Lambda' + S'_d + \rho A'_m) + g(-\Lambda' + S'_d + \rho A'_m)' z_t, \end{aligned}$$

where the second line takes logs and applies assumption 1.ii. Rearranging yields the $(L+1)$-dimensional system of equations in $A_{0,m}$ and A_m

$$\begin{aligned} f(-\Lambda + S_d + \rho A_m) + \lambda_0 + \kappa + (\rho - 1)A_{0,m} + (-\Lambda' + S_d' + \rho A_m')\mu &= 0 \qquad (.27) \\ g(-\Lambda + S_d + \rho A_m) - (I - \rho F')A_m + F'(-\Lambda + S_d) &= 0. \qquad (.28) \end{aligned}$$

This system does not have an analytical solution in the general case; however, it is relatively straightforward to solve the system numerically.

(Part ii)

The coefficients on the risk-free rate can be calculated by direct computation

$$\begin{aligned} rf_{t+1} &= -\log E_t[\exp(\lambda_0 - \Lambda' z_{t+1})] = -\log E_t[\exp(\lambda_0 + \Lambda'(\mu + F z_t + \varepsilon_{t+1}))] \\ &= -[\lambda_0 - \Lambda'\mu + f(-\Lambda)] - (\Lambda' F + g(-\Lambda)y_t \equiv A_{0,f} + A_f' z_t. \end{aligned}$$

(Part iii)

To calculate the expected excess market return, we calculate its conditional cumulant-generating function for an arbitrary scalar γ

$$\log E_t[\exp(\gamma r_{t+1})] \equiv \gamma[\kappa + (S_d' + \rho A_m')z_t] + f(\gamma[S_d + \rho A_m]) + g(\gamma[S_d + \rho A_m])' z_t \quad (.29)$$

The expected log market return ($E_t[r_{t+1}]$) is the first derivative of (.29) with respect to γ, evaluated at $\gamma = 0$. Since (.29) is affine in z_t, its derivative will also be affine in γ, which establishes the claim. □

3.B Details of Nonparametric Estimation

Robinson (1989) and Cai (2007) consider local constant and local linear approximations of β respectively, but this approach can easily be generalized to accommodate polynomials of arbitrary order. In particular, we approximate the function β_t as a p^{th}-order Taylor expansion about the point $\frac{t}{T}$ (where $p \geq 0$). To this end, define the quantities:

$$\mathbf{W}_{st} = \left(1, \frac{s-t}{T}, \ldots, \left(\frac{s-t}{T}\right)^p\right)', \tag{.30}$$

$$K_{st} = K\left(\frac{s-t}{hT}\right), \tag{.31}$$

$$\mathbf{Q}_{st} = \mathbf{W}_{st} \otimes x_s, \tag{.32}$$

for $s,t = 1,\ldots,T$, where K is a kernel function and $h \equiv h(T)$ is the bandwidth. More formally, $K : [-1,1] \to \mathbb{R}^+$ is a function that is symmetric about 0 and integrates to 1, and $h \in [0,1]$ satisfies $h \to 0$ and $hT \to \infty$ as $T \to \infty$.

The local polynomial estimator $\alpha = \left(\beta_0', \beta_1', \ldots, \beta_p'\right)'$ is obtained by solving

$$\begin{aligned} &\min_{\alpha \in \mathbb{R}^{pd}} \sum_{s=t-\lfloor hT \rfloor}^{t+\lfloor hT \rfloor} K_{st} \left[r_{s+1} - \beta_0' x_s - \beta_1' \left(\frac{s-t}{T}\right) x_s - \ldots - \beta_p' \left(\frac{s-t}{T}\right)^p x_s \right]^2 \\ = &\sum_{s=t-\lfloor hT \rfloor}^{t+\lfloor hT \rfloor} K_{st} \left(r_{s+1} - \alpha' \mathbf{Q}_{st}\right)^2. \end{aligned} \tag{.33}$$

Solving this optimization problem for α gives the solution

$$\hat{\alpha}_t = \left(\sum_{s=t-\lfloor Th \rfloor}^{t+\lfloor Th \rfloor} K_{st} \mathbf{Q}_{st} \mathbf{Q}_{st}' \right)^{-1} \sum_{s=t-\lfloor Th \rfloor}^{t+\lfloor Th \rfloor} K_{st} \mathbf{Q}_{st} r_{s+1}, \tag{.34}$$

where our object of interest, β_t, is the first element of α_t. That is, the estimator of β_t is

given by

$$\hat{\beta}_t = \left(\mathbf{e}_1' \otimes \mathbf{I}_d\right) \hat{\alpha}_t, \tag{.35}$$

where $\mathbf{e}_1$ is the first standard basis vector of $\mathbb{R}^{p+1}$, $\mathbf{I}_d$ is a $(d \times d)$ identity matrix, and d is the dimension of x_t. This can also be thought of as the OLS estimator of α_0 in the transformed model

$$K_{st}^{1/2} y_{s+1} = K_{st}^{1/2} x_s' \sum_{q=0}^{p} \alpha_q + \varepsilon_{s+1}. \tag{.36}$$

The asymptotic properties of these estimators are studied in Robinson (1989) and Cai (2007). Under various regularity conditions, it can be shown that the estimator $\hat{\alpha}_t$ in (.35) is consistent and asymptotically normal.

The main empirical results adopt a local constant (Nadarya-Watson) estimation procedure and so set $p = 0$. The motivation behind this choice is that the nonparametric procedures require very large amounts of data to perform well in finite samples and every additional degree of approximation requires that we estimate dT additional parameters. However, we also repeated the analysis using local linear models ($p = 1$) and found very similar results.

Bibliography

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

Aguirregabiria, Victor, and Pedro Mira. 2007. "Sequential Estimation of Dynamic Discrete Games". *Econometrica* 75 (1): 1–53.

Aït-Sahalia, Yacine, and Michael W Brandt. 2001. "Variable Selection for Portfolio Choice". *The Journal of Finance* 56 (4): 1297–1351.

Aiyagari, S. Rao. 1995. "Optimal Capital Income Taxation with Incomplete Markets, Borrowing Constraints, and Constant Discounting". *Journal of Political Economy* 103 (6): 1158–1175.

— . 1994. "Uninsured Idiosyncratic Risk and Aggregate Saving". *Quarterly Journal of Economics* 109 (3): 659–684.

Aruoba, S Borağan, Francis X Diebold, Jeremy Nalewaik, Frank Schorfheide, and Dongho Song. 2016. "Improving GDP Measurement: A Measurement-Error Perspective". *Journal of Econometrics* 191 (2): 384–397.

Aruoba, S. Borağan, Jesús Fernández-Villaverde, and Juan F. Rubio-Ramírez. 2006. "Comparing Solution Methods for Dynamic Equilibrium Economies". *Journal of Economic Dynamics and Control* 30 (12): 2477–2508.

Avramov, Doron, and Russ Wermers. 2006. "Investing in mutual funds when returns are predictable". *Journal of Financial Economics* 81 (2): 339–377.

Baker, Malcolm, and Jeffrey Wurgler. 2006. "Investor sentiment and the cross-section of stock returns". *The Journal of Finance* 61 (4): 1645–1680.

— . 2007. "Investor sentiment in the stock market". *The Journal of Economic Perspectives* 21 (2): 129–151.

Bakry, Dominique, Xavier Milhaud, and Pierre Vandekerkhove. 1997. “Statistique de Chaînes de Markov Cachées à Espace d’États Fini. Le Cas Non Stationnaire”. *Comptes Rendus de l’Académie des Sciences - Series I - Mathematics* 325 (2): 203–206.

Banegas, Ayelen, Ben Gillen, Allan Timmermann, and Russ Wermers. 2013. “The cross section of conditional mutual fund performance in European stock markets”. *Journal of Financial economics* 108 (3): 699–726.

Bansal, Ravi, Dana Kiku, and Amir Yaron. 2012. “An Empirical Evaluation of the Long-Run Risks Model for Asset Prices”. *Critical Finance Review* 1 (1): 183–221. ISSN: 2164-5744. doi:10.1561/104.00000005. http://dx.doi.org/10.1561/104.00000005.

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

Barro, Robert J. 2006. “Rare Disasters and Asset Markets in the Twentieth Century”. *Quarterly Journal of Economics* 121 (3): 823–866.

Barro, Robert J, and José F Ursúa. 2008. “Macroeconomic Crises Since 1870”. *Brookings Papers on Economic Activity* 2008 (1): 255–350.

Bauer, Michael D., and Glenn D. Rudebusch. 2016. “Monetary Policy Expectations at the Zero Lower Bound”. *Journal of Money, Credit, and Banking* 48 (7): 1439–1465.

Baum, Leonard E, and Ted Petrie. 1966. “Statistical Inference for Probabilistic Functions of Finite State Markov Chains”. *Annals of Mathematical Statistics* 37 (6): 1554–1563.

Bickel, Peter J, and Ya’Acov Ritov. 1996. “Inference in Hidden Markov Models I: Local Asymptotic Normality in the Stationary Case”. *Bernoulli* 2 (3): 199–228.

Bickel, Peter J, Yaacov Ritov, and Tobias Ryden. 1998. “Asymptotic Normality of the Maximum-Likelihood Estimator for General Hidden Markov Models”. *Annals of Statistics* 26 (4): 1614–1635.

Bonomo, Marco, René Garcia, Nour Meddahi, and Roméo Tédongap. 2011. “Generalized Disappointment Aversion, Long-run Volatility Risk, and Asset Prices”. *Review of Financial Studies* 24 (1): 82–122.

Borwein, Jonathan M., and Adrian S. Lewis. 1991. “Duality Relationships for Entropy-like Minimization Problems”. *SIAM Journal on Control and Optimization* 29 (2): 325–338.

Bucy, Richard S. 1969. “Bayes Theorem and Digital Realizations for Non-Linear Filters”. *Journal of the Astronautical Sciences* 17:80–94.

Bucy, Richard S, and Kenneth D Senne. 1971. “Digital Synthesis of Non-Linear Filters”. *Automatica* 7 (3): 287–298.

Burnside, Craig. 1999. “Discrete State-Space Methods for the Study of Dynamic Economies”. Chap. 5 in *Computational Methods for the Study of Dynamic Economies*, ed. by Ramon Marimon and Andrew Scott, 95–113. Oxford: Oxford University Press.

— . 1998. “Solving Asset Pricing Models with Gaussian Shocks”. *Journal of Economic Dynamics and Control* 22 (3): 329–340.

Cai, Zongwu. 2007. “Trending Time-Varying Coefficient Time Series Models with Serially Correlated Errors”. *Journal of Econometrics* 136 (1): 163–188.

Caldara, Dario, Jesús Fernández-Villaverde, Juan F. Rubio-Ramírez, and Wen Yao. 2012. “Computing DSGE Models with Recursive Preferences and Stochastic Volatility”. *Review of Economic Dynamics* 15 (2): 188–206.

Campbell, John Y. 1987. “Stock Returns and the Term Structure”. *Journal of Financial Economics* 18 (2): 373–399.

Campbell, John Y, and John H Cochrane. 1999. “By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior”. *Journal of Political Economy* 107 (2): 205–251.

Campbell, John Y, and Robert J Shiller. 1988. “The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors”. *Review of financial studies* 1 (3): 195–228.

Campbell, John Y, and Samuel B Thompson. 2008. “Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?” *Review of Financial Studies* 21 (4): 1509–1531.

Campbell, John Y, and Luis M Viceira. 1999. “Consumption and Portfolio Decisions when Expected Returns are Time Varying”. *The Quarterly Journal of Economics* 114 (2): 433–495.

Cecchetti, Stephen G., Pok-sang Lam, and Nelson C. Mark. 1993. “The Equity Premium and the Risk-free Rate: Matching Moments”. *Journal of Monetary Economics* 31 (1): 21–45.

Chen, Bin, and Yongmiao Hong. 2012. "Testing for Smooth Structural Changes in Time Series Models via Nonparametric Regression". *Econometrica* 80 (3): 1157–1183.

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

Clark, Todd E, and Kenneth D West. 2007. "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models". *Journal of Econometrics* 138 (1): 291–311.

Clenshaw, C. W., and A. R. Curtis. 1960. "A Method for Numerical Integration on an Automatic Computer". *Numerische Mathematik* 2 (1): 195–207.

Collard, Fabrice, and Michel Juillard. 2001. "Accuracy of Stochastic Perturbation Methods: The Case of Asset Pricing Models". *Journal of Economic Dynamics and Control* 25 (6-7): 979–999.

Dangl, Thomas, and Michael Halling. 2012. "Predictive Regressions with Time-Varying Coefficients". *Journal of Financial Economics* 106 (1): 157–181.

David, Alexander, and Pietro Veronesi. 2013. "What Ties Return Volatilities to Price Valuations and Fundamentals?" *Journal of Political Economy* 121 (4): 682–746.

Dávila, Julio, Jay H. Hong, Per Krusell, and José-Víctor Ríos-Rull. 2012. "Constrained Efficiency in the Neoclassical Growth Model with Uninsurable Idiosyncratic Shocks". *Econometrica* 80 (6): 2431–2467.

Davis, Philip J., and Philip Rabinowitz. 1984. *Methods of Numerical Integration*. Second. Orlando, FL: Academic Press.

Diebold, Francis X, and Robert S Mariano. 1995. "Comparing Predictive Accuracy". *Journal of Business & Economic Statistics* 13 (3): 253–263.

Douc, Randal, Eric Moulines, Jimmy Olsson, and Ramon Van Handel. 2011. "Consistency of the Maximum Likelihood Estimator for General Hidden Markov Models". *Annals of Statistics* 39 (1): 474–513.

Douc, Randal, Eric Moulines, and Tobias Ryden. 2004. "Asymptotic Properties of the Maximum Likelihood Estimator in Autoregressive Models with Markov Regime". *Annals of Statistics* 32 (5): 2254–2304.

Elliott, Graham, and Ulrich K Müller. 2006. "Efficient Tests for General Persistent Time Variation in Regression Coefficients". *The Review of Economic Studies* 73 (4): 907–940.

Fama, Eugene F, and Kenneth R French. 1989. “Business Conditions and Expected Returns on Stocks and Bonds”. *Journal of Financial Economics* 25 (1): 23–49.

— . 1988. “Dividend Yields and Expected Stock Returns”. *Journal of Financial Economics* 22 (1): 3–25.

Farmer, Leland E. 2017. “The Discretization Filter: A Simple Way to Estimate Nonlinear State Space Models”. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2780166.

Farmer, Leland E, and Alexis Akira Toda. 2016. “Discretizing Nonlinear, Non-Gaussian Markov Processes with Exact Conditional Moments”. *Quantitative Economics, forthcoming*.

Ferson, Wayne E, and Rudi W Schadt. 1996. “Measuring Fund Strategy and Performance in Changing Economic Conditions”. *The Journal of Finance* 51 (2): 425–461.

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

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

Gabaix, Xavier. 2009. “Linearity-Generating Processes: A Modelling Tool Yielding Closed Forms for Asset Prices”. Working Paper, New York University.

— . 2012. “Variable Rare Disasters: An Exactly Solved Framework for Ten Puzzles in Macro-Finance”. *The Quarterly Journal of Economics* 127:645–700.

Galindev, Ragchaasuren, and Damba Lkhagvasuren. 2010. “Discretization of Highly Persistent Correlated AR(1) Shocks”. *Journal of Economic Dynamics and Control* 34 (7): 1260–1276.

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

Gourio, François. 2012. “Disaster Risk and Business Cycles”. *American Economic Review* 102 (6): 2734–2766.

Green, Jeremiah, John RM Hand, and Mark T Soliman. 2011. “Going, Going, Gone? The Apparent Demise of the Accruals Anomaly”. *Management Science* 57 (5): 797–816.

Gürkaynak, Refet S, Brian Sack, and Jonathan H Wright. 2007. "The U.S. Treasury Yield Curve: 1961 to the Present". *Journal of Monetary Economics* 54 (8): 2291–2304.

Guvenen, Fatih. 2009. "A Parsimonious Macroeconomic Model for Asset Pricing". *Econometrica* 77 (6): 1711–1750.

Guvenen, Fatih, Serdar Ozkan, and Jae Song. 2014. "The Nature of Countercyclical Income Risk". *Journal of Political Economy* 122 (3): 621–660.

Haliassos, Michael, and Alexander Michaelides. 2003. "Portfolio Choice and Liquidity Constraints". *International Economic Review* 44 (1): 143–177.

Hamilton, James D. 1989. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle". *Econometrica* 57 (2): 357–384.

— . 1994. *Time Series Analysis*. Princeton University Press, Princeton NJ.

Hautsch, Nikolaus, and Yangguoyi Ou. 2008. "Stochastic Volatility Estimation Using Markov Chain Simulation". Chap. 12 in *Applied Quantitative Finance*, 249–274. Springer.

Heaton, John, and Deborah J. Lucas. 1996. "Evaluating the Effects of Incomplete Markets on Risk Sharing and Asset Pricing". *Journal of Political Economy* 104 (3): 443–487.

Heiss, Florian, and Viktor Winschel. 2008. "Likelihood Approximation by Numerical Integration on Sparse Grids". *Journal of Econometrics* 144 (1): 62–80.

Henkel, Sam James, J Spencer Martin, and Federico Nardari. 2011. "Time-Varying Short-Horizon Predictability". *Journal of Financial Economics* 99 (3): 560–580.

Herbst, Edward P, and Frank Schorfheide. 2015. *Bayesian Estimation of DSGE Models*. Princeton University Press, Princeton NJ.

Hong, Harrison, Jeremy C Stein, and Jialin Yu. 2007. "Simple Forecasts and Paradigm Shifts". *The Journal of Finance* 62 (3): 1207–1242.

Jensen, Jens Ledet, and Niels Væver Petersen. 1999. "Asymptotic Normality of the Maximum Likelihood Estimator in State Space Models". *Annals of Statistics* 27 (2): 514–535.

Johannes, Michael, Arthur Korteweg, and Nicholas Polson. 2014. "Sequential Learning, Predictability, and Optimal Portfolio Returns". *The Journal of Finance* 69 (2): 611–644.

Judd, Kenneth L. 1992. “Projection Methods for Solving Aggregate Growth Models”. *Journal of Economic Theory* 58 (2): 410–452.

Judd, Kenneth L., Felix Kubler, and Karl Schmedders. 2011. “Bond Ladders and Optimal Portfolios”. *Review of Financial Studies* 24 (12): 4123–4166.

Kandel, Shmuel, and Robert F Stambaugh. 1996. “On the Predictability of Stock Returns: An Asset-Allocation Perspective”. *The Journal of Finance* 51 (2): 385–424.

Keim, Donald B, and Robert F Stambaugh. 1986. “Predicting Returns in the Stock and Bond Markets”. *Journal of Financial Economics* 17 (2): 357–390.

Kitamura, Yuichi. 2007. “Empirical Likelihood Methods in Econometrics: Theory and Practice”. Chap. 7 in *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, ed. by Richard Blundell, Whitney Newey, and Torsten Persson, 3:174–237. Econometric Society Monographs. New York: Cambridge University Press.

Kopecky, Karen A., and Richard M. H. Suen. 2010. “Finite State Markov-Chain Approximations to Highly Persistent Processes”. *Review of Economic Dynamics* 13 (3): 701–714.

Krueger, Dirk, and Felix Kubler. 2004. “Computing Equilibrium in OLG Models with Stochastic Production”. *Journal of Economic Dynamics and Control* 28 (7): 1411 –1436.

Kullback, Solomon, and Richard A. Leibler. 1951. “On Information and Sufficiency”. *Annals of Mathematical Statistics* 22 (1): 79–86.

Leroux, Brian G. 1992. “Maximum-Likelihood Estimation for Hidden Markov Models”. *Stochastic Processes and their Applications* 40 (1): 127–143.

Lettau, Martin, and Sydney C Ludvigson. 2010. “Measuring and Modeling Variation in the Risk-Return Tradeoff”. In *Handbook of Financial Econometrics, vol. 1*, ed. by Yacine Aït-Sahalia and Lars-Peter Hansen, 617–690.

Lettau, Martin, and Stijn Van Nieuwerburgh. 2008. “Reconciling the Return Predictability Evidence”. *Review of Financial Studies* 21 (4): 1607–1652.

Levintal, Oren. 2014. “Fifth Order Perturbation Solution to DSGE Models”. http://ssrn.com/abstract=2364989.

Lindvall, Torgny. 1992. *Lectures on the Coupling Method*. Wiley, New York, NY.

Maliar, Lilia, and Serguei Maliar. 2015. "Merging Simulation and Projection Approaches to Solve High-Dimensional Problems with an Application to a New Keynesian Model". *Quantitative Economics* 6 (1): 1–47.

McLean, R David, and Jeffrey Pontiff. 2016. "Does Academic Research Destroy Stock Return Predictability?" *The Journal of Finance* 71 (1): 5–32.

Mehra, Rajnish, and Edward C. Prescott. 1985. "The Equity Premium: A Puzzle". *Journal of Monetary Economics* 15 (2): 145–161.

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

Owen, Art B. 2001. *Empirical Likelihood*. Monographs on Statistics and Applied Probability 92. Chapman & Hall/CRC.

Pástor, L'uboš, and Robert F Stambaugh. 2009. "Predictive Systems: Living with Imperfect Predictors". *The Journal of Finance* 64 (4): 1583–1628.

Paye, Bradley S, and Allan Timmermann. 2006. "Instability of Return Prediction Models". *Journal of Empirical Finance* 13 (3): 274–315.

Pesaran, M Hashem, and Allan Timmermann. 1995. "Predictability of Stock Returns: Robustness and Economic Significance". *The Journal of Finance* 50 (4): 1201–1228.

Pettenuzzo, Davide, Allan Timmermann, and Rossen Valkanov. 2014. "Forecasting Stock Returns Under Economic Constraints". *Journal of Financial Economics* 114 (3): 517–553.

Pohl, Walter, Karl Schmedders, and Ole Wilms. 2014. "Solving Asset-Pricing Models with Recursive Preferences". https://site.stanford.edu/sites/default/files/solvingassetpricingmodels.pdf.pdf.

Rapach, David E, Jack K Strauss, and Guofu Zhou. 2010. "Out-of-sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy". *Review of Financial Studies* 23 (2): 821–862.

Rapach, David E, and Mark E Wohar. 2006. "Structural Breaks and Predictive Regression Models of Aggregate US Stock Returns". *Journal of Financial Econometrics* 4 (2): 238–274.

Rapach, David E, and Guofu Zhou. 2013. "Forecasting Stock Returns". In *Handbook of Economic Forecasting, vol. 2A*, ed. by Graham Elliott and Allan Timmermann, 328–383.

Rietz, Thomas A. 1988. "The Equity Risk Premium: A Solution". *Journal of Monetary Economics* 22 (1): 117–131.

Robinson, Peter M. 1989. "Nonparametric Estimation of Time-Varying Parameters". In *Statistical Analysis and Forecasting of Economic Structural Change*, 253–264. Springer.

Rossi, Alberto G, and Allan Timmermann. 2015. "Modeling Covariance Risk in Merton's ICAPM". *Review of Financial Studies*: hhv015.

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

Schmitt-Grohé, Stephanie, and Martín Uribe. 2004. "Solving Dynamic General Equilibrium Models Using a Second-Order Approximation to the Policy Function". *Journal of Economic Dynamics and Control* 28 (4): 755–775.

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

Tanaka, Kenichiro, and Alexis Akira Toda. 2013. "Discrete Approximations of Continuous Distributions by Maximum Entropy". *Economics Letters* 118 (3): 445–450.

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

Tauchen, George, and Robert Hussey. 1991. "Quadrature-Based Methods for Obtaining Approximate Solutions to Nonlinear Asset Pricing Models". *Econometrica* 59 (2): 371–396.

Taylor, S.J. 1982. "Financial Returns Modelled by the Product of Two Stochastic Processes: A Study of the Daily Sugar Prices 1961-75". In *Time Series Analysis: Theory and Practice*, 1:203–226. North-Holland, Amsterdam.

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

Timmermann, Allan. 2000. “Moments of Markov Switching Models”. *Journal of Econometrics* 96 (1): 75–111.

Trefethen, Lloyd N. 2008. “Is Gauss Quadrature Better than Clenshaw-Curtis?” *SIAM Review* 50 (1): 67–87.

Tsao, Min. 2004. “Bounds on Coverage Probabilities of the Empirical Likelihood Ratio Confidence Regions”. *Annals of Statistics* 32 (3): 1215–1221. doi:10.1214/009053604000000337.

Tsao, Min, and Fan Wu. 2013. “Empirical Likelihood on the Full Parameter Space”. *Annals of Statistics* 41 (4): 2176–2196.

Tsionas, Efthymios G. 2003. “Exact Solution of Asset Pricing Models with Arbitrary Shock Distributions”. *Journal of Economic Dynamics and Control* 27 (5): 843–851.

Vavra, Joseph. 2014. “Inflation Dynamics and Time-Varying Volatility: New Evidence and an *Ss* Interpretation”. *Quarterly Journal of Economics* 129 (1): 215–258.

Wachter, Jessica A. 2013. “Can Time-Varying Risk of Rare Disasters Explain Aggregate Stock Market Volatility?” *The Journal of Finance* 68 (3): 987–1035.

Welch, Ivo, and Amit Goyal. 2008. “A Compehensive Look at the Empirical Performance of Equity Premium Prediction”. *Review of Financial Studies* 21 (4): 1455–1508.

Wu, Jing Cynthia, and Fan Dora Xia. 2016. “Measuring the Macroeconomic Impact of Monetary Policy at the Zero Lower Bound”. *Journal of Money, Credit and Banking* 48 (2-3): 253–291.

Zhang, Lu. 2005. “The Value Premium”. *Journal of Finance* 60 (1): 67–103.