UNCERTAINTY AND LEARNING IN DYNAMIC FINANCIAL ECONOMETRICS

Paul Sangrey

A DISSERTATION

in

Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2019

Supervisor of Dissertation

Francis X. Diebold
Professor of Economics

Co-Supervisor of Dissertation

Frank Schorfheide
Professor of Economics

Graduate Group Chairperson

Jesús Fernández-Villaverde
Professor of Economics

Dissertation Committee

Amir Yaron, Professor of Finance

Xu Cheng, Associate Professor of Economics

UNCERTAINTY AND LEARNING IN DYNAMIC FINANCIAL ECONOMETRICS

Paul Sangrey

2019

For my parents, who taught me to love learning and how to teach myself.

ACKNOWLEDGMENT

I am indebted to my advisors, Francis X. Diebold and Frank Schorfheide, as well as the other members of my committee, Xu Cheng and Amir Yaron, for your invaluable guidance and encouragement. Without your help, I would have been unable to write this dissertation. In particular, I want to thank Francis X. Diebold who spent many hours helping me make my research approachable and encouraging me each step along the way. I am deeply beholden to each of my coauthors: Minsu Chang, Xu Cheng, and Eric Renault. Your contributions to our joint work are fundamental, and I hope that I taught you a tenth as much as you taught me. Thank you also to Jaap Abbring, Karun Adusumilli, Hengjie Ai, Benjamin Connault, Francis J. DiTraglia, Winston Wei Dou, Jesús Fernández-Villaverde, and Laura Liu for your invaluable feedback.

I also want to thank my cohort- and office-mates: in particular, Minsu Chang, Yiran Chen, Amanda Chuan, Matt Davis, Ryan Fackler, Mallick Hossain, Nitin Krishnan, and Hanna Wang. It has been a great ride. I do not know how I could have made it through without you. May the next step in your journey be as memorable as preliminary examinations, the job market, and everything in between have been. I am sure you will tackle it with the fortitude you displayed here. I also want to thank the friends I have made in Philadelphia, especially those from my community group for providing encouragement and rest from the stress of research.

Last, but certainly not least, I must thank my family. To Rebekah, Daniel, Anna, and my nieces for joining me here in Philadelphia and helping to make it home. To my parents for your patience, encouragement, and providing a home to visit. To Priscilla, Colin, Charles, Stephen, and Joshua for much-needed love and support. I owe each one of you a great deal.

ABSTRACT

UNCERTAINTY AND LEARNING IN DYNAMIC FINANCIAL ECONOMETRICS

Paul Sangrey

Francis X. Diebold and Frank Schorfheide

Every day the news reminds us that we live in a complex, ever-changing world. Against that background, this dissertation studies the econometrics of the interaction between time-varying uncertainty and learning. In particular, it develops parsimonious nonparametric methods for estimating risk in real time. The first two chapters develop tractable models and estimators for entire densities. The third chapter provides identification-robust inference for the prices of market and volatility risk when volatility exhibits complex dynamics.

The first chapter, "Jumps, Realized Densities, and News Premia," studies how jumps affect asset prices. It derives both a tractable nonparametric continuous-time representation for the price jumps and an implied sufficient statistic for their dynamics. This statistic — jump volatility — is the instantaneous variance of the jump part and measures news risk. It also develops estimators for the volatilities and nonparametrically identifies continuous-time jump dynamics and associated risk premia. It also provides a detailed empirical application to the S&P 500, showing that the jump volatility commands a smaller premium than the diffusion volatility does.

The second chapter, "Bypassing the Curse of Dimensionality: Feasible Multivariate Density Estimation," is coauthored with Minsu Chang and studies nonparametrically estimating multivariate densities. Most economic data are multivariate and estimating their densities is a classic problem. However, the curse of dimensionality makes nonparametrically estimating the data's density infeasible when there are many series. This chapter does not seek to provide estimators that perform well all of the time (it is impossible) but instead adapts ideas from the Bayesian compression literature to provide estimators that perform well most of the time.

The third chapter, "Identification-Robust Inference for Risk Prices in Structural Stochastic Volatility Models," is coauthored with Xu Cheng and Eric Renault and studies the identification problems inherent to measuring compensation for risk in stochastic volatility asset pricing models. Disentangling the channels by which risk affects expected returns is difficult and poses a subtle identification problem that invalidates standard inference. We adapt the conditional quasi-likelihood ratio test Andrews and Mikusheva (2016) develop in a GMM framework to a minimum distance framework to provide uniformly valid confidence sets.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

Chapter 1

INTRODUCTION

Our world is one of ever-increasing complexity. Investors and policymakers continually gain access to messy new data they must learn from and react to in real-time. These reactions both drive and are driven by the nonlinearity and non-Gaussianity that characterize financial and macroeconomic data. For example, investors see hundreds of news releases on a Bloomberg terminal each day. This new information causes their beliefs to jump. Understanding how investors and policymakers solve these learning problems poses many empirical challenges.

This dissertation is composed of three substantive chapters, along with this introduction, that address some of these challenges. Chapter 2, "Jumps, Realized Densities, and News Premia," analyzes the effects of jumps in high-frequency data. Equities and other assets that investors trade in continuous-time exhibit complex dynamics. In particular, prices jump hundreds of times per day. This paper develops a nonparametrically-identified parsimonious representation for jumps in price processes and uses it to analyze the jumps' stylized features and effects on expected returns. Chapter 3, "Bypassing the Curse of Dimensionality: Feasible Multivariate Density Estimation," is coauthored with Minsu Chang and studies feasible nonparametric density estimation. Estimating multivariate densities is very difficult when the number of series is large because of the curse of dimensionality. We use ideas from the Bayesian compression literature to develop a nonparametric Bayesian estimator that works well most of the time. Chapter 4, "Identification Robust Inference for Risk Prices in Structural Stochastic Volatility Models," is coauthored with Xu Cheng and Eric Renault. This chapter provides identification-robust inference for the key parameters governing investors' risk aversion in highly nonlinear, heteroskedastic environments that typify modern asset pricing.

To delve further into "Jumps, Realized Densities, and News Premia," about fif-

teen years ago, Barndorff-Nielsen and Shephard (2002) and Andersen et al. (2003) substantially enhanced our understanding of the continuous-time information structure of asset returns in the (conditionally Gaussian) diffusion case. They did this by providing the nonparametric Realized Volatility estimator for the integrated diffusion volatility and showing the diffusion volatility entirely determines the short-horizon price dynamics. Volatility can be time-aggregated in closed form, providing closed-form expressions for discrete-time distributions. Another series of classic papers shows that the instantaneous covariance between prices and marginal utility determine risk premia, (Merton 1973; Bollerslev, Engle, and Wooldridge 1988).

However, hundreds of quantitatively relevant news releases strike financial markets every day and cause the prices to jump. Aït-Sahalia and Jacod (2009a, 2009b, 2012) even show that models with infinitely many jumps fit the data better than models with only finitely many jumps. Also, simple covariance-based characterizations of risk premia no longer hold.

This paper derives a parsimonious representation for prices with nonparametrically identified jump dynamics. It provides both a tractable continuous-time representation and an implied sufficient statistic for the jump dynamics. This statistic — jump volatility — is the instantaneous variance of the jump part and measures news risk. The resulting realized density then depends, exclusively, on the diffusion and jump volatilities in continuous-time. In other words, volatilities control all of the distribution's short-horizon dynamics. This paper time-aggregates this representation and derives closed-form representations for the discrete-time densities and volatilities.

It then develops an estimator for the instantaneous jump volatility, thereby showing that high-frequency data nonparametrically identifies short-horizon jump dynamics. It also provides estimators for all of the other volatilities and the realized density. It applies these estimators to high-frequency data on the S&P 500, providing several new stylized facts. This paper then nonparametrically characterizes continuous-time risk-premia in the presence of recursive utility and jumps. In particular, it shows that the jump volatility is economically and statistically significantly less than the diffusion volatility premium. This result shows investor's preferences are not time-separable and that we need at least two factors that move at high-frequency to explain movements in risk premia.

The second chapter, "Bypassing the Curse of Dimensionality: Feasible Multivariate Density Estimation," with Minsu Chang also studies density estimation. Real financial data often displays substantial nonlinearity and non-Gaussianity. Also, nonparametrically estimating the densities of multivariate data becomes infeasible when the number of series, D, is larger than 2 or 3. This phenomenon is called the *curse of dimensionality*.

Nonparametric estimators simultaneously solve two problems. First, they approximate the density. Second, they estimate the parameters that govern this approximation. The original curse of dimensionality papers, such as Stone (1980, 1982), relate this approximation problem to the previously existing deterministic function approximation literature. They show that requiring the estimators to be consistent causes the estimator and the deterministic approximation to use the same number of terms asymptotically. Deterministic approximations can be viewed as subdividing a D-dimensional hypercube into hypercubes of width $1/T$, where T is the number of periods. This procedure requires T^D terms, and so convergence rates must decline exponentially-fast in D.[1] More recently, the Bayesian compression literature has studied random function approximations and relates them to approximating balls in high dimensions, (Klartag and Mendelson 2005; Talagrand 2014). High-dimensional random variables cluster on balls instead of hypercubes and balls have substantially less volume than hypercubes have in high-dimensions, and so this leads to significantly more parsimonious approximations.

Thus far, the Bayesian compression literature has focused on the function approximation problem and the closely-related data compression problem. We apply these ideas to estimating multivariate densities. In particular, we develop a dynamic generalization of the infinite-mixture representation commonly used in the Bayesian nonparametric literature, (Ghosal and van der Vaart 2017) and show how draws from the posterior can be viewed as random approximations. Because infinite-mixtures can approximate a broad class of densities, this procedure only requires a few assumptions on the data generating process (DGP). Also, we can estimate both unconditional and transition densities for both i.i.d. and Markov data.

For any finite T, we construct a bound for the number of mixture components

[1] The particular way in which D enters into the exponent for a particular sieve depends upon the smoothness of the class of functions being considered.

as a function of T that holds with high probability with respect to the random approximation algorithm. We then relate this random approximation algorithm to the prior. This argument lets us convert bounds on the mixture's complexity into convergence rates for the estimators. Our estimators' convergences rates — $\sqrt{\log(T)}/\sqrt{T}$ in the unconditional case and $\log(T)/\sqrt{T}$ in the conditional case — depend on D only through the constant term.

The third chapter, "Robust Inference for Risk Prices in Structural Stochastic Volatility Models," which is coauthored with Xu Cheng and Eric Renault, considers how investors optimally trade off risk and return in environments with complex volatility dynamics. Some seminal early papers propose a static trade-off between risk and expected return, most notably the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965). In practice, volatility varies over time. A significant strand of the recent literature examines the dynamic tradeoff between volatility and returns, including structural stochastic volatility models such as Christoffersen, Heston, and Jacobs (2013), Bansal et al. (2014), and Dew-Becker et al. (2017). In nonlinear models such as these where uncertainty changes in complex ways, investors care not just about how an asset's returns co-move with the volatility but also about how they co-move with changes in volatility.

Consequently, changes in volatility affect risk premia through two channels: (1) the investor's willingness to tolerate high volatility in order to get high expected returns as measured by the market return risk price, and (2) the investor's direct aversion to changes in future volatility as measured by the volatility risk price. We adopt the discrete-time exponentially affine model of Han, Khrapov, and Renault (2018). They show that the identification of the volatility risk price depends on a substantial leverage effect, which is the correlation between innovations to returns and volatility. However, this leverage effect is difficult to estimate, and often small, (Aït-Sahalia, Fan, and Li 2013; Bandi and Renò 2012). This low signal-to-noise ratio, which we model using weak identification, makes the asymptotic approximations perform poorly in finite samples.

This paper provides confidence sets for the risk prices that are robust to this weak identification by developing a minimum distance criterion that uses link functions between the structural parameters and a set reduced-form parameters whose distri-

butions can be approximated well using standard asymptotics. These link functions are well-behaved in terms of the reduced-form parameters, but not the structural parameters. We use this minimum distance criterion to construct a uniformly valid confidence set by inverting a conditional quasi-likelihood ratio (QLR) test. The critical value is constructed by conditioning on a sufficient statistic for an infinite-dimensional nuisance parameter. We adapt this test from Andrews and Mikusheva (2016) who developed it in a GMM framework. We show it works in the minimum distance context considered here, provide conditions for its asymptotic validity, and provide a detailed simulation algorithm to compute it.

Chapter 2

JUMPS, REALIZED DENSITIES, AND NEWS PREMIA

By Paul Sangrey

Announcements and other news continuously barrage financial markets, causing asset prices to jump hundreds of times each day. If price paths are continuous, the diffusion volatility nonparametrically summarizes the return distributions' dynamics, and risk premia are instantaneous covariances. However, this is not true in the empirically-relevant case involving price jumps. To address this impasse, I derive both a tractable nonparametric continuous-time representation for the price jumps and an implied sufficient statistic for their dynamics. This statistic — jump volatility — is the instantaneous variance of the jump part and measures news risk. The realized density then depends, exclusively, on the diffusion volatility and the jump volatility. I develop estimators for both and show how to use them to nonparametrically identify continuous-time jump dynamics and associated risk premia. I provide a detailed empirical application to the S&P 500 and show that the jump volatility premium is less than the diffusion volatility premium.

2.1 Introduction

The study of how individuals' react to time-varying risk forms the core of modern finance and macroeconomics. Asset pricing, portfolio allocation, and performance evaluation all require investors to asses the risk they face in real time. Moreover,

optimal financial regulation requires trading off risk and return at the societal level, and real-time risk measures form its core as well. The most general measure of this risk is the distribution of future returns as a function of the information available.

About fifteen years ago, Barndorff-Nielsen and Shephard (2002) and Andersen, Bollerslev, Diebold, and Labys (2003) substantially enhanced our understanding of the volatility by providing the nonparametric Realized Volatility estimator for the integrated diffusion volatility. Moreover, they showed that as long as price paths are continuous (that is, they are stochastic volatility diffusions) the diffusion volatility entirely determines the continuous-time martingale dynamics. They also derived closed-form expressions for the discrete-time distributions as functions of integrated diffusion volatility by time-aggregating the continuous-time measures. Another series of classic papers shows that the instantaneous covariance between prices and investors' stochastic discount factors determine risk premia, (Merton 1973; Breeden 1979; Bollerslev, Engle, and Wooldridge 1988a).

However, hundreds of quantitatively relevant news releases strike financial markets every day and cause the prices to jump. Aït-Sahalia and Jacod (2009a, 2009b, 2012) even show that models with infinitely many jumps fit the data better than models with only finitely many jumps. Meanwhile, various papers, such as Drechsler and Yaron (2011) and Ai and Bansal (2018), show the parsimonious covariance-based characterizations of risk premia mentioned above fail when prices jump.

At present, however, no parsimonious representation with nonparametrically identified dynamics exists for jump processes. To address this impasse, I derive both a tractable nonparametric continuous-time representation for the price jumps and an implied sufficient statistic for their dynamics. This statistic — jump volatility — is the instantaneous variance of the jump part and measures news risk. The resulting realized density then depends, exclusively, on the diffusion and jump volatilities in continuous-time. In other words, volatilities control all of the distribution's short-horizon dynamics. I then time-aggregate this representation and derive closed-form representations for the discrete-time densities and volatilities.

To enable taking this theory to the data, I develop an estimator for the instantaneous diffusion volatility by extending Jacod et al. (2009). I nonparametrically identify the jump part of the dynamics, in the presence of stochastic diffusion volatil-

ity, by deriving the first estimator for instantaneous jump volatility. I time-aggregate both estimators to provide estimators for the daily diffusion and jump volatilities. I then apply these estimators to high-frequency data on the S&P 500. This provides several new stylized facts. First, diffusion and jump volatility are highly positively correlated. Second, like diffusion volatility, jump volatility is highly persistent, remaining high for extended periods of time during recessions.

I then connect jump volatility to consumption-based asset pricing by nonparametrically characterizing continuous-time risk-premia in the presence of recursive utility and jumps. My characterization shows how jump and diffusion volatility jointly determine risk premia and requires both terms in general. I then take my estimators to the data and show that the diffusion volatility commands an economically and statistically significant premium, as in Brandt and Kang (2004) and Lettau and Ludvigson (2010). I further show that the jump volatility is substantially less than the diffusion volatility premium. I show that this implies that investor's preferences are not time-separable and the data require at least two factors that move at high-frequency to explain movements in risk premia.

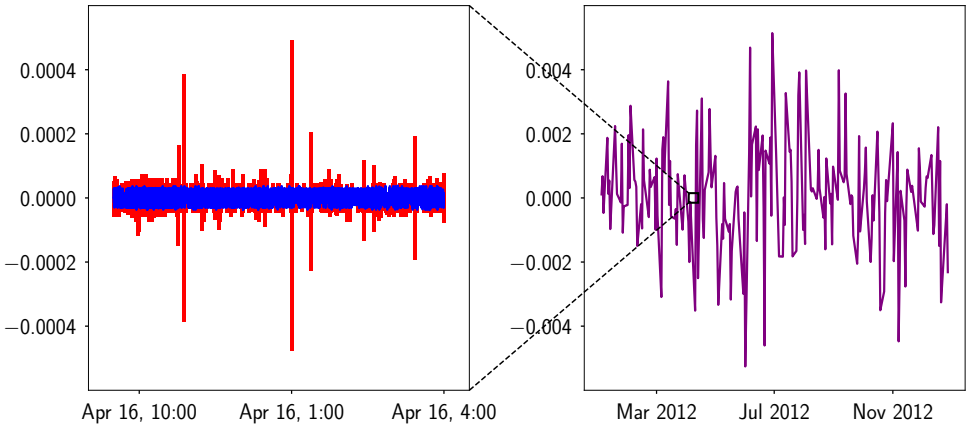
I lay out the paper as follows. The remainder of the introduction fixes ideas and explains the close connection between discontinuous information flows and jumps in asset prices. Section 2.2 relates my paper to the literature. Section 2.3 lays out the data generating process I use, while Section 2.4 proves the main representation theorem. Section 2.5 derives the estimators, and Section 2.6 characterizes their finite-sample performance in simulations. Section 2.7 describes my dataset, and Section 2.8 provides a series of new stylized facts concerning the jump volatility dynamics. I derive risk premia in the presence of recursive utility and jumps in Section 2.9 and show that the jump volatility premium is less than the diffusion volatility premium in Section 2.10. Section 2.11 concludes. The appendices contain the proofs and robustness checks.

Stylized Features of the Data

I motivated this project by claiming that prices jump extremely often and that news frequently and dramatically affect asset prices. The literature has shown this, but it is helpful to investigate the matter ourselves to fix ideas. We need high-frequency data

to identify these jumps, and so I start there. The data show jumps in price processes are ubiquitous and form a large portion of the price's variation. For example in Figure 2.1, I plot the daily log-return on the S&P 500 during 2012 and then zoom in on the 1-second return on April 16. The red lines are jumps in the prices identified by sampling the data once per second, and the blue lines contain the diffusion part of the process and jumps that are too small to identify easily. The behavior in this graph is entirely typical. I purposefully chose April 16, 2012, because it was a completely normal day in the markets.

Figure 2.1: S&P 500 Log-Return



(a) 1 Second (b) Daily

As we can see in Figure 2.1, prices jump extremely often and drive a great deal of the variation in the price. Estimates range from as low as $\approx 7\,\%$ to as high as $\approx 80\,\%$, (Pan 2002; Huang and Tauchen 2005; Santa-Clara and Yan 2010; Ornthanalia 2014). In particular, Aït-Sahalia and Jacod (2009a) find jumps drive $\approx 40\,\%$ of the squared variation in individual equities and $\approx 10\,\%$ of the variation in the market index using a ratio of bipower-type estimators. This wide divergence between various estimates likely arises from the difficulty in disentangling the infinite-activity jumps from the diffusive part. The precise percentage is not important for this paper. I estimate this proportion below, (Figure 2.7). Rather, the important takeaway is that jumps occur frequently enough to be important, and even $7\,\%$ of the variation in the market is economically meaningful.

Almost every paper that explicitly tests for the degree of activity finds infinitely-active jumps, or at the very minimum a massive number, (Aït-Sahalia, Mykland, and Zhang 2005; Bakshi, Carr, and Wu 2008; Aït-Sahalia and Jacod 2009a).[2] From both a modeling and pricing perspective, a large number of jumps and infinitely many are essentially equivalent in practice, as shown in detail below. Even if the literature has not reached a consensus on the precise number and magnitude of the jumps, jumps are clearly ubiquitous and crucial to understanding price dynamics.

What Causes Jumps?

To understand Figure 2.1a, we need to understand what precisely a jump is. There are two equivalent characterizations. First, a jump is a discontinuity in the price process. The price changes by such a large amount over such a small period that we cannot draw a continuous line through it. However, this is a mathematical definition; we would like an economic characterization. What are jumps economically?

Various authors, such as Andersen, Bollerslev, Diebold, and Vega (2003, 2007), Beechey and Wright (2009), and Lahaye, Laurent, and Neely (2011), argue that jumps are responses of prices to news releases. Most of these papers consider the effects of macroeconomic announcements on prices. They start with a series of news items that they a priori believe to be important and show that the prices react effectively instantaneously.[3] However, in general, many different sources cause discontinuities in investor's information sets. Other sources include Congressional decisions, a startup announcing a new product line on Twitter, effectively anything in a Bloomberg or Associated Press feed relevant for asset pricing, even private communications between financiers. The last example highlights the utter impossibility of listing all the potentially relevant events. We cannot construct investors' actual information sets. (Note, this paper uses *news* quite broadly. It refers to any discontinuous change in information, not just traditional news sources such as newspapers.) As these examples illustrate, news often come at unpredictable times and only a few investors may observe it, and so a priori choosing which news items are relevant necessarily excludes

[2]The single exception is Christensen, Oomen, and Podolskij (2014), which I discuss in Section 2.8.

[3]By far the most commonly studied announcements are the Federal Open Market (FOMC) announcements.

many relevant items. Besides, there is no reason to assume that the resultant price change is in any way substantial. Many news items cause a small, but measurable, impact on the prices.

The connection between news and jumps is rather intuitive, and the empirics in the papers mentioned substantiate it. However, the connection is even more fundamental. Delbaen and Schachermayer (1994) show no-arbitrage implies prices are semimartingales.[4] In that framework, which is standard in high-frequency econometrics, jump times are times when the information contained in prices jumps. In other words, jump times are times when the representative investor's information set evolves discontinuously.

To make this claim precise, consider the following. Let $P(t)$ be a price process, and $\mathcal{F}_t^p$ be its natural filtration.[5] $\mathcal{F}_t^p$ contains the events that are known at time t to anyone observing the history of prices up to t. In other words, it is the part of the representative investor's information set relevant for pricing. Then, $P(t)$ jumps at τ if and only if $\mathcal{F}_t^p$ jumps at τ. Since standard economic intuition implies that causality runs from information to prices, $P(t)$ jumps whenever the available information evolves discontinuously, that is a news item is released. This relationship implies that we can identify news shocks by looking for jumps in the prices. Consequently, since the jump volatility is a sufficient statistic for jumps dynamics, it measures news risk.

Theorem 2.1 (Jump Times are News Times)**.** *Consider a stopping time τ. Let $P(t)$ be a price process satisfying no-arbitrage. Then its natural filtration — $\mathcal{F}_t^p$ — contains all of the information in the representative investor's information set relevant for asset pricing, and $\mathcal{F}_\tau^p \neq \mathcal{F}_{\tau-}^p$ if and only if $P(t)$ jumps at τ, where $\mathcal{F}_{t-}^p$ is the associated predictable filtration.*

This result also explains why not all price changes are jumps. Prices do not always reflect new information instantaneously. Some information takes time to process before the market participants can use it effectively. For example, after a firm announces its earnings, the headline reveals much of the information. However, many

[4]Throughout this paper, I use no-arbitrage to refer to no-free-lunch with vanishing risk as is standard in continuous-time finance.

[5]This paper uses functional notation to refer stochastic processes and subscript notation to refer to discrete-time objects, e.g., $P(t)$ is the price process, and P_t is the price at t. I time-index objects by the first time they enter the representative investor's information set.

articles still analyze what each release implies about both the stock in question and other related assets. As various investors update their beliefs and buy or sell accordingly, other market participants see the information that is now revealed by the prices and buy or sell themselves. This process changes the asset's price, and it takes time. Gürkaynak, Kısacıkoğlu, and Wright (2018) make exactly this distinction and that it substantially improves forecasting performance.

2.2 Literature Review

Since questions concerning volatility, news, and risk-return trade-offs are central to finance and economics, a few different literatures study the questions considered in this paper. Consequently, I cannot hope to survey the literature adequately. I can only cover a few of the closest related papers.

Jumps in Asset Prices

The first literature that I build upon is the econometrics literature that studies jumps in asset prices. Barndorff-Nielsen and Shephard (2006) develop the bipower variation estimator to disentangle jumps and diffusive variation. Since then, several authors have shown that jumps are both frequent and economically important, including Andersen, Bollerslev, and Diebold (2007), Bollerslev, Law, and Tauchen (2008), and Aït-Sahalia and Jacod (2009b). The critical difference between my estimates of jump variation and previous bipower variation estimates is that I measure ex-ante jump variation, while previous papers measure ex-post variation. This difference is essential for two reasons. First, my density characterization relies upon an ex-ante characterization. Second, investors price ex-ante risk, and so my measure is a core object in pricing, while ex-post jump variation cannot be priced. Other authors have argued they are not just statistically significant, but economically as well. For example, we also need them to price derivatives, such as (Pan 2002; Branger, Schlag, and Schneider 2008; Todorov 2010, 2011).

In Section 2.1, I discuss the literature that measures the magnitude of jump variation and the jump intensities. I will not repeat that discussion here except to recall the twofold consensus. First, asset prices contain a vast number of jumps. Jumps are

likely infinitely-active, or, at a minimum, have a very high intensity. Second, jumps constitute an economically and statistically significant portion of the price variation.

I rely on these results in three ways. First, as motivation for the project. Second, as evidence that my empirical results are reasonable. Third, and most importantly, I rely heavily on these empirical facts in that I assume prices have infinitely-active jumps. This assumption is somewhat unusual, but not unique. For example, Gallant and Tauchen (2018) considers a similar class of processes.

Gallant and Tauchen (2018) is arguably the closest related paper in the econometrics literature. It is the only other paper that nonparametrically relates jump variation to the distribution of returns. It is a fascinating paper and provides useful estimates for the intensity of jump processes. However, their representation relies on Todorov and Tauchen (2014) and so can only handle small jumps.

Representing Price Processes

The second literature that this paper builds upon is the stochastic process representation literature. The main contribution to this literature is Theorem 2.4 and its corollaries. This theorem provides general conditions under which jump processes are stochastic volatility variance-gamma processes. The variance-gamma process is a Lévy process first introduced by Madan, Carr, and Chang (1998).

The first main time-change method for representing price processes is the Dambis, Dubins & Schwarz theorem, (Dambis 1965; Dubins and Schwarz 1965). Theorem 2.4 is the jump analog of that theorem. Epps and Epps (1976) and various subsequent authors relate this time-change to "business-time," that is the speed at which information gets released into the market, creating the mixture-of-distributions hypothesis.

Various authors partially extend these results to the jump case. Monroe (1978) shows that any semimartingale can be embedded into Brownian motion, but did not construct this embedding explicitly. Geman, Madan, and Yor (2002) shows that this embedding is not identified. More recently, Todorov and Tauchen (2014) show how to embed the jump processes' infinitesimal jumps into an α-stable process using bipower-variation. Infinitesimal means, here, that the maximum jump size approaches zero as the increment size approaches zero. By using an ex-ante measure of jump variation,

instead of an ex-post one like Todorov and Tauchen (2014) do, I can handle large jumps as well.

In Section 2.4, I time-aggregate these continuous-time representations to discrete-time under some additional assumptions. In doing this, I follow Barndorff-Nielsen and Shephard (2002) and Andersen, Bollerslev, Diebold, and Labys (2003) who provide analogous results for diffusive processes. Barndorff-Nielsen and Shiryaev (2010) further analyze these representations, providing a useful survey of the current state of the literature.

Pricing Assets with Recursive Utility

The curvature in investors' preferences, i.e., their risk appetite, implies a negative relationship between expected returns and volatility. Consequently, many different papers estimate this relationship, and I cannot comprehensively survey this literature. Surprisingly, the empirical evidence has proven much less conclusive than the theory. Bollerslev, Engle, and Wooldridge (1988b), Harvey (1989), Ghysels, Santa-Clara, and Valkanov (2005), and Lettau and Ludvigson (2010) find a positive relationship between expected returns and volatility. Campbell (1987), Pagan and Hong (1991), Glosten, Jagannathan, and Runkle (1993), and Brandt and Kang (2004) actually find a negative relationship. Besides, many authors argue that the instantaneous correlation, which is often called a "volatility-feedback" or leverage effect is negative, both in continuous-time (Bandi and Renò 2012; Aït-Sahalia, Fan, and Li 2013) and in discrete-time (Engle and Ng 1993; Yu 2005). This negative sign is likely the main reason why estimating the risk premium has proven difficult. The researcher must disentangle two different relationships, risk-premia and volatility feedback, that have opposite signs.

Investors' utility functions are not the only place their preferences can display curvature. Some examples of models with curvature in their certainty equivalence functionals (CEF) include max-min expected utility, (Gilboa and Schmeidler 1989; Epstein and Schneider 2003), models with ambiguity aversion (Hansen and Sargent 2001; Klibanoff, Marinacci, and Mukerji 2005; Ju and Miao 2012), and Epstein-Zin recursive utility (Epstein and Zin 1989; Duffie and Epstein 1992). This additional curvature leads to additional risk-return trade-offs. Ai and Bansal (2018) show pre-

mia for this curvature cause announcements to be priced differently. Hence, simple covariance-based explanations for risk-premia break down.

Arguably the closest related paper in the finance literature, Ai and Bansal (2018), is inspired by a recent surprising stylized fact presented by Lucca and Moench (2015): the majority of the equity premium occurs on the days around when the Federal Open Market Committee (FOMC) makes it announcements. This paper extends Ai and Bansal (2018) by deriving risk-premia in continuous-time for models with recursive utility and jumps. I then show that this additional term is closely related to the jump volatility. This characterization shows that, in general, the theory requires two pricing factors that move at high-frequency.

2.3 Data Generating Process

In this section, I describe the data generating process (DGP). Models of prices differ along two different dimensions. They can be either continuous or discrete, and they can be either in continuous-time or in discrete-time. I write down a continuous-time DGP with jumps and derive the discrete-time representation from it. I also discuss the purely continuous special case that my DGP nests in parallel to provide a point of comparison.

Continuous-Time DGP

We know from Dambis (1965) and Dubins and Schwarz (1965) that continuous Itô semimartingales are stochastic volatility diffusions. That is, for some drift, $\mu(t)$, and diffusion volatility, $\sigma^2(t)$, we can represent the log-price process as

$$dp(t) = \mu(t)\,dt + \sigma(t)\,dW(t), \tag{2.1}$$

where $W(t)$ is a Wiener process. However, as mentioned in the introduction, asset prices are not continuous processes, and so the models considered above cannot fully replicate the stylized facts in the data. For example, because $W(t)$ is a Wiener process, conditional on $\sigma^2(t)$ the price increments are Gaussian variables and so do not have fat tails.

The standard nonparametric way to add jumps to these models is to assume that prices are Itô semimartingales. This representation is quite general because it only requires that prices are semimartingales and each of the components of the process have time-derivatives. The log-price being an Itô semimartingale implies that the jump part is an integral with respect to a Poisson random measure. Let n be a Poisson random measure with associated compensator, ν. The function $\delta(s,x)$ controls the magnitude of the process. In general, the triple (δ, n, ν) is not unique, which allows us to pick a particularly useful representation later.

Definition 2.1. Jump-Diffusion DGP (Grigelionis Form of an Itô Semimartingale)

$$p(t) = p(0) + \int_0^t \mu(s)\,ds + \int_0^t \sigma(s)\,dW(s) + \int_0^t \int_X \delta(s,x)1\{\|\delta(x,s)\| \leq \mathbf{1}\}(n-\nu)(ds,dx) + \int_0^t \int_X \delta(s,x)\mathbf{1}\{\|\delta(x,s)\| > 1\}n(ds,dx)$$

I simplify Definition 2.1 by adding the following assumption.

Assumption Square-Integrable. The process, $p(t)$, is locally-square integrable.

Assumption 2.1 is relatively innocuous in practice as it holds as long as returns themselves have conditional variances. Although many high-frequency papers initially allow for jumps that are so large no compensator exists, they almost always restrict themselves to processes that satisfy Assumption 2.1 when they derive estimators. Making Assumption 2.1 now simplifies notation because it implies that the jump measure has a predictable compensator. I also assume without loss of generality that $p(0) = 0$, giving

$$p(t) = \int_0^t \mu(s)\,ds + \int_0^t \sigma(s)\,dW(s) + \int_0^t \int_X \delta(s,x)(n-\nu)(ds,dx). \tag{2.2}$$

I now assume without loss of generality that n is a standard Poisson random measure. In other words, for finite open sets $A \subset B \subset X$, the event $1\{x \in A \,|\, x \in B\}$ is Poisson distributed with intensity $\int_A \delta(t,x)\,dx / \int_B \delta(t,x)\,dx$, which is the $\Pr(x \in A \,|\, x \in B)$. In particular, the function δ completely controls the process's dynamics.

This representation is quite general and can handle a great variety of different price processes. However, it is rather intractable, and not identified. For each time t, $\delta(t,\cdot)$ is a function of x. For each set A above, we have a Poisson process. It takes

infinitely-many finite-sized open sets to form a valid partition of $\mathbb{R}$. Each of these infinitely-many sets has a time-varying Poisson intensity. The δ function combines these intensities in the appropriate way. To estimate this process, we would have to estimate these infinitely-many intensity parameters for each time τ using only one realization. That is obviously impossible. In other words, we must estimate a entire function using only one datapoint. Also, it is not obvious how to time-aggregate this representation, i.e., parsimoniously map it to discrete-time.

Discrete-Time DGP

Before I relate the discrete- and continuous-time returns, we must know what a discrete-time return is. The discrete-time return is just the change in (an increment of) the price process over some length of time, say a day.[6] Throughout, I use subscripts to refer to daily objects, and functional notation to refer to stochastic processes, as mentioned previously, I index each variable by the time it first becomes known to the investor, i.e., becomes measurable with respect to the filtration induced by the prices. For example, r_t is the daily return on date t, while $p(t)$ is the log-price at time t.

Definition 2.2. Daily Return

$$r_t \coloneqq \int_{t-1}^{t} dp(t).$$

This return has a density — h — in each period given the available information at the end of the day before — $\mathcal{F}_{t-1}$.

Definition 2.3. Daily Density

$$r_t \,|\, \mathcal{F}_{t-1} \sim h\left(r_t \,|\, \mathcal{F}_{t-1}\right).$$

This predictive density fully characterizes the statistical risk that investors face. In particular, any statistical measure of risk, such as Expected Shortfall or Value-at-Risk, is a statistic of this density.

[6] Throughout, I focus on daily returns whose length I normalize to one, but there is nothing special about a day. We could perform the same analysis over any discrete length of time.

Daily returns are not very well-behaved objects in that they are unpredictable and their distributions vary substantially over time. Furthermore, we only observe one observation for each $h(r_t \mid \mathcal{F}_{t-1})$. Since $\mathcal{F}_{t-1}$ grows each day, $h(r_t \mid \mathcal{F}_{t-1})$ is a function-valued time-varying parameter. Modeling such parameters is quite difficult. Hence the literature, e.g., Engle (1982), Bollerslev (1986), and Nelson (1991), focuses on representations for $h(r_t \mid \mathcal{F}_{t-1})$ in terms of a well-behaved sufficient statistic for the dynamics. The most common choice for x_t is some measure of volatility.

They use x_t to separate $h(r_t \mid \mathcal{F}_{t-1})$ into three parts. The first — x_t — is well-behaved and predictable and hence easily forecastable. The second is noise as far as prediction is concerned with associated density — f. It affects the risk investors face but not the density's dynamics. The third part — G — is a process governing x_t's dynamics.

Both f and G are fixed across-time, and G is simple if we chose x_t well. This gives

$$r_t \mid \mathcal{F}_{t-1} \sim h\left(r_t \mid \mathcal{F}_{t-1}\right) = \int_{x_t} f\left(r_t \mid x_t\right) dG\left(x_t \mid \mathcal{F}_{t-1}\right), \tag{2.3}$$

replacing the question how should we model $h(r_t \mid \mathcal{F}_{t-1})$ with three related questions. What should we use for x_t? What should use for f? What should we use for G?

For example, consider the following simple stochastic volatility model:

$$r_t \sim \sigma_t \mathcal{N}(0,1) \tag{2.4}$$

and

$$\log\left(\sigma_t^2\right) = \rho \log\left(\sigma_{t-1}^2\right) + \sigma_\sigma \mathcal{N}(0,1). \tag{2.5}$$

As is standard, this model uses volatility, σ_t^2, as x_t. Here the return is a Gaussian innovation with stochastic volatility — σ_t^2. Hence, f is a Gaussian distribution. The σ_t^2 follows an $AR(1)$ process in logs with persistence ρ and innovation variance σ_σ^2.

Now that we have a discrete-time DGP, we can define the *realized density*.

Definition 2.4 (Realized Density)**.**

$$RD_t \coloneqq f\left(r_t \mid x\right)\Big|_{x=x_t}$$

Just as the realized volatility, RV_t, is the particular value of the volatility that realizes in a given day, the realized density, RD_t, is the conditional density that

realizes that day. For example, in the model given above, the realized density is $f(r_t \mid x_t) = f(r_t \mid \sigma_t^2)$.

The realized density is useful because it separates the dynamic and static parts of the process. In addition, it is precisely the part of the likelihood that high-frequency data identifies, and Section 2.5 provides sufficient conditions for this identification. Once we have RD_t, we only need to model G. In practice, this is much simpler than modeling $h(r_t \mid \mathcal{F}_{t-1})$ directly because x_t is usually well-behaved.

2.4 Modeling Jump Processes

The previous section claimed that the most common choice for a sufficient statistic for the dynamics is some measure of volatility. This section constructs a new measure of volatility. This measure, unlike various realized measures in the literature, is an ex-ante measure. This distinction is fundamental to the representation constructed below.

Jump Volatility

The continuous-time data generating process in (2.1) implicitly defined the instantaneous diffusion volatility, $\sigma^2(t)$. It is the integrand in that representation. However, there is an equivalent characterization going back as far as Merton (1973) that is more useful for our purposes. This characterization gives $\sigma^2(t)$ its interpretation as an instantaneous variance; $\sigma^2(t)$ is the appropriately standardized variance of the diffusion part of the process over a shrinking interval. [7]

Definition 2.5 (Instantaneous Diffusion Volatility)**.**

$$\sigma_t^2 := \frac{1}{\Delta}\mathbb{E}\left[\left|p^D(t+\Delta) - p^D(t)\right|^2 \,\middle|\, \mathcal{F}_{t-}\right]$$

One key subtlety of this definition is that we are only using the information available before time t. Variances are forward-looking operators. This subtlety is not essential in the diffusion case. The ex-ante and ex-post measures coincide, and so the literature has not stressed it. In the jump case, however, it is fundamental.

[7] I use superscript D to refer to the diffusion part of the process.

Volatility's key advantage is that we can time-aggregate it easily. The daily volatility is just the integral (average) of the high-frequency volatility. This aggregation property is precisely what Barndorff-Nielsen and Shephard (2002) and Andersen, Bollerslev, Diebold, and Labys (2003) use to develop the Realized Volatility estimator for σ_t^2.

Definition 2.6 (Integrated Diffusion Volatility)**.**

$$\sigma_t^2 \coloneqq \int_{t-1}^{t} \sigma^2(t)\, ds.$$

The goal moving forward is to construct a sufficient statistic for the jump dynamics that also has this aggregation property. To do this, I define the *jump volatility* — $\gamma^2(t)$. A volatility is a variance, and so we can construct the jump analogue to Definition 2.5. I substitute the diffusion part of the prices — $p^D(t)$ with the jump part — $p^J(t)$. In other words, I define the instantaneous jump volatility as the local variance of the jump part — $p^J(t)$.

Definition 2.7 (Instantaneous Jump Volatility)**.**

$$\gamma^2(t) \coloneqq \frac{1}{\Delta}\mathbb{E}\left[\left|p^J(t+\Delta) - p^J(t)\right|^2 \,\middle|\, \mathcal{F}_{t-}\right].$$

The integrated jump volatility is defined in the obvious way.

Definition 2.8 (Integrated Jump Volatility)**.**

$$\gamma_t^2 \coloneqq \int_{t-1}^{t} \gamma^2(s)\, ds.$$

We can also define $\gamma^2(t)$ in terms of Definition 2.1. The jump volatility is the time-derivative of the predictable quadratic variation of the jump part of the process.

Theorem 2.2 (Jump Volatility and the Predictable Quadratic Variation)**.** *Let $p(t)$ be an Itô semimartingale satisfying Assumption Square-Integrable, then the following holds where $\langle p^J \rangle(t)$ is the predictable quadratic variation (angle-bracket) of $p^J(t)$:*

$$\gamma_t^2 = \int_{t-1}^{t} \gamma^2(s)\, ds = \int_{t-1}^{t}\int_{\mathcal{X}} \delta^2(s,x)\nu(dx,ds) = \langle p^J \rangle(t) - \langle p^J \rangle(t-1).$$

The are three main advantages of γ_t^2 over the jump part of the quadratic variation. First, since jump processes are not absolutely continuous, there is no ex-post analog to $\gamma^2(t)$. We cannot take the time-derivative of the quadratic variation like we can take the predictable quadratic variation's time-derivative. Second, by conditioning on $\gamma^2(t)$, I construct a closed-form nonparametric continuous-time representation for $p(t)$ in Section 2.4. This representation avoids any truncation. Todorov and Tauchen (2014) must truncate all of the jumps above a shrinking threshold in order to derive their results while using an ex-post measure. Third, as Section 2.9 shows, $\gamma^2(t)$ controls risk premia. This result is intuitive because risk-premia are ex-ante objects. As a final advantage, high-frequency data identifies both $\gamma^2(t)$ and γ_t^2. Section 2.5 shows this by constructing consistent estimators for them.

Static Jump Processes (Variance-Gamma Process)

The next section constructs the static model that my model reduces to when there are no dynamics. It will also be the integrator in the general case. I start with a simple jump process where the locations of the jumps are Poisson distributed, and the magnitudes are i.i.d. Gaussian variables and then take limits to construct the general case.

Define $N(t)$ as the process that determines when $p^J(t)$ jumps, i.e., $N(\tau)-N(\tau-) = 1$ if and only if $p(t)$ jumps at τ.

Definition 2.9. Location Process

$$N(t) \coloneqq \sum_{s \leq t} 1\left\{\left|p^J(t) - p^J(t-)\right| > 0\right\}$$

Let $\kappa(t) \coloneqq \left\{p^D(t) \,|\, N(t) \neq N(t-)\right\}$ be a process that controls the jump magnitudes. Note, $\kappa(t)$ is not a Wiener process because its variance does not depend on the length of the interval. It is just an ordered collection of $\mathcal{N}(0,1)$ random variables, one for each t. In this case, the jump part of the price process has the following relatively simple form:

$$p^J(t) = \sum_{s \leq t} \kappa(s)|N(s) - N(s-)|. \tag{2.6}$$

The variability in (2.6) comes from two places: the number of jumps and their magnitudes. Since we are in a time series context, the number of jumps and their

locations carry the same information. Hence, we can rewrite the jump volatility as follows using the law of iterated expectations:

$$\begin{aligned}\gamma_t^2 &= \lim_{\Delta\to 0}\mathbb{E}\left[\frac{|p_{t+\Delta}-p_t|^2}{\Delta}\middle|\mathcal{F}_t\right] \\ &= \lim_{\Delta\to 0}\mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{N(t+\Delta)-N(t)}\frac{\mathbb{V}\mathrm{ar}(\kappa_i(t))}{\Delta}\middle|\mathcal{F}_t, N(t+\Delta)-N(t)\right]\middle|\mathcal{F}_t\right].\end{aligned} \tag{2.7}$$

We then simplify this using (2.6).

$$\gamma_t^2 = \lim_{\Delta\to 0}\mathbb{E}\left[\frac{N(t+\Delta)-N(t)}{\Delta}\right]\mathbb{E}\left[\kappa(t)^2\right] = \frac{1}{\Delta}\Delta = 1.$$

To put (2.7) into words, the variance of the jump process is the intensity multiplied by the magnitude's variance. This characterization holds whenever the intensities and magnitudes are conditionally independent. It combines the variation from the jump locations and the jump magnitude into a single parameter. Changing the jump intensity or expected magnitude alters the variance of $p^J(t)$ in precisely the same way. This irrelevance is useful because the data do not identify the intensity and magnitude functions but do identify the volatility.

This lack of identification no longer affects our results if we take $\mathbb{E}[N(t)]\to\infty$. In taking this limit, we must model the distribution of $\kappa(t)$ properly so that $p(t)$ remains square-integrable.[8] In particular, only finitely many jumps can exceed any fixed $\epsilon>0$ in magnitude; otherwise, the price diverges. Consequently, we must shrink the size of the increments towards zero as we let $\mathbb{E}[N(t)]\to\infty$.

One common pure-jump process — the variance-gamma process — is an infinite-activity "compound Poisson process" with arbitrarily small Gaussian-distributed summands. My model for the jumps reduces to this if it does not have any dynamics. The option pricing literature often uses the variance-gamma process in its models. For example, European option prices are available in closed form, (Madan, Carr, and Chang 1998).

A gamma process — $\Gamma(t)$ — is a process with gamma distributed increments, and a variance-gamma process is a Wiener process time-changed (subordinated) by

[8] Just letting $p(t)$ be an ordered collection of $\mathcal{N}(0,1)$ variables does not work.

a gamma process. Equivalently, a gamma process is a pure-jump Lévy process where the jumps that lie in an interval $[x, x + \Delta x)$ are Poisson distributed with intensity measure $x^{-1}\exp(-x)$ for positive x and Δx small.

Definition 2.10 (Variance-Gamma Process)**.**

$$\text{Variance} - \text{Gamma}(t) \coloneqq W(\Gamma(t))$$

This paper exclusively uses the standard variance-gamma process, which is the variance-gamma process whose increments are mean zero with all the scale parameters equal to one.[9] The exponential distribution is a special case of the Gamma distribution. (Take a Gamma random variable and set all of its scale parameters equal to one.) If we consider a Wiener process time-changed by a gamma process with rate=1 exponentially-distributed increments, we get a standard variance-gamma process. I use the symbol $\mathcal{L}(t)$ to refer to the standard variance-gamma process because the increments of this process are Laplace random variables.

To understand why the increments are Laplace-distributed, consider the following characterization of a standard variance-gamma process. A Laplace distribution is as a Gaussian distribution with random variance, where the random variable is exponentially distributed. This equality gives the following characterization of the Laplace distribution:

$$z \sim \mathcal{L}(\text{mean} = 0, \text{variance} = 1) \iff z \sim \frac{\sigma}{\sqrt{2}}\mathcal{N}(0,1)\,,\ \sigma^2 \sim \exp(1). \tag{2.8}$$

This is the discrete analogue of Definition 2.10. The $\sqrt{2}$ in the expression is an adjustment to convert the standard deviation into a scale parameter. Each increment of a variance-gamma process has two sources of variation: the number of jumps, which "is" exponentially distributed, and the magnitudes, which are Gaussian distributed. This characterization is not quite accurate because exponential random variables are real-valued, not integer-valued. The number of jumps cannot be exponentially distributed.

However, a Poisson process is the process where the waiting time between jumps is exponentially distributed. This characterization is well-defined because the intervals'

[9] I introduce the standard variance-gamma process to facilitate exposition because it aggregates in ways that the general case does not.

lengths take positive real values, and it returns us to the initial discussion of a variance-gamma process as an infinite-intensity "compound Poisson" process.

Jump Process Representation Theorem

Itô Semimartingales

Recall the simplified Grigelionis form of the semimartingale (2.2):

$$p(t) = \int_0^t \mu(s)\, ds + \int_0^t \sigma(s)\, dW(s) + \int_0^t \int_X \delta(s,x)(n-\nu)(ds,dx). \tag{2.9}$$

We can use the variance-gamma processes and the jump volatility discussed above to simplify the representation for the jump part of the process. To do this, I introduce some empirically-innocuous assumptions that are not entirely standard in the literature. First, $p(t)$ must have infinite-activity jumps. In other words, we need at least one jump in every finite interval. This assumption implies two results. First, we do not need to track the probability that there are no jumps in a specific interval. Second, it identifies $\gamma^2(t)$. If we consider an interval without jumps, we obviously cannot estimate $\gamma^2(t)$ because we have no variation to identify it with.

Assumption Infinite-Activity Jumps. The $p(t)$ process has infinite-activity jumps.

Assumption Infinite-Activity Jumps sounds very restrictive at first and contradicts the compound Poisson assumption often used in the literature. However, it is rather innocuous for two reasons. First, the literature is essentially unanimous in arguing that jumps are quite common in the data as discussed in Section 2.1. Second, standard variance-gamma processes are limits of compound Poisson process. As long as we have a sufficient number of jumps, the representation will work well in practice. I discuss this further in Section 2.4.

The last assumption requires jump times to be unpredictable.

Assumption No Predictable Jumps. There does not exist any stopping times τ such that the event $p(\tau) \neq p(\tau-)$ is contained in the information set $\mathcal{F}_{\tau-}$.

Having laid out the assumptions, I state the main theorem. I later prove a more general proposition, Theorem 2.4. However, I have now described the environment sufficiently to make the result understandable. Stating the result now that will likely be used in practice make it easier to see where the paper is headed.

Theorem 2.3 (Locally Square-Integrable Itô Semimartingales as Integrals)**.** *Let* $p(t)$ *be an Itô semimartingale with full-support satisfying Assumptions Square-Integrable, Infinite-Activity Jumps, and No Predictable Jumps. Then we can represent* $p(t)$ *as*

$$p(t) = \int_0^t \mu(s)\, ds + \int_0^t \sigma(s)\, dW(s) + \frac{1}{\sqrt{2}} \int_0^t \gamma(s)\, d\mathcal{L}(s).$$

Proof. We can replace the jump part of (2.2) with an integral with respect to the standard-variance gamma process where the root jump volatility is the integrator using Corollary 2.1. □

This representation replaces the function $\delta(\tau, \cdot)$, with a single scalar $\gamma^2(\tau)$ for each time τ. In addition the integrator is switched from a compensated Poisson random measure, $(n - \nu)$, to a standard variance-gamma process, $\mathcal{L}(t)$.

Time-Change Representation

The proof of Theorem 2.3 relies on Corollary 2.1, which I have not yet proven. In practice, Theorem 2.4 is the fundamental result. The other results, such as Corollary 2.1, are straightforward implications of it. I prove this theorem now. Theorem 2.4 is a time-change representation for jump processes and is closely related to the time-change representations in the diffusion case. Consequently, to ease comprehension let us recall those results.

The validity of the standard diffusion representation for general continuous martingales is implied by the Dambis-Dubins-Schwarz theorem, which shows that any continuous martingale time-changed by its predictable quadratic variation is a Wiener process. To put it in mathematical notation, Dambis (1965) and Dubins and Schwarz (1965) says that

$$p^D(t) \overset{\mathcal{L}}{=} W\left(\left\langle p^D \right\rangle(t)\right), \tag{2.10}$$

where the equals sign with an $\mathcal{L}$ above it refers to equality in law. The right-hand side of (2.10) evaluates the Wiener process at the random-clock $\langle p^D(t) \rangle$.

The crucial difference between the jump part and the continuous part of a semimartingale is that the variation in the continuous part comes from variation in magnitudes, while the jump part has two sources of variation: the magnitudes and the locations. Intuitively, the Dambis-Dubins-Schwarz theorem separates the variation

in any continuous martingale into a predictable part (the volatility) and i.i.d. innovations. By doing this, the martingale becomes a sum of appropriately scaled independent random variables. In other words, it is a "central limit theorem."[10] In fact, one method to prove standard central limit theorems is deriving them from this result.

In the jump case, though, the dynamics are more complicated. Not only do the magnitudes vary, but the locations also vary. When we take the infill asymptotics, both of these sources of variation are still present. In other words, a jump martingale is a sum of a random number of random summands. If the number of summands is geometrically-distributed, various geometric-stable central limit theorems tell us how the sum behaves as the expected number of summands approaches infinity, (Mittnik and Svetlozar 1993; Kozubowski and Svetlozar 1994).

We can generalize the Itô semimartingale assumption in Theorem 2.3 by only requiring that $p^J(t)$ is an integral with respect to a Poisson random measure. In particular, the $p(t)$'s characteristics do not need time-derivatives.

Theorem 2.4 (Time-Changing Jump Martingales)**.** *Let $p^J(t)$ be a purely discontinuous, martingale with full support satisfying Assumptions Square-Integrable, Infinite-Activity Jumps, and No Predictable Jumps that can be represented as $H*(n-\nu)$ where $H(t)$ is a predictable process, n a Poisson random measure, and ν its predictable compensator with Lebesgue base Levy measure.*

Then $p^J(t)$ time-changed by its predictable quadratic variation is a standard variance-gamma process. In other words, $p^J(t) \stackrel{\mathcal{L}}{=} \mathcal{L}\left(\left\langle p^J \right\rangle(t)\right)$.[11]

The proof of this theorem is in Section 2.A. I present the intuition here. The first result we must establish is that the jump locations and magnitudes are conditionally independent. Thankfully, the Poisson random measure representation implies that the location and magnitude risk are independent given $\mathcal{F}_{t-}$.

I condition on the number of jumps and show that the magnitudes are a continuous process in that space. Thus, I can apply the Dambis-Dubins-Schwarz theorem

[10] Technically, this result is a law of large numbers, not a central limit theorem because the convergence here is almost sure instead of in law.

[11] Note, the equality here only holds in law unlike in the Dambis-Dubins-Schwarz theorem, where it holds almost surely.

there, which results in a time-changed Wiener process. The standard representations further imply that each hitting times for each open set of magnitudes is a compound Poisson process. We can time-change these locations by their predictable quadratic variation, getting a standard Poisson process. Since the times between jumps for a Poisson process are exponential random variables, by keeping careful track of how the exponential time-changes aggregate, we get that the time-change coming from the locations is a standard Gamma process. The predictable quadratic variation of $p(t)$ is the composition of quadratic variation arising from each of two time-changes. Therefore, the original process is a time-changed standard variance-gamma process.

Time-changed results are not particularly intuitive, and so we would like an integral representation as well. So we assume that $p(t)$'s characteristics are absolutely continuous.

Corollary 2.1 (Jumps Processes as Integrals)**.** *Let $p^J(t)$ be a Itô semimartingale with full support satisfying Assumptions Square-Integrable, Infinite-Activity Jumps, and No Predictable Jumps. Then $p^J(t) = \frac{1}{\sqrt{2}} \int_0^t \gamma(s)\, d\mathcal{L}(s)$, where $\mathcal{L}$ is a standard variance-gamma process.*

Corollary 2.1 is analogous to how we represent continuous martingales as stochastic volatility diffusion as shown in Dambis-Dubins-Schwarz theorem by assuming the relevant characteristics are absolutely continuous.

Processes with Finite-Activity Jumps

The most controversial assumption I make is Assumption Infinite-Activity Jumps. Various authors have claimed that we have a large, but finite, number of jumps in each period. The natural question is what happens to the distributional result in this case? In any given interval, the price process is a point mass at zero if it does not jump. If the price does jump, we can represent it as done above. In other words, the ex-ante distribution over each interval is a mixture of a point mass at zero and a Laplace distribution where the mixing weights are the probability of the jump in that interval.

Corollary 2.2 (Time-Changing Finite-Activity Jump Martingales)**.** *Let $p^J(t)$ be a purely discontinuous martingale with full support satisfying Assumptions Square-*

Integrable and *No Predictable Jumps* *that can be represented as* $H * (n - \nu)$ *where* $H(t)$ *is a predictable process,* n *a Poisson random measure, and* ν *its predictable compensator with Lebesgue base Levy measure.*

Let $\langle p^J \,|\, n(t) \rangle$ *be the predictable quadratic variation of* p^J *where additionally we condition on all the jumps occurring up to and including at time* t*. Then* $p^J(t)$ *time-changed by* $\langle p^J \,|\, n(t) \rangle$ *is a mixture of the* 0 *process —* δ_0 *— and the standard variance-gamma process where the mixing weights are the intensity of the jump process.*

This theorem uses the same jump locations as the original process (they are controlled by $n(t)$) but treats the jump magnitudes as a scale Gaussian mixture. The scale process may be correlated with the Gaussian part. We cannot integrate out the jump locations here as we did above because the jump locations matter. We have to keep track of the probability of no jump. Also, unlike above the time-change here is not identified, and we have two states that govern the process, not just one.

The main benefit of Corollary 2.2 is that it implies that Theorem 2.4 is the limiting case of a finite-activity process as the intensity approaches infinity. Consequently, the representation in Theorem 2.3 approximates the true DGP well if the intensity is relatively large.

Also, standard Poisson processes satisfy the assumptions of Corollary 2.2. In that case, the corollary does not change the representation. We model the magnitudes there as scale Gaussian mixtures, but the scale is just zero, and so they degenerate into point masses.

Deriving the Realized Density

Having derived the continuous-time representation, we can solve the time-aggregation problem and derive the realized density. Barndorff-Nielsen and Shephard (2002) and Andersen, Bollerslev, Diebold, and Labys (2003) concurrently derived the realized density when prices have continuous paths, although, they did use that name. They showed that if volatility and prices are correlated, σ_t^2 is a sufficient statistic for the dynamics under some technical conditions and that conditional on σ_t^2, the return density is Gaussian.

This conditional Gaussianity separates the daily return distribution into a well-behaved volatility component and a Gaussian noise component. To relate it to the previous discussion, we have the following decomposition for $h\left(r_t \,|\, \mathcal{F}_{t-1}\right)$ if the price is a martingale:

$$f\left(r_t \,\middle|\, \sigma_t^2\right) = f\left|_{x_t=\left[\int_{t-1}^t \sigma^2(s)\,ds\right]} = \mathcal{N}\left(0, \int_{t-1}^t \sigma^2(s)\,ds\right). \tag{2.11}$$

I now discuss the realized density in the general case with jumps. The return has two parts: $dp(t) = \sigma(t)\,dW(t) + \int_X \delta(t,x)(n-\nu)(dx,dt)$. Conditional on the values of $\sigma^2(t)$ and $\delta(t,\cdot)$, the jumps and diffusion parts are independent. Consequently, returns are the sum of two conditionally independent components. Densities of sums of independent components are convolutions of the summands' densities. We know, as discussed above, that the diffusion part is a Gaussian density whose variance equals the integrated diffusion volatility. Hence, we only need to develop a parametric expression for jump part.

Let $\mathcal{L}(0,x)$ refer to the Laplace density with mean zero and variance x and recall that $*$ is the standard convolution symbol. Then we have the following discrete-time representation.

Theorem 2.5 (Realized Density Representation)**.** *Let $p(t)$ be an Itô semimartingale with full support satisfying Assumptions Square-Integrable, Infinite-Activity Jumps, and No Predictable Jumps. Let $\sigma^2(t)$ and $\gamma^2(t)$ be semimartingales whose martingale components are independent of the martingale components of $p(t)$. Then*

$$RD_t = \mathcal{N}\left(\int_{t-1}^t \mu(s)\,ds, \int_{t-1}^t \sigma^2(s)\,ds\right) * \mathcal{L}\left(0, \int_{t-1}^t \gamma^2(s)\,ds\right), \tag{2.12}$$

and the predictive density is

$$h\left(r_t \,|\, \mathcal{F}_{t-1}\right) = \int_{\mu_t,\sigma_t^2,\gamma_t^2} RD_t(\mu_t,\sigma_t^2,\gamma_t^2)\,dG\left(\mu_t,\sigma_t^2,\gamma_t^2 \,\middle|\, \mathcal{F}_{t-1}\right). \tag{2.13}$$

The intuition behind Theorem 2.5 is the following. If $\gamma^2(t)$ was constant, we could pull it out of the integral without affecting the distribution: $\int_{t-1}^t \gamma^2(t-1)\,d\mathcal{L}(s) \overset{\mathcal{L}}{=} \frac{\gamma_{t-1}^2}{\sqrt{2}} \int_{t-1}^t d\mathcal{L}(s)$. Since increments of the standard variance-gamma process are Laplace distributed, the second component is distributed $\mathcal{L}(0,1)$. Consequently, conditionally

on γ_{t-1}^2, we have a Laplace distribution with the specified variance. The $\sqrt{2}$ term arises because the scale of a Laplace distribution is the square root of one-half the variance. We can replace the constant assumption on the volatilities with the independence conditions between the martingale components to recover the general case.

Integrating RD_t out using its distribution G recovers (2.13) from (2.12). In practice, we likely want to model G directly. This model has the same form as the various stochastic volatility / GARCH type models in the diffusion case. Many of those models can be extended straightforwardly to the jump-diffusion case because the stylized features of the γ_t^2 and σ_t^2 are quite similar, as Section 2.8 shows.

The primary assumption that Theorem 2.5 adds is the independence between the martingale components of the various terms. We need this assumption to time-aggregate because we need the marginal and conditional distributions given the volatilities of $p(t)$ to coincide. Importantly, this assumption restricts the leverage effect but does not assume away all dependence. The volatilities and drift can be arbitrarily related.

Since the jump part is purely discontinuous, it is orthogonal to the diffusion part. In other words, if we condition on the one process, the other process is still a martingale. Since we are integrating with respect to Brownian and Laplace motions, the martingale property is sufficient to imply that all the integrators are independent. To time-aggregate, we must separate the volatilities from the martingale components. Consequently, we must assume that the volatilities' martingale components are independent of the martingale components of $p(t)$.

The predictable relationship between the drift and volatilities is entirely unrestricted as is the relationship between the volatilities themselves. As long as it takes a positive amount of time for feedback from the volatilities to affect the level of the prices or vice-versa, this assumption is satisfied. Besides, the observed correlation between the martingale parts is close to zero at high frequency as noticed by Aït-Sahalia, Fan, and Li (2013), who call it "the leverage effect puzzle."[12]

[12]There is some evidence that this is an artifact of the estimation procedure, and so I leave to future work the optimal way of bringing it into my framework. One way to do this is by keeping track of this correlation and using tools similar to those developed by the above paper and by Neuberger (2012) and Kalnina and Xiu (2017) and making Gaussian and Laplacian conditioning arguments.

2.5 Estimation

This section constructs estimators for $\sigma^2(t)$ and $\gamma^2(t)$ and their daily analogs. As is standard, the data do not identify the drift, $\mu(t)$, and so we cannot estimate it. The proposed estimator for $\sigma^2(t)$ is adapted from Jacod and Rosenbaum (2013). I show that their estimator is still valid under my slightly more general assumptions. The estimator for $\gamma^2(t)$ is completely new. In particular, it develops a consistent estimator for $\gamma^2(\tau)$ for any fixed τ. [13] Also, this is the first consistent estimator for any instantaneous measure of jump dynamics. Note, this implies that high-frequency data nonparametrically identifies instantaneous jump dynamics are nonparametrically identified, which was not heretofore known.

Assumptions

To start, I fix some notation and state some assumptions. The way that the instantaneous volatility estimators work is by taking an appropriately defined average over an increasing number of increments over a shrinking interval. In other words, for a given index — n, we have a triangular array of increments. To make the notation even more complicated, we have both a true D.G.P. with time-varying volatility and an approximate D.G.P., whose volatility is locally constant.

This setup implies we must keep track of both triangular arrays as we take limits. I adopt the notation used in Jacod and Protter (2012) for the most part. Specifically, I use $\Delta_i^n p$ to refer to a increment i in process $p(t)$ of length Δ^n, and I take limits with respect to n, that is $\{\Delta_i^n p\}$ is a triangular array of increments of $p(t)$. The assumptions used are very similar to the standard ones used in the literature. When possible, I simplify them using the representation theory developed above.

Assumption HL. 1. $\mu(t)$ is locally bounded.

2. $\sigma(t)$ is càdlàg (or càglàd).

3. $\gamma(t)$ is càdlàg (or càglàd).

[13] In general, much of the theory that I develop can likely be extended to stopping times, but I leave that for future work.

Assumption HL is essentially Jacod and Protter's (2012) Assumption H. The assumption on the jumps is slightly more general and more straightforward. I also slightly modify the literature's Assumption SH. (Here ω indexes the underlying probability space Ω.)

Assumption SHL. We have Assumption HL and there is a constant A such that the following hold for all t and all ω:

$$\|b(t,\omega)\| < A, \ \|\sigma(t,\omega)\| < A, \ \|\gamma(t,\omega)\| < A.$$

These two assumptions are closely related. Assumption HL is the local version of Assumption SHL. Assumption HL only restricts the local behavior of the function, while Assumption SHL make the equivalent conditions globally. Since convergence in the Skorokhod topology only depends upon local behavior, if we prove consistency the first assumption, the estimator automatically converges under the second assumption as well. This result implies that in the proofs below we can assume SHL without loss of generality. To make this statement explicit, we have the following lemma whose proof is in the appendix. The arrow with $\mathcal{L}$-s above it refers to stable convergence in law, which is the type of convergence necessary for confidence intervals to be valid in this setting.

Lemma 2.6 (HL implies SHL)**.** *If an Itô semimartingale $p(t)^n \xrightarrow{\mathcal{L}\text{-s}} p(t)$ under Assumption SHL, then $p(t)^n \xrightarrow{\mathcal{L}\text{-s}} p(t)$ under Assumption HL, and the equivalent statement holds for convergence in probability.*

To reduce notation, I adopt the convention (2.14) from the literature to make processes well-defined over the entire line, not just where we estimate them.

$$i \in \mathbb{Z}, i \leq 0 \implies \Delta_i^n p = 0. \tag{2.14}$$

Its sets the processes equal to zero outside of the relevant window.

To estimate the instantaneous volatility, we must approximate $\sigma^2(\tau-)$ and $\gamma^2(\tau-)$. Thus, we must choose a sequence of i_n, k_n, and Δ_{i_n}, so that we are averaging the variation over smaller and smaller intervals to the left of τ. The k_n term will refer to the number of terms we are averaging over. Consider the following interval:

$$I(i,n) \coloneqq [(i-k_n-1)\Delta^n, (i-1)\Delta^n]. \tag{2.15}$$

Let $\Delta_{i_n}^n p$ denote the change in p in $I(i,n)$. If we choose a sequence $i \to \tau$, the interval approaches τ from the left. Also, as $p(t)$ is one-dimensional, the driving Wiener and variance-gamma processes can be assumed to be one-dimensional without loss of generality.

Instantaneous Volatility Estimators

Having laid out the framework, I state the estimators themselves. The intuition behind their convergence is that we are averaging the volatilities over shrinking intervals that approach τ from the left. As long as the number of increments being averaging over is increasing faster than the length of the interval is shrinking, we precisely estimate the volatility. Since we are estimating the process from the left, we are approximating the value before τ, i.e., we are estimating $\gamma^2(\tau-)$.

I first derive an estimator for $\sigma^2(\tau-)$. There are few such estimators in the literature that do this, including Mancini (2001) and Jacod and Rosenbaum (2013). They do this by that noting that estimating the integrated diffusion volatility — $\langle p^D \rangle(t)$ — is straightforward. We can use the integrated volatilities' sample analogue to estimate $p^D(t)$. Consequently, we can use time-derivative of $\widehat{\langle p^D \rangle}(t)$ to estimate the time-derivative of $\langle p^D \rangle$, $\sigma^2(\tau)$.

The main difficulty in practice is separating the jump and diffusion variation. Following Mancini (2001), I truncate away the large increments, where *large* is defined in terms of an asymptotic rate. Asymptotically, this eliminates large jumps, and the small jumps do not affect the asymptotic value.

Theorem 2.7 (Estimating the Instantaneous Diffusion Volatility)**.** *Let $p(t)$ be an Itô semimartingale satisfying Assumptions HL, Infinite-Activity Jumps, and Square-Integrable. Let k_n, Δ^n satisfy $k_n \to \infty$ and $k_n\sqrt{\Delta_n} \to 0$, and let $0 < \tau < \infty$ be a deterministic time. Define $i_n = i - k_n - 1$. Let $c_1(\Delta^n)^{1/4} < v_1^n < c_2\sqrt{\Delta^n}$ for some constants c_1, c_2 and $v_2^n \to 1$. Then*

$$\widehat{\sigma}_{i_n}^2(k_n, \tau-, p) \coloneqq \frac{1}{k_n\Delta^n}\sum_{m=0}^{k_n-1} v_2^n \left|\Delta_{i_n}^n p\right|^2 1\{\left|\Delta_{i_n}^n p\right| \leq v_1^n\} \xrightarrow{\mathbb{P}} \sigma^2(\tau-).$$

One might think we could use a similar estimation strategy to estimate $\gamma^2(t)$, i.e., form an estimator of $\langle p^J \rangle(t)$ by truncating away the small increments and take the

time derivative of the resulting object. In fact, Jacod and Protter (2012, 256) show that this estimator converges to zero in their proof of the validity of their estimator for $\sigma^2(t)$. Intuitively, by considering a specific time τ, we implicitly condition on τ. Doing this reduces the variation in the locations, and shrinking the window eliminates variation from large jumps. If we also truncate away variation arising from the small jumps, we have no variation left to identify the jump volatility.

Over a fixed interval, the quadratic variation of jump processes and diffusive processes are of the same asymptotic order as we shrink $\Delta_{i_n}^n$, (Jacod, Podolskij, and Vetter 2010). If we consider shrinking intervals, this is no longer the case. Instead, it is the absolute value of the stochastic volatility Laplace and diffusive processes that have similar asymptotic properties. [14] The absolute value of a standard variance-gamma process, $|\mathcal{L}|(t)$, is a well-behaved object, just like the absolute value of a Wiener process, $|W|(t)$, and they vanish at the same asymptotic rate: $\sqrt{\Delta}$.[15] Consequently, the $\lim_{\Delta^n \to 0} \left|\Delta_{i_n}^n p(t)\right|$ contains both $\gamma^2(\tau)$ and $\sigma^2(\tau)$.

Theorem 2.8 (Estimating the Instantaneous Absolute Volatility)**.** *Let $p(t)$ be an Itô semimartingale satisfying Assumptions HL, Infinite-Activity Jumps, and Square-Integrable. Let k_n, Δ^n satisfy $k_n \to \infty$ and $k_n\sqrt{\Delta^n} \to 0$, and let $0 < \tau < \infty$ be a deterministic time. Define $i_n \coloneqq i - k_n - 1$.*

Then the following holds, where $\operatorname{erfcx} \coloneqq \frac{2\exp(x^2)}{\sqrt{\pi}} \int_x^\infty \exp(-s^2)\, ds$*:*[16]

$$\frac{1}{k_n\sqrt{\Delta^n}} \sum_{m=0}^{k_n-1} |\Delta_{i_n+m}^n p| \xrightarrow{\mathbb{P}} \mathbb{E}|\mathcal{N}(0,1)|\sigma(\tau-) + \frac{\gamma(\tau-)}{\sqrt{2}} \operatorname{erfcx}\left(\frac{\sigma(\tau-)}{\gamma(\tau-)}\right).$$

As long as $\sigma^2(t)$ and $\gamma^2(t)$ are locally constant around τ, we can use the implied parametric form to compute the limiting value as a function of $\sigma^2(\tau)$ and $\gamma^2(\tau)$. The expression on the right of the equation in Theorem 2.9 is the mean of the convolution of $|\mathcal{N}\left(0, \sigma_{\tau-}^2\right)|$ and $|\mathcal{L}\left(0, \gamma_{\tau-}^2\right)|$.

We combine this convolution and $\sigma^2(\tau)$ to estimate $\gamma^2(\tau)$. To do this, we must weight the difference between the absolute population moment as a function of $\gamma(\tau)$

[14] It is an interesting open question to what other jump processes this result extends.

[15] As an aside, neither of the processes are martingales. They are semimartingales.

[16] This function, erfcx, is the scaled complementary error function. It is a reparameterization of Mill's ratio. Most scientific programming suites provide efficient, numerically-stable implementations.

and the absolute sample moment. In general, any convex weighting function of the differences will work. I use the absolute value of the difference between the two values because it works well in simulations.

Theorem 2.9 (Estimating the Instantaneous Jump Volatility)**.** *Let $p(t)$ be an Itô semimartingale satisfying Assumptions HL, Infinite-Activity Jumps, and Square-Integrable. Let k_n, Δ^n satisfy $k_n \to \infty$ and $k_n\sqrt{\Delta^n} \to 0$, and let $0 < \tau < \infty$ be a deterministic time. Define $i_n = i - k_n - 1$. Let $\widehat{\sigma}_n(\tau-)$ converge in probability to $\sigma(\tau-)$. Let $\gamma(\tau) > 0$ and g be strictly-increasing, convex, and continuous, then the following holds:*

$$\widehat{\gamma}(k_n, \tau-, p) \coloneqq \underset{\gamma}{\operatorname{argmin}}\, g\left(\left|\frac{1}{k_n\sqrt{\Delta^n}}\sum_{m=0}^{k_n-1}\left|\Delta^n_{i_n+m}p\right| - \mathbb{E}|\mathcal{N}(0,1)|\widehat{\sigma}_n(\tau-) - \frac{\gamma\,\operatorname{erfcx}\left(\frac{\widehat{\sigma}_n(\tau-)}{\gamma}\right)}{2}\right|\right)$$
$$\xrightarrow{\mathbb{P}} \gamma(\tau-).$$

Implementation

We now have estimators for the instantaneous jump and integrated volatilities. The difficult part is estimating the instantaneous volatilities. The integrated volatilities are their averages. In practice, two issues affect the analysis. First, we must remove market microstructure noise. To do this, I adopt the pre-averaging approach argued for in Podolskij and Vetter (2009, Eqn. (3.9)). Define the function:

$$g(x) \coloneqq \left(1 - (2x-1)^2\right)\mathbf{1}\{x >= 0\}\mathbf{1}\{x <= 1\}. \tag{2.16}$$

The pre-averaged data is the rolling average of the true data:

$$\bar{p}_{i_n} \coloneqq \frac{1}{\kappa_n\sqrt{\int_0^1 g^2(s)\,ds}}\sum_{m=1}^{\kappa_n-1} g\left(\frac{m-1}{\kappa_n}\right)\Delta^n_{i_n+m}p. \tag{2.17}$$

The g function corrects for the error introduced by the pre-averaging.

If $\kappa_n \propto 1/\sqrt{\Delta^n}$, we likely achieve the optimal rate in the presence of noise, but the noise leads to an asymptotic bias in most cases, (Jacod, Podolskij, and Vetter 2010). To avoid this, I set $\kappa_n = \lfloor\frac{\theta}{(\Delta^n)^{0.55}}\rfloor$. This rate is useful because we can apply the estimators directly to the pre-averaged data, and it is not obvious exactly what

bias exists when estimating the instantaneous absolute variation.[17] I set $\theta = 0.5$, which is a values recommend by Hautsch and Podolskij (2013), and works well in my simulations as well.

I apply Theorem 2.7 to estimate $\sigma^2(\tau-)$. To do this, we must choose v_2^n to converge to 1, I let $v_2^n = 1$. More importantly, I must choose the truncation threshold v_1^n. We need v_1^n to asymptotically upper bound the absolute diffusion part. In the literature, papers usually set $v_1^n = c\widetilde{\sigma}(\tau-)\Delta_n^{0.49}$ where $\widetilde{\sigma}(\tau-)$ is a preliminary estimator for σ and c is a number of standard deviations chosen by the econometrician.

The tails of the Laplace and Gaussian random variables both decline rapidly. The Gaussian density's tails are proportional to $\exp(-x^2/2)$, while the Laplace density's tails are proportional to $\exp(-x/\sqrt{2})$. Distinguishing these two is quite difficult in practice. Setting $v_1^n \propto \Delta^{.49}$, does not work particularly well in this scenario as I show in Section 2.6. On the other hand, the law of the iterated logarithm tightly bounds the deviations of a Gaussian variable, and so I use $v_1^n = \sqrt{2}\widetilde{\sigma}(\tau-)\sqrt{\Delta_n}\sqrt{\log(\log(1/\Delta_n))}$.

To form a preliminary estimator, I start with the 1.25 times bipower variation and then iterate until convergence. We need to start by overestimating the volatility to avoid incorrectly setting $\widehat{\sigma}(\tau-) = 0$ since that would truncate away all the increments. It is worth noting that this volatility estimator relies on neither $\gamma^2(\tau)$ nor the qualitative properties of the Laplace representation.

In addition, we must choose k_n, where $1/k_n$ controls the length of the interval over which the volatilities are treated as approximately constant. Theory tells us that $k_n \to \infty$ and $k_n\sqrt{\Delta^n} \to 0$, I choose $k_n = \bar{\kappa} + (\Delta^n)^{1/4}$ with $\bar{\kappa} = 1000$ because that seems to work well in the simulations with market microstructure noise.

Now that we can estimate $\sigma^2(\tau-)$, we need an estimator for the local absolute value. I plug the pre-averaged data into Theorem 2.9. It is worth noting that the theory I develop is for the no-noise case; the particular implementation likely is not affected by the noise, but that has not been proven. An interesting extension for future work would be to extend these results to cover the noise case as well and to figure out the various biases arising there.

[17] The transformation creating $\bar{p}_{i_n}$ does not affect the volatilities but does affect the mean.

Integrated Volatilities

We want to estimate discrete increments of the volatilities. To do this, we use the obvious procedure and average the instantaneous estimators each day. The diffusion estimator defined this way coincides with standard diffusion estimators in the literature up to edge effects.

Theorem 2.10 (Consistentcy of the Integrated Estimators)**.** *Let $p(t)$ be Itô semimartingale satisfying Assumptions HL, Infinite-Activity Jumps, and Square-Integrable. Let k_n, Δ^n satisfy $k_n \to \infty$ and $k_n\sqrt{\Delta_n} \to 0$. Define $i_n = i - k_n - 1$. Then*

$$\widehat{\sigma}_t^2 := \frac{1}{\#t_n \in [t-1,t]} \sum_{t-1<t_n\leq t} \widehat{\sigma}^2(k_n, t_n, p) \xrightarrow{\mathbb{P}} \int_{t-1}^{t} \sigma^2(s)\, ds, \tag{2.18}$$

and

$$\widehat{\gamma}_t^2 := \frac{1}{\#t_n \in [t-1,t]} \sum_{t-1<t_n\leq t} \widehat{\gamma}^2(k_n, t_n, p) \xrightarrow{\mathbb{P}} \int_{t-1}^{t} \gamma^2(s)\, ds. \tag{2.19}$$

Proof. I am averaging estimates of $\sigma^2(t)$ and $\gamma^2(t)$. Averages of consistent estimators are consistent by the law of iterated expectations, Jensen's inequality applied to the square, and Chebyshev's theorem. □

Implementing the discrete volatility estimators is straightforward; we can take daily averages of the instantaneous volatilities. To estimate the realized density, plug the daily estimates into (2.12). Since this function is uniformly continuous given a lower bound on the volatilities, the resulting estimators should work well.

2.6 Simulations

One of my representation's key advantages of is t it can be simulated from easily whenever we can simulate the instantaneous volatilities. Perhaps the most commonly used model for the diffusion volatility is the Cox-Ingersoll-Ross (CIR) process. (A diffusion model whose volatility follows a CIR process is the Heston model.) One nice feature of this model is that the volatility itself has volatility, but we only need to simulate one process. The qualitative features of the jump and diffusion volatilities are quite similar, and so I adopt this model for the jump volatility. Once we have

the volatilities, we can simulate the price as the sum of the diffusion and jump parts directly.

Simulation Data Generating Process

The Cox-Ingersoll-Ross (CIR) process, also known as the square-root process, has the following form:

$$dx(t) = \kappa(\theta - (x(t)) + \omega\sqrt{x(t)}\, dW(t), \tag{2.20}$$

where θ is the asymptotic mean, κ is the mean-reversion rate, and ω is a scale parameter.

I simulate a CIR process for both $\gamma^2(t)$ and $\sigma^2(t)$ using the full-truncation scheme of Lord, Koekkoek, and Van Dijk (2010). The parameters are given in Table 2.1. Note, the asymptotic standard deviation for a CIR process equals $\frac{2\theta\omega^2}{\kappa}$. I chose the specific parameter values displayed below to match the discrete-time dynamics of the data.

Table 2.1: Volatility Parameters

Parameter θ	κ	ω	$\frac{2\theta\omega^2}{\kappa}$
$\sigma^2(t)$ 5.00×10^{-5}	1	2.10×10^{-3}	4.60×10^{-4}
$\gamma^2(t)$ 5.00×10^{-5}	1	2.10×10^{-3}	4.60×10^{-4}

Once I obtain $\sigma^2(t)$ and $\gamma^2(t)$, I plug them into the following continuous-time DGP:

$$dp(t) = \sigma(t)\, dW(t) + \frac{\gamma(t)}{\sqrt{2}}\, d\mathcal{L}(t). \tag{2.21}$$

This gives me a sequence of prices, which I use to estimate the volatilities.

Simulation Results

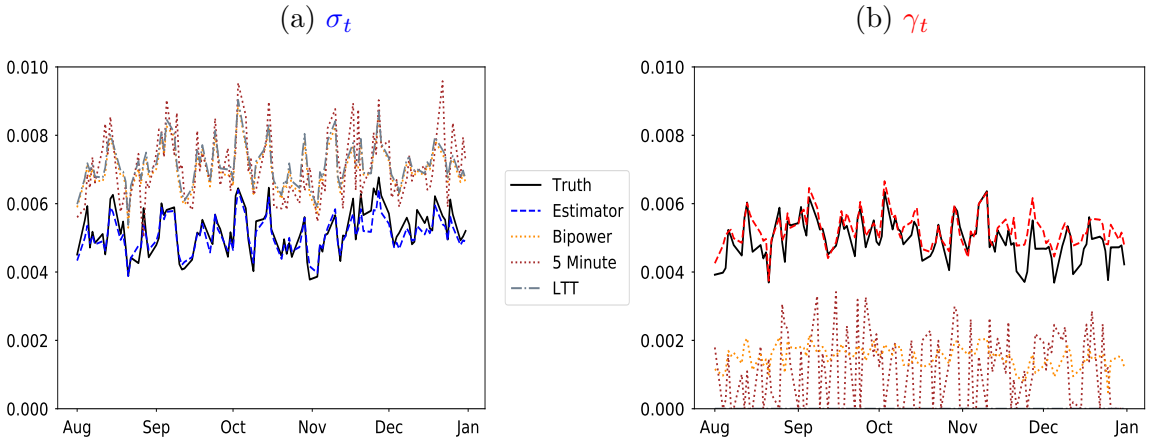
This section focuses on the daily volatility results below as they are sufficient statistics for all of the daily objects, which the applications study.[18] This section also reports the truncation-based estimator used by Li, Todorov, and Tauchen (2017), (LTT), the

[18] Section 2.E reports the report continuous-time results.

bipower estimator of Barndorff-Nielsen and Shephard (2004) and Podolskij and Vetter (2009), (Bipower), and bipower estimators computed on 5 minute data (5 Minute) to provide a point of comparison. In the jump case, these estimators do not converge to γ_t^2 but rather to the jump part of the quadratic variation. However, since γ_t^2 is the predictable quadratic variation, these estimators should still be asymptotically unbiased for γ_t^2.

I first estimate the model using the estimation procedure described in Section 2.5 without the microstructure correction described there. Figure 2.2 reports the results when I sample at the one-second frequency. Some of the jump variation estimators are not easy to see on the plot because I truncated them to zero. In the tables, I report averages over 250 days. It is worth noting that the daily estimates are independent, i.e., I do not smooth across days.

Figure 2.2: Simulation Results without Microstructure Noise



As can be seen in Figure 2.2, the estimators in the literature for σ_t^2 are badly biased upwards in finite-samples when the jump activity is high. This bias even holds in simulations without market microstructure noise at the one-second frequency ($\approx$ 24000 observations per day) This bias for σ_t^2 causes the literature's estimators for γ_t^2 to be severely biased as well.[19] Comapring the jump variation estimators is

[19]This bias likely explains why I find significantly higher jump variation than Christensen, Oomen, and Podolskij (2014) do, which is the only other paper to use pre-averaging and ultra high-frequency returns to measure jump variation. Because they use bipower variation to measure the jump proportion, their estimators for the jump proportion are likely highly-biased downwards.

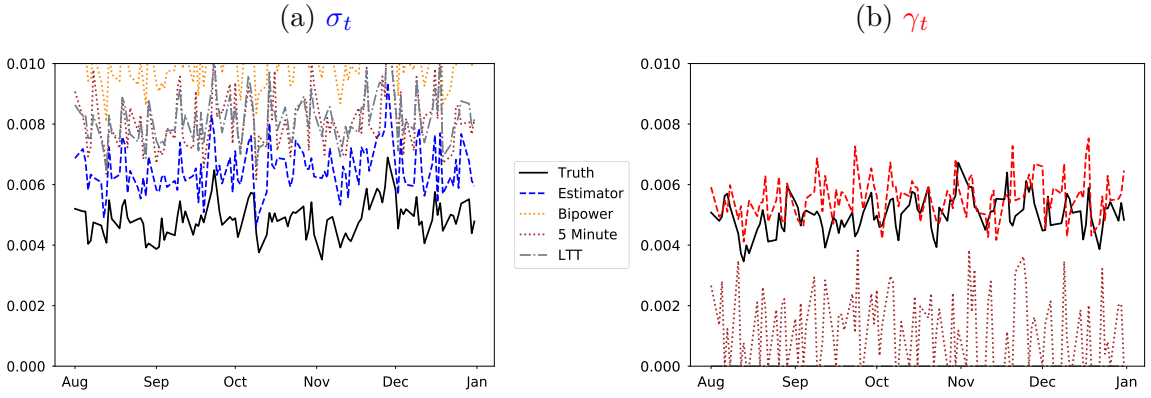
somewhat suspect because they are estimating different objects. However, since the predictable quadratic expectation is the instanteonous expectation of the quadratic variation, their estimators should be unbiased for γ_t^2.

Table 2.2: Relative Simulation Error without Microstructure

Obs. per Min.	$\frac{\mathbb{E}[(\widehat{\sigma_t}-\sigma_t)^2]}{\mathbb{E}[\sigma_t]}$				$\frac{\mathbb{E}[(\widehat{\gamma_t}-\gamma_t)^2]}{\mathbb{E}[\gamma_t]}$			
	Bipower	LTT	5 Minute	Proposed	Bipower	LTT	5 Minute	Proposed
≈ 2	0.37	0.40	0.40	0.46	0.72	1.01	0.80	0.72
≈ 12	0.38	0.40	0.42	0.16	0.70	1.01	0.83	0.21
≈ 60	0.40	0.43	0.45	0.05	0.68	1.01	0.87	0.07
≈ 180	0.39	0.41	0.43	0.07	0.69	1.01	0.85	0.07

The proposed estimators, however, perform quite well at this frequency. Table 2.2 reports the average root mean square errors of a year's worth of various estimators. As can be seen from this table, the proposed estimators outperform the other estimators in the literature by approximately an order of magnitude in this simulation.

Figure 2.3: Simulation Results with Microstructure Noise



The data have substantial market microstructure noise. To mimic its effect, I follow Christensen, Oomen, and Podolskij (2014). They assume we observe $r_{i_n} + u_{i_n}$, and u_{i_n} follows

$$u_{i_n} = \beta u_{i_n-1} + \epsilon_{i_n}, \quad \epsilon_{i_n} \overset{i.i.d.}{\sim} \mathcal{N}(0, \omega^2(1-\beta^2)). \tag{2.22}$$

I set $\omega^2 = 1.00 \times 10^{-10}$ because that is the value obtained from the data using the jump robust noise variance bipower-type estimator of Oomen (2006): $\frac{1}{T}\sum_{t=1}^{T-1}\frac{\Delta^n}{2}\sum_{t-1<i_n,i_n-1<t}\left|\Delta^n_{i_n}p\right|\left|\Delta^n_{i_n-1}p\right|$. I set $\beta = 0.77$, which is the value used in Christensen, Oomen, and Podolskij (2014). They set it to match the trade sign of the S&P 500 futures contract on the day of the 2010 Flash Crash.

I now add the market-microstructure correction described in Section 2.5. I also set $\theta = 0.5$ (the constant for the pre-averaging correction) and $\bar{\kappa} = 1000$ (the constant for the instantaneous estimator), which are the values used in the actual estimation. I chose these values because they appeared to work well in the simulated data. As we can see in Figure 2.3, the estimators are slightly biased upwards in this scenario, especially the estimators for σ_t^2.

Even though they are slightly biased upwards, the proposed estimators perform reasonably well in practice. This claim does not hold for the other estimators in the literature. In Table 2.2, I report the mean-square error of the previous estimates averaged over a year's worth of simulations. Here I have approximately 1/2 the average error in estimating σ_t^2 and 1/5 the error in estimating γ_t^2. Again, although the jump variation estimators in the literature are not consistent for γ_t^2, they should be asymptotically unbiased. In large finite-samples, they appear both biased and inconsistent.

Table 2.3: Relative Simulation Error with Microstructure

Obs. per Min.	$\frac{\mathbb{E}[(\widehat{\sigma_t}-\sigma_t)^2]}{\mathbb{E}[\sigma_t]}$				$\frac{\mathbb{E}[(\widehat{\gamma_t}-\gamma_t)^2]}{\mathbb{E}[\gamma_t]}$			
	Bipower	LTT	5 Minute	Proposed	Bipower	LTT	5 Minute	Proposed
≈ 2	0.74	0.41	0.42	1.00	1.01	1.01	0.83	0.65
≈ 12	0.82	0.46	0.46	0.36	1.01	1.01	0.82	0.41
≈ 60	1.11	0.69	0.69	0.36	1.01	1.01	0.84	0.21
≈ 180	1.58	1.06	1.06	0.85	1.01	1.09	0.81	0.18

The last simulation result I report is a simulation with only a few jumps. To be precise, I simulate the volatilities using the DGP described in Table 2.1. Then instead of simulating the prices using (2.21), I follow Huang and Tauchen (2005) and assume the jump locations follow a time-invariant Poisson distribution and the magnitudes

are Gaussian distributed. I set the Poisson's intensity to result in an average of one jump per day. I set the variance of the magnitude so that the jump process has the volatility given by γ_t^2. This DGP should be quite difficult for my procedure because there are very few jumps. It drastically violates the infinite-activity assumption. I also add the microstructure noise as described in (2.22).

Table 2.4: Relative Simulation Error with Microstructure and Poisson Jumps

Obs. per Min.	$\frac{\mathbb{E}[(\widehat{\sigma_t^2}-\sigma_t^2)^2]}{\mathbb{E}[\sigma_t^2]}$				$\frac{\mathbb{E}[(\widehat{\gamma_t^2}-\gamma_t^2)^2]}{\mathbb{E}[\gamma_t^2]}$			
	BNS	LTT	5 Minute	Proposed	BNS	LTT	5 Minute	Proposed
≈ 2	0.88	0.12	0.20	0.88	1.01	1.01	0.78	0.34
≈ 12	0.95	0.13	0.21	0.51	1.01	1.01	0.79	0.32
≈ 60	1.17	0.32	0.41	0.09	1.01	1.01	0.80	0.39
≈ 180	1.55	0.77	0.85	0.58	1.01	1.01	0.75	0.36

As we can see in Table 2.4, the proposed method works well even in cases with only a few jumps. It uniformly dominates the other methods by a significant amount when estimating γ_t^2. The results are closer if we look at estimating σ_t^2. In that case, the estimation error is only smaller when we sample quite finely. However, for liquid stocks, this is the empirically-relevant case. Also, the proposed method outperforms by more when estimating the γ_t^2 then it underperforms when estimating σ_t^2. To see, this note that the average RMSE over the two targets is lowest with the proposed method at all frequencies. Consequently, I suggest using the proposed method in all cases except when we are both sure the price rarely jumps and we cannot sample very finely.[20]

[20] If we consider the situation without microstructure noise, the results are similar. The previous methods do perform better when estimating σ_t^2, but the proposed method still substantially outperforms in estimating γ_t^2.

2.7 Data

The methods developed in this paper require high-frequency data. For the analysis to be interesting, we need a dataset that faces a dense stream of relevant news. I chose SPY, (SPDR S&P 500 ETF Trust), which is an exchange-traded fund that mimics the S&P 500. I obtain it from the Trade and Quotes (TAQ) database at Wharton Research Data Services (WRDS). The S&P 500 is arguably the most important index of financial activity. It is likely the most closely watched equity index and several heavily subscribed index funds (including SPY) track it directly. Consequently, the economics and finance literature has studied it extensively, often using it as a proxy for the market.

Since this paper only use one asset, and SPY is one of the most liquid assets traded, we can essentially choose the frequency at which we want to observe the underlying price. In order to balance market-microstructure noise, computational cost, and efficiency of the resultant estimators I sample at the 1-second frequency. The data used starts in 2003 and ends in September 2017. Since the asset is only traded during business hours, this leads to 3713 days of data with an average of $\approx 24\,000$ observations per day. The dataset takes up about 4.4 GiB of memory. It is also worth noting that SPY is by far the most liquid exchange-traded fund, especially in recent years, reducing the effect of market microstructure, such as bid-ask spreads, bounces, and rounding error.

This market microstructure causes the asset to fail to be a semimartingale in practice. Thankfully, a substantial literature has developed to deal with precisely this issue. The two leading methods are sampling rather sparsely, for example at a 5-minute frequency as Liu, Patton, and Sheppard (2015) argue for, and pre-averaging, where one takes appropriately weighted averages of the price over small (shrinking) intervals. We must separate the jump volatility from the diffusion volatility, and so we must sample much more finely than once very 5-minutes. This requirement arises because the only information the estimators use to separate the jumps and diffusive component come from the tails, and tails by definition are times without much data. Consequently, any deconvolution procedure we use here is inherently low-powered.

Consequently, I preprocess the data using the pre-averaging approach as in Podolskij and Vetter (2009) and Aït-Sahalia, Jacod, and Li (2012). This procedure is known

not to affect the consistency of the estimation procedure. The basic idea is rather simple. We average the price over a small interval to remove the noise. If we pick the rates at which we shrink the interval to appropriately balance averaging away the noise and estimating the instantaneous variation, the estimators will be consistent even in the presence of market microstructure noise.

2.8 Volatility: Empirics

I separate this empirical part into three subsections. The first section characterizes the static properties of the volatilities. The second characterizes their dynamic properties. In particular, it shows that both volatilities are highly persistent, displaying long-memory. The third section introduces a new measure of jump variability — $\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}$ — in order to isolate the effect of γ_t^2 in the presence of σ_t^2. This ratio is a measure for the proportion of the investors' new information driven by news.

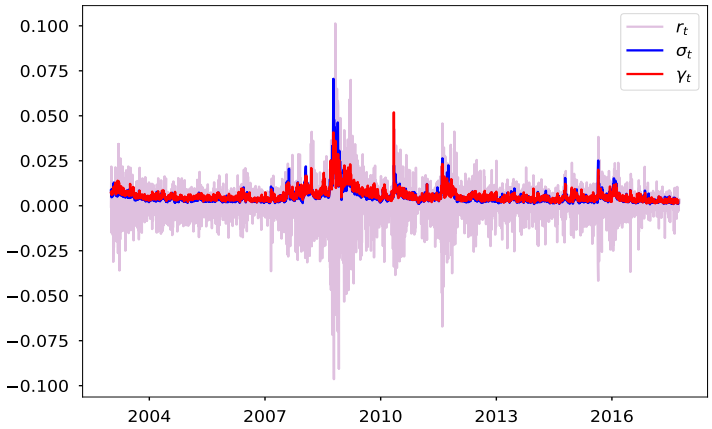
Statics

The results concerning σ_t^2 are broadly consistent with previous work on the topic. Since this paper introduces γ_t^2, the stylized facts regarding its features are new. Thankfully, in practice, σ_t^2 and γ_t^2 have very similar dynamics, and so much of the intuition regarding σ_t^2 can be directly translated to γ_t^2.

As can be seen in Figure 2.4, the volatilities are very closely related; their correlation coefficient equals 0.93. As one would expect from previous volatility measures, they both significantly increase during crises/recessions. Interestingly, σ_t^2 spiked more than γ_t^2 during the Financial Crisis and seems to spike more during recessions.

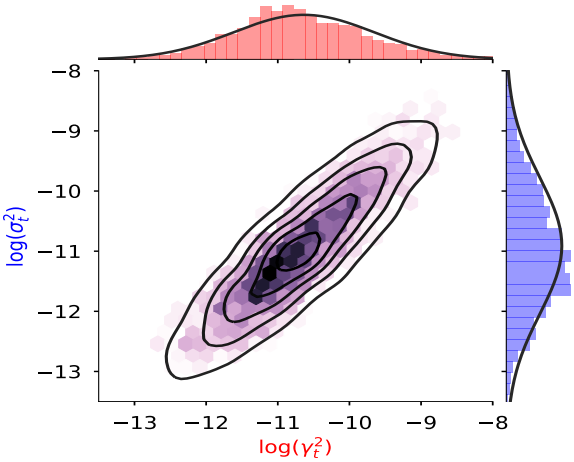
Figure 2.5 plots the two log-volatility distributions along with their joint distribution. As can be seen from the graph both marginal distributions are skewed right, and the joint distribution is just as skewed as the marginal densities. It is worth noting that being skewed right means that the volatilities are more likely to take on abnormally large values than take on abnormally small ones. Volatilities usually spike during crises, and so the distributions are skewed in a direction that increases the investors' risk relative to an unskewed distribution. This fact is particularly note-

Figure 2.4: Root Volatilities



worthy as these are distributions of log-volatilities, and taking the logarithm already removes a large amount of skewness.

Figure 2.5: Log-Volatility Densities



A few of the original realized volatility papers, (Andersen, Bollerslev, Diebold, and Labys 2001; Andersen, Bollerslev, Diebold, and Ebens 2001), argue that realized volatilities are approximately log-Gaussian. One might expect this to continue to hold in this case. The black lines in Figure 2.5 are Gaussian densities fit to the data for comparison purposes. At a qualitative level, the log-volatilities are roughly

log-Gaussian. They are slightly skewed and slightly kurtotic, even after taking logs, which we can also see in Table 2.5.[21]

Table 2.5: Volatility Summary Statistics

	σ_t^2	γ_t^2	$\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}$	$\log(\sigma_t^2)$	$\log(\gamma_t^2)$	$\log(\sigma_t^2+\gamma_t^2)$	$\log\left(\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}\right)$
Mean	4.47×10^{-5}	3.68×10^{-5}	0.56	−10.91	−10.64	−13.15	−2.17
Std. Dev.	1.52×10^{-4}	9.12×10^{-5}	0.12	1.13	0.98	1.03	0.22
Skew.	15.65	11.81	−0.18	0.71	0.55	0.72	−0.95
Kurt.	376.55	250.23	2.92	4.12	3.81	4.10	4.88

We are interested not just in the univariate dynamics, but also their relationships. We know from Figure 2.5 that the two volatilities move together. To investigate this further, Table 2.6 reports the correlations between the various volatility measures and daily excess returns.

Table 2.6 also includes an indicator — $\mathbf{1}\{\text{FOMC}\}_t$ — for days when the Federal Open Market Committee (FOMC) releases its announcements. As discussed in the literature review, much of the previous literature on the effect of news on asset prices has focused on the effect of FOMC announcements.

Table 2.6: Volatility Correlations

	σ_t^2	γ_t^2	$\sigma_t^2+\gamma_t^2$	$\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}$	rx_t	$\mathbf{1}\{\text{FOMC}\}_t$
σ_t^2	1.00	0.74	0.96	−0.29	−0.11	0.01
γ_t^2	0.74	1.00	0.89	−0.10	−0.13	0.06
$\sigma_t^2+\gamma_t^2$	0.96	0.89	1.00	−0.23	−0.13	0.05
$\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}$	−0.11	−0.13	−0.13	1.00	0.12	0.05

Clearly, σ_t^2 and γ_t^2 are highly positively correlated. Table 2.6 also reports the correlations between the logarithms of the parameters above because Pearson's correlation coefficients only measure linear relationships. On the other hand, $\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}$ is weakly negatively correlated with the other volatility measures. Note, this is possible because it is a non-linear transformation of γ_t^2 and $\sigma_t^2+\gamma_t^2$. Interestingly, $\mathbf{1}\{\text{FOMC}\}_t$

[21] The only reason that the diffusion density might appear to be skewed left is that it is plotted sideways.

is positively correlated with all of the volatility measures even though they are not all positively correlated with each other. The standard negative contemporaneous relationship between volatility and returns also holds.

Since the volatilities are closer to log-Gaussian than they are to Gaussian, Table 2.7 reports the correlations reported in Table 2.6 with the volatilities measured in terms of their logarithms.

Table 2.7: Log Volatility Correlations

	$\log(\sigma_t^2)$	$\log(\gamma_t^2)$	$\log(\sigma_t^2+\gamma_t^2)$	$\log\left(\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}\right)$	rx_t	$\mathbf{1}\{\text{FOMC}\}_t$
$\log(\sigma_t^2)$	1.00	0.90	0.97	−0.50	−0.18	0.06
$\log(\gamma_t^2)$	0.90	1.00	0.98	−0.08	−0.14	0.09
$\log(\sigma_t^2+\gamma_t^2)$	0.97	0.98	1.00	−0.29	−0.16	0.08
$\log\left(\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}\right)$	−0.29	−0.08	−0.29	1.00	0.13	0.04

The signs of the relationships are the same in both tables, but the magnitudes are larger in Table 2.7.

Dynamics

Having considered the data's static properties, I now consider the dynamic properties starting with the univariate case. Throughout, I focus on the log-volatilities because they are closer to Gaussian as shown in Section 2.8, and so the true conditional expectations are likely closer to approximately linear. I first replicate the standard stylized features for the diffusion volatility and show that the jump volatility behaves similarly. I then perform a joint analysis.

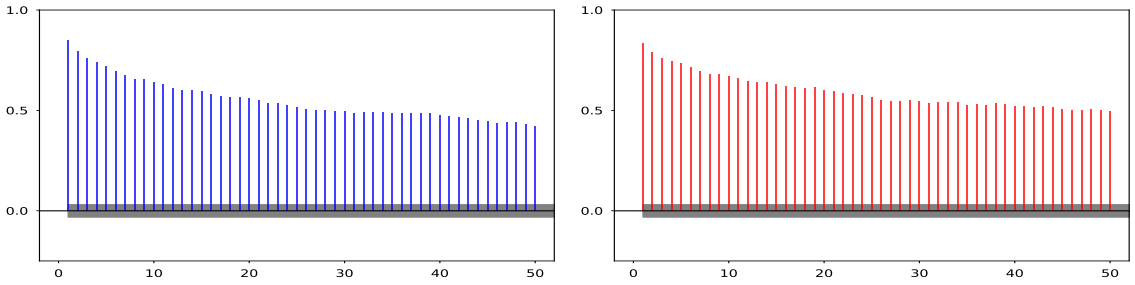
Measuring the Persistence

Figure 2.6 plots the volatilities' autocorrelation functions. Both series are extremely persistent.[22] We can also see that both series have a similar univariate autocorrelation structure.

[22]The gray bars are the standard Bartlett bands, i.e., confidence sets for the null of independent and identically distributed data.

Figure 2.6: Autocorrelation Functions

(a) $\log(\sigma_t^2)$ (b) $\log(\gamma_t^2)$



Since the series are so persistent, one might wonder if they have a unit root. Table 2.8 rejects this hypothesis. In particular, the standard Augmented Dickey-Fuller test rejects at any reasonable level of significance, (Dickey and Fuller 1981). Since the volatilities do not have a unit root, one might think that they are short memory processes; that is, their autocorrelation functions decay geometrically. Perhaps less surprisingly given Figure 2.6, the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test also rejects this hypothesis, (Kwiatkowski et al. 1992).

Those readers familiar with the empirical volatility literature should not find this result particularly surprising. The diffusion volatility's long memory is a key stylized fact in the literature (Andersen, Bollerslev, Diebold, and Labys 2003). Perhaps more surprisingly, the jump volatility also has long memory. Table 2.8 reports estimates for the long-memory coefficient, (d), using the Geweke Porter-Hudak (GPH) estimator (Geweke and Porter-Hudak 1983). The smoothed periodogram estimator developed by Reisen (1994) gives almost identical results.

Notably, the point estimates for d are in the infinite-variance region ($d > 1/2$). These estimates imply the volatility itself has an infinite unconditional variance.[23] However, we cannot reject the hypothesis that the $d < 1/2$ in any of the cases.

Univariate Dynamics

Table 2.9 reports independent $AR(1)$ regressions on each volatility to gain some high-level understanding of the dynamics. Both series are quite persistent and predictable.

[23]Having an infinite unconditional variance does not imply that the volatilities have an infinite conditional variance. A process can be locally square-integrable even if has infinite variance.

Table 2.8: Persistence Statistics

	$\log(\sigma_t^2)$	$\log(\gamma_t^2)$	$\log\left(\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}\right)$
	p-value		
ADF Test (Unit-Root Null)	1.90×10^{-5}	3.61×10^{-4}	8.01×10^{-15}
KPSS Test (Short-Memory Null)	$\ll 1\,\%$	$\ll 1\,\%$	$\ll 1\,\%$
	statistic		
1st-Order Autocorrelation	0.85	0.83	0.26
Fractional Integration Coefficient (d)	0.57 (0.45, 0.79)	0.66 (0.50, 0.82)	0.47 (0.31, 0.64)

However, we still have economically significant innovations.[24]

I now consider univariate autoregressive models for both series. I use Schwarz Information Criterion (SIC) to select the lag order.[25] This is not the ideal thing to do as it assumes away the long-memory that I just demonstrated. However, it still is useful to understand the short-memory dynamics of the two series. The two series both exhibit substantial autocorrelation as shown above, with the AR coefficients declining slowly. SIC chooses 9 lags for both series. The two series are both quite predictable in terms of $\mathbb{R}^2$ as well. The regressions chosen by SIC give an $\mathbb{R}^2$ of 76 % for $\log(\sigma_t^2)$ and 69 % for $\log(\gamma_t^2)$. These numbers are likely higher than that found in the literature, which are often in the neighborhood of 40 % to 50 %, because I am doing a better job at separating out the diffusion and jump volatilities, (Bollerslev, Patton, and Quaedvlieg 2016, 8). Effectively, my variables have less measurement error than is commonly used in the literature. In addition, volatility appears to be more predictable during the Great Recession, which is a large portion of my sample.

[24] This section's results come with the significant caveat that I am using estimated regressors and do not correct for this in my statistical results. For the most part, the evidence is so overwhelming the conclusions should not be affected, but, in some of the more borderline cases, it may be an issue.

[25] Other selection criteria such as Akaike information criteria (AIC) choose similar models. As one would expect, AIC chooses a few more lags.

Table 2.9: Univariate Autoregressive Models

	$\log \sigma_t^2$			$\log \gamma_t^2$		
	AR(1)					
Intercept	−1.63	(−1.82,	−1.45)	−1.78	(−1.97,	−1.59)
	0.85	(0.83,	0.87)	0.83	(0.82,	0.85)
$\mathbb{R}^2$	72 %			69 %		
	AR(BIC)					
Intercept	−0.68	(−0.88,	−0.48)	−0.62	(−0.81,	−0.42)
Lag 1	0.54	(0.51,	0.58)	0.46	(0.43,	0.49)
Lag 2	0.15	(0.11,	0.18)	0.17	(0.13,	0.21)
Lag 3	0.06	(0.02,	0.09)	0.05	(0.02,	0.09)
Lag 4	0.07	(0.04,	0.11)	0.08	(0.04,	0.11)
Lag 5	0.04	(0.00,	0.08)	0.09	(0.05,	0.13)
Lag 6	0.00	(−0.03,	0.04)	0.01	(−0.02,	0.05)
Lag 7	−0.00	(−0.04,	0.04)	0.01	(−0.03,	0.05)
Lag 8	−0.00	(−0.04,	0.04)	−0.02	(−0.05,	0.02)
Lag 9	0.08	(0.04,	0.11)	0.08	(0.05,	0.12)
$\mathbb{R}^2$	76 %			74 %		
Innovation Variance	0.31			0.25		

Joint Dynamics

The joint analysis starts by considering whether the two volatility series Granger cause each other. Standard tests conclusively reject the null of no-causality in either direction. The sum-of-squared residuals (SSR) test for $\log(\gamma_t^2)$ Granger-causing $\log(\sigma_t^2)$ with one lag returns a $\chi^2(df = 1)$ value of 298. Conversely, the SSR test for $\log(\sigma_t^2)$ Granger-causing $\log(\gamma_t^2)$ with one lag returns a $\chi^2(df = 1)$ value of 398. These results are robust to the number of lags chosen and to the specific version of the test. The tests overwhelmingly reject no-causality in every case. In other words, adding information about the jumps helps us to predict the diffusive variation, and vice-versa.

To make this operational, consider a vector autoregression (VAR). Here, the Schwarz Information Criterion (SIC) chooses 6 lags. Table 2.14, which is in Sec-

tion 2.F, reports the results. Table 2.10 reports the results for a VAR(1). The results for the more general specification are consistent with these results. The results are consistent with the Granger-causality results above. Both volatilities depend on the lags of both volatilities.

Table 2.10: VAR(1) Results

	Intercept	$\log\left(\sigma^2_{t-1}\right)$	$\log\left(\gamma^2_{t-1}\right)$	Innovation Variance	$\mathbb{R}^2$
$\log \sigma^2_t$	-0.84	0.56	0.38	0.33	74 %
$\log \gamma^2_t$	-1.80	0.34	0.48	0.27	72 %

The correlation between the innovations equals 0.63. Since both the unconditional correlation and the innovation correlation between the two series are high, there appears to be a shared component that drives a large amount of the variation in both series.

Jump Proportion

The previous sections showed that the volatilities share a component that drives a large portion of each of their variations. We would like to isolate the effect of the jumps and examine its dynamics directly. (This will be quite important when we consider the pricing implications.) To do this, define the jump proportion — $\frac{\gamma^2_t}{\sigma^2_t+\gamma^2_t}$ — which the previous sections briefly alluded to but did not investigate in detail.

To understand $\frac{\gamma^2_t}{\sigma^2_t+\gamma^2_t}$, Figure 2.7 plots its variation over time. Its mean and the Great Recession are plotted for reference. Figure 2.7 also displays the rolling average to visualize the series' low-frequency variation better. Clearly, $\frac{\gamma^2_t}{\sigma^2_t+\gamma^2_t}$ has substantial low- and high-frequency variation.

Figure 2.8a plots $\frac{\gamma^2_t}{\sigma^2_t+\gamma^2_t}$'s histogram. The red line is a kernel density estimate, and the back line is a Gaussian distribution fit to the data. As we can see, $\frac{\gamma^2_t}{\sigma^2_t+\gamma^2_t}$'s density is roughly log-Gaussian.

Since Figure 2.7 plots daily data, and the dataset spans several years, the graph is at too low a resolution to be easily comprehensible. Hence, Figure 2.9 plots $\frac{\gamma^2_t}{\sigma^2_t+\gamma^2_t}$ and $\mathbf{1}\{\text{FOMC}\}_t$ on the same graph in the most interesting sub-period in the data: 2008–2009. As we might expect given previous work, such as Andersen, Bollerslev, Diebold,

Figure 2.7: Time-Varying Jump Proportion

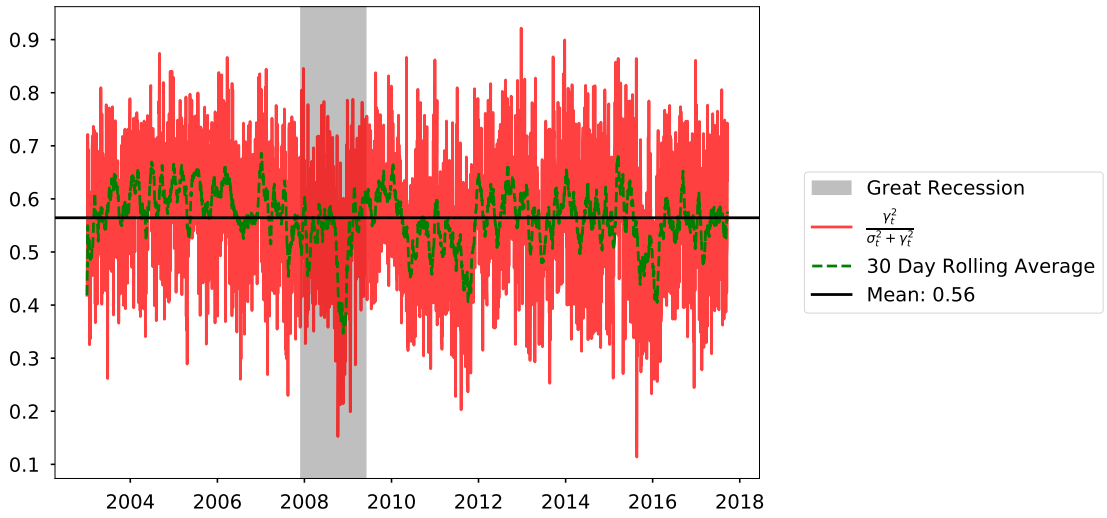
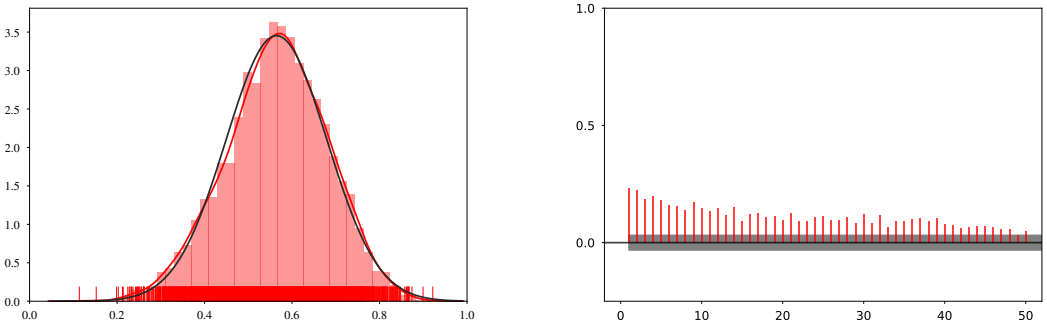


Figure 2.8: $\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}$

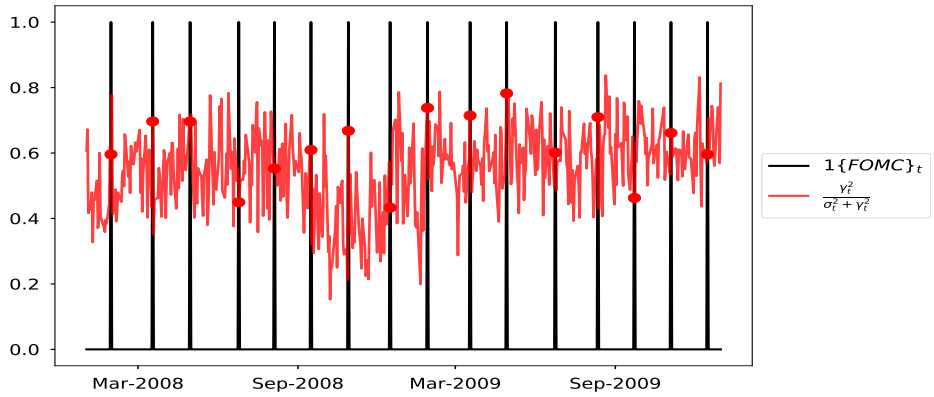


(a) Density

(b) Autocorrelation Function

and Vega (2003), Faust et al. (2007), and Beechey and Wright (2009), $\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}$ often spikes when the FOMC makes its announcements. However, $\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}$ varies significantly more than $\mathbf{1}\{\text{FOMC}\}_t$ does. If you regress $\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}$ on $\mathbf{1}\{\text{FOMC}\}_t$, the resulting coefficient is 0.50 with associated t-statistic equal to 7.90. Even though this relationship is highly statistically significant, The $\bar{\mathbb{R}}^2$ from this regression is only 0.78 %.

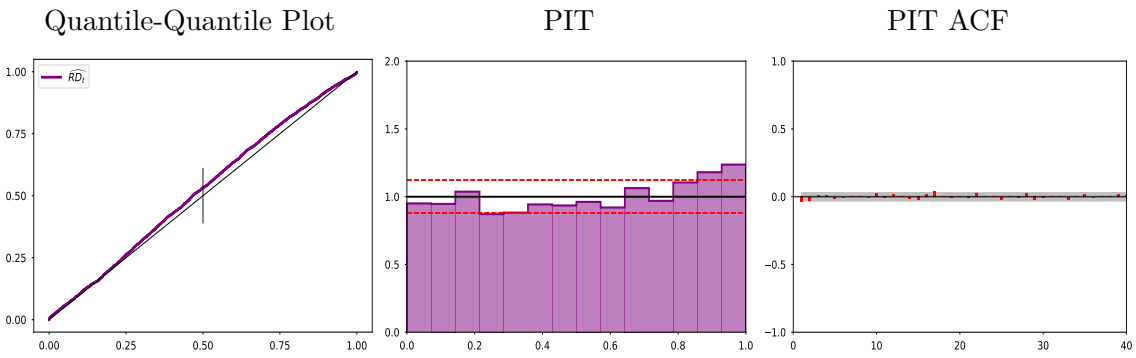
Figure 2.9: $\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}$ versus $\mathbf{1}\{\text{FOMC}\}_t$



Realized Density Evaluation

In Section 2.6, I showed you that the estimators work well in simulations. It would be useful to know if they worked well in the data as well. Besides, perhaps the assumptions justifying the integrated-Laplace representation do not hold in practice. Thankfully, Theorem 2.5 is a valid conditional density, and we can consistently estimate the conditioning variables. Consequently, techniques developed to analyze conditional densities work well here.

Figure 2.10: Realized Density Evaluation



Each day, I take the $\widehat{\sigma_t^2}$ and $\widehat{\gamma_t^2}$ and compute $\widehat{RD_t}$. I can draw from $\widehat{RD_t}$ easily, and so I compute its inverse-CDF through simulation. I then apply this inverse-CDF to the demeaned daily return. This procedure jointly evaluates the density representation, the time-aggregation procedure, and the estimation of σ_t^2 and γ_t^2.

As can be seen in Figure 2.10, the PIT is close to uniform. The only deviation is in the far right tail. I did not correct for the skewness in the data when I computed RD_t. We can see this in the graph. However, the deviation is not large, and for most risk-measures, we are far more concerned about the left-tail. I am estimating that tail almost perfectly. It is also worth noting that I needed to assume this symmetry in the discrete-time representation, but not the continuous-time one. The deviations here do not invalidate that representation at all. If we look at the Figure 2.10, we can see that the far right tail is also measured relatively well. It is only the 80th to 95th percentiles that I am missing. Furthermore, Figure 2.10 the deviations are most perfectly uncorrelated across time. This lack of correlation implies the densities dynamics are estimated quite well.

2.9 News Premia: Theory

Discontinuous prices or information flows that this paper considers break the derivation of CAPM-style results where risk-premia are instantaneous covariances with marginal utility. In particular, risk premia are no longer proportional to the integrated diffusive covariation between prices and stochastic discount factors. For example, in Ai and Bansal's (2018) world, we have an announcement SDF (A-SDF) whose covariation with returns is also priced, while in Tsai and Wachter's (2018) world, it is the covariation between the SDF and returns during extreme events that matters. This section unifies these two theories by decomposing the covariation between the prices and the investor's pricing kernels into their predictable and unpredictable components. In particular, it shows that risk premia have two components in the general case. Contrary to the discussion in Tsai and Wachter (2018, 31), both of these terms are, appropriately-defined, covariances, and so we would expect returns to exhibit a factor structure.

Preferences

To avoid introducing more notation than necessary, I characterize the investor's decision problem over a short period Δ and take $\Delta \to 0$. Following Ai and Bansal (2018), I adopt the intertemporal preferences represented as in Strzalecki (2013). Let

$V(t)$ be the representative investor's value function at time t, and $u(\cdot)$ be the associated flow utility over current period consumption — $C(t)$. Let κ denote the rate of time-preference. I assume that κ is constant for notation convenience, but this can be easily generalized.

Definition 2.11 (Certainty Equivalence Functional)**.**

$$V(t) = \int_t^{t+\Delta} u\left(c\left(t\right)\right) dt + \mathbb{I}\left[\int_{s\geq t+\Delta} \exp\left(-\kappa\left(s-t\right)\right) u(C(s))\, ds \,\middle|\, \mathcal{F}_{t-}\right].$$

I immediately specialize to the form given in Definition 2.11, which is (Ai and Bansal 2018, observation ii, p. 1401). Most of the results still go through in the general case, albeit with a loss of interpretability. In this case, investors preferences are represented as

$$\mathbb{I}\left[V(t) \,|\, \mathcal{F}_{t-}\right] = \phi^{-1}\left(\mathbb{E}\left[\phi(V(t)) \,|\, \mathcal{F}_{t-}\right]\right), \tag{2.23}$$

for some strictly increasing function ϕ. Preferences having this form include the recursive utility of Kreps and Porteus (1978) and Epstein and Zin (1989) and the second-order expected utility of Ergin and Gul (2009). If ϕ is the identity function, preferences are time-separable. They also cleanly characterize the problem at hand. Ai and Bansal (2018) show that these form of preferences lead to an announcement premium if and only if ϕ is concave.

I make filtration, $\mathcal{F}_t$, explicit in (2.23) to emphasize that certainty equivalence functionals map information sets to utility. In particular, $\mathcal{I}$ is just the expectations operator, $\mathbb{E}\left[\cdot \,|\, \mathcal{F}_{t-}\right]$, if preferences are time-separable.

Note, if $V(t)$ is continuous, it is predictable, i.e., $V(t) \in \mathcal{F}_{t-}$, then

$$\mathbb{I}\left[V(t) \,|\, \mathcal{F}_{t-}\right] = \phi^{-1}\left(\mathbb{E}[\phi(V(t)) \,|\, \mathcal{F}_{t-}\right) = \phi^{-1}\left(\phi(V(t))\mathbb{E}[1 \,|\, \mathcal{F}_{t-}\right) = \phi^{-1}(\phi(V(t))) = V(t). \tag{2.24}$$

In other words, the recursive utilities can be appropriately reparameterized in terms of a time-separable preferences. Now, there is no reason to assume the reparameterization is particularly convenient for analysis, just that it exists. This validity of this reparameterization is why Tsai and Wachter (2018, Theorem 5) find a single risk price is sufficient even in the presence of recursive utility as long as the underlying shocks are continuous.

To make this more concrete, consider the example of Epstein-Zin preferences. I adopt the notation used in Bansal and Yaron (2004). Let ρ denote risk aversion and ψ denote to the intertemporal elasticity of substitution (IES). Then Epstein-Zin Utility can be represented as:

$$U_t = \left[C_t^{1-1/\psi} + \exp(-\kappa\Delta)\mathbb{E}\left[U_{t+\Delta}^{1-\rho} \,\middle|\, \mathcal{F}_t \right]^{\frac{1-1/\psi}{1-\rho}} \right]^{\frac{1}{1-1/\psi}} \tag{2.25}$$

The formulation in (2.25) is not in the form given in (2.23), and so is not particularly useful for our purposes. Define $V_t \coloneqq U_t^{1-1/\psi}$ and reparameterize (2.25) as

$$V_t = \left[C_t^{1-1/\psi} + \exp(-\kappa\Delta)\mathbb{E}\left[V_{t+\Delta}^{\frac{1-\rho}{1-1/\psi}} \,\middle|\, \mathcal{F}_t \right]^{\frac{1-1/\psi}{1-\rho}} \right]. \tag{2.26}$$

Let $\phi(V) \coloneqq \frac{1-\rho}{1-1/\psi} V^{\frac{1-\rho}{1-1/\psi}}$ and $U(C(t)) \coloneqq C_t^{1-1/\psi}$, then we have: [26]

$$V_t = \left[u(C_t) + \exp(-\kappa\Delta)\phi^{-1}\left(\mathbb{E}\left[\phi\left(V_{t+\Delta} \right) \middle| \mathcal{F}_t \right] \right) \right]. \tag{2.27}$$

The Investor's Portfolio Optimization Problem

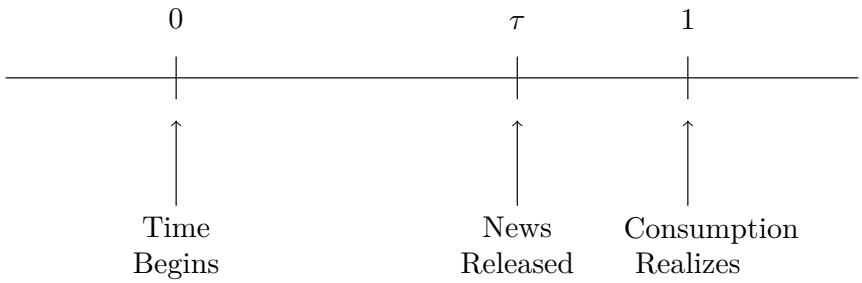
To fix intuition, consider a one-period version of the model. The investor can continually trade between time 0 and time 1 and consumes her wealth at time 1 as displayed in Figure 2.11. At some time $\tau \in (0, 1)$, a news item is released, on which the investor can trade. The investor's preferences satisfy the following utility recursion for any time $\tau > t$.

$$V_t(W_t) = u(C_t) + \phi^{-1}\left(\mathbb{E}\left[\phi\left(V_\tau(W_\tau) \right) \middle| \mathcal{F}_{t-} \right] \right). \tag{2.28}$$

Assume that the investor has access to three assets. 1) A risk-less asset, $\chi_{f,t}$, that pays off 1 unit in every period, and whose price equals 1 because investors do not discount the future. Essentially, it is a costless storage technology. 2) An asset, ζ_t, whose payout, R_ζ, is announced by the news release. 3) An asset ξ_t whose payout R_ξ realizes as a Brownian motion, i.e., its variance and mean are both proportional to the length of the remaining interval. Figure 2.11 displays the timing. I maintain the convention where the time subscript refers to when the variable first enters the investor's information set.

[26] The constant in front cancels between ϕ and ϕ^{-1}, and so does not affect the level of utility. It is there to ensure that ϕ is an increasing function regardless of the values of the parameters.

Figure 2.11: Timing



Since this is a finite-horizon problem, we can solve it by working backward. At time 1, all uncertainty has been resolved, and the representative agent eats all of her wealth:

$$V_1(W_1) = u(W_1). \tag{2.29}$$

Let $\tau < t < 1$, then the investor can trade ξ_t and the risk-less asset χ_t. However, since the news was already released, we know the payout of ζ_t, and so it is a risk-free asset. Consequently, the value function satisfies

$$V_t(W_t) = \max_{\xi} \phi^{-1}(\mathbb{E}\left[\phi\left(V_1\left(W_1\right)\right) | \mathcal{F}_t\right]), \text{where } W_1 = W_t + (1-t)(R_\xi - 1)\xi_t, \tag{2.30}$$

because she gets return 1 from the risk-less asset and R_ξ from the risky asset over the course of the entire interval. By substituting the constraint into the problem, and noting that $V_1(W_1) = u(W_1)$, this simplifies to

$$V_t(W_t) = \max_{\xi} \phi^{-1}\left(\mathbb{E}\left[\phi\left(u\left(W_t + (1-t)\xi_t(R_\xi - 1)\right)\right) | \mathcal{F}_t\right]\right). \tag{2.31}$$

The first-order condition is

$$0 = \mathbb{E}\left[\phi'(u(W_1))u'(W_1)(1-t)(R_{\xi_t} - 1) \,|\, \mathcal{F}_t\right], \tag{2.32}$$

since the term arising from the derivative of ϕ^{-1} is always positive.

The investor's problem for some time t in $(0, \tau)$ has similar structure except now she trades both assets.

$$V_t(W_t) = \max_{\zeta_t, \xi_t} \phi^{-1}(\mathbb{E}\left[\phi(V_\tau(W_\tau)) \,|\, \mathcal{F}_t\right]) \tag{2.33}$$

$$W_\tau = W_t + (R_\zeta - 1)\zeta_t + (\tau - t)(R_\xi - 1)\xi_t \tag{2.34}$$

Substituting the constraints into the problem gives

$$V_t(W_t) = \max_{\zeta_t, \xi_t} \phi^{-1}\left(\mathbb{E}\left[\phi\left(V_\tau\left(W_t + (R_\zeta - 1)\zeta_t + (\tau - t)(R_\xi - 1)\xi_t\right)\right) \middle| \mathcal{F}_t\right]\right). \tag{2.35}$$

Taking first-order conditions with respect to the x for $x \in \{\zeta, \xi\}$ and simplifying gives

$$0 = \mathbb{E}\left[\phi'\left(V_\tau\left(W_\tau\right)\right) V_\tau'(W_\tau)\left(R_x - 1\right) \middle| \mathcal{F}_t\right]. \tag{2.36}$$

Consider some time immediately before τ, $\tau-$, and some time right after τ, $\tau+$. Then, substitute (2.32) into (2.35) and consider the derivative with respect to ξ_t:[27]

$$0 = \mathbb{E}\left[\mathbb{E}\left[\phi'\left(V_\tau\left(W_\tau\right)\right) V_\tau'(W_\tau) \middle| \mathcal{F}_{t+}\right](d\xi)(R_\xi - 1) \middle| \mathcal{F}_{\tau-}\right]. \tag{2.37}$$

Since ξ is continuous, it is orthogonal to all discontinuous process, and so we can replace the expectation with respect to τ^+ with an expectation with respect to τ^-:

$$0 = \mathbb{E}\left[\mathbb{E}\left[\phi'\left(V_\tau\left(W_\tau\right)\right) V_\tau'(W_\tau) \middle| \mathcal{F}_{t-}\right]\left(R_\xi - 1\right) \middle| \mathcal{F}_{\tau-}\right] \tag{2.38}$$

Consequently, the investor only cares about the predictable part of the co-variation. In order for the returns to have finite variances, this must be proportional to the length of the interval. Their variance is proportional to $(\tau^+) - (\tau^-) \approx 0$. Hence, the Brownian asset ξ is risk-less over short enough intervals.

I substitute (2.32) into (2.35), and consider the derivative with respect to ζ_t. The jump asset, ξ_t, is not risk-less over short enough intervals:

$$0 = \mathbb{E}\left[\mathbb{E}\left[\phi'\left(V_\tau\left(W_\tau\right)\right) V_\tau'(W_\tau)(R_\zeta - 1)(d\zeta) \middle| \mathcal{F}_{\tau-}\right] \middle| \mathcal{F}_{\tau-}\right]. \tag{2.39}$$

We cannot pull R_ζ outside of the inner expectation because it is not predictable, i.e., it is not contained in the $\mathcal{F}_{\tau-}$ information set. Hence, there is no reason to expect the ex-post variation to be proportional to the length of the interval.

To facilitate comparing the two equations, isolate the unpredictable variation in the SDF by multiplying and dividing through by its left-limit, which we can pull through the inner expectation:

$$0 = \mathbb{E}\left[\phi'\left(V_{\tau-}\left(W_{\tau-}\right)\right) V_{\tau-}'(W_{\tau-})\mathbb{E}\left[\frac{\phi'\left(V_\tau\left(W_\tau\right)\right) V_\tau'(W_\tau)}{\phi'\left(V_{\tau-}\left(W_{\tau-}\right)\right) V_\tau'(W_\tau)}(R_\zeta - 1)(d\zeta) \middle| \mathcal{F}_{t-}\right] \middle| \mathcal{F}_{\tau-}\right] \tag{2.40}$$

Note, $\phi'(V_{\tau-}(W_{\tau-}))V_{\tau-}'(W_{\tau-})$ is predictable, while $\frac{\phi'(V_\tau(W_\tau))V_\tau'(W_\tau)}{\phi'(V_{\tau-}(W_{\tau-}))V_\tau'(W_\tau)}$ is purely unpredictable.

[27] I am taking limits here with respect to time loosely here to provide intuition. I make the statements rigorous in the theorems below.

Deriving the Asset-Pricing Equation

The market environment is mostly standard. We need a series of technical conditions that ensure that preferences are reasonable and first-order conditions uniquely characterize the optimum.

Assumption. Market Environment

1. Both u and ϕ are Lipschitz continuous with Lipschitz derivatives.
2. $u : \mathbb{R} \to \mathbb{R}$ has strictly positive first-order derivatives, and ϕ is increasing with first- and second-order derivatives that are bounded away from zero and infinity.
3. A representative investor prices all assets.
4. Consumption — $C(t)$ — is an Itô semimartingale.
5. All of the stochastic processes do not contain predictable jumps.

The fourth assumption is the principal distinction from the setup in Ai and Bansal (2018). Assumption 2.6 generalize their assumptions by allowing consumption to jump. I will discuss later how my results slightly simplify if we require consumption to be continuous. The third assumption is likely unnecessarily restrictive, most of the results in this section would go through in terms of a marginal investor's preferences. I make this assumption to simplify the exposition.

Consider an representative investor with preferences given by (2.23). She has access to a (potentially infinite) vector of assets $\Xi(t) \coloneqq \xi_1(t), \ldots$. Assume for simplicity that she has no other sources of income. Over some small length of time Δ, the investor's problem is as follows. She enters into the period with asset allocation $\Xi(t - \Delta)$, and prices are $p(t)$.[28] She need to solve for consumption $C(t)$ and an asset allocation $\Xi(t)$. The results are reported cum-dividend to avoid introducing even more notation. The extension to the ex-dividend case is straightforward.[29]

[28] The timing notation may seem somewhat strange here because it maintains the convention used elsewhere in the paper where time arguments denote when the objects first enter the representative investor's information set.

[29] Cum-dividend means before dividend. Assets here behave like Bitcoin or gold and never pay out dividends.

Problem 2.1. Consumer's Portfolio Allocation

$$V\left(\Xi(t-\Delta), P(t)\right) = \max_{C(t),\Xi(t)} \int_t^{t+\Delta} u(C(s))\, ds + \phi^{-1}\left(\mathbb{E}\left[\frac{\phi\left(V\left(\Xi(t), P(t+\Delta)\right)\right)}{\exp\left(\kappa\Delta\right)} \,\middle|\, \mathcal{F}_t\right]\right)$$
$$C(t) + \sum_i P_i(t)\xi_i(t-\Delta) = \sum_i P_i(t)\xi_i(t)$$

The continuous-time problem is the limit of Problem 2.1 as Δ approaches 0. The trade-offs are slightly easier to see in the discrete-time problem. The investor must purchase consumption, $C(t)$, and assets, $\Xi(t)$, at prices, $P_i(t)$, using wealth, $\sum_i P_i(t)\Xi(t-\Delta)$. Let $\widetilde{P}(t) = \exp\left(-\kappa(t)\right) p(t)$ be the appropriately discounted price. We are interested in excess returns, not returns themselves. Then we can derive the following result, where $p(t)$ refers to the price.[30]

Theorem 2.11 (Asset-Pricing Equation)**.** *Let Assumption 2.6 hold, prices be Itô semimartingales, and the representative consumer face Problem 2.1 as* $\Delta \to 0$*. Assume preferences are such that optimal consumption is strictly positive. Define*

$$M^{UP}(t) \coloneqq \frac{\phi'(V\left(W(t)\right))}{\phi'(V\left(W(t-)\right))} \text{ and } M(t) \coloneqq \frac{\phi'(V\left(W(t-)\right)}{\phi'\left(\phi^{-1}\left(\mathbb{E}\left[\phi(V(W(t)))\,|\,\mathcal{F}_{t-}\right]\right)\right)} \frac{V'\left(W(t)\right)}{u'(c(t-))}.$$

Then $M^{UP}(t)$ *is a purely discontinuous martingale, and for all stopping times* $\tau > t$*,*[31]

$$\widetilde{P}(t) = \mathbb{E}\left[M(\tau)M^{UP}(\tau)\widetilde{P}(\tau) \,\middle|\, \mathcal{F}_t\right]$$

Conceptually, Theorem 2.11 is straightforward. Prices are semimartingales, and so we have a pricing kernel — $\mathcal{M}(t)$ — that prices all assets:

$$\widetilde{P}(t) = \mathbb{E}\left[\mathcal{M}(\tau)\widetilde{P}(\tau) \,\middle|\, \mathcal{F}_t\right]. \tag{2.41}$$

However, $\mathcal{M}(t) \not\propto V'(W(t))$. Instead, it has two parts: $M(t)$, which reflects compensation for consumption risk and $M^{UP}(t)$ which reflects compensation for discontinuities in the investor's information set.

[30] This is a generalization of Ai and Bansal (2018, Theorem 2) to allow for jumps in consumption.
[31] I use the UP superscript because M^{UP} is an unpredictable process.

Deriving Risk Premia

The end of the previous section is essentially where Ai and Bansal (2018) stop. Section 2.10 estimates risk premia, and so I must derive risk premia from Theorem 2.11. If prices were continuous, Itô's formula lets us solve for the expected log-return in terms of the covariance between the $M(t)$ and $p(t)$. However, the generalized Itô's formula in the literature that applies to general semimartingale does not have a simple form in terms of covariances. To resolve this impasse, I derive a generalized Itô's formula in terms of predictable quadratic covariation, (integrated diffusive and jump volatilities) that has the standard form but applies to jump processes.

Lemma 2.12 (An Itô's Formula for the Expectation of a Square Integrable Semimartingale)**.** *Let f be a twice-differentiable function and $\widetilde{Z}$ be a vector-valued semimartingale with locally bounded predictable $\langle Z\rangle(t)$. Then the differential of f satisfies*

$$d\mathbb{E}\left[f(\widetilde{Z})\,\middle|\,\mathcal{F}_{t-}\right] = \mathbb{E}\left[f'(\tilde{Z}(t-))\,d\tilde{Z}(t)\,\middle|\,\mathcal{F}_{t-}\right] + \frac{1}{2}f''(\tilde{Z}(t-))\,d\langle\tilde{Z}\rangle(t).$$

The assumptions and conclusion in Lemma 2.12 are both weaker than Itô's formula for continuous processes. We do not need continuous processes, but the equality only holds in expectation. However, this is sufficient for our purposes as risk premia are expectations. Importantly, the convexity correction has the same form as it does in the standard Itô's formula.

I now compute risk premia by applying Lemma 2.12 to the logarithm. Let $m(t) \coloneqq \log(M(t))$ and $m^{UP}(t) \coloneqq \log(M^{UP}(t))$. Recall that throughout, $p(t)$ refers the log-price. Let P_f denote the price of the risk-free asset.

Theorem 2.13 (Asset-Pricing Equation)**.** *Let the assumptions in Assumption 2.6 hold, $P_i(t)$ be an Itô semimartingales, and the representative consumer face Problem 2.1 as $\Delta \to 0$. Assume that preferences are such that optimal consumption is strictly positive. Then risk-premia for some asset i is*

$$\mathbb{E}\left[\frac{dP_i(t)}{P_i(t-)} - \frac{dP_f(t)}{P_f(t-)}\,\middle|\,\mathcal{F}_{t-}\right] = -\,d\langle m, p^D + p^J\rangle(t) - d\langle m^{UP}, p^J\rangle(t).$$

The cost of the assumptions' generality is that Theorem 2.13 is rather abstract. To make the representation more concrete, consider a few specializations. First,

assume that preferences are time-separable and consumption is continuous. Then $M^J(t)$ is identically one, and $M(t) = \frac{u'(C(t))}{u'(C(t-))}$ by the envelop theorem. Consequently, $\mathbb{E}\left[\frac{dP_i(t)}{P_i(t-)} - \frac{dP_f(t)}{P_f(t-)} \middle| \mathcal{F}_{t-}\right] = -\sigma_m' \sigma_{p_i}$. This is the consumption-CAPM model of Breeden (1979).

In addition, since $m(t)$ is continuous, which is implied by $u'(\cdot)$ being a smooth function and $C(t)$ being continuous, jumps do not command a premium. This is because time-separability implies $m^{UP}(t)$ is identically zero and so $d\langle m, p^D + p^J\rangle(t) + d\langle m^{UP}, p^J\rangle(t) = d\langle m, p^D\rangle$.

If we allow for jumps and recursive utility, Theorem 2.13 generalizes Tsai and Wachter (2018, Theorem 5). The key difference is that it is apparent that the second term is a covariance. Even in the presence of jumps, risk premia can be split into a risk price and risk quantity. It is not immediately apparent in Tsai and Wachter's (2018) environment that their formula is a covariance, but as long as returns have finite variance, we can rewrite their expression as a predictable quadratic variation.[32]

If we are in a world like Ai and Bansal (2018), where consumption is continuous and the envelop theorem holds (which implies $V'(W(t))$ is a continuous process), the equation in Theorem 2.13 simplifies to

$$\mathbb{E}\left[\frac{dP_i(t)}{P_i(t-)} - \frac{dP_f(t)}{P_f(t-)} \middle| \mathcal{F}_{t-}\right] = -d\langle m, p^D\rangle(t) - d\langle m^{UP}, p^J\rangle(t). \tag{2.42}$$

If we further assume that high-frequency consumption movements can be ignored, i.e., $u'(c(t-)) = u'(C(t))$, we can combine the two terms using the law of iterated expectations.

$$\mathbb{E}\left[\frac{dP_i(t)}{P_i(t-)} - \frac{dP_f(t)}{P_f(t-)} \middle| \mathcal{F}_{t-}\right] = -d\left\langle \frac{\phi'(V(Wt))}{\phi'(V(W(t-)))}, p(t)\right\rangle. \tag{2.43}$$

Again, we have a single risk-price and risk-premia equal $-\sigma_m' \sqrt{(\sigma_t^2 + \gamma_t^2)_i}$. Doing this is equivalent to assuming that the market wealth portfolio is the only risk factor and ignoring movements in the wealth-consumption ratio. However, as will be shown below, the data require a two-factor model.

We could instead assume that consumption is continuous and the envelop theorem holds, but high-frequency movements in consumption cannot be ignored. In that case,

[32] This is implied by Lemma 2.12.

the risk premia equation is

$$\mathbb{E}\left[\frac{dP_i(t)}{P_i(t-)} - \frac{dP_f(t)}{P_f(t-)}\,\middle|\,\mathcal{F}_{t-}\right] = -\,d\langle m, p^D\rangle(t) - d\langle m^{UP}, p^J\rangle(t). \tag{2.44}$$

In this case, we could isolate the effects of each of the two factors by estimating the risk-premia as a bivariate function of σ_t^2 and γ_t^2. In that case, we could view the second term as a measure of announcement premia, similar to how Ai and Bansal (2018) use the excess returns on FOMC days. However, then the regressions done below imply that ϕ is convex because γ_t^2 predicts lower risk premia once we condition in σ_t^2. Regardless, the data demand we have two factors that move at high-frequency. They also require the news risk premium to be less than the diffusion volatility premium.

2.10 News Premia: Empirics

Recall the formula for risk-premia in the presence of recursive utility and jumps derived in Theorem 2.13:

$$\mathbb{E}\left[\frac{dP_i(t)}{P_i(t-)} - \frac{dP_f(t)}{P_f(t-)}\,\middle|\,\mathcal{F}_{t-}\right] = -\,d\langle m, p^D\rangle(t) - d\langle m^{UP}, p^J\rangle(t). \tag{2.45}$$

In general, we must to specify a full model for both $m(t)$ and $m^{UP}(t)$ in order to take (2.45) to the data. There are two leading cases. One, make a CAPM-style approximation that assumes market wealth is the only factor, i.e., $V_t = V(W_t)$.[33] In this case, we have

$$\mathbb{E}\left[\frac{dP_i(t)}{P_i(t-)} - \frac{dP_{rf}(t)}{P_{rf}(t-)}\,\middle|\,\mathcal{F}_{t-}\right] = \beta_1(\sigma_t^2 + \gamma_t^2) + \beta_2\gamma_t^2 \tag{2.46}$$

because $V(t)$ is perfectly correlated with $p(t)$. Two, assume that the news structure and underlying productivity shocks are continuous as in Ai and Bansal (2018). Then $V(t)$ is a continuous process and $\langle m, p^J\rangle = 0$. This gives

$$\mathbb{E}\left[\frac{dP_i(t)}{P_i(t-)} - \frac{dP_{rf}(t)}{P_{rf}(t-)}\,\middle|\,\mathcal{F}_{t-}\right] = \beta_1\sigma_t^2 + \beta_2\gamma_t^2. \tag{2.47}$$

[33] Under the assumptions in Assumption 2.6, this implies V_t jumps since W_t does.

Excess Return and Volatility: Contemporaneous Relationship

The question now facing us is how should we estimate (2.46) and (2.46). In practice, $\sigma_t^2 + \gamma_t^2$ and γ_t^2 are very heavily correlated, (89 %), and so regressing on them does not lead to robust results. Moreover, interaction terms in those regressions are often significant. To isolate the effect of the jumps, I use $\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}$ instead of γ_t^2. To make the results more Gaussian and avoid the need for interaction terms, I report elasticities, i.e., I apply a log transformation. Hence, the preferred specification is

$$rx_t = \beta_0 + \beta_1 \log\left(\sigma_t^2 + \gamma_t^2\right) + \beta_2 \log\left(\frac{\gamma_t^2}{\sigma_t^2 + \gamma_t^2}\right) + \epsilon_t. \tag{2.48}$$

I report robustness results in Section 2.D. The results in the other specifications either agree with the main specification or are insignificant.

Consider the contemporaneous relationship between the volatility and the return. This section starts by replicating the standard result that volatility and returns are contemporaneously negatively correlated, (Lettau and Ludvigson 2010). The crucial difference between the results reported here and those in the literature is that Table 2.11 splits contemporaneous relationship up into relationships with $\sigma_t^2 + \gamma_t^2$ and with $\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}$.

The analysis below uses the daily excess return, rx_t, to make the results more easily comparable with those in the literature. I construct rx_t by taking r_t and subtracting the log yield on the 10 year treasury bill, which is obtained from FRED. I annualize rx_t (multiplied it by 252) to make the results more interpretable. I use Newey-West heteroskedasticity and autocorrelation (HAC) robust standard errors and report t-statistics in the square brackets. I use Bartlett's kernel with the optimal bandwidth, per Newey and West (1994).

As can clearly be seen in Table 2.11, $\log(\sigma_t^2 + \gamma_t^2)$ and rx_t are strongly negatively correlated. This is what the literature has found Brandt and Kang (2004) and Lettau and Ludvigson (2010). The unconditional positive relationship between $\log(\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2})$ and rx_t is new, however.

Weighted least squares is more efficient than ordinary least squares if we choose the weights appropriately. Section 2.D reports weighted regressions that weight each datapoint by the inverse of that day's total volatility. This weighting is optimal up to

Table 2.11: $\mathbb{E}\left[rx_t \,\middle|\, \sigma_t^2 + \gamma_t^2, \frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}\right]$ (OLS)

Regressors	Specifications			
Intercept	−4.55 [5.81]	1.02 [6.48]	−3.17 [−3.94]	−1.27 [−0.54]
$\log(\sigma_t^2 + \gamma_t^2)$	−0.46 [−5.58]		−0.39 [−4.12]	−0.19 [−0.85]
$\log\left(\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}\right)$		1.65 [5.81]	1.13 [4.06]	3.91 [1.07]
$\log(\sigma_t^2 + \gamma_t^2)\log\left(\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}\right)$				0.29 [0.80]
$\bar{\mathbb{R}}^2$	2.67 %	1.61 %	3.35 %	3.42 %

unpredictable terms because $\sigma_t^2 + \gamma_t^2$ equals variance of the martingale part of $p(t)$ in expectation. This martingale part is the innovation in (2.48). In order to deal with any residual heteroskedasticity, Section 2.D continues to use robust standard errors.

It is also worth noting that since the regressions are contemporaneous, the $\bar{\mathbb{R}}^2$'s that Table 2.11 reports are reasonable. The volatility explains a notable, but small, part of the variation in the excess return.

News Premia

The regressions in Table 2.11 are contemporaneous, and so they conflate risk premia and volatility feedback effects. If we try to interpret the coefficients as measures of risk premia, we have the classic endogenous regressors problem because the regressors and error terms are correlated.

Risk premia are forward-looking by definition, and so we must isolate the predictable variation in the regressors. Intuitively, we want to regress returns on expected volatilities.[34] The most common way of handling endogenous regressors is using instrumental variables, and that is what I do.

In particular, I use the lagged regressors as instruments. This procedure gives better estimates than regressing on the lagged volatilities directly for three reasons.

[34]This is equivalent to regressing expected returns on volatilities, but we do not observe expected returns.

First, $\sigma_t^2 + \gamma_t^2$ and $\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}$ are not AR(1) processes. Hence, regressing directly on $\sigma_{t-1}^2 + \gamma_{t-1}^2$ and $\frac{\gamma_{t-1}^2}{\sigma_{t-1}^2+\gamma_{t-1}^2}$ unnecessarily throws away useful information. Second, the coefficients from these this regression conflate predictability of the volatilities and risk premia. Consequently, they only identify the sign, not the magnitude of the risk premia. Third, as discussed in Section 2.10, returns are highly heteroskedastic. Since I consistently estimate $\sigma_t^2 + \gamma_t^2$, I can adjust for heteroskedasticity in the instrumented contemporaneous relationship. It is not obvious how to do this appropriately if you regress rx_t on $\sigma_{t-1}^2 + \gamma_{t-1}^2$ and $\frac{\gamma_{t-1}^2}{\sigma_{t-1}^2+\gamma_{t-1}^2}$.

The lagged volatilities are valid instruments. First, they explain a large amount of the variation in the regressors. I adopt an approximate heterogeneous autoregressive (HAR) specification to choose lags used as instruments, (Corsi 2009). To be precise, I use $\sigma_{t-l}^2 + \gamma_{t-l}^2$, $\frac{\gamma_{t-l}^2}{\sigma_{t-l}^2+\gamma_{t-l}^2}$ for $l \in \{1, 2, 5, 25\}$ as instruments. I report the results from the first-stage regressions in Table 2.17. The $\bar{\mathbb{R}}^2$ for the $\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}$ regression equals 14.63 % with an associated F-statistic of 248. The $\bar{\mathbb{R}}^2$ for the $\sigma_t^2 + \gamma_t^2$ regression equals 79.43 % with an associated F-statistic of 20,140. Both of these are comfortably within the strong instruments region. Second, they are predetermined. Consequently, they are, by definition, independent of the date-t innovation. Innovations are unpredictable.

I consider two specifications. The leading specification uses $\log(\sigma_t^2 + \gamma_t^2)$ as my first regressor and $\log(\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2})$ as my second regressor. I also consider a specification with $\log(\sigma_t^2)$ as the first regressor and $\log(\sigma_t^2 + \gamma_t^2)$ as the second regressor.

Table 2.12: News Premia Estimates

Regressors	Specifications					
Intercept	2.95 [6.61]	−2.45 [−5.12]	−5.04 [−0.58]	3.27 [7.25]	2.95 [6.07]	2.32 [4.15]
$\log(\sigma_t^2 + \gamma_t^2)$	0.24 [6.61]		0.14 [2.68]			
$\log\left(\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}\right)$		−5.01 [−5.86]	−4.15 [−4.93]			
$\log(\sigma_t^2)$				0.25 [6.53]		1.86 [5.18]
$\log(\gamma_t^2)$					0.23 [5.40]	−1.74 [−4.53]

The question facing us is how do we interpret the coefficients reported in Table 2.12. If $\log(V_t)$ is proportional to wealth and both consumption and wealth move at high-frequency, then the coefficient on $\sigma_t^2 + \gamma_t^2$ measures risk aversion. The coefficient on $\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}$ measures the curvature of the CEF as parameterize by ϕ. Per the discussion in Section 2.9, this implies the CEF is convex, and if investors have Epstein-Zin preferences they prefer late resolution of uncertainty.

Conversely, if consumption is the only factor and is continuous, the coefficient on $\log(\sigma_t^2)$ measures risk aversion. The coefficient on $\log(\gamma_t^2)$ measures the curvature of the CEF. Here, though, the sign on that term changes if we include $\log(\sigma_t^2)$.

The obvious question is why is this? This is likely because the univariate regression on γ_t^2 suffers from the classic endogenous regressors problem. I showed in Table 2.7 that $\log(\sigma_t^2)$ and $\log(\gamma_t^2)$ are highly positively correlated. Since risk aversion implies that σ_t^2 commands a premia and the two volatilities are highly correlated, the univariate regression misattributes risk premia driven by risk aversion to news premia.[35]

This implies that the correctly specified regressions are the bivariate ones. In both cases we have that risk aversion results in diffusive risk commanding a large, positive premium. The news risk premium is substantially smaller in both cases.

We want to interpret the magnitude of the coefficients, not just the sign. Since I regress annualized excess log-return on the log total volatility and log jump proportion, the estimates are elasticities. These elasticities are highly statistically and economically significant. For example, consider the first row. The elasticity of rx_t with respect to $\sigma_t^2 + \gamma_t^2$ is 0.24. In other words, a 1 % increase in $\sigma_t^2 + \gamma_t^2$ for the course of an entire year increases the expected yearly return by 0.24 %.[36] For comparison, the average year-to-year difference in average $\sigma_t^2 + \gamma_t^2$ in my sample is $\approx$ 50 %. It increased by $\approx$ 150 % between 2007 and 2008.

The average annual absolute difference in $\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}$ is only 6.13 %, but the regression coefficient is significantly larger. A 1 % change in $\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}$ over the course of changes expected yearly returns by -5.01 %. In both cases, the implied movements in risk pre-

[35] The regression in terms of $\log(\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2})$ does not suffer from this exogeneity problem to as near a large extent because it is not nearly as heavily correlated.

[36] The reason that I only considered a 1 % change is that the approximation of log-differences as percent differences only holds for small changes.

mia from year to year are very large. I am not the first researcher to find movements in risk premia that are this large; Martin (2017) reports similarly-sized changes.

I consider several other specification in Section 2.D. The volatility coefficients are robust to the heteroskedasticity correction and the particular instruments chosen, (Table 2.18). Results from running the regression over a subsample either agree with the main results or are not statistically significant, (Table 2.20).

Robustness Checks

Estimating risk premia is difficult because the signal-to-noise ratio is quite low. The literature has pointed out some issues that can bias the empirical estimates. Perhaps the most important is the Stambaugh bias, (Stambaugh 1999). He shows that a finite-sample bias can inflate coefficient estimates if the regressors are stochastic. However, the regressions are run at the daily frequency, not the monthly frequency as is commonly done. Hence, I have approximately 3700 datapoints. Since this bias decreases at a $\frac{1}{\#\text{ datapoints}}$ rate, this bias should not noticeably affect my estimates.

The other significant sources of bias noted in the literature are also not nearly as significant here because I use daily data. For example, regressing long-horizon returns on persistent regressors causes the $\mathbb{R}^2$ to spuriously increase with the horizon under certain conditions. However, I am not using long-horizon returns, and so this does not apply. Various authors also have used overlapping returns to increase their effective sample size, which can invalidate the inference. I do not use overlapping returns, and so this also does not apply.

There is one primary source of error that is worth pointing out. The regressors that I use are estimated from high-frequency data. Consequently, we may have an error-in-regressors problem. This problem should not be a significant issue for three reasons. First, since I have a great deal of intraday data, the regressors should be estimated precisely. Second, the main empirical source of estimation error is separating the diffusion and jump components, and this should be independent of the expected returns because it only depends on the magnitude of the high-frequency returns, not their sign. Besides, it does not even affect estimating $\sigma_t^2 + \gamma_t^2$. Third, and most importantly, as I am instrumenting for the volatilities by their lags and the estimation error is likely independent across time, both the coefficient estimates and

their standard errors should be asymptotically valid.

2.11 Conclusion

This paper investigates how jumps affect investors' risk. I first show that standard no-arbitrage based pricing theory implies that jumps are price responses to news shocks. When a news shock hits causing the representative investor's information set to jump, she responds by pricing assets differently. Having done that, I introduce jump volatility — γ_t^2 — which is a sufficient statistic for the jump part of price dynamics. I then introduce the realized density, RD_t, to reduce tracking the returns' predictive density — $h\left(r_t \,|\, \mathcal{F}_{t-1}\right)$ – to forecasting γ_t^2 and σ_t^2. I do this by providing a new representation for infinite-activity jump processes as integrals with respect to a variance-gamma process. I then develop nonparametric estimators for the instantaneous and integrated jump and diffusion volatilities and for the realized density to enable taking these representations to the data.

I apply these estimators to the S&P 500 using high-frequency data from SPY. I find that jumps drive approximately one-half of the ex-post squared variation and that this proportion varies substantially over time. I also evaluate the performance of the estimators in simulations and find that my estimators perform well in estimating the volatilities. I then consider the behavior of these estimators in the data providing several new stylized facts. I show that the jump volatility is relatively well-behaved and has a bell-shaped distribution after applying a logarithmic transformation. In other words, the volatilities are roughly log-Gaussian. Finally, I show that γ_t^2 is both very persistent, having long-memory, and highly correlated with σ_t^2.

I next analyze how jumps affect expected returns. In particular, I show that risk premia have the following form — $-d\langle m, p\rangle(t) - d\left\langle m^{UP}, p^J\right\rangle$ — where $m(t)$ is the predictable part of the log-SDF and $m^{UP}(t)$ is the unpredictable part. I further relate $m(t)$ and $m^{UP}(t)$ to the curvature of the utility function and the certainty equivalence functional. The theory requires two factors that move at high-frequency in general. I show that the premium associated with γ_t^2 is statically and economically significantly less than the one associated with σ_t^2. This divergence implies that investors preferences are not time-separable and that the data require two factors that move

at high-frequency as well.

As this work introduces the jump volatility, a great deal of work still needs to be done. One prominent question is how to generalize the theory and empirics to higher dimensions. Can we derive a similar multivariate representation and estimators for the jump processes? Doing this will require figuring out what the appropriate multivariate Laplace distribution is. Presently, several multivariate Laplace distributions exist, but it is not apparent any of them have the proper relationship to Poisson and Gaussian processes. Moreover, this paper shows the proposed estimators are consistent in the noise-free case. Deriving the relevant inference theory in the presence of market microstructure noise would be quite useful.

Second, previous authors have shown that the stylized features of σ_t^2 are relatively stable across different assets. Is this also true for γ_t^2? For example, people have argued that news risk is fundamental in understanding foreign exchange markets. How does γ_t^2 act in those environments? Third, on the financial side, a great deal more empirical and theoretical work is needed to fully understand the relationship between the premia associated with σ_t^2 and γ_t^2. A fully specified general equilibrium model that determines the correct underlying risk factors would be useful to rationalize the new empirical evidence.

Fourth, since this paper reduces forecasting returns' distributions to forecasting the volatilities, it greatly simplifies tracking time-varying tail risk. Consequently, building a joint dynamic model for both volatilities and the drift and analyzing the resulting models' performance in tracking tail risk would be extremely useful.[37]

References

Ai, Hengjie, and Ravi Bansal. 2018. "Risk Preferences and the Macro Announcement Premium." *Econometrica* 86 (4): 1383–1430.

Aït-Sahalia, Yacine, Jianqing Fan, and Yingying Li. 2013. "The Leverage Effect Puzzle: Disentangling Sources of Bias at High Frequency." *Journal of Financial Economics* 109 (1): 224–249.

[37] This is the project I tackle in "Jumps, Tail Risk, and the Distribution of Stock Returns."

Aït-Sahalia, Yacine, and Jean Jacod. 2009a. "Estimating the Degree of Activity of Jumps in High Frequency Data." *The Annals of Statistics* 37 (5A): 2202–2244.

———. 2009b. "Testing for Jumps in a Discretely Observed Process." *The Annals of Statistics* 37 (1): 184–222.

———. 2012. "Analyzing the Spectrum of Asset Returns: Jump and Volatility Components in High Frequency Data." *Journal of Economic Literature* 50 (4): 1007–1050.

Aït-Sahalia, Yacine, Jean Jacod, and Jia Li. 2012. "Testing for Jumps in Noisy High Frequency Data." *Journal of Econometrics* 168:207–222.

Aït-Sahalia, Yacine, Per A. Mykland, and Lan Zhang. 2005. "How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise." *Review of Financial Studies* 18 (2): 351–416.

Andersen, Torben G., Tim Bollerslev, and Francis X. Diebold. 2007. "Roughing It Up: Including Jump Components in the Measurement, Modeling, and Forecasting of Return Volatility." *The Review of Economics and Statistics* 89 (4): 701–720.

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Heiko Ebens. 2001. "The Distribution of Realized Stock Return Volatility." *Journal of Financial Economics* 61 (1): 43–76.

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Paul Labys. 2001. "The Distribution of Realized Exchange Rate Volatility." *Journal of the American Statistical Association* 96 (453): 42–55.

———. 2003. "Modeling and Forecasting Realized Volatility." *Econometrica* 71 (2): 579–625.

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Clara Vega. 2003. "Micro Effects of Macro Announcements: Real-Time Price Discovery in Foreign Exchange." *The American Economic Review* 93 (1): 38–62.

———. 2007. "Real-time Price Discovery in Global Stock, Bond and Foreign Exchange Markets." *Journal of International Economics* 73 (2): 251–277.

Bakshi, Gurdip, Peter Carr, and Liuren Wu. 2008. "Stochastic Risk Premiums, Stochastic Skewness in Currency Options, and Stochastic Discount Factors in International Economies." *Journal of Financial Economics* 87 (1): 132–156.

Bandi, Federico M., and Roberto Renò. 2012. "Time-varying Leverage Effects." *Journal of Econometrics* 169 (1): 94–113.

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

Barlow, Martin T. 1978. "Study of a Filtration Expanded to Include an Honest Time." *Probability Theory and Related Fields* 44 (4): 307–323.

Barndorff-Nielsen, Ole E., and Neil Shephard. 2002. "Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 64 (2): 253–280.

———. 2004. "Power and Bipower Variation with Stochastic Volatility and Jumps." *Journal of Financial Econometrics* 2 (1): 1–37.

———. 2006. "Econometrics of Testing for Jumps in Financial Economics Using Bipower Variation." *Journal of Financial Econometrics* 4 (1): 1–30.

Barndorff-Nielsen, Ole E., and Albert Nikolaevich Shiryaev. 2010. "Change of Time and Change of Measure." In *Advanced Series on Statistical Science & Applied Probability,* edited by Ole E. Barndorff-Nielsen, vol. 13. Toh Tuck Link, Singapore: World Scientific.

Beechey, Meredith J., and Jonathan H. Wright. 2009. "The High-Frequency Impact of News on Long-Term Yields and Forward Rates: Is it Real?" *Journal of Monetary Economics* 56 (4): 535–544.

Bollerslev, Tim. 1986. "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics* 31 (3): 307–327.

Bollerslev, Tim, Robert F. Engle, and Jeffrey M. Wooldridge. 1988a. *Journal of Political Economy* 96 (1): 116–131.

Bollerslev, Tim, Robert F. Engle, and Jeffrey M. Wooldridge. 1988b. "A Capital Asset Pricing Model with Time-Varying Covariances." *Journal of Political Economy* 96 (1): 116–131.

Bollerslev, Tim, Tzuo Hann Law, and George Tauchen. 2008. "Risk, Jumps and Diversification." *Journal of Econometrics* 144:234–256.

Bollerslev, Tim, Andrew J. Patton, and Rogier Quaedvlieg. 2016. "Exploiting the Errors: A Simple Approach for Improved Volatility Forecasting." *Journal of Econometrics* 192 (1): 1–18.

Brandt, Michael W., and Qiang Kang. 2004. "On the Relationship Between the Conditional Mean and Volatility of Stock Returns: A Latent VAR Approach." *Journal of Financial Economics* 72 (2): 217–257.

Branger, Nicole, Christian Schlag, and Eva Schneider. 2008. "Optimal Portfolios when Volatility Can Jump." *Journal of Banking & Finance* 32 (6): 1087–1097.

Breeden, Douglas T. 1979. "An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities." *Journal of Financial Economics* 7 (3): 265–296.

Campbell, John Y. 1987. "Stock Returns and the Term Structure." *Journal of Financial Economics* 18 (2): 373–399.

Christensen, Kim, Roel C.A. Oomen, and Mark Podolskij. 2014. "Fact or friction: Jumps at ultra high frequency." *Journal of Financial Economics* 114 (3): 576–599.

Corsi, Fulvio. 2009. "A Simple Approximate Long-Memory Model of Realized Volatility." *Journal of Financial Econometrics* 7 (2): 174–196.

Dambis, K. E. 1965. "On the Decomposition of Continuous Submartingales." *Theory of Probability and its Applications* 10 (3): 401–10.

Delbaen, Freddy, and Walter Schachermayer. 1994. "A General Version of the Fundamental Theorem of Asset Pricing." *Mathematische Annalen* 300 (1): 463–520.

Dickey, David A., and Wayne A. Fuller. 1981. "Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root." *Econometrica:* 1057–1072.

Drechsler, Itamar, and Amir Yaron. 2011. "What's Vol Got to Do with It." *The Review of Financial Studies* 24 (1): 1–45.

Dubins, Lester E., and Gideon Schwarz. 1965. "On Continuous Martingales." *Proceedings of the National Academy of Sciences of the United States of America* 53 (5): 913–916.

Duffie, Darrell, and Larry G. Epstein. 1992. "Stochastic Differential Utility." *Econometrica* 60 (2): 353–394.

Engle, Robert F. 1982. "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation." *Econometrica:* 987–1007.

Engle, Robert F., and Victor K. Ng. 1993. "Measuring and Testing the Impact of News on Volatility." *The Journal of Finance* 48 (5): 1749–1778.

Epps, Thomas W., and Mary Lee Epps. 1976. "The Stochastic Dependence of Security Price Changes and Transaction Volumes: Implications for the Mixture-of-Distributions Hypothesis." *Econometrica* 44 (2): 305–321.

Epstein, Larry G., and Martin Schneider. 2003. "Recursive Multiple-Priors." *Journal of Economic Theory* 113 (1): 1–31.

Epstein, Larry G., and Stanley E. Zin. 1989. "Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework." *Econometrica* 57 (4): 937–969.

Ergin, Haluk, and Faruk Gul. 2009. "A Theory of Subjective Compound Lotteries." *Journal of Economic Theory* 144 (3): 899–929.

Faust, Jon, John H. Rogers, Shing-Yi B. Wang, and Jonathan H. Wright. 2007. "The High-Frequency Response of Exchange Rates and Interest Rates to Macroeconomic Announcements." *Journal of Monetary Economics* 54 (4): 1051–1068.

Gallant, A. Ronald, and George Tauchen. 2018. "Exact Bayesian Moment Based Inference for the Distribution of the Small-time Movements of an Itô Semimartingale." Indirect Estimation Methods in Finance and Economics, *Journal of Econometrics* 205 (1): 140–155.

Geman, Hélyette, Dilip B. Madan, and Marc Yor. 2002. "Stochastic Volatility, Jumps and Hidden Time Changes." *Finance and Stochastics* 6 (1): 63–90.

Geweke, John, and Susan Porter-Hudak. 1983. "The Estimation and Application of Long Memory Time Series Models." *Journal of Time Series Analysis* 4 (4): 221–238.

Ghysels, Eric, Pedro Santa-Clara, and Rossen Valkanov. 2005. "There is a Risk-Return Trade-off after All." *Journal of Financial Economics* 76 (3): 509–548.

Gilboa, Itzhak, and David Schmeidler. 1989. "Maxmin Expected Utility with Non-Unique Prior." *Journal of Mathematical Economics* 18 (2): 141–153.

Glosten, Lawrence R., Ravi Jagannathan, and David E. Runkle. 1993. "On the Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks." *The Journal of Finance* 48 (5): 1779–1801.

Gürkaynak, Refet S., Burçin Kısacıkoğlu, and Jonathan H. Wright. 2018. *Missing Events in Event Studies: Identifying the Effects of Partially-Measured News Surprises.* Working Paper, Working Paper Series 25016. National Bureau of Economic Research, September.

Hansen, Lars Peter, and Thomas J. Sargent. 2001. "Robust Control and Model Uncertainty." *The American Economic Review* 91 (2): 60–66.

Harvey, Campbell R. 1989. "Time-Varying Conditional Covariances in Tests of Asset Pricing Models." *Journal of Financial Economics* 24 (2): 289–317.

Hautsch, Nikolaus, and Mark Podolskij. 2013. "Preaveraging-based Estimation of Quadratic Variation in the Presence of Noise and Jumps: Theory, Implementation, and Empirical Evidence." *Journal of Business & Economic Statistics* 31 (2): 165–183.

Huang, Xin, and George Tauchen. 2005. "The Relative Contribution of Jumps to Total Price Variance." *Journal of Financial Econometrics* 3 (4): 456–499.

Jacod, Jean, Yingying Li, Per A. Mykland, Mark Podolskij, and Mathias Vetter. 2009. "Microstructure Noise in the Continuous Case: The Pre-averaging Approach." *Stochastic Processes and their Applications* 119 (7): 2249–2276.

Jacod, Jean, Mark Podolskij, and Mathias Vetter. 2010. "Limit Theorems for Moving Averages of Discretized Processes Plus Noise." *The Annals of Statistics* 38 (3): 1478–1545.

Jacod, Jean, and Phillip Protter. 2012. "Discretization of Processes." In *Stochastic Modelling and Appplied Probability,* edited by Boris Rozovksiĭ and Peter W. Glynn, vol. 67. Berlin, Germany: Springer-Verlag.

Jacod, Jean, and Mathieu Rosenbaum. 2013. "Quarticity and Other Functionals of Volatility: Efficient Estimation." *Annals of Statistics* 41 (4): 1462–1484.

Ju, Nengjiu, and Jianjun Miao. 2012. "Ambiguity, Learning, and Asset Returns." *Econometrica* 80 (2): 559–591.

Kalnina, Ilze, and Dacheng Xiu. 2017. "Nonparametric Estimation of the Leverage Effect: A Trade-Off Between Robustness and Efficiency." *Journal of the American Statistical Association* 112 (517): 384–396.

Klibanoff, Peter, Massimo Marinacci, and Sujoy Mukerji. 2005. "A Smooth Model of Decision Making under Ambiguity." *Econometrica* 73 (6): 1849–1892.

Kozubowski, Tomasz J., and Rachev T. Svetlozar. 1994. "The Theory of Geometric Stable Distributions and Its Use in Modeling Financial Data." *European Journal of Operational Research* 74 (2): 310–324.

Kreps, David M., and Evan L. Porteus. 1978. "Temporal Resolution of Uncertainty and Dynamic Choice Theory." *Econometrica* 46 (1): 185–200.

Kwiatkowski, Denis, Peter C.B. Phillips, Peter Schmidt, and Yongcheol Shin. 1992. "Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root: How Sure Are We That Economic Time Series Have a Unit Root?" *Journal of Econometrics* 54 (1–-3): 159–178.

Lahaye, Jérôme, Sébastien Laurent, and Christopher J. Neely. 2011. "Jumps, Cojumps and Macro Announcements." *Journal of Applied Econometrics* 26 (6): 893–921.

Lettau, Martin, and Sydney C. Ludvigson. 2010. "Measuring and Modeling Variation in the Risk-Return Tradeoff." Chap. 11, edited by Yacine Aït-Sahalia and Lars Peter Hansen, 1:617–690. Handbook of Financial Econometrics. Elsevier.

Li, Jia, Viktor Todorov, and George Tauchen. 2017. "Jump Regressions." *Econometrica* 85 (1): 173–195.

Liu, Lily Y., Andrew J. Patton, and Kevin Sheppard. 2015. "Does Anything Beat 5-minute RV? A Comparison of Realized Measures Across Multiple Asset Classes." *Journal of Econometrics* 187 (1): 293–311.

Lord, Roger, Remmert Koekkoek, and Dick Van Dijk. 2010. "A Comparison of Biased Simulation Schemes for Stochastic Volatility Models." *Quantitative Finance* 10 (2): 177–194.

Lucca, David O., and Emanuel Moench. 2015. "The Pre-FOMC Announcement Drift." *The Journal of Finance* 70 (1): 329–371.

Madan, Dilip B., Peter P. Carr, and Eric C. Chang. 1998. "The Variance Gamma Process and Option Pricing." *Review of Finance* 2 (1): 79.

Mancini, Cecilia. 2001. "Disentangling the Jumps of the Diffusion in a Geometric Jumping Brownian Motion." *Giornale dell'Istituto Italiano degli Attuari* 64:19–47.

Martin, Ian. 2017. "What is the Expected Return on the Market?" *The Quarterly Journal of Economics* 132 (1): 367–433.

Medvegyev, Peter. 2007. *Stochastic Integration Theory.* Edited by R. Cohen, S.K. Donaldson, S. Hildebrant, T.J. Lyons, and M.J. Taylor. Oxford Graduate Texts in Mathematics. New York: Oxford University Press.

Merton, Robert C. 1973. "An Intertemporal Capital Asset Pricing Model." *Econometrica* 41 (5): 867–887.

Mittnik, Stefan, and Rachev T. Svetlozar. 1993. "Modeling Asset Returns with Alternative Stable Distributions." *Econometric Reviews* 12 (3): 261–330.

Monroe, Itrel. 1978. "Processes that can be Embedded in Brownian Motion." *The Annals of Probability* 6 (1): 42–56.

Nelson, Daniel B. 1991. "Conditional Heteroskedasticity in Asset Returns: A New Approach." *Econometrica:* 347–370.

Neuberger, Anthony. 2012. "Realized Skewness." *The Review of Financial Studies* 25 (11): 3423.

Newey, Whitney K., and Daniel McFadden. 1994. "Large Sample Estimation and Hypothesis Testing." Chap. 36, edited by Robert Engle and Daniel McFadden, 4:2111–2245. Handbook of Econometrics. Elsevier.

Newey, Whitney K., and Kenneth D. West. 1994. "Automatic Lag Selection in Covariance Matrix Estimation." *The Review of Economic Studies* 61 (4): 631–653.

Nikeghbali, Ashkan. 2007. "Non-stopping Times and Stopping Theorems." *Stochastic Processes and their Applications* 117 (4): 457–475.

Oomen, Roel C. A. 2006. "Comment." *Journal of Business & Economic Statistics* 24 (2): 195–202.

Ornthanalia, Chayawat. 2014. "Lévy Jump Risk: Evidence from Options and Returns." *Journal of Financial Economics* 112 (1): 69–90.

Pagan, Adrian R., and Y.S. Hong. 1991. "Nonparametric Estimation and the Risk Premium." Chap. 2 in *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics,* edited by William A. Barnett, James Powell, and George Tauchen, 51–75. Cambridge University Press.

Pan, Jun. 2002. "The Jump-Risk Premia Implicit in Options: Evidence from an Integrated Time-Series Study." *Journal of Financial Economics* 63 (1): 3–50.

Podolskij, Mark, and Mathias Vetter. 2009. "Bipower-type Estimation in a Noisy Diffusion Setting." *Stochastic Processes and Their Applications* 119 (9): 2803–2831.

Reisen, Valderio A. 1994. "Estimation of the Fractional Difference Parameter in the ARIMA(p, d, q) Model Using the Smoothed Periodogram." *Journal of Time Series Analysis* 15 (3): 335–350.

Santa-Clara, Pedro, and Shu Yan. 2010. "Crashes, Volatility, and the Equity Preimum: Lessons from S&P 500 Options." *The Review of Economics and Statistics* 92 (2): 435–451.

Stambaugh, Robert F. 1999. "Predictive Regressions." *Journal of Financial Economics* 54 (3): 375–421.

Strzalecki, Tomasz. 2013. "Temporal Resolution of Uncertainty and Recursive Models of Ambiguity Aversion." *Econometrica* 81 (3): 1039–1074.

Todorov, Viktor. 2010. "Variance Risk-Premium Dynamics: The Role of Jumps." *Review of Financial Studies* 23 (1): 345–383.

———. 2011. "Econometric Analysis of Jump-Driven Stochastic Volatility Models." *Journal of Econometrics* 160 (1): 12–21.

Todorov, Viktor, and George Tauchen. 2014. "Limit Theorems for the Empirical Distribution Function of Scaled Increments of Itô Semimartingales at High Frequencies." *The Annals of Applied Probability* 24, no. 5 (October): 1850–1888.

Tsai, Jerry, and Jessica A. Wachter. 2018. *Pricing Long-Lived Securities in Dynamic Endowment Economies.* Working Paper, Working Paper Series 24641. National Bureau of Economic Research, May.

Yu, Jun. 2005. "On Leverage in a Stochastic Volatility Model." *Journal of Econometrics* 127 (2): 165–178.

2.A Representation Theorems

Theorem 2.2 (Jump Volatility and the Predictable Quadratic Variation)**.** *Let $p(t)$ be an Itô semimartingale satisfying Assumption Square-Integrable, then the following holds where $\langle p^J \rangle(t)$ is the predictable quadratic variation (angle-bracket) of $p^J(t)$:*

$$\gamma_t^2 = \int_{t-1}^{t} \gamma^2(s)\, ds = \int_{t-1}^{t} \int_{\mathcal{X}} \delta^2(s,x)\nu(dx,ds) = \langle p^J \rangle(t) - \langle p^J \rangle(t-1).$$

Proof.

$$[p]^J(t) = \sum_{s \le t} \Delta p(s)^2 = \int_0^t \int_{\mathcal{X}} \delta^2(s,x)\mu(ds,dx) \tag{2.49}$$

This comes from the view of the jumps as integrals with respect to Poisson random measures and there being no predictable jumps. Intuitively, the compensator ν does not jump and realizations of μ are equal 1 which does not change when squared.

$$\begin{aligned} \implies \langle p \rangle^J(t) = \mathbb{E}\left[[p]^j(t) \,\middle|\, \mathcal{F}_{t-} \right] &= \mathbb{E}\left[\int_0^t \int_X \delta^2(s,x)\mu(ds,dx) \,\middle|\, \mathcal{F}_{t-} \right] \\ &= \int_0^t \int_X \delta^2(s,x)\nu(ds,dx) \end{aligned} \tag{2.50}$$

We also need to show that the limit in the expectation form approaches $\gamma^2(t)$.

Define $\gamma^2(t) \coloneqq \int_X \delta^2(t,x)\nu(dx,dt)$, then

$$\lim_{\Delta \to 0} \frac{1}{\Delta} \mathbb{E}\left[\left| p^J(t+\Delta) - p^J(t) \right|^2 \,\middle|\, \mathcal{F}_{t-} \right] = \lim_{\Delta \to 0} \frac{1}{\Delta} \mathbb{E}\left[\left| \int_t^{t+\Delta} \delta(s,x)(\mu - \nu)(ds\, dx) \right|^2 \,\middle|\, \mathcal{F}_{t-} \right]. \tag{2.51}$$

By the Itô Isometry, we can rewrite (2.51) as

$$= \lim_{\Delta \to 0} \frac{1}{\Delta} \mathbb{E}\left[\int_t^{t+\Delta} \delta^2(s,x)\mu(ds\, dx) \,\middle|\, \mathcal{F}_{t-} \right]. \tag{2.52}$$

Then by choosing δ so that dx, ds are independent, and the projection of ν onto the Lebesgue measure is constant. We have

$$= \lim_{\Delta \to 0} \frac{1}{\Delta} \mathbb{E}\left[\int_t^{t+\Delta} \int_{\mathcal{X}} \delta^2(s,x)(dx\,dx) \,\middle|\, \mathcal{F}_{t-} \right] \tag{2.53}$$

$$= \lim_{\Delta \to 0} \frac{1}{\Delta} \mathbb{E}\left[\Delta\gamma^2(t) + \int_t^{t+\Delta} \left(\gamma^2(t) - \gamma^2(s)\right) ds \,\middle|\, \mathcal{F}_{t-} \right]. \tag{2.54}$$

We can split this into the value of the jump volatility at t and deviations from it:

$$= \lim_{\Delta \to 0} \gamma^2(t) + \frac{1}{\Delta} \Delta O\left(\mathbb{E}\left[\left| \sup_{t \le s \le t+\Delta} \gamma^2(t) - \gamma^2(s) \right| \,\middle|\, \mathcal{F}_{t-} \right] \right) = \gamma^2(t). \tag{2.55}$$

□

Theorem 2.4 (Time-Changing Jump Martingales)**.** *Let $p^J(t)$ be a purely discontinuous, martingale with full support satisfying Assumptions Square-Integrable, Infinite-Activity Jumps, and No Predictable Jumps that can be represented as $H * (n - \nu)$ where $H(t)$ is a predictable process, n a Poisson random measure, and ν its predictable compensator with Lebesgue base Levy measure.*

Then $p^J(t)$ time-changed by its predictable quadratic variation is a standard variance-gamma process. In other words, $p^J(t) \overset{\mathcal{L}}{=} \mathcal{L}\left(\langle p^J \rangle(t)\right)$.[38]

Proof. To prove the result, I start with a representation of a purely-discontinuous martingale as an integral with respect to a Poisson random measure. This is a two-dimensional representation of the jump process with all of the dynamics contained in the predictable process H. There are two key parts to the theorem above. First, we must handle the dynamics contained in H, and second we need to reduce the two-dimensional representation to a one-dimensional one.

We know that there are only finitely-many jumps in any strip that is bounded away from 0, but infinitely-many in any interval containing 0. To maintain this intuition, I switch the base Levy measure to one that has this property. Second, I make a time-change argument in each strip to deal with its dynamics. Third, I switch from an integral with respect to infinitely-many Poisson processes to one with respect

[38] Note, the equality here only holds in law unlike in the Dambis-Dubins-Schwarz theorem, where it holds almost surely.

to a Poisson random measure by taking the appropriate sum of these processes. I use capital letters to refer to processes as is standard in the literature. Since I do not discretize, there should be no confusion. Define $1^z = 1\{x \in [z, z + dz]\}$ where $z \in \mathbb{R}$, and $dz \in \mathbb{R}_+$, where I suppress dz in the notation. Similarly, for a process X define $X^z = X * 1^z$. In words, X^z is the process X restricted to the strip $[z, z = dz]$.

Denote $p(t)$ by $Y(t)$. I now switch the representation of Y as an integral with respect to a Poisson random measure with more intuitive properties. Y is locally-square integrable, and hence $\langle Y \rangle$ is well-defined, that is for any stopping-time τ, the stopped-process $\langle Y \rangle^\tau$ is almost surely finite. Since Y is a purely-discontinuous process, Y^z is a two-dimensional sum. To put in mathematical notation, $(Y^z)^\tau = \sum_{s \leq t, x \in [z, z+dz]} \delta(x, s)$, where δ is a predictable Dirac delta. Also, define $\langle X \rangle^{-1}(t) = \inf\{\tau : \langle X \rangle = t\}$ for any process X. This is the standard inverse definition when the process may be zero, and is innocuous here because if $\langle X \rangle \overset{a.s.}{=} 0 \implies X \overset{a.s.}{=} 0$.

Recall that I assumed that the base measure of μ was the Lebesgue measure λ. $\left(\frac{1}{z}\right)$ is an infinite-measure and is absolutely continuous with respect to Lebesgue measure in any interval not containing zero. Let $\tilde{\mu}$ be a Poisson random measure with associated Levy measure $\left(\frac{1}{z}\right)$. Throughout the rest of proof, I use tilde's to refer to measures associated with this random measure. For example, $\tilde{\nu}$ is its associated compensator. Note, since I am using compensated random measures, each strip $[z, z + dz]$ is a martingale.

The benefit of using this representation is that it implies the associated predictable integrator, $\widetilde{H}$, is $O_p(1)$. In the original case, the local square-integrability of Y implies that $H(x, t)$ as a function of x is $O_p\left(\frac{1}{x}\right)$. Effectively, I am moving the necessary reduction in the intensity of the process as the jump size increases into the Poisson random measure instead of the integrator.

It is worth noting that in general we cannot choose $\tilde{H}$ to be proportional to a constant; it might be zero. However, since we have an infinite-activity process, we can without loss of generality. In addition, the Poisson processes formed by restricting the Poisson random measure to a strip in $\mathbb{R}$, $\tilde{X}^z$, have intensity measures, $\nu(x) = x^{-1} \exp(-x)\, dx$, which I use in the sequel.

I now turn to using a time-change argument to handle the dependence of Y^z, or equivalently, H^z. Since $\tilde{X}^z$ is a finite-activity Poisson process, its intrinsic filtration

is the filtration generated by the jump locations. Let $t^{\tilde{X}^z}$ be a jump time for the process $\tilde{X}^z$, and consider the set $\{t < t^{\tilde{X}^z}\}$. This set is optional, but not predictable, and its ending time $\hat{t}$, is not a stopping time with respect to the predictable filtration. (It is what is known in the literature as an honest time.) This allows us to define the minimal enlargement of the filtration of Y^z, $\mathcal{F}_t^{Y^z}$ so that the $\hat{t}$ are stopping times.

$$\hat{\mathcal{F}}_t^{X^n} := \cap_{\epsilon>0} \mathcal{F}_{t-\epsilon}^{\tilde{X}^z} \cup \sigma(\{\rho < t\}) \tag{2.56}$$

It is worth noting that when you progressively enlarge a filtration with an honest time, semimartingales with respect to the original filtration are still semimartingales with respect to the new filtration (Barlow 1978). However, this enlargement does not necessarily preserve the martingale structure. Since I am doing this almost surely only finitely many times and jumps of the original process are almost surely unique, it is without loss of generality to consider the case with only one jump.

Consider X stopped at some time ρ that is a stopping time with respect to the expanded filtration, not to the original one. It is worth noting that we are expanding the predictable filtration $\mathcal{F}_{t-}$, not the original filtration. So using the Nikeghbali (2007, eqn 2.3), we can define the martingale on the new space. Then $X(t)$ has the following form, where $Z_t^\rho := \Pr[\rho > t \,|\, \mathcal{F}_{t-}]$, is chosen by to be cádlág. The $\mathcal{F}_t$ dual optional projection of the process $1\{\rho \leq t\}$ is denoted by $A^\rho(t)$. Let $\hat{X}$ be a martingale with respect to $\widehat{\mathcal{F}}_t$. Also, define $\mu^\rho(t) = \mathbb{E}[A^\rho(\infty) \,|\, \mathcal{F}_{t-}] = A^\rho(t) + Z^\rho(t)$. Then

$$X(t) = \widehat{X}(t) + \int_0^{t\wedge\rho} \frac{d\langle X, \mu^\rho(t)\rangle(s)}{Z^\rho(s-)} - \int_\rho^t \frac{d\langle X, A^\rho(t) + Z^\rho(t)\rangle(s)}{1 - Z^\rho(s-)}. \tag{2.57}$$

Since $\mu^\rho(t)$ is $\mathcal{F}_{t-}$ measurable, and the jumps are distributed according to a Poisson process, $\mu^\rho(t)$ is a constant. Consequently, the predictable quadratic variation terms in (2.57) terms are almost surely zero.

Consider the process $\hat{X}^z$, where $\tilde{X}^z$ and $\hat{X}^z$ are equal pathwise, but we change the filtration from $\mathcal{F}_t$ to $\hat{\mathcal{F}}_t$. Since the stopping times of $\widetilde{X}^z$ are sufficient to generate its filtration, and $\widetilde{H}$ is predictable, we can choose $\widehat{\mathcal{F}}_t^{X^z}$ to be generated by the predictable σ-algebra. Equivalently, it is generated by the continuous processes. As a result, for any process adapted to this filtration there exists a continuous process that is equal to it in probability. Since equality in distribution is weaker than equality in probability, it

is without loss of generality to assume that the process is continuous in this filtration, and so I do so.

By the Dambis, Dubins & Schwarz theorem, we know that a continuous process is a Wiener process when time-changed by its quadratic variation. Therefore, $\hat{Y}^z([Y^z]) \stackrel{\mathcal{L}}{=} W$, where W is the standard Wiener process. Intuitively, we can view the jump magnitudes as appropriately rescaled Gaussian random variables.

However, this is not the filtration generated by the data, and so we need to consider what this representation implies about the original filtration. We start by considering the precise relationship between the predictable and quadratic variations both within and between each of the filtrations.

$\langle \hat{Y}^z \rangle \stackrel{a.s.}{=} [\hat{Y}^z]$ because all of the adapted processes in $\hat{\mathcal{F}}_t$ are predictable. In addition, changing the filtration does not change the quadratic variation because the process is optional and adapted, and all the change of filtration is doing is turning optional processes into predicable ones.

Therefore, the key question is what is the relationship between the $\langle Y^z \rangle$ and $[Y^z]$ in the original filtration. The quadratic variation of an integral with respect to a finite-activity Poisson process is $[H^z * X^z] = \sum_{s \leq t} (H^Z)^2(\hat{\tau})$, where the $\hat{\tau}$ are the jump locations.

Since $\tilde{X}^z$ is a Poisson process, the amount of time between jumps, that is the length of the intervals define above, is an exponential random variable with intensity $\hat{\nu}^z$. Since $\hat{\nu}^z$ is a deterministic function, $\hat{F}^n$ is an exponential-time change of $\tilde{\mathcal{F}}$. Therefore, $Y^z = H^z * X^z$ is Wiener process after both an exponential time-change and then a continuous-time change in the transformed space.

There are two main limitations of this result. First, the exponential time-change is not identified, and so we cannot use it for inference. Second, we want an expression for Y not just for each of the Y^z.

The first problem can be resolved by recalling that if the expectations of a sufficiently general class of functions are the same between two processes, then the processes equal in distribution. A sequence of nested expectations does not change if we reorder the nesting as long as the σ-algebras we are conditioning on are independent. However, because the exponential-time change was with respect to a Poisson process with a deterministic compensator and the other time-change was with respect

to a predictable process, the relevant filtrations are independent here. Consequently, we have that if we time-change Y^z by $\langle Y^z \rangle$, then we have a Wiener process with an exponential subordinator.

To resolve the second problem, that is aggregate over the strips correctly, note what happens if we aggregate all of the μ^z together. $\tilde{\mu}^z$ is a Poisson random variable with intensity measure $\tilde{\nu}(z) = z^{-1}\exp(-z)dz$. However, the definition of the Gamma process is that its intensity measure over strips is precisely the expression above.

For a countable partition of $\mathbb{R}$, $z_1, z_2 \ldots$, $Y = \sum_{z_i} Y^{z_i}$, and $\tilde{H} = \sum_{z_i} \tilde{H}^{z_i}$, and $\tilde{\mu} = \sum_{z_i} \tilde{\mu}^{z_i}$. Furthermore, Wiener processes are stable under countable sums as long as the variance remains finite, which it will in this context because the initial process is locally-square integrable. Consequently, we can do the following.

$$\lim_{I\to\infty} \sum_{i\leq I} Y^z \left(\langle Y^z \rangle^{-1}\right) \stackrel{\mathcal{L}}{=} \lim_{I\to\infty} W(\exp^{\sum \nu^i}) = W(\Gamma(t)) = \mathcal{L} \tag{2.58}$$

To wrap it up, if we time-change a purely-discontinuous, jump process with infinite-variation by its predictable quadratic variation, we get the variance-gamma process, also known as a standard variance-gamma process.

□

Corollary 2.2 (Time-Changing Finite-Activity Jump Martingales)**.** *Let $p^J(t)$ be a purely discontinuous martingale with full support satisfying Assumptions Square-Integrable and No Predictable Jumps that can be represented as $H * (n - \nu)$ where $H(t)$ is a predictable process, n a Poisson random measure, and ν its predictable compensator with Lebesgue base Levy measure.*

Let $\langle p^J \,|\, n(t) \rangle$ be the predictable quadratic variation of p^J where additionally we condition on all the jumps occurring up to and including at time t. Then $p^J(t)$ time-changed by $\langle p^J \,|\, n(t) \rangle$ is a mixture of the 0 process — δ_0 — and the standard variance-gamma process where the mixing weights are the intensity of the jump process.

Proof. Since $p^J(t)$ is a finite-activity jump process, $(\mu - \nu) * 1^z$ is almost-surely zero as a function of z for all but a finite-subset of $\mathbb{R}$. For a segment of time when there are no jumps, the process is identically 0. You cannot time-change a process by the 0 process. Therefore, the proof of the main theorem where we take the limit of the number of strips to infinity is no longer valid.

Consider some interval Δ. Partition the event-space Ω into spaces where $[p^J]$ is positive when $t \in \Omega$, and spaces where it is not. In the first subset, we can make the same argument I made above extending the space if necessary. In the second space, we have the 0 process, δ_0. Since we have a finite-activity process driven by a Poisson random measure, there exists a compound Poisson process, $N(t)$ driving this process where $N(t) = 1$ if and only if one of the processes jumps.[39] Define $\left\langle p^J \,\middle|\, N(t) \right\rangle$ to be the predictable quadratic variation of p^J conditional on the times τ that $N(\tau) = 1$ for $\tau <= t$. Note, this is well-defined because it is well-defined for any interval Δ, and we can use Kolmogorov's extension theorem. In the space of intervals Δ that contain a jump, we can make the argument I made in the infinite-activity case. Otherwise, we just have the zero process.

In addition, it is innocuous to time-change the zero process because diving zero by a positive number is still zero.

$$p^J(t) = \begin{cases} \mathcal{L}\left(\left\langle p^J \,\middle|\, N(t) \right\rangle(t)\right) & \text{with intensity } \nu \\ \delta_0(t) & \text{with intensity } 1 - \nu \end{cases} \tag{2.59}$$

□

Corollary 2.1 (Jumps Processes as Integrals)**.** *Let $p^J(t)$ be a Itô semimartingale with full support satisfying Assumptions Square-Integrable, Infinite-Activity Jumps, and No Predictable Jumps. Then $p^J(t) = \frac{1}{\sqrt{2}} \int_0^t \gamma(s)\, d\mathcal{L}(s)$, where $\mathcal{L}$ is a standard variance-gamma process.*

Proof. Since $Y(t)$ is an Itô semimartingale, $Y(t) = \int_0^t \int_{\mathbb{R}} \delta(s,x)\, ds$, where I use standard notation. This implies that its predicable quadratic variation, $K(t)$, equals $\int_0^t \int_{\mathbb{R}} \delta(s,x)^2\, dx\, ds$, with time-derivative $k(t)$ equal to $\int_{\mathbb{R}} \delta^2(s,x)\, dx$.

Let $J(t)$ be the purely-discontinuous martingale part of $Y(t)$, then Theorem 2.4 implies that $J(\langle Y \rangle^{-1})(t) \overset{\mathcal{L}}{=} \mathcal{L}(t)$, or equivalently, $J(t) \overset{\mathcal{L}}{=} \int_0^{\langle Y \rangle^{-1}(t)} d\mathcal{L}(s)$. Then since $\frac{k}{\sqrt{2}}\mathcal{L}(1) = \mathcal{L}(k^2)$, where $\mathcal{L}(1)$ is a standard Laplace random variable, and $k(t)$ is a predicable process (and so independent of $\mathcal{L}$), $J(t) = \int_0^t k(s)\, d\mathcal{L}(s)$. This is completely

[39] Since the Gaussian distribution is a stable distribution and the Laplace distribution is geometrically-stable distribution, conditioning on one jump and multiple jumps is equivalent.

analogous to how the time-changed theorem for continuous processes and absolute continuity together imply the integral representation of continuous martingales.

□

Theorem 2.5 (Realized Density Representation)**.** *Let $p(t)$ be an Itô semimartingale with full support satisfying Assumptions Square-Integrable, Infinite-Activity Jumps, and No Predictable Jumps. Let $\sigma^2(t)$ and $\gamma^2(t)$ be semimartingales whose martingale components are independent of the martingale components of $p(t)$. Then*

$$RD_t = \mathcal{N}\left(\int_{t-1}^{t} \mu(s)\,ds, \int_{t-1}^{t} \sigma^2(s)\,ds\right) * \mathcal{L}\left(0, \int_{t-1}^{t} \gamma^2(s)\,ds\right), \tag{2.12}$$

and the predictive density is

$$h\left(r_t \,|\, \mathcal{F}_{t-1}\right) = \int_{\mu_t,\sigma_t^2,\gamma_t^2} RD_t(\mu_t, \sigma_t^2, \gamma_t^2)\, dG\left(\mu_t, \sigma_t^2, \gamma_t^2 \,\middle|\, \mathcal{F}_{t-1}\right). \tag{2.13}$$

Proof. Consider the diffusion part of the process.

$$h\left(p^D(t) - p^D(t-1) \,\middle|\, \mathcal{F}_{t-1}\right) = h\left(\sum_{n\in\frac{1}{\Delta},\ldots,0} \int_{t-n\Delta}^{t-(n+1)\Delta} \sigma(s)\,dW(s) \,\middle|\, \mathcal{F}_t\right) \tag{2.60}$$

If Δ is small enough, we can pull $\sigma^2(t)$ out of the integral because requiring the integrand to be predictable does not affect the value of the process.

$$= h\left(\sum_{n\in 1,\frac{1}{\Delta}} \sigma(t-n\Delta) \int_{t-n\Delta}^{t-(n+1)\Delta} dW(s) \,\middle|\, \mathcal{F}_{t-1}\right) \tag{2.61}$$

Since the martingale components of $\sigma^2(t)$ are independent of W, we can condition on the entire path of $\sigma^2(t)$ without affecting the distribution of the increments of W.

$$\overset{\mathcal{L}}{=} h\left(\sqrt{\int_{t-1}^{t} \sigma^2(s)\,ds} \int_{t-1}^{t} dW(s) \,\middle|\, \mathcal{F}_{t-1}\right) \tag{2.62}$$

$$\overset{\mathcal{L}}{=} \int_{\sigma_t^2} N\left(0, \int_{t}^{t+\Delta} \sigma^2(s)\,ds\right) dG\left(\sigma_t^2 \,\middle|\, \mathcal{F}_{t-1}\right) \tag{2.63}$$

The argument for the jump volatility follows mutatis mutandis. The only real difference is that the scale (the expectation of the absolute deviation) of the Laplace

distribution is the square-root of one-half the variance. Consequently, when you pull the variance outside of the integral, you get an additional $\sqrt{2}$ in the denominator.

You can just carry the mean through the analysis, and then add it back in when you are done. To combine the jump and diffusion realized densities, note that density of independent variables are convolutions of the densities. The integrators are pure-jump and diffusive martingales, and so they are automatically orthogonal. Consequently, the jump and diffuse parts are independent conditional on the drift and the volatilities. Also, I derived RD_t in the argument above because it is simply the function inside the integral.

□

2.B Volatility Estimation

Lemma 2.6 (HL implies SHL)**.** *If an Itô semimartingale* $p(t)^n \xrightarrow{\mathcal{L}\text{-s}} p(t)$ *under Assumption SHL, then* $p(t)^n \xrightarrow{\mathcal{L}\text{-s}} p(t)$ *under Assumption HL, and the equivalent statement holds for convergence in probability.*

Proof. Let $U^n(p)(t)$ and $U(p)(t)$ refer to two processes that are defined in terms of $p(t)$. In the first step, I define a process in terms of $p(t)$ that satisfies Assumptions SHL and Infinite-Activity Jumps and characterize its relationship to $p(\tau)$. In the second step, I show that if that $p(t)$ satisfies Assumptions HL and Infinite-Activity Jumps, then $U^n(p)(t) \xrightarrow{\mathcal{L}\text{-s}} U(p)(t)$ under Assumption SHL implies $U^N(p)(t) \xrightarrow{\mathcal{L}\text{-s}} U(p)(t)$ under Assumption HL. I then show that Assumption Infinite-Activity Jumps is unnecessary, and similar statements hold for convergence in probability and convergence of stopped processes.

Step 1

Let $\omega \in \Omega$ index the event space. We can assume without loss of generality that $\mu(0) = 0$, and so there is a localizing sequence τ_j such that $\|\mu(t)\| \leq j$ if $0 \leq t \leq \tau_j$. Define the stopping times $R_j = \inf(t : \|p(t)\| + \|\sigma(t)\| \geq p)$ and the stopping times $Q_j = \inf(t : \|p(t)\| + \|\gamma(t)\| \geq p)$. These increase to $+\infty$ as well. Therefore, we can set $S_j = \tau_j \wedge R_j \wedge Q_j$.

Then we can define the following processes:

$$\mu^{(j)}(t) = \mu(t \wedge S_j),\ \sigma^{(j)}(t) = \sigma(t \wedge S_j),\ \gamma^{(j)}(t) = \gamma(t \wedge S_j) \tag{2.64}$$

and

$$p^{(j)}(t) = \begin{cases} 0 & \text{if } S_j = 0 \\ p(0) + \int_0^t \mu^{(j)}(s)\, ds + \int_0^t \sigma^{(j)}(s)\, dW(s) + \int_0^t \gamma^{(j)}\, d\mathcal{L}(s) & \text{if } S_j > 0. \end{cases} \tag{2.65}$$

Now, local characteristics of $p^{(j)}$ agree when $t < S_j$ as they are defined to be the same. If $S_j = 0$, then $\|p(t)\| = 0$, and so we are equal there as well. Furthermore, if we use the same driving measures $W(t)$ and $\mathcal{L}(t)$ to represent both processes, the equality is not just in distribution, but ω by ω, where the original processes are defined relative to an event space Ω. In addition, $p^{(j)}(t)$ satisfies Assumption SHL, since $\|p^{(j)}(t)\| \leq 3p$.

Step 2

By the proof of Jacod and Protter 2012, Lemma 4.4.9, the above statement is sufficient to show that the estimators defined above imply convergence stably-in-law. Then this holds for any process, and so it clearly holds for the stopped versions above. In addition, convergence stably-in-law implies convergence in probability if the two processes are defined on the same probability space, which we do not change above. So if the original result was for convergence in probability, the new one is as well.

If $p(t)$ does not satisfy Assumption Infinite-Activity Jumps, then it is locally a convolution of a Laplacian mixture and the zero process. Replacing part of the sample path with 0 does not violate any boundedness conditions. Therefore, we can replace $p^{(j)}(t)$ with the 0 process when necessary, and so the result even holds if Assumption Infinite-Activity Jumps does not hold.

□

Theorem 2.8 (Estimating the Instantaneous Absolute Volatility)**.** *Let $p(t)$ be an Itô semimartingale satisfying Assumptions HL, Infinite-Activity Jumps, and Square-Integrable. Let k_n, Δ^n satisfy $k_n \to \infty$ and $k_n\sqrt{\Delta^n} \to 0$, and let $0 < \tau < \infty$ be a deterministic time. Define $i_n \coloneqq i - k_n - 1$.*

Then the following holds, where $\text{erfcx} := \frac{2\exp(x^2)}{\sqrt{\pi}} \int_x^\infty \exp(-s^2)\, ds$*:*[40]

$$\frac{1}{k_n\sqrt{\Delta^n}} \sum_{m=0}^{k_n-1} |\Delta^n_{i_n+m} p| \xrightarrow{\mathbb{P}} \mathbb{E}|\mathcal{N}(0,1)|\sigma(\tau-) + \frac{\gamma(\tau-)}{\sqrt{2}} \text{erfcx}\left(\frac{\sigma(\tau-)}{\gamma(\tau-)}\right).$$

Proof. This proof is divided into a number of steps. I start by deriving the mean of the absolute volatility under an assumption that $\sigma(t)$ and $\gamma(t)$ are locally constant. I then show that the estimator in that situation converges to its mean. I then relax the assumption of locally-constant volatility.

Step 1

In this section, I start by applying Itô's Formula for convex functions to $|p|(t)$ to separate its variation into its jump and continuous components. Recall the left-derivative of the absolute value function:

$$f'_- = \text{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x \leq 0. \end{cases} \tag{2.66}$$

Using Medvegyev (2007, Theorem 6.65), where $A(t)$ is a finite-valued increasing process, we can rewrite $||p(t)$ as

$$|p(t)| = \int_0^t \text{sign}(p(s-))\, dp(s) + A(t) = \int_0^t \text{sign}(p(s-))\, dW(s) + \int_0^t \text{sign}(p(s-))\, d\mathcal{L}(s) + A(t). \tag{2.67}$$

This $A(t)$ is a finite-valued increasing process and so it can be absorbed into the drift term of $p(t)$ and vanishes as $\Delta \to 0$. If the Laplace part and the diffusion parts have the same sign, $|p|(t) - A(t)$ is the sum of the absolute values of the two processes. Since the innovation processes are independent and symmetric, this occurs one-half of the time.

If they have different signs, the situation is more difficult. In that case, $\text{sign}(p(s-))$ is the same as the sign of the larger, in magnitude, of the two processes. Since the two processes have different signs, the smaller process has the opposite sign. Consequently,

[40] This function, erfcx, is the scaled complementary error function. It is a reparameterization of Mill's ratio. Most scientific programming suites provide efficient, numerically-stable implementations.

the part of $|p(t)| - A(t)$ where the two process has different signs can be rewritten as follows. Let $\Omega^{\mathcal{L}}$ be the set where the Laplace part in magnitude is larger and Ω^{W} the part where the diffusion part is.

$$
\begin{aligned}
&|p(t)| - A(t) \,|\, \text{the signs differ} \qquad (2.68)\\
&= \int_0^t \operatorname{sign}(W(s-))1_{\Omega^W}(s-)\sigma(s)\,dW(s) - \int_0^t \operatorname{sign}(\mathcal{L}(s-))1_{\Omega^W}(s-)\gamma(s)\,d\mathcal{L}(s)\\
&+ \int_0^t \operatorname{sign}(\mathcal{L}(s-))1_{\Omega^{\mathcal{L}}}(s-)\gamma(s)\,d\mathcal{L}(s) - \int_0^t \operatorname{sign}(W(s-))1_{\Omega^{\mathcal{L}}}(s-)\sigma(s)\,dW(s)
\end{aligned}
$$

Let Δ be the length of an interval over which $\gamma(t)$ and $\sigma(t)$ are constant, and let $|\psi|$ and $|\phi|$ denote the densities of the absolute values of a Laplace and Gaussian variables, respectively. Then we can rewrite an increment of (2.68) as follows condition on the signs differing as follows.[41]

$$
\int_0^\infty \int_x^\infty (y-x)\psi_{\gamma,\Delta}(x)|\phi|_{\sigma,\Delta}(y)\,dx\,dy + \int_0^\infty \int_y^\infty (x-y)\psi_{\gamma,\Delta}(x)|\phi|_{\sigma,\Delta}\,dy\,dx \qquad (2.69)
$$

$$
= \frac{\sqrt{\Delta}}{\sqrt{2}}\left(-\gamma + \frac{2}{\sqrt{\pi}}\sigma + \gamma\,\operatorname{erfcx}\left(\frac{\sigma}{\gamma}\right)\right) + \frac{\gamma\sqrt{\Delta}}{\sqrt{2}}\operatorname{erfcx}\left(\frac{\sigma}{\gamma}\right) \qquad (2.70)
$$

$$
= \sqrt{\Delta}\left(m_1\sigma + \frac{\gamma}{\sqrt{2}}\left(2\operatorname{erfcx}\left(\frac{\sigma}{\gamma}\right) - 1\right)\right) \qquad (2.71)
$$

In the part where they both have the same sign, the absolute value is just the sum of the absolute values and so we can rewrite (2.68) given that the signs equal as

$$
m_1\sigma\sqrt{\Delta} + \frac{\gamma}{\sqrt{2}}\sqrt{\Delta}. \qquad (2.72)
$$

By taking the average of the (2.71) and (2.72), we can solve for (2.68):

$$
\mathbb{E}|p(t)| - A(t) = m_1\sigma\sqrt{\Delta} + \frac{\gamma\sqrt{\Delta}}{\sqrt{2}}\operatorname{erfcx}\left(\frac{\sigma}{\gamma}\right). \qquad (2.73)
$$

The first part of this equation is the expectation of the absolute value of the diffusion part. If $\operatorname{erfcx}(\sigma/\gamma)$ were replaced with 1, the second part would be the absolute

[41] A standard computer algebra system can be used to perform the requisite integration.

value of the jump part. Consequently, $\text{erfcx}(\sigma/\gamma)$ reweights the jumps appropriately. It is also worth noting that $\lim_{x\to 0}\text{erfcx}(x) = 1$, and $\lim_{x\to\infty}\text{erfcx}(x) = 0$. Consequently, as σ vanishes we recover the mean of absolute value of the jumps, while as γ vanishes we recover the mean of the absolute value of the diffusion part, as it should.

Step 3

This section considers the asymptotic behavior of the estimator. It proves convergence in mean-square, which implies convergence in probability. Let Ω_n be the set where the two increments have the same sign and let λ_n be its accompanying Lebesgue measure.

Since $\sigma(t)$ and $\gamma(t)$ are step functions, there exists a sequence $\{\tau_j\}$ such that $\sigma(t)$ and $\gamma(t)$ are constant over the intervals between the various τ_j. Hence,

$$p(t) = \sum_j \int_{\tau_j}^{\tau_{j+1}} \sigma\left(\tau_j\right) dW(s) + \int_{\tau_j}^{\tau_{j+1}} \gamma\left(\tau_j\right) d\mathcal{L}(s) \tag{2.74}$$

Consider the squared norm of the difference between the estimator and its expectation. It is worth noting that as k_n gets large we are averaging over times earlier and earlier with reference to τ, which is why the bottom part of the integral is growing with k_n, not the top part. We can assume without loss of generality $\sigma(t)$ and $\gamma(t)$ are constant over $\tau - k_n\Delta^n, \tau$ by taking $k_n\Delta^n$ to 0 faster than the mesh of τ goes to zero, (which it may not at all). Consequently, we let τ depend upon n in our notation.

We can rewrite the sample and population difference as

$$\mathbb{E}\left[\left\|\frac{1}{k_n\Delta^n}\sum_{m=0}^{n}\left|\Delta_{i_n+m}^n p\right| - \left|m_1\sigma(\tau_n) + \frac{\gamma(\tau_n)}{\sqrt{2}}\,\text{erfcx}\left(\frac{\sigma(\tau_n)}{\gamma(\tau_n)}\right)\right|\right\|^2\right] \tag{2.75}$$

$$= \frac{1}{k_n^2\Delta^n}\mathbb{E}\left[\left\|\sum_{m=0}^{k_n}\left|\int_{\tau(n,m+1)}^{\tau(n,m)}\sigma(\tau_n)\,dW(s) + \int_{\tau(n,m+1)}^{\tau(n,m)}\gamma(\tau_n)\,d\mathcal{L}(s)\right|\right.\right. \tag{2.76}$$

$$\left.\left. - k_n\sqrt{\Delta^n}\left|m_1\sigma(\tau_n) + \frac{\gamma(\tau_n)}{\sqrt{2}}\,\text{erfcx}\left(\frac{\sigma(\tau_n)}{\gamma(\tau_n)}\right)\right|\right\|^2\right].$$

By applying (2.73), we have

$$= \frac{1}{k_n^2 \Delta^n} \mathbb{E}\left[\left\| \left| k_n \sqrt{\Delta^n} \right| m_1 \sigma(\tau_n) + \frac{\gamma(\tau_n)}{\sqrt{2}} \operatorname{erfcx}\left(\frac{\sigma(\tau_n)}{\gamma(\tau_n)} \right) \right| + A(t) \right. \tag{2.77}$$
$$\left. - k_n \sqrt{\Delta^n} \left| m_1 \sigma(\tau_n) + \frac{\gamma(\tau_n)}{\sqrt{2}} \operatorname{erfcx}\left(\frac{\sigma(\tau_n)}{\gamma(\tau_n)} \right) \right| \right\|^2 \Bigg].$$

Simplifying implies this equals

$$O_p\left(\frac{k_n^2 (\Delta^n)^2}{k_n^2 \Delta^n} \right) + O_p(\Delta^n), \tag{2.78}$$

since $A(t)$ is a finite-variation term.

Step 5

To finish deriving the theorem, we show that approximating the volatility functions by step functions is innocuous. Consider a sequence $\tau_n \to \tau$, and define $\tilde{\sigma}(t) = \sigma(\max \tau_n : \tau_n \leq t)$, and similarly for $\tilde{\gamma}(t)$. Define $\gamma_x^2(t) = \sup_{s_1, s_2 < t \wedge \tau} |x(s_1) - \tilde{x}(s_2)|^2$ for x equal to σ and γ, while let $\gamma_b^2(t) = \sum_{s_1, s_2 < t \wedge \tau} |b(s_1) - b(s_2)|$. These functions exist and are almost surely finite by localization since σ, γ, and b are locally-bounded. Now, consider the squared distance between any semimartingale satisfying our assumptions and the one used in (2.74). Let $t_1, t_2 < \tau$. Consider:

$$\mathbb{E}\left[\left\| \int_{t_1}^{t_2} \mu(s)\, ds + \int_{t_1}^{t_2} \sigma(s)\, dW(s) + \frac{1}{2} \int_{t_1}^{t_2} \gamma(s) d\mathcal{L}(s) \right. \right. \tag{2.79}$$
$$\left. \left. - \left(\int_{t_1}^{t_2} \tilde{\sigma}(s)\, dW(s) + \frac{1}{2} \int_{t_1}^{t_2} \tilde{\gamma}(s)\, d\mathcal{L}(s) \right) \right\|^2 \right].$$

Increasing the range is valid because all of the integrands are positive:

$$\leq \mathbb{E}\left[\int_{t_1}^{\tau} \mu(s)^2(s) + \int_{t_1}^{\tau} |\tilde{\sigma}(s) - \sigma(s)|^2\, ds + \frac{1}{2} \int_{t_1}^{\tau} |\tilde{\gamma}(s) - \gamma(s)|^2\, ds \right]. \tag{2.80}$$

Then we can bound each of the terms:

$$= (O(1)\gamma_b^2(\tau) + O(1)\gamma_\sigma^2(\tau) + O(1)\gamma_\gamma^2(\tau))(\tau - t_2). \tag{2.81}$$
$$= O(1)(\tau - t_2). \tag{2.82}$$

In other words, if we choose a sequence of meshes so that the supremum of their magnitudes $\Delta^n \to 0$ and the minimal value $\tau - k_n\Delta^n \to 0$, the entire square converges. As one might expect from the definition of integration, approximating the integrands by step functions is innocuous.

Finally, we combine the preceding parts to bound the entire process. Note, since variances of sums can be written in terms of variance of the original parts and their covariance, the asymptotic rate at which the quadratic variation decreases towards zero equals the larger of the asymptotic rates at which its constituent components do. Let $Y'(t)$ be the absolute value of the process derived in (2.74). Consider the mean-square deviation of the estimator from its limiting value:

$$\frac{1}{k_n^2\Delta_n}\mathbb{E}\left[\left\|\sum_{m=0}^{k_n-1}|\Delta^n_{i_n+m}p| - Y'(t) + Y'(t) - \left(m_1\sigma(\tau-)k_n\sqrt{\Delta^n} + \frac{\gamma(\tau-)}{\sqrt{2}}\operatorname{erfcx}\left(\frac{\sigma(\tau-)}{\gamma(\tau-)}\right)k_n\sqrt{\Delta^n}\right)\right\|^2\right]. \tag{2.83}$$

By splitting the term into two parts and using the bounds from (2.78) and (2.82).

$$= \frac{1}{k_n^2\Delta_n}\left(O(\Delta k_n^2) + O(\Delta k_n)\right) \to 0. \tag{2.84}$$

□

Theorem 2.7 (Estimating the Instantaneous Diffusion Volatility)**.** *Let $p(t)$ be an Itô semimartingale satisfying Assumptions HL, Infinite-Activity Jumps, and Square-Integrable. Let k_n, Δ^n satisfy $k_n \to \infty$ and $k_n\sqrt{\Delta_n} \to 0$, and let $0 < \tau < \infty$ be a deterministic time. Define $i_n = i - k_n - 1$. Let $c_1(\Delta^n)^{1/4} < v_1^n < c_2\sqrt{\Delta^n}$ for some constants c_1, c_2 and $v_2^n \to 1$. Then*

$$\widehat{\sigma}^2_{i_n}(k_n, \tau-, p) := \frac{1}{k_n\Delta^n}\sum_{m=0}^{k_n-1} v_2^n\left|\Delta^n_{i_n}p\right|^2 1\{\left|\Delta^n_{i_n}p\right| \le v_1^n\} \xrightarrow{\mathbb{P}} \sigma^2(\tau-).$$

Proof. The intuition behind the proof is straightforward. We separate the large jumps from the continuous part by truncating, and then note that the small jumps do not matter asymptotically because by squaring the remainder they get pushed even closer to zero. Consequently, we only pick up the middle range of the distribution, which is

dominated by the continuous variation. Effectively, we are considering $\lim_{s\to 0} \widehat{IV}(\tau) - \widehat{IV}(\tau - s)$, and since we are estimating the left-limit of its time-derivative, $\sigma^2(\tau-)$, this works.

By localization we can strengthen some assumptions. Specifically, we can replace Assumption HL with Assumption SHL. In addition, the jump martingale part of the process is a sum of an integral with respect to Laplace motion $\mathcal{L}(t)$ and the zero process $\delta_0(t)$ where the weights depend upon the intensity of the jumps by Corollary 2.2 The jump increments of that part are almost surely zero, and so if we separate the space into parts where $\mathcal{L}(t)$ is active and where $\delta_0(t)$ is active, we only have to deal with the first section. Consequently, we can assume that the jump part is an integral with respect to $\mathcal{L}(t)$. The part of the proof regarding the continuous part of the process will not change in either part.

Step 1

I proceed by showing convergence in mean square, which implies convergence in probability. Note, $\left|\Delta^n_{i_n+m} p\right|^2 = O_p(\Delta^n)$ for all i, since $p(t)$ is an integral with bounded integrands and integrators whose quadratic variation is proportional to Δ^n. Consider the jump part of the variation. To prove consistency of the original process, I must show that the jump part converges to zero.

Following Jacod and Protter (2012, 258), for all $w, x, y, z \in R$, $\epsilon \in (0, 1]$, and $v \geq 1$,

$$\left|(x+y+z+w)1\{|x+y+z+w| < v\} - x^2\right| \leq K\frac{|x|^4}{v^2} + \epsilon x^2 + \frac{K}{\epsilon}((v^2 \wedge y^2) + z^2 + w^2). \tag{2.85}$$

Define the following four processes, where I split the process up. The continuous variation is split into two parts, one with locally constant volatility and the other being the additional deviation coming from the change in the volatility:

$$Y^n(t) \coloneqq \sigma(\tau_n)(W_t - W_{\tau_n})1\{\tau_n \leq t\} \tag{2.86}$$

$$Y'^n(t) \coloneqq \int_{\tau_n \wedge t}^{t} (\sigma(s) - \sigma(\tau_n)\, dW(s) \tag{2.87}$$

$$Z^n(t) \coloneqq \int_{\tau_n \wedge t}^{t} \gamma(s)\, d\mathcal{L}(s) \tag{2.88}$$

$$B^n(t) \coloneqq \int_{\tau_n \wedge t}^{t} \mu(s)\, ds. \tag{2.89}$$

Note, $p(\tau_n \wedge t) = Y^n(t) + Y'^n(t) + Z^n(t) + B^n(t)$. Now, we can use (2.85), with $x = \frac{\Delta^n_{i_n+m} Y^n}{\sqrt{\Delta^n}}$, $y = \frac{\Delta^n_{i_n+m} Z^n_{i_n+m}}{\sqrt{\Delta^n}}$, and $w = \frac{\Delta^n_{i_n+m} B^n_{i_n+m}}{\sqrt{\Delta^n}}$. The main issue here is showing that all of the parts except for $Y^n(t)$ converge to zero because then we are essentially just taking the variance of that part. Take $v = \frac{v_n}{\sqrt{\Delta^n}} = \omega_n$, where $\omega_n = o_p(1/\Delta^n)$ and $1/\omega_n$ is $o_p(\sqrt{\Delta})$. Then we have the following inequality:

$$\begin{aligned}\frac{1}{k_n \Delta^n} \sum_{m=0}^{k_n-1} \left|(Y^n_t)^2 - (p^n_t)^2\right| \leq \frac{1}{k_n} \sum_{m=0}^{k_n-1} \Bigg(& \frac{K}{\omega_n^2} \left| \frac{\Delta^n_{i_n+m} Y^n}{\sqrt{\Delta^n}} \right|^4 + \epsilon \left| \frac{\Delta^n_{i_n+m} Y^n}{\sqrt{\Delta^n}} \right|^2 \\ & + \frac{K}{\omega_n^2 \epsilon} \left| \frac{\Delta^n_{i_n+m} Z}{\sqrt{\Delta^n}\omega_n} \right|^2 + \frac{K}{\epsilon} \left| \frac{\Delta^n_{i_n+m} Y'^n}{\sqrt{\Delta^n}} \right|^2 + \frac{K}{\epsilon} \left| \frac{\Delta^n_{i_n+m} B}{\sqrt{\Delta^n}} \right|^2 \Bigg).\end{aligned} \tag{2.90}$$

Set $\gamma_n = \sum_{s \in |\tau_n, \tau_n + (k_n+2)\Delta^n|} |\sigma(s) - \sigma(\tau_n)|^2$, which is bounded and converges to zero, and $\phi_n = \sum_{s \in |\tau_n, \tau_n + (k_n+2)\Delta^n|} |\gamma(s)|^2$ The key hard part is bounding $\Delta^n_{i_n+m} Z$. Clearly, $E\left|\Delta^n_{i_n+m}\right| \leq \phi_n \sqrt{\Delta^n}$. Consider the part of the variation in $Z(t)$ that comes from jumps smaller than 1 in magnitude, where 1 is an arbitrary constant picked for the sake of simplicity:

$$\mathbb{E}|\mathcal{L}(0, \phi_n) \wedge 1| = \phi_n \sqrt{\Delta^n} - \exp\left(-\frac{1}{\phi_n \sqrt{\Delta^n}}\right)(\phi_n \sqrt{\Delta^n} + 1) \leq O\left(\frac{1}{\sqrt{\Delta^n}}\right) \exp\left(-\frac{1}{\phi_n \sqrt{\Delta^n}}\right). \tag{2.91}$$

In addition, since τ_n is a stopping time, the probability that a jump exceeds 1 in the previous k_n periods declines to 0 almost surely with Δ^n. Consequently, $\frac{1}{\omega_n^2} \frac{\Delta^n_{i_n+m} Z}{\sqrt{\Delta^n}\omega_n} \overset{a.s.}{\in} O_p\left(\frac{1}{\Delta^n \omega_n^3}\right) \exp\left(-\frac{1}{\phi_n \sqrt{\Delta^n}}\right) = o_p(1)$ as exponential functions decay faster than polynomials increase.

I use K to refer to an arbitrary constant here, which may change. $\Delta^n_{i_n+m} B$ is the drift term, and so $\left|\Delta^n_{i_n+m} B\right| \leq K\Delta^n$. $\mathbb{E}[\left|\Delta^n_{i_n+m} Y^n\right|^4 | \mathcal{F}_{(i_n+m-1)\Delta^n}] \leq K(\Delta^n)^2$. $\mathbb{E}[\left|\Delta^n_{i_n+m} Y'^n\right|^n | \mathcal{F}_{(i_n+m-1)\Delta^n}] \leq K\Delta^2 \mathbb{E}[\gamma_n | \mathcal{F}_{(i_n+m-1)\Delta^n} \leq K\Delta^n$. As a consequence, we have the following where ξ_n is some sequence converging to zero:

$$\mathbb{E}\left[\left|(Y^n_t)^2 - (p^n_t)^2\right|\right] \leq K\epsilon + \frac{K}{\epsilon}\left(o_p(1) + o_p(1) + \mathbb{E}[\gamma_n]\right). \tag{2.92}$$

If we take $n \to \infty$, and then $\epsilon \to 0$, the left hand side of the above equation converges to zero.

Step 2

To complete the proof, we have to consider what $\lim_{n\to\infty} \frac{1}{k_n \Delta^n} \sum_{m=0}^{k_n-1} |Y_t^n|^2$ is. If we recall its definition, we note that it converges to the variance of the increment:

$$\frac{1}{k_n \Delta^n} \sum_{m=0}^{k_n-1} |\sigma_{\tau_n}(W_t - W_{\tau_n}) 1\{\tau_n \leq t\}|^2 = \sigma(\tau_n)^2 \frac{1}{k_n} \sum_{m=0}^{k_n-1} \left| \frac{\Delta_{i_n+m}^n W}{\sqrt{\Delta^n}} \right|^2 \to \sigma(\tau_n)^2 \quad (2.93)$$

Since the square is a convex function, we can combine these two previous limits, and we get that the original expression converges to $\sigma(\tau_n)^2$. However, this is the local integrated volatility evaluated at τ_n, which was the object of interest. Clearly, if we multiply the expression by a value that is almost surely converging to 1, none of the results change, and we are done.

□

Theorem 2.9 (Estimating the Instantaneous Jump Volatility)**.** *Let $p(t)$ be an Itô semimartingale satisfying Assumptions HL, Infinite-Activity Jumps, and Square-Integrable. Let k_n, Δ^n satisfy $k_n \to \infty$ and $k_n \sqrt{\Delta^n} \to 0$, and let $0 < \tau < \infty$ be a deterministic time. Define $i_n = i - k_n - 1$. Let $\widehat{\sigma}_n(\tau-)$ converge in probability to $\sigma(\tau-)$. Let $\gamma(\tau) > 0$ and g be strictly-increasing, convex, and continuous, then the following holds:*

$$\widehat{\gamma}(k_n, \tau-, p) \coloneqq \operatorname*{argmin}_{\gamma} g\left(\left| \frac{1}{k_n \sqrt{\Delta^n}} \sum_{m=0}^{k_n-1} \left| \Delta_{i_n+m}^n p \right| - \mathbb{E}|\mathcal{N}(0,1)| \widehat{\sigma}_n(\tau-) - \frac{\gamma \operatorname{erfcx}\left(\frac{\widehat{\sigma}_n(\tau-)}{\gamma} \right)}{2} \right| \right)$$
$$\xrightarrow{\mathbb{P}} \gamma(\tau-).$$

Proof. In the following proof, I use 0 subscripts to denote population objects. Consider

$$\widehat{Q}_n(\gamma) \coloneqq g\left(\left| \frac{1}{k_n \sqrt{\Delta}} \sum_{m=0}^{k_n-1} \left| \Delta_{i_n+m}^n p \right| - m_1 \hat{\sigma}(\tau-) - \gamma \operatorname{erfcx}\left(\frac{\sigma_0}{\gamma \sqrt{2}} \right) \right| \right). \quad (2.94)$$

We can start by noting that $\widehat{Q}_n(\gamma)$ is implicitly a continuous function of $\hat{\sigma}_n(\tau-)$. However, since, by assumption, $\hat{\sigma}_n(\tau-) \xrightarrow{\mathbb{P}} \sigma_0$, we can suppress that dependence in our notation and plug in σ_0. In addition, g is an increasing function and both g and the absolute value are convex, continuous functions, we can use the continuous mapping theorem to derive the limiting value of $\widehat{Q}_n(\gamma)$.

$$Q_0(\gamma) \coloneqq g\left(\left|\gamma_0 \operatorname{erfcx}\left(\frac{\sigma_0}{\gamma\sqrt{2}}\right) - \gamma \operatorname{erfcx}\left(\frac{\sigma_0}{\gamma\sqrt{2}}\right)\right|\right). \tag{2.95}$$

Clearly, this equals zero when $\gamma = \gamma_0$. Moving forward, we will show that both $\widehat{Q}_n(\gamma)$ and $Q_0(\gamma)$ are both strictly convex, which will imply the minimum is unique. Define $A(\sigma, \gamma) \coloneqq \gamma \operatorname{erfcx}\left(\frac{\sigma}{\gamma\sqrt{2}}\right)$. Showing $A(\sigma, \gamma)$ is strictly increasing for all σ is sufficient to show this convexity because of properties assumed about g and the absolute-value function. Consider

$$\frac{\partial}{\partial\gamma}\gamma \operatorname{erfcx}\left(\frac{\sigma}{\gamma\sqrt{2}}\right) = \operatorname{erfcx}\left(\frac{\sigma}{\gamma\sqrt{2}}\right) - \frac{\sigma}{\gamma^2\sqrt{2}}\frac{\partial}{\partial x}\operatorname{erfcx}(x)\Big|_{x=\frac{\sigma}{\gamma\sqrt{2}}}. \tag{2.96}$$

Since erfcx is a decreasing function, the last term is negative, and so the entire equation is strictly positive. This implies that $\widehat{Q}_n(\gamma)$ and $Q_0(\gamma)$ are both strictly convex as functions of γ, which then implies the minimum given above is strict.

Since we assumed that $\gamma_0 > 0$, γ_0 is in the interior of a convex set. Consequently, by Newey and McFadden (1994, Theorem 2.7), $\hat{\gamma}_n$ is well-defined in the sense of being a unique minimizer, and $\hat{\gamma}_n \xrightarrow{\mathbb{P}} \gamma_0$.

□

2.C News Premia Theorems

Theorem 2.1 (Jump Times are News Times)**.** *Consider a stopping time τ. Let $P(t)$ be a price process satisfying no-arbitrage. Then its natural filtration — $\mathcal{F}_t^p$ — contains all of the information in the representative investor's information set relevant for asset pricing, and $\mathcal{F}_\tau^p \neq \mathcal{F}_{\tau-}^p$ if and only if $P(t)$ jumps at τ, where $\mathcal{F}_{t-}^p$ is the associated predictable filtration.*

Proof. Since, $P(t)$ satisfies no-arbitrage in the sense of no-free lunch with vanishing risk, by Delbaen and Schachermayer (1994), it is a semimartingale. First we prove

that if $P(t)$ jumps at τ, then the two filtrations are not equal. Note, $\mathcal{F}^p_{t-} = \cup_{s<t}\mathcal{F}^p_s$. Clearly, $p(\tau) \notin \mathcal{F}^p_s$ for all $s < t$, and so it is not contained in their union, and so $\mathcal{F}_{\tau-} \neq \mathcal{F}_\tau$.

To prove the reverse direction, let ${}^pP(t)$ be the predictable projection of $P(t)$, but then since ${}^pP(t)$ is pre-visible, ${}^pP(\tau)$ is measurable with respect to $\mathcal{F}^p_{\tau-}$, but $p(\tau)$ is not by assumption. Hence, it cannot equal its predictable projection with probability 1. However, this implies that τ is a jump time of $P(t)$.

The only other thing that we need to prove is that $\mathcal{F}^p_t$ contains all of the information that the representative investor knows that is relevant for asset pricing. Assume not. Then there exists an event $\mathcal{E}$ contained in the representative investor's information set $\mathcal{F}^r_t$ that is relevant for asset pricing, but is not measurable with respect to $\mathcal{F}^p_t$. Let $P(t)$ be the price, and $\mathcal{M}(t)$ be the representative investor's pricing kernel.

Then we know the following, where $P(t)$ is the cum-dividend price:

$$P(t) = \mathbb{E}\left[\mathcal{M}(\tau)P(\tau) \,|\, \mathcal{F}^r_t\right] \forall \tau \geq t. \tag{2.97}$$

Since $\mathcal{E}$ is relevant for asset pricing there exists a stopping time τ such that the following inequality holds:

$$\mathbb{E}\left[\mathcal{M}(\tau)P(\tau) \,|\, \mathcal{F}^r_t\right] \neq \mathbb{E}\left[\mathcal{M}(\tau)P(\tau) \,|\, \mathcal{F}^p_t\right]. \tag{2.98}$$

However, $P(t)$ is measurable with respect to $\mathcal{F}^p_t$ by definition, and it equals the value on the left. This is a contradiction.

□

Lemma 2.12 (An Itô's Formula for the Expectation of a Square Integrable Semimartingale)**.** *Let f be a twice-differentiable function and $\widetilde{Z}$ be a vector-valued semimartingale with locally bounded predictable $\langle Z\rangle(t)$. Then the differential of f satisfies*

$$d\mathbb{E}\left[f(\widetilde{Z}) \,\middle|\, \mathcal{F}_{t-}\right] = \mathbb{E}\left[f'(\tilde{Z}(t-))\, d\tilde{Z}(t) \,\middle|\, \mathcal{F}_{t-}\right] + \frac{1}{2} f''(\tilde{Z}(t-))\, d\langle \tilde{Z}\rangle(t).$$

Proof. The argument below is a standard application of Itô's formula for non-continuous processes applied to processes of bounded variation. In addition, the notation below should be interpreted in vector form. For example, $d\tilde{Z}(t)$ is the vector of dZ_i for all i, and $\langle \tilde{Z}^D\rangle$ is a matrix. We start by writing expanding the differential inside

the expectation using Itô's formula for non-continuous semimartingales, (Medvegyev 2007, Theorem 6.46):

$$d\mathbb{E}\left[f(\tilde{Z}(t))\,\middle|\,\mathcal{F}_{t-}\right] = d\mathbb{E}\left[\sum_{i=1}^{d}\frac{\partial f}{\partial z_i}\tilde{Z}(t-))\,d\tilde{Z}_i(t) + \frac{1}{2}\sum_{i,j}\frac{\partial f}{\partial z_i \partial z_j}f(\tilde{z}(t-))\langle \tilde{z_i}^D, \tilde{z_j}^D\rangle(t)\right. \tag{2.99}$$
$$\left. + \left(\Delta f(\tilde{Z}(t)) - \sum_{i=1}^{d}\frac{\partial f}{\partial z_i}f(\tilde{Z}(t-))\Delta\tilde{Z}_i(t)\right)\,\middle|\,\mathcal{F}_{t-}\right].$$

Rearranging and combining terms, we have.

$$= d\mathbb{E}\left[f'(\tilde{Z}(t-))\,d\tilde{Z}(t) + \left(\Delta f(\tilde{Z}(t)) + f'(\tilde{Z}(t-))\Delta\tilde{Z}(s)\right)\right. \tag{2.100}$$
$$\left. - \frac{1}{2}f''(\tilde{Z}(t-))\left\langle\tilde{Z}^D\right\rangle(t)\,\middle|\,\mathcal{F}_{t-}\right].$$

Then by Taylor's theorem, canceling terms and noting that continuity implies bounded for the derivatives of f as long as $\tilde{Z}$ is bounded:

$$= d\mathbb{E}\left[f'(\tilde{Z}(t-))\,d\tilde{Z}(t) + \frac{1}{2}f''(\tilde{Z}(t-))\,d\left\langle\tilde{Z}^D\right\rangle(t)\,\middle|\,\right] \tag{2.101}$$
$$+ \frac{1}{2}\,d\mathbb{E}\left[f''(\tilde{Z}(t-))\Delta\tilde{Z}(t)^2 + O((\Delta\tilde{Z}(t)^3)\,\middle|\,\mathcal{F}_{t-}\right]. \tag{2.102}$$

Since the quadratic variation and the predictable quadratic variation coincide for continuous processes:

$$= d\mathbb{E}\left[f'(\tilde{Z}(t-))\,d\tilde{Z}(t) + \frac{1}{2}f''(\tilde{Z}(t-))\,d[\tilde{Z}](t)\,\middle|\,\mathcal{F}_{t-1}\right] \tag{2.103}$$
$$+ \mathbb{E}\left[O((\Delta\tilde{Z}(t)^3))\,\middle|\,\mathcal{F}_{t-}\right]. \tag{2.104}$$

By the Davis-Burkholder-Gundy inequality, for some constant c:

$$= d\mathbb{E}\left[f'(\tilde{Z}(t-))\,d\tilde{Z}(t)\,\middle|\,\mathcal{F}_{t-1}\right] + \frac{1}{2}\mathbb{E}\left[f''(\tilde{Z}(t-))\,d[\tilde{Z}](t)\,\middle|\,\mathcal{F}_{t-}\right] \tag{2.105}$$
$$+ \mathbb{E}\left[c_1 O([\tilde{Z}]^{3/2})\,\middle|\,\mathcal{F}_{t-}\right].$$

Since we are considering local changes in time,

$$= d\mathbb{E}\left[f'(\tilde{Z}(t-))\,d\tilde{Z}(t)\,\middle|\,\mathcal{F}_{t-}\right] + \frac{1}{2}f''(\tilde{Z}(t-))\,d\left\langle\tilde{Z}\right\rangle(t). \tag{2.106}$$

□

Theorem 2.11 (Asset-Pricing Equation)**.** *Let Assumption 2.6 hold, prices be Itô semimartingales, and the representative consumer face Problem 2.1 as* $\Delta \to 0$*. Assume preferences are such that optimal consumption is strictly positive. Define*

$$M^{UP}(t) \coloneqq \frac{\phi'(V(W(t))}{\phi'(V(W(t-))} \text{ and } M(t) \coloneqq \frac{\phi'(V(W(t-))}{\phi'\left(\phi^{-1}\left(\mathbb{E}\left[\phi(V(W(t)))\,|\,\mathcal{F}_{t-}\right]\right)\right)}\frac{V'(W(t))}{u'(c(t-))}.$$

Then $M^{UP}(t)$ *is a purely discontinuous martingale, and for all stopping times* $\tau > t$,[42]

$$\widetilde{P}(t) = \mathbb{E}\left[M(\tau)M^{UP}(\tau)\widetilde{P}(\tau)\,\middle|\,\mathcal{F}_t\right]$$

Proof. Define the discounted price: $\widetilde{P}(t) \coloneqq \exp(-\kappa t)P(t)$. This is a concave maximization problem and so first-order conditions characterize the optimum. Assume, for now, that the investor can only adjust his portfolio at a discrete grid of points $t, t+\Delta, t+2\Delta, \ldots$ Then consumption and prices are effectively constant within each period, and the investor is faced with the following problem:

$$V(W(t)) = \max_{\Xi(t),C(t)} u(C(t)) + \exp(-\kappa\Delta)\phi^{-1}\left(\left[\phi\left(V(W(t+\Delta))\right)\,|\,\mathcal{F}_t\right]\right) \tag{2.107}$$

$$C(t) + \sum_i P_i(t)\xi_i(t) = W(t) \tag{2.108}$$

$$W(t+\Delta) = \sum_i P_i(t+\Delta)\xi_i(t) \tag{2.109}$$

Submitting in the constraints gives

$$V(W(t)) = \max_{\Xi(t)} u(W(t) - \sum_i P_i(t+\Delta) + \exp(-\kappa\Delta)\phi^{-1}\left(\left[\phi\left(V(\sum_i P_i(t+\Delta)\xi_i(t))\right)\,\middle|\,\mathcal{F}_t\right]\right). \tag{2.110}$$

[42] I use the UP superscript because M^{UP} is an unpredictable process.

The discounted and original prices coincide at t, and we can equate $\exp(-\kappa\Delta)P_i(t+\Delta)$ with $\widetilde{P}_i(t+\Delta 0$. Hence, by using chain rule, and the formula for the derivative of an inverse, the first-order condition for (2.110) is

$$u'(c(t))\widetilde{P}_i(t) = \mathbb{E}\left[\frac{\phi'(V\left(W(t+\Delta)\right)}{\phi'\left(\phi^{-1}\left(\mathbb{E}\left[\phi(V(W(t+\Delta)))\,|\,\mathcal{F}_t\right]\right)\right)}V'\left(W(t+\Delta)\right)\widetilde{P}_i(t+\Delta)\,\middle|\,\mathcal{F}_t\right], \tag{2.111}$$

at the optimal level of consumption and optimal asset shares. We can rearrange (2.111) as:

$$\widetilde{P}_i(t) = \mathbb{E}\left[\frac{\phi'(V\left(W(t+\Delta)\right)}{\phi'\left(\phi^{-1}\left(\mathbb{E}\left[\phi(V(W(t+\Delta)))\,|\,\mathcal{F}_t\right]\right)\right)}\frac{V'\left(W(t+\Delta)\right)}{u'(c(t))}\widetilde{P}_i(t+\Delta)\,\middle|\,\mathcal{F}_t\right]. \tag{2.112}$$

If we plug in the risk-free rate $\widetilde{P}_f(t)$ in to (2.112), the prices on each side of the equal side are the same, and we can divide through by them. This gives

$$1 = \mathbb{E}\left[\frac{\phi'(V\left(W(t+\Delta)\right)}{\phi'\left(\phi^{-1}\left(\mathbb{E}\left[\phi(V(W(t+\Delta)))\,|\,\mathcal{F}_t\right]\right)\right)}\frac{V'\left(W(t+\Delta)\right)}{u'(c(t))}\,\middle|\,\mathcal{F}_t\right]. \tag{2.113}$$

In other words, the two terms in the inside the expectation are a martingale. Consequently, prices are a martingale with respect to the change of measure they induce. We can take limits as $\Delta \to 0$ in (2.112), which gives

$$\widetilde{P}_i(t) = \mathbb{E}\left[\frac{\phi'(V\left(W(t)\right)}{\phi'\left(\phi^{-1}\left(\mathbb{E}\left[\phi(V(W(t)))\,|\,\mathcal{F}_{t-}\right]\right)\right)}\frac{V'\left(W(t)\right)}{u'(c(t-))}\widetilde{P}_i(t)\,\middle|\,\mathcal{F}_{t-}\right] \tag{2.114}$$

Now, we want to separate these two terms into a pure jump component and the remainder.

To do this, multiply and divide the first expression by $\phi'(V\left(W(t-)\right)$:

$$\widetilde{P}_i(t) = \mathbb{E}\left[\frac{\phi'(V\left(W(t)\right)}{\phi'(V\left(W(t-)\right)}\frac{\phi'(V\left(W(t-)\right)}{\phi'\left(\phi^{-1}\left(\mathbb{E}\left[\phi(V(W(t)))\,|\,\mathcal{F}_{t-}\right]\right)\right)}\frac{V'\left(W(t)\right)}{u'(c(t-))}\widetilde{P}_i(t)\,\middle|\,\mathcal{F}_{t-}\right] \tag{2.115}$$

Note, the first term here is simply $M^{UP}(t)$. Claim: $M^{UP}(t)$ is purely discontinuous. By Ai and Bansal (2018, Theorem 1), we know that the value function is a differentiable, and hence continuous, function of wealth. In addition, I am taking limits locally in time, and ϕ' is strictly positive. Consider $\lim_{\Delta\to 0} M^{UP}(t-\Delta)$.

$$\lim_{\Delta\to 0}\frac{\phi'\left(V\left(W(t-\Delta)\right)\right)}{\phi'\left(V\left(W((t-\Delta)-)\right)\right)}=\frac{\lim_{\Delta\to 0}\phi'\left(V\left(W(t-\Delta)\right)\right)}{\lim_{\Delta\to 0}\phi'\left(V\left(W((t-\Delta)-)\right)\right)}=\frac{\phi'\left(V\left(\lim_{\Delta\to 0}W(t-\Delta)\right)\right)}{\phi'\left(\lim_{\Delta\to 0}V\left(W((t-\Delta)-)\right)\right)} \tag{2.116}$$

$$=\frac{\phi'(V\left(W(t-)\right)}{\phi'(W(t-))}=1$$

This implies that $M^{UP}(t)$ is a pure-jump process because its continuous part is identically one. In addition, I assumed there were no-predictable jumps, hence any drift (finite-variation, predictable) terms in the environment must be continuous. Consequently, $M^{UP}(t)$ is a pure-jump martingale.

□

Theorem 2.13 (Asset-Pricing Equation)**.** *Let the assumptions in Assumption 2.6 hold, $P_i(t)$ be an Itô semimartingales, and the representative consumer face Problem 2.1 as $\Delta\to 0$. Assume that preferences are such that optimal consumption is strictly positive. Then risk-premia for some asset i is*

$$\mathbb{E}\left[\frac{dP_i(t)}{P_i(t-)}-\frac{dP_f(t)}{P_f(t-)}\middle|\mathcal{F}_{t-}\right]=-d\langle m,p^D+p^J\rangle(t)-d\langle m^{UP},p^J\rangle(t).$$

Proof. The goal here is to replace the asset pricing equation in Theorem 2.11 with a stochastic logarithm of $P_i(t)$. Let $\widetilde{M}(\tau)=\exp(-\kappa(\tau))M(\tau)$ be the discounted stochastic discount factor. In this derivation, it is more useful to place the deterministic discounting into the discount factor than into the prices.

Then the asset-pricing equation is:

$$\widetilde{P}_i(t)=\mathbb{E}\left[\widetilde{M}(\tau)M^{UP}(\tau)\widetilde{P}(\tau)\middle|\mathcal{F}_t\right]. \tag{2.117}$$

Since $M(t)M^{UP}(t)$ given $\mathcal{F}_t$ equal 1, we can pre-multiply by it,

$$\widetilde{M}(t)M^{UP}(t)\widetilde{P}(t)=\mathbb{E}\left[\widetilde{M}(\tau)M^{UP}(\tau)\widetilde{P}(\tau)\middle|\mathcal{F}_t\right]. \tag{2.118}$$

In other words, $\widetilde{M}(t)M^{UP}P(t)$ is a martingale. This is the standard SDF type result. Discounted prices are martingales. I now take the stochastic logarithm of both sides. Taking the stochastic logarithm (as opposed to the regular logarithm) is useful

because it preserves the martingale property. (The stochastic logarithm — $\mathcal{L}og(X)$ — is the inverse of the Doléans-Dade exponential.)

Before, I do this, it is useful to consider a few of the stochastic logarithms' properties. First, the following holds: $\mathcal{L}og(X \cdot Y) = \mathcal{L}og(X) + \mathcal{L}og(Y) + [\mathcal{L}og(X), \mathcal{L}og(Y)]$. We can also handle triple-products. You just need to apply the expression twice, and note that finite-variation terms do not affect the quadratic variation.

$$\mathcal{L}og(X \cdot Y \cdot Z) = \mathcal{L}og(X) + \mathcal{L}og(Y) + \mathcal{L}og(Z) + [\mathcal{L}og(X), \mathcal{L}og(Z)] + [\mathcal{L}og(X), \mathcal{L}og(Z)] \\ + [\mathcal{L}og(Y), \mathcal{L}og(Z)] \tag{2.119}$$

As noted above, since (2.118) is a martingale its stochastic logarithm is as well.

$$0 = \mathbb{E}\left[\int_t^\tau d\mathcal{L}og\left(MM^{UP}P\right)(s) \middle| \mathcal{F}_t\right]. \tag{2.120}$$

We can expand this equation using (2.119). We can also replace the integrals with differentials without loss of generality because τ is arbitrary:

$$\implies 0 = \mathbb{E}\Bigg[d\mathcal{L}og(\widetilde{M})(t) + d\mathcal{L}og(M^{UP})(t) + d\mathcal{L}og(P)(t) \\ + d[\mathcal{L}og(\widetilde{M}), \mathcal{L}og(P)](t) + d[\mathcal{L}og(M^{UP}), \mathcal{L}og(P)](t) + d[\mathcal{L}og(\widetilde{M}), \mathcal{L}og(M^{UP})](t) \Bigg| \mathcal{F}_{t-}\Bigg] \tag{2.121}$$

The stochastic logarithm equals the regular logarithm up to finite-variation terms:

$$= \mathbb{E}\Big[d\mathcal{L}og(\widetilde{M})(t) + d\mathcal{L}og(M^{UP})(t) + d\mathcal{L}og(P)(t) \\ + d[\log(\widetilde{M}), \log(P)](t) + d[\log(M^{UP}), \log(P)](t) + d[\mathcal{L}og(\widetilde{M}), \mathcal{L}og(M^{UP})](t) \Big| \mathcal{F}_{t-}\Big]. \tag{2.122}$$

We can combine M and M^{UP} together:

$$= \mathbb{E}\Bigg[d\mathcal{L}og\left(M \cdot M^{UP}\right)(t) + d\mathcal{L}og(P)(t) + d[\log(\widetilde{M}), \log(P)](t) \\ + d[\log(M^{UP}), \log(P)](t) \Bigg| \mathcal{F}_{t-}\Bigg]. \tag{2.123}$$

The stochastic logarithm satisfies the following stochastic differential equation:

$$\mathcal{L}og(X)(t) = \int_0^t \frac{1}{X(s-)}\, dX(s). \tag{2.124}$$

Consequently, we can rewrite (2.123) as follows, where I replace the quadratic variation terms with predictable quadratic variation terms,

$$0 = \mathbb{E}\left[\frac{d\left(M \cdot M^{UP}\right)(t)}{\widetilde{M}(t-)M^{UP}(t-)} + \frac{dP(t)}{P(t-)}\,\middle|\,\mathcal{F}_{t-}\right] + d\langle \log(\widetilde{M}), \log(P)\rangle(t) + d\langle \log(M^{UP}), \log(P)\rangle(t). \tag{2.125}$$

If $M^{UP}(t)$ is identically 1, the all of the terms containing it disappear, which gives the standard asset pricing equation:

$$\mathbb{E}\left[\frac{dP(t)}{P(t-)} + \frac{d\widetilde{M}(t)}{\widetilde{M}(t-)}\,\middle|\,\mathcal{F}_{t-}\right] = -\, d\langle m, p\rangle(t), \tag{2.126}$$

where $m = \log(M)$. We can ignore the discounting because it only cases a mean shift, and so will not affect quadratic covariation terms.

In the recursive case with jumps through, it is more complicated. An announcement SDF term is a pure-jump process so it only have non-zero covariation with the jump part of the prices:

$$\mathbb{E}\left[\frac{dP(t)}{P(t-)} + \frac{d\left(\widetilde{M} \cdot M^{UP}\right)(t)}{\widetilde{M}(t-)M^{UP}(t-)}\,\middle|\,\mathcal{F}_{t-}\right] = -\, d\langle m, p\rangle(t) - d\langle m^{UP}, p\rangle(t), \tag{2.127}$$

where $m^{UP}(t) = \log(M^{UP}(t))$. Since (2.127) prices all assets, if we consider a risk-neutral asset, we have all of the of the quadratic variation terms equaling zero:

$$\frac{dP_f(t)}{P_f(t-)} = -\mathbb{E}\left[\frac{d\left(\widetilde{M} \cdot M^{UP}\right)(t)}{\widetilde{M}(t-)M^{UP}(t-)}\,\middle|\,\mathcal{F}_{t-}\right]. \tag{2.128}$$

Consequently, the risk premium on a asset i with discounted price P_i is

$$\frac{dP_i(t)}{P_i(t-)} - \frac{dP_f(t)}{P_f(t-)} = -\, d\langle m, p\rangle(t) - d\langle m^{UP}, p\rangle(t) \tag{2.129}$$

Since $M^{UP}(t)$ and hence $m^{UP}(t)$ are purely discontinuous processes, the second quadratic variation does not depend upon $p^D(t)$. That is

$$\mathbb{E}\left[\frac{dP_i(t)}{P_i(t-)} - \frac{dP_f(t)}{P_f(t-)}\,\middle|\,\mathcal{F}_{t-}\right] = -\, d\langle m, p^D + p^J\rangle(t) - d\langle m^{UP}, p^J\rangle(t). \tag{2.130}$$

□

2.D News Premia: Empirical Results

Table 2.13: $\mathbb{E}\left[rx_t \,\middle|\, \sigma_t^2 + \gamma_t^2, \frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}, \mathbf{1}\{\text{FOMC}\}_t\right]$ (OLS)

Intercept	$\mathbf{1}\{\text{FOMC}\}_t$	$\log(\sigma_t^2 + \gamma_t^2)$	$\log\left(\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}\right)$	$\log(\sigma_t^2 + \gamma_t^2)\log\left(\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}\right)$	$\bar{\mathbb{R}}^2$
0.01	0.88				0.26 %
[0.23]	[2.94]				
−4.55		−0.46			2.67 %
[−5.45]		[−5.58]			
1.02			1.65		1.61 %
[5.81]			[6.48]		
−3.17		−0.39	1.13		3.35 %
[−3.94]		[−4.12]	[4.06]		
−1.27		−0.19	3.91	0.29	3.42 %
[−0.54]		[−0.85]	[1.07]	[0.80]	
−4.73	1.09	−0.47			3.09 %
[−5.88]	[3.55]	[−6.11]			
−3.40	1.00	−0.40	1.07		3.70 %
[−3.96]	[3.38]	[−5.08]	[3.94]		
	0.98	−0.23	3.55	0.26	3.74 %
[−0.68]	[3.30]	[−0.95]	[0.88]	[0.65]	

2.E Simulation Results

Figure 2.12: Continuous-Time Simulation Results without Microstructure

(Average every 5 minutes)

(a) $\mathbb{E}[|p(t)|]$ (b) $\sigma(t)$ (c) $\gamma(t)$

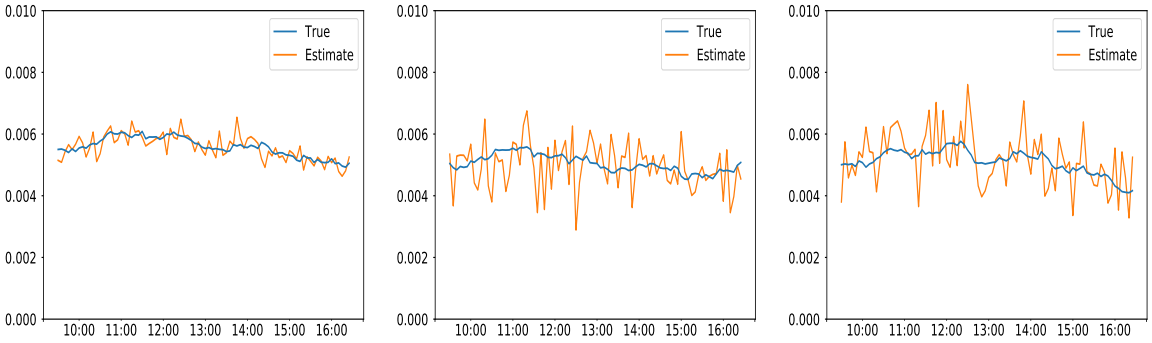
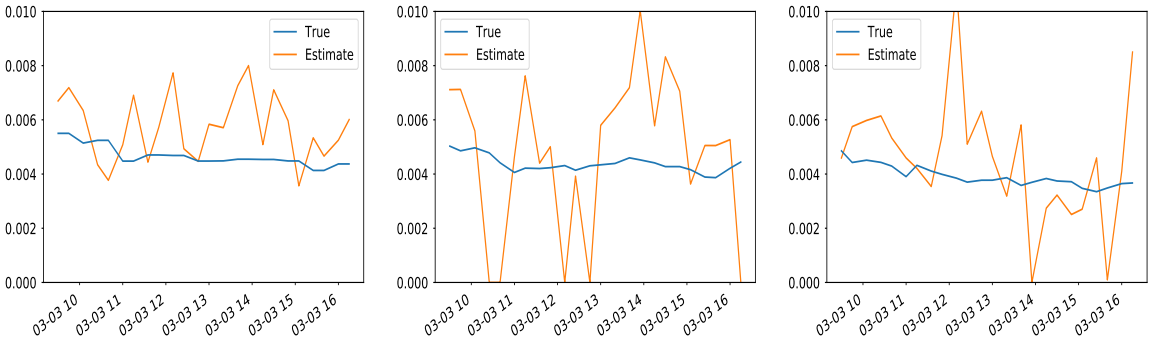


Figure 2.13: Continuous-Time Simulation Results with Microstructure

(Average every 5 minutes)

(a) $\mathbb{E}[|p(t)|]$ (b) $\sigma(t)$ (c) $\gamma(t)$



2.F Volatility: Empirical Results

Table 2.14: Vector Autoregression Models

	$\log(\sigma_t^2)$		$\log(\gamma_t^2)$	
VAR(1)				
Intercept	−0.84	(−1.04,−0.64)	−1.80	(−1.98,−1.62)
$\log(\sigma_{t-1}^2)$	0.56	(0.52, 0.59)	0.34	(0.31, 0.38)
$\log(\gamma_{t-1}^2)$	0.38	(0.33, 0.42)	0.48	(0.44, 0.52)
$\mathbb{R}^2$	24 %		72 %	
Innovation Covariance	$\begin{pmatrix} 0.33 & 0.19 \\ 0.19 & 0.27 \end{pmatrix}$			
VAR(6) — Chosen by SIC				
Intercept	−0.33	(−0.54,−0.11)	−0.88	(−0.73,−0.69)
$\log(\sigma_{t-1}^2)$	0.40	(0.36, 0.44)	0.24	(0.24, 0.27)
$\log(\gamma_{t-1}^2)$	0.25	(0.20, 0.29)	0.30	(0.19, 0.34)
$\log \sigma_{t-2}^2$	0.11	(0.07, 0.16)	0.01	(−0.06, 0.04)
$\log \gamma_{t-2}^2$	0.01	(−0.03, 0.06)	0.13	(0.09, 0.17)
$\log(\sigma_{t-3}^2)$	0.05	(0.01, 0.09)	−0.01	(−0.12, 0.03)
$\log(\gamma_{t-3}^2)$	−0.00	(−0.05, 0.04)	0.06	(0.02, 0.10)
$\log \sigma_{t-4}^2$	0.07	(0.02, 0.11)	−0.03	(−0.08, 0.01)
$\log \gamma_{t-4}^2$	0.01	(−0.03, 0.06)	0.11	(0.05, 0.15)
$\log(\sigma_{t-5}^2)$	0.03	(−0.01, 0.07)	−0.02	(−0.04, 0.02)
$\log(\gamma_{t-5}^2)$	0.04	(−0.00, 0.09)	0.14	(−0.01, 0.18)
$\mathbb{R}^2$	76 %		75 %	
Innovation Covariance	$\begin{pmatrix} 0.31 & 0.17 \\ 0.17 & 0.24 \end{pmatrix}$			

Table 2.15: $\mathbb{E}\left[rx_t \,\middle|\, \mathbf{1}\{\text{FOMC}\}_t, \sigma_t^2+\gamma_t^2, \frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}\right]$ (WLS)

Intercept	$\mathbf{1}\{\text{FOMC}\}_t$	$\log(\sigma_t^2+\gamma_t^2)$	$\log\left(\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}\right)$	$\log(\sigma_t^2+\gamma_t^2)\log\left(\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}\right)$	$\bar{\mathbb{R}}^2$
0.34	0.26				0.05 %
[14.43]	[1.44]				
−2.68		−0.28			1.83 %
[−6.93]		[−7.77]			
0.68		0.59			0.53 %
[8.49]		[4.14]			
−2.29		−0.27	0.51		2.20 %
[−5.46]		[−7.19]	[4.59]		
2.62		0.18	8.62	0.75	2.85 %
[2.48]		[1.89]	[4.92]	[4.63]	
−2.76	0.41	−6.76			1.93 %
[−6.76]	[2.13]	[−7.47]			
−2.37	0.39	−0.28	0.51		2.28 %
[−5.51]	[2.06]	[−7.17]	[3.60]		
2.38	0.36	0.17	8.50	0.74	2.91 %
[1.66]	[2.55]	[1.21]	[3.65]	[3.39]	

Table 2.16: News Premia Estimates Extended Results

Regressors			
Intercept	$\mathbf{1}\{\text{FOMC}\}_t$	$\log(\sigma_t^2 + \gamma_t^2)$	$\log\left(\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}\right)$
0.34	0.26		
[14.43]	[1.44]		
2.95		0.24	
[6.61]		[5.88]	
−2.45			−5.01
[−5.12]			[−5.86]
−5.04		0.14	−4.15
[−0.58]		[2.68]	[−4.93]
2.95	0.15	0.24	
[6.62]	[0.85]	[5.90]	
−5.10	0.33		−5.02
[−5.10]	[1.54]		[−5.81]
0.14	0.25	0.16	−3.52
[0.20]	[1.25]	[3.66]	[−5.35]

Table 2.17: Instrument Variables: First Stage Regression

$$\psi_t := \log\left(\sigma_t^2 + \gamma_t^2\right),\ \phi_t := \log\left(\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}\right)$$

Regressand	Intercept	ϕ_{t-1}	ϕ_{t-2}	ϕ_{t-5}	ϕ_{t-25}	ψ_{t-1}	ψ_{t-2}	ψ_{t-5}	ψ_{t-25}	$\psi_{t-1}\phi_{t-1}$	$\bar{\mathbb{R}}^2$	$\widehat{F}$
	−0.44	0.26									6.58 %	62.2
	[−25.84]	[7.88]										
$\log\left(\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2}\right)$	−0.28	0.18	0.16	0.12	0.06						11.53 %	110.0
	[−9.48]	[8.95]	[7.14]	[6.37]	[3.39]							
	−0.61	0.14	0.13	0.10	0.07	−0.06	−0.00	0.01	0.02		14.63 %	248.1
	[−8.39]	[8.38]	[6.86]	[5.51]	[4.07]	[−6.89]	[−0.25]	[1.91]	[3.14]			
	−0.23	0.73	0.11	0.10	0.07	−0.02	−0.00	0.01	0.02	0.06	15.49 %	525.4
	[−2.35]	[7.05]	[6.51]	[5.43]	[3.90]	[−1.51]	[−0.07]	[1.87]	[3.34]	[5.73]		
	−2.10					0.19					66.28 %	1986.4
	[−10.85]					[44.57]						
$\log\left(\sigma_t^2 + \gamma_t^2\right)$	−0.57					0.61	0.17	0.13	0.04		79.19 %	7712.2
	[−4.94]					[26.42]	[8.96]	[8.55]	[4.07]			
	−0.59	−0.15	0.00	0.07	0.06	0.60	0.16	0.13	0.05		79.27 %	9517.4
	[−4.99]	[−3.47]	[0.11]	[1.79]	[1.65]	[27.95]	[8.80]	[8.92]	[4.40]			
	−1.31	−1.23	0.02	0.07	0.07	0.53	0.16	0.13	0.05	−0.11	79.43 %	20 140
	[−5.84]	[−5.05]	[0.63]	[1.91]	[1.92]	[18.91]	[8.70]	[8.92]	[4.70]	[−4.43]		

Table 2.18: News Premia Estimates: Other Instruments

Regressors				Instruments			
Intercept	$\mathbf{1}\{\text{FOMC}\}_t$	$\log(\sigma_t^2+\gamma_t^2)$	$\log(\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2})$	$\mathbf{1}\{\text{FOMC}\}_t$	$\log(\frac{\gamma_{t-l}^2}{\sigma_{t-l}^2+\gamma_{t-l}^2})\ldots$	$\log(\sigma_{t-l}^2+\gamma_{t-l}^2)\ldots$	$\log(\frac{\gamma_{t-l}^2}{\sigma_{t-l}^2+\gamma_{t-l}^2})\log(\sigma_{t-l}^2+\gamma_{t-l}^2)$
				$l \in \{1,2,5,25\}$			
3.05		0.25			✓		
[6.83]		[6.07]					
3.01		0.25		✓	✓	✓	✓
[6.03]		[6.77]					
−2.02			−4.21			✓	
[−3.99]			[−4.63]				
−2.05			−4.28	✓	✓	✓	✓
[−5.28]			[−6.20]				
0.25		0.17	−3.47	✓	✓	✓	✓
[0.36]		[3.83]	[−5.01]				
0.11	0.25	0.16	−3.57	✓	✓	✓	✓
[0.14]	[1.24]	[3.59]	[−5.02]				
				$l = 1$			
3.10		0.25		✓			
[6.85]		[6.10]					
−2.71			−5.44			✓	
[−2.86]			[−3.20]				
−1.43		0.11	−5.22	✓	✓		
[−0.84]		[1.35]	[−2.94]				
−1.03	0.29	0.12	−4.76	✓	✓	✓	✓
[−0.76]	[1.36]	[1.68]	[−3.56]				

Table 2.19: News Premia Estimates in Levels

(Volatility is measured in yearly terms. (252 * daily)).
$l \in \{1, 2, 5, 25\}$.

Regressors					Instruments			
Intercept	$\mathbf{1}\{\text{FOMC}\}_t$	σ_t^2	γ_t^2	$(\sigma_t^2)(\gamma_t^2)$	$\mathbf{1}\{\text{FOMC}\}_t$	$\sigma_{t-l}^2 \ldots$	$\gamma_{t-l}^2 \ldots$	$(\sigma_{t-l}^2)(\gamma_{t-l}^2)$
0.28		0.08				✓		
[8.85]		[3.11]						
0.28		0.08			✓	✓	✓	✓
[8.80]		[2.74]						
0.27			0.07		✓			
[7.52]			[2.53]					
0.24			0.10		✓	✓	✓	✓
[6.63]			[3.33]					
0.31		0.16	−0.09			✓	✓	
[7.25]		[1.35]	[−0.77]					
0.24		0.02	0.09		✓	✓	✓	✓
[5.57]		[0.22]	[1.05]					
0.23		0.36	−0.50	−0.00	✓	✓	✓	✓
[5.75]		[1.39]	[−1.14]	[1.53]				
0.26	0.59	−0.34	−0.01		✓	✓	✓	✓
[5.69]	[2.78]	[−2.00]	[−3.18]					
0.27	0.34	0.71	−0.45	−0.01	✓	✓	✓	✓
[5.97]	[1.72]	[3.07]	[−2.41]	[−3.38]				

Table 2.20: News Premia Estimates: Robustness

	Regressors				Instruments			
	Intercept	$\mathbf{1}\{\text{FOMC}\}_t$	$\log(\sigma_t^2+\gamma_t^2)$	$\log(\frac{\gamma_t^2}{\sigma_t^2+\gamma_t^2})$	$\mathbf{1}\{\text{FOMC}\}_t$	$\log(\frac{\gamma_{t-l}^2}{\sigma_{t-l}^2+\gamma_{t-l}^2})\ldots$	$\log(\sigma_{t-l}^2+\gamma_{t-l}^2)\ldots$	$\log(\frac{\gamma_{t-l}^2}{\sigma_{t-l}^2+\gamma_{t-l}^2})\cdot\log(\sigma_{t-l}^2+\gamma_{t-l}^2)$
Sub-period Analysis								
2003–2007	−2.59	0.69	0.08	−6.93	✓	✓	✓	✓
	[−0.67]	[1.58]	[0.36]	[−2.05]				
2008–2012	0.17	0.79	0.06	−1.56	✓	✓	✓	✓
	[0.10]	[1.82]	[0.55]	[−1.33]				
2013–2007/9	3.71	−0.26	0.40	−1.94	✓	✓	✓	✓
	[2.50]	[−1.05]	[4.40]	[−1.66]				
Unweighted Analysis								
	0.63		0.06		✓	✓	✓	✓
	[0.82]		[0.77]					
	0.03			−0.04	✓	✓	✓	✓
	[0.06]			[−0.04]				
	−0.77		−0.02	−1.14	✓	✓	✓	✓
	[−0.62]		[−0.17]	[−1.50]				
	−1.42	0.93	−0.09	−0.81	✓	✓	✓	✓
	[−1.13]	[3.02]	[−0.99]	[−0.89]				

Chapter 3

BYPASSING THE CURSE OF DIMENSIONALITY: FEASIBLE MULTIVARIATE DENSITY ESTIMATION

By Minsu Chang and Paul Sangrey

Most economic data are multivariate making estimating multivariate densities a classic problem in the literature. However, given vector-valued data — $\{x_t\}_{t=1}^T$ — the *curse of dimensionality* makes nonparametrically estimating the data's density infeasible when the number of series, D, is large. Hence, we do not seek to provide estimators that perform well all of the time (it is impossible), but rather seek to provide estimators that perform well most of the time. We adapt the ideas in the Bayesian compression literature to density estimation by randomly binning the data. The binning randomly determines both the number of bins and which observation is placed in which bin. This novel procedure induces a simple mixture representation for the data's density. For any finite number of periods, T, the number of mixture components used is random. We construct a bound for this variable as a function of T that holds with high probability. We adopt the nonparametric Bayesian framework and construct a computationally efficient density estimator using Dirichlet processes. Since the number of mixture components is the key determinant of our model's complexity, our estimator's convergence rates — $\sqrt{\log(T)}/\sqrt{T}$ in the un-

conditional case and $\log(T)/\sqrt{T}$ in the conditional case — depend on D only through the constant term. We then analyze our estimators' performance in a monthly macroeconomic panel. Our procedure performs well in capturing the data's stylized features such as time-varying volatility and skewness.

3.1 Introduction

Estimating multivariate densities is a classic problem across econometrics, statistics, and computer science. Researchers often find parametric assumptions restrictive and their models sensitive to deviations from these assumptions. On the other hand, given vector-valued data — $\{x_t\}_{t=1}^T$ — nonparametrically estimating the data's density is infeasible when the number of series, D, is large. This phenomenon is called the *curse of dimensionality*.

Nonparametric estimators simultaneously solve two problems. First, they approximate the density. Second, they estimate the parameters that govern this approximation. The original curse of dimensionality papers, such as Stone (1980, 1982), examine the approximation problem through the lens of the deterministic approximation literature. They show that requiring the estimators to be consistent causes the estimator and the deterministic approximation to use the same number of terms asymptotically. Solving this deterministic problem requires $T^{g(D)}$ terms for some g that depends upon the set of functions under consideration. To understand this, consider creating a multidimensional histogram. Dividing a D-dimensional hypercube into small hypercubes with width $1/T$ requires T^D terms. The various deterministic approximations essentially form these high-dimensional histograms asymptotically. The precise form of the "histogram" being formed depends on the application. In the years since Stone (1980), the minimax estimation literature has focused on mapping various applications to these high-dimensional histograms, e.g., Yang and Barron (1999) and Ichimura and Todd (2007).

Over the same period, various other authors have studied how random approximations behave in high dimensions,e.g., Johnson and Lindenstrauss (1984), Klartag and Mendelson (2005), Boucheron, Lugosi, and Massart (2013), and Talagrand (2014).

This Bayesian compression literature develops parsimonious random approximations to high-dimensional datasets. Since high-dimensional random variables cluster on balls instead of hypercubes, the question is how should we approximate high-dimensional balls, not high-dimensional hypercubes. (We provide intuition below on both why random data tends to cluster on balls and why this dramatically simplifies the problem.)

Thus far, the Bayesian compression literature has focused on the approximation problem and the closely related data compression problem. For example, Koop, Korobilis, and Pettenuzzo (2019) compress hundreds of variables and compute vector autoregressions on the compressed data. However, unlike the deterministic approximation case, no one has yet applied these ideas to density estimation. We apply these ideas and develop parsimonious high-dimensional approximations to feasibly estimate multivariate densities.

In particular, we develop a dynamic generalization of the infinite-mixture representation commonly used in the Bayesian nonparametric literature, (Ghosal and van der Vaart 2017), as an alternative to current Bayesian conditional density estimators, e.g., Geweke and Keane (2007), Norets (2010), and Pati, Dunson, and Tokdar (2013). Infinite mixtures are commonly used to flexibly approximate cross-sectional densities, (Ghosal, Ghosh, and van der Vaart 2000; van der Vaart and van Zanten 2008). Because infinite-mixtures can approximate a broad class of densities, this procedure only requires a few assumptions on the data generating process (DGP). We can estimate both unconditional and transition densities for both i.i.d. and Markov data.

We apply the results from the Bayesian compression literature to nonparametric density estimation in a series of steps. First, we construct a novel method for approximating high-dimensional balls that bins the data and endogenously determines both the number of bins and which vector — x_t — goes into which bin. Second, we show that this random binning induces an approximating mixture representation that is close to the true density.

For any finite T, we construct a bound for the number of mixture components as a function of T that holds with high probability. It is impossible to create a nonparametric estimator that is always parsimonious. This probability is with respect to the data-agnostic procedure that determines the number of mixture components. We convert these bounds on the mixture's complexity into convergence rates for the

estimators. Our estimators' convergences rates — $\sqrt{\log(T)}/\sqrt{T}$ in the unconditional case and $\log(T)/\sqrt{T}$ in the conditional case — depend on D only through the constant term.

To summarize, we show that our estimator converges rapidly — it does not require many mixture components even when D is large — with arbitrarily high probability. We do this by tolerating a small chance of our estimator's converging slowly. Even though we cannot beat the minimax rate in general, we show that our estimators will perform well even when D is large and the true distribution is not smooth. In particular, we show that distance between the induced mixture representation and the data's true distribution, as measured by standard divergences, such as Hellinger and Kullback-Leibler, is small even when we take supremum over the set of true DGPs and D is large.

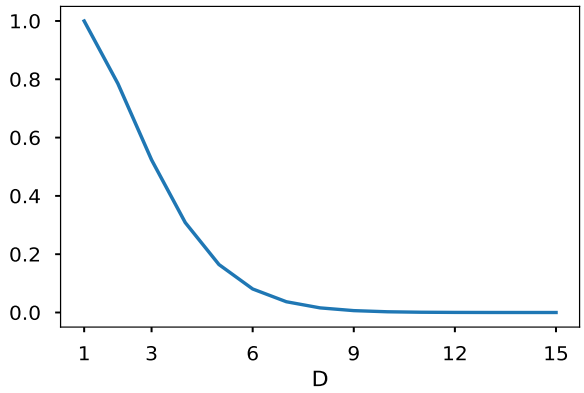
We organize the paper as follows. Section 3.2 provides the intuition underlying the results in the Bayesian compression literature, and consequently our results. Section 3.3 describes the data generating process. Section 3.4 constructs the sieve and provides conditions under which it approximates the true density well. Section 3.5 proves our estimators converge at the rates given above. Section 3.6 provides a computationally efficient Gibbs sampling algorithm to estimate our sieve. Section 3.7 introduces the data and the prior we use for the empirical analysis. Section 3.8 uses our method to empirically analyze a monthly macroeconomic panel showing our method works well in practice. Section 3.9 concludes. The appendices contain the proofs and additional empirical results.

3.2 Intuition

The convergence rates discussed above likely seem surprising, so we now explain why they are reasonable. We do this by discussing the intuition that drives the results in the Bayesian compression literature. As discussed above, the standard convergence rates are consequences of the number of bins of width $1/T$ required to fill a D-dimensional hypercube equaling T^D. The Bayesian compression algorithms use fewer terms than the deterministic approximations do by exploiting two facts. First, random data tend to cluster in balls. Second, the volume of a D-dimensional

ball grows exponentially slower with the number of series than the hypercube does as shown in Figure 3.1. We exploit this behavior by constructing a sieve for the D-dimensional ball instead of constructing a sieve for the D-dimensional hypercube as the literature usually does. Since the volume of the ball grows more slowly, our sieve requires far fewer terms, especially when D is large.

Figure 3.1: Volume of a Ball Relative to a Hypercube [43]



The goal in this paper is to exploit this simplicity to bound the number of terms required to estimate a density, instead of just compressing the data. In other words, similar to how Stone (1980) created a sieve for the hypercube, we construct a sieve for the ball. Previous methods have shown how to compress the data while only slightly perturbing the data's first two sample moments. We construct a sparse discretization operator (i.e., we bin the data) that does not significantly perturb the data's first two sample moments. To convert this distance between the sample moments into a distance between densities, we use the fact that as long as a process is locally asymptotically normal, its first two moments form a sufficient statistic for the density. Consequently, if the first two moments are close, the densities must be close as well.

We build a Dirichlet mixture process and adopt the standard Bayesian mixture framework. The number of mixture components determines the complexity of a Gaussian mixture and the estimator's convergence rate. For any fixed sample-size T, this is a random variable. Hence, we have a series of distributions indexed by T.

[43] The ratio between the volume of a ball of hypercube with the same diameter equals $\frac{\pi^{\frac{D}{2}}}{2^D \Gamma\left(\frac{D}{2}+1\right)}$, where Γ denotes the Gamma function.

As mentioned above, we can view the distance between the estimator and the truth itself as a random variable whose distribution is indexed by T. The critical difference between our results and the previous ones in the literature is that we only require the convergence rate to hold in a $1-2\delta$ probability region with respect to the prior. In other words, we want our estimator to converge rapidly "most of the time" where "most" means with probability at least $1-2\delta$ and this probability is only taken with respect to the prior. We still require the convergence rate to be uniform with respect to the likelihood.

Because the previous literature requires the convergence rate to be uniform with respect to randomness in the prior, they cannot exploit the smoothness that the prior induces in deriving their convergence rates. At a technical level, for any fixed T our sieve is not a measurable function of the data and so the Stone (1980) bounds do not apply.

3.3 Data Generating Process

Consider a D-dimensional time series: $X_T \coloneqq \{x_t\}_{t=1}^T$. Assume that X_T is first-order hidden Markov and is a Gaussian process. That is, there exists a latent state z_t, such that (x_t, z_t) are jointly Markov. The z_t may be a constant. We want to estimate x_t's conditional densities for $t = 1, \ldots, T$. Let $\mathcal{F}_t$ denote the time-t information set. We denote the true distribution P_T and the approximating distribution Q_T. They have associated densities p_T and q_T.

Definition 3.1 (Data Generating Process)**.**

$$p_T\left(x_t \,|\, \mathcal{F}_{t-1}\right) \coloneqq \sum_{k=1}^{\infty} \Pi_{t-1,k} \phi\left(x_t \,|\, x_{t-1}\beta_{k,t}, \Sigma_{k,t}\right). \tag{3.1}$$

Since X_T is Gaussian process, the conditional density — $p_T(x_t \,|\, F_{t-1})$ — has an infinite Gaussian mixture representation for each time period. Each mixture component has associated mixture probability, $\Pi_{t-1,k}$, and component-specific parameters, $\beta_{k,t}$, and $\Sigma_{k,t}$. We assume that all x_t have finite means and variances. The Gaussian process assumption is quite general allowing, for example, for both multiple modes and fat tails. We let the true DGP depend upon T because at this point we are only

approximating the density for a fixed T. Of course, the Markov implies a consistency condition across p_T for all T.

Definition 3.2 (Approximating Model)**.**

$$q_T\left(x_t \mid \mathcal{F}_{t-1}\right) \coloneqq \sum_{k=1}^{K_T} \Pi_{t-1,k} \phi\left(x_t \mid x_{t-1}\beta_k, \Sigma_k\right). \tag{3.2}$$

The approximating model is a Gaussian mixture with K_T components, and so K_T governs the complexity of the model. As one would expect, K_T grows with T. Second, each cluster's components, (β_k, Σ_k), no longer have time t subscripts. The idea is that we can reuse the latent variables $(\beta_{k,t}, \Sigma_{k,t})$ across time without loss of generality. If two separate periods have sufficiently similar dynamics, we group them into one component with the same parameters. Since the clusters are defined differently in the true and approximating models, no simple relationship the parameters exists in general.

Throughout, we use μ_T to refer to the $T \times D$ mean matrix. We also consider the rescaled data:

$$\widetilde{X}_T \coloneqq \frac{X_T - \mu_T}{\sqrt{\|X_T - \mu_T\|_{L_2}}} \in S^{TD-1} = \left\{x \in \mathbb{R}^{TD} \,\middle|\, \|X\|_{L_2} = 1\right\}, \tag{3.3}$$

where $\|\cdot\|_{L_2}$ is the L_2-norm. Since we are on the unit hypersphere, we are in a compact space for any fixed T. Since $X_T - \mu_T$ is a zero-mean Gaussian process, its $TD \times TD$ covariance matrix completely determines its distribution. We define the densities of $\widetilde{X}_T$ as we did for X_T above and denote them $\widetilde{p}_T$ and $\widetilde{q}_T$.

3.4 Sieve Construction

Setting up the Problem

We construct a sieve in this section that approximates a wide variety of data generating processes while still being as simple as possible. By simple, we mean that the metric entropy of these approximating models grows slowly with the number of datapoints. This property is useful because metric entropy controls the rate at which posteriors converge as shown by Ghosal, Ghosh, and van der Vaart (2000) and Shen

and Wasserman (2001). It also controls the minimax rate at which estimators can converge, (Wong and Shen 1995; Yang and Barron 1999).

We approximate a marginal density in the space of densities over $\mathbb{R}^D$ — $\mathscr{P}(\mathbb{R}^D)$ — and a transition density which lies in associated the product space — $\mathscr{P}(\mathbb{R}^D)\times\mathscr{P}(\mathbb{R}^D)$. These approximations problems are not well-posed because multiple equivalent representations exist for each density given X_T that satisfy some bound on the distance to p_T in some metric on $\mathscr{P}\left(R^D\right)$. We can exploit this multiplicity by choosing a representation that is particularly amenable to estimation for each T. We want the most parsimonious density that still approximates well.

We construct our sieve as follows. Given some $\epsilon > 0$, we construct a mapping Θ_T that takes the TD-dimensional hypersphere and maps it onto a KD-dimensional hypersphere, where $K \ll T$. This mapping only perturbs the norms of the individual elements by at most ϵ. In other words, we construct an ϵ-isometry.

We then show the densities are also not perturbed significantly in Theorem 3.2. This result is true whenever the norm of the data matrix is a locally sufficient statistic for the density. In other words, we can use bounds on divergences between the norms, $\{\|\widetilde{x}_t\|_{L_2}\}$, to bound divergences between the densities in $\mathscr{P}(\widetilde{X}_T)$.

Bounding the Norm Perturbation

We construct our approximate sufficient statistic for $\widetilde{X}_T$ by "projecting" it onto a lower-dimensional space. The only reason this projection intuition is not exact is that the target space is not a subspace of the original space. We need the compressed data to have a mixture distribution. Hence, the compression operator Θ_T must be a discretization operator. A mixture distribution for some collection of data $\widetilde{X}_T$ is a random binning of the data where the data in each bin has the same parametric distribution. The question is how to construct the bins.

A standard discretization operator with K bins is a $T \times K$ matrix where each row θ_t contains exactly one 1 and the rest of the elements equal zero. A variable x_t is in bin k if and only if $\theta_{t,k} = 1$, i.e., Θ_T has a 1 in row t column k. We cannot use a standard discretization operator for two reasons. First, since all of the elements are weakly positive $\mathbb{E}[\theta_{t,k}] \neq 0$. Second, once we see a 1, the rest of the columns in the

row must be identically zero. This property makes the columns too dependent for our results to hold.

Fixing the first issue is relatively straightforward. We let $\theta_{t,k}$ take on values from $\{-1, 0, 1\}$. Each x_t is in bin k if $\theta_{t,k} = 1$ and in bin $K + k$ if $\theta_{t,k} = -1$. There is no reason the elements of θ must be positive. The second issue is more problematic. We let each row have as many 1's and -1's as necessary. Once we do this, seeing a 1 in column k gives us no information about columns $k + 1$ through K. It does complicate the analysis slightly, however. We are letting each period be in more than one component simultaneously. In other words, we do not just create a mixture distribution across periods but also create one in each period.

To make the discussion in the previous few paragraphs more formal, we now define the random operator Θ_T. We use a stick-breaking process to construct Θ_T, adapting the form often used to construct Dirichlet processes, as proposed by Sethuraman (1994).[44]

Definition 3.3 (Θ_T Operator)**.** Let b be a Bernoulli random variable with $\Pr(b = 1) \in (0, 1)$. Draw another random variable $\chi \in \{-1, 1\}$ with probability 1/2 each. Let $T \in \mathbb{N}$ be given. Draw T variables $\theta \coloneqq \chi \cdot b$ independently of all of the previous values, and form them into a column-vector — Θ_1. Form another column vector Θ_2 the same way and append it to the right of Θ_T. Continue this until all of the rows of Θ_T contain at least one nonzero element.

We form the Θ_T operator this way so that $\mathbb{E}[\theta] = 0$ and $\mathbb{V}\text{ar}(\theta) = \mathbb{E}[|\theta|] = \Pr(b = 1)$. Furthermore, its rows are independent and its columns form a martingale-difference sequence. The only dependence between the columns of Θ_T arises through the stopping rule, and stopped martingales are still martingales. In addition, Θ_T is independent of $\widetilde{X}_T$. Since Θ_T is discrete, Θ_T implicitly clusters $\widetilde{X}_T$. Consider some row θ_t of Θ_T. For each column of θ_t, define a bin as $|\theta_{t,k}| \times \text{sign}(\theta_{t,k})$. Clearly, if Θ_T has K_T columns, there are $2K_T$ possible total bins.

Our analysis requires a tight bound on the tail behavior of K_T. To create such a bound, we must understand its distribution. By Lemma 3.11, the probability density

[44] In Section 3.4, we show that a Dirichlet process can replace Θ_T without affecting our results. The intuition behind this is that we can construct both of them using similar stick-breaking processes. Consequently, they are mutually absolutely continuous, and so we a density that converts the measures exits.

function of K_T is

$$\Pr(K_T \leq \widetilde{K}) \propto (1 - (1 - \Pr(b = 1))^{\widetilde{K}})^T. \tag{3.4}$$

Furthermore, we show in Lemma 3.12 that $K_T \propto \log(T)$ with high probability. The intuition behind this is that to get $K = \widetilde{K}$ the Bernoulli random variable must have $\widetilde{K}$ failures. The probability of this occurring declines exponentially fast in $\widetilde{K}$. This logarithmic growth is relied upon extensively in what follows.

We claimed above that Θ_T constructs an approximate sufficient statistic by binning $\widetilde{X}_T$. In other words, we are compressing the data. Equation (3.4) quantifies the amount by which we compress the data. Instead of considering each of the T values of x_t separately, we can bin them into K_T bins, and we can treat the values in each bin identically. Since $K_T \propto \log(T) \ll T$ this substantially reduces the complexity.

We also must show that Θ_T preserves the $\widetilde{x}_t$'s densities. It is not a sufficient statistic if we lose any necessary information. We turn to this now.

Theorem 3.1 (Bounding the Norm Perturbation)**.** *Let Θ_T be constructed as in Definition 3.3 with the number of columns denoted by K_T. Let $\epsilon > 0$ be given. Let $0 < \delta < 1$ be given such that $0 < \log(\frac{1}{\delta}) < c_1 \epsilon^2 K_T$ for some constant c_1. Let $\widetilde{X}_T$ be in the unit hypersphere in $\mathbb{R}^{TD-1}$. Then with probability greater than $1 - 2\delta$ with respect to Θ_T, there exists a constant c_2 such that for any $\epsilon > \sqrt{\frac{\log T}{K_T}}$,*

$$\sup_t \left| \|\theta_t \widetilde{x}_t\|_{L_2} - \|\widetilde{x}_t\|_{L_2} \right| < c_2 \left(1 + \log\left(\frac{1}{\delta}\right)\right) \epsilon.$$

Theorem 3.1 implies that if we choose Θ_T with the number of columns $K \propto \log(T)$, applying Θ_t perturbs the norms of $\widetilde{x}_t$ by at most ϵ. This result holds with probability at least $1 - 2\delta$ with respect to the distribution over Θ_T. Since $\widetilde{X}_T \in S^{TD-1}$, we can map S^{TD-1} onto a smaller space $S^{K_T D-1}$, with $K \ll T$, without perturbing the individual elements' norms significantly.

The basic idea here is that we are pre-multiplying the data by a martingale, i.e., a process which expectation equal to one. This does not affect the mean or the variance. This increases randomness "smooths" the data. To gain intuition, one can think about the average value. Koop, Korobilis, and Pettenuzzo (2019) do precisely this, focusing on Bayesian model averaging. This allows us to get very tight bounds on the tails of the distribution with high probability. Since we have not changed the

first two population moments and can tightly bound the tails of the distribution, we can place strong bounds on how we moved the sample moments. This is precisely what Theorem 3.1 does.

Distances on the Space of Densities

In the previous section, we showed that Θ_T does not affect $\widetilde{x}_t$'s norms significantly. These norms are not inherently interesting objects. Instead, they are interesting because they form a sufficient statistic for the Gaussian process. To show the densities are close, we must convert the distances between the norms into distances on $\mathscr{P}(\widetilde{X}_T)$.

The compressed data, $\Theta_T'\widetilde{X}_T$, has a distribution conditional on Θ_T. Since $\widetilde{X}_T$ is a normalized Gaussian process and Θ_T is a matrix, this process is Gaussian. Hence, there exists a distribution for $\widetilde{X}_T$ constructed by integrating out Θ_T. This integration creates an approximating distribution for $\widetilde{X}_T$: $\widetilde{Q}_T$.

Since Θ_T is almost surely discrete, this approximating distribution is a mixture, as in Definition 3.2. We represent it as an integral with respect to a latent mixing measure — G_t^Q — for each t. The parameters in each component are means and covariances, and so the G_t^Q measure is over the space of means and covariances. Because Θ_T can have more than one non-zero element, G_t^Q is a mixture distribution in each period, even conditional on Θ_T.

Let G^Q be the latent mixing measure over the space of G_t^Q. That is, each G_t^Q is a draw from G^Q. Since latent mixing measures are almost surely discrete, the G_t^Q share the same atoms. This dependence regularizes the mixing measures across time, i.e., it "smooths" the approximating model. However, since the atoms of G^Q are left arbitrary, it does not restrict the set of DGPs that we can approximate well.

Let δ_t^Q denote the mixture identity that determines which cluster contains Σ_t. Let $\phi(\cdot \,|\, \delta_t^Q)$ denote the mean-zero multivariate Gaussian density with covariance Σ_t. Then $\widetilde{Q}_T$ can be expressed as

$$\widetilde{q}_T(\widetilde{\mathcal{X}}) = \int_G \int_{G_t} \phi\left(\widetilde{x}_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)\, dG^Q(dG_t^Q). \tag{3.5}$$

Likewise, if we replace q with p, we write the true model's density, $\widetilde{p}_T$, as

$$\widetilde{p}_T(\widetilde{\mathcal{X}}) = \int_G \int_{G_t} \phi\left(\widetilde{x}_t \,\middle|\, \delta_t^P\right) dG_t^P(\delta_t^P)\, dG^P\left(dG_t^P\right), \tag{3.6}$$

with its associated latent mixing measures and mixture identities. Note, The approximating cluster identities, $\{\delta^Q{}_t\}_{t=1}^T$, are different than the true cluster identities, $\{\delta^P{}_t\}_{t=1}^T$, because Θ_T induces Q's clustering. It is not induced by the underlying true clustering.

Since the densities are mixtures parameterized by their covariances, we must convert a clustering in $\widetilde{x}_t$-space into a clustering in Σ_t-space in order to convert the bounds above into bounds on the densities. The norms of $\widetilde{x}_t$ and $\widetilde{x}_{t^*}$ being close does not imply that the associated matrix norms for Σ_t and Σ_{t^*} are close. Consequently, we cluster $\Sigma_t^{-1/2}\widetilde{x}_t$ directly.

The error bound Theorem 3.1 provides is independent of δ_t^P and so it does not depend on Σ_t. In other words, for times t, t^* such that the associated $\widetilde{x}_t$ and $\widetilde{x}_{t^*}$ are contained in the same cluster, δ_k^Q, the following holds:[45]

$$\sup_{t,t^* \in \delta_k^Q} \left| \widetilde{x}_t \Sigma_t^{-1} \widetilde{x}_t - \widetilde{x}'_{t^*} \Sigma_{t^*}^{-1} \widetilde{x}_{t^*} \right| < \epsilon. \tag{3.7}$$

Here ϵ is independent of t, t^*, and the cluster identity. The right-hand side of (3.7) is a "distance" on the space of covariance matrices. That is, we introduce the following semimetric on the space of covariance matrices.[46]

Definition 3.4 (Weighted-L_2 Semimetric)**.**

$$\delta_{wl_2}(\Sigma_k, \Omega_k) \coloneqq \sup_{t,t^* \in \delta_k^Q} \left| \widetilde{x}'_t \Sigma_k^{-1} \widetilde{x}_t - \widetilde{x}'_{t^*} \Omega_k^{-1} \widetilde{x}_{t^*} \right|. \tag{3.8}$$

Note, δ_{wl_2} is compatible with, and weaker than, the max-norm.[47] The max-norm is equivalent to the L_2-norm up to a scale transformation, and the relevant scale is a constant since we only consider full-rank matrices. Hence, the space of covariance matrices forms a Polish space because the space of $D \times D$ matrices is isomorphic to $\mathbb{R}^{D \times D}$ and we are choosing an open subset of that space. In other words, δ_{wl_2} constructs a set of equivalence classes over the space of covariance matrices, where

[45] We abuse notation slightly and use $t \in \delta_k^Q$ if the cluster identity associated with x_t equals δ_k^Q.

[46] It is a semimetric because we can have $\Sigma \neq \Omega$ but $\delta_{wl_2}(\Sigma, \Omega) = 0$. The two matrices may differ that cannot be identified by the set $x \in$ cluster k.

[47] If x, y in $x\Sigma^{-1}y$ are (possibly) different unit selection vectors we can pick out the maximum absolute deviation between elements in the two matrices. This difference is clearly at least as big as the δ_{wl_2} because that semimetric requires x, y to be the same.

two sample covariances are equivalent if the implied second-moment behavior of the $\{\widetilde{x}_t \in \delta_k^Q\}$ is indistinguishable.

Definition 3.4 converts bounds on the norms into $\widetilde{x}_t$ into bounds on covariances. We must convert this bound to a bound on densities. The distance we use here is the Hellinger distance.

Definition 3.5. Hellinger Distance

$$h(p,q) \coloneqq \frac{1}{\sqrt{2}}\sqrt{\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx}. \tag{3.9}$$

The Hellinger distance is useful because it is a valid norm on the space of densities. Since the covariance matrix is a sufficient statistic for a centered Gaussian process, we can convert bounds between the covariances into bounds in this distance. Instead of applying this directly to the joint distribution, we take the supremum over the conditional distributions.

Definition 3.6 (Supremum Hellinger Distance)**.**

$$h_\infty^2(p,q) \coloneqq \sup_{\mathcal{F}_{t-1}^P, \mathcal{F}_{t-1}^Q, 1\le t\le T} h^2\left(p\left(\cdot \,\middle|\, \mathcal{F}_{t-1}^P\right), q\left(\cdot \,\middle|\, \mathcal{F}_{t-1}^Q\right)\right). \tag{3.10}$$

The supremum Hellinger distance will prove useful because it is stronger than both the Hellinger distance and the Kullback-Leibler divergence applied to the joint density. As a consequence, once we bound h_∞, we can directly deduce other bounds as necessary.

Representing the Joint Density

We now show that the approximating distribution of $\widetilde{X}_T$ induced by Θ_T is close to the true distribution $\widetilde{P}_T$ using h_∞. We can do this whenever the rescaled trace is a locally sufficient statistic for the density. Hence, we can use bounds on divergences in $\widetilde{\mathcal{X}}$ to bound divergences in $\mathcal{P}(\widetilde{\mathcal{X}})$.

Theorem 3.2 (Representing the Joint Density)**.** *Let $\widetilde{X}_T \coloneqq \widetilde{x}_1, \ldots, \widetilde{x}_T$ be a D-dimensional Gaussian process with period-t finite stochastic means, μ_t, and covariances, Σ_t. Let Σ_t be positive-definite for all t. Let Θ_T be the generalized selection*

matrix constructed in Definition 3.3. Let $\widetilde{P}_T$ denote the distribution of $\widetilde{X}_T$. Then given $\epsilon > 0$ and $\delta \in (0,1)$, the approximating distribution, Q_T, which is the mixture distribution over $\widetilde{\mathcal{X}}$ that Θ_T induces, satisfies the following with probability at least $1 - 2\delta$ with respect to Θ_T for some constant C:

$$h_\infty\left(\widetilde{P}_T(\widetilde{\mathcal{X}}), \widetilde{Q}_T(\widetilde{\mathcal{X}})\right) < C\left(1 + \log\left(\frac{1}{\delta}\right)\right)\epsilon.$$

We represent the joint density as follows. Since $\widetilde{\mathcal{X}}$ lives in S^{TD-1}, we start by mapping S^{TD-1} onto a smaller space $S^{K_T D-1}$ where $K_T \ll T$. This argument is very similar to the various projection arguments that the literature makes when it projects S^{TD-1} into a "smaller" space. However, the operator Θ_T we use does not form a projection because it is not mapping the space onto itself. The unit sphere in $\mathbb{R}^{K_T D}$ is not a subset of the one in $\mathbb{R}^{TD}$.

Unlike the previous compression operators in the literature, Θ_T is discrete, and so it clusters $\widetilde{x}_t$. This property implies that the density of $\widetilde{x}_t$ is a process with respect to a discrete measure. That is, Q_T is a mixture distribution. Also, we show in Section 3.4, that we can assume that this latent measure is Dirichlet without loss of generality. In other words, our method represents the $\widetilde{X}_T$ process as an integral with respect to a Dirichlet process. Consequently, since $\widetilde{X}_T$ is a Gaussian process, and hence locally Gaussian, we can represent $\widetilde{X}_T$ using a Gaussian mixture process whose latent mixing measure is a Dirichlet process.

The leading issue that remains is that we bounded the rescaled $\widetilde{X}_T$, not X_T. As one might expect, estimating the true joint density of X_T is impossible. Since $\|X_T\|^2 \propto T$, the bound we have is of the order $\sqrt{T}\epsilon$, which is useless. Instead, we consider simpler quantities such as X_T's marginal density (Section 3.4) and transition density (Section 3.4). We show that sample means of the marginal and transition densities converge to those implied by Q_T, and hence those implied by P_T. This convergence occurs because sample means converge to population means.

Representing the Marginal Density

We now derive a representation for the marginal density of X_T from the representation for the joint density. We first consider the case where the true density has a

product form, i.e., the data are independent. The intuition behind the proof is that Theorem 3.2 implies that $T\epsilon^2$ bounds the maximum deviation of the approximating density. Standard arguments about the convergence of means for product measures give a $\frac{1}{T}$ term. Hence, the deviation between the the the means is bounded by ϵ^2. We use the Hellinger distance here instead of the sup-Hellinger distance because there is no conditioning information we need to take the supremum over.

Theorem 3.3 (Representing the Marginal Density)**.** *Let $x_1, \ldots, x_T$ be drawn independently from P_T where each x_t has a infinite-Gaussian mixture representation. Let Θ_T be constructed as in Theorem 3.2 for each t. Let ϵ be given. Construct Q_T by using the Θ_T operator to group the data and letting the data be Gaussian distributed within each component with component-wise means and covariances given by their conditional expectations. Then, with probability $1-2\delta$ with respect to Θ_T, there exists a constant C such that the following holds uniformly over T*

$$h\left(\int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^P\right) dG_t(\delta_t^P), \int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t(\delta_t^Q)\right) < C\left(1 + \log\left(\frac{1}{\delta}\right)\right)\epsilon.$$

We now extend Theorem 3.3 to the non-i.i.d. case. The hidden Markov assumption implies that the transitions are conditionally i.i.d. and this conditioning does not affect the convergence rate because we have a supremum-norm bound on the deviations in the joint density. Uniform ergodicity implies that the sample marginal density converges to the true marginal density. Consequently, using hidden Markov data instead of independent data does not affect the approximation results.

Corollary 3.1 (Representing the Marginal Density with Markov Data)**.** *Theorem 3.3 continues to hold when the x_t form a uniformly ergodic hidden Markov chain instead of being fully independent.*

Representing the Transition Density

We now show our model approximates transition densities well. Since the data are Markov, we can construct the sample transition density as an average of the transitions in the data. Component by component, we solve for the correct conditional distributions in the approximating model. Similar to above, we relate the error in the transition densities and the error for the joint densities. We can consider the space of

transitions as the product space: $\widetilde{X}_T \otimes \widetilde{X}_T$. We can construct the marginal density in the space. As in Section 3.4, the approximate product form gives us a $1/T$ term in the convergence rate. Theorem 3.2 gives us a $T\epsilon^2$ term. The T terms cancel, and so ϵ^2 bounds the distance between the densities.

Theorem 3.4 (Transition Density Representation)**.** *Let $x_1, \ldots, x_T \in R^{T \times D}$ be a uniformly ergodic Markov Gaussian process with density p_T. Let $\epsilon > 0$ be given. Let $K \geq c \log(T)^2/\epsilon$ for some constant c. Let δ_t be the cluster identity at time t. Then there exists a mixture density q_T with K clusters with the following form:*

$$q_T\left(x_t \,|\, x_{t-1}, \delta_{t-1}\right) \coloneqq \sum_{k=1}^{K} \phi\left(\beta_k x_{t-1}, \Sigma_k\right) \Pr\left(\delta_t = k \,|\, \delta_{t-1}\right).$$

Construct $q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^{Q}\right)$ from $q_T\left(x_t \,|\, x_{t-1}, \delta_{t-1}\right)$ by integrating out δ_{t-1} using $\Pr(\delta_{t-1} \,|\, X_T)$. Then with probability $1 - 2\delta$ with respect to the prior

$$h_\infty\left(p_T\left(x_t \,|\, \mathcal{F}_{t-1}^{P}\right), q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^{Q}\right)\right) < C\sqrt{1 + \log\left(\frac{1}{\delta}\right)}\epsilon.$$

Replacing Θ_T with a Dirichlet Process

The previous subsections use Θ_T to construct an approximating representation that is arbitrarily close to the truth. We want to construct an estimator that takes this representation to the data. (We do not claim that the representation is unique.) Here we argue that Θ_T can be chosen to be a Dirichlet process without loss of generality.

Consider the Θ_T process as in Definition 3.3 except we no longer stop when we no longer need columns. Then we can replace Θ_T with a Dirichlet process without altering the results. By doing this we can use standard Dirichlet-based samplers to estimate the sieve. In particular, it shows that the nonparametric Bayesian marginal density estimators in the literature satisfy the requirements of our theory, (Ghosal, Ghosh, and van der Vaart 2000; Walker 2007).

Lemma 3.5 (Replacing Θ_T with a Dirichlet Process)**.** *Let Q be a mixture distribution representable as an integral with respect to the Θ_T process defined in Definition 3.2. Then Q has a mixture representation as an integral with respect to the Dirichlet process.*

The intuition behind Lemma 3.5 is as follows. Theorem 3.2 shows that we can represent the density as an integral with respect to the random measure generated by Θ_T with probability $1 - 2\delta$. In other words, there exists a subset Θ_T space with $\Pr(\text{that subset}) = (1 - 2\delta)$ such that the representation above holds. Since each realization Θ'_T in Θ'_T-space is a consistent sequence of categorical random variables, we can extend the probability space for these realizations by using a Dirichlet process. Intuitively, we are placing a Dirichlet prior on these categorical random variables.

To use the same notation we used to construct Q_T, we can view G_t^Q as a draw from G^Q and assume that both processes are Dirichlet, i.e., we are using a hierarchical Dirichlet process. Again by using the normalized completely random measure property of Dirichlet processes, this implies that the implied prior for the transition densities is Dirichlet.

3.5 Bayesian Nonparametrics and Convergence Rates

Problem Setup

We now use the sieve and associated bounds constructed in the previous section to derive the convergence rates of the associated estimators. We adopt a standard Bayesian nonparametric framework and show how fast the posteriors contract to the true model.

We start by recalling this setup. We assume the data $\{x_t\}_{t=1}^T$ are drawn from some distribution P_T which is parameterized $P_T(\cdot \,|\, \xi)$, for $\xi \in \Xi$. This parameter set is equipped with the Borel σ-algebra $\mathscr{B}$ with associated prior distribution $\mathcal{Q}_0(\xi)$. We further assume that there exists a regular conditional distribution of X_T given ξ — $P(X_T \,|\, \xi)$ — on the sample space $(\mathcal{X}, \mathscr{X})$. This implicitly defines a joint distribution over $(\mathcal{X} \times \Xi, \mathscr{X} \times \mathscr{B})$:

$$\Pr(X_T \in A, \xi \in B) = \int_B \Pr(A \,|\, \xi)\, d\mathcal{Q}_0(\xi). \tag{3.11}$$

Under some technical conditions, we can define a regular version of the conditional distribution of ξ given X_T, i.e., a Markov kernel from $(\mathcal{X}, \mathscr{X})$ into $(\Xi, \mathscr{B})$, which is

called the posterior.

Definition 3.7. Posterior Distribution

$$\mathcal{Q}_T\left(B \mid X_T\right) \coloneqq \Pr\left(\{\xi \in B\} \mid X_T\right), B \in \mathscr{B}. \tag{3.12}$$

Posterior contraction rates characterize the speed at which the posterior distribution become close to the true value in a distributional sense. They are useful for two reasons. First, it puts upper bound on the convergence rate of point estimators such as the mean. Second, it tells you the speed at which inference using the estimated posterior distribution becomes valid. Our definition of this rate comes from Ghosal and van der Vaart (2017, Theorem 8.2).

Definition 3.8. Contraction Rate A sequence ϵ_T is a *posterior contraction rate* at parameter ξ^P with respect to the semimetric d if $\mathcal{Q}_T\left(\{\xi \mid d(\xi^P, \xi) \geq M_T \epsilon_T\} \mid X_T\right) \to 0$ in $P_T\left(X_T \mid \xi^P\right)$-probability for every $M_T \to \infty$.

To bound the asymptotic behavior of ϵ_T, we must simultaneously bound two separate quantities. First, we must show that our approximating model is close to the true density in the appropriate distance. We did this in the previous section. Second, we must bound the complexity (entropy) of our model, showing that it does not grow too rapidly.

We start by defining some notation that we use in deriving our theorems for the contraction rates. The concepts we use here are standard in the Bayesian nonparametrics literature. First, we define the metric (Kolmogorov) entropy for some small distance ϵ, some set Ξ, and some semimetrics, d_T and e_T. (One can, of course, use the same semimetric for both d_T and e_T.)

Definition 3.9. Metric Entropy $N(C\epsilon, d_T(\xi, \xi^P), e_T)$ is the function whose value for $\epsilon > 0$ is the minimum number of balls of radius $C\epsilon$ with respect to the d_T semimetric (i.e., d_T-balls of radius $C\epsilon$) needed to cover an e_T-ball of radius ϵ around the true parameter ξ^P.

The logarithm of this number — the *Le Cam Dimension* — is the relevant measure of the model's complexity, and hence the "size" of the sieve, and controls the minimax rate under some technical conditions. We define a ball with respect to the minimum

of the Kullback-Leibler divergence and some related divergence measures. We also adopt the following two concepts used in Ghosal and van der Vaart (2007).

First, $V_{k,0}$ is "essentially" the k$^{\text{th}}$-centered moment of the Kullback-Leibler divergence between two densities f, g, and associated distributions F, G:

$$V_{k,0}(f,g) \coloneqq \int \left|\log(f/g) - \mathrm{D}_{\mathrm{KL}}\left(f \,||\, g\right)\right|^k dF. \tag{3.13}$$

Having defined $V_{k,0}(f,g)$, we define the relevant balls. $f_T(X \,|\, \xi)$ is the density of the length T data sequence X_T associated with parameter ξ. The ball is defined thus:

$$B_T(\xi^P, \epsilon, k) \coloneqq \left\{ \xi \in \Xi \,\middle|\, \begin{array}{l} \mathrm{D}_{\mathrm{KL}}\left(f\left(X_T \,|\, \xi^P\right) \,||\, f\left(X_T \,|\, \xi\right)\right) \leq T\epsilon^2, \\ V_{k,0}\left(f\left(X_T \,|\, \xi^P\right), f\left(X_T \,|\, \xi^P\right)\right) \leq T\epsilon^2 \end{array} \right\}. \tag{3.14}$$

We now quote Ghosal and van der Vaart (2007, Theoerem 1). This theorem provides general conditions for convergence of posterior distributions even if the data are not i.i.d.. It extends the results in Ghosal, Ghosh, and van der Vaart (2000), which is the most common way to derive convergence rates in the literature, to cover dependent data.

Theorem 3.6 (Ghosal and van der Vaart (2007) Theorem 1)**.** *Let d_T and e_T be semimetrics on Ξ. Let $\epsilon_T > 0, \epsilon_T \to 0$, $\left(\frac{1}{T\epsilon^2}\right)^{-1} \in O(1)$. $C_1 > 1$, $\Xi_T \in \Xi$ be such that for sufficient large $n \in \mathbb{N}$.*

1. *There exist exponentially consistent tests Υ_T as in Lemma 3.7 with respect to d_T.*

2. $$\sup_{\epsilon_T > \epsilon} \log N\left(\frac{C_2}{2}\epsilon, \left\{\xi \in \Xi_T \,\middle|\, d_T(\xi, \xi^P) \leq \epsilon\right\}, e_T\right) \leq T\epsilon_T^2 \tag{3.15}$$

3. $$\frac{\mathcal{Q}_T\left(\left\{\xi \in \Xi_T \,\middle|\, n\epsilon_T < d_T(\xi, \xi^P) \leq 2n\epsilon_T\right\} \,\middle|\, X\right)}{\mathcal{Q}_T\left(B_T(\xi^P, \epsilon_T, C_1) \,|\, X\right)} \leq \exp\left(\frac{C_2 T\epsilon_T^2 n^2}{2}\right) \tag{3.16}$$

Then for every $M_T \to \infty$, we have that

$$P_T\left(\mathcal{Q}_T\left(\left\{\xi \in \Xi_T \,\middle|\, d_T(\xi, \xi^P) \geq M_T\epsilon_T\right\} \,\middle|\, X\right) \,\middle|\, \xi^P\right) \to 0. \tag{3.17}$$

Contraction Rates

We now show that uniformly consistent tests exist with respect to the semimetric that we use: h_∞. This metric is stronger than the average squared Hellinger distance, which is usually used in the Bayesian nonparametric estimation of Markov transition densities, (Ghosal and van der Vaart 2017, 542).

Note, h_∞^2 should be interpreted as a distance on the joint distributions because we can always factor a joint distribution as

$$f(X_T) = f\left(x_T \,|\, \mathcal{F}_{T-1}\right) \cdot f\left(x_{T-1} \,|\, \mathcal{F}_{T-2}\right) \cdots f\left(x_2 \,|\, \mathcal{F}_1\right) \cdot f\left(x_1 \,|\, \mathcal{F}_0\right), \tag{3.18}$$

where $\mathcal{F}_0$ denotes information that is always known, as is standard.

It is worth noting that h_∞^2 is a function of T even though we suppress it in the notation. We are only considering deviations between the densities over length-T sequences. The first goal is to show that consistent tests exit to separate two distributions in h_∞^2. To do this, we provide the following lemma.

Lemma 3.7 (Exponentially consistent tests exist with respect to h_∞)**.** *There exist tests Υ_T and universal constants $C_2 > 0$, $C_3 > 0$ satisfying for every $\epsilon > 0$ and each $\xi_1 \in \Xi$ and true parameter ξ^P with $h_\infty(\xi_1, \xi^P)$:*

1.
$$P_T\left(\Upsilon_T \,\middle|\, \xi^P\right) \leq \exp(-C_2 T \epsilon^2) \tag{3.19}$$

2.
$$\sup_{\xi \in \Xi,\, e_n(\xi_1, \xi) < \epsilon C_3} P_T\left(1 - \Upsilon_T \,\middle|\, \xi^P\right) \leq \exp(-C_2 T \epsilon^2) \tag{3.20}$$

Having shown there exist the appropriate tests, we now show that (3.15) and (3.16) hold. As noted in Ghosal and van der Vaart (2007, 197), the numerator is trivially bounded by 1, as long as $T\epsilon_T \to \infty$ which it does in this case. We do this by proving a proposition that covers both the marginal and transition dentin cases. We can deduce the mains theorems as results of it.

Proposition 3.8 (Bounding the Posterior Divergence)**.** *Let P_T be a uniformly ergodic Hidden Markov Gaussian process, i.e., $p_T \coloneqq \sum_k \Pi_{k,t} \phi(x_t \,|\, \mu_t, \Sigma_t)$ with finite means and finite positive-definite covariances. Let $\Xi_T \subset \Xi$ and $T \to \infty$. Let Q_T be a mixture approximation with $\frac{K_T^i}{\eta_T}$ components. Assume the following condition holds*

with probability $1 - 2\delta$ *for* $\delta > 0$ *and constants* C *and* $i \in \mathbb{N}$*:*

$$\sup_t h\left(q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^Q\right), p_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^P\right)\right) < C\eta_T. \tag{3.21}$$

Let $\epsilon_{i,T} := \frac{\log(T)^{\sqrt{i}}}{\sqrt{T}}$*. Then the following two conditions hold with probability* $1 - 2\delta$ *with respect to the prior*

$$\sup_{\epsilon_i \geq \epsilon_{T,i}} \log N\left((\epsilon_i, \left\{\xi \in \Xi_T \,\middle|\, h_\infty(\xi, \xi^P) \leq \epsilon_i\right\}, h_\infty\right) \leq T\epsilon_{T,i}^2, \tag{3.22}$$

and

$$\mathcal{Q}_T\left(B_T\left(\xi^P, \epsilon_{T,i}, 2\right) \,\middle|\, X_T\right) \geq C \exp\left(-C_0 T \epsilon_{T,i}^2\right). \tag{3.23}$$

We can apply Proposition 3.8 to the transition density by taking $i = 2$. We use the representation for the transition density we proved in Theorem 3.4. As a consequence, by Theorem 3.6, the following result holds.

Theorem 3.9 (Contraction Rate of the Transition Density)**.** *Let* P_T *be a uniformly ergodic Hidden Markov Gaussian process, i.e.,* $p_T := \sum_k \Pi_{t,k} \phi(x_t \,|\, \mu_t, \Sigma_t)$ *with finite means and finite positive-definite covariances. Let* $T \to \infty$*, then the following holds with* $\epsilon_T := \sqrt{\frac{\log(T)^2}{T}}$ *with probability* $1 - 2\delta$ *with respect to the prior. There exists a constant* C *independent of* T *such that the posterior over the transition densities constructed above and the true transition density satisfies*

$$P_T\left(\mathcal{Q}_T\left(\sup_{\mathcal{F}_{t-1}^P, \mathcal{F}_{t-1}^Q} h\left(p_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^P\right), q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^Q\right)\right) \geq C\epsilon_T \,\middle|\, X_T\right)\right) \to 0.$$

We also bound for the convergence rate of the marginal density. This should not be too surprising. Estimating the Markov transition density with respect to h_∞ is strictly harder than estimating the marginal distribution. You can integrate out the marginal distribution by using the stationary distribution. (In this context, the stationary and marginal distributions are the same.) Also, since i.i.d. data is trivially a uniformly ergodic Markov process, we cover the i.i.d. case as well.

Theorem 3.10 (Contraction Rate of the Marginal Density)**.** *Let* P_T *be a uniformly ergodic Hidden Markov Gaussian process, i.e.,* $p_T := \sum_k \pi_k \phi(\cdot \,|\, \mu_t, \Sigma_t)$ *with finite mean and finite variance. Let* $T \to \infty$*, then the following holds with* $\epsilon_T = \sqrt{\frac{\log(T)}{T}}$

and probability $1-2\delta$ *with respect to the prior. There exists a constant* C *independent of* T *such that the posterior over the transition densities constructed above and the true transition density satisfies*

$$P_T\left(\mathcal{Q}_T\left(h\left(p_T\left(x_t\right), q_T\left(x_t\right)\right) \geq C\epsilon_T \mid X\right)\right) \to 0.$$

3.6 Estimation Strategy

We estimate our model using Bayesian methods. So far, the discussion has been rather abstract. We have focused on proving theoretical results about our estimation strategy. We now construct a Gibbs sampler to estimate the model. Recall the definition of the approximating model for the transition density:

$$q_T\left(x_t \mid \mathcal{F}_{t-1}\right) = \sum_{k=1}^{K_T} \Pi\left(k = \delta_t \mid \delta_{t-1}\right) \phi\left(\beta_k x_{t-1}, \Sigma_k\right). \tag{3.24}$$

We must place a prior on each of the components in this model. We start by using a Dirichlet process to place a prior on $\Pi_{t,k} \coloneqq \Pi(\delta_t = k \mid \delta_{t-1})$ and, hence, implicitly on K_T. We then construct priors for β_k and Σ_k.

A substantial literature exists on efficiently estimating Dirichlet mixture models, e.g., Ishwaran and James (2001), Papaspiliopoulos and Roberts (2008), and Griffin and Walker (2011). We use Walker's (2007) slice sampler to handle the potentially infinite number of clusters without truncation and compute a valid upper bound for K_T. Conditional on K_T we draw the δ_t's marginal distribution. This is straightforward because (3.24) is almost a standard Gaussian mixture model.[48] We then update the transition matrix Π so it has the correct marginal distributions, and draw the $\{\delta_t\}_{t=1}^T$. Given $\delta_t = k$ and the hyperparameters, we apply standard Bayesian regression methods to obtain β_k and Σ_k. We use a conditionally conjugate hierarchical prior and draw from the hyperparameters' posterior. We present the procedure in Algorithm 3.1.

[48] Conditional on δ_{t-1} it is.

Algorithm 3.1 Gibbs Sampler

1. **Posterior of** $\{\delta_t\}_{t=1}^T$
   a) Use Walker (2007) to determine the number of clusters K_T.
   b) Draw the new marginal probabilities, π, and update the transition matrix, Π.
   c) Given K_T and $\{x_t\}_{t=1}^T$, use multinomial sampling to draw δ_t with
   $$\Pr(\delta_t = k) \propto \phi\left(x_t \,|\, \beta_k, \Sigma_k\right) \Pi_{t,k}.$$
2. **Posterior of** π
   a) Estimate the posterior of Π conditional on $\{\delta_t\}_{t=1}^T$:
   $$\Pi_{k,j} = \frac{\mathcal{Q}_0(\delta_{t-1} = k)\mathcal{Q}_0(\delta_t = j) + \sum_{t=2}^T \mathbf{1}(\delta_{t-1} = k)\mathbf{1}(\delta_t = j)}{\mathcal{Q}_0(\delta_{t-1} = k) + \sum_{t=2}^T \mathbf{1}(\delta_{t-1} = k)}.$$ [49]
3. **Posterior of Component-Specific Parameters**
   a) Given each cluster k, use Bayesian regression to draw $\{\beta_k, \Sigma_k\}$.
4. **Posterior of Hyperparameters**
   a) Draw the hyperparameters governing $\{\beta_k, \Sigma_k\}$ from their conjugate posteriors.
5. **Iterate**

Posterior of $\{\delta_t\}_{t=1}^T$

Bounding K_T

In each period, the approximating model and implied marginal density are Dirichlet mixtures. Consequently, we draw the cluster identities by adapting algorithms from the literature. We are in the standard situation, except our prior varies over time.

Sampling Dirichlet mixtures is difficult for two reasons. First, the prior allows for infinitely many clusters, and so we cannot sum the probabilities and cannot compute the resulting marginal cluster probabilities. This inability arises because we cannot

numerically solve the probability of cluster k: $\Pr(k) = 1 - \sum_{k^* \neq k} \Pr(k^*)$. All Dirichlet mixture models share this property and so several authors have developed ingenious ways to deal with this issue. We adopt the algorithm developed by Walker (2007) because this algorithm is exact (we do not need to truncate the distribution) and computationally efficient. He does this by introducing a random variable — u_t — so that, conditional on u_t, the distributions are available in closed form.

Given the cluster parameters, we can write the distribution of x_t as

$$q_T(x_t) = \sum_{k=1}^{\infty} \Pi_{t,k} \phi\left(x_t \,|\, \beta_k, \Sigma_k\right). \tag{3.25}$$

As mentioned above, we introduce a latent variable $u_t \sim U(0, \Pi_{t,k})$ so we can rewrite (3.25) as

$$q_T(x_t) = \sum_{k=1}^{\infty} \mathbf{1}\left(u_t < \Pi_{t,k}\right) \phi\left(x_t \,|\, \beta_k, \Sigma_k\right) = \sum_{k=1}^{\infty} \Pi_{t,k} U\left(u_t \,|\, 0, \Pi_{t,k}\right) \phi\left(x_t \,|\, \beta_k, \Sigma_k\right). \tag{3.26}$$

Consequently, with probability $\Pi_{t,k}$, x_t and u_t are independent, and so the marginal density for u_t is

$$\Pr\left(u_t \,\middle|\, \{\Pi_{t,k}\}_{k=1}^{K}\right) = \sum_{k=1}^{\infty} \Pi_{t,k} U\left(u_t \,|\, 0, \Pi_{t,k}\right) = \sum_{k=1}^{\infty} \mathbf{1}\left(u_t < \Pi_{t,k}\right). \tag{3.27}$$

Then we can condition on $\{u_t\}_{t=1}^{T}$ as a vector, but not on $\Pi_{t,k}$.

$$\Pr\left(\{v_k\}_{k=1}^{K} \,\middle|\, \{\delta_t\}_{t=1}^{T}\right) = \mathcal{Q}_0\left(\{v_k\}_{k=1}^{K}\right) \prod_{t=1}^{T} \mathbf{1}\left(v_{k=\delta_t} \prod_{\kappa<\delta_t} (1 - v_\kappa) > u_{k=\delta_t}\right), \tag{3.28}$$

where the v_k are the sticks in the stick-breaking representation of the prior.

The dependence between the u_t does not affect (3.28) because the v_k do not depend upon t. Hence, the v_k are conditionally independent given $\{u_t\}_{t=1}^{T}$. Exploiting this independence and the stick-breaking representation of the prior, we can draw v_k from (3.28); it only shows up once in the product. By adopting the prior for the sticks implied by standard Dirichlet process — $\text{Beta}(1, \alpha)$, we use (3.28) to draw v_k. As shown by Papaspiliopoulos and Roberts (2008), this implies v_k are distributed:

$$v_k \sim \text{Beta}\left(1 + \sum_{t=1}^{T} \mathbf{1}(\delta_t = k), T - \sum_{\kappa=1}^{k} \sum_{t=1}^{T} \mathbf{1}(\delta_t = \kappa) + \alpha\right) \tag{3.29}$$

for $k = 0, 1, \ldots$. We only need to do this for the v_k where that $k \leq \max(\delta_t)$. These sticks are the only sticks at affect the likelihood. We can calculate the marginal cluster probabilities π_k:

$$\pi_k = v_k \prod_{\kappa=1}^{k} (1 - v_\kappa). \tag{3.30}$$

Correcting Π to have the Correct Marginal Distribution

If the data were i.i.d., we could convert the v_k into π_k, and then compute the set of possible δ_t. This step is precisely the references above use. However, the data are not i.i.d. because $\Pi_{t,k}$ depends on δ_{t-1}. The question at hand is how to transform the algorithm to update the marginal distribution in the presence of i.i.d. data into one that does not change the dependence structure in non-i.i.d. data.

We must construct a probability matrix such that the relationship between two clusters, k and k^*, remain the same as they did in the previous draw of the sampler, but the marginal distribution is updated appropriately. We know that Markov transition matrices and their associated marginal distributions have the following relationship for each cluster k:[50]

$$\pi_k = \sum_{j=1}^{\infty} \Pi_{k,j} \pi_j. \tag{3.31}$$

Let $\widetilde{\pi}$ be a new marginal distribution that is equivalent (in the measure-theoretic sense) to π. Define a transition matrix $\widetilde{\Pi}$ whose elements satisfy $\widetilde{\Pi}_{j,k} = \Pi_{j,k} \frac{\widetilde{\pi}_k}{\pi_k} \frac{\pi_j}{\widetilde{\pi}_j}$. We now show that $\widetilde{\pi}$ is the marginal distribution associated with $\widetilde{\Pi}$ by showing it satisfies (3.31).[51]

$$\widetilde{\pi}_k = \pi_k \frac{\widetilde{\pi}_k}{\pi_k} = \sum_{j=1}^{\infty} \Pi_{j,k} \pi_j \frac{\widetilde{\pi}_k}{\pi_k} = \sum_{j=1}^{\infty} \Pi_{j,k} \frac{\widetilde{\pi}_k}{\pi_k} \frac{\pi_j}{\widetilde{\pi}_j} \widetilde{\pi}_j = \widetilde{\pi}_k \sum_{j=1}^{\infty} \widetilde{\Pi}_{k,j} \widetilde{\pi}_j \tag{3.32}$$

We constructed a matrix $\widetilde{\Pi}$ that induces the correct marginal distributions. In doing this, we only changed the marginal distribution. The relative probabilities between

[50] This condition holding for all k is the standard condition that a stationary distribution is a left-eigenvector of the transition matrix.

[51] The multiplication and division in (3.32) is the scalar version.

different states has not been affected, i.e., for all states j, k, the relative probabilities $\Pi(k \mid \delta_t)/\Pi\pi(k \mid \delta_t) = \pi_k/\widetilde{\pi}_j$.

To run a Gibbs sampler, we view the operation in (3.32) as a draw from a conditional posterior. We condition on all but the first left eigenvector (the one associated with the eigenvector 1) of the transition matrix, Π and replace it with the one associated with $\widetilde{\Pi}$. We then calculate the resulting transition matrix. Transition matrices associated with irreducible Markov chains have exactly one stationary distribution, and that stationary distribution is the first left eigenvector. So this algorithm computes the unique new transition matrix.

Conditionally Drawing the $\{\delta_t\}_{t=1}^T$

If the new stationary distribution, $\widetilde{\pi}$, has more clusters than the previous one, π, did, we use the prior for Π to draw them. We do not have to transform them to have the appropriate dynamics because they contain no datapoints under Π, implying that π and $\tilde{\pi}$ coincide as they have the same prior.

From $\widetilde{\Pi}$ can compute $\Pi_{t,k}$ for each t by drawing the first cluster identity, δ_0 from the stationary distribution, and then using using the Markov property of δ_{t-1} for $t > 1$, and iterating forward. We can now compute $\{k \mid \Pi_{t,k} > u_t\}$ for each t. Then the posterior of δ_t is

$$\Pr\left(\delta_t = k \mid \Pi_{t,k}, u_t, x_t, \beta_k, \Sigma_k\right) \propto \mathbf{1}\left(k \in \{k \mid \Pi_{t,k} > u_t\}\right) \phi\left(x_t \mid \beta_k x_{t-1}, \Sigma_k\right). \tag{3.33}$$

This is a finite set with known probabilities, and the δ_t are categorical variables. These can sampled directly.

Posterior on the Transition Matrix

We place the Dirichlet process prior over these cluster identities in each period to allow for an arbitrary number of clusters. By stacking the Dirichlet processes over time, we obtain a Dirichlet process over the (δ_{t-1}, δ_t) product space. Intuitively, we are constructing the transition matrix, Π, as a Dirichlet-distributed infinite-dimensional square matrix as noted by Lin, Grimson, and Fisher (2010).

Given the cluster identities, δ_t, which we drew in Section 3.6, we draw the transition matrices. We do this by noting that the prior probability of a transition is the

product of the unconditional probabilities normalized appropriately. We can update this by counting the proportion of realized transitions:

$$\Pi_{k,j} = \frac{\mathcal{Q}_0(\delta_{t-1} = k)\mathcal{Q}_0(\delta_t = j) + \sum_{t=2}^{T} \mathbf{1}(\delta_{t-1} = k)\mathbf{1}(\delta_t = j)}{\mathcal{Q}_0(\delta_{t-1} = k) + \sum_{t=2}^{T} \mathbf{1}(\delta_{t-1} = k)}.$$

Each element, $\Pi_{k,j}$, determines the probability of transitions in (δ_{t-1}, δ_t) and is updated by counting the number of transitions from k to j.

Identification Strategy and Cluster Labeling Problem

The other problem endemic to mixture models is that the cluster identities are not uniquely identified. In particular, we have a label switching problem. A model with clusters labeled 0 and 1 is the same model as one with those clusters labeled 1 and 0. This lack of uniqueness is particularly problematic in i.i.d. environments because there is no natural way to order the clusters.

In time series environments, like the one we consider here, we can label the clusters by when they first appear. The first period is always in cluster zero. The second cluster to arrive is always labeled cluster two. This labeling procedure has two nice features relative to the existing methods of labeling the clusters by their probability ordering. First, it imposes a strict order of the clusters. We have no ties, such as occur in probability-based labeling when two probabilities are equal. Second, the ordering is invariant to estimation uncertainty. We do not have to estimate which datapoint comes first in time, and so it is easy to maintain the same ordering across draws.

In order to enforce this identification restriction, we re-order the cluster identities immediately before returning the next posterior draw so that they always arrive in time order. This reordering does not solve the identification problem, but it does reduce the amount of multi-modality in our posterior.

Posterior for the Coefficient Parameters

The component-specific likelihood is given in Definition 3.10 where $X_k \coloneqq \{x_t \,|\, t \in \delta_k\}$, $Y_t \coloneqq \{x_t \,|\, t-1 \in \delta_k\}$, and T_k is the number of datapoints in cluster k. We are factoring the likelihood into the component specific components.

Definition 3.10. Component-Specific Likelihood

$$\{x_t\}_{t=1}^T \mid \{\delta_t\}_{t=1}^T, \{\beta_k, \Sigma_k\}_{k=1}^K \sim \prod_{k=1}^K \frac{|\Sigma_k|^{-T_k/2}}{(2\pi)^{T_k/2}} \exp\left(-\frac{1}{2}\operatorname{tr}\{(Y_k - X_k\beta_k)\,\Sigma_k^{-1}\,(Y_k - X_k\beta_k)'\}\right),$$

We can factor the likelihood into component-specific parts, and estimate the parameters component by component. Because the components have varying amounts of data, we cannot assume that the number of datapoints in each of the components approaches infinity. Also, when we forecast, we sometimes must add more components. To do this effectively, we want to use all of the information the observed data gives us. We cannot condition on the data in the new component because there is none. Consequently, we specify a hierarchical model to pool information across components.

The first level is the standard Gaussian Inverse-Wishart prior that is conjugate to the prior specified in Definition 3.10.[52] The only difference is that we parameterize the innovation covariance distribution in terms of its mean: Ω.[53] If we need to add a new component during the course of the algorithm we draw from the distribution of β_k, Σ_k conditional on the $\bar{\beta}, U, \Omega, \mu_1$. We cannot condition on the data in the new component because none exists.

Definition 3.11. Component-Specific Parameters' Prior

$$\{\beta_k\}_{k=1}^K \mid \Sigma_k, \bar{\beta}, U \sim \mathcal{MN}\left(\bar{\beta}, \Sigma_k, U\right) \tag{3.34}$$

$$\{\Sigma_k\}_{k=1}^K \mid \Omega \sim \mathcal{W}^{-1}\left((\mu_1 - 2)\Omega, \mu_1 + D - 1\right) \tag{3.35}$$

This prior is the conjugate prior for the likelihood in Definition 3.10, and so we can use the standard formulas to estimate it. This gives the following marginal posterior

[52] Throughout, we use the parametric formulas given in the Wikipedia pages for the distribution. For example, the Matrix-Normal distribution is parameterized as it is at https://en.wikipedia.org/wiki/Matrix_normal_distribution.

[53] The scale parameter and the degrees of freedom parameter are chosen in the appropriate way to make Ω the mean matrix: $\mathbb{E}[\Sigma_k] = \text{Scale}/(\text{Degrees of freedom} - D - 1) = (\mu_1 - 2)\Omega/(\mu_1 + D - 1 - D - 1) = \Omega$.

for the Σ_k:

$$\Sigma_k \,|\, X_k, Y_k \sim \mathcal{W}^{-1}\Bigg(\bar{\beta}' U^{-1}\bar{\beta} + (\mu_1 - 2)\Omega + Y_k' Y_k \\ - \left(U^{-1}\bar{\beta} + X_k' Y_k\right)' \left(U^{-1} + X_k' X_k\right)^{-1} \left(U^{-1}\bar{\beta} + X_k' Y_k\right), \mu_1 + D - 1 + T_k\Bigg). \tag{3.36}$$

We can also compute the following conditional posterior for β_k given Σ_k:

$$\bar{\beta}, \Sigma_k \,|\, X_k, Y_k \sim \mathcal{MN}\left(\left(U^{-1} + X_k' X_k\right)^{-1}\left(U^{-1}\bar{\beta} + X_k' Y_k\right), \Sigma_k, \left(U^{-1} + X_k' X_k\right)^{-1}\right) \tag{3.37}$$

We now specify the prior and posterior for the hyperparameters. As is common in the literature, we draw $\bar{\beta}$ and U from their posteriors to control level of smoothing in a data-dependent way by placing prior distributions on the hyperparameters and estimating them. As we did above, we place a conjugate matrix-normal prior on the coefficient matrix and an Inverse-Wishart prior on the covariance matrix.

Definition 3.12. Coefficient Hyperparameters' Prior

$$\bar{\beta}, U \sim \mathcal{MN}(\beta^\dagger, \mathbb{I}_D, U)\mathcal{W}^{-1}(\Psi_U, \nu_U)$$

The product of the priors for β_k's given in (3.34) now behaves as the likelihood. Since we have Gaussian priors and likelihoods this a fairly standard posterior calculation. The only complication is that the $\{\beta_k\}_{k=1}^K$ are heteroskedastic.[54] Consequently, we provide the derivation in Section 3.E:

$$U \,|\, \{\Sigma_k, \beta_k\}_{k=1}^K \sim \mathcal{W}^{-1}\Bigg(\beta^\dagger \beta^{\dagger\prime} + \Psi_U + \sum_{k=1}^K \beta_k \Sigma_k^{-1} \beta_k' - \left(\beta^\dagger + \sum_{k=1}^K \beta_k \Sigma_k^{-1}\right)\left(\sum_{k=1}^K \Sigma_k^{-1} + \mathbb{I}_D\right)^{-1} \\ \left(\beta^\dagger + \sum_{k=1}^K \beta_k \Sigma_k^{-1}\right)', \nu_U + (K+1)D\Bigg), \tag{3.38}$$

and

$$\bar{\beta} \,|\, U, \{\Sigma_k, \beta_k\}_{k=1}^K \sim \mathcal{MN}\left(\left(\beta^\dagger + \sum_{k=1}^K \beta_k \Sigma_k^{-1}\right)\left(\sum_{k=1}^K \Sigma_k^{-1} + \mathbb{I}_D\right)^{-1}, \left(\sum_{k=1}^K \Sigma_k^{-1} + \mathbb{I}_D\right), U\right). \tag{3.39}$$

[54] They must be in order for the prior in (3.34) to be conjugate with its likelihood because the likelihood is heteroskedastic itself.

To draw Ω from its posterior, we adapt the hierarchical prior Huang and Wand (2013) construct. We deviate from them to allow Ω to have off-diagonal elements. Our covariance matrices are i.i.d. in expectation, but the prior for a new covariance matrix is not necessarily i.i.d. Also, Huang and Wand's (2013) model does not necessarily have a density with respect to Lebesgue measure for the covariance matrix itself. We only allow for the hyperparameters to take on values where Σ_k's distribution has a both mean and density.

In particular, we parameterize the hierarchy for the Σ_k as follows. We have two degree of freedom parameters, μ_1 and μ_2, a mean matrix, $\Omega = \mathbb{E}[\Sigma_k]$, and D scale parameters for Ω: $a_1, \ldots, a_D$.

Definition 3.13 (Prior for the Covariances)**.**

$$\Omega \sim \mathcal{W}\left(\frac{\text{diag}(a_1, \ldots, a_D)}{\mu_2 + D - 1}, \mu_2 + D - 1\right)$$

If we send $\mu_2 \to \infty$, the implied prior for the prior for Ω becomes fully dogmatic. If $\nu_2 = 1/2$ and $D = 1$, the root diagonal elements — $\sqrt{(\Sigma_k)_{dd}}$ — have half-t distributions. In general, the $(\Sigma_k)_{dd}$ have appropriately scaled F-distributions.[55] If the off-diagonal elements of Ω almost surely equal to 0, the diagonal elements satisfy $(\Sigma_k)_{dd} \sim \Gamma^{-1}\left(\mu_1/2, (\frac{\mu_1}{2} - 1)\Omega_{dd}\right)$. This is why we let the number of degrees of freedom in (3.35) depend upon D. In general, the mean of these elements is the same, but the distribution is different since the off-diagonal elements of Ω affect the distribution of $(\Sigma_k)_{dd}$.

Obviously, conditional on Ω, everything is independent. The posterior distribution of Σ_k given $\Omega, \{x_t \,|\, \delta_t = k\}$ is

$$\Omega \,|\, \{\Sigma_k\}_{k=1}^K \sim \mathcal{W}\Bigg(K(\mu_1 + D - 1) + (\mu_2 + D - 1), \left(\text{diag}(a_1, \ldots a_D)^{-1} + (\mu_1 - 2)\sum_{k=1}^K \Sigma_k^{-1}\right)^{-1}\Bigg). \tag{3.40}$$

[55] $\sigma^2 \sim F(1, \mu_1 + D - 1) \implies \sigma \sim \text{half-}t(\mu_1 + D - 1)$. In the one dimensional case, $\mu_1 + D - 1_2 = 1/2$ implies that $\sigma^2 \sim F(1, \mu_1 + D - 1)$. This result is not feasible in the multivariate case while maintaining a density with respect to Lebesgue measure. If we let $\mu_1 \to 2$, we recover this expression. However, Ω is not well-defined in this case.

As noted by Huang and Wand (2013), if Ω is almost surely diagonal, then the correlation parameters in Σ_k have a prior density of the form $p(\rho_{ij}) \propto (1-\rho^{ij})^{\mu_1/2-1}$, $-1 < \rho_{ij} < 1$. Note, this implies that as $\mu_1 \to 2$, then the distribution of these off-diagonal elements approaches $U(-1,1)$. Conversely, as $\mu_1 \to \infty$, the distribution of these off-diagonal elements converges to point masses at the off-diagonal elements of Ω. The off-diagonal elements of Ω are normal variance-mean mixtures where the mixing density is a χ^2-distribution, as is standard for Wishart priors.

3.7 Data and Prior

We downloaded monthly data on real consumption (DPCERAM1M225NBEA), the personal consumption expenditure price index (PCEPI), industrial production (INDPRO), housing supply (MSACSR), the M2 measure of money supply (M2), unemployment rate (UNRATE), and 10-year Government bond yields (IRLTLT01USM156N) from the Federal Reserve Bank of Saint Louis economic database, (FRED). We chose these data series because they are several of the fundamental economic series underlying the macroeconomy, and they span much of the interesting variation.

All of the data are seasonally-adjusted by FRED. We convert to approximate percent changes by log-differencing all of the data except for the consumption measure, which is already measured in percent changes, the unemloyment rate and the long-term interest rate. We then demean the data and rescale them so they have standard deviations equal to 1. This is useful because it puts all of the data on the same scale.

The data covers the January 1963 to December 2018. The time dimension is 671, and the cross-sectional dimension is 7. Figure 3.2 shows the standardized monthly macroeconomic data used in this subsection. The gray bars are the NBER recessions.

We use the same prior for both datasets and for the simulation, as in Table 3.1 to make our results more easily interpretable. The prior we use for the component coefficients has a Kronecker structure, and so we specify prior beliefs over the relationship between regressands and regressors separately. In particular, the parameters are a priori independent across different regressands.

The prior we use for the component parameters and base Dirichlet measure is rather flat. We are not imposing a great deal of a priori structure. In addition, the

Figure 3.2: Monthly Macroeconomic Series

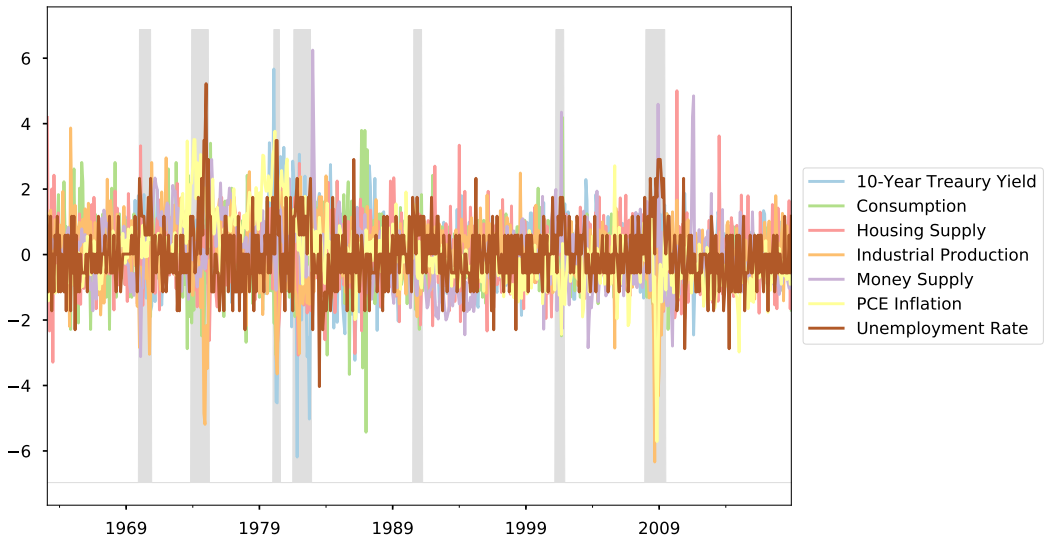


Table 3.1: Prior

Degrees of freedom for the hierarchical prior	5
Expected Number of Components	5
Component Coefficients	
Intercept	0
Expected Diagonal Autocorrelation	0.9
Expected Off-Diagonal Autocorrelation	0
Component Covariances	
Mean	$.25^2\mathbb{I}_D$
μ_1	3
μ_2	3

theory tells us it will not matter asymptotically.

3.8 Empirical Results

Monthly Macroeconomic Series

Using the macroeconomic data, we obtain the posterior draws from our sampler which are summarized in Figure 3.3. In Figure 3.3a, we see that the conditional mean tracks the dynamics of data quite well. We can divide the conditional variance in each period into two components using the law of total volatility:

$$\mathbb{V}\mathrm{ar}\left(x_t \mid \mathcal{F}_{t-1}\right) = \mathbb{V}\mathrm{ar}\left(\mathbb{E}\left[x_t \mid \delta_t\right] \mid \mathcal{F}_{t-1}\right) + \mathbb{E}\left[\mathbb{V}\mathrm{ar}\left(x_t \mid \delta_t\right) \mid \mathcal{F}_{t-1}\right]. \tag{3.41}$$

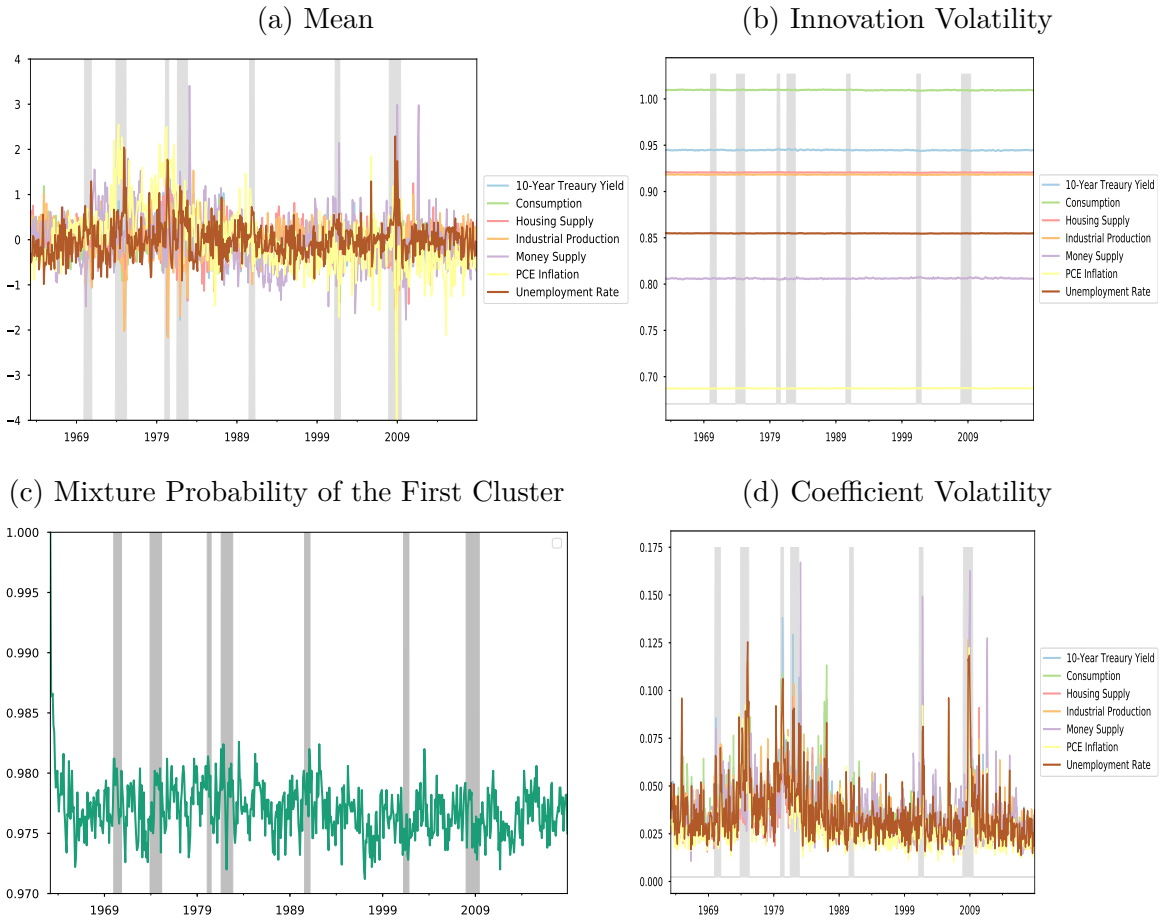
Since the model is linear conditional on the cluster identity δ_t, the first term comes from variation in $\beta_k x_t$, while the second arises from variation in the innovations. Figure 3.3d shows the volatility associated with autoregressive coefficients, whereas Figure 3.3b shows the volatility associated with innovations. The total volatility, which we graph for consumption in Figure 3.6a, is the sum of the two. Interestingly, most of the variation arises from the variation in the conditional means, not variation in the conditional variances.

Comparing these two volatilities, we observe bigger changes in dynamics for the coefficient volatility. This implies that the stochastic volatility in macroeconomic data studied in papers such as Fernández-Villaverde and Rubio-Ramírez (2010) and Fernández-Villaverde et al. (2015) can be more parsimoniously modeled using variation in the conditional mean than by using stochastic volatility.

Figure 3.3c shows the mixture probability of the first cluster in each period. From the empirical results, we see that 5 clusters become active in our sample but the mixture probability of the first cluster is very high. Hence, our model is very parsimonious, which we did not impose. We can also see that the mixture probabilities fluctuates at the monthly frequency.

Since our estimator uses a mixture representation, we can link it to regime-switching models by viewing the mixtures as regimes. Our model provides more flexibility than the standard regime-switching models, because it endogenously determines the number of regimes and multiple regimes with different probabilities can be active in each period. We are not interested in identifying the underlying regimes, (which is impossible) but only in approximating the true density. Unlike many regime-switching models, we do not have a "recession" regime and "normal-times" regimes.

Figure 3.3: Empirical Results with Monthly Macroeconomic Series



To show that our algorithm works reasonably well in practice, we display the conditional density forecast for consumption in Figure 3.4. Predictive Densities and PIT's for the other series are provided in Section 3.D. The qualitative performance of the estimator and its its relationship to the VAR estimator holds for all of the series considered. If the model works perfectly, the probability integral transform (PIT) should be independent and distributed $U[0, 1]$. As we can see, it is roughly independent and distributed approximately uniformly. This implies it is picking up the underlying data's time-varying volatility and skewness. It is worth noting that since the main objective of this paper is density estimation, not forecasting, we report in-sample fits.

The dynamics of the data in Figure 3.4a are not obviously non-Gaussian or non-

Figure 3.4: 1-Period Ahead Conditional Forecasts: Consumption Expenditure

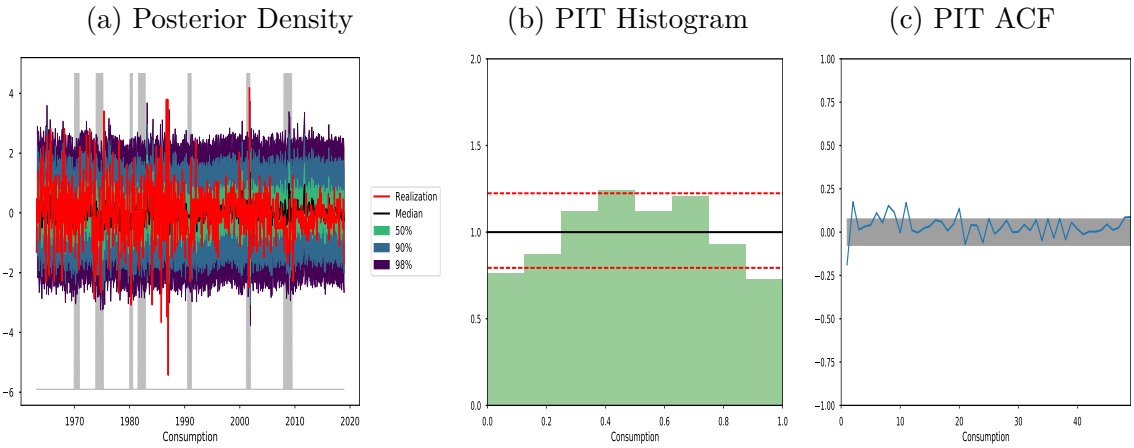
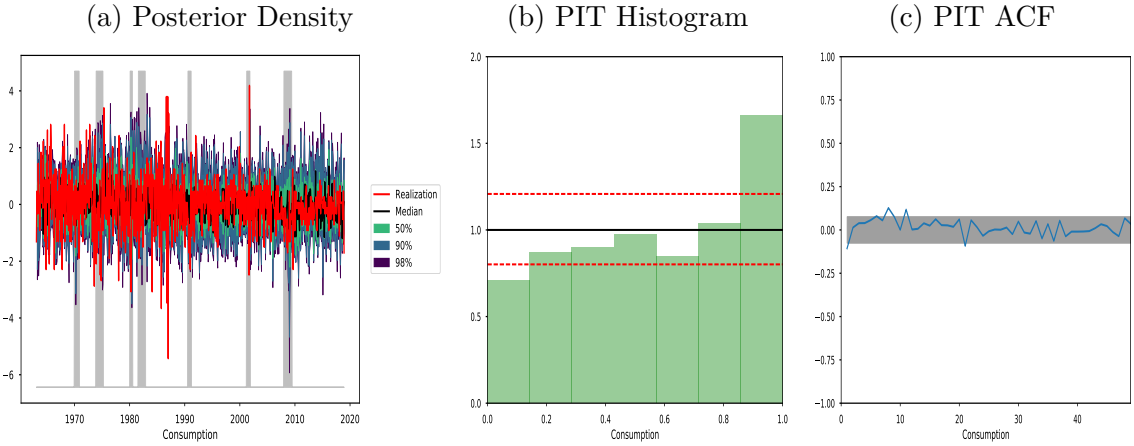


Figure 3.5: 1-Period Ahead Conditional Forecasts: Consumption Expenditure (VAR(1))

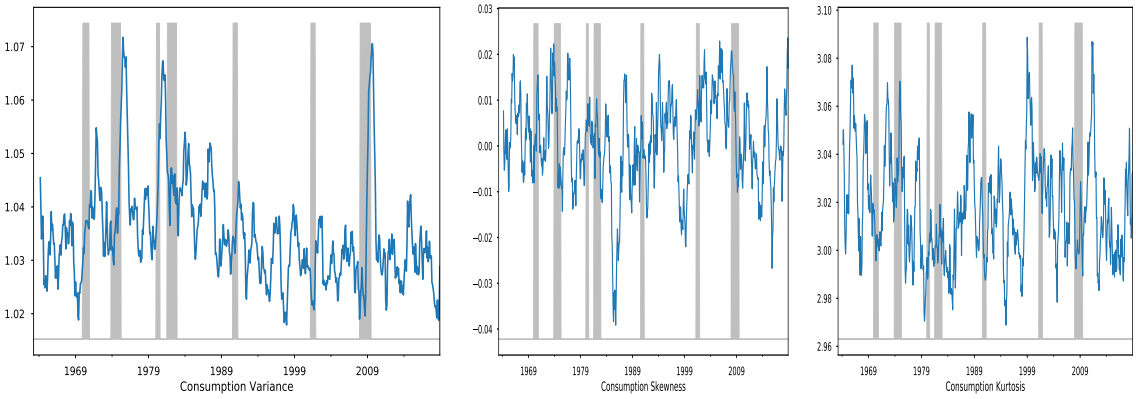


linear. Are we effectively just estimating a VAR? The answer is no. First of all, we can compare the 1-period ahead conditional forecasts from our model and those from Bayesian VAR(1). We can clearly see in Figure 3.5b that the PIT for the VAR is far from being uniformly distributed. For example, the underlying true DGP is skewed, but the VAR is not. In addition, by examining the conditional variance, (Figure 3.6a), we see that it spikes a great deal in recessions when we compute the rolling averages over 1 year. In other words, we can see stochastic volatility in consumption data that varies with the business cycle which would not be captured by VAR(1). Interestingly, although skewness (Figure 3.6b) and kurtosis (Figure 3.6c) exhibit significant time-

variation, this time variation does not obviously co-move with the business cycle. For instance, skewness of consumption fluctuated more in magnitude in the 70s and 80s than it did in the period after 2000.

Figure 3.6: Consumption Variability

(a) Variance (b) Skewness (c) Kurtosis



This time-variation in volatility but not in higher-moments is interesting on a number of dimensions. For example, similar to Schorfheide, Song, and Yaron (2018), we find stochastic volatility for consumption growth at business cycle frequencies using purely macroeconomic data. Conversely, disaster models such as Barro and Jin (2011) and Tsai and Wachter (2016) predict that kurtosis should either always be high (not approximately 3) or increase substantially during disasters.

3.9 Conclusion

In this paper, we show how to practically estimate marginal and transition densities of multivariate processes. This is a classic question in econometrics because most economic datasets are multivariate and parametric approximations often perform poorly. Furthermore, even outside of economics, other data-based disciplines face the same issues. We develop a Dirichlet Gaussian mixture model to estimate a wide variety of processes quite rapidly. Our method scales to a more series than the literature has thus far been able to handle and performs reasonably well in practice.

We provide new theory that shows, under some general assumptions, the posterior distribution of our estimators converges more rapidly than the previous literature

has been able to achieve. In particular, we exploit the tail behavior of probability distributions in high dimensions to show that our estimator for the marginal densities converges at a $\sqrt{\log(T)/T}$ rate and our estimator for the transition densities converge at a $\log(T)/\sqrt{T}$ rate with high probability. They are noteworthy because they are the parametric rate up to a logarithmic term. These rates are remarkable because they do not depend on the number of series except through the constant term.

We show that this estimation strategy performs well in simulations and when applied to various macroeconomic and financial data. In the empirical applications, we show that macroeconomic and financial data's dynamics are often far from Gaussian and the dynamic structure moves across the business cycle. We further find that our proposed representation requires more than one mixture component, but only a few, to handle the data's dynamics well.

References

Barro, Robert J., and Tao Jin. 2011. "On the Size Distribution of Macroeconomic Disasters." *Econometrica* 79 (5): 1567–1589.

Birgé, Lucien. 2013. "Robust Tests for Model Selection." In *From Probability to Statistics and Back: High-Dimensional Models and Processes — A Festscrift in Honor of Jon A. Wellner,* edited by M. Banerjee, F. Bunea, J. Huang, M. Koltchinksii, and M. H. Maathius, 9:47–68. IMS Collections. Institute of Mathematical Statistics.

Boucheron, Stéphane, Gábor Lugosi, and Pascal Massart. 2013. *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford, UK: Oxford University Press.

de la Peña, Victor H. 1999. "A General Class of Exponential Inequalities for Martingales and Ratios." *The Annals of Probability* 27 (1): 537–564.

Fernández-Villaverde, Jesús, Pablo Guerrón-Quintana, Keith Kuester, and Juan Rubio-Ramírez. 2015. "Fiscal Volatility Shocks and Economic Activity." *American Economic Review* 105 (11): 3352–84.

Fernández-Villaverde, Jesús, and Juan Rubio-Ramírez. 2010. *Macroeconomics and Volatility: Data, Models, and Estimation.* Working Paper 16618. National Bureau of Economic Research, December.

Geweke, John, and Michael Keane. 2007. "Smoothly Mixing Regressions." 50th Anniversary Econometric Institute, *Journal of Econometrics* 138 (1): 252–290.

Ghosal, Subhashis, Jayanta K. Ghosh, and Aad W. van der Vaart. 2000. "Convergence Rates of Posterior Distributions." *The Annals of Statistics* 28 (2): 500–531.

Ghosal, Subhashis, and Aad W. van der Vaart. 2007. "Convergence Rates of Posterior Distributions for Non-i.i.d. Observations." *The Annals of Statitics* 35:192–223.

———. 2017. *Fundamentals of Nonparametric Bayesian Inference.* Vol. 44. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Griffin, Jim E., and Stephen G. Walker. 2011. "Posterior Simulation of Normalized Random Measure Mixtures." *Journal of Computational and Graphical Statistics* 20 (1): 241–259.

Huang, Alan, and Matt P. Wand. 2013. "Simple Marginally Noninformative Prior Distributions for Covariance Matrices." *Bayesian Analysis* 8, no. 2 (June): 439–452.

Ichimura, Hidehiko, and Petra E. Todd. 2007. "Implementing Nonparametric and Semiparametric Estimators." Chap. 74, edited by James J. Heckman and Edward E. Leamer, vol. 6, Part B, 5369–5468. Handbook of Econometrics. Elsevier.

Ishwaran, Hemant, and Lancelot F. James. 2001. "Gibbs Sampling Methods for Stick-Breaking Priors." *Journal of the American Statistical Association* 96 (453): 161–173.

Johnson, William B., and Joram Lindenstrauss. 1984. "Extensions of Lipschitz Maps into a Hilbert Space." *Contemporary Mathematics* 26:189–206.

Klartag, B., and S. Mendelson. 2005. "Empirical Processes and Random Projections." *Journal of Functional Analysis* 225 (1): 229–245.

Koop, Gary, Dimitris Korobilis, and Davide Pettenuzzo. 2019. "Bayesian Compressed Vector Autoregressions." Annals Issue in Honor of John Geweke "Complexity and Big Data in Economics and Finance: Recent Developments from a Bayesian Perspective", *Journal of Econometrics* 210 (1): 135–154.

Lin, Dashua, Eric Grimson, and John Fisher. 2010. "Construction of Dependent Dirichlet Processes Based on Poisson Processes." In *Advances in Neural Information Processing Systems,* edited by J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, 23:1396–1404. Curran Associates, Inc.

Nguyen, XuanLong. 2016. "Borrowing Strengh in Hierarchical Bayes: Posterior Concentration of the Dirichlet Base Measure." *Bernoulli* 22, no. 3 (August): 1535–1571.

Norets, Andriy. 2010. "Approximation of Conditional Densities by Smooth Mixtures of Regressions." *The Annals of Statistics* 38, no. 3 (June): 1733–1766.

Papaspiliopoulos, Omiros, and Gareth O. Roberts. 2008. "Retrospective Markov Chain Monte Carlo Methods for Dirichlet Process Hierarchical Models." *Biometrika* 95 (1): 169–186.

Pati, Debdeep, David B. Dunson, and Surya T. Tokdar. 2013. "Posterior Consistency in Conditional Distribution Estimation." *Journal of Multivariate Analysis* 116:456–472.

Schorfheide, Frank, Dongho Song, and Amir Yaron. 2018. "Identifying Long-Run Risks: A Bayesian Mixed-Frequency Approach." *Econometrica* 86 (2): 617–654.

Sethuraman, Jayaram. 1994. "A Constructive Definition of Dirichlet Priors." *Statistica Sinica* 4 (2): 639–650.

Shen, Xiaotong, and Larry Wasserman. 2001. "Rates of Convergence of Posterior Distributions." *The Annals of Statistics* 29 (3): 687–714.

Stone, Charles J. 1980. "Optimal Rates of Convergence for Nonparametric Estimators." *The Annals of Statistics* 8 (6): 1348–1360.

Stone, Charles J. 1982. "Optimal Global Rates of Convergence for Nonparametric Regression." *The Annals of Statistics* 10 (4): 1040–1053.

Talagrand, Michel. 1996. "Majorizing Measures: The Generic Chaining." *The Annals of Probability* 24 (3): 1049–1103.

———. 2014. *Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems.* Vol. 60. Springer Science & Business Media.

Tsai, Jerry, and Jessica A. Wachter. 2016. "Rare Booms and Disasters in a Multisector Endowment Economy." *The Review of Financial Studies* 29 (5): 1113–1169.

van der Vaart, Aad W., and Harry J. van Zanten. 2008. "Rates of Contraction of Posterior Distributions based on Gaussian Process Priors." *The Annals of Statistics* 36, no. 3 (June): 1435–1463.

Walker, Stephen G. 2007. "Sampling the Dirichlet Mixture Model with Slices." *Communications in Statistics – Simulation and Computation* 36 (1): 45–54.

Wong, Wing Hung, and Xiaotong Shen. 1995. "Probability Inequalities for Likelihood Ratios and Convergence Rates of Sieve MLEs." *The Annals of Statistics* 23 (2): 339–362.

Yang, Yuhong, and Andrew Barron. 1999. "Information-Theoretic Determination of Minimax Rates of Convergence." *The Annals of Statistics* 27 (5): 1564–1599.

3.A Measure Concentration

Generic Chaining

We start with recalling a few definitions and fixing some notation. Recall the definition of a γ-functional, where the infimum is taken with respect to all subsets $\mathcal{X}_s \subset \mathcal{X} \subset \mathbb{R}^{T\times D}$ such that the cardinality $|\mathcal{X}_s| \leq 2^{2^s}$ and $|\mathcal{X}_0| = 1$, and d is a metric: $\gamma_\alpha(\mathcal{X}, d) = \inf\sup_{x\in\mathcal{X}} \sum_{s=0}^{\infty} 2^{s/\alpha} d(s, \mathcal{X}_s)$. This $\gamma_2(\mathcal{X}, d)$ functional is useful because it controls the expected size of a Gaussian process by the majorizing measures theorem, (Talagrand 1996).

Recall the definition of the Orlicz norm of order n: $\psi_n \coloneqq \inf_{C>0} \mathbb{E}\left[\exp\left(\frac{|X|^n}{C^n} - 1\right) \leq 1\right]$. This is useful because a standard argument shows if X has a bounded ψ_n norm then the tail of X decays faster than $2\exp\left(-\frac{x^n}{\|x\|_{\psi_n}^n}\right)$. Hence, if x has a finite ψ_2-norm, it is subgaussian.

Definition and Properties of the Θ_T-operator

Lemma 3.11. *Let K be the number of columns of Θ_T as defined in Definition 3.3. Then its probability density function has the following form, where $\mu \coloneqq \Pr(b = 1)$.*

$$\Pr(K \leq \widetilde{K}) = \left(1 - (1-\mu)^{\widetilde{K}}\right)^T \tag{3.42}$$

Proof. Let θ_t denote a row of Θ_T. Then

$$\Pr(K \leq \widetilde{K}) = \Pr(1 \in \theta_t \text{ for all } t) = \Pr(1 \in \theta_t)^T = (1 - \Pr(1 \notin \theta_t))^T = \left(1 - (1-\mu)^{\widetilde{K}}\right)^T. \tag{3.43}$$

□

Lemma 3.12. *There exists a constant $\gamma \in (0,1)$ and constants c_1, c_2, such that with probability at least γ, the following holds.*

$$c_1 \log(T) \leq K \leq c_2 \log(T) \tag{3.44}$$

Proof. Let $B \coloneqq \exp(\widetilde{K})$ We set the cumulative distribution function equal to $1-\gamma$, i.e. the survival function equal to γ:

$$(1-\gamma) = (1 - (1-\mu)^{\widetilde{K}})^T \implies \log(1-\gamma)/T = \log(1 - (1-\mu)^{\tilde{K}}). \tag{3.45}$$

Note, for positive a and b, $a^{\log(b)} = b^{\log(a)}$, which can be verified by taking logs of both sides.

$$\log(1-\gamma)/T = \log\left(1 - \left(\frac{1}{1-\mu}\right)^{-\log B}\right) = \log\left(1 - \left(\frac{1}{B}\right)^{-\log(1-\mu)}\right) = \log\left(1 - B^{\log(1-\mu)}\right). \tag{3.46}$$

Taking the Taylor series approximation of the logarithm function around 1 gives

$$-\log(1-\gamma)/T \approx B^{\log(1-\mu)} \implies T \propto B^{-\log(1-\mu)} \implies B \propto T^{-1/\log(1-\mu)}. \tag{3.47}$$

This implies

$$K \propto -\frac{1}{\log(1-\mu)} \log(T) \propto \log(T). \tag{3.48}$$

We can bound this in the opposite direction by replacing $1-\gamma$ with γ since $\gamma \in (0,1)$.

□

Relationship between the Orlicz and L_2 norms.

We use the following lemma in our proof of Theorem 3.1. We need it to bound the tail deviations using a bound on the 2nd moment deviations.

Lemma 3.13. *Let Θ_T be a matrix constructed as in Definition 3.3. Let $\{x_t\}_{t=1}^T$ be a sequence of known random vectors of length D. Then we have the following.*

1. *The squared L_2-norm of x is equivalent to $\mathbb{E}\left[\langle \Theta_k, x\rangle^2\right]$.*
2. *The squared L_2-norm of x, $\|x\|_{L_2}^2$ dominates the 2nd-order Orlicz norm.*

Proof.

Part 3.1. First, we start by showing Item 1. Let Θ_k denote a column of the matrix. The root of the proof follows from realizing that Θ_T is a generalized selection matrix, and covariances are dominated by variances:

$$\mathbb{E}_\Theta\left[X'\Theta_k\Theta_k'X\right] = \mathbb{E}_\Theta\left[\sum_{t=1}^T x_t\theta_{t,k}\theta_{t,k}x_t'\right] = \mathbb{E}_{\Theta_k}\left[\sum_{t=1}^T |\theta_{t,k}|x_t x_t'\right] = \frac{1}{K}\sum_{t=1}^T x_t x_t', \tag{3.49}$$

where the last line follows by the independence of the rows of Θ_k.

Consider $\mathbb{E}_\Theta\left[X'\Theta\Theta'X\right]$. Since the columns of Θ_T are a martingale difference sequence, variances of sums are sums of variances:

$$\mathbb{E}_\Theta\left[X'\Theta\Theta'X\right] = \sum_{k=1}^K \mathbb{E}_{\Theta_k}\left[X'\Theta_k\Theta_k'X\right] = \sum_{t=1}^T x_t x_t'. \tag{3.50}$$

Part 3.2. Now that we have shown Item 1, we must show that L_2-norm dominates the ψ_2-norm. This is useful because it implies that if we can control the variance of the distribution, we automatically control the tails as well:

$$\inf\left\{C>0\,\middle|\,\mathbb{E}\left[\exp\left(\frac{|\langle\Theta_k,x\rangle|^2}{C^2}\right)\right]-1\leq 1\right\} \tag{3.51}$$

$$=\inf\left\{C>0\,\middle|\,\mathbb{E}\left[\exp\left(\frac{\sum_{t=1}^T|\theta_{t,k}|x_t'x_t+2\sum_{t,\tau\neq t}\theta_{t,k}\theta_{\tau,k}x_t'x_\tau}{C^2}\right)\right]\leq 2\right\}. \tag{3.52}$$

Since the cross-terms are proportional to squares, and the Θ_k are generalized selection vectors this bounded by

$$\inf\left\{C>0\,\middle|\,\mathbb{E}\left[\exp\left(\frac{2\sum_{t=1}^T|\theta_{t,k}|x_t'x_t}{C^2}\right)\right]\leq 2\right\}. \tag{3.53}$$

By the definition of the exponential function, $|\theta_{t,k}|\in\{0,1\}$, and the multinomial theorem, this equals

$$\inf\left\{C>0\,\middle|\,\mathbb{E}\left[\sum_{h=0}^\infty\frac{2^h\left(\sum_{t=1}^T|\theta_{t,k}|x_t'x_t\right)^h}{C^{2h}h!}\right]\leq 2\right\} \tag{3.54}$$

$$=\inf\left\{C>0\,\middle|\,\mathbb{E}\left[\sum_{h=0}^\infty\frac{2^h\sum_{\sum k_t=h}\binom{h}{k_1,k_2,\ldots k_T}\prod_{t=1}^T|\theta_{t,k}|(x_t'x_t)^{k_t}}{C^{2h}h!}\right]\leq 2\right\}. \tag{3.55}$$

Since everything is absolutely convergent, we can interchange expectations and infinite sums, and so this equals

$$\inf\left\{C>0\,\middle|\,\sum_{h=0}^\infty\frac{2^h\sum_{\sum k_t=h}\binom{h}{k_1,k_2,\ldots,k_T}\prod_{t=1}^T\frac{1}{K}(x_t'x_t)^{k_t}}{C^{2h}h!}\leq 2\right\}. \tag{3.56}$$

Then we can use the multinomial theorem and the formula for the exponential function in the reverse direction, implying this equals

$$\inf\left\{C>0\,\middle|\,\frac{1}{K}\exp\left(\frac{2\|x\|_{L_2}^2}{C^2}\right)\leq 2\right\}=\inf\left\{C>0\,\middle|\,\frac{2\|x\|_{L_2}^2}{C^2}=\log(2K)\right\}\leq\frac{\sqrt{2}\|x\|_{L_2}}{\sqrt{\log(2)}}, \tag{3.57}$$

where the last inequality follows because $K\geq 1$. Hence, we have that the L_2-norm dominates the ψ_2-norm.

□

Norm Equivalence

In the section below we reproduce Klartag and Mendelson (2005, Prososition 2.2). The one change that we make is that we spell out one of the constants as a function of its arguments. We do this because we will need to take limits with respect to δ on what follows.

Proposition 3.14 (Klartag and Mendelson (2005) Proposition 2.2)**.** *Let $(\mathcal{X}, d)$ be a metric space and let $\{Z_x\}_{x\in\mathcal{X}}$ be a stochastic process. Let $K > 0, \Upsilon : [0,\infty) \to \mathbb{R}$ and set $W_x \coloneqq \Upsilon(|Z_x|)$ and $\epsilon \coloneqq \frac{\gamma_2(\mathcal{X},d)}{\sqrt{K}}$. Assume that for some $\eta > 0$ and $\exp(-c_1(\eta)K) < \delta < \frac{1}{4}$, the following hold.*

1. *For any $x, y \in \mathcal{X}$ and $u < \delta_0 \coloneqq \frac{4}{\eta}\log\frac{1}{\delta}$,*

$$\Pr\left(|Z_x - Z_y| > ud(x,y)\right) < \exp\left(-\frac{\eta}{\delta_0}Ku^2\right)$$

2. *For any $x, y \in \mathcal{X}$ and $u > 1$*

$$\Pr\left(|W_x - W_y| > ud(x,y)\right) < \exp\left(-\eta K u^2\right)$$

3. *For any $x \in \mathcal{X}$, with probability larger than $1-\delta$, $|Z_x| < \epsilon$.*

4. *Υ is increasing, differentiable at zero and $\Upsilon'(0) > 0$.*

Then, with probability larger than $1-2\delta$, with $C(\Upsilon,\delta,\eta) \coloneqq \left(c(\Upsilon)c(\eta)(\frac{2}{\eta}(\log\frac{1}{\delta}+1))\right) > 0$, and both $c(\Upsilon)$ and $c(\eta)$ depend solely on their arguments.

$$\sup_{x\in\mathcal{X}}|Z_x| < C(\Upsilon,\delta,\eta)\epsilon.$$

Here we quote a version of Bernstein's inequality for martingales due to (de la Peña 1999, Theorem 1.2A), which we use later.

Theorem 3.15 (Bernstein's Inequality for Martingales)**.** *Let $\{x_i, \mathcal{F}_i\}$ be a martingale difference sequence with $\mathbb{E}[x_i \,|\, \mathcal{F}_{i-1}] = 0, \mathbb{E}[x_i^2 \,|\, \mathcal{F}_{i-1}] = \sigma_i^2, v_k = \sum_{i=1}^k \sigma_i^2$. Furthermore, assume that $\mathbb{E}[|x_i|^n \,|\, \mathcal{F}_{i-1}] \leq \frac{n!}{2}\sigma_i^2 M^{n-2}$ almost everywhere. Then, for all $x, y > 0$,*

$$\Pr\left(\left\{\left|\sum_{i=1}^{k} x_i\right| \geq u, v_k \leq y \text{ for some } k\right\}\right) \geq 2\exp\left(-\frac{u^2}{2(y+uM)}\right). \tag{3.58}$$

If we choose c small enough, this implies

$$\Pr\left(\left\{\left|\frac{1}{k}\sum_{i=1}^{k} x_i\right| \geq u, v_k \leq y \text{ for some } k\right\}\right) \geq 2\exp\left(-c\min\left\{\frac{u^2k^2}{v}, \frac{uk}{M}\right\}\right). \tag{3.59}$$

Theorem 3.1 (Bounding the Norm Perturbation)**.** *Let Θ_T be constructed as in Definition 3.3 with the number of columns denoted by K_T. Let $\epsilon > 0$ be given. Let $0 < \delta < 1$ be given such that $0 < \log(\frac{1}{\delta}) < c_1\epsilon^2 K_T$ for some constant c_1. Let $\widetilde{X}_T$ be in the unit hypersphere in $\mathbb{R}^{TD-1}$. Then with probability greater than $1-2\delta$ with respect to Θ_T, there exists a constant c_2 such that for any $\epsilon > \sqrt{\frac{\log T}{K_T}}$,*

$$\sup_t \left|\|\theta_t\widetilde{x}_t\|_{L_2} - \|\widetilde{x}_t\|_{L_2}\right| < c_2\left(1+\log\left(\frac{1}{\delta}\right)\right)\epsilon.$$

Proof. We mimic the proof of Klartag and Mendelson 2005, Theorem 3.1, verifying the conditions of Proposition 3.14. Similar to them we use $\Upsilon(t) \coloneqq \sqrt{1-t}$. Our conclusion is stated in terms of the logarithm of the sample size — T. This conclusion is weaker than theirs as $\gamma_2\left(\widetilde{\mathcal{X}}, \|\cdot\|_{L_2}\right) < C\sqrt{\log(T)}$. We can see this by combining the majorizing measure theorem, (Talagrand 2014, Theorem 2.4.1), and the minoration theorem, (Lemma 2.4.2). Effectively, we have an upper bound for the supremum of a Gaussian process and tighter upper bound for the same process.

We start by fixing some notation. Let $x, y \in \mathcal{X}$. We use the functional notation $x(\theta_k)$ to refer $\sum_{d=1}^{D} \theta'_k x_d$.

$$Z_x^K \coloneqq \frac{1}{K}\sum_{k=1}^{K} x^2(\theta_k) - \|x\|_{L_2}^2 \tag{3.60}$$

Consider $Z_x^K - Z_y^K$.

$$Z_x^K - Z_y^K = \frac{1}{K}\sum_{k=1}^{K} x^2(\theta_k) - y^2(\theta_k) = \frac{1}{K}\sum_{k=1}^{K}(x-y)(\theta_k)(x+y)(\theta_k) \tag{3.61}$$

Part 3.3. Let $Y_k \coloneqq x^2(\theta_k) - y^2(\theta_k)$, then

$$\begin{aligned}
&\Pr(|Y_k| > 4u\|x-y\|_{\psi_2}\|x+y\|_{\psi_2}) \qquad (3.62)\\
&\quad \leq \Pr(|x(\theta_k) - y(\theta_k)| > 2\sqrt{u}\|x-y\|_{\psi_2}) + \Pr(|x(\theta_k) + y(\theta_k)| > 2\sqrt{u}\|x+y\|_{\psi_2})\\
&\quad \leq 2\exp(-u),
\end{aligned}$$

where the last inequality comes from the sub-exponential tails of $\theta_{t,k}$ and the first by the union bound. This implies that $\|Y_k\|_{\psi_1} \leq c_1\|x-y\|_{\psi_2}\|x+y\|_{\psi_2} \leq c_2\|x-y\|_{\psi_2}$. We do not need the β used by Klartag and Mendelson because the entries in our θ operator are uniformly bounded by 1 in absolute value.

The Y_k are a martingale difference sequence, and so we can apply Theorem 3.15. They are a martingale difference sequences because the expectation in the next period is either the current value because the increments are mean zero if the sum does not stop or identically zero if they do. If we set $v = 4K\|Y_k\|_{\psi_1}^2$ we can use Bernstein's inequality for martingales mentioned above. $\sum_{k=1}^K \sigma_k^2 \leq v$ with probability 1 because this variance is either the same as it is in the independent case or zero. Consequently, by Theorem 3.15, we have the following if set $v \coloneqq 4K\|\theta\|_{\psi_1}^2$ and $M = \|\theta\|_{\psi_1}$:

$$\Pr\left(\left\{\left|\frac{1}{K}\sum_{k=1}^K \theta_k\right| > u\right\}\right) \leq 2\exp\left(-cK\min\left\{\frac{u^2}{\|\theta\|_{\psi_1}^2}, \frac{u}{\|\theta\|_{\psi_1}}\right\}\right) \qquad (3.63)$$

Then by applying (3.63) to $\Pr(|z_x^k - z_y^k| > u)$, we have the following.

$$\Pr\left(\left|Z_x^k - Z_y^k\right| > u\right) \leq 2\exp\left(-c\min\left\{\frac{u^2}{\|x-y\|_{L_2}^2}, \frac{u}{\|x-y\|_{L_2}}\right\}\right) \qquad (3.64)$$

The estimate for $\Pr\left(\left|Z_x^k\right| > u\right)$ follows from the same method, but we define $Y_k \coloneqq x^2(\theta_k) - 1$, and use the fact that $\|x(\theta)\|_{\psi_2} \leq 1$, which we verified in Lemma 3.13. The L_2-norm is bounded above by 1 because we are using rescaled data.

We fix $\eta \leq c$. Assume that $u < \delta_0 = 4\frac{1}{\eta}\log\frac{1}{\delta}$. Then we have

$$\Pr\left(\left|Z_x^k - Z_y^k\right| > 2\|x-y\|_{L_2}\right) \leq 2\exp\left(\eta K\min\left\{u, u^2\right\}\right) < \exp\left(-\eta K\frac{u^2}{\delta_0}\right). \qquad (3.65)$$

Part 3.4. By the triangle inequality,

$$|W_x - W_y| = \left| \left(\frac{1}{K} \sum_{k=1}^{K} x^2(\theta_i) \right)^{1/2} - \left(\frac{1}{K} \sum_{k=1}^{K} y^2(\theta_i) \right)^{1/2} \right| \leq \left(\frac{1}{K} \sum_{k=1}^{K} (x-y)^2(\theta_i) \right)^{1/2}. \tag{3.66}$$

Applying (3.63) for $u > 1$:

$$\begin{aligned} \Pr\left(|W_x - W_y| > u\|x-y\|_{\psi_2} \right) &\leq \Pr\left(\frac{1}{K} \sum_{k=1}^{K} (x-y)^2(\theta_k) > u^2 \|x-y\|_{\psi_2}^2 \right) \\ &\leq \Pr\left(\frac{1}{K} \sum_{k=1}^{K} (x-y)^2(\theta_k) > u^2 \left\|(x-y)^2\right\|_{\psi_1} \right) \\ &< \exp\left(-cku^2\right). \end{aligned} \tag{3.67}$$

Since $\eta < c$, this is bounded by $\exp(-\eta K u^2)$.

Part 3.5. For any $x \in \mathcal{X}$ by (3.63),

$$\Pr(|Z_x| > \epsilon) < \exp(-\eta K \epsilon^2) < \delta. \tag{3.68}$$

Part 3.6. We can bound the derivative of Υ:

$$\Upsilon'(0) = 1/2 > 0. \tag{3.69}$$

□

3.B Representation Theory

The Joint Density

Lemma 3.16 (Bouding Ratio of Sums by Max Ratio)**.** *Let x_t, y_t be a sequence of numbers whose sum is absolutely convergent. Then the ratio of the sums is bounded by the supremum of the ratios, i.e.,*

$$\frac{\sum x_t}{\sum y_t} \leq \sup_t \frac{x_t}{y_t}.$$

Proof. Clearly, if $\#t = 1$, the result holds. Assume $\#t = 2$. Assume the claim is false. Then

$$\frac{x_1 + x_2}{y_1 + y_2} > \max\left\{\frac{x_1}{y_1}, \frac{x_2}{y_2}\right\} \implies x_1 + x_2 > \max\left\{x_1 + \frac{x_1 y_2}{y_1}, x_2 + \frac{x_2 y_1}{y_2}\right\} \tag{3.70}$$
$$\implies x_1 > \frac{x_2 y_1}{y_2} \text{ and } x_2 > \frac{x_1 y_2}{y_1} \implies x_1 > \frac{y_1}{y_2}\frac{x_1 y_2}{y_1} \implies x_1 > x_1.$$

This is a contradiction. To see the general case we proceed by induction,

$$\frac{\sum_t x_t}{\sum_t y_t} \leq \max\left\{\frac{\sum_{t \neq T} x_t}{\sum_{t \neq T}}, \frac{x_T}{y_T}\right\} \leq \cdots \leq \max\left\{\frac{x_t}{y_t}\right\}, \tag{3.71}$$

where the first inequality holds by the first step. Clearly, as long as everything convergent, this still holds if we take limits. □

Lemma 3.17. *Consider the ratio of the densities between p_T and q_T. Let δ_k^q be a clustering of x_t with respect to q_T. Let these clusters δ_k^q satisfy the following, where $\mu_k^q = \mathbb{E}_{P_T}\left[x_t \,|\, t \in \delta_k^q\right]$ and $\Sigma_k^q = \mathbb{C}\mathrm{ov}_{P_T}\left[x_t \,|\, x_t \in \delta_t^q\right]$:*

$$\sup_{\delta_k^q} \sup_{x_t \in \delta_k^q} \left| (x_t - \mu_t)' \Sigma_t^{-1} (x_t - \mu_t) - (x_t - \mu_k^q)' \left(\Sigma_k^q\right)^{-1} (x_t - \mu_k^q) \right| < C(\delta)\epsilon. \tag{3.72}$$

Then the log-divergence satisfies

$$\sup_{x_t, x_t^*} \left| (x_t - \mu_t)' \Sigma_t^{-1} (x_t - \mu_t) - (x_{t^*} - \mu_{t^*})' \Sigma_{t^*}^{-1} (x_{t^*} - \mu_{t^*}) \right| < \epsilon \implies \sup_{x_t, x_t^*} \left| \log\left(\frac{p_T(x_t)}{p_T(x_{t^*})}\right) \right| \propto \epsilon. \tag{3.73}$$

Proof. Consider the log-ratio of Gaussian kernels, by assumption

$$\sup_{\delta_k^q} \sup_{x_t \in \delta_k^q} \left| (x_t - \mu_t)' \Sigma_t^{-1} (x_t - \mu_t) - (x_t - \mu_k^q)' \left(\Sigma_k^q\right)^{-1} (x_t - \mu_k^q) \right| \leq \epsilon. \tag{3.74}$$

Consider the ratio of the proportionality constants χ^p and χ^q associated with the kernels k^p, k^q above:

$$\chi^p = \int_{\mathcal{X}} k^p(x)\, dx, \quad \chi^q = \int_{\mathcal{X}} k^q(x)\, dx. \tag{3.75}$$

By the definition of proportionality constant, we can write

$$\log\left(\frac{\chi^q}{\chi^p}\right) = \log\left(\frac{\sum k^q(x)\, dx}{\sum k^p(y)\, dy}\right) = \log\left(\frac{\sum k^q(x)/p_T(x)\, dP_T(x)}{\sum k^p(y)/p_T(y)\, dP_T(y)}\right), \tag{3.76}$$

where we can change measures to P_T. By Lemma 3.16, this is bounded by the supremum of the ratios, since we are integrating over the same space in both sums:

$$\leq \sup_x \log \left(\frac{k^q(x)/p_T(x)}{k^p(x)/p_T(x)} \right) \leq \sup_x \log \left(\frac{k^q(x)}{k^p(x)} \right), \tag{3.77}$$

because the Jacobian terms cancel. We can bound the inverse-ratio of the proportionality constants — $\frac{\mu_q}{\mu_p}$ — in the same way. We just interchange the labels on the kernels. Consequently, the proportionality constants satisfy

$$\left| \log \frac{\mu_1}{\mu_2} \right| = \frac{1}{2} c(\delta) \epsilon \tag{3.78}$$

because the $k^{\cdot}(x)$ are Gaussian kernels, and we bounded the log-ratio in (3.74). The total deviation is the sum of the deviation in the constants and in the kernels. The result holds by combining (3.78) and (3.74).

□

Proposition 3.18 (Bounding the Supremum of the Rescaled Data)**.** *Let $\widetilde{X} := \widetilde{x}_1, \ldots, \widetilde{x}_T$ be a D-dimensional Gaussian process with finite stochastic means μ_t and covariances Σ_t, where Σ_t is positive-definite for all t. Let Θ_T be the generalized selection matrix defined in Definition 3.3. Let $\widetilde{P}_T$ denote the distribution of $\widetilde{X}$. Then given $\epsilon > 0$ and for some $\delta \in (0, 1)$, the approximating distribution $\widetilde{Q}_T$, which is the mixture distribution over $\{\widetilde{\Sigma}_t^{-1/2} \widetilde{x}_t\}_{t=1}^T$ defined by the clustering induced by Θ_T satisfies the following with probability at least $1 - 2\delta$ with respect to Θ_T.*

$$\sup_t h^2 \left(\int_{G_t} \phi \left(\widetilde{x}_t \,\middle|\, \delta_t^P \right) dG_t^P(\delta_t^P), \int_{G_t} \phi \left(\widetilde{x}_t \,\middle|\, \delta_t^Q \right) dG_t^Q(\delta_t^Q) \right) < c \left(1 + \log \frac{1}{\delta} \right)^2 \epsilon^2 \tag{3.79}$$

Proof. In this proof, we drop the tilde's over the x_t because all of the terms have them. Let G^P and G^Q be the associated mixing measures of the covariances. Let $\mathcal{K}$ be a coupling from between the space of G^P and G^Q. Consider the supremum of the squared Hellinger distance — h^2 — between P_T and Q_T:

$$\sup_t h^2 \left(\int_{G_t} \phi \left(x_t \,\middle|\, \delta_t^P \right) dG_t^P(\delta_t^P), \int_{G_t} \phi \left(x_t \,\middle|\, \delta_t^Q \right) dG_t^Q(\delta_t^Q) \right). \tag{3.80}$$

Combining the integrals with respect to the marginals (G_t^P, G_t^Q) into a integral with respect to the joint, and exploiting the convexity of the supremum and of the squared Hellinger distance gives:

$$\leq \int_{G_t^P \times G_t^Q} \sup_t h^2 \left(\phi\left(x_t \,\middle|\, \delta_t^P\right), \phi\left(x_t \,\middle|\, \delta_t^Q\right) \right) d\mathcal{K}(G_t^P, G_t^Q). \tag{3.81}$$

We expand the definition of h^2 using its formula as an f-divergence:

$$\leq \int_{G_t^P \times G_t^Q} \sup_t \int_{\mathbb{R}^D} \left| \left(\frac{\phi\left(x_t \,\middle|\, \delta_t^P\right)}{\phi\left(x_t \,\middle|\, \delta_t^Q\right)} \right)^{1/2} - 1 \right|^2 d\Phi\left(x_t \,\middle|\, \delta_t^Q\right) d\mathcal{K}(G_t^P, G_t^Q). \tag{3.82}$$

Since we are only considering the density for one period within the integral:

$$= \int_{G_t^P \times G_t^Q} \int_{\mathbb{R}^D} \sup_t \left| \left(\frac{\phi\left(x_t \,\middle|\, \delta_t^P\right)}{\phi\left(x_t \,\middle|\, \delta_t^Q\right)} \right)^{1/2} - 1 \right|^2 d\Phi\left(x_t \,\middle|\, \delta_t^Q\right) d\mathcal{K}(G_t^P, G_t^Q). \tag{3.83}$$

By Lemma 3.17 and a first-order Taylor series of the exponential function around the logarithm of the original argument, after pulling the square-root inside

$$\leq C_1 \int_{G_t^P \times G_t^Q} \int_{\mathbb{R}^D} \sup_t \left| (x_t - \mu_t^P)' \Sigma_t^P (x_t - \mu_t^P) - (x_t - \mu_t^Q) \Sigma_t^Q (x_t - \mu_t^Q) \right| d\Phi\left(x_t \,\middle|\, \delta_t^Q\right) d\mathcal{K}(G_t^P, G_t^Q). \tag{3.84}$$

Since Q_T was defined through applying Θ_T to $(\Sigma_t^P)^{-1/2}(x_t - \mu_t^P)$, by Theorem 3.1 this norm perturbation is bounded by ϵ^2; we just have to square the constant:

$$\leq C \left(1 + \log \frac{1}{\delta}\right)^2 \int_{G_t^P \times G_t^Q} \int_{\mathbb{R}^D} |\epsilon|^2 \, d\Phi\left(x_t \,\middle|\, \delta_t^Q\right) d\mathcal{K}(G_t^P, G_t^Q) = C \left(1 + \log \frac{1}{\delta}\right)^2 \epsilon^2, \tag{3.85}$$

where the last equality holds because all of the integrals integrate to 1. □

Theorem 3.2 (Representing the Joint Density)**.** *Let $\widetilde{X}_T := \widetilde{x}_1, \ldots, \widetilde{x}_T$ be a D-dimensional Gaussian process with period-t finite stochastic means, μ_t, and covariances, Σ_t. Let Σ_t be positive-definite for all t. Let Θ_T be the generalized selection matrix constructed in Definition 3.3. Let $\widetilde{P}_T$ denote the distribution of $\widetilde{X}_T$. Then*

given $\epsilon > 0$ and $\delta \in (0,1)$, the approximating distribution, Q_T, which is the mixture distribution over $\widetilde{\mathcal{X}}$ that Θ_T induces, satisfies the following with probability at least $1-2\delta$ with respect to Θ_T for some constant C:

$$h_\infty\left(\widetilde{P}_T(\widetilde{\mathcal{X}}), \widetilde{Q}_T(\widetilde{\mathcal{X}})\right) < C\left(1+\log\left(\frac{1}{\delta}\right)\right)\epsilon.$$

Proof. Let G^P, G^Q be the associated mixing measures of the associated covariances. Let $\mathcal{K}$ be a coupling from between the space of G^P and G^Q, and the space of such couplings be $\mathcal{T}(G^P, G^Q)$. Consider the squared supremum Hellinger distance — h^2_∞ — between P_T and Q_T. The proof here is based on a combination of proofs of Nguyen (2016, Lemma 3.1) and Nguyen (2016, Lemma 3.2). Let δ_t be the latent mixture identity that tells you which cluster μ_t, Σ_t is in.

We can represent both densities succinctly as follows. Importantly, we do not require that the G^P_t are independent:

$$p_T(\widetilde{\mathcal{X}}) = \int_G \int_{G_t} \phi\left(x_t \,\middle|\, \delta^P_t\right) dG^P_t\left(\delta^P_t\right) dG^P\left(dG^P_t\right). \tag{3.86}$$

We represent q_T in the same fashion replacing the P's in the expression above with Q's:

$$q_T(\widetilde{\mathcal{X}}) = \int_G \int_{G_t} \phi\left(x_t \,\middle|\, \delta^Q_t\right) dG^Q_t\left(\delta^Q_t\right) dG^Q\left(dG^Q_t\right). \tag{3.87}$$

Then the squared sup-Hellinger distance between the two measures has the following form:

$$\begin{aligned} &h^2_\infty\left(p_T(\widetilde{\mathcal{X}}), q_T(\widetilde{\mathcal{X}})\right) \qquad (3.88)\\ &= h^2_\infty\left(\int\int \phi\left(x_t \,\middle|\, \delta^P_t\right) dG^P_t(\delta^P_t)\, dG^P(dG^P), \int\int \phi\left(x_t \,\middle|\, \delta^Q_t\right) dG^Q_t(\delta^Q_t)\, dG^Q(dG^Q_t)\right). \end{aligned}$$

Letting $\mathcal{K}(G^P, G^Q)$ be any coupling between the two densities, we can combine G^P and G^Q into one process. We want to integrate with respect to their joint density:

$$\begin{aligned} &= h^2_\infty\Bigg(\int_G \int_{G_t} \phi\left(x_t \,\middle|\, \delta^P_t\right) dG^P_t(\delta^P_t)\, d\mathcal{K}(dG^P_t, dG^Q_t), \qquad (3.89)\\ &\qquad \int_G \int_{G_t} \phi\left(x_t \,\middle|\, \delta^Q_t\right) dG^Q_t(\delta^Q_t)\, d\mathcal{K}(dG^P_t, dG^Q_t)\Bigg). \end{aligned}$$

Since supremum of squared Hellinger distance is convex as is the supremum, by Jensen's inequality that is bounded

$$\leq \int_{G\times G} \sup_t h^2 \left(\int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^P\right) dG_t^P(\delta_t^P), \int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q) \right) d\mathcal{K}(dG_t^P, dG_t^Q). \tag{3.90}$$

If we can bound the supremum of the deviations over the periods, we have bounded the joint. This is true even in the dependent case.

We can place the bound obtained in Proposition 3.18 inside (3.90). Since we are integrating $C\epsilon^2$ over a joint density, the density is bounded above by 1, and we are done.

In other words, we have with probability $1 - 2\delta$:

$$h_\infty^2(P_T(\widetilde{\mathcal{X}}), Q_T(\widetilde{\mathcal{X}})) < C \left(\log \frac{1}{\delta}\right)^2 \epsilon^2. \tag{3.91}$$

□

Lemma 3.19. *Let f, g be two densities of locally asymptotically normal (LAN) processes with respect to the sample size T.*[56] *Squared Hellinger distance and Kullback-Leibler divergence are equivalent.*

Proof. Consider the following decomposition of the Hellinger distance:

$$\int (\sqrt{f/g} - 1)\, dG = \int \left(\exp\left(\frac{1}{2}(\log f - \log g) \right) - 1 \right) dG. \tag{3.92}$$

Taking a Taylor expansion of the exponential function:

$$= \int \left(1 + \frac{1}{2} \log\left(\frac{f}{g}\right) + O\left(\log\left(\frac{f}{g}\right)^2 \right) - 1 \right) dG \tag{3.93}$$

$$= \int \frac{1}{2} \log\left(\frac{f}{g}\right) dG + O\left(\int \log\left(\frac{f}{g}\right)^2 dG \right). \tag{3.94}$$

Consider one-half the Kullback-Leibler divergence:

$$\frac{1}{2} \int \log\left(\frac{f}{g}\right) \frac{f}{g}\, dG = \frac{1}{2} \int \log\left(\frac{f}{g}\right) \exp\left(\log\left(\frac{f}{g}\right) \right) dG. \tag{3.95}$$

[56] This trivially covers all Gaussian processes with finite-means and variances.

Taking a 1st-order Taylor expansion of the exponential function:

$$= \frac{1}{2}\int \log\left(\frac{f}{g}\right)\left(1+\log\left(\frac{f}{g}\right)\right) dG = \frac{1}{2}\int \log\left(\frac{f}{g}\right) dG + O\left(\int \log\left(\frac{f}{g}\right)\log\left(\frac{f}{g}\right) dG\right). \tag{3.96}$$

The first terms in (3.93) and (3.96) are the same. Consequently, is of the same-asymptotic order as $\log(f/g)^2$. By the locally asymptotically normal assumption $\log f(x) \propto (x-\mu_f)'\Sigma_f^{-1}(x-\mu_f) + o(T)$ Choose $\epsilon \propto \frac{1}{T}$. Let z denote the deviation above. By the convexity of the square function and Jensen's inequality, it is sufficient to bound the value inside the integral:

$$\int \log(f/g)^2\, dG \leq \int |z|^2\, dG + O(\epsilon) \leq \int |z|\, dG + O(\epsilon) = \int \log(f/g)\, dG + O(\epsilon), \tag{3.97}$$

where the first inequality holds by the LAN property, the second inequality holds since $|z| < 1$, and the third-inequality holds by the LAN property. By (3.93) and (3.96), the last term in (3.97) is bounded by both the Hellinger and Kullback-Leibler divergences.

□

Representing the Marginal Density

Theorem 3.3 (Representing the Marginal Density)**.** *Let $x_1, \ldots, x_T$ be drawn independently from P_T where each x_t has a infinite-Gaussian mixture representation. Let Θ_T be constructed as in Theorem 3.2 for each t. Let ϵ be given. Construct Q_T by using the Θ_T operator to group the data and letting the data be Gaussian distributed within each component with component-wise means and covariances given by their conditional expectations. Then, with probability $1 - 2\delta$ with respect to Θ_T, there exists a constant C such that the following holds uniformly over T*

$$h\left(\int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^P\right) dG_t(\delta_t^P), \int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t(\delta_t^Q)\right) < C\left(1 + \log\left(\frac{1}{\delta}\right)\right)\epsilon.$$

Proof. We start by comparing the Hellinger distance between the joint densities, which are both product measures. We want to compare the difference between the marginal densities in terms of the difference between the joint densities. In particular, we show that the difference between the marginal densities is $1/T$ times the difference between the joint densities if the joint densities have a product form. By Theorem 3.2, we know that is bounded by $T\epsilon^2$, and so we have the desired result. The unusual thing is that we are trying to bound the difference between the joint density and its components in the opposite direction as is usually done. We want to bind the component distance in terms of the joint density distance instead of the other way around.

We can write the squared Hellinger distance between the joint distributions as follows. Let G_m be the marginal distribution over δ_t. Note, the following holds:

$$\prod_{t=1}^{T} \int_{G_t} \phi\left(x_t \,|\, \delta_t\right) dG_t(\delta_t) = \prod_{t=1}^{T} \int_{G_m} \phi\left(x_t \,|\, \delta_t\right) dG_m(\delta_t). \tag{3.98}$$

All (3.98) is saying is that the joint T independent draws from the marginal are the same as T independent draws from a sequence $G_1, \ldots G_T$, which is drawn from G. By assumption G has a product form. The Kullback-Leibler

divergence between the two joint distributions is

$$\mathrm{D}_{\mathrm{KL}}\left(q_T \,||\, p_T\right) = \int_{\mathbb{R}^{T\times D}} \log\left(\frac{q_T}{p_T}\right) dP_T = \int_{\mathbb{R}^{T\times D}} \log\left(\frac{\prod_{t=1}^{T}\int_{G_t}\phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)}{\prod_{t=1}^{T}\int_{G_t}\phi\left(x_t \,|\, \delta_t^P\right) dG_t^P(\delta_t^P)}\right) dP_T. \tag{3.99}$$

Ratios of products are products of ratios, and logs of products are sums of logs, and we can substitute in the definition of the marginal distribution, (3.98), giving

$$= \int_{\mathbb{R}^{T\times D}} \sum_{t=1}^{T} \log\left(\frac{\int_{G_t}\phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)}{\int_{G_t}\phi\left(x_t \,|\, \delta_t^P\right) dG_t^P(\delta_t^P)}\right) dP_T = \int_{\mathbb{R}^{T\times D}} \sum_{t=1}^{T} \log\left(\frac{\int_{G_m}\phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_m^Q(\delta_t^Q)}{\int_{G_m}\phi\left(x_t \,|\, \delta_t^P\right) dG_m^P(\delta_t^P)}\right) dP_T. \tag{3.100}$$

We can rewrite P_T in terms of its mixture representation:

$$\int_{G_t}\int_{\mathbb{R}^{T\times D}} \sum_{t=1}^{T} \log\left(\frac{\int_{G_m}\phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)}{\int_{G_m}\phi\left(x_t \,|\, \delta_t^P\right) dG_t^P(\delta_t^P)}\right) \prod_{t=1}^{T} \phi\left(x_t \,|\, \delta_t\right) dx\, dG_m^P(\delta_t). \tag{3.101}$$

The only interactions between the two terms are the x_t:

$$= \sum_{t=1}^{T}\left(\left(\int_{G_t}\int_{\mathbb{R}^D} \log\left(\frac{\int_{G_m}\phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)}{\int_{G_m}\phi\left(x_t \,|\, \delta_t^P\right) dG_t^P(\delta_t^P)}\right) \phi\left(x_t \,|\, \delta_t\right) dx\, dG_m^P(\delta_t)\right)\left(\int_{\mathbb{R}^{(T-1)\times D}} \prod_{\tau\neq t} \phi\left(x_\tau \,|\, \delta_\tau\right) dx\, dG_m^P(\delta_\tau)\right)\right). \tag{3.102}$$

The second integrals all equal 1, and so their product does as well, giving

$$= \sum_{t=1}^{T}\left(\int_{G_t}\int_{\mathbb{R}^D} \log\left(\frac{\int_{G_m}\phi\left(x_t \,\middle|\, \delta_t^P\right) dG_t^P(\delta_t^P)}{\int_{G_m}\phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)}\right) \phi\left(x_t \,|\, \delta_t\right) dx\, dG_m^P(\delta_t)\right). \tag{3.103}$$

The term inside the sum is the Kullback-Leibler divergence between the two marginal distributions, which does not depend upon t:

$$= \sum_{t=1}^{T} \mathrm{D}_{\mathrm{KL}} \left(\int_{G_m} \phi \left(x_t \,\middle|\, \delta_t^Q \right) dG_m^Q(\delta_t^Q) \,\middle\|\, \int_{G_m} \phi \left(x_t \,\middle|\, \delta_t^P \right) dG_m^P(\delta_t^P) \right) \tag{3.104}$$

$$= T \mathrm{D}_{\mathrm{KL}} \left(\int_{G_m} \phi \left(x_t \,\middle|\, \delta_t^Q \right) dG_m^Q(\delta_t^Q) \,\middle\|\, \int_{G_m} \phi \left(x_t \,\middle|\, \delta_t^P \right) dG_m^P(\delta_t^P) \right). \tag{3.105}$$

In other words, the distance between the joint densities is at least T times the distance between the distance marginal densities. Also, by Lemma 3.19 this is proportional to squared Hellinger distance. In other words, the difference between the joint densities is at least T times the distance between the distance between the marginal densities. We know by Theorem 3.2 that this is bounded above by $CT\epsilon^2$. The T arises because we are no longer using the rescaled data, and $\|X\|^2 \propto T$. This gives

$$h^2 \left(\int_{G_m} \phi \left(x_t \,\middle|\, \delta_t^Q \right) dG_m^Q(\delta_t^Q), \int_{G_m} \phi \left(x_t \,\middle|\, \delta_t^P \right) dG_m^P(\delta_t^P) \right) \leq \frac{1}{T} h^2(q_T, p_T) \leq C \frac{T}{T} \epsilon^2 = C\epsilon^2. \tag{3.106}$$

□

Corollary 3.1 (Representing the Marginal Density with Markov Data)**.** *Theorem 3.3 continues to hold when the x_t form a uniformly ergodic hidden Markov chain instead of being fully independent.*

Proof. Let z_1 be a latent variable such that (x_t, z_t) forms Markov sequence. Consider a reshuffling $(\tilde{x}_1, \tilde{z}_1), \ldots (\tilde{x}_T, \tilde{z}_T)$. Now both of these sequences clearly have the same marginal distribution. (They likely do not have the same joint distribution.) Hence, by Theorem 3.3 the result follows since the reshuffled data has a product density.

□

Representing the Transition Density

Theorem 3.4 (Transition Density Representation)**.** *Let $x_1, \ldots, x_T \in R^{T \times D}$ be a uniformly ergodic Markov Gaussian process with density p_T. Let $\epsilon > 0$ be given. Let $K \geq c \log(T)^2/\epsilon$ for some constant c. Let δ_t be the cluster identity at time t. Then there exists a mixture density q_T with K clusters with the following form:*

$$q_T\left(x_t \,|\, x_{t-1}, \delta_{t-1}\right) \coloneqq \sum_{k=1}^{K} \phi\left(\beta_k x_{t-1}, \Sigma_k\right) \Pr\left(\delta_t = k \,|\, \delta_{t-1}\right).$$

Construct $q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^{Q}\right)$ from $q_T\left(x_t \,|\, x_{t-1}, \delta_{t-1}\right)$ by integrating out δ_{t-1} using $\Pr(\delta_{t-1} \,|\, X_T)$. Then with probability $1 - 2\delta$ with respect to the prior

$$h_\infty\left(p_T\left(x_t \,|\, \mathcal{F}_{t-1}^{P}\right), q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^{Q}\right)\right) < C\sqrt{1 + \log\left(\frac{1}{\delta}\right)}\epsilon.$$

Proof. We need the conditional density of $\widetilde{x}_t \,|\, \widetilde{x}_{t-1}, \delta_{t-1}$. By Theorem 3.2, there exists a generalized selection matrix Θ_T satisfying the statement of the theorem. Conditional on Θ_T, the distribution is Gaussian. So consider the following where θ_t is the t^{th} row of Θ_T. (Throughout, we will implicitly prepend a 1 to $\tilde{x}_{t-1}$ to allow for a non-zero mean as is standard in regression notation.)

By the linearity of Gaussian conditioning in $\theta_t\widetilde{x}_t, \theta_{t-1}\widetilde{x}_{t-1}$ space, for some $\beta_{k,k'}$, $\Sigma_{k,k'}$.

$$\theta_t\widetilde{x}_t \,|\, \widetilde{x}_{t-1}, \theta_t, \theta_{t-1} \overset{\mathcal{L}}{=} \theta_t\widetilde{x}_t \,|\, \theta_{t-1}\widetilde{x}_{t-1}, \theta_t, \theta_{t-1} \overset{\mathcal{L}}{=} \phi(\beta_{k,k'}\theta_{t-1}\widetilde{x}_{t-1}, \Sigma_{k,k'}) \overset{\mathcal{L}}{=} \phi(\beta_{k,k'}\widetilde{x}_{t-1}, \Sigma_{k,k'}). \tag{3.107}$$

The first equality holds because the elements in each cluster have the same Gaussian distribution under Q_T. The last equality holds because the elements of θ_{t-1} are in $\{-1, 0, 1\}$, we can absorb the θ_{t-1} into the $\beta_{k,k'}$ without increasing the number of clusters more than two-fold. This is because the vectors θ_{t-1} that contain at most one non-zero element form a convex hull, and we take the weighted averages over them in (3.108).

We want the distribution of $\widetilde{x}_t$ given $\theta_{t-1}, \widetilde{x}_{t-1}$. We do not want to condition on θ_t. So we can just integrate over θ_t using its distribution. Its predictive distribution does not depend upon $\widetilde{x}_{t-1}$ because we construct Θ_T independently of $\widetilde{x}$:

$$\widetilde{x}_t \,|\, \theta_{t-1} = k, \widetilde{x}_{t-1} \sim \sum_{k'} \phi(\beta_{k,k'} \widetilde{x}_{t-1}, \Sigma_{k,k'}) \Pr\left(\theta_t = k'\right) \tag{3.108}$$

The last probability — $\Pr(\theta_t = k')$ — does not have any conditioning information because the rows of the Θ_T process are independent except for the stopping rule, which is not relevant here. Define a set of clusters in $(\widetilde{x}_t, \widetilde{x}_{t-1})$ space by grouping the ones whose associated $\{\beta, \Sigma\}$ are equal. In other words, take the Cartesian product of the clusters used in (3.108) and denote the cluster identities by δ_t's. Integrating out the cluster identities gives

$$\widetilde{x}_t \,|\, \widetilde{x}_{t-1}, \delta_{t-1} \sim \sum_{j} \phi(\beta_j \widetilde{x}_{t-1}, \Sigma_j) \Pr\left(\delta_t = j \,|\, \delta_{t-1}\right). \tag{3.109}$$

Clearly, there are $\log(T)^2 = K_T^2$ different clusters.[57]

We now make a similar argument to the one we made in the marginal density case. Again, we must show that the Kullback-Leibler divergence between the joint density is T times an average Kullback-Leibler divergence. The tricky issue is that we no longer have a product distribution. Instead, we must show that appropriately constructed conditional densities satisfy the necessary inequalities. We start by considering the Kullback-Leibler divergence between the joint distributions. We assumed that p_T is a hidden Markov model. That implies there exists a hidden state z_t such that (x_t, z_t) are jointly Markov. We use capital letters to refer to the entire processes, i.e. Δ_T^P is the vector of cluster identities with respect to P_T. Consider the supremum of the deviations in each period:

$$\sup_t \mathrm{D}_{\mathrm{KL}} \left(\int_{G_t} \phi\left(x_t \,|\, \delta_t^P\right) dG_t^P(\delta_t^P) \,\middle\|\, \int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q) \right). \tag{3.110}$$

[57] The number of clusters used here is of the same asymptotic order as in the prior. This bound may no longer be tight.

We can rewrite this as follows by the definition of filtration, since we can condition on only past events without loss of generality, where we G_M to refer to a Markov density:

$$= \sup_{\mathcal{F}_{t-1}^P, \mathcal{F}_{t-1}^Q} \sup_t \mathrm{D}_{\mathrm{KL}} \left(\int_{G_M} \phi\left(x_t \,\middle|\, \delta_t^P\right) dG_M^P\left(\delta_t^P \,\middle|\, \mathcal{F}_{t-1}^P\right) \middle|\middle| \int_{G_M} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_M^Q\left(\delta_t^Q \,\middle|\, \mathcal{F}_{t-1}^Q\right) \right). \tag{3.111}$$

The goal is to show that the integral of (3.111) with respect to P_T can be rewritten as the sum of the individual conditionals. We start by considering the Kullback-Leibler divergence between the joint distributions. We assumed p_T is a hidden Markov model. This implies there exists a hidden state z_t such that (x_t, z_t) are jointly Markov. We use capital letters to refer to the entire processes, i.e. Δ_T^P is the vector of cluster identities with respect to P_T. The Kullback-Leibler divergence is

$$\mathrm{D}_{\mathrm{KL}}\left(P_T \,||\, Q_T\right) = \mathrm{D}_{\mathrm{KL}} \left(\prod_{t=1}^T p_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^P\right) \middle|\middle| \prod_{t=1}^T q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^Q\right) \right). \tag{3.112}$$

Consider the Kullback-Leibler divergence. We can rewrite the density period-by-period in terms of the transitions, the hidden Markov assumption implies that the G_t from the are constant functions of $\mathcal{F}_{t-1}$. Since the filtrations are measurable with respect to $x_1, \ldots, x_{t-1}$:

$$\int_{\mathbb{R}^{T\times D}} \sup_t \log \frac{q_T(X)}{p_T(X)} dP_T = \int_{\mathbb{R}^{T\times D}} \left(\sup_{x_t, \mathcal{F}_{t-1}^P, \mathcal{F}_{t-1}^Q} \log \frac{\int_{G_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta_t^Q\right) dG_M^Q\left(\delta_t^Q \,\middle|\, \mathcal{F}_{t-1}^Q\right)}{\int_{G_M} \phi\left(x_t \,|\, x_{t-1}, \delta_t^P\right) dG_P\left(\delta_t \,\middle|\, \mathcal{F}_{t-1}^P\right)} \right) dP_T. \tag{3.113}$$

Clearly, the supremum with respect to x_t is greater than the average with respect to the x_t, and so this is

$$\geq \int_{\mathbb{R}^{T\times D}} \left(\sup_{\mathcal{F}_{t-1}^P, \mathcal{F}_{t-1}^Q} \log \frac{\int_{G_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta_t^Q\right) dG_M^Q\left(\delta_t^Q \,\middle|\, \mathcal{F}_{t-1}^Q\right)}{\int_{G_M} \phi\left(x_t \,|\, x_{t-1}, \delta_t^P\right) dG_P\left(\delta_t \,\middle|\, \mathcal{F}_{t-1}^P\right)} \right) dP_T. \tag{3.114}$$

Let $\mathcal{K}\left(dG^P(\Delta^P), dG^Q(\Delta^Q)\right)$ be a coupling between the joint distributions of Δ^P and Δ^Q. Note, this a coupling over the entire sequence of δ_t^P and δ_t^Q. Then we can rewrite above as

$$\int_{\mathbb{R}^{T\times D}} \int_{G^P\times G^Q} \left(\sup_{\mathcal{F}_{t-1}^P, \mathcal{F}_{t-1}^Q} \log \frac{\int_{G_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta_t^Q\right) dG_M^Q\left(\delta_t^Q \,\middle|\, \mathcal{F}_{t-1}^Q\right)}{\int_{G_M} \phi\left(x_t \,|\, x_{t-1}, \delta_t^P\right) dG_P\left(\delta_t \,\middle|\, \mathcal{F}_{t-1}^P\right)} \right) d\mathcal{K}\left(dG^P(\Delta^P), dG^Q(\Delta^Q)\right) dP_T. \tag{3.115}$$

Conditional on both Δ^P and Δ^Q, $\mathcal{F}_{t-1}^P$ and $\mathcal{F}_{t-1}^Q$ contain no information regarding δ_t^P and δ_t^Q. By the law of iterated expectations, we can rewrite this as integral with respect to the joint distribution as we did above. The reason that we can pull the logarithm through integral is because conditional on δ_t^Q, and δ_t^Q, the integral contains only a single element. We then factor $\mathcal{K}$:

$$= \int_{\mathbb{R}^{T\times D}} \sup_{\mathcal{F}_{t-1}^Q, \mathcal{F}_{t-1}^P} \log \left(\prod_{t=1}^{T} \int_{G_t^P\times G_t^Q} \frac{\phi\left(x_t \,\middle|\, x_{t-1}, \delta_t^Q\right)}{\phi\left(x_t \,|\, x_{t-1}, \delta_t^P\right)} d\mathcal{K}\left(dG_t^P(\delta_t^P), dG_t^Q(\delta_t^Q) \,\middle|\, \mathcal{F}_{t-1}^Q, \mathcal{F}_{t-1}^P\right) \right) dP_T. \tag{3.116}$$

The Markov assumption on x_t, z_t implies that the δ_t^P and δ_t^Q will be hidden Markov as well. In addition since the δ_t are almost surely discrete, we can assume without loss of generality that the hidden state that makes x_t be a hidden Markov is almost surely discrete.

$$\int_{\mathbb{R}^{T\times D}} \sup_{\mathcal{F}_{t-1}^Q, \mathcal{F}_{t-1}^P} \sum_{t=1}^{T} \log \left(\int_{G_M^P\times G_M^Q} \frac{\phi\left(x_t \,\middle|\, x_{t-1}, \delta_t^Q\right)}{\phi\left(x_t \,|\, x_{t-1}, \delta_t^P\right)} d\mathcal{K}\left(dG_M^P(\delta_t^P), dG_M^Q(\delta_t^Q) \,\middle|\, \delta_{t-1}^Q, \delta_{t-1}^P, x_{t-1}, z_{t-1}\right) \right) dP_T. \tag{3.117}$$

We apply Jensen's inequality to the logarithm, and pull the supremum through the sum because it does not depend upon t:

$$\geq \int_{\mathbb{R}^{T\times D}} \sum_{t=1}^{T} \sup_{\mathcal{F}_{t-1}^Q, \mathcal{F}_{t-1}^P} \left(\int_{G_M^P\times G_M^Q} \log \frac{\phi\left(x_t \,\middle|\, x_{t-1}, \delta_t^Q\right)}{\phi\left(x_t \,|\, x_{t-1}, \delta_t^P\right)} d\mathcal{K}\left(dG_M^P(\delta_t^P), dG_M^Q(\delta_t^Q) \,\middle|\, \delta_{t-1}^Q, \delta_{t-1}^P, x_{t-1}, z_{t-1}\right) \right) dP_T. \tag{3.118}$$

The terms inside the sum are all the same after interchanging the sum and the integral:

$$= T \int_{\mathbb{R}^{T\times D}} \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \left(\int_{G^P_M \times G^Q_M} \log \frac{\phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right)}{\phi\left(x_t \mid x_{t-1}, \delta^P_t\right)} \, d\mathcal{K}\left(dG^P_M(\delta^P_t), dG^Q_M(\delta^Q_t) \,\middle|\, \delta^Q_{t-1}, \delta^P_{t-1}, x_{t-1}, z_{t-1} \right) \right) dP_T. \quad (3.119)$$

since couplings preserve marginals, and the δ_t are almost surely discrete, by the law of iterated expectations

$$= T \int_{\mathbb{R}^{T\times D}} \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \log \frac{\int_{G^Q_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right) dG^Q_M\left(\delta^Q_t \,\middle|\, \mathcal{F}^Q_{t-1}\right)}{\int_{G^P_M} \phi\left(x_t \mid x_{t-1}, \delta^P_t\right) dG^P_M\left(\delta^P_t \,\middle|\, \mathcal{F}^P_{t-1}\right)} \, dP_T. \quad (3.120)$$

We can factor P_T, and pull the supremum outside of the expectation, because we have finitely many terms. Given $\mathcal{F}^P_{t-1}$ and $\mathcal{F}^Q_{t-1}$, the only place where the two terms share a value is x_t. The other terms integrated to one.

$$= T \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \int_{\mathbb{R}^D} \cdots \int_{R^D} \log \frac{\int_{G^Q_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right) dG^Q_M\left(\delta^Q_t \,\middle|\, \mathcal{F}^Q_{t-1}\right)}{\int_{G^P_M} \phi\left(x_t \mid x_{t-1}, \delta^P_t\right) dG^P_M\left(\delta^P_t \,\middle|\, \mathcal{F}^P_{t-1}\right)} \prod_{t=1}^{T} p_T\left(x_t \,\middle|\, \mathcal{F}^P_{t-1}\right) dx_t \quad (3.121)$$

$$= T \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \int_{\mathbb{R}^D} \log \frac{\int_{G^Q_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right) dG^Q_M\left(\delta^Q_t \,\middle|\, \mathcal{F}^Q_{t-1}\right)}{\int_{G^P_M} \phi\left(x_t \mid x_{t-1}, \delta^P_t\right) dG^P_M\left(\delta^P_t \,\middle|\, \mathcal{F}^P_{t-1}\right)} p_T\left(x_t \,\middle|\, \mathcal{F}^P_{t-1}\right) dx_t \quad (3.122)$$

This is the formula for the Kullback-Leibler divergence between the conditional expectations.

$$= T \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \mathrm{D}_{\mathrm{KL}}\left(q_T\left(x_t \,\middle|\, \mathcal{F}^Q_{t-1}\right) \,\middle\|\, p_T\left(x_t \,\middle|\, \mathcal{F}^P_{t-1}\right) \right) \quad (3.123)$$

By Lemma 3.19 the supremum of the Kullback-Leibler divergences is proportional to squared Hellinger distance. By Proposition 3.18 the initial equation is bounded above by $CT\epsilon^2$: (The T comes from using non-rescaled data.)

$$\sup_t h^2\left(q_T\left(x_t \,\middle|\, \mathcal{F}^Q_{t-1}\right), p_T\left(x_t \,\middle|\, \mathcal{F}^P_{t-1}\right) \right) \leq \sup_t \frac{1}{T} h^2(q_T(X), p_T(X)) \leq C\frac{T}{T}\epsilon^2 = C\epsilon^2. \quad (3.124)$$

□

Lemma 3.5 (Replacing Θ_T with a Dirichlet Process)**.** *Let Q be a mixture distribution representable as an integral with respect to the Θ_T process defined in Definition 3.2. Then Q has a mixture representation as an integral with respect to the Dirichlet process.*

Proof. We can represent a Dirichlet process as $\Pr(x) = \sum_{i=1}^{\infty} \beta_i \delta_{x_i}(x)$, where δ_{x_i} is a indicator function with $\delta_{x_i}(x_i) = 1$, and the β_i satisfy a stick-breaking process. In other words, $\beta_i = \beta'_i \prod_{j=1}^{i=1}(1-\beta'_j)$ with $\beta'_j \sim \text{Beta}(1, \alpha)$ for some positive scalar α. Consider the probability mass function of a row of Θ_T, θ_t. Then $\Pr(|i| = 1) = b \prod_{j=1}^{j-1}(1-b)$. Since draws from the beta distribution lie in $(0, 1)$ with probability 1, these two stick-breaking processes are clearly mutually absolutely continuous. If we take $x \in \{-1, 1\}$ with the probability $1/2$ each as the Dirichlet base measure, the process are mutually absolutely continuous after possibly extending the space so that the Beta random variables are well-defined.

Because these two processes are mutually absolutely continuous, a Radon-Nikodym derivative exists because both measures are σ-finite. Since the rows are independent, and Dirichlet processes are normalized random measures, (Lin, Grimson, and Fisher 2010), we can extend this to the entire Θ_T process. Consequently, any process that is representable as an integral with respect to Θ_T can be represented as an integral with respect to to a Dirichlet process. □

3.C Contraction Rates

Constructing Exponentially Consistent Tests with Respect to h_∞

Lemma 3.7 (Exponentially consistent tests exist with respect to h_∞)**.** *There exist tests Υ_T and universal constants $C_2 > 0$, $C_3 > 0$ satisfying for every $\epsilon > 0$ and each $\xi_1 \in \Xi$ and true parameter ξ^P with $h_\infty(\xi_1, \xi^P)$:*

1. $$P_T\left(\Upsilon_T \,\middle|\, \xi^P\right) \leq \exp(-C_2 T\epsilon^2) \tag{3.19}$$

2. $$\sup_{\xi \in \Xi,\, e_n(\xi_1,\xi) < \epsilon C_3} P_T\left(1 - \Upsilon_T \,\middle|\, \xi^P\right) \leq \exp(-C_2 T\epsilon^2) \tag{3.20}$$

Proof. As done in the proof of the representing the Markov data, we can represent the joint density as a product density conditionally on a sequence of latent mixing measures G_t:

$$f\left(X_T \,|\, G_1, \ldots G_T\right) = \prod_{t=1}^{T} \int_{G_t^f} \phi\left(x_t \,\middle|\, \delta_t^f\right) dG_t^f(\delta_t^f). \tag{3.125}$$

Since we are letting G_t differ every period, we can do this for both Q_T and P_T. We can define a distance between these conditional densities as the sum of the squared Hellinger distances between each period. This is not the same as the Hellinger distance between the joint measures:

$$\begin{aligned} &h^2_{\text{avg}}\left(f\left(X \,\middle|\, \{G_t^f\}\right), g\left(X \,|\, \{G_t^g\}\right)\right) \\ &\quad := \frac{1}{T}\sum_{t=1}^{T} h^2\left(\int_{G_t^f} \phi\left(x_t \,\middle|\, \delta_t^f\right) dG_t^f(\delta_t^f), \int_{G_t^g} \phi\left(x_t \,|\, \delta_t^g\right) dG_t^g(\delta_t^g)\right). \end{aligned} \tag{3.126}$$

Then by (Birgé 2013, Corollary 2), there exists a test ϕ_T that satisfies the following:[58]

$$\begin{aligned} &\Pr_T\left(\phi_T(X) \,\middle|\, \left\{G_t^f, G_t^g\right\}\right) \\ &\quad \leq \exp\left(-\frac{1}{3} T h^2_{\text{avg}}\left(\int_{G_t^f} \phi\left(x_t \,\middle|\, \delta_t^f\right) dG_t^f(\delta_t^f), \int_{G_t^g} \phi\left(x_t \,|\, \delta_t^g\right) dG_t^g(\delta_t^g)\right)\right) \end{aligned} \tag{3.127}$$

[58] To map his notation into ours, take his $z = 0$, and take his measure R equal to P. Equation (3.127) is obvious then, and (3.128) follows by taking the exponential of both sides in the inequality inside the probability and rearranging.

and

$$\Pr_T\left(1-\phi_T(X)\,\middle|\,\left\{G_t^f,G_t^g\right\}\right) \tag{3.128}$$
$$\leq \exp\left(-\frac{1}{3}Th_{\text{avg}}^2\left(\int_{G_t^f}\phi\left(x_t\,\middle|\,\delta_t^f\right)dG_t^f(\delta_t^f),\int_{G_t^g}\phi\left(x_t\,|\,\delta_t^g\right)dG_t^g(\delta_t^g)\right)\right).$$

The issue with these equations is that they are not in terms of h_∞ and only hold conditionally. The reason that we can get around this is because they hold for all G_t^f and for all G_t^g. Consequently, we can take the infimum of both sides, and bound the right-hand side of both equations by

$$\frac{T}{3}\sup_{\{(G_t^f,G_t^g)\}}h_{\text{avg}}^2\left(\int_{G_t^f}\phi\left(x_t\,\middle|\,\delta_t^f\right)dG_t^f(\delta_t^f),\int_{G_t^g}\phi\left(x_t\,|\,\delta_t^g\right)dG_t^g(\delta_t^g)\right) \tag{3.129}$$

for any length T sequence. This equals the least favorable G_t^f and G_t^g repeated T times. This joint distribution exists in our set because we are not placing any restrictions on the dynamics besides ergodicity, and stationary distribution are clearly ergodic. Hence, this equals

$$=\frac{T}{3}\frac{1}{T}\sum_{t=1}^{T}h^2\left(\int_{G_{sup}^f}\phi\left(x_t\,\middle|\,\delta_t^f\right)dG_{sup}^f(\delta_t^f),\int_{G_{sup}^g}\phi\left(x_t\,|\,\delta_t^g\right)dG_{sup}^g(\delta_t^g)\right). \tag{3.130}$$

The terms inside the sum are all the same:

$$=\frac{T}{3}h^2\left(\int_{G_{sup}^f}\phi\left(x_t\,\middle|\,\delta_t^f\right)dG_{sup}^f(\delta_t^f),\int_{G_{sup}^g}\phi\left(x_t\,|\,\delta_t^g\right)dG_{sup}^g(\delta_t^g)\right) \tag{3.131}$$
$$=\frac{T}{3}\sup_{(G_t^f,G_t^g)}h^2\left(\int_{G_t^f}\phi\left(x_t\,\middle|\,\delta_t^f\right)dG_t^f(\delta_t^f),\int_{G_t^g}\phi\left(x_t\,|\,\delta_t^g\right)dG_t^g(\delta_t^g)\right) \tag{3.132}$$
$$=\frac{T}{3}h_\infty^2\left(\int_{G_t^f}\phi\left(x_t\,\middle|\,\delta_t^f\right)dG_t^f(\delta_t^f),\int_{G_t^g}\phi\left(x_t\,|\,\delta_t^g\right)dG_t^g(\delta_t^g)\right). \tag{3.133}$$

Taking the supremum over G_t^f and G_t^g is equivalent to taking supremum over $\mathcal{F}_{t-1}^f$ and $\mathcal{F}_{t-1}^g$ because the G_t^f and G_t^g are measurable functions of the later, and we are taking the supremum outside of the integral. They both span the same information sets:

$$=\frac{T}{3}h_\infty^2\left(\int_{G_t^f}\phi\left(x_t\,\middle|\,\delta_t^f\right)dG_t^f(\delta_t^f),\int_{G_t^g}\phi\left(x_t\,|\,\delta_t^g\right)dG_t^g(\delta_t^g)\right). \tag{3.134}$$

Since we can bound the error probabilities in both directions, using exponentially consistent tests, we have shown both items in Lemma 3.7 hold. □

Bounding the Posterior Divergence

Proposition 3.8 (Bounding the Posterior Divergence)**.** *Let P_T be a uniformly ergodic Hidden Markov Gaussian process, i.e., $p_T \coloneqq \sum_k \Pi_{k,t}\phi(x_t \,|\, \mu_t, \Sigma_t)$ with finite means and finite positive-definite covariances. Let $\Xi_T \subset \Xi$ and $T \to \infty$. Let Q_T be a mixture approximation with $\frac{K_T^i}{\eta_T}$ components. Assume the following condition holds with probability $1 - 2\delta$ for $\delta > 0$ and constants C and $i \in \mathbb{N}$:*

$$\sup_t h\left(q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^Q\right), p_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^P\right)\right) < C\eta_T. \tag{3.21}$$

Let $\epsilon_{i,T} \coloneqq \frac{\log(T)^{\sqrt{i}}}{\sqrt{T}}$. Then the following two conditions hold with probability $1 - 2\delta$ with respect to the prior

$$\sup_{\epsilon_i \geq \epsilon_{T,i}} \log N\left((\epsilon_i, \left\{\xi \in \Xi_T \,\middle|\, h_\infty(\xi, \xi^P) \leq \epsilon_i\right\}, h_\infty\right) \leq T\epsilon_{T,i}^2, \tag{3.22}$$

and

$$\mathcal{Q}_T\left(B_T\left(\xi^P, \epsilon_{T,i}, 2\right) \,\middle|\, X_T\right) \geq C\exp\left(-C_0 T\epsilon_{T,i}^2\right). \tag{3.23}$$

Proof. We are looking at locally asymptotically normal models, as discussed in Lemma 3.19, and we bind the Hellinger distance and Kullback-Leibler divergence in terms of $(x_t - \mu_t)'\Sigma_t^{-1}(x_t - \mu_t)$. In addition, the supremum of the deviations is clearly greater than the average of the deviations, and so the h_∞-norm forms smaller balls than both $\mathrm{D}_{\mathrm{KL}}(f \,||\, g)$ and $V_{k,0}$. Consequently, we can replace $B_T(\xi_0, \epsilon_T, 2)$ with $\{\xi \in \Xi \,|\, h_\infty^2(\xi, \xi_0) < T\epsilon_T^2\}$. We use 2 as the last argument of B because we are using $V_{2,0}$, i.e., effectively the 2nd-moment of the Kullback-Leibler divergence.

To prove the result we need to find a sequence $\epsilon_{T,i} \to 0$ that satisfies the following two conditions:

$$\sup_{\epsilon_i > \epsilon_{T,i}} \log N\left(\epsilon_i, \left\{\xi \in \Xi_T \,\middle|\, h_\infty(\xi, \xi_0) \leq \epsilon_i\right\}, h_\infty\right) \leq T\epsilon_{T,i}^2 \tag{3.135}$$

and

$$\mathcal{Q}_T\left(\left\{\xi \in \Xi \,\middle|\, h_\infty^2(\xi, \xi_0) < \epsilon_{T,i}\right\}\right) \geq C\exp\left(-T\epsilon_{T,i}^2\right). \tag{3.136}$$

These two conditions work in opposite directions. The first criterion is easier to satisfy the larger $\epsilon_{T,i}$ is, but to achieve a fast rate of convergence we want a small $\epsilon_{T,i}$ in the second condition.

By assumption, there exists a covering with $\frac{K_T^i}{\eta_T}$ components such that the following holds:

$$\sup_t h\left(q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^Q\right), p_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^P\right)\right) < C\sqrt{1+\log\frac{1}{\delta}}\eta_T. \tag{3.137}$$

Equation (3.136) is satisfied if

$$\eta_T^2 \geq \frac{C_0}{1+\log(1/\delta)}\exp\left(-T\epsilon_{T,i}^2\right) \propto \exp\left(-T\frac{\log(T)^i}{T}\right) = \frac{1}{T^i}. \tag{3.138}$$

To satisfy (3.135), h_∞^2 must be bounded below and decline exponentially fast. The expressions above hold for any $\eta_T^* \geq \eta_T$. Let $\eta_T^* = \frac{\log(T)^n}{T^n}$. We know there exists a covering with $K_T = \frac{\log(T)^i}{\eta_T^*}$ components. This implies that

$$K_T = \frac{\log(T)^i}{\eta_T^*} = \frac{\log(T)^i}{\log(T)^i/T^i} = T^i. \tag{3.139}$$

This K_T is proportional to the number of terms we are using, and the bracketing number is proportional to the covering number:

$$N\left(\epsilon_n, \left\{\xi \in \Xi \,\middle|\, h_\infty^2(\xi,\xi_0)\right\} \leq \epsilon_i, h_\infty^2\right) \leq T^i = \exp\left(\log\left(T^i\right)\right) = \exp\left(T\epsilon_{T,i}^2\right). \tag{3.140}$$

Taking logarithms of both sides of (3.140) finishes the proof.

□

Contraction Rate of the Marginal Density

Theorem 3.10 (Contraction Rate of the Marginal Density)**.** *Let P_T be a uniformly ergodic Hidden Markov Gaussian process, i.e., $p_T := \sum_k \pi_k \phi(\cdot \,|\, \mu_t, \Sigma_t)$ with finite mean and finite variance. Let $T \to \infty$, then the following holds with $\epsilon_T = \sqrt{\frac{\log(T)}{T}}$ and probability $1-2\delta$ with respect to the prior. There exists a constant C independent of T such that the posterior over the transition densities constructed above and the true transition density satisfies*

$$P_T\left(\mathcal{Q}_T\left(h\left(p_T\left(x_t\right), q_T\left(x_t\right)\right) \geq C\epsilon_T \,|\, X\right)\right) \to 0.$$

Proof. To prove this result, note that the existence of exponentially consistent tests with respect to the average Hellinger metric for independent data is well-known (Ghosal and van der Vaart 2017, 540). We can represent the density as product density by a resampling argument as we did in the construction of the sieve.

Having done that we can verify the conditions in Proposition 3.8. If we take $i = 1$ in (3.21), Theorem 3.3 implies the necessary bound on the sieve complexity exists. In addition, since h_∞ is bounded above by the Hellinger distance, h, the conclusions of Proposition 3.8 trivially go through in this weaker topology.

This verifies the three conditions in Theorem 3.6 on a set with with probability $1 - 2\delta$ with respect to the prior. This then gives us the posterior contraction rate $\epsilon_T = \sqrt{\frac{\log T}{T}}$.

□

Contraction Rate of the Transition Density

Theorem 3.9 (Contraction Rate of the Transition Density)**.** *Let P_T be a uniformly ergodic Hidden Markov Gaussian process, i.e., $p_T \coloneqq \sum_k \Pi_{t,k} \phi(x_t \,|\, \mu_t, \Sigma_t)$ with finite means and finite positive-definite covariances. Let $T \to \infty$, then the following holds with $\epsilon_T \coloneqq \sqrt{\frac{\log(T)^2}{T}}$ with probability $1 - 2\delta$ with respect to the prior. There exists a constant C independent of T such that the posterior over the transition densities constructed above and the true transition density satisfies*

$$P_T \left(\mathcal{Q}_T \left(\sup_{\mathcal{F}^P_{t-1}, \mathcal{F}^Q_{t-1}} h \left(p_T \left(x_t \,|\, \mathcal{F}^P_{t-1} \right), q_T \left(x_t \,\middle|\, \mathcal{F}^Q_{t-1} \right) \right) \geq C \epsilon_T \,\middle|\, X_T \right) \right) \to 0.$$

Proof. The proof of this is essentially identical to the marginal density case, mutatis mutandis. Lemma 3.7 implies the that h_∞ has the required exponentially consistent tests. We verify the conditions in Proposition 3.8. If we take $i = 2$ in (3.21), Theorem 3.4 implies the necessary bound on the sieve complexity exists.

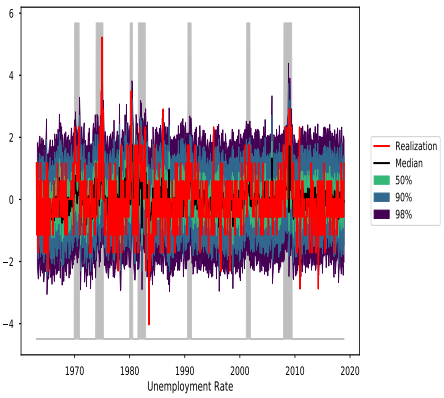
This verifies the three conditions in Theorem 3.6 on a set with with probability $1 - 2\delta$ with respect to the prior. This then gives us the posterior contraction rate $\epsilon_T = \sqrt{\frac{\log(T)^2}{T}}$.

□

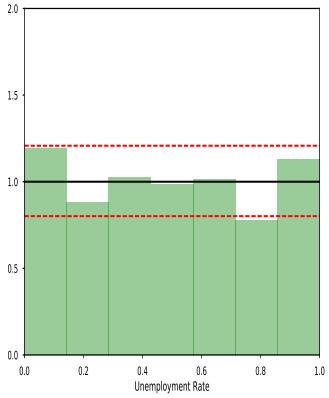
3.D Macroeconomic Empirical Results

Figure 3.7: Unemployment Rate

(a) Posterior Density



(b) PIT Histogram



(c) PIT ACF

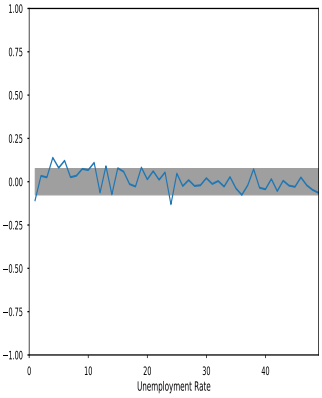
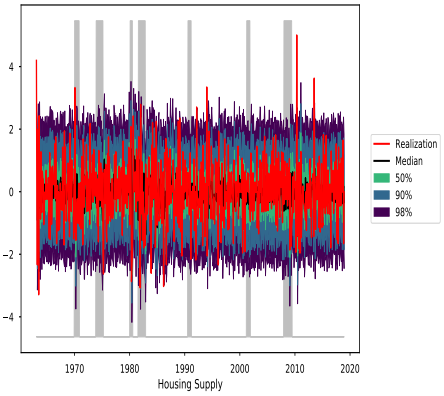
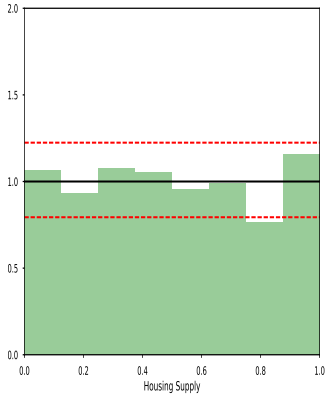


Figure 3.8: Housing Supply

(a) Posterior Density



(b) PIT Histogram



(c) PIT ACF

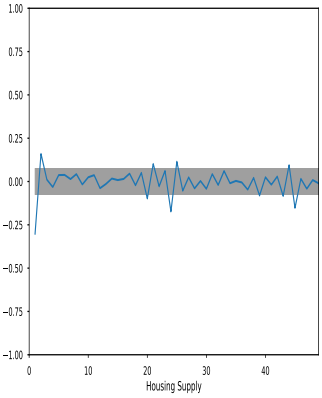
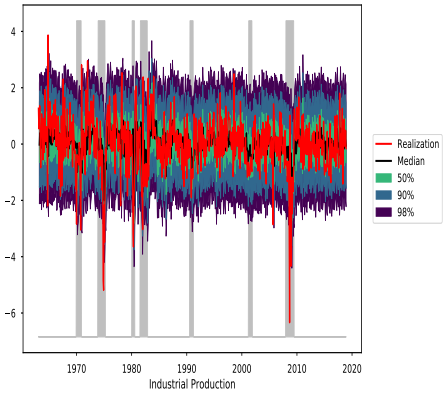
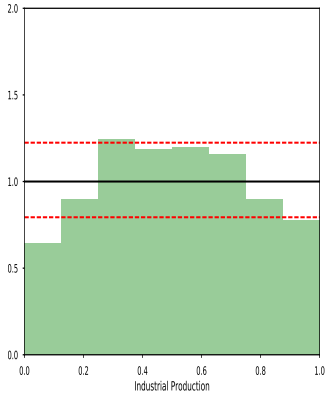


Figure 3.9: Industrial Production

(a) Posterior Density



(b) PIT Histogram



(c) PIT ACF

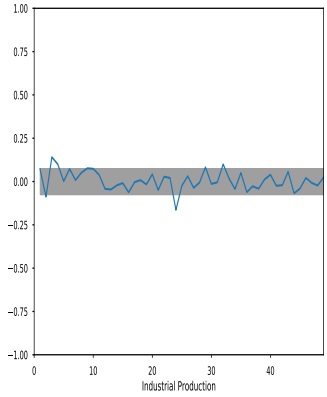
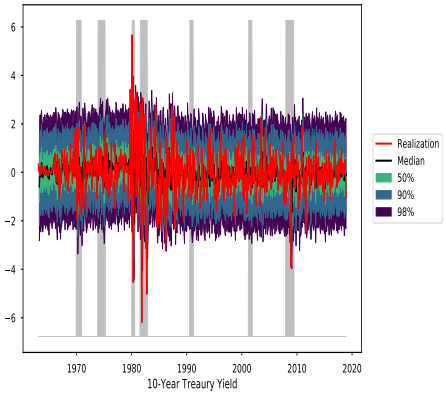
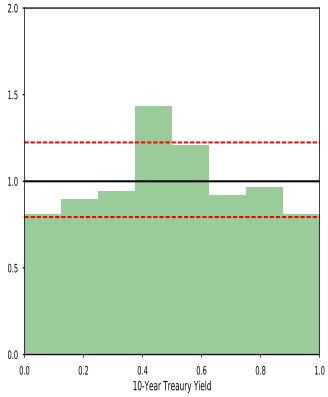


Figure 3.10: Long-Term Interest Rate

(a) Posterior Density



(b) PIT Histogram



(c) PIT ACF

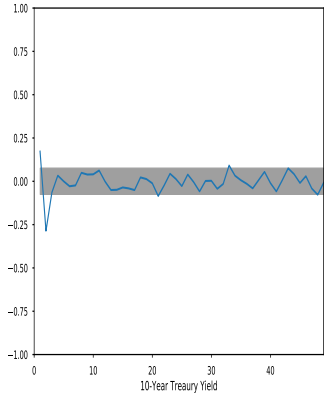


Figure 3.11: Money Supply (M2)

(a) Posterior Density (b) PIT Histogram (c) PIT ACF

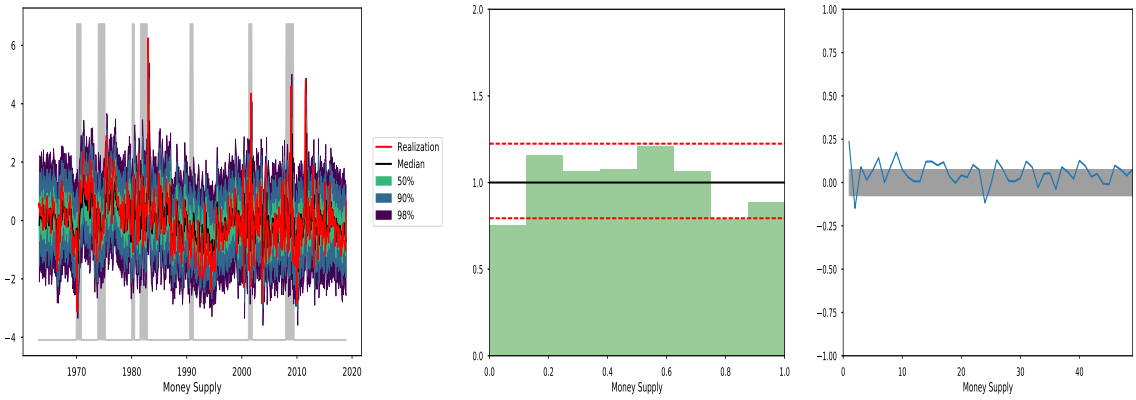
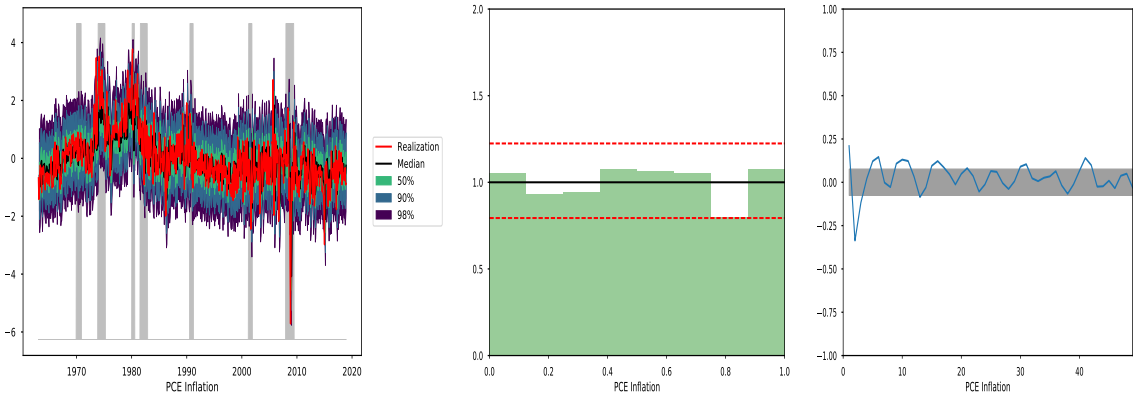


Figure 3.12: Personal Consumption Expenditures (PCE) Inflation

(a) Posterior Density (b) PIT Histogram (c) PIT ACF



3.E Posterior Derivations

Component Coefficient Posterior

Let X_k be the $T_k \times N$ vector and Y_k be the $T_k \times D$ vector of data in component K. This implies that Σ_k is a $D \times D$ matrix and β_k is an $N \times D$ matrix.[59] Meanwhile, V is a $D \times D$ matrix and U is a $N \times N$ matrix.

The joint density is

$$\Pr\left(Y_k, \beta_k, \Sigma_k \,|\, X_k\right) = \exp\left(-\frac{1}{2}\operatorname{tr}\{V^{-1}\left(\beta_k - \bar{\beta}\right)' U^{-1}\left(\beta_k - \bar{\beta}\right)\}\right) \exp\left(-\frac{1}{2}\operatorname{tr}\{\left(Y_k - X_k\beta_k\right)\Sigma_k^{-1}\left(Y_k - X_k\beta_k\right)'\}\right)$$
$$\frac{|\Sigma_k|^{-T_k/2}}{(2\pi)^{T_k/2}} \frac{1}{\sqrt{(2\pi)^{ND}|V|^N|U|^D}} \frac{|(\mu_1 - 2)\Omega|^{\nu/2}}{\sqrt{2^{\nu D}}\Gamma_D(\frac{\nu}{2})} |\Sigma_k|^{-\frac{\nu+D+1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\{(\mu_1 - 2)\Omega\Sigma_k^{-1}\}\right) \tag{3.141}$$

By the additivity and circular commutativity of the trace, and associativity of matrix multiplication:

$$\propto |\Sigma_k|^{-\frac{\nu+D+T+1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\{V^{-1}\left(\beta_k - \bar{\beta}\right)' U^{-1}\left(\beta_k - \bar{\beta}\right)\}\right) \exp\left(-\frac{1}{2}\operatorname{tr}\{\left(\left(Y_k - X_k\beta_k\right)'\left(Y_k - X_k\beta_k\right) + (\mu_1 - 2)\Omega\right)\Sigma_k^{-1}\}\right). \tag{3.142}$$

Combining the two kernels of β_k and expanding gives

$$\propto |\Sigma_k|^{-\frac{\nu+D+T+1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\{V^{-1}\left(\left(\beta_k - \bar{\beta}\right)' U^{-1}\left(\beta_k - \bar{\beta}\right)\right) + \left(\left(Y_k - X_k\beta_k\right)'\left(Y_k - X_k\beta_k\right) + (\mu_1 - 2)\Omega\right)\Sigma_k^{-1}\}\right) \tag{3.143}$$
$$= |\Sigma_k|^{-\frac{\nu+D+T+1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\{V^{-1}\left(\beta_k' U^{-1}\beta_k - 2\beta_k' U^{-1}\bar{\beta} + \bar{\beta}' U^{-1}\bar{\beta}\right) + \Sigma_k^{-1}\left(Y_k' Y_k - 2\beta_k' X_k' Y_k + \beta_k' X_k' X_k\beta_k + (\mu_1 - 2)\Omega\right)\}\right). \tag{3.144}$$

[59] The likelihood in (3.141) is correct because the trace is the sum of the diagonal elements.

Isolating the terms that have a β in them:

$$= \exp\left(-\frac{1}{2}\operatorname{tr}\{V^{-1}\left(-2\beta_k' U^{-1}\bar{\beta} + \beta_k' U^{-1}\beta_k\right) + \Sigma_k^{-1}\left(-2\beta_k' X_k' Y_k + \beta_k' X_k' X_k \beta_k\right) + V^{-1}\bar{\beta}' U^{-1}\bar{\beta} + \Sigma_k^{-1}\left(Y_k' Y_k + (\mu_1 - 2)\Omega\right)\}\right). \\ |\Sigma_k|^{-\frac{\nu+D+T+1}{2}} \tag{3.145}$$

Rewriting the traces in terms of the vectorization operator:

$$= \exp\left(-\frac{1}{2}\left(\operatorname{tr}\{V^{-1}(-2\beta_k' U^{-1}\bar{\beta})\} + \operatorname{vec}\{\beta_k\}' \operatorname{vec}\left\{U^{-1}\beta_k V^{-1}\right\} \operatorname{tr}\{\Sigma_k^{-1}(-2\beta_k' X_k' Y_k)\} + \operatorname{vec}\{\beta_k\}' \operatorname{vec}\left\{X_k' X_k \beta_k \Sigma_k^{-1}\right\}\right)\right) \\ \exp\left(-\frac{1}{2}\operatorname{tr}\{V^{-1}\bar{\beta}' U^{-1}\bar{\beta} + \Sigma_k^{-1}\left(Y_k' Y_k + (\mu_1 - 2)\Omega\right)\}\right) |\Sigma_k|^{-\frac{\nu+D+T+1}{2}}.$$

Exploiting the relationship between vectorization and the Kronecker product and then combining squared terms:

$$\propto \exp\left(\operatorname{tr}\{\beta_k'\left(U^{-1}\bar{\beta}V^{-1} + X_k' Y_k \Sigma_k^{-1}\right)\} - \frac{1}{2}\operatorname{tr}\{\left(\left(V^{-1}\otimes U^{-1}\right) + \left(\Sigma_k^{-1}\otimes X_k' X_k\right)\right)\operatorname{vec}\{\beta_k\}\operatorname{vec}\{\beta_k\}'\}\right) \\ \exp\left(-\frac{1}{2}\operatorname{tr}\{V^{-1}\bar{\beta}' U^{-1}\bar{\beta} + \Sigma_k^{-1}\left(Y_k' Y_k + (\mu_1 - 2)\Omega\right)\}\right) |\Sigma_k|^{-\frac{\nu+D+T+1}{2}}. \tag{3.146}$$

If we assume that $V = \Sigma_k$, we can simplify this as

$$= \exp\left(\operatorname{tr}\{\beta_k'\left(U^{-1}\bar{\beta} + X_k' Y_k\right)\Sigma_k^{-1}\} - \frac{1}{2}\operatorname{tr}\{\left(\Sigma_k^{-1}\otimes\left(U^{-1} + X_k' X_k\right)\right)\operatorname{vec}\{\beta_k\}\operatorname{vec}\{\beta_k\}'\}\right) \tag{3.147}$$

$$\exp\left(-\frac{1}{2}\operatorname{tr}\{\Sigma_k^{-1}\left(\bar{\beta}' U^{-1}\bar{\beta} + Y_k' Y_k + (\mu_1 - 2)\Omega\right)\}\right) |\Sigma_k|^{-\frac{\nu+D+T+1}{2}}$$

$$= \exp\left(\operatorname{vec}\{\beta_k\}' \operatorname{vec}\left\{\left(U^{-1}\bar{\beta} + X_k' Y_k\right)\Sigma_k^{-1}\right\} - \frac{1}{2}\operatorname{vec}\{\beta_k\}'\left(\Sigma_k^{-1}\otimes\left(U^{-1} + X_k' X_k\right)\right)\operatorname{vec}\{\beta_k\}\right)$$

$$\exp\left(-\frac{1}{2}\operatorname{tr}\{\Sigma_k^{-1}\left(\bar{\beta}' U^{-1}\bar{\beta} + Y_k' Y_k + (\mu_1 - 2)\Omega\right)\}\right) |\Sigma_k|^{-\frac{\nu+D+T+1}{2}}. \tag{3.148}$$

We now use the multivariate completion of squares: $u'Au - 2\alpha'u = (u - A^{-1}\alpha)'A(u - A^{-1}\alpha) - \alpha'A^{-1}\alpha$. Let $Z_k \coloneqq (U^{-1}\bar{\beta} + X_k'Y_k)$ and $W_k \coloneqq (U^{-1} + X_k'X_k)$. We can now rewrite (3.148) as

$$= \exp\left(-\frac{1}{2}\left(\text{vec}\{\beta_k\} - (\Sigma_k^{-1} \otimes W_k)^{-1} Z_k \Sigma_k^{-1}\right)' (\Sigma_k^{-1} \otimes W_k) \left(\text{vec}\{\beta_k\} - (\Sigma_k^{-1} \otimes W_k)^{-1} Z_k \Sigma_k^{-1}\right) \right)$$
$$\exp\left(\frac{1}{2}\Sigma_k^{-1} Z_k' (\Sigma_k^{-1} \otimes W_k)^{-1} Z_k \Sigma_k^{-1}\right) \exp\left(-\frac{1}{2}\text{tr}\{\Sigma_k^{-1}\left(\bar{\beta}'U^{-1}\bar{\beta} + Y_k'Y_k + (\mu_1 - 2)\Omega\right)\}\right) |\Sigma_k|^{-\frac{\nu+D+T+1}{2}}. \tag{3.149}$$

I now eliminate all of the Kronecker products:

$$= \exp\left(-\frac{1}{2}\text{vec}\left\{\beta_k - W_k^{-1}Z_k\right\}' \text{vec}\left\{W_k\left(\beta_k - W_k^{-1}Z_k\right)\Sigma_k^{-1}\right\}\right) \tag{3.150}$$
$$\exp\left(\frac{1}{2}\text{vec}\left\{(U^{-1}\bar{\beta} + X_k'Y_k)\Sigma_k^{-1}\right\}' \text{vec}\left\{W_k^{-1}Z_k\right\} - \frac{1}{2}\text{tr}\{\Sigma_k^{-1}\left(\bar{\beta}'U^{-1}\bar{\beta} + Y_k'Y_k + (\mu_1 - 2)\Omega\right)\}\right) |\Sigma_k|^{-\frac{\nu+D+T+1}{2}}. \tag{3.151}$$

We rewrite this in terms of the traces, reorder some of the terms, and substitute the definitions of Z_k and W_k back in:

$$= \exp\left(-\frac{1}{2}\text{tr}\{\Sigma_k^{-1}\left(\beta_k - \left(U^{-1} + X_k'X_k\right)^{-1}\left(U^{-1}\bar{\beta} + X_k'Y_k\right)\right)' \left(U^{-1} + X_k'X_k\right)\left(\beta_k - \left(U^{-1} + X_k'X_k\right)^{-1}\left(U^{-1}\bar{\beta} + X_k'Y_k\right)\right)\}\right) \tag{3.152}$$
$$\exp\left(-\frac{1}{2}\text{tr}\{\Sigma_k^{-1}\left(\left(\bar{\beta}'U^{-1}\bar{\beta} + Y_k'Y_k + (\mu_1 - 2)\Omega\right) - \left(U^{-1}\bar{\beta} + X_k'Y_k\right)'\left(U^{-1} + X_k'X_k\right)^{-1}\left(U^{-1}\bar{\beta} + X_k'Y_k\right)\right)\}\right) |\Sigma_k|^{-\frac{\nu+D+T+1}{2}}.$$

The first expression is kernel of a matrix-normal distribution. The mean is $\left(U^{-1} + X_k'X_k\right)^{-1}\left(U^{-1}\bar{\beta} + X_k'Y_k\right)$, and the two covariance parameters are Σ_k, and $\left(U^{-1} + X_k'X_k\right)^{-1}$. The second expression is the kernel of a Inverse-Wishart distribution. Its scale parameter is $\left(\bar{\beta}'U^{-1}\bar{\beta} + Y_k'Y_k + (\mu_1 - 2)\Omega\right) - \left(U^{-1}\bar{\beta} + X_k'Y_k\right)'\left(U^{-1} + X_k'X_k\right)^{-1}\left(U^{-1}\bar{\beta} + X_k'Y_k\right)$. It has $\mu_1 + D - 1 + T_k$ degrees of freedom. To see the intuition behind this, note that if U^{-1} and Ω both equal zero, this equals $Y_k'Y_k - Y_k'X_k'(X_k'X_k')^{-1}X_kY_k$, i.e., the sum of squared residuals. Since the β_k parameter does not show up in the second expression, we can draw from the posterior by drawing the Σ_k from its marginal posterior, and then drawing from the posterior of β_k conditional on Σ_k .

Hypermean Posterior with Heteroskedastic Data

We now compute the posterior of the hierarchical mean for the coefficients conditional on the covariance matrices, $\{\Sigma_k\}_{k=1}^{K_T}$:

$$\Pr\left(\{\beta\}_{k=1}^{K}, \bar{\beta}, \{\Sigma\}_{k=1}^{K}\right) = \exp\left(-\frac{1}{2}\operatorname{tr}\{V^{-1}\left(\bar{\beta}-\beta^{\dagger}\right)'U^{-1}\left(\bar{\beta}-\beta^{\dagger}\right)\}\right)\exp\left(\sum_{k=1}^{K}-\frac{1}{2}\operatorname{tr}\{\Sigma_k^{-1}\left(\beta_k-\bar{\beta}\right)'U^{-1}\left(\beta_k-\bar{\beta}\right)\}\right)$$
$$\sqrt{(2\pi)^{ND}|U|^{D}}|U|^{-\frac{\nu_U+N+1}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}\{\Psi_U U^{-1}\}\right)\prod_{k=1}^{K}\frac{1}{\sqrt{(2\pi)^{ND}|\Sigma_k|^{N}|U|^{D}}} \tag{3.153}$$

Dropping all of the terms that contain neither $\bar{\beta}$ nor U:

$$\propto |U|^{-\frac{\nu_U+N+(K+1)D+1}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}\{V^{-1}\left(\bar{\beta}-\beta^{\dagger}\right)'U^{-1}\left(\bar{\beta}-\beta^{\dagger}\right)+\sum_{k=1}^{K}\Sigma_k^{-1}(\bar{\beta}-\beta_k)'U^{-1}(\bar{\beta}-\beta_k)\}\right)\exp\left(-\frac{1}{2}\operatorname{tr}\{\Psi_U U^{-1}\}\right). \tag{3.154}$$

Expanding out the terms and dropping terms that do not involve $\bar{\beta}$ or U:

$$\propto \exp\left(-\frac{1}{2}\operatorname{tr}\{V^{-1}\bar{\beta}'U^{-1}\bar{\beta}-2V^{-1}\beta^{\dagger\prime}U^{-1}\bar{\beta}+V^{-1}\beta^{\dagger\prime}U^{-1}\beta^{\dagger}+\sum_{k=1}^{K}\Sigma_k^{-1}(\bar{\beta}'U^{-1}\bar{\beta}-2\beta_k'U^{-1}\bar{\beta}+\beta_k'U^{-1}\beta_k\}\right) \tag{3.155}$$
$$|U|^{-\frac{\nu_U+N+(K+1)D+1}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}\{\Psi_U U^{-1}\}\right).$$

Exploiting properties of the trace and vectorization, where $B \coloneqq \operatorname{vec}\{\bar{\beta}\}$:

$$\propto \exp\left(-\frac{1}{2}\operatorname{vec}\{\beta^{\dagger}\}'\left(V^{-1}\otimes W^{-1}\right)B+\operatorname{vec}\left\{W^{-1}\beta^{\dagger\prime}V^{-1}\right\}'B-\frac{1}{2}\sum_{k=1}^{K}\operatorname{tr}\{(\Sigma_k^{-1}\otimes U^{-1})BB'\}+\operatorname{vec}\left\{\sum_{k=1}^{K}U^{-1}\beta_k\Sigma_k^{-1}\right\}'B\right)$$
$$|U|^{-\frac{\nu_U+N+(K+1)D+1}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}\{V^{-1}\beta^{\dagger\prime}U^{-1}\beta^{\dagger}+\sum_{k=1}^{K}\Sigma_k^{-1}\beta_k'U^{-1}\beta_k+\Psi_U U^{-1}\}\right). \tag{3.156}$$

We can simplify using the circular commutativity of the trace:

$$\propto \exp\left(-\frac{1}{2}\operatorname{vec}\{\bar{\beta}\}'\left(\left(\sum_{k=1}^{K}\Sigma_k^{-1}\right)\otimes U^{-1}+V^{-1}\otimes U^{-1}\right)\operatorname{vec}\{\bar{\beta}\}+\operatorname{vec}\left\{U^{-1}\beta^{\dagger}V^{-1}+\sum_{k=1}^{K}U^{-1}\beta_k\Sigma_k^{-1}\right\}'\operatorname{vec}\{\bar{\beta}\}\right) \tag{3.157}$$

$$|U|^{-\frac{\nu_U+N+(K+1)D+1}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}\{\beta^{\dagger}V^{-1}\beta^{\dagger\prime}U^{-1}+\sum_{k=1}^{K}\beta_k\Sigma_k^{-1}\beta_k'U^{-1}+\Psi_UU^{-1}\}\right).$$

Collecting terms:

$$\propto \exp\left(-\frac{1}{2}\operatorname{vec}\{\bar{\beta}\}'\left(\left(\sum_{k=1}^{K}\Sigma_k^{-1}+V^{-1}\right)\otimes U^{-1}\right)\operatorname{vec}\{\bar{\beta}\}+\operatorname{vec}\left\{U^{-1}\left(\beta^{\dagger}V^{-1}+\sum_{k=1}^{K}\beta_k\Sigma_k^{-1}\right)\right\}'\operatorname{vec}\{\bar{\beta}\}\right) \tag{3.158}$$

$$|U|^{-\frac{\nu_U+N+(K+1)D+1}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}\{\left(\beta^{\dagger}V^{-1}\beta^{\dagger\prime}+\sum_{k=1}^{K}\beta_k\Sigma_k^{-1}\beta_k'+\Psi_U\right)U^{-1}\}\right)$$

$$\propto \exp\left(-\frac{1}{2}\operatorname{tr}\{\left(\sum_{k=1}^{K}\Sigma_k^{-1}+V^{-1}\right)\bar{\beta}'U^{-1}\bar{\beta}+\left(\beta^{\dagger}V^{-1}+\sum_{k=1}^{K}\beta_k\Sigma_k^{-1}\right)'U^{-1}\bar{\beta}\}\right) \tag{3.159}$$

$$|U|^{-\frac{\nu_U+N+(K+1)D+1}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}\{\left(\beta^{\dagger}V^{-1}\beta^{\dagger\prime}+\sum_{k=1}^{K}\beta_k\Sigma_k^{-1}\beta_k'+\Psi_U\right)U^{-1}\}\right).$$

We now vectorize the first line of (3.159) after using the circular commutativity of the trace to the square term. We drop the second line for now to simplify the exposition. We will bring it back in later. This gives

$$\exp\left(-\frac{1}{2}\operatorname{vec}\{\bar{\beta}\}'\left(\left(\sum_{k=1}^{K}\Sigma_k^{-1}+V^{-1}\right)\otimes U^{-1}\right)\operatorname{vec}\{\bar{\beta}\}-2\operatorname{vec}\left\{U^{-1}\left(\beta^{\dagger}V^{-1}+\sum_{k=1}^{K}\beta_k\Sigma_k^{-1}\right)\right\}'\operatorname{vec}\{\bar{\beta}\}\right) \tag{3.160}$$

We then apply the multivariate equation of squares, and let $Z \coloneqq (\beta^\dagger V^{-1} + \sum_{k=1}^K \beta_k \Sigma_k^{-1})$ and $W \coloneqq (\sum_{k=1}^K \Sigma_k^{-1} + V^{-1})$:

$$= \exp\left(-\frac{1}{2}\left(\mathrm{vec}\{\bar{\beta}\} - \left(W \otimes U^{-1}\right)^{-1} \mathrm{vec}\{U^{-1}Z\}\right)\left(W \otimes U^{-1}\right)\left(\mathrm{vec}\{\bar{\beta}\} - \left(W \otimes U^{-1}\right)^{-1} \mathrm{vec}\{U^{-1}Z\}\right)\right)$$
$$\exp\left(\frac{1}{2}\mathrm{vec}\{U^{-1}Z\}'\left(Z \otimes U^{-1}\right)^{-1}\mathrm{vec}\{U^{-1}Z\}\right) \tag{3.161}$$

We can simplify the vectorization.

$$= \exp\left(-\frac{1}{2}\mathrm{vec}\{\bar{\beta} - ZW^{-1}\}\left(W \otimes U^{-1}\right)\mathrm{vec}\{\bar{\beta} - ZW^{-1}\}\right)\exp\left(\frac{1}{2}\mathrm{tr}\{U^{-1}ZW^{-1}Z'\}\right) \tag{3.162}$$

We can replace the vectorizations with traces.

$$= \exp\left(-\frac{1}{2}\mathrm{tr}\{U^{-1}\left(\bar{\beta} - ZW^{-1}\right)W\left(\bar{\beta} - ZW^{-1}\right)\}\right)\exp\left(\frac{1}{2}\mathrm{tr}\{U^{-1}ZW^{-1}Z'\}\right) \tag{3.163}$$

Equation (3.163) is the kernel of a matrix normal distribution given the covariance matrices. We substitute the definitions of W and Z back in. The row matrix covariance is U, the column posterior covariance is $(\sum_{k=1}^K \Sigma_k^{-1} + V^{-1})$, and the mean is $(\beta^\dagger V^{-1} + \sum_{k=1}^K \beta_k \Sigma_k^{-1})(\sum_{k=1}^K \Sigma_k^{-1} + V^{-1})^{-1}$ Note, there is no reason here that β_k cannot itself be a matrix.

To compute the distribution of U, we combine the last lines of (3.159) and (3.163). This gives

$$|U|^{-\frac{\nu_U + N + (K+1)D + 1}{2}} \exp\Bigg(-\frac{1}{2}\mathrm{tr}\Bigg\{U^{-1}\Bigg(\beta^\dagger V^{-1}\beta^{\dagger\prime} + \sum_{k=1}^K \beta_k \Sigma_k^{-1}\beta_k' + \Psi_U \tag{3.164}$$
$$- \left(\beta^\dagger V^{-1} + \sum_{k=1}^K \beta_k \Sigma_k^{-1}\right)\left(\sum_{k=1}^K \Sigma_k^{-1} + V^{-1}\right)^{-1}\left(\beta^\dagger V^{-1} + \sum_{k=1}^K \beta_k \Sigma_k^{-1}\right)'\Bigg)\Bigg\}\Bigg)$$

Clearly, U is marginally inverse-Wishart. It has $\nu_U + (K+1)D$ degrees of freedom, and its scale matrix equals $\beta^\dagger V^{-1}\beta^{\dagger\prime} + \sum_{k=1}^K \beta_k \Sigma_k^{-1}\beta_k' + \Psi_U - (\beta^\dagger V^{-1} + \sum_{k=1}^K \beta_k \Sigma_k^{-1})(\sum_{k=1}^K \Sigma_k^{-1} + V^{-1})^{-1}(\beta^\dagger V^{-1} + \sum_{k=1}^K \beta_k \Sigma_k^{-1})'$.

Derivation of the Posterior of the Innovation Covariances' Mean

The product of the relevant likelihood and prior is

$$\Omega \,|\, \{\Sigma_k,\}_{k=1}^K \propto \prod_{k=1}^K |\Omega|^{\frac{\mu_1+D-1}{2}} \exp\left(-\frac{\mu_1-2}{2}\operatorname{tr}\{\Omega\Sigma_k^{-1}\}\right) \cdot |\Omega|^{\frac{\mu_2-2}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\{\operatorname{diag}(a_1,\ldots,a_D)^{-1}\Omega\}\right). \tag{3.165}$$

Since matrix multiplication distributes over matrix addition:

$$= |\Omega|^{\frac{K(\mu_1+D-1)}{2}} \exp\left(-\frac{\mu_1-2}{2}\sum_{k=1}^K \operatorname{tr}\{\Omega\Sigma_k^{-1}\}\right) \cdot |\Omega|^{\frac{\mu_2-2}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\{\operatorname{diag}(a_1,\ldots,a_D)^{-1}\Omega\}\right) \tag{3.166}$$

$$= |\Omega|^{\frac{K(\mu_1+D-1)+\mu_2-2}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\{\left(\operatorname{diag}(a_1,\ldots,a_D)^{-1} + (\mu_1-2)\sum_{k=1}^K \Sigma_k^{-1}\right)\Omega\}\right). \tag{3.167}$$

This is the kernel of a Wishart distribution. That is

$$\Omega \,|\, \{\Sigma_k\}_{k=1}^K \sim \mathcal{W}\left(K(\mu_1+D-1)+(\mu_2+D-1), \left(\operatorname{diag}(a_1,\ldots,a_D)^{-1} + (\mu_1-2)\sum_{k=1}^K \Sigma_k^{-1}\right)^{-1}\right). \tag{3.168}$$

Chapter 4

IDENTIFICATION ROBUST INFERENCE FOR RISK PRICES IN STRUCTURAL STOCHASTIC VOLATILITY MODELS

By Xu Cheng, Eric Renault, and Paul Sangrey

In structural stochastic volatility asset pricing models, changes in volatility affect risk premia through two channels: (1) the investor's willingness to bear high volatility in order to get high expected returns as measured by the market return risk price, and (2) the investor's direct aversion to changes in future volatility as measured by the volatility risk price. Disentangling these channels is difficult and poses a subtle identification problem that invalidates standard inference. We adopt the discrete-time exponentially affine model of Han, Khrapov, and Renault (2018), which links the identification of the volatility risk price to the leverage effect. In particular, we develop a minimum distance criterion that links the market return risk price, the volatility risk price, and the leverage effect to well-behaved reduced-form parameters that govern the return and volatility's joint distribution. The link functions are almost flat if the leverage effect is close to zero, making estimating the volatility risk price difficult. We translate the conditional quasi-likelihood ratio test Andrews and Mikusheva (2016) develop in a nonlinear GMM framework to a minimum distance framework. The resulting conditional quasi-likelihood ratio test is uniformly valid. We invert this test to derive robust confidence sets that provide correct coverage for the risk prices regardless of the leverage effect's magnitude.

4.1 Introduction

A fundamental question in finance is how investors optimally trade off risk and return. Economic theory predicts investors demand a higher return as compensation for bearing more risk. Hence, we should expect a positive relationship between the mean and volatility of returns. Some seminal early papers proposed a static trade-off between risk and expected return, most notably the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965). In practice, volatility varies over time. Consequently, a significant strand of the recent literature examines the dynamic tradeoff between volatility and returns, including structural stochastic volatility models such as Christoffersen, Heston, and Jacobs (2013), Bansal et al. (2014), and Dew-Becker et al. (2017). In nonlinear models like these, investors care not just about how an asset's returns co-move with the volatility but also care how they co-move with changes in volatility.

In these structural stochastic volatility models, changes in volatility affect risk premia through two channels: (1) the investor's willingness to tolerate high volatility in order to get high expected returns as measured by the market return risk price, and (2) the investor's direct aversion to changes in future volatility as measured by the volatility risk price. We adopt the discrete-time exponentially affine model of Han, Khrapov, and Renault (2018), who represent the market return risk price and the volatility risk price by two structural parameters. In this model, Han, Khrapov, and Renault (2018) establish the significant result that the identification of the volatility risk price depends on a substantial leverage effect, which is the negative contemporaneous correlation between returns and volatility.

Although the leverage effect is theoretically less than zero, it is difficult to quantify empirically, and its estimate usually is small (Aït-Sahalia, Fan, and Li 2013). When the leverage effect is small, the data only provide a limited amount of information about the volatility risk, compared to the finite-sample noise in the data. This low signal-to-noise ratio, as modeled by weak identification, invalidates standard inference based on the generalized method of moments (GMM) estimator; see Stock and Wright (2000) and Andrews and Cheng (2012).

We provide an identification-robust confidence set for the structural parameters that measure the market return risk price, the volatility risk price, and the leverage

effect. The robust confidence set provides correct asymptotic coverage, uniformly over a large set of models and allows for any magnitude of the leverage effect. This uniform validity is crucial for the confidence set to have good finite-sample coverage (Mikusheva 2007; Andrews and Guggenberger 2010). In contrast, standard confidence sets based on the GMM estimator and its asymptotic normality do not have uniform validity in the presence of a small leverage effect. This issue affects all of the structural parameters because they are estimated simultaneously.

We achieve robust inference in two steps. First, we establish a minimum distance criterion using link functions between the structural parameters and a set of reduced-form parameters that determine the joint distribution of the return and volatility. The structural model implies that the link functions are zero when evaluated at the true values of the structural parameters and the reduced-form parameters. Identification and estimation of these reduced form parameters are standard and are not affected by the presence of a small leverage effect. However, the link functions are almost flat in one of the structural parameters when the leverage effect is small, resulting in weak identification. Second, given this minimum distance criterion, we invert the conditional quasi-likelihood ratio (QLR) test by Andrews and Mikusheva (2016) to construct a robust confidence set. The key feature of this test is that it treats the flat link functions as an infinite-dimensional nuisance parameter. The critical value is constructed by conditioning on a sufficient statistic for this nuisance parameter, and it is known to yield a valid test regardless of the nuisance parameter's value. Andrews and Mikusheva (2016) develop this test in a GMM framework. We show it works in minimum distance contexts such as the one considered here and provide conditions for its asymptotic validity. For practitioners, we provide a detailed algorithm for the construction of this simulation-based robust confidence set.

Our empirical results relates to the empirical analysis of the effect of volatility on risk premia. As Lettau and Ludvigson (2010) mention, the evidence here is inconclusive. Bollerslev, Engle, and Wooldridge (1988), Harvey (1989), Ghysels, Santa-Clara, and Valkanov (2005), Bali and Peng (2006), and Ludvigson and Ng (2007) find a positive relationship, while Campbell (1987), Breen, Glosten, and Jagannathan (1989), Pagan and Hong (1991), Whitelaw (1994), and Brandt and Kang (2004) find a negative relationship. Also, some papers use both a market return risk factor and a

variance risk factor to explain the risk premia dynamics, including Christoffersen, Heston, and Jacobs (2013), Feunou et al. (2014), and Dew-Becker et al. (2017). In related strand of the literature, Bollerslev, Law, and Tauchen (2008) and Drechsler and Yaron (2011) document a substantial positive variance risk premium. We contribute to this literature by providing the first method for making valid inference for the market return risk price and the volatility risk price. This new confidence set not only allows for both effects but also takes into account the potential identification issue.

To have a non-linear relationship between changes in volatility and expected returns, we need either volatility of volatility (as used by Drechsler and Yaron (2011)) or jumps (as used by Drechsler (2013)). Since we are working in discrete-time, it is far more natural to work volatility than with jumps. This is because discontinuities are not well-defined in discrete-time; all functions are continuous in the discrete-topology. The most straightforward models that allow for closed-form expressions for the risk prices are exponentially-affine (not affine) models. This is why they are frequently used in the option-pricing literature. In order to avoid complicating the analysis, we use such a model. We focus on the time-series behavior of the index because as a complement to, not a subtitle to, estimating the risk prices from cross-sectional or options pricing data. Using market-level variation over time to examine this non-linear relationship is a common approach in the literature, used by both the variance-premium literature and Han, Khrapov, and Renault (2018).

The weak identification issue studied in this paper is relevant in many economic applications, ranging from linear instrumental variable models (Staiger and Stock 1997) to nonlinear structural models (Mavroeidis, Plagborg-Møller, and Stock 2014; Andrews and Mikusheva 2015). This paper is the first one to study this issue in structural asset pricing models with stochastic volatility.

Moreira (2003) introduces the conditional inference approach to the linear instrumental variable model creating the conditional likelihood-ratio (CLR) test, and Kleibergen (2005) applies it to the nonlinear GMM problem. Magnusson and Mavroeidis (2010) and Magnusson (2010) extend Kleibergen's (2005) results to the minimum-distance case. The key issue with these papers is that they rely exclusively upon local behavior of the moment-conditions. This is inherently under-powered in some envi-

ronments. Andrews and Mikusheva (2016) resolve this issue by developing a global approach in the GMM case by proposing conditional inference for nonlinear GMM problems with an infinite-dimensional nuisance parameter. Their method is known to be the most-powerful in some special cases. We develop a global weak-identification robust inference method for minimum distance estimation by extending Andrews and Mikusheva (2016). We bear the same relationship to Magnusson and Mavroeidis (2010) and Magnusson (2010) that Andrews and Mikusheva (2016) bears to Kleibergen (2005). We also extend the scope of the application of these weak-identification robust methods to a new type of asset pricing model with substantial non-linearity and heteroskedasticity.

The rest of the paper is organized as follows. Section 4.2 provides the model and its parameterization. Section 4.3 provides model-implied restrictions and use them to derive the link function. Section 4.4 provides the asymptotic distribution of the reduced-form parameter and robust confidence sets for the structural parameter. A detailed algorithm to construct the robust confidence set is given in Section 4.4. Section 4.5 show that the method works well in simulation, and Section 4.6 applies the methods to data on the S&P 500 providing estimates of the risk prices. Section 4.7 concludes. Proofs are given in the appendix.

4.2 Model

This section provides a parametric structural model with stochastic volatility, following Han, Khrapov, and Renault (2018). They extend the discrete-time exponentially-affine model of Darolles, Gouriéroux, and Jasiak (2006), and their model is a natural discrete-time analog of the Heston (1993) model. We specify this model using a stochastic discount factor (SDF), also called the pricing kernel, and the physical measure, which gives the joint distribution of the return and volatility dynamics.[60] We first define the SDF and parameterize it as an exponential affine function with unknown parameters. We then provide parametric distribution for the physical measure.

Let P_t be the price of the asset under consideration. Let $r_{t+1} = \log(P_{t+1}/P_t) - r_f$ denote the log excess return minus the risk-free rate and σ^2_{t+1} denote the associated

[60] The risk-neutral measure is unobserved due to the lack of option data.

volatility. The observed data is $W_t = (r_t, \sigma_t^2)$ for $t = 1, \ldots, T$. Let $\mathcal{F}_t$ be the representative investor's information set at time t .

Stochastic Discount Factor and Its Parameterization

The prices of all assets satisfy the following asset pricing equation in terms of the SDF:

$$P_t = \mathbb{E}\left[M_{t,t+1} \exp\left(-r_f\right) P_{t+1} \,|\, \mathcal{F}_t\right]. \tag{4.1}$$

Following the definition of r_{t+1}, the pricing equation implies that for all assets

$$1 = \mathbb{E}\left[M_{t,t+1} \exp\left(r_{t+1}\right) \,|\, \mathcal{F}_t\right]. \tag{4.2}$$

We start by parameterizing the SDF by the exponential affine model. Let π be the price of volatility risk and κ be the market return risk price. They are both considered as structural parameters.

Definition 4.1. Parameterizing the Stochastic Discount Factor

$$M_{t,t+1}(\pi, \kappa) = \exp\left(m_0 + m_1 \sigma_t^2 - \pi \sigma_{t+1}^2 - \kappa r_{t+1}\right). \tag{4.3}$$

Throughout we assume that the two risks that command nonzero prices are the market return risk price and the volatility risk price. These two risks are closely related to the first two moments of r_{t+1}. Consequently, we only use variation in the first two moments of the data to estimate these parameters.

Parameterizing the Volatility and Return Dynamics

Next, we parameterize the joint distribution of $\{W_t : t = 1, \ldots, T\}$. Following Han, Khrapov, and Renault (2018), we make the following assumptions. First, the return r_t and volatility σ_t^2 are first-order Markov. Second, there is no Granger-causality from the return to the volatility. Third, returns are independent across time given the volatility. We do allow σ_t^2 and r_t to be contemporaneously correlated, as they are in the data.

Under these assumptions, the volatility drives all of the dynamics of the process. The only relevant information in the information set $\mathcal{F}_t$ for time $t+1$-measurable

variables is contained in σ_t^2. In general, σ_t^2, σ_{t+1}^2, and r_{t+1} form a sufficient statistic for $\mathcal{F}_{t+1}$.

We adopt the conditional autoregressive gamma process as in Gouriéroux and Jasiak (2006) and Han, Khrapov, and Renault (2018) for the volatility process. The model is parameterized in terms of the Laplace transform:

$$\mathbb{E}\left[\exp(-x\sigma_{t+1}^2) \,\middle|\, \mathcal{F}_t\right] = \exp\left(-A(x)\sigma_t^2 - B(x)\right) \tag{4.4}$$

for all $x \in \mathbb{R}$. The function $A(x)$ and $B(x)$ are parameterized as follows.

Definition 4.2. Parameterize the Volatility Dynamics

$$A(x) \coloneqq \frac{\rho x}{1 + cx}, \tag{4.5}$$

$$B(x) \coloneqq \delta \log(1 + cx), \tag{4.6}$$

with $\rho \in [0, 1 - \epsilon]$, $c > \epsilon$, $\delta > \epsilon$ for some $\epsilon > 0$.

In this specification, ρ is a persistence parameter, c is a scale parameter, and δ is a level parameter. We can see this clearly in the following conditional mean and variance formulas for σ_{t+1}^2.

Remark 4.1 (Volatilty Moment Conditions)**.**

$$\mathbb{E}\left[\sigma_{t+1}^2 \,\middle|\, \sigma_t^2\right] = \rho\sigma_t^2 + c\delta, \tag{4.7}$$

$$\mathbb{V}\mathrm{ar}\left[\sigma_{t+1}^2 \,\middle|\, \sigma_t^2\right] = 2c\rho\sigma_t^2 + c^2\delta. \tag{4.8}$$

Next, we model the return dynamics. Similar to the volatility dynamics, the distribution of r_t given both σ_{t+1}^2 and σ_t^2 is specified in terms of the Laplace transform:

$$\mathbb{E}\left[\exp(-xr_{t+1}) \,\middle|\, \mathcal{F}_t, \sigma_{t+1}^2\right] = \exp\left(-C(x)\sigma_{t+1}^2 - D(x)\sigma_t^2 - E(x)\right) \tag{4.9}$$

for all $x \in \mathbb{R}$. The function $C(x)$, $D(x)$, and $E(x)$ are parameterized as follows such that the return has a conditional Gaussian distribution.

Definition 4.3. Parameterizing the Return Dynamics

$$C(x) \coloneqq \psi x - \frac{1 - \phi^2}{2}x^2, \tag{4.10}$$

$$D(x) \coloneqq \beta x, \tag{4.11}$$

$$E(x) \coloneqq \gamma x, \tag{4.12}$$

with $\phi \in [-1 + \epsilon, 0]$ for some $\epsilon > 0$.

Under this specification, we have the following representation of the conditional mean and variance for r_{t+1}.

Remark 4.2 (Return Moment Conditions)**.**

$$\mathbb{E}\left[r_{t+1} \,\middle|\, \sigma_t^2, \sigma_{t+1}^2\right] = \psi\sigma_{t+1}^2 + \beta\sigma_t^2 + \gamma, \tag{4.13}$$

$$\mathbb{V}\text{ar}\left[r_{t+1} \,\middle|\, \sigma_t^2, \sigma_{t+1}^2\right] = (1-\phi^2)\sigma_{t+1}^2. \tag{4.14}$$

The parameter ϕ represents the leverage effect because it measures the reduction in return's volatility caused by conditioning on the volatility path.

4.3 Link Functions

So far, we have introduced the following parameters: (m_0, m_1, κ, π) in SDF, (ρ, c, δ) for the volatility dynamics, and $(\psi, \beta, \gamma, \phi)$ for the return dynamics. Next, we explore restrictions among these parameters that are consistent with this model. In other words, not all of these parameters can change freely under the structural model.

We use these restrictions to construct link functions between a set of reduced-form parameters and a set of structural parameters. These link functions play an important role in separating the regularly behaved reduced-form parameters from the structural parameters. They also are used to conduct identification robust inference for the structural parameters based on a minimum distance criterion. All of these restrictions are also imposed in the GMM estimation in Han, Khrapov, and Renault (2018). However, because the volatility risk price is weakly identified, they calibrate it instead of estimating it. Given this calibrated value, they proceed to estimate all other parameters with GMM.

Pricing Equation Restrictions

We first explore restrictions implied by the pricing equation $\mathbb{E}[M_{t,t+1}\exp(r_{t+1}) \,|\, \mathcal{F}_t] = 1$. We start with a simple result stating that the constants m_0 and m_1 are normalization constants implied by all the other parameters. Thus, m_0 and m_1 are not free parameters to be estimated. Instead, they should take the values given below, once other parameters are specified. These restrictions on m_0 and m_1 are obtained by

applying the restriction $\mathbb{E}[M_{t,t+1}\exp(r_{t+1})|\mathcal{F}_t] = 1$ to the risk free asset. Applying the same argument to any other asset, we also obtain another set of two restrictions, which can be written in terms of the coefficients β and γ under the linear form of $D(x)$ and $E(x)$.

Lemma 4.1. *Given the parameterization in the model, the pricing equation* $\mathbb{E}[M_{t,t+1}\exp(r_{t+1})\,|\,\mathcal{F}_t] = 1$ *implies that*[61]

$$m_0 = E(\kappa) + B\left(\pi + C\left(\kappa\right)\right),$$
$$m_1 = D\left(\kappa\right) + A\left(\pi + C\left(\kappa\right)\right),$$

and

$$\gamma = B\left(\pi + C\left(\kappa - 1\right)\right) - B\left(\pi + C\left(\kappa\right)\right),$$
$$\beta = A\left(\pi + C\left(\kappa - 1\right)\right) - A\left(\pi + C\left(\kappa\right)\right).$$

The two equalities on β and γ link them to the market return risk price, κ, and the volatility risk price, π, through the functions $A(\cdot), B(\cdot), C(\cdot)$, which also involve the parameters $(\rho, c, \delta, \psi, \phi)$. We treat these two equalities as link functions in the minimum distance criterion specified below.

Leverage Effect Restrictions

Following Han, Khrapov, and Renault (2018), we parameterize ψ as

$$\psi = \frac{\phi}{\sqrt{2c}} - \frac{1-\phi^2}{2} + (1-\phi^2)\kappa. \tag{4.15}$$

The first part $\phi/\sqrt{2c}$ measures the leverage effect arising from the instantaneous correlation between r_{t+1} and σ^2_{t+1}. The second part is the traditional Jensen effect term that arises from taking expectation of a log-Gaussian random variable. The third term arises from risk-aversion, which is why it is proportional to κ.

[61] Proof 4.A

Structural and Reduced-Form Parameters

Because ϕ is the leverage effect parameter, we group it together with market return risk price, κ, and the volatility risk price, π, and call $\theta \coloneqq (\kappa, \pi, \phi)'$ structural parameters. These structural parameters are estimated by restrictions from this structural model. In contrast, the other parameters in the conditional mean and variance of the return and volatility, see Remark 4.1 and Remark 4.2, are simply estimated using these moments, without any model restrictions. As such, we call them the reduced-form parameters. Because $1 - \phi^2$ shows up in the conditional variance of r_{t+1}, see (4.14), we define $\zeta = 1 - \phi^2$ as a reduced-form parameter and link it to the structural parameter ϕ through this relationship. To sum up, the reduced-form parameters are $\omega \coloneqq (\rho, c, \delta, \psi, \beta, \gamma, \zeta)'$.

Using ζ as a reduced-form parameter has the additional benefit of avoiding estimating ϕ directly. Estimating ϕ when its true value is close to 0 results in an estimator with a non-standard asymptotic distribution due to the boundary constraint. The inference procedure below does not require estimation of ϕ and is uniform over ϕ even if its true value is on or close to the boundary 0. It is worth noting that this boundary condition gives us additional information in estimating ϕ in some cases. The estimator for ϕ may converge quite rapidly; however, it is almost certainly not asymptotically approximately Gaussian. In addition, we cannot recover asymptotic Gaussianity by removing this constraint. Even though, ϕ could conceivably be greater than 0, ϕ^2 cannot conceivably be less than 0. The ϕ parameter enters the last link in (4.16) through ϕ^2. This is where the non-standard behavior arises. Economically, we saying that the variance of r_{t+1} must reduce when we condition on more information. Although, this is clearly in population, it may not hold for the sample variances.

The link functions between the structural parameter θ and the reduced-form parameter ω are collected together in

$$g(\theta, \omega) = \begin{pmatrix} \gamma - [B(\pi + C(\kappa - 1)) - B(\pi + C(\kappa))] \\ \beta - [A(\pi + C(\kappa - 1)) - A(\pi + C(\kappa))] \\ \psi - (1 - \phi^2)\kappa + \frac{1}{2}(1 - \phi^2) - 1/(2c)^{1/2}\phi \\ \zeta - (1 - \phi^2) \end{pmatrix}. \tag{4.16}$$

For the inference problem studied below, we know $g(\theta_0, \omega_0) = 0$ when evaluated at the true value of θ and ω.

Identification

One of the important contributions of Han, Khrapov, and Renault (2018) is to establish the relationship between the identification of the volatility risk price and the leverage effect. In particular, they show that when the leverage effect parameter $\phi = 0$, the volatility risk price π is not identified. To see this result, note that the only source of identification information on π are the first two link functions in $g(\theta_0, \omega_0) = 0$, which come from Lemma 4.1. Clearly, these two equations are independent of π if $C(\kappa) = C(\kappa - 1)$. Using the definition of $C(\cdot)$ and (4.15), we have

$$C(\kappa) - C(\kappa - 1) = \psi - (1 - \phi^2)\left(\kappa - \frac{1}{2}\right) = \frac{\phi}{\sqrt{2c}}. \tag{4.17}$$

Clearly, the strength of identification is governed by the strength of the leverage effect. In other words, we need $\phi \neq 0$ to identify the volatility risk price π.

Even if $\phi \neq 0$, we do not know it. In practice, with a finite-sample size and different types of noise in the data, such as measurement errors and omitted variables, a substantial leverage effect is required to obtain a standard identification situation where the noise in the data is negligible compared to the information to identify π. However, if only a small leverage effect is found, as in Bandi and Renò (2012) and Aït-Sahalia, Fan, and Li (2013), or the magnitude of the leverage effect is completely unknown, an identification robust procedure is needed to conduct inference in this problem. In addition, standard minimum-distance estimators do not provide valid inference when some of the first-stage parameters are either asymptotically non-Gaussian or the link functions are ill-behaved. In our case, we should not expect ϕ to be asymptotically Gaussian even though it is well identified. We provide a procedure that is robust to both non-standard issues now.

4.4 Robust Inference for Risk Prices

Asymptotic Distribution of the Reduced-Form Parameter

Write $\omega = (\omega_1, \omega_2, \omega_3)'$, where $\omega_1 = (\rho, c, \delta)' \in O_1$, $\omega_2 = (\gamma, \beta, \psi)' \in O_2$, and $\omega_3 = \zeta \in O_3$. The parameter space for ω is $O = O_1 \times O_2 \times O_3 \subset R^{d_\omega}$. The true value of ω is assumed to be in the interior of the parameter space.

Below we describe the estimator $\widehat{\omega} \coloneqq (\widehat{\omega}_1, \widehat{\omega}_2, \widehat{\omega}_3)'$ and provide its asymptotic distribution. We estimate these parameters separately because ω_1 only shows up in the conditional mean and variance of σ_{t+1}^2, ω_2 only shows up in the conditional mean of r_{t+1}, and ω_3 only shows up in the conditional variance of r_{t+1}.

We first estimate $\omega_1 = (\rho, c)'$ based on the conditional mean and variance of σ_{t+1}^2, which can be equivalently written as

$$\begin{aligned} &E[\sigma_{t+1}^2|\sigma_t^2] = A \text{ and } E[\sigma_{t+1}^4|\sigma_t^2] = B, \text{ where} \\ &\qquad A = \rho\sigma_t^2 + c\delta \text{ and } B = A^2 + \left(2c\rho\sigma_t^2 + c^2\delta\right). \end{aligned} \tag{4.18}$$

Because the conditional mean of σ_{t+1}^2 and σ_{t+1}^4 are linear and quadratic functions, respectively, of the conditioning variable σ_t^2, they can be transformed to the unconditional moments

$$E[h_t(\omega_{10})] = 0, \text{ where } h_t(\omega_1) = [(1, \sigma_t^2) \otimes (\sigma_{t+1}^2 - A), (1, \sigma_t^2, \sigma_t^4) \otimes (\sigma_{t+1}^4 - B)]', \tag{4.19}$$

and ω_{10} represents the true value of ω_1. The two-step GMM estimator of ω_1 is

$$\widehat{\omega}_1 = \underset{\omega_1 \in O_1}{\arg\min} \left(T^{-1} \sum_{t=1}^{T} h_t(\omega_1) \right)' \widehat{V}_1 \left(T^{-1} \sum_{t=1}^{T} h_t(\omega_1) \right), \tag{4.20}$$

where $\widehat{V}_1$ is a consistent estimator of $V_1 \coloneqq \sum_{m=-\infty}^{\infty} \mathbb{C}\text{ov}[h_t(\omega_{10}), h_{t+m}(\omega_{10})]$.

We estimate ω_2 by the generalized least squares (GLS) estimator because the conditional mean of r_{t+1} is a linear function of the conditioning variable σ_t^2 and σ_{t+1}^2 and the conditional variance is proportional to σ_{t+1}^2. The GLS estimator of ω_2 is

$$\begin{aligned} &\widehat{\omega}_2 = \left(\sum_{t=1}^{T} x_t x_t' \right)^{-1} \sum_{t=1}^{T} x_t y_t, \text{ where} \\ &x_t = \sigma_{t+1}^{-1}(1, \sigma_t^2, \sigma_{t+1}^2)' \text{ and } y_t = \sigma_{t+1}^{-1} r_{t+1}. \end{aligned} \tag{4.21}$$

We estimate ω_3 by the sample variance estimator:

$$\widehat{\omega}_3 = T^{-1} \sum_{t=1}^{T} (y_t - \widehat{y}_t)^2, \text{ where } \widehat{y}_t = x_t' \widehat{\omega}_2. \tag{4.22}$$

Let P denote the distribution of the data $\{W_t = (r_{t+1}, \sigma_{t+1}^2) : t \geq 1\}$ and $\mathcal{P}$ denote the parameter space of P. Note that the true values of the structural parameter and

the reduced-form parameters are all determined by P. We allow P to change with T. For notational simplicity, the dependence on P and T is suppressed.

Let

$$f_t(\omega) = \begin{pmatrix} h_t(\omega_1) \\ x_t(y_t - x_t'\omega_2) \\ (y_t - x_t'\omega_2)^2 \end{pmatrix} \in R^{d_f} \text{ and } V = \sum_{m=-\infty}^{\infty} \mathbb{C}\text{ov}\left[f_t(\omega_0), f_{t+m}(\omega_0)\right]. \tag{4.23}$$

The estimator $\widehat{\omega}$ defined above is based on the first moment of $f_t(\omega)$. Thus, the limiting distribution of $\widehat{\omega}$ relates to the limiting distribution of $T^{-1/2}\sum_{t=1}^{T}(f_t(\omega_0) - \mathbb{E}[f_t(\omega_0])$ following from the central limit theorem. Furthermore, because ω_1 is the GMM estimator based on some nonlinear moment conditions, we need uniform convergence of the sample moments and their derivatives to show the consistency and asymptotic normality of $\widehat{\omega}_1$. These uniform convergence follows from the uniform law of large numbers. Because $\widehat{\omega}_2$ is a simple OLS estimator by regressing y_t and x_t, we need the regressors to not exhibit multicollinearity. We make the necessary assumptions below. All of them are easily verifiable with weakly dependent time series data.

Let $\widehat{V}$ denote a heteroskedasticity and autocorrelation consistent (HAC) estimator of V. The estimator $\widehat{V}_1$ is a submatrix of $\widehat{V}$ associate with V_1. Let $H_t(\omega_1) = \partial h_t(\omega_1)/\partial\omega_1'$.

Assumption R. The following conditions hold uniformly over $P \in \mathcal{P}$, for some fixed $0 < C < \infty$.

1. $T^{-1}\sum_{t=1}^{T}(h_t(\omega_1) - \mathbb{E}[h_t(\omega_1)) \to_p 0$ and $T^{-1}\sum_{t=1}^{T}(H_t(\omega_1) - \mathbb{E}[H_t(\omega_1)]) \to_p 0$, $\mathbb{E}[H_t(\omega_1)]$ is continuous in ω_1, all uniformly over the parameter space of ω_1.
2. $T^{-1}\sum_{t=1}^{T}(x_t x_t' - \mathbb{E}[x_t x_t']) \to_p 0$.
3. $V^{-1/2}\{T^{-1/2}(\sum_{t=1}^{T} f_t(\omega_0) - \mathbb{E}[f_t(\omega_0)]\} \to_d N(0, I)$ and $\widehat{V} - V \to_p 0$.
4. $C^{-1} \le \lambda_{\min}(A) \le \lambda_{\max}(A) \le C$ for $A = V, \mathbb{E}[H_t\left(\omega_{1,0}\right)' H_t\left(\omega_{1,0}\right)]), \mathbb{E}[x_t x_t'], \mathbb{E}[z_t z_t']$, where $z_t = (1, \sigma_t^2, \sigma_t^4)'$.[62]

[62] We use λ(matrix) to denote the eigenvalue of the matrix.

Let $H(\omega_1) = \mathbb{E}[H_t(\omega_1)]$ and $\overline{H}(\omega_1) = T^{-1}\sum_{t=1}^{T} H_t(\omega_1)$. Define

$$\mathcal{B} = \text{diag}\{[H(\omega_{10})V_1^{-1}H(\omega_{10})]^{-1}H(\omega_{10})V_1^{-1}, \mathbb{E}[x_t x_t']^{-1}, 1\},$$
$$\widehat{\mathcal{B}} = \text{diag}\{[\overline{H}(\widehat{\omega}_1)'\widehat{V}_1^{-1}\overline{H}(\widehat{\omega}_1)]^{-1}\overline{H}(\widehat{\omega}_1)'\widehat{V}_1^{-1}, [T^{-1}\sum_{t=1}^{T} x_t x_t']^{-1}, 1\}. \tag{4.24}$$

The following lemma provides the asymptotic distribution of the reduced-form parameter and a consistent estimator of its asymptotic covariance. Note, we put the asymptotic covariance on the left side of the convergence to allow the distribution of the data to change with sample size T.

Lemma 4.2. *Suppose Assumption R holds. The following results hold uniformly over* $P \in \mathcal{P}$.[63]

$$\xi_T := \Omega^{-1/2}T^{-1/2}(\widehat{\omega} - \omega_0) \to_d \xi \sim N(0, I), \text{ where } \Omega = \mathcal{B}V\mathcal{B}',$$

and

$$\widehat{\Omega} - \Omega \to_p 0, \text{ where } \widehat{\Omega} = \widehat{\mathcal{B}}\widehat{V}\widehat{\mathcal{B}}'.$$

Weak Identification

The true value of the structural parameter θ and the reduced-form parameter ω satisfy the link function $g(\theta_0, \omega_0) = 0$. In a standard problem without any identification issues, we can estimate θ_0 by the minimum distance estimator $\widehat{\theta} = (\widehat{\kappa}, \widehat{\pi}, \widehat{\phi})'$, which minimizes $Q_T(\theta) = g(\theta, \widehat{\omega})'W_T g(\theta, \widehat{\omega})$ for some weighting matrix W_T, and construct tests and confidence sets for θ_0 using an asymptotically normal approximation for $T^{1/2}(\widehat{\theta} - \theta_0)$. However, this standard method does not work in the present problem when π_0 is only weakly identified. In this case, $g(\theta, \widehat{\omega})$ is almost flat in π and the minimum distance estimator of $\widehat{\pi}$ is not even consistent. To make the problem even more complicated, the inconsistency of $\widehat{\pi}$ has a spillover effect on $\widehat{\kappa}$ and $\widehat{\phi}$, making the distribution of $\widehat{\kappa}$ and $\widehat{\phi}$ non-normal even in large samples.

Before presenting the robust confidence set, we first introduce some useful quantities and provide a heuristic discussion of the identification problem and its consequences. Let $G(\theta, \omega)$ denote the partial derivative of $g(\theta, \omega)$ with respect to (w.r.t.) ω.

[63] Proof 4.A

Let $g_0(\theta) = g(\theta, \omega_0)$ and $G_0(\theta) = G(\theta, \omega_0)$ be the link function and its derivative evaluated at ω_0 and $\widehat{g}(\theta) = g(\theta, \widehat{\omega})$ and $\widehat{G}(\theta) = G(\theta, \widehat{\omega})$ be the same quantities evaluated at the estimator $\widehat{\omega}$. The delta method gives

$$\eta_T(\theta) := T^{1/2}\left[\widehat{g}(\theta) - g_0(\theta)\right] = G_0(\theta)\Omega^{1/2} \cdot \xi_T + o_p(1), \tag{4.25}$$

where $\xi_T \to_d N(0, I)$ following Lemma 4.2. Thus, $\eta_T(\cdot)$ weakly converges to a Gaussian process $\eta(\cdot)$ with covariance function $\Sigma(\theta_1, \theta_2) = G_0(\theta_1)\Omega G_0(\theta_2)'$.

Following (4.25), we can write $T^{1/2}\widehat{g}(\theta) = \eta_T(\theta) + T^{1/2}g_0(\theta)$, where $\eta_T(\theta)$ is the noise from the reduced-form parameter estimation and $T^{1/2}g_0(\theta)$ is the signal from the link function. Under weak identification, $g_0(\theta)$ is almost flat in θ, modeled as the signal $T^{1/2}g_0(\theta)$ being finite even for $\theta \neq \theta_0$ and $T \to \infty$. Thus, the signal and the noise are of the same order of magnitude, yielding an inconsistent minimum distance estimator $\widehat{\theta}$. This is in contrast with the strong identification scenario, where $T^{1/2}g_0(\theta) \to \infty$ for $\theta \neq \theta_0$ as $T \to \infty$ and $g_0(\theta_0) = 0$. In this case, the signal is strong enough that the minimum distance estimator is consistent.

The identification strength of θ_0 is determined by the function $T^{1/2}g_0(\theta)$. However, this function is unknown and cannot be consistently estimated (due to $T^{1/2}$). Thus, we take the conditional inference procedure as in Andrews and Mikusheva (2016) and view $T^{1/2}g_0(\theta)$ as an infinite dimensional nuisance parameter for the inference of θ_0. The goal is to construct robust confidence set for θ_0 that has correct size asymptotically regardless of this unknown nuisance parameter.

Conditional QLR Test

We construct a confidence set for $\theta \in \Theta := [0, M_1] \times [-M_2, 0] \times [1 - \epsilon, 0]$ by inverting the test $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$, where M_1 and M_2 are large positive constants and ϵ is a small positive constant. The test statistic is a QLR statistic that takes the form

$$QLR(\theta_0) := T\widehat{g}(\theta_0)'\widehat{\Sigma}(\theta_0, \theta_0)^{-1}\widehat{g}(\theta_0) - \min_{\theta \in \Theta} T\widehat{g}(\theta)'\widehat{\Sigma}(\theta, \theta)^{-1}\widehat{g}(\theta), \tag{4.26}$$

where $\widehat{\Sigma}(\theta_1, \theta_2,) = \widehat{G}(\theta_1)\widehat{\Omega}\widehat{G}(\theta_2)'$ and $\widehat{\Omega}$ is the consistent estimator of Ω defined above.

Andrews and Mikusheva (2016) provide the conditional QLR test in a nonlinear GMM problem, where $\widehat{g}(\theta)$ is replaced by a sample moment. The same method can be

applied to the present nonlinear minimum distance problem. Following Andrews and Mikusheva (2016), we first project $\widehat{g}(\theta)$ onto $\widehat{g}(\theta_0)$ and construct a residual process

$$\widehat{r}(\theta) = \widehat{g}(\theta) - \widehat{\Sigma}(\theta, \theta_0)\widehat{\Sigma}(\theta_0, \theta_0)^{-1}\widehat{g}(\theta_0). \tag{4.27}$$

The limiting distributions of $\widehat{r}(\theta)$ and $\widehat{g}(\theta_0)$ are Gaussian and independent. Thus, conditional on $\widehat{r}(\theta)$, the asymptotic distribution of $\widehat{g}(\theta)$ no longer depends on the nuisance parameter, $T^{1/2}g_0(\theta)$, making the procedure robust to any identification strength.

Specifically, we obtain the $1-\alpha$ conditional quantile of the QLR statistic, denoted by $c_{1-\alpha}(r, \theta_0)$, as follows. For $b = 1, \ldots, B$, we take independent draws $\eta_b^* \sim N(0, \widehat{\Sigma}(\theta_0, \theta_0))$ and produce a simulated process,

$$g_b^*(\theta) \coloneqq \widehat{r}(\theta) + \widehat{\Sigma}(\theta, \theta_0)\widehat{\Sigma}(\theta_0, \theta_0)^{-1}\eta_b^*, \tag{4.28}$$

and a simulated statistic,

$$QLR_b^*(\theta_0) \coloneqq T\eta_b^{*\prime}\widehat{\Sigma}(\theta_0, \theta_0)^{-1}\eta_b^* - \min_{\theta \in \Pi} T g_b^*(\theta)'\widehat{\Sigma}(\theta, \theta)^{-1}g_b^*(\theta). \tag{4.29}$$

Let $b_0 = \lceil (1-\alpha)B \rceil$, the smallest integer greater than or equal to $(1-\alpha)B$. Then the critical value $c_{1-\alpha}(r, \theta_0)$ is the b_0^{th} smallest value among $\{QLR_b^*, b = 1, \ldots, B\}$. We execute the steps reported in Algorithm 4.1 to form a robust confidence set for θ.

To obtain confidence intervals for each element of θ_0, one simple solution is to project the confidence set constructed above to each axis. The resulting confidence interval also has correct coverage. An alternative solution is to first concentrate out the nuisance parameters before applying the conditional inference approach above, see Andrews and Mikusheva (2016, Section 5). However, this concentration approach only works when the nuisance parameter is strongly identified. In the present set-up, this approach does not work for κ and ϕ because the nuisance parameter π is weakly identified.

Assumption S. The following conditions hold over $P \in \mathcal{P}$, for any θ in its parameter space, and any ω in some fixed neighborhood around its true value, for some fixed $0 < C < \infty$.

1. $g(\theta, \omega)$ is partially differentiable in ω, with partial derivative $G(\theta, \omega)$ that satisfies $||G(\theta_1, \omega) - G(\theta_2, \omega)|| \leq C||\theta_1 - \theta_2||$ and $||G(\theta, \omega_1) - G(\theta, \omega_2)|| \leq C||\omega_1 - \omega_2||$.

Algorithm 4.1 Construing the Confidence Set

1. Estimate the reduced-form parameter $\widehat{\omega} = (\widehat{\omega}_1, \widehat{\omega}_2, \widehat{\omega}_3)'$ following the estimators defined in (4.20), (4.21), and (4.22).
2. Obtain a consistent estimator of $\widehat{\omega}$'s asymptotic covariance $\widehat{\Omega} = \widehat{\mathcal{B}}\widehat{V}\widehat{\mathcal{B}}'$, where $\widehat{\mathcal{B}}$ is defined in (4.24) and $\widehat{V}$ is a HAC estimator of V.
3. For $\theta_0 \in \Theta$,
   a) Construct the QLR statistic $QLR(\theta_0)$ in (4.26) using $g(\theta, \omega)$, $G(\theta, \omega)$, $\widehat{\omega}$, and $\widehat{\Omega}$.
   b) Compute the residual process $\widehat{r}(\theta)$ in (4.27).
   c) Given $\widehat{r}(\theta)$, compute the critical value $c_{1-\alpha}(r, \theta_0)$ described above.
4. Repeat these steps for different values of θ_0. Construct a confidence set by collecting the null values that are not rejected, i.e., the nominal level $1 - \alpha$ confidence set for θ_0 is

$$CS_T = \{\theta_0 : QLR_T(\theta_0) \leq c_{1-\alpha}(r, \theta_0)\}.$$

2. $C^{-1} \leq \lambda_{\min}(G(\theta, \omega)'G(\theta, \omega)) \leq \lambda_{\max}(G(\theta, \omega)'G(\theta, \omega)) \leq C$.

Theorem 4.3. *Suppose Assumption R and Assumption S hold. Then,*

$$\liminf_{T\to\infty} \inf_{P\in\mathcal{P}} \Pr(\theta_0 \in CS_T) \geq 1 - \alpha.^{64}$$

This theorem states that the confidence set constructed by the conditional QLR test has correct asymptotic size. Uniformity is important for this confidence set to cover the true parameter with a probability close to $1 - \alpha$ in finite-samples. This uniform result is established over a parameter space $\mathcal{P}$ that allows for weak identification of the structural parameter θ.

4.5 Simulations

In this section, we investigate the finite-sample performance of the proposed test and show that the asymptotic approximations derived above work well in practice.

[64] Proof 4.A

We also compare it with the standard test that assumes all parameters are strongly identified. The standard test is known to be invalid under weak identification but its degree of distortion is unknown in general. We simulate the data with the parametric model above where the true values of the parameters are given in Table 4.1 based on the values used by Han, Khrapov, and Renault (2018). To investigate the robustness of the procedure with respect to various identification strengths, we vary both ϕ and T. Specifically, we consider $\phi \in \{-0.40, -0.10, -0.01\}$ and $T \in \{2,000; 10,000\}$. The number of data points in the empirical section is $3,700$.

Table 4.1: Simulation Set-up

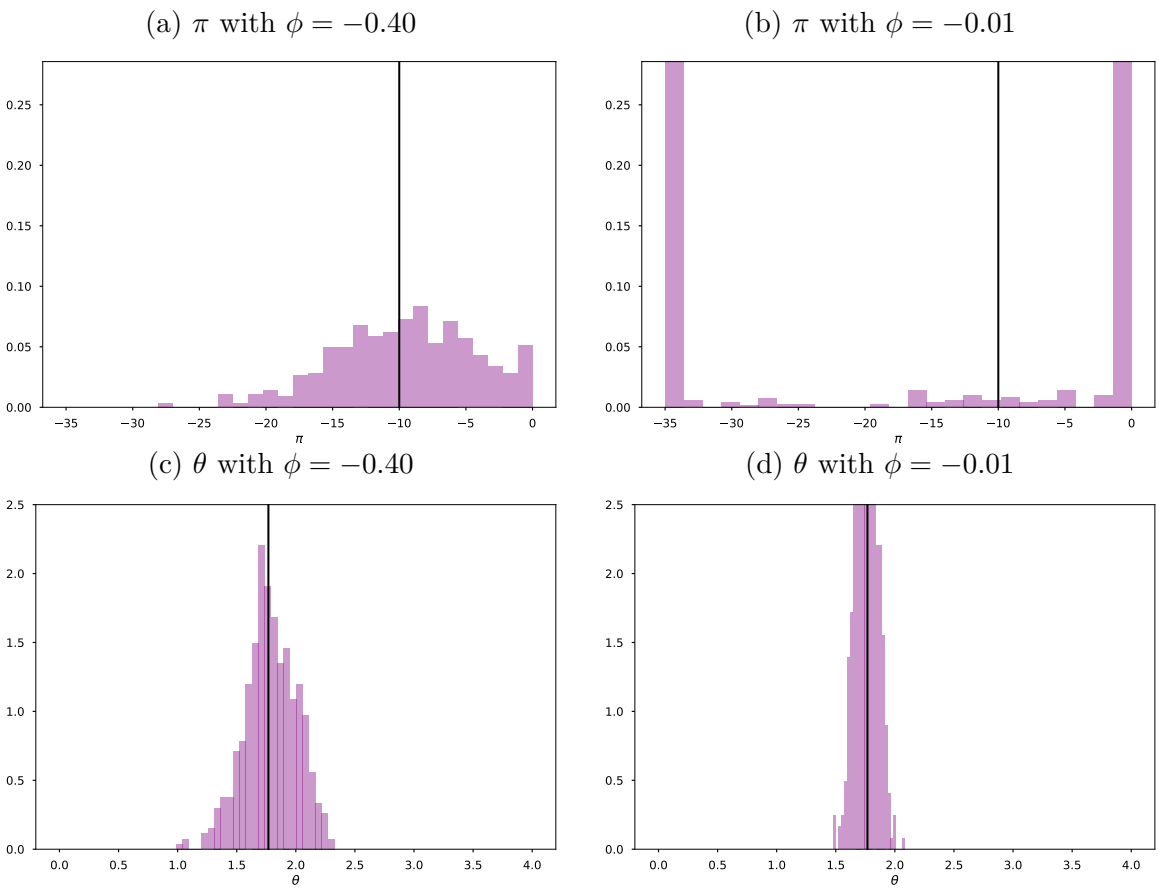
δ	ρ	c	π	κ
Parameter Values used by Han, Khrapov, and Renault (2018)				
0.6475	0.50	3.94×10^{-3}	-10	1.7680

To avoid boundary issues with respect to the estimate of c and δ in finite-sample, we reparameterize the moment conditions and link functions in terms of $\log(c)$, $\log(c)+\log(\delta)$, and $\text{logit}(\rho)$. This reparameterization forces the scale parameters to be positive and ρ to lie in $(0,1)$. We find that the resulting estimates for the transformed reduced-form parameters are better approximated by the Gaussian distribution for a given finite sample.

To show the effect of various identification strength, we first vary the true value of ϕ and plot the distribution of $\widehat{\pi}$ and $\widehat{\theta}$ in Figure 4.1. The reported result is based on $10,000$ observations and 500 simulation repetitions. The black lines in the middle of the figures are the true parameter values. Clearly, the estimators sometimes pile up at the boundaries of the parameter space. As expected, this simulation shows that the Gaussian distribution is not a good approximation for the finite-sample distribution of either of the estimators.

Next, we study the finite-sample size of in the standard QLR test and the proposed conditional QLR test for a joint test for the three structural parameters. The nominal level of the test is 5 %. The critical value of the standard QLR test is the 95 % quantile of the χ^2-distribution with 3 degree of freedom. The critical value of the conditional QLR test is obtained by the stimulation-based procedure in Algorithm 4.1, with 250

Figure 4.1: Parameter Estimates' t-Statistics



simulation repetitions to approximate the quantile of the conditional distribution. The finite-sample size is based on 250 simulation repetitions.

The standard test is no longer valid under weak identification because the QLR statistic does not have a χ^2-distribution in this case. However, it is not clear whether the standard QLR test under-rejects or over-rejects in finite-sample and how large is the difference from 5 %.

Simulation results show that the standard QLR test under-rejects in finite-sample. This is most severe when the identification is weak, e.g., for $\phi = -0.01$ and $T = 10,000$, the rejection rate is 1.60 %. If we have enough data and ϕ is large enough in magnitude, the standard test static does okay. However, this is not the empirically relevant case. The proposed test, however, has approximately uniform coverage and, hence, is much more trustworthy.

Table 4.2: Finite-Sample Size of the Standard and Proposed Tests

	Standard %	Proposed %	Standard %	Proposed %
ϕ	$T = 2{,}000$		$T = 10{,}000$	
-0.01	2.00	5.20	1.60	4.40
-0.40	2.40	5.60	6.00	6.40

4.6 Data and Empirical Results

For the empirical application, we use the daily return on the S&P 500 for r_{t+1} and the associated realized volatility computed with high-frequency data for σ^2_{t+1}. The data is obtained from SPY (SPDR S&P 500 ETF Trust), an exchange-traded fund that mimics the S&P 500. This gives us a market index whose risk is not easily diversifiable and can be used to estimate the prices of risk that investors face in practice. We use the procedure Sangrey (2018) develops to estimate the integrated total volatility, i.e, the instantaneous expectation of the price variance. This measure reduces to the integrated diffusion volatility if prices have continuous paths and it works well in the presence of market microstructure noise.

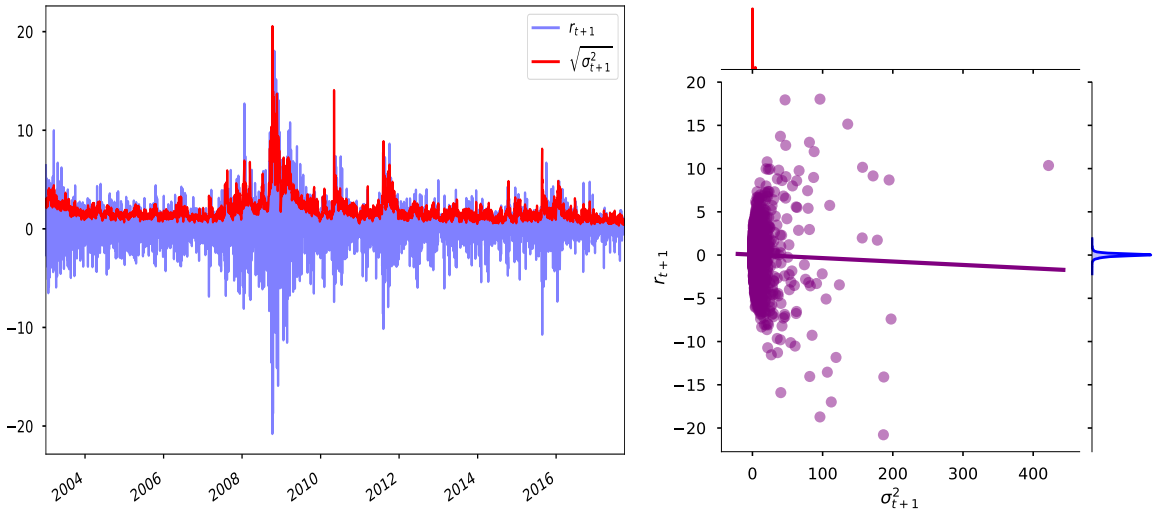
Since SPY is one of the most liquid assets traded, we can choose the frequency at which we sample the underlying price. To balance market-microstructure noise, computational cost, and efficiency of the resultant estimators, we sample at the 1-second frequency. We annualize the data by multiplying r_{t+1} by 252 and σ^2_{t+1} by 252^2. The data starts in 2003 and ends in September 2017. Since the asset is only traded during business hours, this leads to 3713 days of data with an average of approximately 24 000 observations per day. We compute r_{t+1} as the daily return from the open to the close of the market, the interval over which we can estimate the volatility. This avoids specifying the relationship between overnight and intra-day returns. We preprocess the data using the pre-averaging approach as in Podolskij and Vetter (2009) and Aït-Sahalia, Jacod, and Li (2012).

To see how the data move over time, we plot their time series in Figure 4.2. We also plot the joint unconditional distribution in Figure 4.2 to see the static relationship between the two series. The volatility has a long-right tail, a typical gamma-type distribution. The returns has a bell-shaped distribution. They are slightly negatively

Figure 4.2: S&P 500 Volatility and Log-Return

(a) Time Series

(b) Joint Distribution



correlated, as shown by the regression line in the joint plot. This corroborates the work by Bandi and Renò (2012) and Aït-Sahalia, Fan, and Li (2013). We also report a series of summary statistics.

Table 4.3: Summary Statistics

	r_{t+1}	σ^2_{t+1}
Mean	0.02	5.62
Standard Deviation	2.35	14.46
Skewness	−0.31	12.21
Kurtosis	13.07	243.40
Correlation	−0.02	

We now report the estimates and confidence intervals for the reduced-form parameters c, δ, and ρ. The confidence intervals reported here use the Gaussian limiting theory, i.e., the point estimates ± 1.96 standard errors. We first obtain confidence intervals for $\log(c)$ and $\log(c) + \log(\delta)$, and transform them into confidence intervals for c and δ. Similarly, we create the confidence interval for ρ by inverting the interval for $\text{logit}(\rho)$.

For confidence intervals of the three structural parameters, we first compute their

Table 4.4: Parameters that Govern the Volatility Process

	Point Estimate	95 % Confidence Interval
c	3.07	(1.38, 6.79)
δ	37.98	(17.65, 81.72)
ρ	0.77	(0.67, 0.85)

joint confidence set based on the conditional QLR test and then project it to each of the components. We also plot a joint confidence sets for the two risk prices, after projecting out ϕ. We use 500 simulations to compute the quantile for the QLR statistic.

Table 4.5: Structural Parameters

	95 % Confidence Interval
ϕ	(-0.33, -0.27)
π	(-30.97, 0.00)
κ	(0.00, 2.00)

The results in Table 4.5. have a few notable features. First, we can reject the null hypothesis $\phi = 0$. We cannot, however, reject the hypothesis that $\pi = 0$ at the 5 % level. We cannot reject the null hypotheses that $\kappa = 0$. This should not be particularly surprising given the difficulty in precisely estimating this parameter documented in the previous literature, (Lettau and Ludvigson 2010). Although not recorded on the table, we can reject the hypothesis that both $\kappa = \pi = 0$. The procedure can determine that investors demand compensation for risk, just not what combination of risks they demand compensation for. The risk price associated with the market return is covered by $(0.00, 2.00)$ with at least 95 % probability. The volatility risk price is covered by $(-30.97, 0.00))$ with at least 95 % probability. Our confidence intervals for both parameters are reasonable given other values that previous authors have found. For example, Han, Khrapov, and Renault (2018) preferred a value of $\pi = -10$, for example.

4.7 Conclusion

In structural stochastic volatility models as the one considered here, changes in the volatility affect returns through two channels. On the one hand, investors are willing to tolerate high volatility to get high expected returns as measured by the price of market return risk. On the other hand, investors are directly averse to changes in future volatility, as measured by the price of volatility risk. Han, Khrapov, and Renault (2018) shows how to disentangle these two channels by exploiting information arising from the leverage effect in an exponentially-affine pricing model. However, standard inference for this structural model is invalid because the volatility risk price is only weakly identified when the leverage effect is mild. This paper propose an identification robust inference procedure that provides reliable confidence sets for the risk prices regardless of the magnitude of the leverage effect. We take this procedure to the data on the S&P 500. The robust inference procedure provides reliable yet informative confidence intervals for the risk prices associated with the market return and the volatility.

References

Aït-Sahalia, Yacine, Jianqing Fan, and Yingying Li. 2013. "The Leverage Effect Puzzle: Disentangling Sources of Bias at High Frequency." *Journal of Financial Economics* 109 (1): 224–249.

Aït-Sahalia, Yacine, Jean Jacod, and Jia Li. 2012. "Testing for Jumps in Noisy High Frequency Data." *Journal of Econometrics* 168:207–222.

Andrews, Donald W.K., and Xu Cheng. 2012. "Estimation and Inference With Weak, Semi-Strong, and Strong Identification." *Econometrica* 80 (5): 2153–2211.

Andrews, Donald W.K., and Patrik Guggenberger. 2010. "Asymptotic Size and a Problem with Subsampling and with the M out of N Bootstrap." *Econometric Theory* 26 (2): 426–468.

Andrews, Isaiah, and Anna Mikusheva. 2015. "Maximum Likelihood Inference in Weakly Identified Dynamic Stochastic General Equilibrium Models." *Quantitative Economics* 6 (1): 123–152.

———. 2016. "Conditional Inference With a Functional Nuisance Parameter." *Econometrica* 84 (4): 1571–1612.

Bali, Turan G., and Lin Peng. 2006. "Is There a Risk–Return Trade-off? Evidence from High-Frequency Data." *Journal of Applied Econometrics* 21 (8): 1169–1198.

Bandi, Federico M., and Roberto Renò. 2012. "Time-varying Leverage Effects." *Journal of Econometrics* 169 (1): 94–113.

Bansal, Ravi, Dana Kiku, Ivan Shaliastovich, and Amir Yaron. 2014. "Volatility, the Macroeconomy, and Asset Prices." *The Journal of Finance* 69 (6): 2471–2511.

Bollerslev, Tim, Robert F. Engle, and Jeffrey M. Wooldridge. 1988. *Journal of Political Economy* 96 (1): 116–131.

Bollerslev, Tim, Tzuo Hann Law, and George Tauchen. 2008. "Risk, Jumps and Diversification." *Journal of Econometrics* 144:234–256.

Brandt, Michael W., and Qiang Kang. 2004. "On the Relationship Between the Conditional Mean and Volatility of Stock Returns: A Latent VAR Approach." *Journal of Financial Economics* 72 (2): 217–257.

Breen, William, Lawrence R. Glosten, and Ravi Jagannathan. 1989. "Economic Significance of Predictable Variations in Stock Index Returns." *The Journal of Finance* 44 (5): 1177–1189.

Campbell, John Y. 1987. "Stock Returns and the Term Structure." *Journal of Financial Economics* 18 (2): 373–399.

Christoffersen, Peter, Steven Heston, and Kris Jacobs. 2013. "Capturing Option Anomalies with a Variance-Dependent Pricing Kernel." *The Review of Financial Studies* 26 (8): 1963–2006.

Darolles, Serge, Christian Gouriéroux, and Joann Jasiak. 2006. "Structural Laplace Transform and Compound Autoregressive Models." *Journal of Time Series Analysis* 27 (4): 477–503.

Dew-Becker, Ian, Stefano Giglio, Anh Le, and Marius Rodriguez. 2017. "The Price of Variance Risk." *Journal of Financial Economics* 123 (2): 225–250.

Drechsler, Itamar. 2013. "Uncertainty, Time-Varying Fear, and Asset Prices." *The Journal of Finance* 68 (5): 1843–1889.

Drechsler, Itamar, and Amir Yaron. 2011. "What's Vol Got to Do with It." *The Review of Financial Studies* 24 (1): 1–45.

Feunou, Bruno, Jean-Sébastien Fontaine, Abderrahim Taamouti, and Roméo Tédongap. 2014. "Risk Premium, Variance Premium, and the Maturity Structure of Uncertainty." *Review of Finance* 18 (1): 219–269.

Ghysels, Eric, Pedro Santa-Clara, and Rossen Valkanov. 2005. "There is a Risk-Return Trade-off after All." *Journal of Financial Economics* 76 (3): 509–548.

Gouriéroux, Christian, and Joann Jasiak. 2006. "Autoregressive Gamma Processes." *Journal of Forecasting* 25 (2): 129–152.

Han, Hyojin, Stanislav Khrapov, and Eric Renault. 2018. *The Leverage Effect Puzzle Revisited: Identification in Discrete Time.* Working Paper. Brown University.

Harvey, Campbell R. 1989. "Time-Varying Conditional Covariances in Tests of Asset Pricing Models." *Journal of Financial Economics* 24 (2): 289–317.

Heston, Steven L. 1993. "A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options." *The Review of Financial Studies* 6 (2): 327–343.

Kleibergen, Frank. 2005. "Testing Parameters in GMM Without Assuming That They Are Identified." *Econometrica* 73 (4): 1103–1123.

Lettau, Martin, and Sydney C. Ludvigson. 2010. "Measuring and Modeling Variation in the Risk-Return Tradeoff." Chap. 11, edited by Yacine Aït-Sahalia and Lars Peter Hansen, 1:617–690. Handbook of Financial Econometrics. Elsevier.

Lintner, John. 1965. "Security Prices, Risk, and Maximal Gains From Diversification." *The Journal of Finance* 20 (4): 587–615.

Ludvigson, Sydney C., and Serena Ng. 2007. "The Empirical Risk–Return Relation: A Factor Analysis Approach." *Journal of Financial Economics* 83 (1): 171–222.

Magnusson, Leandro M. 2010. "Inference in Limited Dependent Variable Models Robust to Weak Identification." *The Econometrics Journal* 13, no. 3 (September): S56–S79.

Magnusson, Leandro M., and Sophocles Mavroeidis. 2010. "Identification-Robust Minimum Distance Estimation of the New Keynesian Phillips Curve." *Journal of Money, Credit and Banking* 42 (2/3): 465–481.

Mavroeidis, Sophocles, Mikkel Plagborg-Møller, and James H. Stock. 2014. "Empirical Evidence on Inflation Expectations in the New Keynesian Phillips Curve." *Journal of Economic Literature* 52, no. 1 (March): 124–88.

Mikusheva, Anna. 2007. "Uniform Inference in Autoregressive Models." *Econometrica* 75 (5): 1411–1452.

Moreira, Marcelo J. 2003. "A Conditional Likelihood Ratio Test for Structural Models." *Econometrica* 71 (4): 1027–1048.

Pagan, Adrian R., and Y.S. Hong. 1991. "Nonparametric Estimation and the Risk Premium." Chap. 2 in *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics,* edited by William A. Barnett, James Powell, and George Tauchen, 51–75. Cambridge University Press.

Podolskij, Mark, and Mathias Vetter. 2009. "Bipower-type Estimation in a Noisy Diffusion Setting." *Stochastic Processes and Their Applications* 119 (9): 2803–2831.

Sangrey, Paul. 2018. *Jumps, Realized Densities, and News Premia.* Working Paper. University of Pennsylvania.

Sharpe, William F. 1964. "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk." *The Journal of Finance* 19 (3): 425–442.

Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65 (3): 557–586.

Stock, James H., and Jonathan H. Wright. 2000. "GMM with Weak Identification." *Econometrica* 68 (5): 1055–1096.

Whitelaw, Robert F. 1994. "Time Variations and Covariations in the Expectation and Volatility of Stock Market Returns." *The Journal of Finance* 49 (2): 515–541.

4.A Proofs

Proof of Lemma 4.1

Proof. For the risk-free asset, the excess return $r_{t+1} = 0$. Therefore, we have

$$\begin{aligned}
1 &= E\left[\exp\left(m_0 + m_1\sigma_t^2 - \pi\sigma_{t+1}^2 - \theta r_{t+1}\right) \middle| \mathcal{F}_t\right] \\
&= \exp(m_0 + m_1\sigma_t) E\left[\exp\left(-\pi\sigma_{t+1}^2\right) E\left[\exp\left(-\theta r_{t+1}\right) \middle| \mathcal{F}_t, \sigma_{t+1}^2\right] \middle| \mathcal{F}_t\right] \\
&= \exp(m_0 - E\left(\theta\right) + m_1\sigma_t - D\left(\theta\right)\sigma_t^2) E\left[\exp\left(-\pi\sigma_{t+1}^2 - C\left(\theta\right)\sigma_{t+1}^2\right) \middle| \mathcal{F}_t\right] \\
&= \exp(m_0 - E\left(\theta\right) + m_1\sigma_t - D\left(\theta\right)\sigma_t^2 - A\left(\pi + C\left(\theta\right)\right)\sigma_t^2 - B\left(\pi + C\left(\theta\right)\right)),
\end{aligned}$$

where the first equality follows from the pricing equation, the second equality follows from the law of iterated expectations, the third equation uses the Laplace transform for r_{t+1} in (4.9), and the last equality follows from the Laplace transform for σ_{t+1}^2 in (4.4). Since $M_{t,t+1}$ must integrate to 1, the constant term and coefficient for σ_t^2 must equal 0, which gives the claimed result for m_0 and m_1.

We can apply the same argument above to any asset r_{t+1}. This gives the same result, except θ is replaced by $\theta - 1$ throughout. This implies that the two equalities for m_0 and m_1 also hold with θ replaced by $\theta - 1$. Therefore,

$$\begin{aligned}
E(\theta - 1) + B\left(C\left(\theta - 1\right) + \pi\right) &= E(\theta) + B\left(C\left(\theta\right) + \pi\right), \\
D\left(\theta - 1\right) + A\left(C\left(\theta - 1\right) + \pi\right) &= D\left(\theta\right) + A\left(C\left(\theta\right) + \pi\right).
\end{aligned}$$

The claimed results for γ and β follow from $\gamma = \mathbb{E}(\theta) - E(\theta-1)$ and $\beta = D(\theta) - D(\theta-1)$ under the linear specification of $E(x) = \gamma x$ and $D(x) = \beta x$.

□

Proof of Lemma 4.2

Proof. Under the assumption that (i) $\mathbb{E}(z_t z_t')$ has the smallest eigenvalue bounded away from 0 and (ii) $c > \varepsilon$ and $\delta > \varepsilon$ for some $\varepsilon > 0$, we not only have ω_{10} as an unique minimizer of $||\mathbb{E}[h_t(\omega_1)]||$ but also have a uniform positive lower bound for $\|\mathbb{E}[h_t(\omega_1)]\|$ for $\|\omega_1 - \omega_{10}\| \geq \varepsilon$. Thus, consistency of $\widehat{\omega}_1$ follows from standard arguments for the consistency of a GMM estimator under an uniform convergence of the criterion under Assumption R (1) and (2).

Let $\overline{h}(\omega_1) = T^{-1}\sum_{t=1}^{T} h_t(\omega_1)$ and $\overline{H}(\omega) = T^{-1}\sum_{t=1}^{T} H_t(\omega_1)$. By construction, the estimator satisfies the first order condition

$$\begin{aligned} 0 &= \begin{pmatrix} \overline{H}(\widehat{\omega}_1)'\widehat{V}_1^{-1}\overline{h}(\widehat{\omega}_1) \\ T^{-1}\sum_{T=1}^{T} x_t(y_t - x_t'\widehat{\omega}_2) \\ \widehat{\omega}_3 - T^{-1}\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 \end{pmatrix} \\ &= \begin{pmatrix} \overline{H}(\widehat{\omega}_1)'\widehat{V}_1^{-1}\overline{h}(\omega_{10}) + \overline{H}(\widehat{\omega}_1)'\widehat{V}_1^{-1}\overline{H}(\widetilde{\omega}_1)(\widehat{\omega}_1 - \omega_{10}) \\ T^{-1}\sum_{t=1}^{T} x_t(y_t - x_t'\omega_{20}) - T^{-1}\sum_{t=1}^{T} x_t x_t'(\widehat{\omega}_2 - \omega_{20}) \\ (\widehat{\omega}_3 - \omega_3) + \omega_3 - T^{-1}\sum_{t=1}^{T}(y_t - x_t\widehat{\omega}_2)^2 \end{pmatrix}, \end{aligned} \tag{4.30}$$

where the second equality follows from a mean value expansion of $\overline{h}(\widehat{\omega}_1)$ around ω_{10}, with $\widetilde{\omega}_1$ between ω_{10} and $\widehat{\omega}_1$. Let

$$\widetilde{\mathcal{B}} = \text{diag}\left\{ [\overline{H}(\widehat{\omega}_1)'\widehat{V}_1^{-1}\overline{H}(\widetilde{\omega}_1)]^{-1}\overline{H}(\widehat{\omega}_1)'\widehat{V}_1^{-1}, [T^{-1}\sum_{t=1}^{T} x_t x_t']^{-1}, 1 \right\}. \tag{4.31}$$

Then (4.30) implies that

$$\begin{aligned} T^{1/2}(\widehat{\omega} - \omega) &= \widetilde{\mathcal{B}} \cdot T^{-1/2}\sum_{t=1}^{T} \begin{pmatrix} -h_t(\omega_{10}) \\ x_t(y_t - x_t'\omega_{20}) \\ (y_t - x_t\widehat{\omega}_2)^2 - \omega_3 \end{pmatrix} \\ &= \widetilde{\mathcal{B}} \cdot T^{-1/2}\sum_{t=1}^{T} \begin{pmatrix} -h_t(\omega_{10}) \\ x_t(y_t - x_t'\omega_{20}) \\ (y_t - x_t'\omega_{20})^2 - \mathbb{E}\left[(y_t - x_t'\omega_{20})^2\right] \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \varepsilon_T \end{pmatrix}, \end{aligned} \tag{4.32}$$

where the second equality uses $\omega_3 = \mathbb{E}[(y_t - x_t'\omega_{20})^2]$ by definition and

$$\begin{aligned}\varepsilon_T &= T^{-1/2}\sum_{t=1}^{T}\left[(y_t - x_t'\widehat{\omega}_2)^2 - (y_t - x_t'\omega_{20})^2\right] \\ &= 2T^{-1}\sum_{t=1}^{T}(y_t - x_t'\omega_{20})\,x_t'\left[T^{1/2}\left(\widehat{\omega}_2 - \omega_{20}\right)\right] + o_p(1) \\ &= o_p(1) \end{aligned} \tag{4.33}$$

because $T^{-1}\sum_{t=1}^{T}(y_t - x_t'\omega_{20})\,x_t' \to_p 0$ and $T^{1/2}(\widehat{\omega}_2 - \omega_{20}) = O_p(1)$ following Assumption R. In addition,

$$\widetilde{\mathcal{B}} \to_p \mathcal{B} \tag{4.34}$$

following from the consistency of $\widehat{\omega}_1$ and Assumption R. Finally, the desirable result follows from (4.32)–(4.34) and Assumption R. The consistency of $\widehat{\Omega}$ follows from the consistency of $\widehat{\mathcal{B}}$ and $\widehat{V}$.

□

Proof of Theorem 4.3

Proof. We obtain this result by applying Andrews and Mikusheva (2016, Theorem 1). We now verify Assumptions 1–3 in Andrews and Mikusheva (2016). To show weak convergence $\eta_T(\cdot)$ to $\eta(\cdot)$ uniformly over $\mathcal{P}$, note that by a second-order Taylor expansion,

$$\begin{aligned}&\eta_T(\lambda) \coloneqq T^{1/2}\left[\widehat{g}(\lambda) - g_0(\lambda)\right] = G_0(\lambda)\Omega^{1/2}\xi_T + \delta_T, \text{ where} \\ &\quad \xi_T = \Omega^{-1/2}T^{1/2}\left(\widehat{\omega} - \omega_0\right), \text{ and } \delta_T = \left(G(\lambda,\widetilde{\omega}) - G(\lambda,\omega_0)\right)T^{1/2}(\widehat{\omega} - \omega_0)\end{aligned} \tag{4.35}$$

and $\widetilde{\omega}$ is between $\widehat{\omega}$ and ω_0. Because $\|G(\lambda,\widetilde{\omega}) - G(\lambda,\omega_0)\| \leq C\|\widetilde{\omega} - \omega_0\|$, $\delta_T = o_p(1)$ uniformly over $\mathcal{P}$ following Lemma Lemma 4.2. To show $G_0(\lambda)\Omega^{1/2}\xi_T$ weakly converges to $\eta(\cdot)$, it is sufficient to show (i) the pointwise convergence

$$\begin{pmatrix} G_0(\lambda_1)\Omega^{1/2}\xi_T \\ G_0(\lambda_2)\Omega^{1/2}\xi_T \end{pmatrix} \to_d \left(\begin{pmatrix} \eta(\lambda_1) \\ \eta(\lambda_2) \end{pmatrix}, \right. \tag{4.36}$$

which follows from Lemma 4.2, and (ii) the stochastic equicontinuity condition, i.e., for every $\varepsilon > 0$ and $\xi > 0$, there exists a $\delta > 0$ such that

$$\limsup_{T\to\infty} \Pr\left(\sup_{P\in\mathcal{P}} \sup_{\|\lambda_1-\lambda_2\|\leq\delta} \left\| G_0(\lambda_1)\Omega^{1/2}\xi_T - G_0(\lambda_2)\Omega^{1/2}\xi_T \right\| > \varepsilon \right) < \xi. \tag{4.37}$$

For some $C < \infty$, we have $\|G_0(\lambda_1) - G(\lambda_2)\| \leq C\|\lambda_1 - \lambda_2\|$ under a uniform bound for the derivative in Assumption S, and we have $\|\Omega^{1/2}\| \leq C$ under Assumption R because F and V both have bounded largest eigenvalue. Thus,

$$\begin{aligned} &\limsup_{T\to\infty} \Pr\left(\sup_{P\in\mathcal{P}} \sup_{\|\lambda_1-\lambda_2\|\leq\delta} \left\| G_0(\lambda_1)\Omega^{1/2}\xi_T - G_0(\lambda_2)\Omega^{1/2}\xi_T \right\| > \varepsilon \right) \\ &\leq \limsup_{T\to\infty} \Pr\left(C^2 \sup_{P\in\mathcal{P}} \|\xi_T\| > \frac{\varepsilon}{\delta} \right). \end{aligned} \tag{4.38}$$

Because $\xi_T = O_p(1)$ uniformly over $P \in \mathcal{P}$, there exists δ such that ε/δ is large enough to make the right hand side of the inequality in (4.38) smaller than ξ.

Assumptions 2 and 3 of Andrews and Mikusheva (2016, Theorem 1) follow from Assumption R. □