CIRJE-F-678

Computing Densities: A Conditional Monte Carlo Estimator

Richard Anton Braun
University of Tokyo

Huiyu Li
Graduate School of Economics, University of Tokyo

John Stachurski
Kyoto University

October 2009

Computing Densities: A Conditional Monte Carlo Estimator*

Richard Anton Braun † Huiyu Li ‡ John Stachurski §

October 12, 2009

Abstract

We propose a generalized conditional Monte Carlo technique for computing densities in economic models. Global consistency and functional asymptotic normality are established under ergodicity assumptions on the simulated process. The asymptotic normality result allows us to characterize the asymptotic distribution of the error in density space, and implies faster convergence than nonparametric kernel density estimators. We show that our results nest several other well-known density estimators, and illustrate potential applications.

Keywords: Distributions, numerical methods, simulation
JEL Classification Codes: C15, C63

1 Introduction

The Monte Carlo method is routinely used by economists and econometricians to extract information on probabilities from their models. In some cases, the random

*This paper has benefitted from the insights of participants at the 15th International Conference on Computing in Economics and Finance, Sydney 2009, the 4th Workshop on Macroeconomic Dynamics, Singapore 2009, the 2009 Society for Economic Dynamics Annual Meeting in Istanbul, and the 2009 Far East and South Asia Meeting of the Econometric Society in Tokyo.

†Faculty of Economics, The University of Tokyo email:toni@e.u-tokyo.ac.jp

‡Faculty of Economics, The University of Tokyo email:tohuiyu@gmail.com

§Institute of Economic Research, Kyoto University email: john@kier.kyoto-u.ac.jp

variables of interest have distributions that can be described by densities, and the researcher seeks to recover, via simulation, an approximation to these densities. In these cases, after the relevant variables have been simulated, there remains the problem of how to produce density estimates from the simulated observations. This paper describes an efficient solution to that problem, based on the use of conditioning information.

The Monte Carlo density estimation problem arises frequently in econometric modelling, such as when Markov chain Monte Carlo is used to calculate posterior densities, or simulated maximum likelihood is used to fit a models where the likelihood function is intractable (see, e.g., Danielsson, 1994; Brandt and Santa-Clara, 2002; Durham and Gallant, 2002; or Chib, Nardari and Shepard, 2002).[1] Simulation of densities also arises in a wide range of economic applications. For example, decision makers compute posterior densities of model parameters when making Bayesian or robust control decisions in real time (e.g., Hansen and Sargent, 2007). Additionally, densities are computed when solving for equilibria in a variety of economic models. For instance, simulations methods are often used to compute the density of wealth in heterogeneous agent economies.

While the random variables in these models can usually be simulated, the estimation problem associated with recovering densities from the simulated observations is nontrivial. In general, parametric estimation techniques do not apply. Instead, many economists use methods such as nonparametric kernel density estimation, which converges to the density of any random variable given sufficient observations. These nonparametric kernel methods typically use no information beyond the Monte Carlo sample (and some notion of smoothness depending on the choice of kernel).

However, for this class of estimation problems, more information is available than just the sample itself. The researcher also has at hand the model that was used to generate the sample, and some aspects of that model may be tractable. If so, one might be able to exploit this information to obtain a more efficient estimator. For example, in many models, some *conditional* densities relating different variables to one another are either known or inexpensive to compute. Moreover, unconditional densities can be expressed as averages of conditional densities. When estimating the unconditional densities, the information provided by

[1] Many other examples of Monte Carlo density estimation can be found in econometrics. For example, Diebold and Schuerman (1996) describe a Monte Carlo approach for computing the density of initial observations when conducting exact (i.e., unconditional) maximum likelihood. Poon and Granger (2003) compute forecast densities to assess alternative models of conditional volatility.

conditional densities can be exploited by averaging over these densities across simulated realizations of the conditioning variable.

This idea has appeared in the literature under different names in a number of specialized settings. For example, Santa-Clara (1995) and Pedersen (1995) independently proposed a method of simulated maximum likelihood based on this conditioning method. Other examples include the density estimator based on Gibbs sampling proposed in the famous paper of Gelfand and Smith (1990), and the look-ahead estimator of Glynn and Henderson (2001), used for computing marginal state densities of discrete time Markov processes.[2]

The objective of this paper is to connect these separate ideas and expand the range of potential applications by providing a general theory of conditional Monte Carlo density estimation.[3] To this end, we pose the estimation problem in a general setting, and prove global consistency and a functional central limit theorem. In order to accommodate standard economic and econometric environments, these results are established without requiring that the sampled data be independent, nor identically distributed. Instead, global consistency is proved using ergodicity, while the central limit theorem requires so-called V-uniform ergodicity and a restriction on second conditional moments.

The assumption of V-uniform ergodicity is weaker than the classical uniform ergodicity condition used in many studies of asymptotic normality in the Markov setting. For example, the stationary linear AR(1) model with Gaussian shocks is V-uniformly ergodic but not uniformly ergodic. Nishimura and Stachurski (2005) establish that the standard Brock–Mirman model satisfies V-uniform ergodicity. In the econometric setting, Kristensen (2008) establishes specific conditions under which V-uniform ergodicity holds for a broad range of linear and nonlinear time series models. More generally, Meyn and Tweedie (2009) provide a range of sufficient conditions under which V-uniform ergodicity is known to hold.

The conditional Monte Carlo density estimator (CMCDE) we study is unbiased,

[2]The idea of averaging over conditional distributions is used to compute not only densities but also expectations. This procedure is referred to as conditional Monte Carlo, or Rao-Blackwellization. We also consider the problem of computing expectations in a related paper (Braun, Li and Stachurski, 2009).

[3]We focus on densities rather than distributions, because density estimates are typically more useful than estimates of cumulative distribution functions or probability measures. One reason is that the process of computing distributions from estimated densities is numerically stable, while that of computing densities from empirical distributions is typically ill-posed. (In essence, this is because the first operation involves integration, while the second requires differentiation.) Second, densities play a central role in many applications, such as those involving maximum likelihood or Bayesian statistics. See Braun, Li and Stachurski (2009) for an application of similar ideas to the problem of computing expectations.

and exhibits a parametric rate of convergence independent of the dimension of the state space. In contrast, nonparametric kernel density estimators (NPKDEs), which make no use of conditioning information, converge at a rate that is slower than the parametric rate, and sensitive to the dimension of the state space. In numerical experiments, we also found excellent small sample properties. For example, in numerical simulations of a GARCH(1,1) model, we report L_1 norm errors for the CMCDE up to 42% lower than a Gaussian NPKDE.

The generality of our convergence results open up new avenues for exploiting existing ideas on conditional Monte Carlo. For example, as noted above, the CMCDE nests the Gibbs sampling technique for computing marginal densities proposed by Gelfand and Smith (1990). It also provides an asymptotic justification for Monte Carlo sampling schemes that relax their IID structure. In particular, our asymptotic theory provides a justification for computing the unconditional marginal density from a single time series draw, instead of n independent paths generated by the Gibbs sampler.

The CMCDE can be applied to estimate densities in both stationary and nonstationary environments. For instance, a cross-sectional version of the CMCDE can be used to compute t-step ahead distributions in settings where these distributions are far from the stationary distribution (as in, e.g., the Brock-Mirman model with a low initial capital stock), or in environments where there is no well-defined asymptotic distribution (such as a unit root process).

The remainder of the paper is organized as follows. In Section 2 we define the general CMCDE and derive its asymptotic properties. Section 3 relates the CMCDE to the previous literature. In Section 4 we discuss a number of potential new applications. Section 5 concludes.

2 Conditional Monte Carlo Density Estimation

In this section, we begin by providing an intuitive motivation for the conditional Monte Carlo density estimation method. We then state definitions and give a formal treatment of the problem.

2.1 Motivation

Let ψ be a density on some set $\mathbb{Y}$ that is defined by a given model. For example, ψ might represent a posterior density given prior and conditional densities,

it might represent a cross-sectional distribution of asset holdings in a macroeconomic model, or it might represent a long run marginal density of returns in a stochastic volatility model. Even if the model itself is analytically tractable, the density ψ may not be. And if the model itself is not tractable, then ψ (which is defined in terms of the model) is almost certainly intractable. Thus, we consider the numerical problem of computing an approximation ψ_n to the density ψ.

One simulation-based approach to computing this approximation can be described as follows. Suppose that the model in question defines a second random variable X such that X is correlated with Y, and that the conditional density of Y given $X = x$ is $q(y \mid x)$. By definition, the density ψ of Y and the conditional density q must satisfy[4]

$$\psi(y) = \mathbb{E}\, q(y \mid X) \qquad \text{for all } \; y \in \mathbb{Y} \tag{1}$$

This equation suggests the following strategy: sample from the distribution of X, use the conditional density q of Y given X to infer the probabilities for Y, and, in this way, build an approximation ψ_n to ψ. In particular, suppose that we are able to generate IID draws $X_1, \ldots, X_n$ of the random variable X, and that q can be evaluated numerically. Then an estimator of ψ is provided by the (random) density

$$\psi_n(y) = \frac{1}{n} \sum_{i=1}^{n} q(y \mid X_i) \tag{2}$$

We refer to this estimator as the conditional Monte Carlo density estimator (CMCDE). The pointwise properties of the CMCDE are readily apparent. Fixing $y \in \mathbb{Y}$, the strong law of large numbers yields

$$\lim_{n\to\infty} \psi_n(y) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} q(y \mid X_i) = \mathbb{E}\, q(y \mid X) = \psi(y) \qquad \mathbb{P}\text{-a.s.} \tag{3}$$

where the last equality is due to (1). Hence $\psi_n(y)$ is strongly consistent for $\psi(y)$. A simple calculation along the same lines shows that $\psi_n(y)$ is also unbiased (i.e., $\mathbb{E}\psi_n(y) = \psi(y)$). Furthermore, if the second moment condition $\mathbb{E}q(y \mid X)^2 < \infty$ is satisfied, then, by the central limit theorem,

$$\sqrt{n}\{\psi_n(y) - \psi(y)\} = \sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^{n} q(y \mid X_i) - \mathbb{E}\, q(y \mid X) \right\}$$

converges in distribution to a centered Gaussian on $\mathbb{R}$. An immediate corollary is that $|\psi_n(y) - \psi(y)| = O_P(n^{-1/2})$. Hence the CMCDE attains the parametric rate of convergence.

[4]Informally, this equation can be regarded as a density equivalent of iterated expectations.

All of the results we have presented so far are pointwise, pertaining to convergence of $\psi_n(y)$ to $\psi(y)$ at some fixed $y \in \mathbb{Y}$. However, in the case of densities, convergence at a single point conveys relatively little information, since densites can be altered at individual points while still representing the same distribution.[5] Moreover, if $\mathbb{Y}$ is uncountable, then (3) does not imply that $\psi_n \to \psi$ pointwise with probability one, since the measure zero set on which (3) fails is indexed by y, and uncountable families of null sets are not generally null.

A more informative measure of deviation between ψ_n and ψ is provided by the $L_p(\mu)$-norm distance. In order to study norm deviation, we view ψ_n as an $L_p(\mu)$-valued random variable. This functional setting also permits us to derive the asymptotic distribution of the error $\psi_n - \psi$. The details are presented in section 2.3.

A further goal of our analysis is to accomodate non-IID environments. To illustrate why this is desirable in economic and econometric settings, consider a stochastic volatility model of the form $r_t = \sigma_t W_t$, where r_t is a measure of returns on investment, σ_t is a stochastic volatility term, and W_t is a standard normal shock independent of σ_t. Let ψ be the (unconditional) density of r_t. The conditional density of r_t given σ_t is $q(r \mid \sigma_t) = N(0, \sigma_t)$, and by (1) we have

$$\psi(r) = \mathbb{E}q(r \mid \sigma_t) = \mathbb{E}\frac{1}{(2\pi)^{1/2}\sigma_t} \exp\left\{-\frac{y^2}{2\sigma_t^2}\right\} \tag{4}$$

Suppose our goal is to compute ψ for a given specification of the process generating σ_t. If σ_t is specified as IID from a fixed distribution, then we can use the CMCDE $\psi_n(r) := n^{-1}\sum_{t=1}^n q(r \mid \sigma_t)$, where $(\sigma_t)_{t=1}^n$ is an IID sample generated via Monte Carlo from that distribution. The previous (pointwise) convergence results then apply, and $\psi_n(r)$ is consistent for $\psi(r)$.

However, for the stochastic volatility model, the data rarely supports an IID specification for σ_t. Instead, $(\sigma_t)_{t\geq 0}$ is usually assumed to follow a (stationary and ergodic) Markov process that induces bursts of volatility. In this case, it is desirable that the CMCDE remains consistent whenever $(\sigma_t)_{t\geq 0}$ is simulated from such a data generating process. We prove that this consistency holds in the general CMCDE setting described above, not only pointwise but globally as well. Moreover, assuming V-uniform ergodicity, we derive a functional asymptotic distribution for the error.

[5] We refer here to densities with respect to Lebesgue measure.

2.2 Definitions

In order to develop the formal theory, we recall a number of definitions and standard theorems. To begin, let $(\mathbb{Y}, \mathscr{Y}, \mu)$ be a σ-finite measure space, and let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. A $\mathbb{Y}$-valued random variable is a measurable map Y from $(\Omega, \mathscr{F})$ into $(\mathbb{Y}, \mathscr{Y})$. A measurable function $g\colon \mathbb{Y} \to \mathbb{R}$ is called a density if $g \geq 0$ μ-almost everywhere and $\int g d\mu = 1$. We say that $\mathbb{Y}$-valued random variable Y has density g if

$$\mathbb{P}\{Y \in B\} = \int_B g \, d\mu \qquad \text{for all } \; B \in \mathscr{Y}$$

For $p \in [1, \infty]$, we let $L_p(\mu) := L_p(\mathbb{Y}, \mathscr{Y}, \mu)$ be the Banach space of p-integrable real-valued functions on $\mathbb{Y}$.[6] The norm on $L_p(\mu)$ is given by

$$\|g\|_p := \left\{ \int g^p d\mu \right\}^{1/p} \qquad (g \in L_p(\mu))$$

with $\|g\|_\infty$ being the essential supremum. If $\mathscr{Y}$ is countably generated,[7] then $L_p(\mu)$ is separable whenever $p < \infty$. If $q \in [1, \infty]$ satisfies $1/p + 1/q = 1$, then $L_q(\mu)$ can be identified (via the Reisz representation theorem) with the norm dual of $L_p(\mu)$. We define

$$\langle g, h \rangle := \int gh \, d\mu =: \int g(x)h(x)\mu(dx) \qquad (g \in L_p(\mu), \; h \in L_q(\mu))$$

In the sequel, we consider random variables taking values in $L_p(\mu)$, where $p \in \{1, 2\}$. An $L_p(\mu)$-valued random variable F is a measurable map from the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ into $L_p(\mu)$.[8] An $L_p(\mu)$-valued random variable G is called *centered Gaussian* if, for every $h \in L_q(\mu)$, the real-valued random variable $\langle G, h \rangle$ is centered Gaussian on $\mathbb{R}$.

Below, we consider data generating processes that produce Markov realizations. In this connection, we recall some facts concerning discrete-time Markov processes taking values in a measurable space $(\mathbb{X}, \mathscr{X})$. A *stochastic kernel* on $\mathbb{X}$ is

[6] As usual, functions equal μ-almost everywhere are identified.

[7] $\mathscr{Y}$ is called countably generated if there exists a countable family $\mathscr{A}$ of subsets of $\mathbb{Y}$ such that $\mathscr{A}$ generates $\mathscr{Y}$.

[8] Measurability requires that $F^{-1}(B) \in \mathscr{F}$ for every Borel subset B of $L_p(\mu)$. If $L_1(\mu)$ and $L_2(\mu)$ are separable, then the Pettis measurability theorem assures us that F will be measurable whenever the real-valued random variable $\langle F, h \rangle$ is measurable with respect to the Borel subsets of $\mathbb{R}$ for every h in the dual space $L_q(\mu)$. This condition is easily verified in the applications that follow, and hence further discussion of measurability issues is omitted.

a function $P\colon \mathbb{X} \times \mathscr{X} \to [0,1]$ such that $B \mapsto P(x,B)$ is a probability measure on $\mathscr{X}$ for all $x \in \mathbb{X}$, and $x \mapsto P(x,B)$ is $\mathscr{X}$-measurable for all $B \in \mathscr{X}$. Informally, $P(x,B)$ gives the probability of moving from state x into set B in one step. The t step transitions are given by P^t, where $P^t(x,B) := \int P^{t-1}(x,dy)P(y,B)$ and $P^1 := P$.

A discrete-time, $\mathbb{X}$-valued stochastic process $(X_t)_{t\geq 0}$ is said to be *P-Markov* if $P(X_t,dy)$ is the conditional distribution of X_{t+1} given X_t. More precisely,

$$\mathbb{P}[X_{t+1} \in B \,|\, \mathscr{F}_t] = P(X_t,B) \qquad \text{for all } \; B \in \mathscr{X}$$

where $\mathscr{F}_t$ is the σ-algebra generated by the history $X_0,\ldots,X_t$. A distribution (i.e., probability measure) ϕ on $\mathscr{X}$ is called *stationary* for P if

$$\phi(B) = \int P(x,B)\phi(dx) \qquad \text{for all } \; B \in \mathscr{X}$$

The kernel P is called *ergodic* if it has a unique stationary distribution ϕ, and, for every P-Markov process $(X_t)_{t\geq 0}$ and every measurable $h\colon \mathbb{X} \to \mathbb{R}$ with $\int |h|d\phi < \infty$, we have

$$\frac{1}{n}\sum_{t=1}^{n} h(X_t) \to \int h\, d\phi \quad \mathbb{P}\text{-almost surely as } \; n \to \infty \tag{5}$$

The kernel P is called *V-uniformly ergodic* if, in addition, there exist a measurable function $V\colon \mathbb{X} \mapsto [1,\infty)$ and nonnegative constants $\lambda < 1$ and $L < \infty$ satisfying

$$\sup_{|h|\leq V} \left| \int h(y)P^t(x,dy) - \int h(y)\phi(dy) \right| \leq \lambda^t LV(x) \quad \text{for all } x \in \mathbb{X},\ t \in \mathbb{N}$$

(If V can be chosen identically equal to 1, then the left-hand side becomes the total variation distance between $P^t(x,dy)$ and ϕ, while the right hand side is independent of x. This is the uniformly ergodic case.) Under the V-uniform ergodicity assumption, the central limit theorem can be established for a broad class of functions. Moreover, V-uniform ergodicity has been shown to hold in a range of economic and econometric applications.[9]

2.3 Main Results

We now give a more formal presentation of the CMCDE defined in (2). The random variable Y is assumed to take values in a measure space $(\mathbb{Y},\mathscr{Y},\mu)$, where the

[9] See, for example, Kristensen (2008) or Nishimura and Stachurski (2005).

σ-algebra $\mathscr{Y}$ is countably generated and μ is σ-finite. We let ψ be the density of Y on $(\mathbb{Y}, \mathscr{Y}, \mu)$. One common case is where $\mathbb{Y} \subset \mathbb{R}^k$, $\mathscr{Y}$ is the Borel sets, and μ is Lebesgue measure. In this case, ψ is a density in the usual sense of the term. Another common case is where $\mathbb{Y}$ is discrete, $\mathscr{Y}$ is the set of all subsets, and μ is the counting measure. In the latter setting, integration corresponds to summation, and ψ is a probability mass function on $\mathbb{Y}$.

We allow X in (1) to take values in an arbitrary measurable space $(\mathbb{X}, \mathscr{X})$. The distribution of X is denoted by ϕ. The conditional density q in (1) is required to be a measurable map from $\mathbb{Y} \times \mathbb{X}$ into $\mathbb{R}_+$ such that $y \mapsto q(y \mid x)$ is a density on $(\mathbb{Y}, \mathscr{Y}, \mu)$ for each $x \in \mathbb{X}$. Existence of q requires that the conditional distribution of Y given $X = x$ is absolutely continuous with respect to μ for every $x \in \mathbb{X}$.

The relationship (1) can be re-expressed in terms of ψ, q and ϕ:

$$\psi(y) = \int q(y \mid x)\phi(dx) \quad \text{for all } y \in \mathbb{Y} \tag{6}$$

We refer to (6) as the *conditional density representation* of ψ. Suppose now that there exists a stochastic kernel P on $(\mathbb{X}, \mathscr{X})$ such that (a) ϕ is the unique stationary distribution of P, and (b) we can simulate P-Markov time series from some initial $X_0 = x_0 \in \mathbb{X}$. In this setting, we define the CMCDE of ψ as

$$\psi_n(y) = \frac{1}{n} \sum_{t=1}^{n} q(y \mid X_t) \quad \text{where } (X_t)_{t=1}^n \text{ is } P\text{-Markov} \tag{7}$$

Example 2.1. Recall the stochastic volatility example presented in section 2.1. The conditional density representation (6) needed to estimate ψ is given in (4). Suppose now that the process for conditional volatility σ_t can be expressed as a stationary Markov process of the form $\sigma_{t+1} = F(\sigma_t, W_{t+1})$, where (W_t) is an IID sequence. Let P be the stochastic kernel that represents this process, in the sense that $P(\sigma, B) = \mathbb{P}\{F(\sigma, W_t) \in B\}$. Let ϕ be the stationary distribution of P. Iterating on the equation $\sigma_{t+1} = F(\sigma_t, W_{t+1})$ produces a P-Markov time series (σ_t), which can be used to construct the CMCDE.

Example 2.2. Suppose that there exists a conditioning variable X with distribution ϕ that admits the conditional density representation (6). Suppose further that direct IID sampling from ϕ is either infeasible or computationally expensive. In this setting, the Markov chain Monte Carlo (MCMC) solution is to construct a kernel P such that ϕ is the stationary distribution of P. By sampling a P-Markov time series, the CMCDE can be constructed.

Example 2.3. Suppose again that there exists a conditioning variable X with distribution ϕ that admits the conditional density representation (6), but suppose

now that IID draws from ϕ are possible and efficient. This can also be regarded as a special case of the CMCDE defined in (7). The stochastic kernel P is defined as $P(x, B) = \phi(B)$. With this definition, P-Markov processes are IID samples from ϕ.

Our interest is in global convergence of ψ_n to ψ, in the sense of norm deviation between the two functions in $L_p(\mu)$. Provided that P is ergodic, global $L_1(\mu)$ consistency of the estimator (7) obtains without any additional assumptions:

Theorem 2.1. *If P is ergodic, then ψ_n is globally consistent for ψ, in the sense that, with probability one, $\psi_n \to \psi$ in $L_1(\mu)$ as $n \to \infty$.*

The $L_1(\mu)$ norm is perhaps the most attractive way to measure the deviation between two densities (Devroye and Lugosi, 2001). For example, Scheffé's identity and theorem 2.1 imply that

$$\sup_{B \in \mathscr{Y}} \left| \int_B \psi_n \, d\mu - \int_B \psi \, d\mu \right| \to 0 \quad \text{as } n \to \infty \quad \mathbb{P}\text{-a.s.}$$

so the maximum deviation in probabilities over all events converges to zero. On the other hand, $L_1(\mu)$ is a Banach space but not a Hilbert space, and, without the Hilbert space property, asymptotic normality is difficult to obtain. To prove asymptotic normality, we now shift our analysis into the Hilbert space $L_2(\mu)$. We also require that P is V-uniformly ergodic for some function V, as well as a second moment condition described below.

Let $Q(x) := q(\cdot, x) - \psi$, and define the linear operator $C \colon L_2(\mu) \to L_2(\mu)$ by

$$\langle g, Ch \rangle = \mathbb{E}\langle g, Q(X_1^*)\rangle\langle h, Q(X_1^*)\rangle + \sum_{t \geq 2}^{\infty} \mathbb{E}\langle g, Q(X_1^*)\rangle\langle h, Q(X_t^*)\rangle + \sum_{t \geq 2}^{\infty} \mathbb{E}\langle h, Q(X_1^*)\rangle\langle g, Q(X_t^*)\rangle$$

for arbitrary $h, g \in L_2(\mu)$, where $(X_t^*)_{t \geq 0}$ is stationary and P-Markov.[10] We can now state the following result:

Theorem 2.2. *If P is V-uniformly ergodic and the second moment condition*

$$\int q(y \mid x)^2 \mu(dy) \leq V(x) \qquad \textit{for all} \quad x \in \mathbb{X} \tag{8}$$

[10] That is, $(X_t^*)_{t \geq 0}$ is P-Markov and X_0^* is drawn from the stationary distribution ϕ. That C is indeed a well-defined operator from $L_2(\mu)$ to itself follows from the proof of theorem 2.2. In fact C is also self-adjoint and compact (i.e., maps bounded subsets of $L_2(\mu)$ into relative compact sets).

holds, then $\sqrt{n}(\psi_n - \psi)$ *converges in distribution to a centered Gaussian* G *on* $L_2(\mu)$ *with covariance operator* C.[11]

Remark 2.1. Since $h \mapsto \|h\|_2$ is continuous on $L_2(\mu)$, the continuous mapping theorem now implies that $\|\psi_n - \psi\|_2 = O_P(n^{-1/2})$.

Remark 2.2. As mentioned above, the IID setting, where IID sampling from ϕ is possible, can be viewed as a special case. Working through the logic or proving it directly, one can show that the strong consistency result in theorem 2.1 holds without any additional conditions on q and ϕ, and that theorem 2.2 holds whenever

$$\int \int q(y \mid x)^2 \mu(dy) \phi(dx) < \infty \quad \text{and} \quad \left\{ \int q(y \mid x)^2 \mu(dy) < \infty \quad \forall x \in \mathbb{X} \right\}$$

The IID case is more important than it might seem, even in time series environments. Below, we will provide examples where we can sample in an IID way from time series which are persistent and possibly nonstationary.

One important aspect of theorems 2.1 and 2.2 is that the simulated P-Markov process $(X_t)_{t=1}^n$ used to construct ψ_n need only be *asymptotically stationary*, rather than stationary *per se*. In particular, when simulating $(X_t)_{t=1}^n$, the initial condition X_0 can be any arbitrarily chosen $x_0 \in \mathbb{X}$ (as opposed to being drawn from ϕ). In applications this is often essential, as the stationary distribution ϕ is typically intractable, and no direct method for sampling from it is available.

However, in the case where it is possible to generate time series with common distribution ϕ, the CMCDE ψ_n is also unbiased, both locally and globally:

Lemma 2.1. *If* (X_t) *is identically distributed with* $X_t \sim \phi$ *for all* t*, then* ψ_n *is an unbiased estimator of* ψ*. In particular,* $\mathbb{E}\psi_n(y) = \psi(y)$ *for all* $y \in \mathbb{Y}$*, and* $\mathcal{E}\psi_n = \psi$.[12]

Before continuing, let us make a brief comparison with standard results for nonparametric kernel density estimation. This comparison is easiest if we limit attention to the IID case. In order to define the NPKDE, we need to restrict attention to the case where $\mathbb{Y} \subset \mathbb{R}^k$, as opposed to the general measure space setting of

[11] A centered Gaussian G has covariance operator C if $\mathbb{E}\langle g, G\rangle\langle h, G\rangle = \langle Cg, h\rangle$ for every $g, h \in L_2(\mu)$. Also, convergence in distribution is defined in the obvious way: Let $\mathscr{C}$ be the continous, bounded, real-valued functions on $L_2(\mu)$, where continuity is with respect to the norm topology. Let $(G_n)_{n\geq 0}$ be $L_2(\mu)$-valued random variables. Then $G_n \to G_0$ in distribution if $\mathbb{E}h(G_n) \to \mathbb{E}h(G)$ for every $h \in \mathscr{C}$.

[12] Here $\mathcal{E}$ is the *functional* expectation of ψ_n, defined in terms of the Dunford–Pettis integral in $L_1(\mu)$. For details, see the technical appendix.

the CMCDE. We must also suppose that one can generate IID samples $Y_1, \ldots, Y_n$ from ψ. The NPKDE f_n is then defined in terms of a kernel (i.e., density) K on $\mathbb{Y}$ and a "bandwidth" parameter δ_n:

$$f_n(y) := \frac{1}{n\delta_n} \sum_{i=1}^{n} K\left(\frac{y - Y_i}{\delta_n}\right) \tag{9}$$

The estimate is f_n known to be consistent, in the sense that $\mathbb{E}\|f_n - \psi\|_1 \to 0$ whenever $\delta_n \to 0$ and $n\delta_n^k \to \infty$ (Devroye and Lugosi, 2001). However, rates of convergence are slower than the parametric rate obtained for the CMCDE. For example, if we fix $y \in \mathbb{Y}$ and take ψ to be twice differentiable, then, for suitable choice of K, it can be shown that

$$|f_n(y) - \psi(y)| = O_P[(n\delta_n^k)^{-1/2}] \text{ when } n\delta_n^k \to \infty \text{ and } (n\delta_n^k)^{1/2}\delta_n^2 \to 0$$

Thus, even with this smoothness assumption on ψ—which may or may not hold in practice—the convergence rate $O_P[(n\delta_n^k)^{-1/2}]$ of the NPKDE is slower than the rate $O_P(n^{-1/2})$ of the CMCDE. Moreover, in contrast to the CMCDE, the rate of convergence slows as the dimension k of the state space increases.

3 Special Cases of CMCDE in the Literature

In this section, we show how the CMCDE nests and extends conditional Monte Carlo results that arise in distinct strands of the exisiting literature.

3.1 Density Estimation under Gibbs Sampling

Gelfand and Smith (1990) propose the following technique for computing marginal densities. Suppose we have a pair of real-valued random variables (X, Y) with joint density $f_{X,Y}$ (for simplicity, we are restricting attention to the bivariate case). Suppose further that both conditional densities $f_{Y|X}$ and $f_{X|Y}$ are available, and can be sampled from. Our objective is to compute the marginal density f_Y of Y, which is assumed to be intractable. The Gibbs sampler proceeds by choosing X_0 arbitrarily, and then simulating $(X_t, Y_t)_{t=0}^n$ by repeated drawing

$$Y_t \sim f_{Y|X}(\cdot \,|\, X_t) \quad \text{and then} \quad X_{t+1} \sim f_{X|Y}(\cdot \,|\, Y_t)$$

Gelfand and Smith (1990) suggest fixing T to be a large number, and repeating the above sampling method to generate n independent replications $X_T^1, \ldots, X_T^n$ of

X_T. They then form the density estimate f_Y^n of f_Y given by

$$f_Y^n(y) = \frac{1}{n}\sum_{i=1}^{n} f_{Y|X}(y \mid X_T^i)$$

The sequence (X_t, Y_t) produced by Gibbs sampling forms a Markov process. Letting g_T denote the marginal density of Y_T for this process, observe that the overall error $\|f_Y^n - f_Y\|_1$ can be decomposed as follows:

$$\|f_Y^n - f_Y\|_1 \leq \|f_Y^n - g_T\|_1 + \|g_T - f_Y\|_1 \tag{10}$$

Under mild assumptions, Gibbs sampling is sufficiently ergodic to ensure that the second term on the right-hand side of (10) is small for large T. Regarding the first term on the right-hand side , Gelfand and Smith (1990) use the IID structure of the sampling data to show that $\|f_Y^n - g_T\|_1$ converges to zero in probability as $n \to \infty$. Our asymptotic theory shows that this convergence occurs almost surely (theorem 2.1). Moreover, theorem 2.2 implies that the deviation $f_Y^n - g_T$ is asymptotically Gaussian, and that the L_2 norm of this deviation is $O_P(n^{-1/2})$.

One limitation of Gelfand and Smith's approach is that f_Y^n is consistent for g_T, but not for f_Y. Our asymptotic theory suggests an alternative approach which is consistent for f_Y. Instead of replicating the Gibbs sampling procedure n times to generate independent draws of X_T, a single replication can be used to produce the estimator

$$\psi_n(y) = \frac{1}{n}\sum_{t=1}^{n} f_{Y|X}(y \mid X_t)$$

Here $(X_t)_{t=1}^n$ is one time series produced by the the Gibbs sampler. Although $(X_t)_{t=1}^n$ is not IID, under certain restrictions on the conditional densities it can be shown to be V-uniformly ergodic (Rosenthal, 1995, lemma 7), and theorems 2.1 and 2.2 apply. In this case, ψ_n is L_1-consistent for f_Y with probability one, and the L_2 deviation is $O_P(n^{-1/2})$.

It is beyond the scope of this paper to provide a detailed comparative efficiency analysis of the two estimators. However, we note an additional attractive property of the second estimator: It uses all observations produced in the Gibbs sampling process to construct the estimate of f_Y. The first estimator, in contrast, discards $(T-1)n$ observations.

3.2 The Look-Ahead Estimator

Another estimator that can be studied as a special case of the CMCDE defined in this paper is the look-ahead estimator of Glynn and Henderson (2001). Con-

sider a $\mathbb{Y}$-valued P-Markov process $(X_t)_{t\geq 0}$, where the stochastic kernel P satisfies $P(x, B) = \int_B q(y \mid x)\mu(dy)$ for some conditional density $q\colon \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}_+$. Suppose that a unique stationary distribution exists. In this setting, the stationary distribution can be represented by a density ψ on $\mathbb{Y}$, and

$$\psi(y) = \int q(y \mid x)\psi(x)\mu(dx) \quad \text{for all } y \in \mathbb{Y} \tag{11}$$

Comparison of (11) and (6) reveals that a conditional density representation of ψ can be obtained by setting $\phi(dx) = \psi(x)\mu(dx)$, and the CMCDE can be applied. Since ϕ is then the stationary distribution of the kernel P in question, we simulate a P-Markov time series $(X_t)_{t=1}^n$, and form the CMCDE as $\psi_n(y) = n^{-1}\sum_{t=1}^n q(y \mid X_t)$.

The CMCDE ψ_n so derived is precisely the *look-ahead estimator* of Glynn and Henderson (2001). Hereafter, we refer to this as the time-series look-ahead estimator (TSLAE). Stachurski and Martin (2008) showed that if P is V-uniformly ergodic and $\int q(y \mid x)^2\mu(dy) \leq V(x)$ for all $x \in \mathbb{Y}$, then the TSLAE converges to ψ in $L_2(\mu)$ at rate $O_P(n^{-1/2})$. This is a special case of theorem 2.2,

The CMCDE also nests another form of the look-ahead estimator considered by Glynn and Henderson (2001). Consider a nonstationary (time-inhomogenous) Markov process of the form $X_{t+1} = F_t(X_t, W_{t+1})$, where (W_t) is an IID process. Let $q_{t+1}(y \mid x)$ be the conditional density of X_{t+1} given $X_t = x$, and let $\psi_t(y \mid x_0)$ be the conditional density of X_t given $X_0 = x_0$. Suppose that we wish to estimate the conditional density $\psi_T(y \mid x_0)$ for some fixed $T \in \mathbb{N}$. From the Markov property we have the recursion

$$\psi_T(y \mid x_0) = \int q_T(y \mid x)\psi_{T-1}(x \mid x_0)\,\mu(dx) \qquad \text{for all } y \in \mathbb{Y} \tag{12}$$

Comparison of (12) and (6) shows that ψ_T admits a conditional density representation, with $\phi(dx) = \psi_{T-1}(x \mid x_0)\,\mu(dx)$. Given that the model is nonstationary, a different approach is required to sample from $\psi_{T-1}(x \mid x_0)\,\mu(dx)$. A natural approach is as follows: Starting at $X_0 = x_0$, draw successive shock values for W_t and iterate on $X_{t+1} = F_t(X_t, W_{t+1})$ to generate a random variable X_{T-1} with distribution $\psi_{T-1}(x \mid x_0)\mu(dx)$. Replicating this process n times gives n *independent* observations from this distribution. This allows us to construct the CMCDE as a special case of example 2.3. Specifically, combining the IID draws with the conditional density q_T yields the estimator

$$\psi_T^n(y \mid x_0) := \sum_{i=1}^n q_T(y \mid X_{T-1}^i) \tag{13}$$

This CMCDE corresponds to the cross-sectional look-ahead estimator (CSLAE) proposed by Glynn and Henderson (2001). Given that the samples are IID, the estimator is unbiased (lemma 2.1). Since IID samples are V-uniformly ergodic, theorems 2.1 and 2.2 also apply, and the estimator is globally consistent and asymptotically normal.

3.3 Simulated Maximum Likelihood

Another area where the idea of density simulation via conditioning has been used is simulated maximum likelihood estimation (SML). As an illustration, we consider the problem of estimating continous time diffusions, as studied by Pedersen (1995) and Brandt and Santa-Clara (2001).[13] Suppose that dynamics of a process (Y_t) are described by a stochastic differential equation of the form

$$dY_t = \mu(Y_t)\,dt + \sigma(Y_t)\,dW_t \tag{14}$$

where (W_t) is a Wiener process. The contiuous time likelihood function associated with a discrete sequence (Y_t) of observations is given by

$$L(Y_0, \ldots, Y_M) = \psi(Y_0) \prod_{t=0}^{M-1} p(\Delta, Y_t, Y_{t+1}) \tag{15}$$

where $p(t, x, y)$ is the transition probability function associated with (14), and Δ is the (common) duration between observations.[14] In many cases of interest, p is intractible, and terms of the form $p(\Delta, Y_t, Y_{t+1})$ cannot easily be computed. In order to approximate them, Pedersen (1995) and Brandt and Santa-Clara (2001) propose the following method: Let T be any integer, and let $\delta := \Delta/T$. The Euler discretization of (14) is given by

$$\hat{Y}_{t+\delta} = \hat{Y}_t + \mu(\hat{Y}_t)\delta + \sigma(\hat{Y}_t) Z_t\, \delta^{1/2} \tag{16}$$

where $(Z_t) \overset{\text{IID}}{\sim} N(0,1)$. Let $\psi_T(y \mid x)$ be time T state density associated with (16), in the sense that $\psi_T(\cdot \mid x)$ is the density of $\hat{Y}_{\delta T} = \hat{Y}_\Delta$ given $\hat{Y}_0 = x$. Brandt and Santa-Clara (2001) provide sufficient conditions under which $\psi_T(y \mid x)$ converges to $p(\Delta, x, y)$ as $T \to \infty$ (and $\delta \to 0$). Hence, the problem of approximating the likelihood function (15) reduces to approximating $\psi_T(y \mid x)$ for large T.

[13] We consider only the implementation of the SML described in these papers. Further refinements such as importance sampling have been shown to enhance the performance of the method (see, e.g., Durham and Gallant, 2002).

[14] For notational simplicity, we are describing the case where the intervals between observations are equal. The techniques presented below can also be applied to irregular observations.

The strategy proposed by Pedersen (2001) and Brandt and Santa-Clara (2001) for approximating $\psi_T(y \mid x)$ is a special case of the CSLAE we described in section 3.2. (To the best of our knowledge, this connection has not been previously made, and the ideas were developed independently.) As discussed above, the CSLAE is, in turn, a special case of the general CMCDE studied in this paper.

4 Potential Applications

The previous section discussed how several existing results could be seen as special cases of the CMCDE. In this section, we show how our results broaden the scope of potential applications to a larger class of models.

4.1 GARCH

The first potential application that we treat is computation of the stationary density of returns in a GARCH environment. There are a variety of reasons why researchers are interested in the stationary density of returns, including density forecasting, value at risk, exact likelihood estimation and model assessment. [cites] After illustrating the method, we go on to document the small sample properties of the CMCDE. We show that these small sample properties compare very favorably to those of a standard nonparametric alternative.

Consider a GARCH(1,1) process of the form

$$r_t = \sigma_t W_t, \quad \text{where } (W_t) \overset{\text{IID}}{\sim} N(0,1) \tag{17}$$

$$\sigma_{t+1}^2 = \alpha_0 + \beta\sigma_t^2 + \alpha_1 r_t^2 \tag{18}$$

where all parameters are strictly positive and $\alpha_1 + \beta < 1$. Let us consider how to compute the stationary marginal density of returns r_t. Letting ψ be this density and letting ϕ be the stationary distribution of $X_t := \sigma_t^2$, equation (17) implies that $r_t = \sqrt{X_t} W_t$, and hence the conditional density $q(r \mid x)$ of r_t given $X_t = x$ is centered Gaussian with variance x. For this q we have $\psi(r) = \int q(r \mid x)\phi(dx)$, which is a version of (6). The process $(X_t)_{t \geq 0}$ can be expressed as

$$X_{t+1} = \alpha_0 + \beta X_t + \alpha_1 X_t W_t^2$$

Given this series, the CMCDE of ψ is the density

$$\psi_n(r) = \frac{1}{n}\sum_{t=1}^{n} q(r \mid X_t) = \frac{1}{n}\sum_{t=1}^{n} (2\pi X_t)^{-1/2} \exp\left\{-\frac{r^2}{2X_t}\right\} \tag{19}$$

Using the sufficient conditions of Meyn and Tweedie (2009), it can be shown that $(X_t)_{t\geq 0}$ is V-uniformly ergodic on $\mathbb{X} := [\alpha_0/(1-\beta), \infty)$ for $V(x) = x + c$, where c is any constant in $[1, \infty)$. (More generally, Kristensen (2008) establishes V-uniform ergodicity for a large class of GARCH formulations.)

Regarding the moment condition (8) in theorem 2.2, we have

$$\int q(r \mid x)^2 dr = \int (2\pi x)^{-1} \exp\left\{-\frac{r^2}{x}\right\} dr = (4\pi x)^{-1/2} \leq \left\{\frac{4\pi\alpha_0}{1-\beta}\right\}^{-1/2}$$

Recall that $V(x) = x + c$, where c can be chosen arbitrarily large. For large enough c, then, we have $\int q(r \mid x)^2 dr \leq c \leq x + c = V(x)$, and (8) is satisfied. As a result, both theorems 2.1 and 2.2 apply.

The fast convergence for the CMCDE ψ_n implied by theorem 2.2 is illustrated in figure 1. For the exercise, we set $n = 500$.[15] The left panel of the figure contains the true density ψ drawn in bold, as well as 50 replications of the NPKDE. Each NPKDE replication uses a simulated time series $(r_t)_{t=1}^n$, combined with standard default settings (a Gaussian kernel and bandwidth calculated according to Silverman's rule).[16] The right panel of figure 1 repeats the exercise, but this time using the CMCDE instead of the NPKDE. Each replication of the CMCDE is calculated according to (19).

It is very clear from the figure that the small sample properties of the CMCDE are favorable to those of the NPKDE, at least for this application. The replications are more tightly clustered around the true distribution both at the center of the distribution and at the tails. One reason the figure is interesting is that the NPKDE foregoes an unbiased estimate in order to obtain lower variance. Nevertheless, fixing y at any value, we see that the variance $\text{Var}\,\psi_n(y)$ of the CMCDE is significantly smaller than that of the NPKDE.

Another metric for comparing density estimators is to look at L_1-norm deviations from the true density ψ. We computed average L_1 deviations over 1000 replications. For $n = 500$, the ratio of the CMCDE L_1 deviation to the NPKDE L_1 deviation was 0.5854, indicating that the average errors for the NPKDE were almost twice those of the CMCDE. If instead we set $n = 2000$, the ratio was 0.7431.

[15] The parameters are $\alpha_0 = \alpha_1 = 0.05$ and $\beta = 0.9$.

[16] The density marked as "true" in the figure is in fact an approximation, calculated by simulation with $n = 10^7$. For such a large n there is no visible variation of the density over different realization, or different methods of simulation.

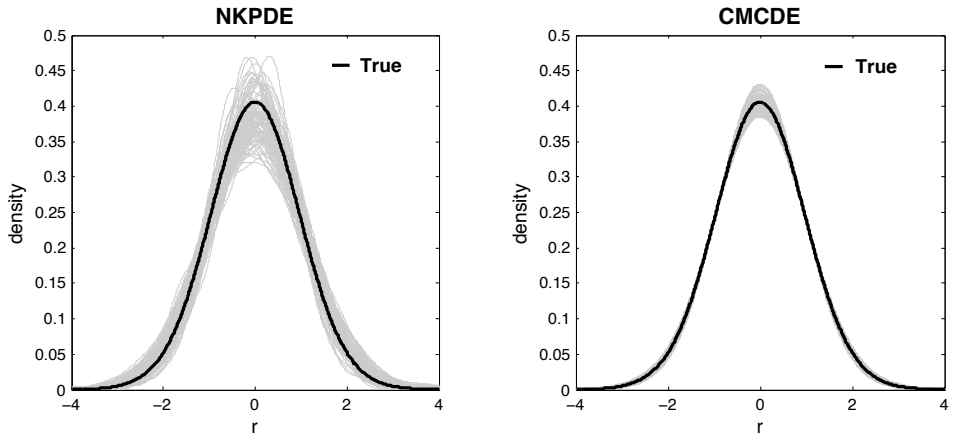

Figure 1: Relative performance, $n = 500$

4.2 The Aiyagari Model

So far we have considered primarily econometric applications. The CMCDE can be also be applied to computing densities in dynamic economic models. To illustrate this point, consider the incomplete market economy of Aiyagari (1994). We begin with the consumption smoothing problem of an individual who insures against idiosyncratic earnings risk by saving at a risk free rate of r. Both the real interest rate and the wage rate w are constant, but total earnings are random due to idiosyncratic variations in labor productivity. Each individual also faces a borrowing constraint that rules out uncollateralized borrowing. The household's problem can be described by the Bellman equation

$$V(a,z) = \max_{c,a'} \left\{ u(c) + \beta \mathbb{E}[V(a',z') \mid z] \right\}$$

subject to

$$0 \leq a' \leq wz + a(1+r) - c$$

The serially correlated shock to labor productivity evolves according to

$$\ln z' = \rho \ln z + \sigma \sqrt{1-\rho^2}\epsilon, \quad \epsilon \sim N(0,1) \tag{20}$$

A common approach to solving this problem is to discretize the productivity process using Tauchen's method (1986), obtaining a grid $\{z_1, ..., z_M\}$ and an $M \times M$ stochastic matrix R that represents the dynamics of (20) on the grid.[17] Assuming

[17] In particular, $\mathbb{P}\{z_{t+1} = z' \mid z_t = z\} = R(z,z')$ for any $z, z' \in \{z_1, \ldots, z_M\}$.

that assets also lie on a finite grid $\{a_1, ...a_L\}$, the solution to the Bellman equation is an $L \times M$ matrix of saving policies $a' = g(a, z)$.

Given the solution to the individual's problem, we are interested in making inferences about the stationary labor productivity-asset distribution ψ. The path of assets and productivity shocks evolves according to $a_{t+1} = g(a_t, z_t)$ with $(z_t)_{t\geq 0}$ generated by R. Taking $X_t := (a_t, z_t)$ as the state variable, the transition probabilities from state $x = (a, z)$ to state $y = (a', z')$ are given by

$$q(y \mid x) :=: q((a', z') \mid (a, z)) := \mathbb{1}\{g(a, z) = a'\}R(z, z') \tag{21}$$

The problem of computing ψ can be expressed in terms of the conditional density representation (6). In this discretized setting, the state space is finite, and all probability distributions can be represented by probability mass functions. Probability mass functions can be regarded as densities with respect to the counting measure. In particular, $q(\cdot \mid x)$ is a density with respect to the counting measure for all x. Moreover, ψ is also a density, and, given that ψ is the stationary distribution of the Markov process with transition probabilities $q(y \mid x)$, we have

$$\psi(y) = \sum_x q(y \mid x)\psi(x)$$

for all y. This is a version of the conditional density representation (6). Hence, the CMCDE of ψ is

$$\psi_n(a', z') = \frac{1}{n}\sum_{t=1}^{n} q((a', z') \mid (a_t, z_t)) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{g(a_t, z_t) = a'\}R(z_t, z') \tag{22}$$

Implementation proceeds by drawing a sequence of n realizations of labor productivity from the matrix R, and using the policy function g to derive n associated values of assets. The resulting sequence $(a_t, z_t)_{t=1}^n$ is then inserted into (22), and this expression is evaluated at all $a' \in \{a_1, \ldots, a_L\}$ and $z' \in \{z_1, \ldots, z_M\}$.

For standard parameterizations, the Markov process associated with the matrix R is ergodic. In the finite state case, ergodicity and V-uniform ergodicity are equivalent, and hence the conditions of theorems 2.1 and 2.2 are satisfied. (Since the state space is finite, the second moment conditions required for theorem 2.2 are automatically satisfied.)

5 Conclusion

We have covered only a few of the potential applications for conditional Monte Carlo density estimation. There are many other substatitive applications where

the density of interest admits a conditional density representation. These include computation of the initial distribution when conducting exact maximum likelihood estimation in time series or panel environments, density forecasting, Bayesian decision making and/or econometrics, and robust control.

At present, this conditioning information is not always used. Instead, densities are simulated and computed via nonparametric kernel density methods or histograms. We have shown that exploiting conditioning information yields estimators with better asymptotic and small-sample properties.

This paper has documented how conditioning information can be used when computing densities. The same insight can be applied to other problems as well. In Braun et al. (2009), we study how conditioning information can be used to produce efficient estimators of moments and other expectations.

6 Technical Appendix

The *expectation* of an $L_p(\mu)$-valued random variable F is defined as the element $\mathcal{E}F$ of $L_p(\mu)$ such that

$$\mathbb{E}\langle F, h\rangle = \langle \mathcal{E}F, h\rangle \text{ for every } h \in L_q(\mu)$$

where $\mathbb{E}$ is the usual scalar expectation. It follows from the Reisz representation theorem that if $\mathbb{E}\|F\|_p$ is finite, then $\mathcal{E}F$ exists and is unique. The Banach-space law of large numbers (cf., e.g., Bosq, 2000, Theorem 2.4) implies that if $(F_i)_{i\geq 1}$ is an IID sequence in $L_p(\mu)$ with expectation $\mathcal{E}F$, then

$$\frac{1}{n}\sum_{i=1}^{n} F_i \to \mathcal{E}F \quad \mathbb{P}\text{-almost surely as } n \to \infty \tag{23}$$

where convergence is with respect to the norm on $L_p(\mu)$.

In the case $p = 2$, the space $L_2(\mu)$ is a Hilbert space, and a version of the central limit theorem holds. In particular, if $(F_i)_{i\geq 1}$ is an IID sequence in $L_2(\mu)$ such that $\mathbb{E}\|F\|_2^2$ is finite, then

$$n^{1/2}\left\{\frac{1}{n}\sum_{i=1}^{n} F_i - \mathcal{E}F\right\} \tag{24}$$

converges in distribution to a centered Gaussian on $L_2(\mu)$.

Lemma 6.1. *If X has distribution ϕ, then $\mathcal{E}q(\cdot \mid X) = \psi$ in $L_1(\mu)$.*

Proof of lemma 6.1. Let us consider the $L_1(\mu)$ case first. Because $\mathbb{E}\|q(\cdot \mid X)\|_1 = 1 < \infty$, the expectation $\mathcal{E}q(\cdot \mid X)$ is well-defined. To show that $\mathcal{E}q(\cdot \mid X) = \psi$, we must prove that $\mathbb{E}\langle q(\cdot \mid X), h\rangle = \langle \psi, h\rangle$ for all $h \in L_\infty(\mu)$. Fixing $h \in L_\infty(\mu)$, Fubini's theorem and (1) yield

$$\mathbb{E}\langle q(\cdot \mid X), h\rangle = \mathbb{E}\int q(y \mid X)h(y)\mu(dy) = \int \mathbb{E}q(y \mid X)h(y)\mu(dy)$$

By (1) this equals $\int \psi h d\mu = \langle \psi, h\rangle$, as was to be shown. □

Proof of lemma 2.1. Assume the hypotheses of the lemma. To see that ψ_n is locally unbiased, fix any $y \in \mathbb{Y}$. Then

$$\mathbb{E}\psi_n(y) = \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} q(y \mid X_t)\right] = \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}q(y \mid X_t) = \psi(y)$$

That ψ_n is globally unbiased follows from linearity of $\mathcal{E}$ and lemma 6.1. □

Proof of theorem 2.1. As in the statement of the theorem, let P be an ergodic stochastic kernel on $(\mathbb{X}, \mathscr{X})$ with stationary distribution ϕ. Let $(X_t)_{t\geq 0}$ be P-Markov and let $X^* \sim \phi$. Define $Q(x) := q(\cdot \mid x) - \psi$, which is a measurable function from $\mathbb{X}$ to $L_1(\mu)$. Note that $\mathcal{E}Q(X^*) = 0$ by lemma 6.1. We need to show that

$$\lim_{n\to\infty}\|\psi_n - \psi\| = \lim_{n\to\infty}\left\|\frac{1}{n}\sum_{t=1}^{n} Q(X_t)\right\| = 0 \qquad (\mathbb{P}\text{-almost surely}) \tag{25}$$

Our proof is an extention of that for the IID Banach space LLN, as given in Bosq (2000, thm. 2.4). To begin, fix $\epsilon > 0$. Since $L_1(\mu)$ is separable, we can choose a partition $\{B_j\}_{j\in\mathbb{N}}$ of $L_1(\mu)$ such that each B_j has diameter less than ϵ. For any $L_1(\mu)$-valued random variable U, we let $L_J U := \sum_{j=1}^{J} b_j \mathbb{1}\{U \in B_j\}$, where, for each j, b_j is a fixed point in B_j. Thus, $L_J U$ is a simple random variable that approximates U. In particular, we have the following result, a proof of which can be found in Bosq (2000, pp. 27-28):

$$\exists\, J \in \mathbb{N} \text{ with } \mathbb{E}\|Q(X^*) - L_J Q(X^*)\| < 2\epsilon \tag{26}$$

Our first claim is that

$$\lim_{n\to\infty}\left\|\frac{1}{n}\sum_{t=1}^{n} L_J Q(X_t) - \mathcal{E}L_J Q(X^*)\right\| = 0 \qquad (\mathbb{P}\text{-almost surely}) \tag{27}$$

To establish (27), we can use the real ergodic law (5) to obtain

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n} L_J Q(X_t) &= \sum_{j=1}^{J} b_j \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{Q(X_t) \in B_j\} \\ &\to \sum_{j=1}^{J} b_j \mathbb{P}\{Q(X^*) \in B_j\} = \mathcal{E} L_J Q(X^*)\end{aligned}$$

almost surely, where the last equality follows immediately from the definition of $\mathcal{E}$. Thus (27) is established.

Returning to (25), we have

$$\begin{aligned}\left\| \frac{1}{n}\sum_{t=1}^{n} Q(X_t) \right\| \leq \frac{1}{n}\sum_{t=1}^{n} \|Q(X_t) - L_J Q(X_t)\| \\ + \left\| \frac{1}{n}\sum_{t=1}^{n} L_J Q(X_t) - \mathcal{E} L_J Q(X^*) \right\| + \|\mathcal{E} L_J Q(X^*)\|\end{aligned}$$

Using real-valued ergodicity again, as well as (27), we get

$$\limsup_{n\to\infty} \left\| \frac{1}{n}\sum_{t=1}^{n} Q(X_t) \right\| \leq \mathbb{E}\|Q(X^*) - L_J Q(X^*)\| + \|\mathcal{E} L_J Q(X^*)\|$$

But the fact that $\mathcal{E} Q(X^*) = 0$ now gives

$$\|\mathcal{E} L_J Q(X^*)\| = \|\mathcal{E} Q(X^*) - \mathcal{E} L_J Q(X^*)\| \leq \mathbb{E}\|Q(X^*) - L_J Q(X^*)\|$$

In view of (26) we then have

$$\limsup_{n\to\infty} \left\| \frac{1}{n}\sum_{t=1}^{n} Q(X_t) \right\| \leq 4\epsilon \qquad (\mathbb{P}\text{-almost surely})$$

Since ϵ is arbitrary, the proof of (25) is now done. □

Proof of theorem 2.2. This theorem can be established using the Hilbert space CLT of Stachurski (2009, theorem 3.1), where $x \mapsto q(\cdot \mid x)$ corresponds to T_0 in that theorem. The only point that needs checking vis-a-vis that CLT is that $\mathcal{E} q(\cdot \mid X) = \psi$ in $L_2(\mu)$ when X is drawn from the stationary distribution ϕ. This result was was already established for the $L_1(\mu)$ case in the proof of lemma 6.1. The proof of the $L_2(\mu)$ case is similar. We verify only that $\mathbb{E}\|q(\cdot \mid X)\|_2 < \infty$, in which case the expectation $\mathcal{E} q(\cdot \mid X)$ is well-defined in $L_2(\mu)$. For this, it suffices to show that

$$\mathbb{E}\|q(\cdot \mid X)\|_2^2 = \mathbb{E} \int q(y \mid X)^2 \mu(dy)$$

is finite. In view of (8), this quantity is bounded above by $\mathbb{E}V(X)$. Finiteness of $\mathbb{E}V(X)$ follows from V-uniform ergodicity. Indeed, for every V-uniformly ergodic process with stationary distribution ϕ, the term $\int V d\phi$ is always finite (Meyn and Tweedie, 2009). □

References

[1] Aiyagari, S Rao, 1994. "Uninsured Idiosyncratic Risk and Aggregate Saving," *The Quarterly Journal of Economics*, Vol. 109(3), pages 659-84.

[2] Bosq, Denis, 2000. *Linear Processes in Function Space,* Springer.

[3] Brandt, Michel W. and Pedro Santa-Clara (2002): "Simulated Likelihood Estimation Of Diffusions With An Application To Exchange Rate Dynamics In Incomplete Markets," *Journal of Financial Economics,* 63 (2), 161–210.

[4] Braun, R. A., H. Li and J. Stachurski (2009): "Computing Stationary Expectations for Markov Models," mimeo, Tokyo University.

[5] Brock, W. A. and L. Mirman, 1972, "Optimal Economic Growth and Uncertainty: The Discounted Case," *Journal of Economic Theory*, Vol. 4, pages 479-513.

[6] Chib, S., F. Nardari and N. Shepard (2002): "Markov chain Monte Carlo methods for stochastic volatility models," *Journal of Econometrics,* 108 (2), pp. 281-316.

[7] Danielsson, Jon (1994): "Stochastic volatility in asset prices estimation with simulated maximum likelihood," *Journal of Econometrics,* 64 (1-2), pages 375–400.

[8] Devroye Luc and Gabor Lugosi, 2001 "Combinatorial Methods in Density Estimation" Springer-Verlag, New York.

[9] Durham, Garland B. and A. Ronald Gallant (2002): "Numerical Techniques for Maximum Likelihood Estimation of Continuous-Time Diffusion Processes," *Journal of Business and Economic Statistics,* 20 (3) pp. 297-338.

[10] Gelfand, Alan E. and Adrian F.M. Smith (1990) "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association*, Vol. 85, pg.398-409.

[11] Glynn, Peter W. and Shane G. Henderson, 2001, "Computing densities for Markov chains via simulation," *Mathematics of Operations Research* Vol. 26, pages 375-400.

[12] Hansen, Lars P. and Thomas J. Sargent (2007) "Robustness" Princeton University Press. Poon S. and Clive Granger (2003) "Forecasting volaility in financial markets: a review." *Journal of Economic Literature* Vol. 41 pg.478-539.

[13] Kristensen, Dennis, 2008. "Uniform Ergodicity of a Class of Markov Chains with Applications to Time Series Models," mimeo, Columbia University.

[14] Meyn, S. and R. L. Tweedie (2009): *Markov Chains and Stochastic Stability,* 2nd Edition, Cambridge University Press.

[15] Nishimura, Kazuo and John Stachurski, 2005, "Stability of Stochastic Optimal Growth Models: A New Approach," *Journal of Economic Theory,* 122 (1), pp. 100–118.

[16] Pedersen, A. (1995): "A New Approach to Maximum Likelihood Estimation for Stochastic Differential Equations Based on Discrete Observations," *Scandinavian Journal of Statistics,* 22, 55–71.

[17] Rosenthal, Jeffrey S., 1995, "Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo," *Journal of the American Statistical Association,* 90, pages 558–566.

[18] Stachurski, John and Vance Martin, 2008, "Computing the Distributions of Economic Models via Simulation," *Econometrica*, Vol. 76(2), pages 443-450.

[19] Stachurski, John, 2009, "A Hilbert Space Central Limit Theorem for Geometrically Ergodic Markov Chains," mimeo, Kyoto University.

[20] Tauchen, George, 1986, "Finite State Markov Chain Approximations to Univariate and Vector Autoregressions." *Economic Letters*, Vol. 20, pages 507-532.