Parametric Conditional Monte Carlo Density Estimation

Yin Liao[*]
John Stachurski[†]

ANU Working Papers in Economics and Econometrics
No. 562

October 2011:

ISBN: 0 86831 562 1

[*] School of Finance, Actuarial Studies and Applied Statistics, Australian National University
E-mail address: yin.liao@anu.edu.au

[†] Research School of Economics, Australian National University
E-mail address: john.stachurski@anu.edu.au

PARAMETRIC CONDITIONAL MONTE CARLO DENSITY ESTIMATION

YIN LIAO AND JOHN STACHURSKI

ABSTRACT. In applied density estimation problems, one often has data not only on the target variable, but also on a collection of covariates. In this paper, we study a density estimator that incorporates this additional information by combining parametric estimation and conditional Monte Carlo. We prove an approximate functional asymptotic normality result that illustrates convergence rates and the asymptotic variance of the estimator. Through simulation, we illustrate the strength of its finite sample properties in a number of standard econometric and financial applications.

1. INTRODUCTION

The classical problem of estimating the density f of a random vector Y_t is most often studied and carried out using observations of Y_t alone. However, in many practical settings, we have at hand observations of other random variables that are correlated with or otherwise related to Y_t. If we possess a model that gives some identification of the relationship between Y_t and these covariates, then a natural idea is to try to use observations of these covariates as additional data, in order to improve our estimate of the density of Y_t.

The particular setting we consider here is the following: In addition to the original data $Y_1, \ldots, Y_n$, we also observe a vector of covariates $X_1, \ldots, X_n$ where $\{X_t\}$ is a stationary and ergodic stochastic process taking values in set $\mathbb{X}$. Suppose further that Y_t is related to X_t via

$$Y_t = G(X_t, \xi_t, \theta) \qquad \text{for some} \qquad \theta \in \Theta \tag{1}$$

The function G is assumed known up to the vector of parameters θ, while $\{\xi_t\}$ is IID and unobservable. For now $\{X_t\}$ is taken to be fully observable, although more general cases are considered later on. We also assume parametric knowledge about the process generating $\{X_t\}$. In particular, we suppose that

$$X_{t+1} = H(X_t, \eta_{t+1}, \theta) \qquad \text{with} \qquad \{\eta_t\}_{t \geq 1} \overset{\text{IID}}{\sim} \nu \tag{2}$$

Date: October 9, 2011.

The authors thank Varang Wiriyawit for excellent research assistance.

As stated above, the process (2) is taken to be stationary and ergodic.

Given that $\{X_t\}$ and $\{\xi_t\}$ are both stationary, the target process $\{Y_t\}$ is likewise stationary. We let

- $f = f(\cdot, \theta)$ denote the common stationary (i.e., marginal) density of Y_t
- $\phi = \phi(\cdot, \theta)$ denote the common stationary distribution of X_t

The relationship between f and ϕ can be expressed in terms of the conditional density implied by the relationship (1). Letting $p(\cdot \,|\, x, \theta)$ be the density of the random variable $G(x, \xi_t, \theta)$,[1] the conditional density of Y_t given X_t is $p(\cdot \,|\, X_t, \theta)$, and integrating over X_t recovers the marginal density of Y_t. That is,

$$f(y, \theta) = \int p(y \,|\, x, \theta) \phi(dx, \theta) \qquad (y \in \mathbb{Y}) \tag{3}$$

The problem considered here is estimation of the density $f(\cdot, \theta)$ given the observed data $\{(X_t, Y_t)\}_{t=1}^n$. In addition to the data, we make use of the information provided by the parametric relationships (1) and (2).

Some preliminary comments about the formulation of the problem are in order. First, the assumption that $\{X_t\}$ satisfies (2) is far less restrictive than it appears. Most models with extra lagged state variables and correlated shock processes can be expressed in the form of (2) by redefining the state vector. In fact it is well-known that any time homogeneous Markov process on a separable and completely metrizable space can be be expressed in the form of (2) for suitable choice of H and $\{\eta_t\}_{t \geq 1}$.[2] Similarly, the specification (1) can accommodate correlated shocks by redefining X_t to include any non-IID variables.

A second remark is that although the law of motion H in (2) and the function G in (1) share the same parameter vector θ, this is just for notational convenience. In many applications, H will depend on some vector of parameters γ and G will depend on a second and unrelated parameter vector β. However, in this case we take $\theta := (\beta, \gamma)$ and adjust the definitions of G and H accordingly. Thus, θ should simply be regarded as a vector that contains all of the unknown parameters in our setting.

A third remark is that although the stationary distribution ϕ is formally defined by the model (2) for each θ, outside of the linear Gaussian case it is typically intractable. When ϕ is intractable, or even when it is not, the integral in (3) will usually be intractable, and in most cases there will be no analytical

[1] Existence requires that the distribution of $G(x, \xi_t, b)$ is absolutely continuous for all x, b.

[2] For a proof see Bhattacharya and Majumdar (2007, proposition C1.1).

expression available for f. The estimation technique we consider in this paper accommodates this lack of analytical tractability via simulation.

In particular, to estimate f, the procedure we consider is:

(1) Use some estimator $\hat{\theta}_n$ to estimate the parameter vector θ.
(2) Generate m IID draws $\{\eta_t\}_{t=1}^m$ from their distribution v in (2), and compute the simulated process $\{X_t^s\}_{t=1}^m$ via

$$X_{t+1}^s = H(X_t^s, \eta_{t+1}, \hat{\theta}_n), \qquad X_0^s = x \in \mathbb{X} \tag{4}$$

(3) Return the estimate

$$\hat{f}_m(y, \hat{\theta}_n) := \frac{1}{m} \sum_{t=1}^{m} p(y \mid X_t^s, \hat{\theta}_n) \tag{5}$$

Here and below, the superscript "s" is a mnemonic for simulation. In what follows, we refer to the estimator (5) as the parametric conditional Monte Carlo (PCMC) density estimator.[3]

The intuition behind convergence of the PCMC density estimator is as follows: The stationary distribution of (4) is $\phi(dx, \hat{\theta}_n)$ and, since the law of motion (2) is assumed to be ergodic, we have

$$\frac{1}{m} \sum_{t=1}^{m} p(y \mid X_t^s, \hat{\theta}_n) \approx \int p(y \mid x, \hat{\theta}_n) \phi(dx, \hat{\theta}_n) \quad \text{for large } m$$

If $\hat{\theta}_n$ is consistent for the true parameter θ_0 and m is large, then, assuming some degree of continuity, the right-hand size is approximately $\int p(y \mid x, \theta_0) \phi(dx, \theta_0)$. In view of (3) this integral is equal to $f(y, \theta_0)$, the true density of Y_t.

Details of the asymptotic properties of the PCMC density estimator are provided in section 2. We show that when $\hat{\theta}_n$ is $\sqrt{n}$-consistent for the true parameter θ_0, the PCMC density estimator is $\sqrt{n}$-consistent for the true density $f(\cdot, \theta_0)$ in the sense of L_2 deviation, modulo the error caused by simulation. The simulation error is itself of order $O(m^{-1/2})$. These result is established via an approximate functional central limit theorem.

In addition to situations where the density itself is of primary interest (e.g., density forecasting), the PCMC density estimator may be applied to a wide

[3] The choice of x in (4) is arbitrary, and our asymptotic results are valid for any selection.

range of statistical problems where density estimates are an input, such as discriminant (Fix and Hodges, 1951) and cluster (Gordon, 1981) analysis.[4] Density estimators can also be used to address specification testing or model validation problems (e.g., Aït-Sahalia et al., 2009, 2010). Finally, many statistical problems require estimates of functionals of the density, such as quantiles and hazard rates, and these functionals can be evaluated once the density is estimated.

A natural way to place the PCMC density estimator in the literature is to compare it to the density estimator proposed by Zhao (2009). Zhao assumes the same parametric knowledge contained in (1), but makes no parametric assumptions regarding the process $\{X_t\}$. He assumes only that this process is suitably stationary and ergodic. His procedure is to estimate the unknown parameters in (1) using some estimator $\hat{\theta}_n$, and then estimate f via

$$\hat{z}_n(y, \hat{\theta}_n) := \frac{1}{n} \sum_{t=1}^{n} p(y \mid X_t, \hat{\theta}_n) \tag{6}$$

Zhao shows consistency and asymptotic normality of $\hat{z}_n$ under rather general conditions (Zhao, 2009, theorem 1). The difference between $\hat{z}_n$ and our estimator $\hat{f}_m$ is that $\hat{f}_m$ uses a parametric assumption about the $\{X_t\}$ process to produce simulated X-data, while Zhao's estimator uses observed X-data. For $\hat{f}_m$, the simulated data size m will typically be much larger than n (the size of the observed data set).

It goes without saying that the PCMC density estimator and $\hat{z}_n$ are not directly comparable. Zhao's estimator makes no parametric assumptions about the process $\{X_t\}$, and is therefore more robust to misspecification. On the other hand, the PCMC density estimator has the usual finite sample advantages parametric methods enjoy over nonparametric methods when the parametric specification is correct. However, what can be said here is that *if* we do assume that the parametric specification is correct, then several features of the current setting imply that the PCMC density estimator has several important advantages *over and above these usual finite sample advantages*. These points are addressed in detail in section 3.

[4] The basic problem behind discriminant analysis is: Given a sample known to come from a population A, a sample known to come from population B, and a new observation Z, does Z come from population A or B? Density estimates for population A and population B can be used to classify Z. Cluster analysis is used to divide a given population into a number of classes. Density estimation can be used to define a hierarchical structure on a set of samples in order to discover classes.

While the PCMC density estimator introduced in this paper is most closely related to the estimator of Zhao (2010) described above, the idea of using observations of covariates to improve density estimates can be found in a number of other papers. Other parametric density estimates using this idea were proposed by Saaverdra and Cao (2000), Schick and Wefelmeyer (2004, 2007) for linear processes, by Frees (1994) and Gine and Mason (2007) for functions of independent variables, and by Kim and Wu (2007) for nonlinear autoregressive models of order one with constant variance.

The structure of our paper is as follows:

2. Consistency and Asymptotic Normality

In this section we discuss theoretical properties of the PCMC density estimator. To simplify the arguments, we assume throughout this section that the stationary distribution $\phi(dx, \theta)$ of X_t can be expressed as a density (with respect to Lebesgue measure) for all $\theta \in \Theta$. This density will be written as $\phi(x, \theta)dx$. In addition, when our parametric assumptions in (1) and (2) are taken as valid, we let θ_0 represent the true value of the parameter vector θ. It follows that the true density of Y_t is given by

$$f(y, \theta_0) := \int p(y \,|\, x, \theta_0)\phi(x, \theta_0)dx$$

The set Θ is taken to be a subset of $\mathbb{R}^M$. To simplify notation, in the sequel we let

$$d(x, y, \theta) := \phi(x, \theta) \begin{pmatrix} \frac{\partial}{\partial \theta_1} p(y \,|\, x, \theta) \\ \vdots \\ \frac{\partial}{\partial \theta_M} p(y \,|\, x, \theta) \end{pmatrix} + p(y \,|\, x, \theta) \begin{pmatrix} \frac{\partial}{\partial \theta_1} \phi(x, \theta) \\ \vdots \\ \frac{\partial}{\partial \theta_M} \phi(x, \theta) \end{pmatrix}$$

whenever the derivatives exist. In particular, $d(x, y, \theta)$ is the M-vector of partial derivatives obtained by differentiating the product $p(y \,|\, x, \theta)\phi(x, \theta)$ with respect to θ, holding x and y constant.

Below we present an approximate L_2 central limit theorem for the deviation between the PCMC density estimator $\hat{f}_m(\cdot, \hat{\theta}_n)$ and the true density $f(\cdot, \theta_0)$. In what follows, we take $\mathbb{Y}$ to be a Borel subset of $\mathbb{R}^d$, and the symbol $L_2(\mathbb{Y})$ represents the set of (equivalence classes of) Borel measurable functions that are square integrable with respect to Lebesgue measure. As usual, the inner product of two elements g and h of $L_2(\mathbb{Y})$ is defined as $\langle g, h \rangle := \int g(y)h(y)dy$ and the norm is $\|g\| := \sqrt{\langle g, g \rangle}$.

Recall that a random element W of $L_2(\mathbb{Y})$ is called *centered Gaussian* if $\langle h, W\rangle$ is zero-mean Gaussian on $\mathbb{R}$ for all $h \in L_2(\mathbb{Y})$. Alternatively, W is centered Gaussian if its characteristic function has the form

$$\psi_W(h) := \mathbb{E}\exp\{i\langle h, W\rangle\} = \exp\left\{\frac{-\langle h, Ch\rangle}{2}\right\} \qquad (h \in L_2(\mathbb{Y}))$$

for some positive self-adjoint linear self-mapping C on $L_2(\mathbb{Y})$. C is called the covariance operator of W. It also satisfies (and is uniquely defined by) $\langle g, Ch\rangle := \mathbb{E}\langle g, W\rangle \langle h, W\rangle$ for all $g, h \in L_2(\mathbb{Y})$.[5]

To prove the main result of this section, we require some differentiability and ergodicity assumptions. Our differentiability assumption is as follows:

Assumption 2.1. There exists an open neighborhood V of the true parameter θ_0 and a function $g\colon \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ such that

(1) The function g satisfies $\int \left\{\int g(x,y)dx\right\}^2 dy < \infty$.
(2) The maps $\theta \mapsto p(y \mid x, \theta)$ and $\theta \mapsto \phi(x, \theta)$ are continuously differentiable over the neighborhood V for all fixed $(x, y) \in \mathbb{X} \times \mathbb{Y}$.
(3) The vector $d(x, y, \theta)$ of partial derivatives satisfies

$$\sup_{\theta \in V} \|d(x,y,\theta)\|_E \leq g(x,y) \quad \text{for all} \quad (x,y) \in \mathbb{X} \times \mathbb{Y}$$

In assumption 2.1, the symbol $\|\cdot\|_E$ is the euclidean norm on $\mathbb{R}^M$. (The subscript E is used to differentiate the euclidean norm from the L_2 norm $\|\cdot\|$.)

Assumption 2.2. The process $\{X_t\}$ is V-uniformly ergodic, with unique stationary density $\phi(\cdot, \theta)$ on $\mathbb{X}$.

The V-uniform ergodicity condition is quite standard, and a precise definition can be found in Meyn and Tweedie (2009, chapter 16). The condition is attractive because it combines widespread applicability with relatively strong implications (in terms of laws of large numbers and central limit theorems). Kristensen (2008) gives detailed condition for V-uniform ergodicity of many common time series models, including linear and nonlinear state space models, VARMA models, nonlinear ARMA models, random coefficient models, bilinear models, and a variety of univariate and multivariate GARCH models. Nishimura and Stachurski (2005) establish V-uniform ergodicity for the stochastic optimal growth model under weak Inada-type conditions.

Assumption 2.3. The sequence $\{\hat{\theta}_n\}$ is asymptotically normal, in the sense that $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Sigma)$ for some symmetric positive definite matrix $\Sigma = (\sigma_{ij})$.

[5] Further details on Hilbert space valued random variables can be found in Bosq (2000).

Next we present a functional central limit theorem for the error. The theorem states that the distribution of the error is well approximated by a centered Gaussian on $L_2(\mathbb{Y})$ when the simulation and sample sizes are both large.

Theorem 2.1. *Let $\alpha_m(y) := \int d_m(x, y, \theta_0)dx$ for $m = 1, \ldots, M$. If assumptions 2.1–2.3 are valid, then*

$$\sqrt{n}\{\hat{f}_m(\cdot, \hat{\theta}_n) - f(\cdot, \theta_0)\} = \frac{1}{\sqrt{m}} O_P(1) + W_n \tag{7}$$

where W_n converges in distribution to a centered Gaussian in $L_2(\mathbb{Y})$ with covariance operator C defined by

$$\langle g, Ch \rangle = \sum_{i=1}^{M} \sum_{j=1}^{M} \sigma_{ij} \langle \alpha_i, g \rangle \langle \alpha_j, h \rangle \qquad (h, g \in L_2(\mathbb{Y}))$$

In the theorem, $\sqrt{n}\{\hat{f}_m(\cdot, \hat{\theta}_n) - f(\cdot, \theta_0)\}$ is treated as a random element of $L_2(\mathbb{Y})$. The expression $m^{-1/2}O_P(1) + W_n$ on the right-hand side of (7) should be interpreted to mean $m^{-1/2}U_{m,n} + W_n$ where $\{U_{m,n}\}$ is a collection of random elements in $L_2(\mathbb{Y})$ such that $\|U_{m,n}\|$ is bounded in probability over m for every fixed n. Since m is the simulation size and can be made arbitrarily large relative to n, the effect of the term $m^{-1/2}O_P(1)$ will typically be negligible.

The functions α_m in the definition of the covariance operator C are defined by $\alpha_m(y) := \int d_m(x, y, \theta_0)dx$, which is the integral of the partial derivative of $p(y \mid x, \theta)\phi(x, \theta)$ with respect to θ. As shown in section 6, under the conditions of the theorem, the order of the derivative and integral can be reversed, so

$$\alpha_m(y) = \frac{\partial}{\partial \theta_m} \int p(y \mid x, \theta_0)\phi(x, \theta_0)dx = \frac{\partial}{\partial \theta_m} f(y, \theta_0)$$

Using this fact, we can express C as the integral operator with kernel

$$k(y, y') := D_\theta f(y, \theta_0)^\top \Sigma D_\theta f(y', \theta_0)$$

In particular, C is defined from k via

$$\langle g, Ch \rangle := \int \int k(y, y')g(y)h(y')dydy'$$

(In the definition of k, $D_\theta f$ represents the vector of partial derivates of f with respect to θ.) Thus, the asymptotic variance in the density estimator reflects the variance in the parameter estimate $\hat{\theta}_n$ transferred via the slope of the density estimate with respect to the parameters in the neighborhood of the true parameter.

3. Semiparametric vs Parametric Estimation

In this paper, the aim is to estimate f efficiently by exploiting the existence of data on correlated variables via the relationship (3). In order to exploit (3), one must estimate both p and ϕ. The PCMC density estimator estimates p parametrically, and ϕ by a combination of parametric estimation and simulation. For comparison, recall that Zhao's estimator is given by (6). Letting ϕ_n be the empirical distribution of the sample $X_1, \ldots, X_n$, this can also be written as

$$\hat{z}_n(y, \hat{\theta}_n) = \int p(y \,|\, x, \hat{\theta}_n) \phi_n(dx) \tag{8}$$

While p is estimated parametrically, the distribution ϕ is unrestricted, and estimated by the nonparametric empirical distribution ϕ_n. Thus, Zhao's estimator is a semiparametric estimator, differing from the PCMC density estimator only in the way that ϕ is estimated.

As discussed in the introduction, the PCMC density estimator and Zhao's estimator are not directly comparable. Zhao's estimator requires no parametric specification of the dynamics of $\{X_t\}$, and hence is more robust to misspecification. At the same time, since the empirical distribution is globally $\sqrt{n}$-consistent, Zhao's semiparametric estimator retains the $\sqrt{n}$-consistency of the parametric scheme.

On the other hand, if the parametric specification of the $\{X_t\}$ process in the PCMC density estimator is correct, then the additional information embedded in the parametric model can improve finite sample properties. In the current setting, however, there are further advantages of the parametric approach that are deeper, and less immediately apparent. Provided that the estimation procedure is structured to exploit these additional advantages, the gains from the parametric alternative can be large. These ideas are described in the remainder of this section.

3.1. **Preservation of Dependence Structure.**

For the PCMC density estimator, using the parametric model (2) to estimate the data generating process for $\{X_t\}$ provides, in addition to the extra structure from the parametric model, the ability to preserve and exploit the dependence structure in the data $X_1, \ldots, X_n$. This dependence structure is discarded in Zhao's estimator, because the estimator is invariant to the order of the sample. Indeed, the difference between the PCMC density estimator and Zhao's estimator is that the latter estimates ϕ using the empirical distribution, and the empirical distribution is invariant to any reordering of $X_1, \ldots, X_n$.

Of course this is an unfair comparison, since Zhao's estimate of ϕ is nonparametric and hence robust to parametric misspecification. To make a more natural comparison, consider the following estimation technique, which is a direct parametric alternative to Zhao's estimator, and conceptually lies between Zhao's estimator and the PCMC density estimator:

(1) Specify a parametric form $\phi(x, \gamma)$ for the stationary density of X_t.
(2) Estimate the parameters θ in the function (1) and γ in the density $\phi(x, \gamma)$ via some estimators $\hat{\theta}_n$ and $\hat{\gamma}_n$.
(3) Return the estimate

$$\hat{z}_n^P(y) := \int p(y \mid x, \hat{\theta}_n) \phi(x, \hat{\gamma}_n) \tag{9}$$

The difference between Zhao's estimator $\hat{z}_n$ and $\hat{z}_n^P$ is that $\hat{z}^P$ estimates ϕ parametrically. The difference between $\hat{z}^P$ and the PCMC density estimator is that, in the case of $\hat{z}^P$, parametric specification is placed directly on the stationary density ϕ of X_t, rather than specifying a data generating process for $\{X_t\}$ such as (2). (Note also that $\hat{z}^P$ also uses an exact integral in (9), rather than a simulation-based approximation like the PCMC density estimator. In applications this integral will rarely be tractable, and hence $\hat{z}^P$ is not a practical alternative. We are interested in $\hat{z}^P$ only to the extent that it is useful to illustrate the value of preserving dependence structure in $\{X_t\}$.)

Like Zhao's estimator, the estimator $\hat{z}_n^P$ will typically discard information on the dependence structure of $\{X_t\}$, in the sense that it will be invariant to the order of $X_1, \ldots, X_n$. For example, suppose we believe that the process for $\{X_t\}$ is a linear process $X_{t+1} = AX_t + BW_{t+1}$ with $\{W_t\}$ a vector zero mean Gaussian. To implement the estimator $\hat{z}_n^P$, we observe that the stationary distribution ϕ corresponding to this law of motion is of the form $\phi(\cdot, \gamma) = N(0, \gamma)$ for some covariance matrix γ. The maximum likelihood estimator of γ is

$$\hat{\gamma}_n := \frac{1}{n} \sum_{t=1}^{N} (X_t - \bar{X})(X_t - \bar{X})^\top$$

This estimator is invariant to any permutation of the sample $X_1, \ldots, X_n$, and hence the dependence information in this sample is not used to estimate γ, or anywhere else in the estimator $\hat{z}_n^P$.

On the other hand, the PCMC density estimator would start by estimating the unknown parameters in the process $X_{t+1} = AX_t + BW_{t+1}$ directly. A standard estimator such as least squares would not be invariant to a permutation of $X_1, \ldots, X_n$. Thus, the information in the order of the X sample is not discarded. In this sense, the PCMC density estimator uses parametric assumptions to not only provide additional structure, but also to allow the information in the order of $X_1, \dots, X_n$ to be exploited. The gains from exploiting this information are discussed in sections 4.1 and 4.2.

3.2. **Latent Variables.** Another advantage of the parametric approach to our density estimation problem is that it allows us to treat latent variables. A large number of modern time series estimation techniques include some kind of latent variables. Examples include latent state space, latent factor and hidden Markov models, regime switching models, GARCH models and stochastic volatility models. In all these models, the process $\{X_t\}$ is not fully observable, and the empirical distribution of $\{X_t\}_{t=1}^n$ cannot be computed. Thus, the semiparametric estimator z_n in (6) cannot be directly implemented.

On the other hand, the PCMC density estimator can be applied to all of the above examples. For example, if the vector X_t contains a latent volatility term, we can estimate the dynamics of the process using a GARCH or stochastic volatility model. Once the dynamics of the process $\{X_t\}$ are estimated, the process can be simulated and the PCMC density estimator can be constructed. Applications along these lines are presented in sections 4.4 and 4.5.

3.3. **Autoregressive Models.** As discussed above, the PCMC density estimator obtains some important benefits from using a parametric specification for the DGP of $\{X_t\}$. Estimating the parametric specification $X_{t+1} = H(X_t, \eta_{t+1}, \theta)$ allows us to exploit the information contained in the order of the sample $\{X_t\}_{t=1}^n$, and also to work with latent variables. Moreover, if the parametric specification is correct, then the parametric structure will aid estimation of the stationary distribution ϕ in finite samples, particularly helps for the processes which exhibit a strong persistence.

Obviously, there is a cost to using a parametric specification for the DGP of $\{X_t\}$: the risk of misspecification. Misspecification typically lead to a poor estimate of ϕ relative to the empirical distribution used in Zhao's estimator $\hat{z}_n$. However, there is an important special case where this risk is ameliorated: When X_t consists only of lagged valued of Y_t. The reason is that, for either the PCMC density estimator or Zhao's estimator, the first step is to specify and estimate the relationship $Y_t = G(X_t, \xi_t, \theta)$ in (1). Let's assume the specification is correct—a necessary condition for consistency of both estimators. If X_t consists only of lagged valued of Y_t, then estimating this relationship is equivalent to estimating the DGP of $\{Y_t\}$. Moreover, since X_t is lagged valued of Y_t, knowing the DGP of $\{Y_t\}$ means knowing the DGP of $\{X_t\}$. Thus, we have no extra risk of misspecifying the DGP of $\{X_t\}$.

Examples are presented in the applications given below.

4. Applications

In this section, we apply the PCMC density estimator to a number of common models. Using simulation, we examine the finite sample performance of the PCMC estimator relative to other density estimators. In all cases, performance is measured in terms of mean integrated squared error (MISE). The MISE of an estimator $\hat{g}$ of an arbitrary density g has the standard definition $\mathbb{E}\|\hat{g} - g\|^2$ where, as above, $\|\cdot\|$ is the L_2 norm. In all the following simulations, the MISE is approximated by averaging 10^3 realizations of $\|\hat{g} - g\|^2$.[6]

4.1. **Dynamic factor model.** As a relatively simple illustration, consider first a linear dynamic factor model

$$Y_t = \beta^\top X_t + \xi_t$$

with

$$X_{t+1} = \begin{pmatrix} \gamma_1 & 0 & 0 \\ 0 & \gamma_2 & 0 \\ 0 & 0 & \gamma_3 \end{pmatrix} X_t + \eta_{t+1} \quad \text{and} \quad \theta := \begin{pmatrix} \beta \\ \gamma \end{pmatrix} \tag{10}$$

Here $Y_t \in \mathbb{R}$, $X_t \in \mathbb{R}^3$, and all shocks are independent and standard normal. The PCMC density estimator is given by (5), where, in the present case,

$$p(y \mid X_t^s, \hat{\theta}_n) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y - \hat{\beta}_n^\top X_t^s)^2\right\}$$

The sequence $X_1^s, \ldots, X_m^s$ is produced by estimating the parameters in (10) and then simulating from some arbitrary initial condition x.

In order to study the finite sample properties of the PCMC density estimator we run a simulation that computes the MISE of the estimate when $n = 200$, and compares it with several other estimators.[7] Following the asset pricing analysis of He *et al.* (2010), parameters are set to $\beta_1 = 6.26$, $\beta_2 = 1.32$, $\beta_3 =$

[6] If the true density g has no closed form solution, then we compute it by simulation. In the simulation studies below, g corresponds to the true density $f = f(\cdot, \theta_0)$ where θ_0 is the set of parameters chosen for the simulation. We compute f by using the PCMC density estimator, but without the parameter estimation step. In other words, we simulate $\{X_t^s\}$ from $X_{t+1}^s = H(X_t^s, \eta_{t+1}, \theta_0)$ and compute f as $m^{-1} \sum_{t=1}^m p(y \mid X_t^s, \theta_0)$. In all cases we set $m = 10^5$.

[7] For the PCMC density estimator, all parameter estimates use least squares.

PCMC	$\hat{z}_n^P$	$\hat{z}_n$	OPE	NPKDE
1.000	1.965	2.256	2.531	3.014

TABLE 1. MISE for dynamic factor model

-1.09, $\gamma_1 = 0.18$, $\gamma_2 = -0.14$, and $\gamma_3 = 0.21$. Estimates are compared against the true stationary density for Y_t, which in this case is equal to

$$f = N(0, \sigma^2) \quad \text{for} \quad \sigma^2 := \frac{\beta_1^2}{1-\gamma_1^2} + \frac{\beta_2^2}{1-\gamma_2^2} + \frac{\beta_3^2}{1-\gamma_3^2} + 1 \tag{11}$$

For comparison, we also compute the MISE of four alternatives to the PCMC density estimator. One is Zhao's estimator $\hat{z}_n$. (So $\hat{z}_n$ can be implemented, we assume that data on the process $\{X_t\}$ is observable.) The second is the direct parametric alternative $\hat{z}_n^P$ defined in (9), which estimates ϕ parametrically but without exploiting the dependence structure of $\{X_t\}$. The last two are direct estimates of f that use only observations of $\{Y_t\}$. One is an ordinary parametric estimate (OPE), and the second is a nonparametric kernel density estimate (NPKDE). The OPE uses the dynamic factor model to infer that the stationary distribution of Y_t has the form $N(0, \sigma^2)$ obtained in (11), and estimates f as $N(0, \hat{\sigma}_n^2)$ where $\hat{\sigma}$ is the sample standard deviation of $\{Y_t\}$. The NPKDE uses $Y_1, \ldots, Y_n$ as the sample, a standard Gaussian kernel, and Silverman's rule for the bandwidth.

The results of the simulation are shown in Table 1. The MISE value for the PCMC density estimator was 4.606×10^{-4}. In the table, all estimators are expressed relative to this base (i.e., as multiples of this value). The reduction in MISE from the NPKDE to the OPE represents the benefit of imposing parametric structure on the data set $\{Y_t\}$, at least when that parametric specification is correct. The reduction in MISE from the OPE to Zhao's estimator represents the benefit of exploiting the relationship (3) and the second data set $\{X_t\}$. The reduction in MISE from Zhao's estimator to $\hat{z}_n^P$ represents the gains from estimating ϕ parametrically (when the parametric specification is correct). The final reduction in MISE from $\hat{z}_n^P$ to the PCMC represents the gain from exploiting the information contained in the order of the sample $\{X_t\}$.

4.2. **Linear AR(1).** In this section we study another very simple example in order to illustrate conceptual issues: the scalar, linear Gaussian AR(1) model

$$Y_t = \theta Y_{t-1} + \xi_t, \qquad \{\xi_t\} \overset{\text{IID}}{\sim} N(0,1), \quad |\theta| < 1 \tag{12}$$

To estimate the stationary density f of Y_t via the PCMC density estimator, we take $X_t := Y_{t-1}$. In this case (12) implies that $p(y \,|\, x, \hat{\theta}_n) = N(\hat{\theta}_n x, 1)$,

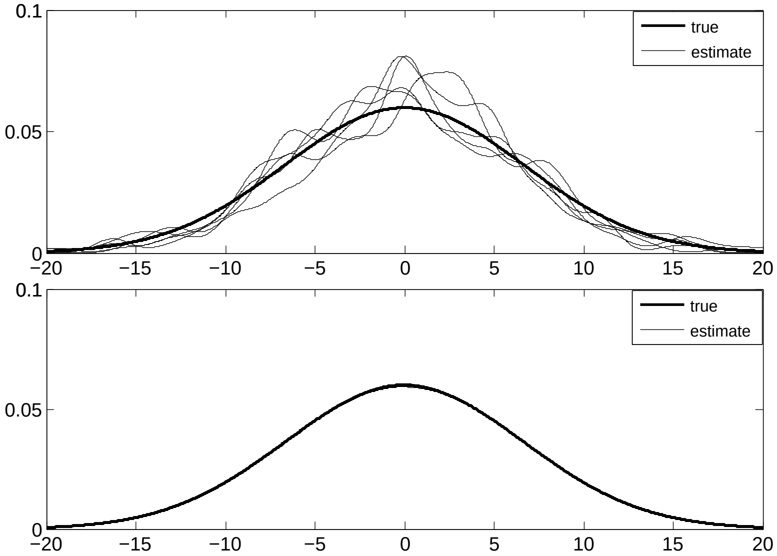

FIGURE 1. Dynamic factor model, $\hat{z}_n$ (top) and PCMC (bottom)

where $\hat{\theta}_n$ is the least squares estimate of θ. Since $X_t = Y_{t-1}$, we can produce the simulated X_t data using (12) as well, by iterating on $X^s_t = \hat{\theta}_n X^s_{t-1} + \xi_t$. Combining this data with $p(y \mid x, \hat{\theta}_n)$ yields the PCMC density estimator in (5).

To study the MISE of the estimator in finite samples, we compute the MISE of the PCMC density estimator when $n = 200$ and $\theta = 0.9$. For comparison we also compute the MISE of Zhao's estimator $\hat{z}_n$, the direct parametric alternative $\hat{z}^P_n$, and the NPKDE. (The ordinary parametric estimate (OPE) is omitted because in the present setting X_t is lagged Y_t, so f and ϕ are equal, and hence the OPE amounts to the same estimator as $\hat{z}^P_n$.) The methods for NPKDE is identical to that used in section 4.1.

The results are reported in table 2. As in table 1, all estimators are expressed as multiples of the MISE for the PCMC density estimator.[8] The ranking of MISE values is similar to that obtained for the dynamic factor model in section 4.1.

[8] The MISE value for the PCMC density estimator was 2.100×10^{-3}.

PCMC	$\hat{z}_n^P$	$\hat{z}_n$	NPKDE
1.000	1.429	3.714	3.827

TABLE 2. MISE for AR(1) model when $\theta = 0.9$

θ	0.1	0.3	0.5	0.7	0.9
PCMC	0.023	0.104	0.309	0.717	2.100
$\hat{z}_n$	0.027	0.194	0.710	2.100	7.800

TABLE 3. MISE for AR(1) model (base 1×10^{-3})

Table 3 is used to illustrate the point made in section 3.1: The PCMC density estimator's use of a parametric model for the DGP $\{X_t\}$ provides the ancillary benefit of exploiting the information contained in the order of the sample $X_1, \ldots, X_n$. The table compares the MISE for the PCMC density estimator to that of Zhao's estimator as θ varies from 0.1 to 0.9. The results are given in table 3. While the MISE for the PCMC density estimator is lower than Zhao's estimator for all values of θ, the difference becomes more pronounced as θ increases (from a factor of 1.2 at $\theta = 0.1$ to a factor of 3.7 at $\theta = 0.9$). Intuitively, when $\theta = 0.1$ the data is almost IID, and preserving the order information in an estimate of ϕ has little value. On the other hand, when $\theta = 0.9$, the data is very persistent, and the value of preserving this order information is higher.

4.3. **Threshold autoregressive model.** As our next application, we replace the linear AR(1) model with the TAR model

$$Y_t = \theta |Y_{t-1}| + \sqrt{1-\theta^2}\xi_t, \qquad \{\xi_t\} \overset{\text{IID}}{\sim} N(0,1), \quad |\theta| < 1$$

The stationary density of Y_t in this model has the skew-normal form $f(y) = 2\psi(y)\Psi(\delta y)$, where $\delta := \theta / \sqrt{1-\theta^2}$, and ψ and Ψ are the standard normal density and cumulative distribution respectively (see Andel et al. (1984)). The parameter θ can be estimated by maximum likelihood.

In the simulation we set $\theta = 0.5$ and $n = 200$. The results are reported in table 4. As before, all estimators are expressed as multiples of the MISE for the PCMC density estimator.[9] Because the TAR model is nonlinear, the target density is more complex, and the finite sample advantages of using correct parametric structure are correspondingly larger. This fact is reflected in the

[9] The MISE value for the PCMC density estimator was 9.345×10^{-7}. The parametric alternative $\hat{z}_n^P$ and ordinary parametric estimate are not available for comparison in this model.

PCMC	$\hat{z}_n^P$	$\hat{z}_n$	NPKDE
1.000	1.572	807.325	2078.429

TABLE 4. MISE for TAR model when $\theta = 0.5$

relative magnitudes of the MISE, which exhibit a much larger gain from using the PCMC density estimator than was the case with the AR(1) model in section 4.2 (compare table 2 and table 4).

4.4. **Markov regime switching model.** Next we consider a Markov regime switching model in order to illustrate how the PCMC estimator is implemented to estimate the density of Y_t in latent variable models. Regime switching models have been used widely in economic and financial applications. The model we consider here is given by

$$Y_t = \mu(X_t) + \sigma(X_t)\xi_t = \mu_{X_t} + \sigma_{X_t}\xi_t$$

where $\{X_t\}$ is a two-state ergodic Markov chain with transition matrix P, and ξ_t is IID normal with zero mean and unit variance. The stationary density of Y_t has a closed form

$$f = N(\mu_1, \sigma_1^2) \times \pi_1 + N(\mu_2, \sigma_2^2) \times \pi_2$$

where π is the stationary distribution of P.

The regime switching model can be estimated using maximum likelihood (see, e.g., Hamilton 1994). Once the model is estimated, the PCMC density estimator can be implemented to obtain an estimate of f. In this case, the conditional density p in (5) is $p(y \mid X_t^s, \hat{\theta}_n) = N(\hat{\mu}_{X_t^s}, \hat{\sigma}^2_{X_t^s})$. The values $\{X_t^s\}$ are simulated from an estimate $\hat{P}$ of the matrix P.

We investigated the finite sample performance of the PCMC estimator by comparing the MISE with that of the NPKDE. (Zhao's estimator is not available for comparison in this model, because the state X_t is latent.) In the simulation, we took $n = 500$. The parameters were set according to Smith and Layton's (2007) business cycle analysis, where $\mu_1 = 0.34$, $\mu_2 = -0.13$, $\sigma_1 = 0.38$, $\sigma_2 = 0.82$, and

$$P = \begin{pmatrix} 0.97 & 0.03 \\ 0.08 & 0.92 \end{pmatrix}$$

From the simulation, the MISE of the PCMC estimator was found to be 9.418×10^{-3}, while that of the NPKDE was 0.015. In other words, the MISE of the NPKDE was roughly 1.6 times larger.

4.5. **Stochastic volatility in mean model.** As another application of the PCMC density estimator in a latent variable setting, we consider the stochastic volatility in mean model

$$Y_t = c\sigma^2 \exp(h_t) + \sigma \exp(h_t/2)\xi_t \tag{13}$$

$$h_t = \kappa h_{t-1} + \sigma_\eta \eta_t \tag{14}$$

Typically, Y_t denotes return on a given asset, and the latent variable h_t denotes underlying volatility. The pair (ξ_t, η_t) is standard normal in $\mathbb{R}^2$ and IID. Parameters in the model can be estimated by simulated MLE (see, e.g., Koopman and Uspensky, 2002). We take h_t as the covariate X_t in the definition of the PCMC density estimator, which then has the form

$$\hat{f}_m(y) = \frac{1}{m} \sum_{t=1}^{m} p(y \,|\, h_t^s, \hat{\theta}_n)$$

where, in view of (13),

$$p(y \,|\, h, \hat{\theta}_n) := N(\hat{c}_n \hat{\sigma}_n^2 \exp(h), \hat{\sigma}_n^2 \exp(h))$$

and $\{h_t^s\}$ is generated by iterating on the estimated version of (14).

As with the Markov switching model, we investigated the finite sample performance of the PCMC estimator by comparing its MISE with that of the NPKDE. (Again, Zhao's estimator is not available for comparison here, because the covariate is latent.) In the simulation we took $n = 500$. We adopted the estimated parameter values in Koopman and Uspensky (2002), where $\kappa = 0.97$, $\sigma_\eta = 0.135$, $\sigma^2 = 0.549$, and $c = 1$.

For these parameters, we calculated the MISE of the PCMC estimator to be 1.524×10^{-4}, while that of the NPKDE was 3.048×10^{-4}. Thus, the MISE of the NPKDE was roughly 2.265 times larger. Typical realizations of the estimators are presented in figure 2.

5. Robustness

Regarding the PCMC density estimator, one concern is that its advantages stem from parametric specification of the DGP of $\{X_t\}$, and this specification may be inaccurate. In this section we take two models and investigate the performance of the PCMC estimator when the DGP is misspecified. The first model is a scalar version of the dynamic factor model in section 4.1, with

$$Y_t = \beta X_t + \xi_t \quad \text{and} \quad X_{t+1} = \gamma X_t + \eta_{t+1} \text{ with } |\gamma| < 1 \tag{15}$$

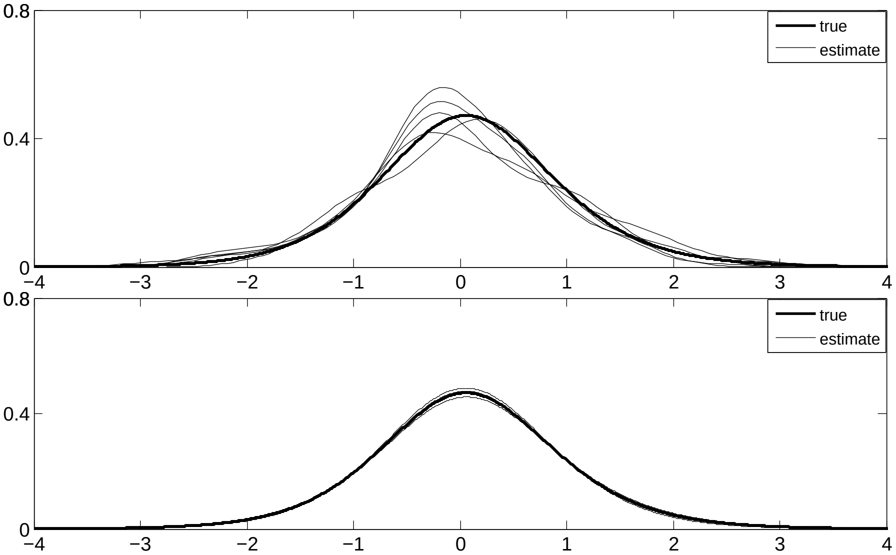

FIGURE 2. Stochastic volatility model, NPKDE (top) and PCMC (bottom)

PCMC (well specified)	PCMC (misspecified)	$\hat{z}_n$	NPKDE
0.0024	0.0067	0.0123	0.0139

TABLE 5. MISE comparison, scalar factor model

where $\beta = 1$, $\gamma = 0.95$, and (ξ_t, η_t) is IID and standard normal in $\mathbb{R}^2$. The second model specifies the DGP of $\{X_t\}$ as the ARMA (1,1) process

$$X_{t+1} = \gamma X_t + \theta \eta_t + \eta_{t+1} \tag{16}$$

In table 5, the well specified PCMC reports the MISE of the PCMC density estimator calculated in the usual way, while the misspecified PCMC is the MISE of the PCMC density estimator when the true process is (15) but the DGP of $\{X_t\}$ is misspecified as (16). While the misspecification of true DGP of X affects the performance of our PCMC estimator, in this example the effect of misspecification is relatively small.

PCMC (well specified)	PCMC (misspecified)	NPKDE
0.2614	0.4512	3.5888

TABLE 6. MISE comparison, Gaussian latent state space model

Next we check robustness in a latent variable model. Consider the Gaussian latent state space model

$$Y_t = g(X_t) + \xi_t, \qquad X_{t+1} = \alpha_0 + \alpha_1 X_t + \eta_{t+1}$$

where X_t is latent, Y_t is observed, (ξ_t, η_t) is IID standard normal in $\mathbb{R}^2$, $\alpha_0 = 0.5$, $\alpha_1 = 0.8$, and $g(x) = 0.25x$. We consider the case where the DGP of $\{X_t\}$ is misspecified as the vector AR(1) process

$$\begin{pmatrix} X_{1t} \\ X_{2t} \end{pmatrix} = \begin{pmatrix} \beta_{10} \\ \beta_{20} \end{pmatrix} + \begin{pmatrix} \beta_{11} & 0 \\ 0 & \beta_{22} \end{pmatrix} \begin{pmatrix} X_{1t-1} \\ X_{2t-1} \end{pmatrix} + \begin{pmatrix} \eta_{1t} \\ \eta_{2t} \end{pmatrix} \tag{17}$$

where (η_{1t}, η_{2t}) is IID and standard normal in $\mathbb{R}^2$. Table 6 reports some results for this robustness check. As in the previous example, misspecification of the DGP increases MISE, but the increase is relatively small.

6. PROOFS

This section contains the proof of theorem 2.1. To simplify notation, let $F(\theta)$ represent the function $f(\cdot, \theta)$. Thus, F is a mapping from Θ into $L_2(\mathbb{X})$ defined by

$$F(\theta) = \int p(\cdot \,|\, x, \theta)\phi(x, \theta)dx \qquad (\theta \in \Theta) \tag{18}$$

Also, let

$$\hat{f}_m(y, \theta) := \frac{1}{m} \sum_{t=1}^{M} p(y \,|\, X_t^\theta, \theta)$$

where $\{X_t^\theta\}$ is the simulated process defined recursively by

$$X_{t+1}^\theta = H(X_t^\theta, \eta_{t+1}, \theta) \quad \text{and} \quad X_0^\theta = x \in \mathbb{X} \tag{19}$$

Here the process $\eta := \{\eta_t\}_{t \geq 1}$ is a simulated copy of the process η in (2). The joint law of η is the infinite product of the common marginal law of η_t, defined on the sequence space D^∞. The joint law will be denoted by v_∞.

Lemma 6.1. *If assumption 2.1 holds, then F is Hadamard differentiable at θ_0, with Hadamard derivative F'_{θ_0} given by*

$$F'_{\theta_0}(\theta) = \int \langle d(x, \cdot, \theta_0), \theta \rangle dx \in L_2(\mathbb{X}) \qquad (\theta \in \mathbb{R}^M) \tag{20}$$

Proof. To verify that F'_{θ_0} is the Hadamard derivative of F at θ_0, we must show that F'_{θ_0} defined in (20) is a bounded linear operator from $\mathbb{R}^M$ to $L_2(\mathbb{X})$ such that

$$\left\| \frac{F(\theta_0 + t_n\theta_n) - F(\theta_0)}{t_n} - F'_{\theta_0}(\theta) \right\| \to 0 \tag{21}$$

for any $\theta \in \Theta$, $t_n \downarrow 0$ and $\theta_n \to \theta \in \Theta$ (cf., e.g., van der Vaart, 1998, p. 296). Evidently F'_{θ_0} is linear. To see that F'_{θ_0} is a bounded operator, observe that, by the Cauchy-Schwartz inequality and assumption 2.1,

$$\begin{aligned} \left| \int \langle d(x,y,\theta_0), \theta \rangle dx \right| &\leq \int |\langle d(x,y,\theta_0), \theta \rangle| dx \\ &\leq \|\theta\|_E \int \|d(x,y,\theta_0)\|_E dx \\ &\leq \|\theta\|_E \int g(x,y) dx \end{aligned}$$

$$\therefore \quad \|F'_{\theta_0}(\theta)\| \leq \|\theta\|_E \left\{ \int \left\{ \int g(x,y)dx \right\}^2 dy \right\}^{1/2}$$

The finiteness of the integral expression is guaranteed by assumption 2.1.

We now turn to the verification of (21). Fix $t_n \downarrow 0$ and $\theta_n \to \theta \in \Theta$. Let

$$\kappa(x,y,\theta) := p(y \,|\, x,\theta)\phi(x,\theta) \qquad (y \in \mathbb{Y},\, x \in \mathbb{X},\, \theta \in \Theta)$$

and

$$g_n(x,y) := \frac{\kappa(x,y,\theta_0 + t_n\theta_n) - \kappa(x,y,\theta_0)}{t_n} - \langle d(x,y,\theta_0), \theta \rangle \tag{22}$$

Since

$$\begin{aligned} \int g_n(x,y)dx &= \int \left[\frac{\kappa(x,y,\theta_0 + t_n\theta_n) - \kappa(x,y,\theta_0)}{t_n} \right] dx - \int \langle d(x,y,\theta_0), \theta \rangle dx \\ &= \frac{\int \kappa(x,y,\theta_0 + t_n\theta_n)dx - \int \kappa(x,y,\theta_0)dx}{t_n} - \int \langle d(x,y,\theta_0), \theta \rangle dx \end{aligned}$$

we have

$$\int g_n(x,\cdot)dx = \frac{F(\theta_0 + t_n\theta_n) - F(\theta_0)}{t_n} - F'_{\theta_0}(\theta)$$

and hence

$$\left\| \frac{F(\theta_0 + t_n\theta_n) - F(\theta_0)}{t_n} - F'_{\theta_0}(\theta) \right\|^2 = \int \left\{ \int g_n(x,y)dx \right\}^2 dy$$

Thus (21) will be established if we can show that

$$\int \left\{ \int g_n(x,y)dx \right\}^2 dy \to 0 \qquad (n \to \infty) \tag{23}$$

As a first step, note that $g_n \to 0$ pointwise on $\mathbb{X} \times \mathbb{Y}$. This first result is almost immediate from the definition of g_n in (22), since, for given x and y, the vector

$d(x, y, \theta_0)$ is the vector of partial derivatives of the function $\theta \mapsto \kappa(x, y, \theta)$. As $\theta \mapsto p(y \,|\, x, \theta)$ and $\theta \mapsto \phi(x, \theta)$ are assumed to be continuously differentiable on V, the map $\theta \mapsto \kappa(x, y, \theta_0)$ is differentiable at θ_0, and the Frechet derivative at θ_0 is the mapping $\theta \mapsto \langle d(x, y, \theta_0), \theta \rangle$. In $\mathbb{R}^M$ the Frechet derivative and the Hadamard derivative coincide, and hence $|g_n(x, y)| \to 0$ by the definition of Hadamard differentiability.

In order to pass the limit through the integrals in (23), we next show that a scalar multiple of the function g defined in assumption 2.1 dominates g_n pointwise on $\mathbb{X} \times \mathbb{Y}$ for all sufficiently large n. To see that this is the case, fix $(x, y) \in \mathbb{X} \times \mathbb{Y}$ and $N \in \mathbb{N}$ such that $\theta_0 + t_n \theta_n \in V$ for all $n \geq N$. Without loss of generality we can choose the neigborhood V to be convex. With convex V, the mean value theorem in $\mathbb{R}^M$ implies existence of a vector $\theta_n^* \in V$ on the line segment between θ_0 and $t_n \theta_n$ with

$$\kappa(x, y, \theta_0 + t_n \theta_n) - \kappa(x, y, \theta_0) = \langle d(x, y, \theta_n^*), t_n \theta_n \rangle$$

Dividing both sides by t_n and using the definition of g_n in (22), we obtain

$$\begin{aligned} |g_n(x, y)| &= |\langle d(x, y, \theta_n^*), \theta_n \rangle - \langle d(x, y, \theta_0), \theta \rangle| \\ &\leq |\langle d(x, y, \theta_n^*), \theta_n \rangle| + |\langle d(x, y, \theta_0), \theta \rangle| \\ &\leq \|d(x, y, \theta_n^*)\|_E \|\theta_n\|_E + \|d(x, y, \theta_0)\|_E \|\theta\|_E \end{aligned}$$

Applying assumption 2.1, we obtain

$$|g_n(x, y)| \leq g(x, y)(\|\theta_n\|_E + \|\theta\|_E)$$

Since θ_n is convergent it is also bounded in n, and hence there exists a constant K with $|g_n(x, y)| \leq K g(x, y)$ for all $n \geq N$.

Returning to the proof of (23), define

$$h_n(y) := \left\{ \int |g_n(x, y)| dx \right\}^2 \quad \text{and} \quad h(y) := \left\{ \int K g(x, y) dx \right\}^2$$

As a first step to proving (23), we claim that $h_n \to 0$ almost everywhere on $\mathbb{Y}$. To see this, observe that assumption 2.1 gives $\int h(y) dy < \infty$, and hence h is finite almost everywhere. For any y such that $h(y)$ is finite, we have $\int K g(x, y) dx < \infty$. In addition, for this same y, we have $|g_n(x, y)| \leq K g(x, y)$ and $g_n(x, y) \to 0$ for all $x \in \mathbb{X}$. It follows from the dominated convergence theorem that $\int g_n(x, y) dx \to 0$, and therefore $h_n(y) \to 0$. This verifies the claim that $h_n \to 0$ almost everywhere on $\mathbb{Y}$.

The final step is to show that $\int h_n(y) dy \to 0$. To see that this is so, observe that, in addition to $h_n \to 0$ almost everywhere, we have $0 \leq h_n \leq h$ for all n, and h is integrable by assumption 2.1. Another application of the dominated convergence theorem now gives $\int h_n(y)dy \to 0$.

The convergence $\int h_n(y)dy \to 0$ is equivalent to (23), completing the proof of lemma 6.1. □

Lemma 6.2. *Under the conditions of theorem 2.1 we have*

$$\sqrt{n}\{f(\cdot,\hat{\theta}_n) - f(\cdot,\theta_0)\} \xrightarrow{d} N(0,C)$$

where $N(0,C)$ is the centered Gaussian defined in theorem 2.1.

Proof of lemma 6.2. Assume the conditions of the lemma. Let V be a random variable on $\mathbb{R}^M$ with $V \sim N(0,\Sigma)$, so that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to V. Let F be as defined in (18). We aim to show that

$$\sqrt{n}\{F(\hat{\theta}_n) - F(\theta_0)\} \xrightarrow{d} N(0,C) \tag{24}$$

in $L_2(\mathbb{Y})$, where C is as defined in theorem 2.1. Lemma 6.1 showed that F is Hadamard differentiable at θ_0, when viewed as a mapping from Θ to $L_2(\mathbb{Y})$. Applying a functional delta theorem (e.g., van der Vaart, 1998, theorem 20.8) we obtain

$$\sqrt{n}\{F(\hat{\theta}_n) - F(\theta_0)\} \xrightarrow{d} F'_{\theta_0}(V)$$

in $L_2(\mathbb{Y})$, where F'_{θ_0} is as defined in (20). Thus, it remains only to show that $F'_{\theta_0}(V) \sim N(0,C)$.

Using the definition of α_m in the statement of theorem 2.1, we have

$$F'_{\theta_0}(V) = \int \langle d(x,\cdot,\theta_0), V\rangle dx = \sum_{m=1}^{M} \int d_m(x,\cdot,\theta_0)dxV_m = \sum_{m=1}^{M} \alpha_m V_m$$

Each α_m is an element of $L_2(\mathbb{Y})$ because

$$|\alpha_m(y)| \leq \int \|d(x,y,\theta_0)\|_E dx \leq \int g(x,y)dx$$

and the right-hand side is square-integrable by assumption 2.1. It follows that $F'_{\theta_0}(V) = \sum_{m=1}^{M} \alpha_m V_m$ is an $L_2(\mathbb{Y})$ valued random variable.

To show that $F'_{\theta_0}(V)$ is Gaussian, we need to prove that the L_2 inner product $\langle F'_{\theta_0}(V), h\rangle$ is Gaussian in $\mathbb{R}$ for each $h \in L_2(\mathbb{Y})$. This follows immediately from the fact that V is multivariate Gaussian, since linear combinations of multivariate Gaussian random variables are univariate Gaussian by definition, and

$$\langle F'_{\theta_0}(V), h\rangle = \sum_{m=1}^{M} \langle \alpha_m, h\rangle V_m \tag{25}$$

To show that the $L_2(\mathbb{Y})$ expectation of $F'_{\theta_0}(V)$ is the zero element, we need to show that the (scalar) expectation of (25) is zero for all $h \in L_2(\mathbb{Y})$. This is true because $\mathbb{E} V_m = 0$ for all m.

Finally, we need to verify that the covariance operator of $F'_{\theta_0}(V)$ is equal to C. In other words, we must show that

$$\mathbb{E}\langle g, F'_{\theta_0}(V)\rangle \, \langle F'_{\theta_0}(V), h\rangle = \langle g, Ch\rangle$$

where the expression for $\langle g, Ch\rangle$ is given in theorem 2.1. Evidently this equality is valid, since

$$\begin{aligned}\mathbb{E}\langle F'_{\theta_0}(V), g\rangle \, \langle F'_{\theta_0}(V), h\rangle &= \mathbb{E}\left(\sum_{i=1}^{M}\langle \alpha_i, g\rangle V_i\right)\left(\sum_{j=1}^{M}\langle \alpha_j, h\rangle V_j\right) \\ &= \mathbb{E}\left(\sum_{i=1}^{M}\sum_{j=1}^{M}\langle \alpha_i, g\rangle \langle \alpha_j, h\rangle V_i V_j\right)\end{aligned}$$

Passing the expectation through the sum yields the expression for $\langle g, Ch\rangle$ given in theorem 2.1. □

Lemma 6.3. *If the conditions of theorem 2.1 hold, then, for any given n, we have*

$$\sqrt{n}\|\hat{f}_m(\cdot, \hat{\theta}_n) - f(\cdot, \hat{\theta}_n)\| = \frac{1}{\sqrt{m}} O_P(1)$$

where $O_P(1)$ indicates that the term is bounded in probability over m.

Proof. Fix $n \in \mathbb{N}$. Since we claim only boundedness in probability, it suffices to show that

$$\sqrt{m}\|\hat{f}_m(\cdot, \hat{\theta}_n) - f(\cdot, \hat{\theta}_n)\| = O_P(1) \tag{26}$$

If we fix $\theta \in \Theta$, then V-uniform ergodicity and theorem 4.2 in Braun *et al.* imply that

$$\sqrt{m}\{\hat{f}_m(\cdot, \theta) - f(\cdot, \theta)\} \xrightarrow{d} W_\theta \sim N(0, S_\theta)$$

for some covariance operator S_θ. It follows from the continuous mapping theorem that

$$Y_m(\theta) := \sqrt{m}\|\hat{f}_m(\cdot, \theta) - f(\cdot, \theta)\| \xrightarrow{d} \|W_\theta\| \tag{27}$$

Let $\pi(\theta, dy)$ denote the distribution of the nonnegative scalar random variable $\|W_\theta\|$. In view of (19), each X_t^θ is a function of θ and the sequence $\eta := \{\eta_t\}_{t\geq 1}$. Since the randomness in $\hat{f}_m(\cdot, \theta)$ comes only through each X_t^θ, we can write $Y_m(\theta)$ as $Y_m(\theta) = G_m(\eta, \theta)$ for some function G_m. Our aim is to show that

$$Y_m(\hat{\theta}_n) = G_m(\eta, \hat{\theta}_n) = O_P(1) \qquad (m \to \infty)$$

Let $h\colon \mathbb{R} \to \mathbb{R}$ be bounded and continuous. Recalling that v_∞ is the joint law of η and letting ν denote the law of $\hat{\theta}_n$, we can write

$$\mathbb{E}\, h \circ Y_m(\hat{\theta}_n) = \mathbb{E}\, h \circ G_m(\eta, \hat{\theta}_n) = \int \int h \circ G_m(z, \theta) v_\infty(dz) \nu(d\theta)$$

where the last equality is due to independence of η and $\hat{\theta}_n$. We saw in (27) that, for fixed θ,

$$\int h \circ G_m(z, \theta) v_\infty(dz) = \mathbb{E}\, h \circ Y_m(\theta) \to \int h(y) \pi(\theta, dy) \qquad (m \to \infty)$$

Since this convergence holds for all θ, and since $\theta \mapsto \int h \circ G_m(z, \theta) v_\infty(dz)$ is uniformly bounded by $\sup_x |h(x)|$, the dominated convergence theorem implies that

$$\mathbb{E}\, h \circ Y_m(\hat{\theta}_n) = \int \int h \circ G_m(z, \theta) v_\infty(dz) \nu(d\theta) \to \int \int h(y) \pi(\theta, dy) \nu(d\theta)$$

as $m \to \infty$. Since h was an arbitrary continuous bounded function, we conclude that $Y_m(\hat{\theta}_n)$ converges in probability to the distribution $\pi(\theta, dy)\nu(d\theta)$. Since it converges in distribution it is also bounded in probability. The claim in (26) is now verified. □

Proof of theorem 2.1. Adding and subtracting $f(\cdot, \hat{\theta}_n)$, we can write

$$\sqrt{n}\{\hat{f}_m(\cdot, \hat{\theta}_n) - f(\cdot, \theta_0)\} = \sqrt{n}\{\hat{f}_m(\cdot, \hat{\theta}_n) - f(\cdot, \hat{\theta}_n)\} + \sqrt{n}\{f(\cdot, \hat{\theta}_n) - f(\cdot, \theta_0)\}$$

The proof of theorem 2.1 now follows from lemma 6.2 and lemma 6.3. □

The next lemma confirms the only technical step needed for deriving the second expression for the covariance operator C in section 2.

Lemma 6.4. *Under the conditions of theorem 2.1, we have*

$$\alpha_m(y) = \frac{\partial}{\partial \theta_m} f(y, \theta_0)$$

for any m in $1, \ldots, M$ and any given $y \in \mathbb{Y}$.

Proof. Fix m in $1, \ldots, M$ and $y \in \mathbb{Y}$. The lemma amounts to the claim that

$$\int \frac{\partial}{\partial \theta_m} p(y \,|\, x, \theta_0) \phi(x, \theta_0) dx = \frac{\partial}{\partial \theta_m} \int p(y \,|\, x, \theta_0) \phi(x, \theta_0) dx$$

Under the standard rules for differentiating under integrals, this statement is valid if there exists an integrable function h on $\mathbb{X}$ such that, for all θ on a neighborhood N of θ_0,

$$|d_m(x, y, \theta)| := \left| \frac{\partial}{\partial \theta_m} p(y \,|\, x, \theta) \phi(x, \theta) \right| \leq h(x) \tag{28}$$

almost everywhere. Take $N = V$ and $h(x) := g(x,y)$ where V and g are as defined in assumption 2.1. By the conditions of assumption 2.1, the function h is integrable, and $\|d(x,y,\theta)\|_E \leq h(x)$ for all $\theta \in N$. This implies the inequality in (28), and lemma 6.4 is proved. □

References

[1] Aït-Sahalia, Y., J. Fan., and J. Jiang, 2010, Nonparametric tests of the Markov hypothesis in continuous-time models. The Annals of Statistics, 38, 3129–3163.

[2] Aït-Sahalia, Y., J. Fan., and H. Peng, 2009, Nonparametric transition-based tests for jump diffusions. Journal of the American Statistical Association, 104, 1102–1116.

[3] Andel, J., I. Netuka, and K. Svara, 1984, On threshold autoregressive processes, Kybernetika, 20, 89–106.

[4] Bosq, D., 2000, Linear Processes in Function Space, Springer-Verlag, New York.

[5] Braun, R. A., H. Li and J. Stachurski (2011): Generalized look-ahead methods for computing stationary densities. Mimeo, Australian National University.

[6] Frees, E. W., 1994, Estimating densities of functions of observations. Journal of the American Statistical Association 89, 517-525.

[7] Fix, E., and J. L. Hodges, 1951, Discriminatory analysis. Nonparametric discrimination; consistency properties. Report Number 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas.

[8] Gine, E. and D. M. Mason, 2007, On local U-statistic processes and the estimation of densities of functions of several sample variables. Annals of Statistics 35, 1105-1145.

[9] Gordon, A. D., 1981, Classification, London: Chapman and Hall.

[10] Hamilton, J. D., 1994, Time Series Analysis, Princeton University Press.

[11] He, L., S. W. Huh, and B. S. Lee, 2010, Dynamic factors and asset pricing. Journal of the American Statistical Association, 45, 707-737.

[12] Kim, K. and W. B. Wu, 2007, Density estimation for nonlinear time series. Mimeo, Michigan State University.

[13] Koopman, S. J. and E. H. Uspensky, 2002, The stochastic volatility in mean model: empirical evidence from international stock markets. Journal of Applied Econometrics, 17, 667–689.

[14] Kristensen, D., 2008, Uniform ergodicity of a class of Markov chains with applications to time series models. Mimeo, Columbia University.

[15] Saavedra, A. and R. Cao, 2000, On the estimation of the marginal density of a moving average process. The Canadian Journal of Statistics 28, 799-815.

[16] Schick, A. and W. Wefelmeyer, 2004, Functional convergence and optimality of plug-in estimators for stationary densities of moving average processes. Bernoulli 10, 889-917.

[17] Schick, A. and W. Wefelmeyer, 2007, Uniformly root-n consistent density estimators for weakly dependent invertible linear processes. Annals of Statistics 35, 815-843.

[18] Smith, D. R., and A. Layton, 2007, Comparing probability forecasts in Markov regime switching business cycle models. Journal of Business Cycle Measurement and Analysis, 479-98.

[19] Zhao, Z., 2010, Density estimation for nonlinear parametric models with conditional heteroskedasticity. Journal of Econometrics, 155, 71-82.

SCHOOL OF FINANCE, ACTUARIAL STUDIES AND APPLIED STATISTICS, AUSTRALIAN NATIONAL UNIVERSITY

E-mail address: yin.liao@anu.edu.au

RESEARCH SCHOOL OF ECONOMICS, AUSTRALIAN NATIONAL UNIVERSITY

E-mail address: john.stachurski@anu.edu.au