Simulation-Based Density Estimation for Time Series using Covariate Data*

Yin Liao[a] and John Stachurski[b]

[a]School of Economics and Finance, Queensland University of Technology

[b]Research School of Economics, Australian National University

August 22, 2014

Abstract

This paper proposes a simulation-based density estimation technique for time series that exploits information found in covariate data. The method can be paired with a large range of parametric models used in time series estimation. We derive asymptotic properties of the estimator and illustrate attractive finite sample properties for a range of well-known econometric and financial applications.

1 Introduction

In this paper we study a parametric density estimation technique for time series that exploits covariate data. While the technique has broad applicability, our motivation

*This paper has benefited from helpful comments by Dennis Kristensen and participants at the 2012 International Symposium on Econometric Theory and Applications in Shanghai, and the 18th International Conference on Computing in Economics and Finance in Prague. We thank Varang Wiriyawit for research assistance and ARC Discovery Outstanding Researcher Award DP120100321 for financial support. Emails: yin.liao@qut.edu.au and john.stachurski@anu.edu.au

stems from applications in econometrics and finance where density estimation is often used for tasks such as analysis of asset returns, interest rates, GDP growth, inflation and so on. For example, the Bank of England routinely estimates densities for inflation and a wide range of asset prices using options data (de Vincent-Humphreys and Noss, 2012). In settings such as this, the primary attraction of density estimates is that they typically provide more information than estimates of central tendancy or a finite set of moments. Depending on the time series in question, density estimates can be used to address a large variety of questions, such as the risk of corporate (or sovereign) default, or the likelihood of recession, or of inflation leaving a target band over a given interval of time.[1]

A variety of techniques for estimating densities have been proposed in the literature. One popular approach is nonparametric kernel density estimation (Rosenblatt, 1956). Nonparametric density estimators have an advantage over parametric methods in terms of robustness and generality, in the sense that asymptotic convergence occurs under very weak assumptions. This makes nonparametric density estimators ideal for certain applications, particularly those where the risk of model misspecification is high.

On the other hand, for some of the more common economic and financial time series, econometricians have spent decades formulating, developing and testing parametric time series models (e.g., ARMA models, GARCH models and their many variations, Markov switching models, stochastic volatility models, dynamic factor models, threshold models, etc.). This research has generated a very substantial body of knowledge on classes of parametric models and how they can be paired with certain time series to effectively represent various data generating processes (for a recent overview, see Martin *et al.*, 2012). In these kinds of settings, it is natural to seek techniques that can exploit this parametric information to construct density estimators. Our paper pursues this idea, with the focus on providing a flexible density estimation scheme that can be used in combination with common time series

[1]For related applications and discussion, see, for example, Aït-Sahalia and Hansen (2009), Calabrese and Zenga (2010) or Polanski and Stoja (2011). In addition to situations where the density itself is of primary interest (e.g., density forecasting), density estimation is also used in a wide range of statistical techniques where density estimates are an input, such as discriminant or cluster analysis. Similarly, density estimators are used to address specification testing or model validation problems (e.g., Aït-Sahalia et al., 2009).

models.

In doing so we confront several problems associated with these kinds of estimation. First, many modern econometric models have nonlinear or non-Gaussian features that make the relevant densities intractable. Hence generating density estimates requires some form of approximation. Second, time series data sets are often (a) smaller than cross-sectional data sets, and (b) contain less information for a given data size, since observations are more likely to be correlated. This problem of information scarcity is compounded in the case of density estimation, since the "point" we are trying to estimate is, in general, an infinite-dimensional object.

The technique we study addresses these problems simultaneously. To accommodate the problem that the densities might be intractable we use a simulation-based approach, which permits construction of density estimates from model primitives in a wide range of settings. To address the issue of limited data, we combine two useful ways to supplement the amount of information available for estimation of a given density: exploitation of parametric structure and incorporation of information from covariates.[2] Our paper pursues these idea within a parametric time series setting. Our simulation studies give a number of examples as to how inclusion of parametric structure combined with covariate data can greatly reduce mean squared error.

From a technical perspective, the method we study in this paper can be understood as a variation on conditional Monte Carlo (see, e.g., Henderson and Glynn, 2001), which is an elegant and effective technique for reducing variance in a variety of Monte Carlo procedures. Here the estimation target is a density. In addition, the primitives for the simulation contain estimated parameters. Accommodating this randomness together with the randomness introduced by the Monte Carlo step, we establish a functional central limit result for the error of the estimator. The theorem also shows that the estimated density converges to the target density in mean integrated squared error.

Following presentation of the theory, we turn to illustrations of the method for a variety of econometric and financial applications, and to Monte Carlo analysis to

[2]The meaning of the second point is as follows: Suppose that we wish to estimate the density f of a random vector Y_t. One possibility is to use only observations of Y_t. If, however, we possess a model that imposes structure on the relationship between Y_t and a vector of covariates X_t, we can use this structure and additional data in order to improve our estimate of f.

investigate finite sample behavior. The case studies include dynamic factor models, linear and nonlinear autoregressive models, Markov regime switching models and stochastic volatility models. In all cases the method exhibits excellent finite sample properties. We give several interpretations of this performance.

Regarding related literature, alternative parametric density estimators using covariates have been proposed by Saaverdra and Cao (2000) and Schick and Wefelmeyer (2004, 2007) for linear processes, by Frees (1994) and Gine and Mason (2007) for functions of independent variables, by Kim and Wu (2007) for nonlinear autoregressive models with constant variance, and by Støve and Tjøstheim (2012) for a nonlinear heterogenous regression model. A related semiparametric approach was proposed by Zhao (2010). In addition, Escanciano and Jacho-Chávez (2012) exhibit a nonparametric estimator of the density of response variables that is $\sqrt{n}$-consistent. Our set up is less specific than the parametric treatments discussed above. For example, we make no assumptions regarding linearity, additive shocks, constant variance and so on, and the density of interest can be vector valued. A Monte Carlo step makes the method viable despite this generality. On the other hand, relative to the nonparametric and semiparametric methods, our estimator puts more emphasis on finite sample properties. These points are illustrated in depth below.

The structure of our paper is as follows: Section 2 gives an introduction to the estimation technique. Section 3 provides asymptotic theory. Section 4 looks at a number of applications and provides some Monte Carlo studies. Section 5 discusses robustness issues. Section 6 gives proofs.

2 Outline of the Method

Our objective is to estimate the density f of random vector Y_t. To illustrate the main idea, suppose that, in addition to the original data $Y_1, \ldots, Y_n$, we also observe a sequence of covariates $X_1, \ldots, X_n$ where $\{X_t\}$ is a stationary and ergodic vector-valued stochastic process. Suppose further that Y_t is related to X_t via $Y_t \,|\, X_t \sim p(\cdot \,|\, X_t)$. That is, $p(\cdot \,|\, x)$ is the conditional density of Y_t given $X_t = x$. Since $\{X_t\}$ and p are assumed to be stationary, the target process $\{Y_t\}$ is likewise stationary.

Letting ϕ be the common stationary (i.e., unconditional) density of X_t, an elementary conditioning argument tells us that the densities f and ϕ are related to one another by

$$f(y) = \int p(y \mid x)\phi(x)dx \qquad (y \in \mathbb{Y}) \tag{1}$$

Example 2.1. Suppose that Y_t denotes returns on a given asset, and let $\{Y_t\}$ obey the GARCH(1,1) model $Y_t = \mu + \sigma_t \epsilon_t$ where $\{\epsilon_t\}$ is IID and $N(0,1)$, and $\sigma_{t+1}^2 = \alpha_0 + \alpha_1 \sigma_t^2 + \alpha_2 Y_t^2$. Assume all parameters are strictly positive and $\alpha_1 + \alpha_2 < 1$. Let $X_t := \sigma_t^2$, so that

$$X_{t+1} = \alpha_0 + \alpha_1 X_t + \alpha_2 (\mu + \sqrt{X_t}\epsilon_t)^2 \tag{2}$$

Let ϕ be the stationary density of this Markov process. In this setting (1) becomes

$$f(y) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi x}} \exp\left\{ -\frac{(y-\mu)^2}{2x} \right\} \phi(x)\, dx \tag{3}$$

Returning to the general case, suppose for the moment that the conditional density p is *known*. Given observations $\{X_t\}_{t=1}^n$, there is a standard procedure for estimating f called conditional density estimation that simply replaces the right-hand side of (1) with its empirical counterpart

$$\hat{r}(y) := \frac{1}{n} \sum_{t=1}^{n} p(y \mid X_t) \tag{4}$$

Since $\{X_t\}$ is assumed to be ergodic with common marginal ϕ, the law of large numbers and (1) yield

$$\hat{r}(y) = \frac{1}{n} \sum_{t=1}^{n} p(y \mid X_t) \to \int p(y \mid x)\phi(x)dx = f(y)$$

as $n \to \infty$. A more complete asymptotic theory is provided in Braun *et. al* (2012).[3]

In a statistical setting, the conditional density p is unknown. If we are prepared to impose parametric assumptions, then we can write

$$Y_t \mid X_t \sim p(\cdot \mid X_t, \beta) \tag{5}$$

[3]The assumption that p is known is appropriate in a computational setting, where all densities are known in principle but f might be intractable and hence require estimation (in the sense of approximation via Monte Carlo). This is the perspective taken in Gelfand and Smith (1990), Henderson and Glynn (2001), Braun *et. al* (2012) and many other papers. The asymptotic theory is well established.

where β is a vector of unknown parameters. In this case, an obvious extension to (4) is to replace the unknown vector β with an estimate $\hat{\beta}_n$ obtained from the data $\{Y_t, X_t\}_{t=1}^n$. This leads to the semiparametric estimator

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

The estimator is semiparametric in the sense that if we let ϕ_n be the empirical distribution of the sample $X_1, \ldots, X_n$, then $\hat{s}$ can be expressed as

$$\hat{s}(y) = \int p(y \mid x, \hat{\beta}_n) \phi_n(dx) \tag{7}$$

Thus, $\hat{s}$ combines the parametric conditional density estimate $p(\cdot \mid \cdot, \hat{\beta}_n)$ with the nonparametric empirical distribution ϕ_n to obtain an estimate of f. The random density $\hat{s}$ is known to be consistent and asymptotically normal for f under certain regularity conditions (Zhao, 2010, theorem 1).

While this semiparametric estimator $\hat{s}$ is natural, there are a number of settings where it cannot be applied, or where its finite sample performance is suboptimal. As a first example, consider a setting where we have a parametric model for the dynamics of $\{X_t\}$ that provides an accurate fit to this data. In such a setting, these estimated dynamics allow us to produce a good estimate for the stationary density ϕ of $\{X_t\}$. The semiparametric estimator $\hat{s}$ fails to exploit this knowledge, replacing it with the (nonparametric) empirical distribution ϕ_n.

Another example along the same lines is a setting such as $Y_t = \beta' X_t + \sigma \epsilon_t$, where $X_t = (Y_{t-1}, \ldots, Y_{t-p})$. Here the covariates are just lagged values of the current state. If we estimate (β, σ), then we can in fact deduce the stationary density ϕ of X_t coinciding with this estimate. It would be inconsistent to discard this stationary density and use the empirical distribution ϕ_n instead.

In addition to the above, there are settings where the semiparametric estimator cannot be used at all. For example, many time series models incorporate latent variables (e.g., 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 hence the semiparametric estimate $\hat{s}$ in (6) can not be computed.

As we now show, all of these issues can be addressed in a setting where we are prepared to add more parametric structure—in particular, parametric information about the process $\{X_t\}$. In what follows, this parametric information is assumed to take the form

$$X_{t+1} \mid X_t \sim q(\cdot \mid X_t, \gamma) \tag{8}$$

where γ is another vector of unknown parameters. (Although we are restricting the model to a first order process, this costs no generality, since any p order process can be reduced to a first order process by reorganizing state variables.)

Given this information, to estimate f, the procedure we consider is:

1. Estimate the parameters in (5) and (8) with $\hat{\beta}_n$ and $\hat{\gamma}_n$ respectively.
2. Simulate $\{X_t^*\}_{t=1}^m$ via

$$X_{t+1}^* \sim q(\cdot \mid X_t^*, \hat{\gamma}_n) \quad \text{with} \quad X_0^* = x \tag{9}$$

3. Return the function $\hat{f}$ defined by

$$\hat{f}(y) := \frac{1}{m} \sum_{t=1}^{m} p(y \mid X_t^*, \hat{\beta}_n) \tag{10}$$

In what follows, we refer to the estimator $\hat{f}$ in (10) as the *parametric conditional Monte Carlo* (PCMC) density estimator. Comparing $\hat{s}$ defined in (6) with the PCMC density estimator $\hat{f}$, the difference is that while $\hat{s}$ uses the observed data $\{X_t\}$ directly on the right hand side of (6), the PCMC density estimator uses simulated data instead. More precisely, instead of using $\{X_t\}$ directly, we use $\{X_t\}$ to estimate the model (8), and then use that model to generate simulated data $\{X_t^*\}$.

Example 2.1 (continued) Consider again the GARCH model presented in example 2.1. Let $\hat{\theta} := (\hat{\mu}, \hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)$ be a consistent estimator of the unknown parameters. By plugging these estimates into (2) we simulate $X_1^*, \ldots, X_m^*$. Recalling (3), the PCMC density estimator is then

$$\hat{f}(y) = \frac{1}{m} \sum_{t=1}^{m} \frac{1}{\sqrt{2\pi X_t^*}} \exp\left\{ -\frac{(y - \hat{\mu})^2}{2X_t^*} \right\} \tag{11}$$

2.1 Discussion

One attractive feature of the PCMC density estimator is that in a single simulation step it bypasses two integration problems that in general fail to have analytical solutions. Without simulation, we would need to (a) take the estimated transition density $q(x' \mid x, \hat{\gamma}_n)$ and calculate from it the corresponding stationary distribution $\phi(\cdot, \hat{\gamma}_n)$ as the solution to an integral equation, and then (b) calculate f as $\int p(y \mid x, \hat{\beta}_n)\phi(x, \hat{\gamma}_n)dx$. Apart from some special cases, neither of these integration problems can be solved in closed form.

Two other attractive features of $\hat{f}$ are as follows: First, the initial value x in (9) can be chosen arbitrarily. This is significant because, as discussed immediately above, the stationary density $\phi(\cdot, \hat{\gamma}_n)$ is typically intractable, and we have no way to sample from it directly. In section 3 we prove that our convergence results do not depend on the choice of x. Second, latent variables in $\{X_t\}$ cause no difficulties in using the PCMC density estimator. As soon as we estimate the transition density $q(x' \mid x, \hat{\gamma}_n)$ we can produce the simulated data $\{X_t^*\}$. In this simulated data, latent variables become observable.[4]

On the other hand, when comparing $\hat{s}$ and $\hat{f}$ in (6) and (10) respectively, the fact that the latter uses simulated data from an estimated model instead of observed data clearly involves some cost. When interpreting this difference, however, it is helpful to bear in mind that in a sense both techniques do use the observed data. What differs is the way in which these data are used. While the semiparametric estimator $\hat{s}$ uses $\{X_t\}$ to construct an empirical distribution (ϕ_n on the right hand side of (7)), the PCMC uses this data to estimate a parametric model. To the extent that the parametric assumptions are correct, this approach can translate into better finite sample properties. These ideas are addressed in section 4 via a range of simulation studies. At the same time, to quantify the estimation effect, we also include information on

[4] A simulation based method for density estimation in the context of parametric time series models with latent states is also considered in Zhao (2011). Zhao's estimator requires a significantly more complex formula, involving the ratio of empirical estimates of a joint and a marginal density. However, this additional complexity arises because he aims to estimate a family of conditional densities providing all the dynamics of the hidden state vector, which can then be tested via a simultaneous confidence envelope constructed using nonparametric methods. In contrast, the PCMC estimates a single density, and we provide a detailed asymptotic theory for that estimate.

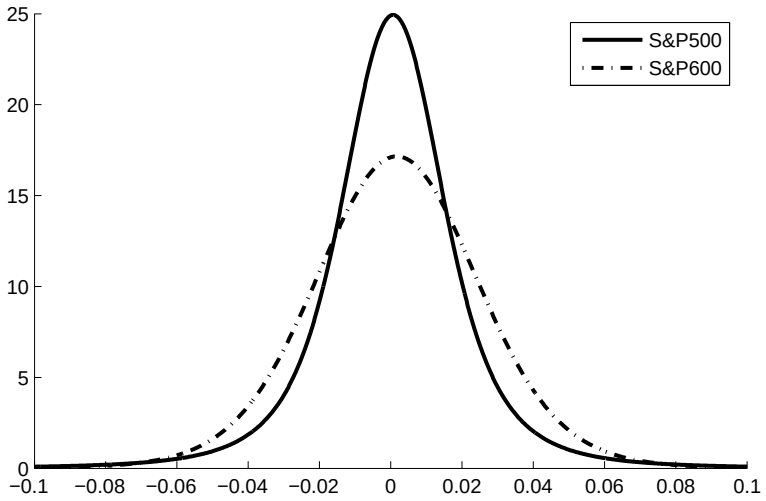

Figure 1: Estimated return densities, small cap vs large cap

the errors associated with both $\hat{f}$ and a version of $\hat{f}$ called $\hat{f}_0$ that simulates from the exact model.

Figure 1 gives an example implementation of the PCMC density estimator. In this application, we estimate two densities: the density of monthly returns on the S&P 500, a stock market index of 500 large companies listed on the NYSE or NASDAQ, and returns on the S&P 600, which is an index of smaller firms. Both densities are estimated using price data from January 2005 to December 2012. The model applied here is the GARCH model of example 2.1. Volatility is latent in this model, but, as discussed above, this presents no difficulty for our procedure.

In interpreting the results, recall that small cap stocks usually have higher average returns than large cap stocks—for example, they are one of the three factors in the Fama-French three factor model. Higher returns are typically associated with higher volatility, since the former is demanded as compensation for the latter by investors. Our estimates conform to this theory. (The mean of the density for the S&P 500 is 0.0006, while that for the S&P 600 is 2.8 times larger.)

2.2 Other Estimators

We have already mentioned several alternatives to the PCMC density estimator. There are several other parametric estimators of f that could be considered here. One is to simply specify a parametric class $\{f(\cdot, \theta)\}_{\theta \in \Theta}$ for f, estimate θ using $\{Y_t\}$ and plug in the result $\hat{\theta}$ to produce $\hat{f} := f(\cdot, \hat{\theta})$. (For example, specify $f = N(\mu, \sigma)$ and plug in the sample mean $\hat{\mu}$ and the sample standard deviation $\hat{\sigma}$.) We refer to this estimator as the ordinary parametric estimator (OPE). In finite samples this estimator is typically inferior to the PCMC density estimator because if fails to exploit the information available in covariates. Section 4 gives an extensive discussion of this point.

A parametric density estimator that *does* exploit covariate data can be obtained by (i) specifying a parametric form $\phi(x, \gamma)$ for the common density of X_t, (ii) estimating β in (5) and γ in the density $\phi(x, \gamma)$ as $\hat{\beta}_n$ and $\hat{\gamma}_n$, and (iii) returning

$$\hat{z}(y) := \int p(y \mid x, \hat{\beta}_n) \phi(x, \hat{\gamma}_n) \tag{12}$$

This estimator is impractical, since it involves integration problems that are not generally tractable. It also has another less obvious deficiency: In contrast to the PCMC density estimator $\hat{f}$, parametric specification is placed directly on the density ϕ of X_t, rather than on the dynamic model (8). As a result, $\hat{z}$ typically fails to exploit the order information in $\{X_t\}$. This can be costly, particularly when $\{X_t\}$ is highly persistent. Sections 4.1 and 4.2 elaborate, comparing $\hat{z}$ and $\hat{f}$ in simulation experiments across a variety of common models.

3 Asymptotic Theory

In this section we clarify assumptions and provide convergence results for the PCMC density estimator introduced in section 2.

3.1 Preliminaries

In this section it will be convenient to let $\theta \in \Theta$ be a vector containing all unknown parameters. Thus $p(\cdot \mid x, \beta)$ in (5) now becomes $p(\cdot \mid x, \theta)$, while $q(\cdot \mid x, \gamma)$ in (8) becomes $q(\cdot \mid x, \theta)$. The set Θ is taken to be a subset of $\mathbb{R}^K$. Let θ_0 represent the true value of the parameter vector θ. It follows that the true density of Y_t is given by $f(y, \theta_0) := \int p(y \mid x, \theta_0)\phi(x, \theta_0)dx$. To simplify notation, in the sequel we let

$$d_k(x, y, \theta) := \phi(x, \theta)\frac{\partial}{\partial \theta_k}p(y \mid x, \theta) + p(y \mid x, \theta)\frac{\partial}{\partial \theta_k}\phi(x, \theta)$$

whenever the derivatives exist. In particular, $d(x, y, \theta)$ is the K-vector obtained by differentiating $p(y \mid x, \theta)\phi(x, \theta)$ 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 $\hat{f}$ and the true density $f(\cdot, \theta_0)$. We take the set $\mathbb{Y}$ in which Y_t takes values to be a Borel subset of $\mathbb{R}^d$, and the symbol $L_2(\mathbb{Y})$ represents the set of (equivalence classes of) real-valued Borel measurable functions on $\mathbb{Y}$ that are square-integrable with respect to Lebesgue measure. As usual, we set $\langle g, h \rangle := \int g(y)h(y)dy$, and $\|g\| := \sqrt{\langle g, g \rangle}$.

A random element W taking values in $L_2(\mathbb{Y})$ is called *centered Gaussian* if $\langle h, W \rangle$ is zero-mean Gaussian for all $h \in L_2(\mathbb{Y})$. Equivalently, W is centered Gaussian if its characteristic function $\psi(h) = \mathbb{E}\exp\{i\langle h, W \rangle\}$ has the form $\psi(h) = \exp\{-\langle h, Ch \rangle / 2\}$ for some self-adjoint linear self-map C on $L_2(\mathbb{Y})$. C is called the covariance operator of W. The operator C satisfies $\langle g, Ch \rangle := \mathbb{E}\langle g, W \rangle \langle h, W \rangle$ for all $g, h \in L_2(\mathbb{Y})$. Covariance operators are themselves often defined by covariance functions. To say that κ is the covariance function for covariance operator C is to say that C is defined by

$$\langle g, Ch \rangle = \int \int \kappa(y, y')g(y)h(y')dydy' \tag{13}$$

for all $g, h \in L_2(\mathbb{Y})$. Further details on Hilbert space valued random variables can be found in Bosq (2000).

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

Assumption 3.1. There exists an open neighborhood U of the true parameter vector θ_0 and a measurable function $g \colon \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ such that

1. g satisfies $\int \left\{ \int g(x,y)dx \right\}^2 dy < \infty$,
2. $\theta \mapsto p(y \mid x, \theta)$ and $\theta \mapsto \phi(x, \theta)$ are continuously differentiable over U for all fixed $(x, y) \in \mathbb{X} \times \mathbb{Y}$, and
3. $d(x, y, \theta)$ satisfies $\sup_{\theta \in U} \|d(x, y, \theta)\|_E \leq g(x, y)$ for all $(x, y) \in \mathbb{X} \times \mathbb{Y}$.

In assumption 3.1, the symbol $\| \cdot \|_E$ is the Euclidean norm on $\mathbb{R}^K$. The subscript E is used to differentiate the Euclidean norm from the L_2 norm $\| \cdot \|$.

The set in which X_t takes values will be denoted by $\mathbb{X}$, a Borel subset of $\mathbb{R}^J$. Regarding the process for $\{X_t\}$ defined in (8), we make the following assumptions:

Assumption 3.2. For each $\theta \in \Theta$, the transition density $x' \mapsto q(x' \mid x, \theta)$ is ergodic, with unique stationary density $\phi(\cdot, \theta)$. Moreover, there exist a measurable function $V \colon \mathbb{X} \mapsto [1, \infty)$ and nonnegative constants $a < 1$ and $R < \infty$ satisfying

$$\sup_{|h| \leq V} \left| \int h(x') q^t(x' \mid x, \theta) dx' - \int h(x') \phi(x', \theta) dx' \right| \leq a^t R V(x)$$

as well as a function $\rho \in L_2(\mathbb{Y})$ such that $p(y \mid x, \theta) \leq \rho(y) V(x)^{1/2}$ for all $y \in \mathbb{Y}$, $x \in \mathbb{X}$, $\theta \in \Theta$ and $t \in \mathbb{N}$.

The main content of assumption 3.2 is that the process $\{X_t\}$ is always V-uniformly ergodic. This is a standard notion of ergodicity, and it generates enough mixing to yield asymptotic normality results under appropriate moment conditions. (The last part of assumption 3.2 is just a moment condition.) It applies to many common time series models under standard stationarity conditions. Details can be found in Meyn and Tweedie (2009, chapter 16).

3.2 Results

We now study the deviation between the true density $f = f(\cdot, \theta_0)$ and the PCMC density estimator $\hat{f}_m(\cdot, \hat{\theta}_n)$ defined in section 2. Since altering the values of densities at individual points does not change the distribution they represent, we focus on global error, treated as an element of $L_2(\mathbb{Y})$. The latter is a natural choice, since the

expectation of the squared norm of the error is then mean integrated squared error (MISE).

To state our main result, suppose now that the sequence of estimators $\{\hat{\theta}_n\}$ is asymptotically normal, with $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Sigma)$ for some symmetric positive definite $\Sigma = (\sigma_{ij})$. The simulation size m for $\hat{f}_m(\cdot, \hat{\theta}_n)$ is taken to be $\tau(n)$ where $\tau \colon \mathbb{N} \to \mathbb{N}$ is a given increasing function. In this setting we have the following result:

Theorem 3.1. *If assumptions 3.1–3.2 are valid and $\tau(n)/n \to \infty$, then*

$$\sqrt{n}\{\hat{f}_{\tau(n)}(\cdot, \hat{\theta}_n) - f(\cdot, \theta_0)\} \xrightarrow{d} W \qquad (n \to \infty) \tag{14}$$

where W is a centered Gaussian in $L_2(\mathbb{Y})$ with covariance function

$$\kappa(y, y') := \nabla_\theta f(y, \theta_0)^\top \, \Sigma \, \nabla_\theta f(y', \theta_0) \tag{15}$$

Here $\nabla_\theta f(y, \theta_0)$ represents the vector of partial derivatives $\partial f(y, \theta_0)/\partial \theta_k$. Thus, the asymptotic variance of 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. A proof of theorem 3.1 can be found in section 6.

In interpreting theorem 3.1 it is useful to note that $\kappa(y, y')$ is in fact the pointwise asymptotic covariance of the function on the left hand side of (14) evaluated at y and y'. In particular, $\kappa(y, y)$ is the pointwise asymptotic variance at y, in the sense that, for all $y \in \mathbb{Y}$,

$$\sqrt{n}\{\hat{f}_{\tau(n)}(y, \hat{\theta}_n) - f(y, \theta_0)\} \xrightarrow{d} N(0, \kappa(y, y)) \qquad (n \to \infty) \tag{16}$$

A consistent estimator for $\kappa(y, y)$ is $\hat{\kappa}(y, y) := \nabla_\theta f(y, \hat{\theta}_n)^\top \, \hat{\Sigma} \, \nabla_\theta f(y, \hat{\theta}_n)$ where $\hat{\Sigma}$ is a consistent estimator of Σ.

One way to understand (16) is to consider an ideal setting where all integrals have analytical solutions. In this case we can take the same estimator $\hat{\theta}_n$ and plug it directly into $f(y, \theta)$. A simple application of the delta method tells us that the asymptotic variance of this estimator $f(y, \hat{\theta}_n)$ is $\nabla_\theta f(y, \theta_0)^\top \, \Sigma \, \nabla_\theta f(y', \theta_0)$, which is the same value $\kappa(y, y)$ obtained by the PCMC estimator. Thus the PCMC estimator obtains the same asymptotic variance as the ideal setting, provided that the simulation sample size grows sufficiently quickly with n.

Example 3.1. As a simple illustration where integrals are tractable, suppose we observe a sequence $\{Y_t\}_{t=0}^n$ from the AR(1) model

$$Y_t = \theta Y_{t-1} + \xi_t, \qquad \{\xi_t\} \overset{\text{IID}}{\sim} N(0,1) \tag{17}$$

To estimate the stationary density f of Y_t via the PCMC density estimator we take $X_t := Y_{t-1}$, so that $Y_t = \theta X_t + \xi_t$, and hence $p(y \mid x, \theta) = N(\theta x, 1)$. Let $\hat{\theta}_n$ be the least squares estimate of θ. Given this estimate we can construct the simulated sequence $\{X_t^*\}$ by iterating on $X_{t+1}^* = \hat{\theta}_n X_t^* + \xi_t^*$. Taking that data and averaging over the conditional density as in (10) produces the PCMC density estimator. Regarding its asymptotic variance, note that both X_t and Y_t share the same stationary density, which in this case is $f(\cdot, \theta) = N(0, 1/(1-\theta^2))$. Also, the asymptotic variance of the OLS estimator $\hat{\theta}_n$ in this setting is $1/\mathbb{E}X_t^2$, which is $1-\theta^2$. Applying (15) and (16), the asymptotic variance of $\hat{f}_{\tau(n)}(y, \hat{\theta}_n)$ is therefore $f'_\theta(y, \theta_0)^2(1-\theta_0^2)$.

Figure 2 compares the pointwise asymptotic variance of the PCMC density estimator with that of the semiparametric estimator (6) when the model is the AR(1) process from example 3.1. The comparison is across different values of θ while holding the point y fixed ($y = 0$ in this case). The asymptotic variance of of the PCMC density estimator is lower at all points. The reason is that the simulation can almost eliminate the variance associated with averaging the conditional density over observations of the covariate X. This dominates the estimation effect associated with averaging over simulated rather than actual data.[5]

Example 3.2. We can also estimate $f(y, \theta)$ for the same AR(1) model via the estimator $\hat{z}(y)$ in (12). Since the parametric form in (17) is known here, it can be inferred that the stationary density $\phi(\cdot, \theta)$ of X_t is the zero-mean Gaussian distribution $N(0, 1/(1-\theta^2))$. We can estimate the unknown variance from $\{X_t\}$ with the sample variance s_n^2. We can then back out an estimate $\bar{\theta}_n$ of θ by solving $1/(1-\theta^2) = s_n^2$ for θ and then plugging this into $\phi(y, \theta)$. Next we compute the integral on the right hand side of (12) to produce an estimate $f(y, \bar{\theta}_n)$. Regarding the asymptotic variance, some elementary analysis shows that the asymptotic variance of $\bar{\theta}_n$ is $(1-\theta_0^2)\zeta$

[5]The comparison is by simulation to facilitate computation of the asymptotic variance of the semiparametric estimator. The simulation is over 1,000 replications with $n = 5,000$ and simulation size for the PCMC estimator set to 100,000.

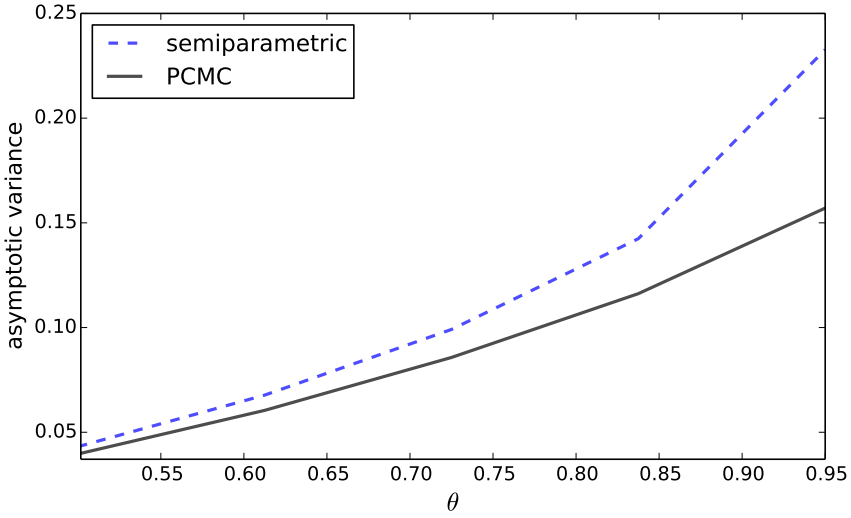

Figure 2: Asymptotic variances in the AR(1) model

where $\zeta := 1 + (1 - \theta_0^2)/(2\theta_0^2)$. Applying the delta method gives the asymptotic variance of $f(y, \bar{\theta}_n)$ as $f'_\theta(y, \theta_0)^2(1 - \theta^2)\zeta$. Since $\zeta > 1$, the pointwise asymptotic variance is larger than that of the PCMC (see example 3.1). The intuition for this was discussed in section 2.2.

4 Simulations

Next we apply the PCMC density estimator to a number of common models and examine its finite sample performance using Monte Carlo. In all cases, performance is measured in terms of mean integrated squared error (MISE).[6]

[6]The MISE of an estimator $\hat{g}$ of the true density $f = f(\cdot, \theta_0)$ is $\mathbb{E}\|\hat{g} - f\|^2$. If f has no closed form solution, then we compute it by conditional Monte Carlo. In all the following simulations, the MISE is approximated by averaging 1,000 realizations of $\|\hat{g} - f\|^2$.

4.1 Dynamic Factor Models

We begin with a simple application intended to illustrate conceptual issues. Consider the linear dynamic factor model $Y_t = \beta^\top X_t + \xi_t$ with

$$X_{t+1} = \Gamma X_t + \eta_{t+1}, \tag{18}$$

Here $Y_t \in \mathbb{R}$, $X_t \in \mathbb{R}^3$, $\Gamma = \text{diag}(\gamma_1, \gamma_2, \gamma_3)$ and all shocks are independent and $N(0,1)$. Following the asset pricing analysis of He et al. (2010), our baseline parameter setting is $\beta_1 = 6.26$, $\beta_2 = 1.32$, $\beta_3 = -1.09$, $\gamma_1 = 0.18$, $\gamma_2 = -0.14$, and $\gamma_3 = 0.21$. Using simulation, we compute the MISE of various estimators of the density f of Y_t. In this case, the true density is equal to

$$f = N(0, \sigma^2) \quad \text{for} \quad \sigma^2 := \frac{\beta_1^2}{1-\gamma_1^2} + \frac{\beta_2^2}{1-\gamma_2^2} + \frac{\beta_3^2}{1-\gamma_3^2} + 1 \tag{19}$$

In order to investigate how the performance of the estimator changes with the degree of persistence in the data, we also consider variations from the baseline. In particular, the baseline values of $\gamma = (\gamma_1, \gamma_2, \gamma_3)$ are multiplied by a scale parameter α, where α varies from 1 to 4. In all simulations, we take the data size $n = 200$.

To compute the PCMC density estimator $\hat{f}$ from any one of these data sets $\{Y_t, X_t\}$, we first estimate β and γ by least squares, producing estimates $\hat{\beta}_n$ and $\hat{\gamma}_n$. Next, $\{X_t^*\}_{t=1}^m$ is produced by simulating from the estimated version of (18), starting at $X_0^* = 0$ and setting $m = 10,000$. We then apply the definition (10) to obtain

$$\hat{f}(y) = \frac{1}{m} \sum_{t=1}^{m} \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}(y - \hat{\beta}_n^\top X_t^*)^2 \right\}$$

For comparison, we also compute the MISE of four alternative estimators: The semiparametric estimator $\hat{s}$ defined in (7), $\hat{z}$ defined in (12), the ordinary parametric estimator (OPE) and the nonparametric kernel density estimator (NPKDE), all of which are discussed earlier. The OPE uses the dynamic factor model to infer that (19) holds, and estimates f as $N(0, \hat{\sigma}_n^2)$ where $\hat{\sigma}$ is the sample standard deviation of $\{Y_t\}$. The NPKDE uses a standard Gaussian kernel and Silverman's rule for the bandwidth. To investigate the estimation effect, we also compute the PCMC density estimator with the true values of β and γ, and label it as $\hat{f}_0$. See the discussion in section 2.1.

The results of the simulation are shown in table 1. The far left-hand column is values of the scaling parameter α, so that higher values of α indicate more persistence in $\{X_t\}$. The remaining columns show MISE values for the six density estimators mentioned above. All are expressed as relative to the PCMC density estimator $\hat{f}$ (i.e., as multiples of this value).[7]

Regarding the outcome, observe that the rank of these estimators in terms of MISE is invariant with respect to data persistence (i.e., the value of α). Of the estimators that can actually be implemented (i.e., excluding $\hat{f}_0$), the PCMC density estimator has lowest MISE, followed by $\hat{z}$, $\hat{s}$, OPE and NPKDE in that order. Our interpretation is as follows: The reduction in MISE from the NPKDE to the OPE represents the benefit of imposing parametric structure on the data set $\{Y_t\}$. The reduction in MISE from the OPE to $\hat{s}$ represents the benefit of exploiting covariate data—in particular, the relationship (1) and the extra data $\{X_t\}$. The reduction in MISE from $\hat{s}$ to $\hat{z}$ represents the gains from estimating ϕ parametrically. The reduction in MISE from $\hat{z}$ to the PCMC estimate $\hat{f}$ represents the additional gain from exploiting the information contained in the order of the sample $\{X_t\}$—see section 2.2 for intuition.

The relatively low MISE for the PCMC density estimator becomes more pronounced as the degree of persistence in the data rises. This is not surprising, given the fact that the PCMC density estimator exploits order information in the sample $\{X_t\}$, and this information becomes more important with higher persistence. At the same time we note that persistence also pushes up the discrepancy between $\hat{f}_0$ and $\hat{f}$. In other words, there is an estimation effect that increases with persistence. This is because more persistence at a given sample size reduces the information content of the data, and makes the underlying parameters harder to estimate.

Further illustration is given in figure 3. The figure shows the true stationary distribution $f(\cdot, \theta_0)$ in bold, one observation of $\hat{s}$ for the same sample size and parameters, and one observation of $\hat{f}$ with the estimated 95% pointwise confidence bands for $f(\cdot, \theta_0)$ under the same sample size and parameters. (See the discussion in section 3.2.)

[7] Actual values for $\hat{f}$ ranged from 4.606×10^{-4} to 5.182×10^{-4}.

α	$\hat{f}_0$	$\hat{f}$	$\hat{z}$	$\hat{s}$	OPE	NPKDE
1.000	0.9998	1.000	1.965	2.256	2.531	3.014
1.150	0.9995	1.000	1.824	2.216	2.470	3.572
1.300	0.9993	1.000	2.031	2.252	2.414	3.614
1.450	0.9991	1.000	2.180	2.324	2.502	3.856
1.600	0.9966	1.000	2.328	2.333	2.514	3.831
1.750	0.9903	1.000	2.353	2.339	2.519	3.929
1.900	0.9890	1.000	2.412	2.438	2.506	4.079
2.050	0.9884	1.000	2.448	2.455	2.517	4.135
2.200	0.9874	1.000	2.467	2.498	2.510	4.079
2.350	0.9840	1.000	2.501	2.537	2.619	4.243
2.500	0.9545	1.000	2.540	2.549	2.798	4.404
2.650	0.9385	1.000	2.671	2.696	2.840	4.717
2.800	0.9162	1.000	2.759	2.785	2.932	4.801
2.950	0.8774	1.000	2.777	2.854	2.940	4.892
3.100	0.8660	1.000	2.835	2.899	2.958	4.897
3.250	0.8011	1.000	2.984	2.994	3.012	5.470
3.400	0.7942	1.000	2.998	3.082	3.116	5.421
3.550	0.7496	1.000	3.159	3.473	3.596	5.744
3.700	0.6994	1.000	3.384	3.528	3.572	6.168
3.850	0.6629	1.000	3.467	3.595	3.635	6.296
4.000	0.6107	1.000	3.563	3.578	3.591	6.424

Table 1: Relative MISE values for the dynamic factor model

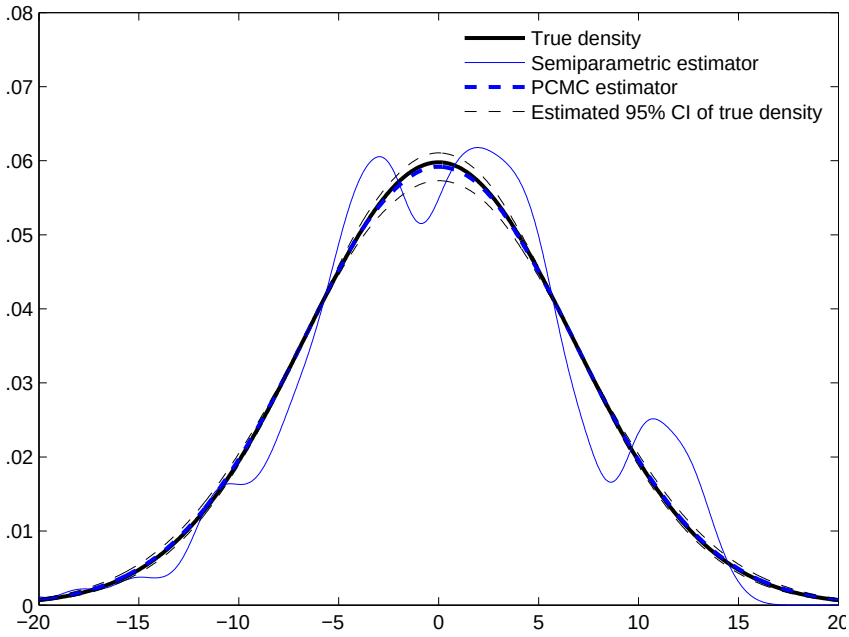

Figure 3: Realizations of $\hat{s}$ and PCMC in dynamic factor model

4.2 Linear AR(1)

In this section we study another simple example in order to further illustrate conceptual issues: the scalar Gaussian AR(1) model from (17). The construction of the PCMC density estimator for this model was discussed in example 3.1. In figure 2 we looked at asymptotic variance at a point. Here we look at finite sample MISE (error over the whole domain).

To study the MISE of the PCMC density estimator of f in finite samples, we compare its MISE with that of the semiparametric estimator $\hat{s}$, the direct parametric alternative $\hat{z}$, the ordinary parametric alternative (OPE) and the NPKDE when $n = 200$. (In this case, the OPE estimates f by observing that $f(y) = N(0, 1/(1-\theta^2))$ and estimating θ by maximum likelihood.) The method for implementing the NPKDE is identical to that used in section 4.1. The correlation coefficient θ is the only parameter, and it is set to 0.1, 0.3, 0.5, 0.7, 0.9, 0.95 and 0.995 in seven separate experiments. We also compare $\hat{f}_0$ with the PCMC density estimator to see how model estimation error influences the estimator under the AR(1) process.

θ	$\hat{f}_0$	$\hat{f}$	$\hat{z}$	$\hat{s}$	OPE	NPKDE
0.1	0.9985	1.000	1.003	1.304	1.721	2.226
0.3	0.9804	1.000	1.021	1.748	2.845	3.882
0.5	0.9517	1.000	1.027	2.437	2.914	4.450
0.7	0.9322	1.000	1.040	2.932	3.217	7.062
0.9	0.9219	1.000	1.400	3.750	4.528	14.127
0.95	0.8873	1.000	2.382	3.971	4.773	36.360
0.995	0.8521	1.000	2.416	4.106	19.217	175.979

Table 2: Relative MISE values for the AR(1) model

Table 2 presents results. All MISE values are expressed as multiples of the MISE for the PCMC density estimator. Of the estimators that can be implemented in practice (i.e., all but $\hat{f}_0$), the MISE for the PCMC density estimator is lowest for all values of θ. Notice also that the differences becomes more pronounced as θ increases. This result reiterates the point made in the previous section: 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. When $\theta = 0.1$, the data is almost IID, and preserving the order information in an estimate of ϕ has relatively little value. The benefit becomes larger when the persistence in the DGP increases.[8] Again, the estimation effect appears fairly weak when the AR(1) model is well estimated.

4.3 Threshold Autoregression

As our next application, we replace the linear AR(1) model with the nonlinear TAR model

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

[8] Actually the preceding intuition best explains the improvement that $\hat{f}$ makes over $\hat{z}$, both of which are parametric. Another factor at work is that more persistent data is in essence less informative than relatively independent observations. Hence the effective data size shrinks as we increase θ. This helps to explain why the nonparametric estimator—which has relatively weak finite sample properties—becomes less competitive.

The stationary density of Y_t has the skew-normal form $f(y) = 2\psi(y)\Psi(sy)$, where $s := \theta/\sqrt{1-\theta^2}$, and ψ and Ψ are the standard normal density and cumulative distribution respectively (see Andel *et al.*, 1984). The parameter θ can be estimated consistently by maximum likelihood. Following the simulations in Zhao (2010), we set $\theta = 0.6$ and $n = 200$. The MISE for the PCMC was found to be 5.836×10^{-4}. The MISE for $\hat{s}$ was 1.945 times larger, while that for the NPKDE was 8.823 times larger.

4.4 Markov Regime Switching

Next we consider a Markov regime switching model, in order to illustrate how the PCMC estimator is implemented in a latent variable model. The model we consider here is

$$Y_t = \mu_{X_t} + \sigma_{X_t}\,\xi_t, \qquad \{\xi_t\} \overset{\text{IID}}{\sim} N(0,1)$$

where $\{X_t\}$ is an unobservable two-state Markov chain with ergodic transition matrix Π. The stationary density of Y_t has the form $f = N(\mu_1, \sigma_1^2) \times \pi_1 + N(\mu_2, \sigma_2^2) \times \pi_2$, where (π_1, π_2) is the stationary distribution of Π. The model is estimated using maximum likelihood. The PCMC density estimator can then be implemented to obtain an estimate of f. In this case, the conditional density p in (10) is $p(y \mid X_t^*, \hat{\theta}_n) = N(\hat{\mu}_{X_t^*}, \hat{\sigma}^2_{X_t^*})$. The values $\{X_t^*\}$ are simulated from a maximum likelihood estimate $\hat{\Pi}$ of the matrix Π.

We investigate the finite sample performance of the PCMC estimator by comparing the MISE with that of the NPKDE when $n = 500$. (The semiparametric estimator is not available for comparison here because the state X_t is latent.) The parameters are set according to Smith and Layton's (2007) business cycle analysis, with $\mu_1 = 0.34$, $\mu_2 = -0.13$, $\sigma_1 = 0.38$, $\sigma_2 = 0.82$, and

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

The MISE of the PCMC estimator was found to be 9.418×10^{-3}, while that of the NPKDE was 0.015. Thus, the MISE of the NPKDE was roughly 1.6 times larger.

4.5 Stochastic Volatility in Mean

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

$$Y_t = c\,\sigma^2 \exp(h_t) + \sigma \exp(h_t/2)\xi_t \tag{20}$$

$$h_t = \kappa h_{t-1} + \sigma_\eta \eta_t \tag{21}$$

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

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

where, in view of (20), $p(y \mid h, \hat{\theta}_n) := N(\hat{c}_n \hat{\sigma}_n^2 \exp(h), \hat{\sigma}_n^2 \exp(h))$, and $\{h_t^*\}$ is generated by iterating on the estimated version of (21).

As with the Markov switching model, we investigate the finite sample performance of the PCMC estimator by comparing its MISE with that of the NPKDE when $n = 500$. We adopt the estimated parameter values in Koopman and Uspensky (2002), with $\kappa = 0.97$, $\sigma_\eta = 0.135$, $\sigma^2 = 0.549$, 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 2.3 times larger. Typical realizations of the estimators and the estimated 95% confidence bands are presented in figure 4.

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

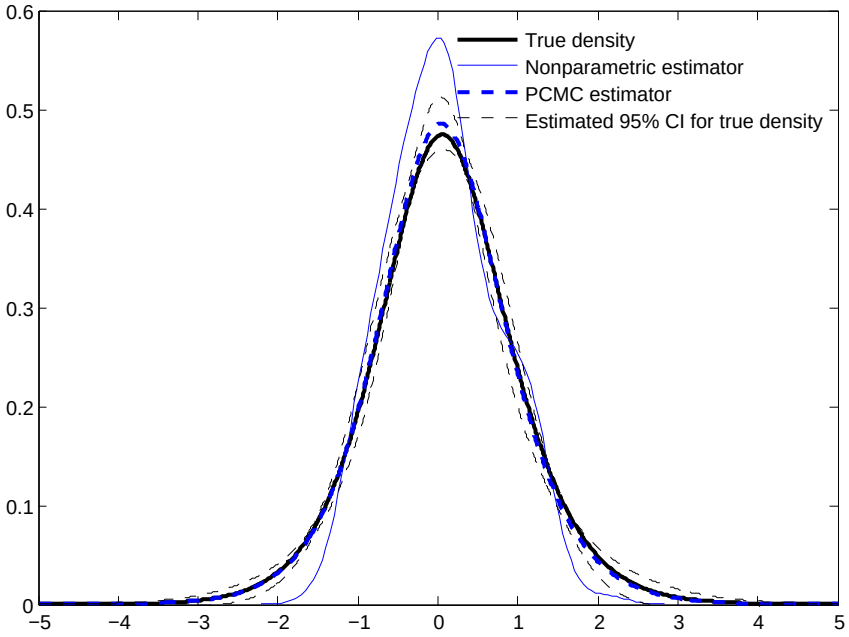

Figure 4: NPKDE and PCMC in the stochastic volatility in mean model

the PCMC estimator when the DGP is slightly misspecified. The first is a scalar model of the form

$$Y_t = \beta X_t + \xi_t \quad \text{and} \quad X_{t+1} = \gamma X_t + \theta \eta_t + \eta_{t+1} \tag{22}$$

where $\beta = 1$, $\theta = 0.1$ and (ξ_t, η_t) is IID and standard normal in $\mathbb{R}^2$. We vary the value of γ from 0.4 to 0.8 in order to investigating the sensitivity of the estimator performance to the degree of persistence of the data. The DGP for $\{X_t\}$ is misspecified as the AR(1) process

$$X_{t+1} = \gamma X_t + \eta_{t+1} \tag{23}$$

Table 3 reports the MISE of the PCMC density estimator calculated in the usual way, the misspecified PCMC density estimator (when the true process is (22) but the DGP of $\{X_t\}$ is misspecified as (23)), the estimator $\hat{s}$ and the NPKDE (all relative to the correctly specified PCMC). While the misspecification affects the performance of the PCMC estimator, in this case the effect is relatively small.

γ	PCMC	misspecified PCMC	$\hat{s}$	NPKDE
0.4	1.000	1.112	3.426	5.868
0.5	1.000	1.280	4.943	6.851
0.6	1.000	1.501	5.706	7.234
0.7	1.000	1.701	6.706	9.520
0.8	1.000	1.235	5.325	8.712

Table 3: MISE comparison, scalar factor model

6 Proofs

We introduce some simplifying notation. First, let $F(\theta)$ represent the function $f(\cdot, \theta)$ regarded as an element of $L_2(\mathbb{X})$. Thus, $\theta \mapsto F(\theta)$ is a mapping from Θ into $L_2(\mathbb{X})$ satisfying

$$F(\theta) = \int p(\cdot \mid x, \theta)\phi(x, \theta)dx \qquad (\theta \in \Theta) \tag{24}$$

Also, regarding the transition density $q(\cdot|x, \theta)$ in (8), recall that a Markov process from such a transition density can always be represented in the form

$$X_{t+1}^{\theta} = H(X_t^{\theta}, \eta_{t+1}, \theta) \quad \text{and} \quad X_0^{\theta} = x \in \mathbb{X} \tag{25}$$

where $\eta := \{\eta_t\}_{t\geq 1}$ is IID with marginal v over shock space D and H is a suitably chosen function (see Bhattacharya and Majumdar (2007, p. 284)). We let $v_\infty := v \times v \times \cdots$ be the joint law for the shocks, defined on the sequence space D^∞.

We begin with a simple lemma regarding the function

$$f_k'(y) := \frac{\partial f(y, \theta_0)}{\partial \theta_k} \tag{26}$$

Lemma 6.1. *Under assumption 3.1 we have* $f_k'(y) = \int d_k(x, y, \theta_0)dx$ *for all* $k = 1, \ldots, K$ *and* $y \in \mathbb{Y}$.

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

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

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,

$$|d_k(x, y, \theta)| := \left| \frac{\partial}{\partial \theta_k} p(y \mid x, \theta)\phi(x, \theta) \right| \leq h(x) \tag{27}$$

almost everywhere. Take $N = U$ and $h(x) := g(x,y)$ where U and g are as defined in assumption 3.1. By the conditions of assumption 3.1, the function h is integrable, and $\|d(x,y,\theta)\|_E \leq h(x)$ for all $\theta \in N$. This implies the inequality in (27), and lemma 6.1 is proved. □

Lemma 6.2. *If assumption 3.1 holds, then F is Hadamard differentiable at θ_0, with Hadamard derivative F'_{θ_0} given by*

$$F'_{\theta_0}(\theta) = \int \langle d(x,\cdot,\theta_0), \theta \rangle dx \in L_2(\mathbb{X}) \qquad (\theta \in \mathbb{R}^K) \tag{28}$$

Proof. Here $\langle \cdot, \cdot \rangle$ is the inner product in $\mathbb{R}^K$. To verify that F'_{θ_0} is the Hadamard derivative of F at θ_0, we must show that F'_{θ_0} defined in (28) is a bounded linear operator from $\mathbb{R}^K$ to $L_2(\mathbb{X})$ such that

$$\left\| \frac{F(\theta_0 + t_n\theta_n) - F(\theta_0)}{t_n} - F'_{\theta_0}(\theta) \right\| \to 0 \tag{29}$$

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 3.1,

$$\left| \int \langle d(x,y,\theta_0), \theta \rangle dx \right| \leq \int |\langle d(x,y,\theta_0), \theta \rangle| dx \leq \|\theta\|_E \int \|d(x,y,\theta_0)\|_E dx \leq \|\theta\|_E \int g(x,y) dx$$

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

The finiteness of the integral expression is guaranteed by assumption 3.1. Boundedness of the operator follows.

We now turn to the verification of (29). Fix $t_n \downarrow 0$ and $\theta_n \to \theta \in \Theta$. Let

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

and

$$\Lambda_n(x,y) := \frac{\zeta(x,y,\theta_0 + t_n\theta_n) - \zeta(x,y,\theta_0)}{t_n} - \langle d(x,y,\theta_0), \theta \rangle \tag{30}$$

Since

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

we have

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

and hence

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

Thus (29) will be established if we can show that

$$\int \left\{ \int \Lambda_n(x,y)dx \right\}^2 dy \to 0 \qquad (n \to \infty) \tag{31}$$

As a first step, note that $\Lambda_n \to 0$ pointwise on $\mathbb{X} \times \mathbb{Y}$. This first result is almost immediate from the definition of Λ_n in (30), since, for given x and y, the vector $d(x,y,\theta_0)$ is the vector of partial derivatives of the function $\theta \mapsto \zeta(x,y,\theta)$. As $\theta \mapsto p(y \mid x,\theta)$ and $\theta \mapsto \phi(x,\theta)$ are assumed to be continuously differentiable on U, the map $\theta \mapsto \zeta(x,y,\theta)$ 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 $\mathbb{R}^K$ the Frechet derivative and the Hadamard derivative coincide, and hence $|\Lambda_n(x,y)| \to 0$ by the definition of Hadamard differentiability.

In order to pass the limit through the integrals in (31), we next show that a scalar multiple of the function g defined in assumption 3.1 dominates Λ_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 $\theta_0 + t_n\theta_n \in U$ for all $n \geq N$. Without loss of generality we can choose the neighborhood U to be convex. With convex U, the mean value theorem in $\mathbb{R}^K$ implies existence of a vector $\theta_n^* \in U$ on the line segment between θ_0 and $t_n\theta_n$ with

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

Dividing both sides by t_n and using the definition of Λ_n in (30), we obtain

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

Applying assumption 3.1, we obtain $|\Lambda_n(x,y)| \leq g(x,y)(\|\theta_n\|_E + \|\theta\|_E)$. Since θ_n is convergent it is also bounded in n, and hence there exists a constant L with $|\Lambda_n(x,y)| \leq Lg(x,y)$ for all $n \geq N$.

Returning to the proof of (31), define

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

As a first step to proving (31), we claim that $h_n \to 0$ almost everywhere on $\mathbb{Y}$. To see this, observe that assumption 3.1 gives $\int h(y)dy < \infty$, and hence h is finite almost everywhere. For any y such that $h(y)$ is finite, we have $\int Lg(x,y)dx < \infty$. In addition, for this same y, we have $|\Lambda_n(x,y)| \leq Lg(x,y)$ and $\Lambda_n(x,y) \to 0$ for all $x \in \mathbb{X}$. It follows from the dominated convergence theorem that $\int \Lambda_n(x,y)dx \to 0$, 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 3.1. 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 (31), completing the proof of lemma 6.2. □

Lemma 6.3. *Under the conditions of theorem 3.1 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 equations* (13) *and* (15).

Proof of lemma 6.3. Let J be a random variable on $\mathbb{R}^K$ with $J \sim N(0,\Sigma)$, so that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to J. Let F be as defined in (24). The claim is then

$$\sqrt{n}\{F(\hat{\theta}_n) - F(\theta_0)\} \xrightarrow{d} N(0,C) \tag{32}$$

in $L_2(\mathbb{Y})$. Lemma 6.2 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 $\sqrt{n}\{F(\hat{\theta}_n) - F(\theta_0)\} \xrightarrow{d} F'_{\theta_0}(J)$ in $L_2(\mathbb{Y})$, where F'_{θ_0} is as defined in (28). Thus, it remains only to show that $F'_{\theta_0}(J) \sim N(0,C)$. Recalling the definition of f'_k in (26) and using lemma 6.1, we have

$$F'_{\theta_0}(J) = \int \langle d(x,\cdot,\theta_0), J\rangle dx = \sum_{k=1}^{K} \int d_k(x,\cdot,\theta_0)dx\, J_k = \sum_{k=1}^{K} f'_k J_k$$

Each f'_k is an element of $L_2(\mathbb{Y})$ because

$$|f'_k(y)| \leq \int \|d(x,y,\theta_0)\|_E \, dx \leq \int g(x,y)dx$$

and the right-hand side is square-integrable by assumption 3.1. It follows that $F'_{\theta_0}(J) = \sum_{k=1}^{K} f'_k J_k$ is an $L_2(\mathbb{Y})$ valued random variable.

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

univariate Gaussian by definition, and

$$\langle F'_{\theta_0}(J), h\rangle = \sum_{k=1}^{K} \langle f'_k, h\rangle J_k \tag{33}$$

To show that the $L_2(\mathbb{Y})$ expectation of $F'_{\theta_0}(J)$ is the zero element, we need to show that the (scalar) expectation of (33) is zero for all $h \in L_2(\mathbb{Y})$. This is true because $\mathbb{E}J_k = 0$ for all m.

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

$$\mathbb{E}\langle g, F'_{\theta_0}(J)\rangle \langle F'_{\theta_0}(J), h\rangle = \langle g, Ch\rangle \tag{34}$$

Regarding the right hand side of (34), by (13) and (15),

$$\langle g, Ch\rangle = \int\int \kappa(y,y')g(y)h(y')dydy' = \int\int \nabla_\theta f(y,\theta_0)^\top \Sigma \nabla_\theta f(y',\theta_0)g(y)h(y')dydy'$$

Using our notation for f'_i above we can reduce this to

$$\langle g, Ch\rangle = \sum_{i=1}^{K}\sum_{j=1}^{K} \sigma_{ij} \langle f'_i, g\rangle \langle f'_j, h\rangle \qquad (h, g \in L_2(\mathbb{Y})) \tag{35}$$

On the other hand, regarding the left hand side of (34), we have

$$\begin{aligned}
\mathbb{E}\langle F'_{\theta_0}(J), g\rangle \langle F'_{\theta_0}(J), h\rangle &= \mathbb{E}\left[\sum_{i=1}^{K}\langle f'_i, g\rangle J_i\right]\left[\sum_{j=1}^{K}\langle f'_j, h\rangle J_j\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{K}\sum_{j=1}^{K}\langle f'_i, g\rangle\langle f'_j, h\rangle J_i J_j\right]
\end{aligned}$$

Passing the expectation through the sum yields (35). In other words, (34) is valid. □

Lemma 6.4. *If the conditions of theorem 3.1 hold, then*

$$\left\|\tau(n)^{1/2}\hat{f}_{\tau(n)}(\cdot,\hat{\theta}_n) - f(\cdot,\hat{\theta}_n)\right\| = O_P(1)$$

Proof. The claim in lemma 6.4 is that the term is bounded in probability over n. A sufficient condition for boundedness in probability is that the square of this term is bounded in expectation, which is to say that

$$\sup_{n\in\mathbb{N}} \mathbb{E}\int \left\{\tau(n)^{-1/2}\sum_{t=1}^{\tau(n)} p(y \mid X^{\hat{\theta}_n}, \hat{\theta}_n) - f(\cdot,\hat{\theta}_n)\right\}^2 dy < \infty$$

or, simplifying notation further,

$$\sup_{n\in\mathbb{N}} \mathbb{E}\int \left\{ \tau(n)^{-1/2} \sum_{t=1}^{\tau(n)} \bar{p}(y \mid X^{\hat{\theta}_n}, \hat{\theta}_n) \right\}^2 dy < \infty \tag{36}$$

where

$$\bar{p}(y \mid x, \theta) := p(y \mid x, \theta) - f(y, \theta)$$

As a first step to proving (36), let $\theta \in \Theta$ be given. We claim that

$$\exists\, M \in \mathbb{N} \quad \text{s.t.} \quad \mathbb{E}\int \left\{ m^{-1/2} \sum_{t=1}^{m} \bar{p}(y \mid X_t^\theta, \theta) \right\}^2 dy \le M \quad \forall\, m \in \mathbb{N} \tag{37}$$

Applying Fubini's theorem and mutiplying and dividing by the square of the function ρ in assumption 3.2, we have

$$\mathbb{E}\int \left\{ m^{-1/2} \sum_{t=1}^{m} \bar{p}(y \mid X_t^\theta, \theta) \right\}^2 dy = \int \mathbb{E} \left\{ m^{-1/2} \sum_{t=1}^{m} \frac{\bar{p}(y \mid X_t^\theta, \theta)}{\rho(y)} \right\}^2 \rho(y)^2 dy$$

By Fatou's lemma we have

$$\limsup_{m\to\infty} \int \mathbb{E} \left\{ m^{-1/2} \sum_{t=1}^{m} \frac{\bar{p}(y \mid X_t^\theta, \theta)}{\rho(y)} \right\}^2 \rho(y)^2 dy \\ \le \int \limsup_{m\to\infty} \mathbb{E} \left\{ m^{-1/2} \sum_{t=1}^{m} \frac{\bar{p}(y \mid X_t^\theta, \theta)}{\rho(y)} \right\}^2 \rho(y)^2 dy$$

From assumption 3.2 we have $\{\bar{p}(y \mid x, \theta)/\rho(y)\}^2 \le V(x)$, and hence, applying the asymptotic normality result in theorem 17.0.1 of Meyn and Tweedie (2009),

$$\limsup_{m\to\infty} \mathbb{E} \left\{ m^{-1/2} \sum_{t=1}^{m} \frac{\bar{p}(y \mid X_t^\theta, \theta)}{\rho(y)} \right\}^2 = \gamma^2$$

where γ^2 is a finite quantity. In consequence,

$$\limsup_{m\to\infty} \mathbb{E}\int \left\{ m^{-1/2} \sum_{t=1}^{m} \bar{p}(y \mid X_t^\theta, \theta) \right\}^2 dy \le \gamma^2 \int \rho(y)^2 dy$$

The right hand side is finite, since $\rho \in L_2(\mathbb{Y})$ by assumption 3.2. The claim in (37) now follows.

To go from (37) to (36), fix $n \in \mathbb{N}$ and condition on $\hat{\theta}_n$ to get

$$\mathbb{E} \int \left\{ \tau(n)^{-1/2} \sum_{t=1}^{\tau(n)} \bar{p}(y \mid X^{\hat{\theta}_n}, \hat{\theta}_n) \right\}^2 dy$$
$$= \mathbb{E} \left[\mathbb{E} \left[\int \left\{ \tau(n)^{-1/2} \sum_{t=1}^{\tau(n)} \bar{p}(y \mid X^{\hat{\theta}_n}, \hat{\theta}_n) \right\}^2 dy \;\middle|\; \hat{\theta}_n = \theta \right] \right]$$

Applying (37) now gives the desired result. □

Proof of theorem 3.1. Adding and subtracting $f(\cdot, \hat{\theta}_n)$, we can write

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

If U, U_n and V_n are L_2 random elements with $U_n \xrightarrow{d} U$ in L_2 and $\|U_n - V_n\| = o_P(1)$ in $\mathbb{R}$, then $V_n \xrightarrow{d} U$ (cf., e.g., Dudley, 2002, lemma 11.9.4). In view this fact and the result in lemma 6.3, it suffices to show that

$$\sqrt{n} \left\| \hat{f}_{\tau(n)}(\cdot, \hat{\theta}_n) - f(\cdot, \hat{\theta}_n) \right\| = o_P(1)$$

To see that this holds, observe that

$$\sqrt{n} \left\| \hat{f}_{\tau(n)}(\cdot, \hat{\theta}_n) - f(\cdot, \hat{\theta}_n) \right\| = \left[\frac{n}{\tau(n)} \right]^{1/2} \left\| \tau(n)^{1/2} \hat{f}_{\tau(n)}(\cdot, \hat{\theta}_n) - f(\cdot, \hat{\theta}_n) \right\|$$

By lemma 6.4 the term on the far right is $O_P(1)$. By assumption we have $\tau(n)/n \to \infty$ and hence $n/\tau(n) \to 0$. The claim follows. □

References

[1] Aït-Sahalia, Y. and L. P. Hansen, 2009, *Handbook of Financial Econometrics*, Volume 1, North Holland.

[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] Bhattacharya, R. N. and M. Majumdar, 2007, Random dynamical systems: theory and applications, Cambridge University Press.

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

[6] Braun, R. A., H. Li and J. Stachurski, 2012, Generalized look-ahead methods for computing stationary densities. *Mathematics of Operations Research,* 37, 489-500.

[7] Calabrese, R. and M. Zenga, 2010, Bank loan recovery rates: Measuring and nonparametric density estimation, *Journal of Banking and Finance,* 34 (5), 903-911.

[8] de Vincent-Humphreys, R. and J. Noss, 2012, Estimating probability distributions of future asset prices: empirical transformations from option-implied risk-neutral to real-world density functions, Bank of England Working Paper No. 455.

[9] Escanciano, J. C., and D. T. Jacho-Chávez, 2012, $\sqrt{n}$-uniformly consistent density estimation in nonparametric regression models. *Journal of Econometrics* 167, 2, 305-316.

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

[11] Gelfand, A. E. and A. F. M. Smith, 1990, Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association,* 85, 398-409.

[12] 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.

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

[14] Henderson, S. G. and P. W. Glynn, 2001, Computing densities for Markov chains via simulation, *Mathematics of Operations Research,* 26, 375-400.

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

[16] 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.

[17] Martin, V., S. Hurn and D. Harris, 2012, *Econometric Modelling with Time Series Specification, Estimation and Testing.* Cambridge University Press.

[18] Meyn, S. and R. W. Tweedie, 2009, *Markov Chains and Stochastic Stability,* 2nd Edition, Cambridge UP.

[19] Polanski, A. and E. Stoja, 2011, Dynamic density forecasts for multivariate asset returns, *Journal of Forecasting,* 30 (6), 523-540.

[20] Rosenblatt, M., 1956, Remarks on some non-parametric estimates of a density function, *The Annals of Mathematical Statistics,* 27, 832-837.

[21] 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.

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

[23] 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.

[24] Støve, B., and D. Tjøstheim, 2012, A Convolution Estimator for the Density of Nonlinear Regression Observations. *Scandinavian Journal of Statistics,* 39, 282304.

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

[26] Zhao, Z., 2011, Nonparametric model validations for hidden Markov models with applications in financial econometrics. *Journal of Econometrics,* 162, 225-239.