Submitted to the Annals of Statistics

GOODNESS OF FIT FOR MARKOV MODELS: A DENSITY APPROACH*

BY VANCE MARTIN YOSHIHIKO NISHIYAMA AND JOHN STACHURSKI

We propose a density-based goodness of fit test suitable for time series data. The test compares the data against a parametric class of models specified in the null hypothesis. Estimation of smoothing parameters is not required, and the test has nontrivial power against $1/\sqrt{n}$ local alternatives.

1. Introduction. A central concern in time series modeling is whether or not a given class of models is able to effectively represent a certain data set. Despite the ongoing popularity of nonparametric methods, in applied time series analysis the classes of models in question are most often parametric. An attractive way to test the fit of a parametric class of models is via goodness of fit tests. These tests are typically characterized by broad applicability and absence of need for parametric structure on the alternative hypothesis.

In the time series setting, a variety of goodness of fit tests for parametric model classes have been proposed in the literature. An elegant example is Koul and Stute (1999), who apply martingale transform techniques to develop a distribution-free test of the autoregression function for real-valued Markov models. Tests for dynamic models focusing on the entire distribution have also been proposed, with representative examples including Aït-Sahalia, (1996), Bai (2003), Neumann and Paparoditis (2008), Chicheportiche and Bouchaud (2011) and Kristensen (2011). Chicheportiche and Bouchaud's (2011) study pertains to Kolmogorov-Smirnov and Cramér-von Mises tests in a time series environment. The tests of Bai (2003) and Neumann and Paparoditis (2008) are based around conditional distribution functions. The density-based tests, such as Aït-Sahalia's (1996) test, use nonparametric estimates in their test statistics, requiring the estimation of smoothing parameters. All of the papers mentioned above treat the case of univariate time series models.

In this paper we propose a different approach. We construct a new density-based goodness of fit test for stationary and ergodic Markov models that has three notable features. First, although

*The authors are grateful to Tim Kane, Yoichi Nishiyama, participants at the 2009 NCER conference at Princeton and the 2011 Australasian Meetings of the Econometric Society, and seminar participants at the Empirical Micro Research seminar at Tokyo University, the Nakanoshima Workshop at University of Osaka, and the 2011 Workshop on Statistical Analysis and Related Topics at Tokyo University. Our research was supported in part by Australian Research Council Grants DP120100321 and DP0987589, and by Japan Society for the Promotion of Science Grants-in-Aid 22330067.

AMS 2000 subject classifications: Primary 62M02; secondary 60J05

Keywords and phrases: Goodness of fit, Markov processes

the test statistic uses densities, there is no need to estimate smoothing parameters. As a result, the asymptotic theory is relatively simple, and the test has nontrivial power against $1/\sqrt{n}$ local alternatives. Second, in terms of theory, the test is no different for univariate data, multivariate data, or even infinite-dimensional data. (In fact the state space for the model requires no algebraic or topological structure at all.) Third, the test statistic combines both the conditional (i.e., transition) density of the process and the stationary density.

The test is not distribution free. Regarding this point it is worth observing that for tests of this nature there is a cost involved with construction of distribution free tests: The test statistic is typically more complicated, and computation of this more complicated test statistic cannot be avoided if one is to implement the test. In the test developed in this paper, the test statistic is relatively simple, with complexity pushed to the asymptotic distribution of the statistic. Given the power of modern computers this can be preferable for applications, since critical values of the asymptotic distribution can be calculated by simulation. In our case, the simulation procedure is straightforward.

In the sequel, we refer to the test proposed in this paper as the LAE test, where LAE stands for look-ahead estimator. The reason is that the original idea for the test came at least partly from study of Henderson and Glynn's (2001) look-ahead estimator, which was proposed as a method for computing intractable stationary densities using simulation. Also, to simplify presentation, in what follows we focus on null hypotheses that have only first order dependency. As will become clear in discussion of the test below, higher order dependencies change little in terms of the basic idea, while requiring a more complicated exposition.

2. Preliminaries. Throughout the paper, we consider stochastic processes taking values in an arbitrary measure space $(\mathbb{X}, \mathscr{X}, \mu)$, where $\mathscr{X}$ is countably generated and μ is σ-finite. To simplify notation, we use symbols such as dx and dy to indicate integration with respect to μ. A *density* on $\mathbb{X}$ is any nonnegative $\mathscr{X}$-measurable function f with $\int f(x)\, dx = 1$. A *density kernel* on $\mathbb{X}$ is a nonnegative $\mathscr{X} \otimes \mathscr{X}$-measurable function p such that $p(x, \cdot)$ is a density on $\mathbb{X}$ for all $x \in \mathbb{X}$. An $\mathbb{X}$-valued stochastic process $\{X_t\}_{t \in \mathbb{N}}$ on probability space $(\Omega, \mathscr{F}, \mathbf{P})$ will be called *p-Markov* if it is stationary and $p(X_t, \cdot)$ is the conditional density of X_{t+1} given $X_t, X_{t-1}, \dots, X_1$.

EXAMPLE 2.1. Let $\mathbb{X} = \mathbb{R}^k$, let $\mathscr{X}$ be the Borel sets, and let μ be Lebesgue measure. Consider a stationary nonlinear AR(1) process $X_{t+1} = g(X_t) + W_{t+1}$ with $\{W_t\}_{t \geq 1} \overset{\text{IID}}{\sim} \phi$, where ϕ is a density on $\mathbb{R}^k$ and g is a measurable function from $\mathbb{R}^k$ to itself. The sequence $\{X_t\}$ defined by this law of motion is p-Markov for $p(x, y) := \phi(y - g(x))$.

Let an arbitrary density kernel p on $\mathbb{X}$ be given, and let $\{X_t\}$ be p-Markov. The conditional distribution of X_t given $X_0 = x$ is represented by the t-th order density $p^t(x, \cdot)$, where $p^1 := p$

and $p^t(x,y) := \int p(x,z)p^{t-1}(z,y)dz$. A density ψ on $\mathbb{X}$ is called *stationary* with respect to p if

$$\int p(x,y)\psi(x)dx = \psi(y) \qquad \forall\, y \in \mathbb{X} \tag{1}$$

In all cases we consider, p will have a unique stationary density ψ. If $\{X_t\}$ is stationary and p-Markov, then $X_t \sim \psi$ for all $t \geq 0$. To simplify notation, in what follows we let

$$\bar{p}(x,y) := p(x,y) - \psi(y) \qquad ((x,y) \in \mathbb{X} \times \mathbb{X}) \tag{2}$$

Let $L_2 := L_2(\mathbb{X}, \mathscr{X}, \mu)$ be the $\mathscr{X}$-measurable functions h from $\mathbb{X}$ to $\mathbb{R}$ such that $\int h(x)^2 dx := \int h(x)^2 \mu(dx)$ is finite. As usual, elements of L_2 equal μ-almost everywhere are identified. The inner product and norm on L_2 are defined by $\langle g, h\rangle := \int g(x)h(x)dx$ and $\|h\| := \langle h, h\rangle^{1/2}$. Since $\mathscr{X}$ is countably generated, the space $(L_2, \|\cdot\|)$ is separable.

An L_2*-valued random variable* F is a measurable map from $(\Omega, \mathscr{F})$ into L_2 paired with its Borel sets. If $\mathbf{E}\|F\| < \infty$, where $\mathbf{E}$ is the ordinary scalar expectation, then the *vector expectation* $\mathcal{E}F$ of F exists and is equal to the unique element of L_2 satisfying $\langle \mathcal{E}F, h\rangle = \mathbf{E}\langle F, h\rangle$ for all $h \in L_2$ (cf., e.g., Bosq, 2000). If $\mathbf{E}\|F\|^2 < \infty$, then the *covariance operator* C of F is the linear operator defined by

$$\langle g, Ch\rangle = \mathbf{E}\langle g, F - \mathcal{E}F\rangle\langle h, F - \mathcal{E}F\rangle \qquad \forall\, g, h \in L_2 \tag{3}$$

An L_2-valued random variable G is called *Gaussian* if $\langle h, G\rangle$ is normally distributed on $\mathbb{R}$ for every $h \in L_2$. We write $G \sim N(m, C)$ if G is Gaussian on L_2 with mean $m = \mathcal{E}G$ and covariance operator C. Letting $\{Z_\ell\}_{\ell \geq 1}$ be an IID sequence of standard normal random variables and $\{\lambda_\ell\}_{\ell \geq 1}$ be the eigenvalues of C, we have the following well-known fact:

LEMMA 2.1. *If $G \sim N(0, C)$ on L_2, then $\|G\|^2$ has the same distribution as $\sum_{\ell=1}^\infty \lambda_\ell Z_\ell^2$.*

Throughout, the symbol $\xrightarrow{\mathscr{D}}$ means convergence in distribution. (If E is a metric space, then $Y_n \xrightarrow{\mathscr{D}} Y$ on E means that $\mathbf{E}g(Y_n) \to g(Y)$ for every continuous bounded $g \colon E \to \mathbb{R}$.)

A density kernel p on $\mathbb{X}$ is called *ergodic* if it has a unique stationary density ψ, and any p-Markov process satisfies the strong law of large numbers (see Meyn and Tweedie, 2009, theorem 17.1.7, and Lindvall, 2002, theorem 21.12 for the many equivalent definitions of ergodicity). *Geometric ergodicity* requires that, in addition, there exist positive constants $\lambda < 1$ and $L < \infty$ and a weight function $V \colon \mathbb{X} \to \mathbb{R}_+$ such that

$$\int V(x)\psi(x)dx < \infty \quad \text{and} \quad \left| \int_B p^t(x,y)dy - \int_B \psi(y)dy \right| \leq \lambda^t L V(x) \tag{4}$$

for all $B \in \mathscr{X}$, $x \in \mathbb{X}$ and $t \in \mathbb{N}$. (Kristensen (2007) gives geometric ergodicity conditions for a number of popular time-series models, including nonlinear ARMA, bilinear, GARCH and random coefficient models. Meyn and Tweedie (2009, chapter 15) provide a general treatment.)

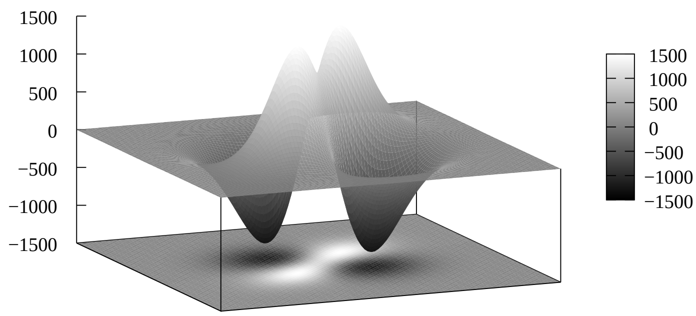

Fig 1: The function $\zeta(y, y')$ for $p(x, \cdot) = N(ax + b, \sigma^2)$

In what follows, we will say that p is *V-mixing* if there exists a function $V\colon \mathbb{X} \to \mathbb{R}_+$ such that p is geometrically ergodic with weight function V, and, in addition, there are nonnegative constants c_0, c_1 and γ with $\gamma < 1$ and

$$\int p(x, y)^2 dy \leq c_0 + c_1 V(x)^\gamma \qquad \forall\, x \in \mathbb{X} \tag{5}$$

Together, geometric ergodicity and (5) provide the mixing and moment conditions necessary for our asymptotic theory to hold. In particular, we will make use of the following central limit result, which is similar to theorem 1 of Stachurski and Martin (2008). The proof can be found in section 7.

THEOREM 2.1. *If p is V-mixing and $\{X_t\}$ is p-Markov, then on L_2 we have*

$$n^{-1/2} \sum_{t=1}^{n} \bar{p}(X_t, \cdot) \xrightarrow{\mathscr{D}} N(0, \Lambda) \qquad (n \to \infty) \tag{6}$$

for the covariance operator Λ satisfying

$$\langle h, \Lambda h \rangle = \mathbf{E}\langle \bar{p}(X_1, \cdot), h\rangle^2 + 2 \sum_{t=2}^{\infty} \mathbf{E}\langle \bar{p}(X_1, \cdot), h\rangle \langle \bar{p}(X_t, \cdot), h\rangle \qquad (h \in L_2) \tag{7}$$

The covariance operator Λ in (7) has the integral representation $\Lambda h(y') := \int \zeta(y, y') h(y) dy$, where the covariance function ζ is given by

$$\zeta(y, y') := \int \bar{p}(x, y)\bar{p}(x, y')\psi(x)dx + \sum_{t=2}^{\infty} \left\{ \int \bar{p}(x, y)\bar{p}^t(x, y')\psi(x)dx + \int \bar{p}(x, y')\bar{p}^t(x, y)\psi(x)dx \right\}$$

A plot of ζ is given in figure 1, corresponding to the density kernel $p(x, \cdot) = N(ax + b, \sigma^2)$. The parameters used in the plot were estimated from US short rate data by Aït Sahalia (1996).

3. Simple Null Hypotheses. To illustrate some of the underlying ideas, we begin discussion of the LAE test by looking at a simple null hypothesis. Let p be a density kernel on $\mathbb{X}$ that is V-mixing with stationary density ψ, and let $\{X_t\}_{t=1}^n$ be an $\mathbb{X}$-valued time series. Our null hypothesis is that this data is p-Markov. To construct a test of this hypothesis, consider the deviation

$$\frac{1}{n}\sum_{t=1}^{n} p(X_t, y) - \psi(y) \tag{8}$$

When the null hypothesis holds, the sequence $\{X_t\}_{t=1}^n$ is stationary and ergodic with common density ψ, and hence, for large n,

$$\frac{1}{n}\sum_{t=1}^{n} p(X_t, y) - \psi(y) \approx \int p(x,y)\psi(x)dx - \psi(y) = 0 \tag{9}$$

where the last equality is by the definition of ψ. Thus, under the null, the deviation in (8) should be small for large n. Since this argument is valid for any given y, we can adopt a functional perspective, regarding

$$\frac{1}{n}\sum_{t=1}^{n} \bar{p}(X_t, \cdot) := \frac{1}{n}\sum_{t=1}^{n} p(X_t, \cdot) - \psi(\cdot) \tag{10}$$

as a random element taking values in L_2, and rejecting the null when its norm is large—that is, when its realization lies outside a certain sphere centered on the origin of L_2.

In practice, we multiply the term in (10) by $\sqrt{n}$ and consider the squared norm rather than norm. In particular, from theorem 2.1, lemma 2.1 and the continuous mapping theorem, we see that if the null hypothesis holds, $\{Z_\ell\}_{\ell \geq 1}$ is an IID sequence of scalar standard normal random variables and $\{\lambda_\ell\}_{\ell \geq 1}$ are the eigenvalues of Λ defined in (7), then

$$T_n := \frac{1}{n}\left\| \sum_{t=1}^{n} \bar{p}(X_t, \cdot) \right\|^2 \xrightarrow{\mathscr{D}} \sum_{\ell=1}^{\infty} \lambda_\ell Z_\ell^2 \qquad (n \to \infty) \tag{11}$$

Thus, if $\alpha \in (0,1)$ and c_α^Λ is the $1-\alpha$ quantile of $\sum_\ell \lambda_\ell Z_\ell^2$, then the test rejecting H_0 when $T_n > c_\alpha^\Lambda$ is asymptotically of size α.

The limiting distribution for the test statistic derived in (11) has a similar structure to that found for the Cramér-von Mises test (cf., e.g., del Barrio *et al.*, 2007). Regarding implementation of the LAE test, the integral in the definition of T_n can be computed numerically. To compute the critical value c_α^Λ, one approach is to approximate the eigenvalues $\{\lambda_\ell\}_{\ell \geq 1}$ of Λ using the covariance function ζ defined in section 2 paired with a numerical technique such as Galerkin projection. However, it is usually simpler to simulate the test statistic T_n under the null and take the $1-\alpha$ quantile. Details are discussed in the technical supplement (Martin *et al.*, 2012).

One immediate comment on the test is that consistency of the test will not hold when the alternative generating the data $\{X_t\}$ is not p-Markov and yet is ergodic and suitably mixing with common marginal density ψ. (The intuition can be see by considering (9), which will again be valid.) However, when we consider a more practical version of this test in section 4 below, we will see that consistency holds under a rather broad range of alternatives.

A second comment is that to implement the test we need to be able to evaluate both $p(x,y)$ and $\psi(y)$. For discrete time models it is common that p can be evaluated but ψ has no analytical solution. In the technical supplement (Martin *et al.*, 2012), we discuss how this problem can be overcome by simulation.

3.1. *Local Alternatives.* The structure of the LAE test implies nontrivial power against $1/\sqrt{n}$ local alternatives under suitable regularity conditions. To clarify this point, consider the test

$$H_0 : \{X_t\}_{t=1}^n \text{ is } p\text{-Markov} \qquad \text{vs} \qquad H_1 : \{X_t\}_{t=1}^n \text{ is } p_n\text{-Markov for all } n$$

where p is V-mixing and $\{p_n\}$ is the sequence of density kernels defined by $p_n(x,y) := p(x,y) + k(x,y)/\sqrt{n}$ for some fixed $k\colon \mathbb{X} \times \mathbb{X} \to \mathbb{R}$. We set $Y_n := n^{-1/2} \sum_{t=1}^n \bar{p}(X_t, \cdot)$, so that Y_n is the random element of L_2 in (6), and $T_n = \|Y_n\|^2$, where T_n is the test statistic in (11). Also, let τ be the element of L_2 defined by

$$\tau(y) := \sum_{t=1}^{\infty} \mathbf{E} \left\{ \bar{p}(X_{t+1}, y) \frac{k(X_1, X_2)}{p(X_1, X_2)} \right\}$$

where the expectation is taken under H_0. We will assume throughout that k and p satisfy the third moment condition

$$\mathbf{E} \sup_{\delta \in [0,1]} \frac{|k(X_1, X_2)|^3}{|p(X_1, X_2) + \delta\, k(X_1, X_2)|^3} < \infty \tag{12}$$

THEOREM 3.1. *If H_1 and (12) hold, then $Y_n \stackrel{\mathscr{D}}{\to} \tau + N(0, \Lambda)$.*

While theorem 2.1 shows that $Y_n \stackrel{\mathscr{D}}{\to} N(0, \Lambda)$ under H_0, theorem 3.1 tells us that under H_1 it converges instead to $N(\tau, \Lambda)$. Since T_n is equal to the squared norm of Y_n, theorem 3.1 implies non-trivial power for the LAE test whenever $\tau \neq 0$. Our proof of theorem 3.1 uses a contiguity argument, based on a Hilbert space extension of Le Cam's third lemma. The proof is long but entirely standard, and hence left to the technical supplement (Martin *et al.*, 2012).

4. The Test with Estimated Parameters. The LAE test discussed above represents a goodness of fit test for individual models. A more practical setting is where we have a parametric class of models, and aim to test the hypothesis that the data are generated by some model in this class. In this case we need to augment our asymptotic theory to accommodate estimated parameters.

4.1. *Asymptotics under H_0.* Let Θ be an open convex subset of $\mathbb{R}^M$, and let $\{p_\theta\}_{\theta\in\Theta}$ be a parametric family of density kernels such that p_θ is V_θ-mixing for each $\theta \in \Theta$. Let ψ_θ be the unique stationary density corresponding to p_θ. When convenient, we write $p(\theta, x, y)$ instead of $p_\theta(x, y)$, and $\psi(\theta, y)$ in place of $\psi_\theta(y)$. In addition, let $\bar{p}(\theta, x, y) := p(\theta, x, y) - \psi(\theta, y)$. We begin with a limit theorem for the distribution of the test statistic in the estimated parameter case. In the assumptions below, $\|\cdot\|_E$ denotes the Euclidean norm in $\mathbb{R}^M$, as opposed to $\|\cdot\|$, the norm in L_2. We suppose the existence of an asymptotically linear and $\sqrt{n}$-consistent sequence of estimators $\{\hat{\theta}_n\}$ for the parameter vector θ. In particular, when $\{X_t\}$ is p-Markov,

ASSUMPTION 4.1. $\hat{\theta}_n$ admits the expansion $\hat{\theta}_n - \theta = n^{-1}\sum_{t=1}^{n-r} g_\theta(X_t, \ldots, X_{t+r}) + o_P(1)$, where $r \in \mathbb{N}$ and $g_\theta \colon \mathbb{R}^{r+1} \to \mathbb{R}^M$ is an influence function such that

1. $\mathbf{E} g_\theta(X_t, \ldots, X_{t+r}) = 0$
2. $\|g_\theta(x_0, \ldots, x_r)\|_E^{2+\delta} \leq \sum_{k=0}^r V_\theta(x_k)$ on $\mathbb{X}^{r+1}$ for some $\delta > 0$

ASSUMPTION 4.2. The vector $D\bar{p}(\theta, x, y)$ of partial derivatives $\frac{\partial}{\partial\theta_m}\bar{p}(\theta, x, y)$ exists and satisfies

$$\int \left\{ \frac{\partial}{\partial\theta_m} \bar{p}(\theta, x, y) \right\}^2 dy \leq V_\theta(x)^{1/2} \quad \text{for all } (x, y) \in \mathbb{X}\times\mathbb{X} \text{ and } \theta \in \Theta$$

ASSUMPTION 4.3. There exists a constant $\alpha > 0$ and function $K_2 \colon \mathbb{X}\times\mathbb{X} \to \mathbb{R}$ such that $\int\int K_2(x, y)^2 dy \psi(\theta, x) dx < \infty$ and

$$\|D\bar{p}(\theta, x, y) - D\bar{p}(\theta', x, y)\|_E \leq K_2(x, y)\|\theta - \theta'\|_E^\alpha \quad \text{for all } (x, y) \in \mathbb{X}\times\mathbb{X} \text{ and } \theta \in \Theta$$

For each $\theta \in \Theta$, the pair (p_θ, g_θ) defines a covariance operator Σ_θ on L_2, the expression for which is presented in (25) below.

THEOREM 4.1. *If assumptions 4.1–4.3 hold and $\{X_t\}$ is p_θ-Markov, then $n^{-1/2}\sum_{t=1}^n \bar{p}(\hat{\theta}_n, X_t, \cdot) \xrightarrow{\mathscr{D}} N(0, \Sigma_\theta)$ on L_2 as $n \to \infty$, and, as a consequence,*

$$\hat{T}_n := \frac{1}{n}\left\| \sum_{t=1}^n \bar{p}(\hat{\theta}_n, X_t, \cdot) \right\|^2 \xrightarrow{\mathscr{D}} \sum_{\ell=1}^\infty \sigma_\ell^\theta Z_\ell^2 \qquad (n \to \infty) \tag{13}$$

where $\{Z_\ell\}_{\ell\geq 1} \overset{\text{IID}}{\sim} N(0, 1)$ and $\{\sigma_\ell^\theta\}_{\ell\geq 1}$ are the eigenvalues of Σ_θ.

Fix $\alpha \in (0, 1)$. Let $c_\alpha^\Sigma(\theta)$ denote the $1 - \alpha$ quantile of the random variable $\sum_{\ell=1}^\infty \sigma_\ell^\theta Z_\ell^2$. Consider the null hypothesis

$$H_0 \colon \text{ the data } \{X_t\}_{t=1}^n \text{ is } p_\theta\text{-Markov for some } \theta \in \Theta \tag{14}$$

When H_0 holds we let $\theta_0 \in \Theta$ denote the true value of θ. In view of (13), under the null hypothesis (14), a test rejecting H_0 when $\hat{T}_n$ exceeds $c_\alpha^\Sigma(\theta_0)$ is asymptotically of size α. Since θ_0 is not observable and $c_\alpha^\Sigma(\theta_0)$ cannot be evaluated, we approximate it with $c_\alpha^\Sigma(\hat{\theta}_n)$. This gives the test

$$\text{reject } H_0 \text{ if } \hat{T}_n > c_\alpha^\Sigma(\hat{\theta}_n) \tag{15}$$

THEOREM 4.2. *If the conditions of theorem 4.1 hold and c_α^Σ is continuous at θ_0, then the test (15) is asymptotically of size α.*

To implement the LAE test in (15), we need a means of evaluating $c_\alpha^\Sigma(\theta)$ for given θ. One possibility is to compute the eigenvalues $\{\sigma_\ell^\theta\}_{\ell\geq 1}$ of Σ_θ, and then the $1-\alpha$ quantile of $\sum_{\ell=1}^\infty \sigma_\ell^\theta Z_\ell^2$. A simpler method is to use simulation of the test statistic under the null. Details are given in the technical supplement (Martin *et al.*, 2012).

4.2. *Related Tests.* One way to interpret the LAE test is as an infinite dimensional version of Hansen's (1982) J-test. The latter begins with a moment restriction of the form $\mathbf{E}\, G(\theta, X_t) = 0$ for some function G taking values in $\mathbb{R}^m$. The null hypothesis of the test is

$$\exists\, \theta \in \Theta \;\text{ such that }\; \mathbf{E} G(\theta, X_t) = 0 \tag{16}$$

The null hypothesis is rejected if

$$\frac{1}{n} \left\| \sum_{t=1}^n G(\hat{\theta}_n, X_t) \right\|_W^2 \tag{17}$$

is large relative to a certain χ^2 distribution, where $\|\cdot\|_W$ is a weighted euclidean norm. To formulate the LAE test in a parallel manner, recall the null hypothesis (14). Under this null, there exists a θ with $X_t \sim \psi(\theta,\cdot)$ for all t, and hence $\mathbf{E}\bar{p}(\theta, X_t, y) = \int p(\theta, X_t, y)\psi(\theta, x)dx - \psi(\theta, y) = 0$. Treating all y simultaneously, we can write this restriction as

$$H_0^J\colon\; \exists\, \theta \in \Theta \;\text{ such that }\; \mathcal{E}\bar{p}(\theta, X_t, \cdot) = 0 \tag{18}$$

where $\mathcal{E}$ is the vector expectation discussed in section 2 and the zero on the right-hand side is the origin of L_2. This is an infinite-dimensional version of (16), and the LAE test statistic in (13) is analogous to (17).

4.3. *Consistency of the Test.* In the preceding section we considered two null hypotheses, H_0 as defined in (14) and H_0^J as defined in (18). As explained in the paragraph preceding (18), the null H_0 implies H_0^J, and hence rejection will be easier on the complement of H_0^J. Indeed, as shown immediately below, the LAE test (15) is consistent for H_0^J, rejecting all alternatives outside H_0^J in the limit with probability one. In applications, alternatives outside H_0 are quite likely to also lie outside H_0^J. However, as we argue below, even for the alternatives in $H_0^J \setminus H_0$, the test will often be consistent. Throughout the following discussion, $\{X_t\}$ is an $\mathbb{X}$-valued stochastic process which we treat as data, Θ is a bounded convex subset of $\mathbb{R}^M$, and $\{p_\theta\}_{\theta\in\Theta}$ is a fixed family of density kernels (treated as the null). We focus on the case where the alternative $\{X_t\}$ is stationary and ergodic. The precise assumptions are as follows:

ASSUMPTION 4.4. $\{X_t\}$ is stationary and $\frac{1}{n}\sum_{t=1}^{n} h(X_t) \xrightarrow{p} \mathcal{E}h(X_t)$ for all $h\colon \mathbb{X} \to L_2$ such that $\mathcal{E}h(X_t)$ exists. The expectation $\mathbf{E}\int p(\theta, X_t, y)^2 dy$ is finite for all $\theta \in \Theta$.

ASSUMPTION 4.5. The sequence $\hat{\theta}_n$ converges in probability to some pseudo-true value θ_2.

ASSUMPTION 4.6. There exist $\eta, \xi \in L_2$ such that, for all $x, y \in \mathbb{X}$ and $\theta, \theta' \in \Theta$,

$$|p(\theta, x, y) - p(\theta', x, y)| \leq \eta(y)\|\theta - \theta'\|_E \quad \text{and} \quad |\psi(\theta, y) - \psi(\theta', y)| \leq \xi(y)\|\theta - \theta'\|_E$$

THEOREM 4.3. *If assumptions 4.4–4.6 hold, then* $\hat{T}_n/n \xrightarrow{p} \| \mathcal{E}\bar{p}(\theta_2, X_t, \cdot) \|$ *as* $n \to \infty$.

As a consequence, if $\|\mathcal{E}\bar{p}(\theta_2, X_t, \cdot)\| > 0$, then $\hat{T}_n \xrightarrow{p} \infty$. Thus, the test (15) is consistent against the set of alternatives such that $\mathcal{E}\bar{p}(\theta_2, X_t, \cdot)$ is not zero. Referring back to (18), it is now immediate that the test is consistent against alternatives in the negation of H_0^J.

In view of theorem 4.3, the LAE test will also be consistent for the alternatives in $H_0^J \setminus H_0$ provided that $\|\mathcal{E}\bar{p}(\theta_2, X_t, \cdot)\|$ is not zero. Typically, this will be the case as long as the estimator $\hat{\theta}_n$ is not chosen to minimize the empirical counterpart of $\|\mathcal{E}\bar{p}(\theta, X_t, \cdot)\|$, which (after taking squares) is

$$\left\| \frac{1}{n}\sum_{t=1}^{n} \bar{p}(\theta, X_t, \cdot) \right\|^2 = \int \left\{ \frac{1}{n}\sum_{t=1}^{n} p(\theta, X_t, y) - \psi(\theta, y) \right\}^2 dy \tag{19}$$

In fact, for the LAE test to be practical, we require that terms of the form $p(\theta, x, y)$ can be evaluated, and in this setting the most natural estimator $\hat{\theta}_n$ is the maximum likelihood estimator. The maximum likelihood estimator minimizes the objective function $-\sum_{t=1}^{n-1}\log p(\theta, X_t, X_{t+1})$. The minimizer of this objective does not in general minimize (19).

It should also be added that the conditions of the theorem are sufficient but not necessary for consistency. For example, while assumption 4.4 requires a stationary and ergodic alternative, intuition suggests that for some choices of H_0 and nonstationary alternatives, the test is likely to reject with high probability when the sample size is large. This point is illustrated in section 5.3.

5. Applications. Next we present applications illustrating properties of the test. Additional details on how the Monte Carlo experiments were run can be found in the technical supplement (Martin *et al.*, 2012).

5.1. *Properties of the Test under* H_0. In the introduction we briefly discussed the relationship between the LAE test proposed in this paper and the test of Aït-Sahalia (1996). Aït-Sahalia's test is based on the L_2 deviation between a theoretical stationary density and a nonparametric kernel density estimate of the stationary density using the data $X_1, \ldots, X_n$. His results have initiated an important line of research. One finding has been that Aït-Sahalia's test statistic might require very large data sizes to attain its asymptotic distribution, causing excessively high rejection rates in finite samples when the asymptotic critical value is adopted (Pritsker, 1998). A possible factor in slow convergence to the asymptotic distribution is the use of nonparametric kernel density estimators in the test statistic. The test proposed here provides a new perspective on this problem. The fact that the LAE test estimates no smoothing parameters and uses a density estimator that is $\sqrt{n}$-consistent under the null suggests that the LAE test might have lower size distortion in small samples.

To investigate this idea, we conducted an experiment to re-examine the discussion of size distortion reported in Pritsker (1998). Pritsker investigated rejection rates for Aït-Sahalia's test under a true null when the sample size is relatively small and the asymptotic critical value is used. Following Pritsker, the underlying model in our experiment was the Vasicek model of interest rates, where the rate of interest X_t follows $dX_t = \kappa(b - X_t)dt + \sigma dW_t$. Here κ, b and σ are parameters, and W_t is standard Brownian motion in $\mathbb{R}$. The transition probability function associated with this process is

$$q(t,x,y) := \{2\pi\, v(t)\}^{-1/2} \exp\left\{ \frac{-[y - m(t,x)]^2}{2v(t)} \right\} \tag{20}$$

where $v(t) := \sigma^2(1 - e^{-2\kappa t})/(2\kappa)$ and $m(t,x) := b + (x - b)e^{-\kappa t}$. If a unit of time corresponds to one year and $\{X_t\}_{t=1}^n$ is a sequence of monthly observations from the model, then $\{X_t\}_{t=1}^n$ is p-Markov for $p(x,y) := q(1/12, x, y)$. The stationary density ψ is $N(b, \sigma^2/(2\kappa))$. Our baseline parameters for our experiment were $\kappa = 0.85837$, $b = 0.089102$ and $\sigma^2 = 0.0021854$, as estimated from US short rate data by Aït Sahalia (1996). For the data generating process (DGP) we used this model, while for H_0 we hypothesized correctly that the data was generated by a Vasicek model for some choice of parameters.

Beginning with Aït-Sahalia's test, we computed the asymptotic critical value of his test at the 5% level, set $n = 264$ (corresponding to 22 years of monthly observations), generated 1,000 time series of length n from the DGP, evaluated Aït-Sahalia's test statistic for each time series, and compared it with the asymptotic critical value. Consistent with the results reported in Pritsker (1998), we found that Aït-Sahalia's test rejects the true null in over 50% of the samples. On the other hand, when we repeated the experiment with the LAE test (15) in place of Aït-Sahalia's test, the LAE test rejected the true null in 4.1% of the samples. Thus, for this particular problem, the size distortion is largely resolved by the LAE test.

5.2. *Power of the Test.* Next we investigated the power of the LAE test in finite samples. To provide context, we began by re-examining a second Monte Carlo experiment of Pritsker (1998), which analyzed the power of Aït-Sahalia's test. For the null hypothesis he took a Vasicek model of interest rates, while for the alternative he used the CIR model of Cox, Ingersoll and Ross (1985). He compared the size-adjusted power of Aït-Sahalia's test against a conditional moment-based specification test, and found that, after size adjustment, the power of

b	Z_t	CvM	AS	LAE	Cond m.
0.090495	$N(0,1)$	0.2487	0.3275	0.3662	1.0000
0.050000	$N(0,1)$	0.4637	0.5887	0.6101	1.0000
0.025000	$N(0,1)$	0.8475	0.9262	0.9375	0.9900
0.090495	t	0.7825	0.8125	0.8052	0.5672

TABLE 1
Rejection frequency with Vasicek null and CIR alternative (test size 0.05)

Aït-Sahalia's test for this null-alternative pair was considerably lower than that of the conditional moment test. He interpreted his findings as implying that Aït-Sahalia's test estimates the stationary density too imprecisely to have good power against this alternative (Pritsker, 1998, p. 462). We conducted a similar experiment to Pritsker, including results for the LAE test and that of the Cramér von Mises test as well. As described below, our results were not consistent with his interpretation.

As in Pritsker, the Vasicek null was paired with a discretized CIR alternative

$$X_{t+1} = X_t + \kappa(b - X_t)\delta + \sigma\sqrt{X_t\delta}Z_t \qquad \{Z_t\} \overset{\text{IID}}{\sim} N(0,1) \tag{21}$$

As in section 5.1, we set $\delta = 1/12$ and $n = 264$ for 22 years of monthly observations. Following Pritsker (1998, p. 460), our baseline parameter values were $\kappa = 0.89218$, $b = 0.090495$ and $\sigma = 0.180947$. For each test we calculated the rejection frequency for size $\alpha = 0.05$ over 1,000 replications. The results of this experiment are shown in the row 1 of table 1. Rows two and three report what happened when we we varied the equilibrium interest rate b from 9% to 5% and 2.5% respectively, while holding κ and σ fixed at the baseline values.

The columns CvM, AS, LAE and Cond m. correspond to the the Cramér von Mises test, a size-adjusted version of Aït-Sahalia's test, the LAE test (test (15)), and the conditional moment test used by Pritsker (1998, p. 462) respectively. As in Pritsker, for the conditional moment test we estimated the model parameters by maximum likelihood under the Vasicek null, and ran the regression $\partial\ell(X_{t+1}, X_t)/\partial\sigma = \beta_0 + \beta_1 X_t + u_{t+1}$, where $\ell(X_{t+1}, X_t)$ is the log likelihood of (X_{t+1}, X_t) under the null. We then conducted a two-sided test of $\beta_1 = 0$, which holds when the Vasicek null hypothesis is true.

In this experiment, the conditional moment test has higher power than the other three tests (table 1, rows 1–3). Pritsker interpreted the low power of Aït-Sahalia's test relative to the conditional moment test as due to the use of nonparametric kernel density estimators, which converge at a less than parametric rate. Our results suggest that the main causes lie elsewhere. Indeed, for this application, the Cramér von Mises and LAE tests are also dominated by the conditional moment test, despite the fact that these tests do not require estimation of smoothing parameters. A more likely explanation for the higher power of the conditional moment test is that this test concentrates a large amount of its power against the CIR alternative (since the

ρ	β	γ	CvM	AS	LAE	Cond m.
0.9	1.0	0.0012	0.271	0.375	0.466	0.077
0.9	1.0	0.0034	0.375	0.653	0.787	0.123
0.9	1.0	0.0056	0.375	0.825	0.925	0.146
0.9	1.0	0.0078	0.348	0.879	0.958	0.154
0.9	1.0	0.0100	0.359	0.885	0.974	0.159

TABLE 2
Rejection frequency with Vasicek null and RSw alternative (test size 0.05)

expected value of the score of the likelihood is linear under CIR). In fact, this test was chosen by Pritsker precisely because of its high power against the CIR alternative.

Because this particular conditional moment test concentrates a large amount of its power against the CIR alternative, the high power of the test shown in rows 1–3 might be fragile in practice, where the alternative is unknown. Even if a CIR alternative is suspected, unknown variations from the CIR alternative can reverse the results, with the conditional moment test having lower power than the other tests. Row 4 of table 1 illustrates this point. Here b returns to the baseline value of row 1, but the Gaussian shock in (21) is replaced by a t-distributed shock with 2.5 degrees of freedom. With this change, the LAE, AS and CvM tests have higher power than the conditional moment test.

To further reinforce this point, next we compare the four tests with another alternative: an AR(1) model with Gaussian shocks and regime switching coefficients. The regime switching alternative (RSw) has the form $X_{t+1} = \beta_t + \rho X_t + Z_{t+1}$, where the shocks are standard normal and $\{\beta_t\}$ follows a discrete Markov process. In particular, the process $\{\beta_t\}$ starts at $\beta_0 = \beta$ where β is a parameter, and then $\beta_{t+1} = \beta_t$ with probability $1 - \gamma$ and $\beta_{t+1} = -\beta_t$ with probability γ. Notice that when $\gamma = 0$, this process reduces to a linear Gaussian AR(1) model. Since the density kernel for the linear Gaussian AR(1) shares the same parametric form as that of the Vasicek density kernel, the null hypothesis is true when $\gamma = 0$. Larger values of γ indicate greater divergence from the null. For the other parameters, we set $\rho = 0.9$, $\beta = 1$ and $n = 500$.

The rejection frequencies for each test over 1,000 replications are shown in table 2 for different values of γ. Other than the alternative, the details of the calculations are the same as the previous section. Power curves for the LAE and conditional moment tests are shown in figure 2. As shown in the table and figure, for the RSw alternative, the conditional moment test was dominated by the other three tests. In this application, the LAE test has uniformly highest power over the values of γ in table 2.

5.3. *Nonstationary Alternatives.* In section 4.3 we made the point that the conditions of theorem 4.3 are sufficient but not necessary for consistency, and, in particular, that for some choices of H_0 and nonstationary alternatives, the test is likely to reject with high probability when the sample size is large. To illustrate, let H_0 be that $\{X_t\}$ is generated by the Vasicek kernel with

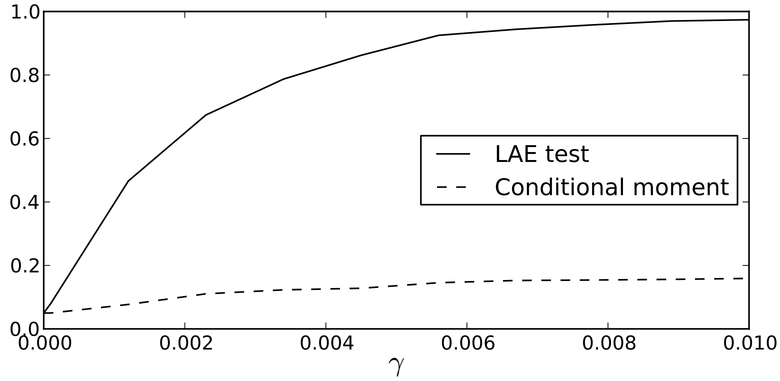

Fig 2: Rejection frequency, regime switching alternative

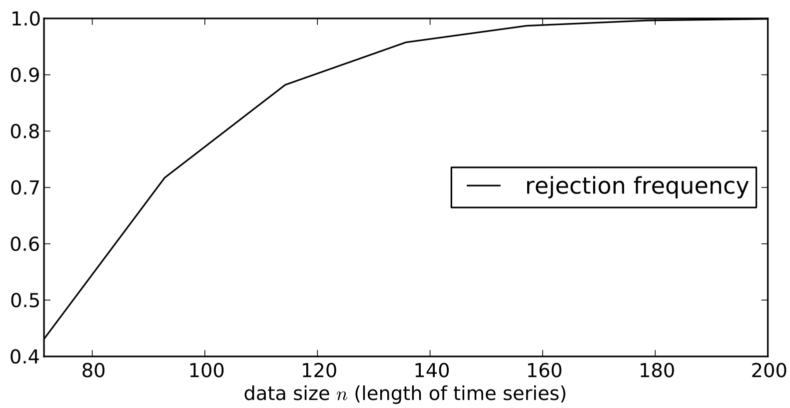

Fig 3: Rejection frequency, nonstationary alternative

baseline parameters defined in section 5.1, and consider a random walk alternative. (In particular, we take the same model as H_0 but with $\kappa = 0$.) The rejection probabilities for data sizes between 50 and 200 are shown in figure 3. By $n = 200$ the rejection probability is one. (Rejection probabilities were calculated by averaging over 1,000 observations.)

5.4. *Empirical Application.* In recent years, nonlinear business cycle models have been studied by many authors (cf., e.g., Hamilton, 1989; Pesaran and Potter, 1997; Harding and Pagan, 2002). In this section, the LAE test is used to test business cycle data for departures from a linear Gaussian AR(1) null. The data $\{y_t\}$ consists of quarterly percentage growth rates in seasonally adjusted GDP for the US and Canada over March 1950–September 2011, using data from the IMF's *International Financial Statistics*. For comparison, results for the LST test of Luukkonen,

	LAE	LST1	LST2	LST3	LST4
Canada	0.113	0.465	0.717	0.499	0.513
US	0.013	0.672	0.688	0.800	0.867

TABLE 3
p-values for LAE and LST tests applied to rates of change in GDP

Saikkonen and Terasvirta (1988) are also presented. The LST test is based on estimating an AR(1) model for y_t in the first stage, and then regressing the residuals from the first stage on a polynomial in y_{t-1}. The three versions of the test we present correspond to a quadratic polynomial (LST 1), a cubic (LST 2), a quartic (LST 3), and a quartic without the cubic term (LST 4). The test statistic is nR^2, where n is the sample size and R^2 is the coefficient of determination from the second stage regression.

The p-values resulting from application of the LAE and LST tests to the data are presented in table 3. In the Canadian data, no tests reject the Gaussian linear null hypothesis at either the 5% or 10% level. In the US data, none of the LST tests reject the null hypothesis at either the 5% or 10% level, while the LAE test rejects the null hypothesis at both 5% and 10%. A likely interpretation is that the departures from the Gaussian linear AR(1) model in the US data are other than those featuring in the LST test. For example, the LAE test can be sensitive to a non-Gaussian error term, as shown in table 1. On the other hand, the LST test is directed only towards nonlinearities in the mean.

6. Conclusion. In this paper we proposed a natural goodness of fit test for ergodic Markov processes. Although the test is based on comparison of densities, the test statistic contains no smoothing parameters, and the test has nontrivial power against $1/\sqrt{n}$ local alternatives. One idea not discussed above is the possibility of using weighting functions to obtain additional power against certain alternatives. This can be done by adjusting the measure μ that defines the integral in the test statistic. The ability to apply different weighting functions should add to the flexibility of the test. Further investigation of this topic is left to future research.

7. Proofs. In the proofs we will use the following facts without comment: If X, X_n and Y_n are L_2 random elements with $X_n \xrightarrow{\mathscr{D}} X$ in L_2 and $\|X_n - Y_n\| = o_P(1)$ in $\mathbb{R}$, then $Y_n \xrightarrow{\mathscr{D}} X$ (cf., e.g., Dudley, 2002, lemma 11.9.4). If X_n is an L_2-valued random variable, then the statement $X_n = o_P(1)$ means that $\|X_n\| = o_P(1)$ in $\mathbb{R}$. If $\alpha_n = o_P(1)$ in $\mathbb{R}$ and $f \in L_2$, then $X_n := \alpha_n f$ is an L_2-valued random variable with $X_n = o_P(1)$. The statement that $G \sim N(m, C)$ on L_2 is equivalent to the statement $\mathbf{E} \exp(i\langle h, G\rangle) = \exp\{i\langle h, m\rangle - \langle h, Ch\rangle/2\}$ for all $h \in L_2$, where i is the imaginary unit (cf., e.g., Parthasarathy, 1967, theorem 6.4.), yielding the characterization

$$G \sim N(m, C) \text{ on } L_2 \iff \langle G, h\rangle \sim N(\langle h, m\rangle, \langle h, Ch\rangle) \text{ on } \mathbb{R} \text{ for all } h \in L_2 \tag{22}$$

Hence, the distribution of G is defined by the values $\langle h, m\rangle$ and $\langle h, Ch\rangle$ over $h \in L_2$.

In what follows, we make repeated use of the following Markov Hilbert space central limit theorem, which is a simple corollary of Stachurski (2012, theorem 3.1). In the statement of the theorem, $\{X_t\}$ is a stationary Markov process on $\mathbb{X}$, the function $F_0 \colon \mathbb{X} \to L_2$ is Borel measurable, and $F := F_0 - \mathcal{E} F_0(X_t)$.

THEOREM 7.1. *If $\{X_t\}$ is geometrically ergodic and, for the function V in (4), there exists nonnegative constants c_0, c_1 and γ such that $\gamma < 1$ and $\|F_0(x)\|^2 \leq c_0 + c_1 V(x)^\gamma$ for all $x \in S$, then $n^{-1/2} \sum_{t=1}^n F(X_t)$ converges to a centered Gaussian on L_2, with covariance operator C satisfying*

$$\langle h, Ch \rangle = \mathbf{E} \langle F(X_1), h \rangle^2 + 2 \sum_{t=2}^{\infty} \mathbf{E} \langle F(X_1), h \rangle \langle F(X_t), h \rangle \tag{23}$$

The following results will also be needed in the proofs below:

LEMMA 7.1. *Let p be a density kernel, and let ψ be its stationary density. If p is V-mixing, then $\psi \in L_2$, $p(x, \cdot) \in L_2$ and $\bar{p}(x, \cdot) \in L_2$ for all $x \in \mathbb{X}$. Moreover, if X is any $\mathbb{X}$-valued random variable, then $p(X, \cdot)$ is an L_2-valued random variable.*

PROOF. Evidently (5) implies that $p(x, \cdot) \in L_2$ for each $x \in \mathbb{X}$. Regarding the claim that $\psi \in L_2$, the definition of stationarity and Jensen's inequality give

$$\int \psi(y)^2 dy = \int \left[\int p(x,y) \psi(x) dx \right]^2 dy \leq \int \int p(x,y)^2 \psi(x) dx \, dy$$

From (5) and (4), we then have

$$\int \psi(y)^2 dy \leq \int \int p(x,y)^2 dy \psi(x) dx \leq c_0 + c_1 \int V(x)^\gamma \psi(x) dx < \infty$$

Since $\gamma < 1$ we can apply Jensen's inequality to obtain

$$\int V(x)^\gamma \psi(x) dx \leq \left[\int V(x) \psi(x) dx \right]^\gamma$$

and this expression is finite by (4). We conclude that $\psi \in L_2$ as claimed. Moreover, we can now see that $\bar{p}(x, \cdot) \in L_2$ for any $x \in \mathbb{X}$, because

$$\|\bar{p}(x, \cdot)\| = \|p(x, \cdot) - \psi(\cdot)\| \leq \|p(x, \cdot)\| + \|\psi\|$$

To show that $p(X, \cdot)$ is an L_2-valued random variable, we need to prove that $\Omega \ni \omega \mapsto p(X(\omega), \cdot) \in L_2$ is also measurable, in the sense that preimages of Borel subsets of L_2 are measurable in Ω. Since L_2 is separable, it follows from the Pettis measurability theorem that any mapping $\Omega \ni \omega \mapsto g(\omega) \in L_2$ is measurable whenever $\Omega \ni \omega \mapsto \langle g(\omega), h \rangle \in \mathbb{R}$ is measurable for each $h \in L_2$. Using this fact, the measurability of $\omega \mapsto p(X(\omega), \cdot)$ is easily verified. This concludes the proof of lemma 7.1. □

LEMMA 7.2. *If p is V-mixing and $\{X_t\}$ is p-Markov, then $\mathcal{E}\bar{p}(X_t, \cdot) = 0$ for all t.*

PROOF. Fixing t and letting $X = X_t$, this amounts to the claim that $\mathbf{E} \int \bar{p}(X, y) h(y) dy = 0$ for any $h \in L_2$. To see this, fix $h \in L_2$. Note that for each $y \in \mathbb{X}$ we have

$$\mathbf{E}\bar{p}(X, y) = \int p(x, y)\psi(x)dx - \psi(y) = \psi(y) - \psi(y) = 0 \tag{24}$$

As a consequence, $\mathbf{E} \int \bar{p}(X, y)h(y)dy = \int \mathbf{E}\bar{p}(X, y)h(y)dy = 0$ whenever Fubini's theorem is valid. Fubini's theorem is valid whenever $\mathbf{E} \int |\bar{p}(X, y)h(y)|dy < \infty$. To check this, observe that, by the Cauchy-Schwartz and triangle inequalities, $\int |\bar{p}(x, y)h(y)|dy \leq \|\bar{p}(x, \cdot)\|\|h\| \leq (\|p(x, \cdot)\| + \|\psi\|)\|h\|$. Hence it suffices to show that $\mathbf{E}\|p(X, \cdot)\|^2 = \int \int p(x, y)^2 dy \psi(x)dx < \infty$. This claim was verified as part of the proof of lemma 7.1. □

PROOF OF THEOREM 2.1. Let p be V-mixing and let $\{X_t\}_{t=1}^n$ be p-Markov. Define $F_0(X_t) := p(X_t, \cdot)$ and let $F(X_t) := \bar{p}(X_t, \cdot) = p(X_t, \cdot) - \psi$. We saw in lemmas 7.1 and 7.2 that $F_0(X_t)$ is an L_2-valued random variable satisfying $\mathcal{E}F_0(X_t) = \psi$. Moreover, $\|F_0(x)\|^2 \leq c_0 + c_1 V(x)^\gamma$ for all $x \in \mathbb{X}$ by (5). Applying theorem 7.1, we then have the weak convergence $n^{-1/2} \sum_{t=1}^n F(X_t) \xrightarrow{\mathscr{D}} N(0, C)$, where C is defined in (23). It is straightforward to check that this expression and (7) are identical, and hence $C = \Lambda$. In summary, $n^{-1/2} \sum_{t=1}^n \bar{p}(X_t, \cdot) \xrightarrow{\mathscr{D}} N(0, \Lambda)$ as claimed. □

Turning to the proof of theorem 4.1, we begin by defining Σ_θ. Given $\theta \in \Theta$ and p_θ-Markov sequence $\{X_t\}$, we let Σ_θ be the operator defined by

$$\langle h, \Sigma_\theta h \rangle = \langle h, \Lambda_\theta h \rangle + 2\mathbf{E}P_1 Q_1 + \mathbf{E}Q_1^2 + 2\sum_{t=2}^{\infty} \mathbf{E}\left\{P_1 Q_t + Q_1 P_t + Q_1 Q_t\right\} \tag{25}$$

for $P_j := \int \bar{p}(\theta, X_j, y)h(y)\, dy$ and $Q_j := \int \mathbf{E}\{D\bar{p}(\theta, X_1, y)\}^\top g(X_j, \ldots, X_{j+r})h(y)\, dy$. Here Λ_θ is the operator (7) corresponding to p_θ.

PROOF OF THEOREM 4.1. Assume the conditions of theorem 4.1. Fix $\theta \in \Theta$ and let $\{X_t\}$ be p_θ-Markov. We need to prove the statement

$$\hat{Y}_n := n^{-1/2} \sum_{t=1}^n \bar{p}(\hat{\theta}_n, X_t, \cdot) \xrightarrow{\mathscr{D}} N(0, \Sigma_\theta) \tag{26}$$

in L_2. Throughout the proof, we use the notation $\rho(y) := \mathbf{E}D\bar{p}(\theta, X_t, y)$ and $\rho_m(y) := \mathbf{E}D_m\bar{p}(\theta, X_t, y)$. By differentiability (assumption 4.2) we can expand p around θ to get

$$\bar{p}(\hat{\theta}_n, x, y) = \bar{p}(\theta, x, y) + D\bar{p}(\theta, x, y)^\top(\hat{\theta}_n - \theta) + R(\hat{\theta}_n, x, y) \tag{27}$$

where R is the remainder term and $\top$ indicates inner product in $\mathbb{R}^M$. We then have

$$\hat{Y}_n(y) = n^{-1/2} \sum_{t=1}^n \left\{\bar{p}(\theta, X_t, y) + D\bar{p}(\theta, X_t, y)^\top(\hat{\theta}_n - \theta) + R(\hat{\theta}_n, X_t, y)\right\}$$

Adding and subtracting $\rho(y)^\top(\hat{\theta}_n - \theta)$, we can write this last expression as

(28) $$\hat{Y}_n(y) = n^{-1/2} \sum_{t=1}^{n} \left\{ \bar{p}(\theta, X_t, y) + \rho(y)^\top(\hat{\theta}_n - \theta) \right\} + I_n(y) + J_n(y)$$

where $I_n := n^{-1/2} \sum_{t=1}^n [D\bar{p}(\theta, X_t, \cdot) - \rho]^\top (\hat{\theta}_n - \theta)$ and $J_n := n^{-1/2} \sum_{t=1}^n R(\hat{\theta}_n, X_t, \cdot)$. As a first step of the proof, we show that $I_n = J_n = o_P(1)$ in L_2. Beginning with I_n, observe that

(29) $$\|I_n\| = \sum_{m=1}^{M} |\hat{\theta}_n^m - \theta^m| \left\| n^{-1/2} \sum_{t=1}^{n} \{D_m \bar{p}(\theta, X_t, \cdot) - \rho_m\} \right\|$$

Fix $m \in \{1, \ldots, M\}$. An application of the definition of $\mathcal{E}$ verifies that $\mathcal{E} D_m \bar{p}(\theta, X_t, \cdot) = \rho_m$. Moreover, assumption 4.2 gives

$$\|D_m \bar{p}(\theta, x, \cdot)\|^2 = \int D_m \bar{p}(\theta, x, y)^2 dy \leq V_\theta(x)^{1/2}$$

As a result, theorem 7.1 applies, and hence $n^{-1/2} \sum_{t=1}^n \{D_m \bar{p}(\theta, X_t, \cdot) - \rho_m\}$ converges in distribution to a centered Gaussian in L_2. Applying the continuous mapping theorem, the norm of this random function also converges in distribution, and hence

$$\left\| n^{-1/2} \sum_{t=1}^{n} \{D_m \bar{p}(\theta, X_t, \cdot) - \rho_m\} \right\| = O_P(1)$$

Since $|\hat{\theta}_n^m - \theta^m| = o_P(1)$ by assumption, we then have

$$|\hat{\theta}_n^m - \theta^m| \left\| n^{-1/2} \sum_{t=1}^{n} \{D_m \bar{p}(\theta, X_t, \cdot) - \rho_m\} \right\| = o_P(1) O_P(1) = o_P(1)$$

for each $m \in \{1, \ldots, M\}$. Returning to (29) we see that $I_n = o_P(1)$ as claimed.

Turning to the case of J_n, we claim that

(30) $$\|J_n\| = \left\| n^{-1/2} \sum_{t=1}^{n} R(\hat{\theta}_n, X_t, \cdot) \right\| = o_p(1)$$

Using the mean value theorem, we can write

$$R(\hat{\theta}_n, X_t, y) = \{D\bar{p}(\tilde{\theta}, X_t, y) - D\bar{p}(\theta, X_t, y)\}^\top (\hat{\theta}_n - \theta)$$

where $\tilde{\theta}$ lies on the line segment between θ and $\hat{\theta}_n$. It follows that

$$n^{-1/2} \sum_{t=1}^{n} R(\hat{\theta}_n, X_t, y) = \left[\frac{1}{n} \sum_{t=1}^{n} \{D\bar{p}(\tilde{\theta}, X_t, y) - D\bar{p}(\theta, X_t, y)\} \right]^\top n^{1/2} (\hat{\theta}_n - \theta)$$

Applying the Cauchy-Schwartz inequality in $\mathbb{R}^M$, we obtain

(31) $$\left| n^{-1/2} \sum_{t=1}^{n} R(\hat{\theta}_n, X_t, y) \right| \leq H_n(y)\, n^{1/2} \|\hat{\theta}_n - \theta\|_E$$

where $\|\cdot\|_E$ is the norm in $\mathbb{R}^M$, and

$$H_n(y) := \left\| \frac{1}{n} \sum_{t=1}^{n} \{ D\bar{p}(\tilde{\theta}, X_t, y) - D\bar{p}(\theta, X_t, y) \} \right\|_E$$

From (31) we obtain the L_2 norm inequality

$$\left\| n^{-1/2} \sum_{t=1}^{n} R(\hat{\theta}_n, X_t, \cdot) \right\| \leq \|H_n\| \cdot O_P(1)$$

Hence, to establish (30), it suffices to prove that $\|H_n\| = o_P(1)$. By the definition of H_n and assumption 4.3, we have

$$\|H_n\| \leq \frac{1}{n} \sum_{t=1}^{n} \left[\int \|D\bar{p}(\tilde{\theta}, X_t, y) - D\bar{p}(\theta, X_t, y)\|_E^2 \, dy \right]^{1/2} \leq \|\tilde{\theta} - \theta\|_E^{\alpha} \frac{1}{n} \sum_{t=1}^{n} \left[\int K_2(X_t, y)^2 \, dy \right]^{1/2}$$

By assumption 4.1, $\|\hat{\theta}_n - \theta\|_E^{\alpha} = o_P(1)$. Moreover, by Jensen's inequality,

$$\mathbf{E} \left[\int K_2(X_t, y)^2 \, dy \right]^{1/2} \leq \left[\mathbf{E} \int K_2(X_t, y)^2 \, dy \right]^{1/2} = \left[\int \int K_2(x, y)^2 \, dy \, \psi_\theta(x) dx \right]^{1/2}$$

This expression is finite by assumption 4.3. Applying the scalar law of large numbers for ergodic Markov processes (e.g., Meyn and Tweedie, theorem 17.1.7), we have

$$\frac{1}{n} \sum_{t=1}^{n} \left[\int K_2(X_t, y)^2 \, dy \right]^{1/2} = O_P(1)$$

We conclude that $\|H_n\| \leq o_P(1) O_P(1) = o_P(1)$, and hence (30) is valid.

Returning now to (28), we have shown that the last two terms on the right-hand side are $o_P(1)$, while assumption 4.1 and simple manipulations show that the first term can be expressed as

$$n^{-1/2} \sum_{t=1}^{n} \left\{ \bar{p}(\theta, X_t, y) + \rho(y)^\top g_\theta(X_t, \ldots, X_{t+r}) \right\} + o_P(1)$$

Define $M_t := (X_t, \ldots, X_{t+r})$,

$$F_0(M_t) := p(\theta, X_t, \cdot) + \rho(\cdot)^\top g_\theta(M_t) \quad \text{and} \quad F(M_t) := \bar{p}(\theta, X_t, \cdot) + \rho(\cdot)^\top g_\theta(M_t)$$

We see that (26) will be established if we can show that

$$n^{-1/2} \sum_{t=1}^{n} F(M_t) := n^{-1/2} \sum_{t=1}^{n} \left\{ \bar{p}(\theta, X_t, \cdot) + \rho^\top g_\theta(M_t) \right\} \xrightarrow{\mathscr{D}} N(0, \Sigma_\theta) \tag{32}$$

We will use theorem 7.1. As a first step, we claim that $\mathcal{E} F_0(M_t) = \psi$. Since $F(M_t) = F_0(M_t) - \psi$, it suffices to show that $\mathcal{E} F(M_t) = 0$. To see that this is so, pick any $h \in L_2$. From the definition and Fubini's theorem we have

$$\begin{aligned} \mathbf{E}\langle F(M_t), h \rangle &= \mathbf{E} \int \bar{p}(\theta, X_t, y) h(y) dy + \mathbf{E} \int \rho(y)^\top g_\theta(M_t) h(y) dy \\ &= \int \mathbf{E}\bar{p}(\theta, X_t, y) h(y) dy + \int \rho(y)^\top \mathbf{E}[g_\theta(M_t)] h(y) dy \end{aligned}$$

Since $\{X_t\}$ is p_θ-Markov, both of these expectations are zero (by the definition of $\bar{p}$ and assumption 4.1 respectively), and hence $\mathcal{E}F(M_t) = 0$ as claimed.

Let $\hat{V}(x_0, \dots, x_r) := \sum_{k=0}^{r} V(x_k)$. It is shown in the technical supplement (Martin *et al.*, 2012) that $\{M_t\}$ is geometrically ergodic with weight function $\hat{V}$. In order to apply theorem 7.1, it remains to show that there exists constants c_0, c_1, γ with $\gamma < 1$ and

$$\|F_0(x_0, \dots, x_r)\|^2 \le c_0 + c_1 \hat{V}(x_0, \dots x_r)^\gamma \qquad \text{for all } (x_0, \dots, x_r) \in \mathbb{X}^{r+1} \tag{33}$$

To establish (33), observe first that

$$\begin{aligned}\|F_0(x_0, \dots, x_r)\|^2 &= \|p(\theta, x_0, \cdot) + \rho(\cdot)^\top g_\theta(x_0, \dots, x_r)\|^2 \\ &\le 2\int p(\theta, x_0, y)^2 dy + 2\int [\rho(y)^\top g_\theta(x_0, \dots, x_r)]^2 dy \\ &\le 2\int p(\theta, x_0, y)^2 dy + 2\int \|\rho(y)\|_E^2 dy \; \|g_\theta(x_0, \dots, x_r)\|_E^2\end{aligned}$$

Note that $\int \|\rho(y)\|_E^2 dy$ is finite. Indeed, using Jensen's inequality and assumption 4.2, we have

$$\begin{aligned}\|\rho(y)\|_E^2 dy &= \sum_{m=1}^{M} \int \{\mathbf{E} D_m \bar{p}(\theta, X_t, y)\}^2 dy \\ &\le \sum_{m=1}^{M} \mathbf{E} \int \{D_m \bar{p}(\theta, X_t, y)\}^2 dy \\ &\le \sum_{m=1}^{M} \mathbf{E}(V_\theta(X_t)^{1/2}) \le \sum_{m=1}^{M} (\mathbf{E} V_\theta(X_t))^{1/2}\end{aligned}$$

The final expression is finite by (4), and hence $\int \|\rho(y)\|_E^2 dy$ is finite as claimed. As a result, combining (5) and assumption 4.1, there are nonnegative constants c_0, a_1, a_2 and $\alpha < 1$ with

$$\begin{aligned}\|F_0(x_0, \dots, x_r)\|^2 &\le c_0 + a_1 V(x_0)^\alpha + a_2 \hat{V}(x_0, \dots x_k)^{2/(2+\delta)} \\ &\le c_0 + a_1 \hat{V}(x_0, \dots x_k)^\alpha + a_2 \hat{V}(x_0, \dots x_k)^{2/(2+\delta)}\end{aligned}$$

Setting $\gamma := \max\{\alpha, 2/(2+\delta)\}$ and $c_1 := \max\{a_1, a_2\}$ yields (33). The conditions of theorem 7.1 are now verified, and from that theorem we obtain $n^{-1/2} \sum_{t=1}^{n} F(M_t) \xrightarrow{\mathscr{D}} N(0, S)$ with

$$\langle h, Sh \rangle = \mathbf{E}\langle F(M_1), h\rangle^2 + 2\sum_{t=2}^{\infty} \mathbf{E}\langle F(M_1), h\rangle \langle F(M_t), h\rangle \tag{34}$$

for arbitrary $h \in L_2$. Thus (32) will be established if we can show that $\langle h, Sh\rangle = \langle h, \Sigma_\theta h\rangle$, which is to say that the right-hand side of (34) agrees with the right-hand side of (25). Observe that $\langle F(M_j), h\rangle = P_j + Q_j$, where P_j and Q_j are defined immediately after (25). As a result, we can write

$$\langle h, Sh\rangle = \mathbf{E}P_1^2 + 2\mathbf{E}P_1 Q_1 + \mathbf{E}Q_1^2 + 2\sum_{t=2}^{\infty} \mathbf{E}\left\{P_1 P_t + P_1 Q_t + Q_1 P_t + Q_1 Q_t\right\}$$

Since $\langle h, \Lambda_\theta h\rangle = \mathbf{E}P_1^2 + 2\sum_{t=2}^{\infty} \mathbf{E}P_1 P_t$ it follows that $\langle h, Sh\rangle = \langle h, \Sigma_\theta h\rangle$ as claimed. Hence (32) is valid, completing the proof of theorem 4.1. □

PROOF OF THEOREM 4.2. The claim in the theorem is that $\lim_{n\to\infty} \mathbf{P}\left\{\hat{T}_n \leq c^{\Sigma}_{\alpha}(\hat{\theta}_n)\right\} \geq 1-\alpha$ under H_0. If H_0 holds, then $\hat{T}_n \xrightarrow{\mathscr{D}} \sum_{\ell} \sigma_{\ell}(\theta_0) Z^2_{\ell}$ and $c^{\Sigma}_{\alpha}(\hat{\theta}_n) \xrightarrow{p} c^{\Sigma}_{\alpha}(\theta_0)$, where the first result is due to (13) and the the second is due to consistency of $\hat{\theta}_n$ and continuity of c^{Σ}_{α} at θ_0. Slutsky's theorem yields $\hat{T}_n - c^{\Sigma}_{\alpha}(\hat{\theta}_n) + c^{\Sigma}_{\alpha}(\theta_0) \xrightarrow{\mathscr{D}} \sum_{\ell} \sigma_{\ell}(\theta_0) Z^2_{\ell}$. As a result,

$$\lim_{n\to\infty} \mathbf{P}\{\hat{T}_n \leq c^{\Sigma}_{\alpha}(\hat{\theta}_n)\} = \lim_{n\to\infty} \mathbf{P}\{\hat{T}_n - c^{\Sigma}_{\alpha}(\hat{\theta}_n) + c^{\Sigma}_{\alpha}(\theta_0) \leq c^{\Sigma}_{\alpha}(\theta_0)\} = \mathbf{P}\left\{\sum_{\ell} \sigma_{\ell}(\theta_0) Z^2_{\ell} \leq c^{\Sigma}_{\alpha}(\theta_0)\right\}$$

By the definition of $c^{\Sigma}_{\alpha}(\theta_0)$ this probability is $1-\alpha$. $\square$

PROOF OF THEOREM 4.3. Assume the conditions of the theorem. The claim is that

$$\frac{\hat{T}_n}{n} = \left\| \frac{1}{n} \sum_{t=1}^{n} \bar{p}(\hat{\theta}_n, X_t, \cdot) \right\| \xrightarrow{p} \left\| \mathcal{E} \bar{p}(\theta_2, X_t, \cdot) \right\| \qquad (n \to \infty) \tag{35}$$

To see this, observe first that the distance between two terms in (35) is bounded above by

$$\left\| \frac{1}{n} \sum_{t=1}^{n} \bar{p}(\hat{\theta}_n, X_t, \cdot) - \mathcal{E} \bar{p}(\hat{\theta}_n, X_t, \cdot) \right\| + \left\| \mathcal{E} \bar{p}(\hat{\theta}_n, X_t, \cdot) - \mathcal{E} \bar{p}(\theta_2, X_t, \cdot) \right\| \tag{36}$$

So that (35) will be established if we can show that both of the terms in (36) are $o_P(1)$. We begin with the first term. In this term, $\bar{p}$ can be replaced with p because the stationary densities cancel. Thus, our aim is to show that

$$\left\| \frac{1}{n} \sum_{t=1}^{n} p(\hat{\theta}_n, X_t, \cdot) - \mathcal{E} p(\hat{\theta}_n, X_t, \cdot) \right\| = o_P(1) \tag{37}$$

The expression in (37) is bounded above by $(I) + (II) + (III)$ where

$$(I) := \left\| \frac{1}{n} \sum_{t=1}^{n} \{p(\hat{\theta}_n, X_t, \cdot) - p(\theta_2, X_t, \cdot)\} \right\|, \quad (II) := \left\| \frac{1}{n} \sum_{t=1}^{n} p(\theta_2, X_t, \cdot) - \mathcal{E} p(\theta_2, X_t, \cdot) \right\|$$

and $(III) := \left\| \mathcal{E} p(\theta_2, X_t, \cdot) - \mathcal{E} p(\hat{\theta}_n, X_t, \cdot) \right\|$. We claim that all of these terms converge to zero. To begin, consider first the term (I). By assumption 4.6, we have

$$|p(\hat{\theta}_n, X_t, y) - p(\theta_2, X_t, y)| \leq \eta(y) \, \|\hat{\theta}_n - \theta_2\|_E \qquad \text{for all } y \in \mathbb{X} \tag{38}$$

Taking the the L_2 norm of this expression we get

$$(I) \leq \frac{1}{n} \sum_{t=1}^{n} \left\| p(\hat{\theta}_n, X_t, \cdot) - p(\theta_2, X_t, \cdot) \right\| \leq \|\eta\| \cdot \|\hat{\theta}_n - \theta_2\|_E = o_P(1)$$

Turning to terms (II) and (III), the claim that (II) is $o_P(1)$ follows directly from assumption 4.4, provided that $\mathcal{E} p(\theta_2, X_t, \cdot)$ exists. This L_2 expectation exists whenever the scalar expectation of the norm of $p(\theta_2, X_t, \cdot)$ is finite. Finiteness of this scalar expectation is a direct consequence of assumption 4.4. Regarding (III), another application of assumption 4.6 gives

$$(III) \leq \mathbf{E} \left\| p(\theta_2, X_t, \cdot) - p(\hat{\theta}_n, X_t, \cdot) \right\| \leq \|\eta\| \, \mathbf{E} \, \|\hat{\theta}_n - \theta_2\|_E \to 0 \tag{39}$$

where the convergence uses $\|\hat{\theta}_n - \theta_2\|_E = o_P(1)$ and the fact that $\|\hat{\theta}_n - \theta_2\|_E^2$ is uniformly bounded as a result of the boundedness of Θ. We conclude that $(I) + (II) + (III) = o_P(1) + o_P(1) + o(1) = o_P(1)$ and hence (37) is valid.

Now we return to the second term in (36), which we claim converges to zero. Evidently

$$\left\|\mathcal{E}\bar{p}(\hat{\theta}_n, X_t, \cdot) - \mathcal{E}\bar{p}(\theta_2, X_t, \cdot)\right\| \leq \left\|\mathcal{E}p(\hat{\theta}_n, X_t, \cdot) - \mathcal{E}p(\theta_2, X_t, \cdot)\right\| + \left\|\mathcal{E}\psi(\hat{\theta}_n, \cdot) - \psi(\theta_2, \cdot)\right\|$$

The first term on the right-hand side was already shown to converge to zero in (39). Regarding the second term, in view of assumption 4.6, we have

$$\left\|\mathcal{E}\psi(\hat{\theta}_n, \cdot) - \psi(\theta_2, \cdot)\right\| \leq \mathbf{E}\left\|\psi(\hat{\theta}_n, \cdot) - \psi(\theta_2, \cdot)\right\| \leq \|\xi\|\,\mathbf{E}\left\|\hat{\theta}_n - \theta_2\right\|_E \to 0$$

where the convergence uses $\|\hat{\theta}_n - \theta_2\|_E = o_P(1)$ and the fact that $\|\hat{\theta}_n - \theta_2\|_E^2$ is bounded. This completes the proof of theorem 4.3. □

REFERENCES

[1] Aït-Sahalia, Y. (1996). Testing continuous-time models of the spot interest rate, *Review of Financial Studies,* 9 (2), 385–426.

[2] Bosq, D. (2000). *Linear Processes in Function Space,* Springer-Verlag.

[3] Chicheportiche, R. and J-P. Bouchaud (2011). Goodness of fit tests with dependent observations, mimeo, Ecole Centrale Paris.

[4] Cox, J.C., J.E. Ingersoll and S.A. Ross (1985). A theory of the term structure of interest rates, *Econometrica*, 53: 385–407.

[5] del Barrio, E., P. Deheuvels and S. van de Geer (2007). *Lectures on Empirical Processes: Theory and Statistical Applications,* European Mathematical Society, EMS Publishing House.

[6] Dudley, R. M. (2002). *Real Analysis and Probability,* Cambridge Studies in Advanced Mathematics No. 74, Cambridge University Press.

[7] Hamilton, J.D. (1989), A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica,* 57, 357–381.

[8] Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators, *Econometrica,* 50, 1029–1054.

[9] Harding A and A.R. Pagan (2002) Dissecting the cycle: A methodological investigation, *Journal of Monetary Economics,* 49, 365–381.

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

[11] Koul, H.L. and Stute, W. (1999). Nonparameteric model checks for time series. *Ann. Statist.* 27 204-236.

[12] Kristensen, D. (2007). Geometric ergodicity of a class of Markov chains with applications to time series models, mimeo, Department of Economics, Columbia University.

[13] Kristensen, D. (2011). Semi-nonparametric estimation and misspecification testing of diffusion models, *Journal of Econometrics,* 164, 382–403.

[14] Lindvall, T. (2002). *Lectures on the Coupling Method,* Dover Publications, Mineola N.Y.

[15] Martin, V. L., Y. Nishiyama and J. Stachurski (2012). Technical supplement to "Goodness of fit for Markov models: a density approach," available from `http://johnstachurski.net/papers/lae_test.html`

[16] Meyn, S. P. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*, Springer-Verlag: London.

[17] Neumann, M. and E. Paparoditis (2008). Goodness-of-fit tests for Markovian time series models: Central limit theory and bootstrap approximations, *Bernoulli*, 14 (1), 14–46.

[18] Parthasarathy, K. R. (1967). *Probability Measures on Metric Spaces*, American Mathematical Society.

[19] Pesaran, M.H. and S.M. Potter (1997). A floor and ceiling model of US output, *Journal of Economic Dynamics and Control*, 21, 661-695.

[20] Pritsker, M. (1998). Nonparametric density estimation and tests of continuous time interest rate models, *Review of Financial Studies*, 11 (3), 449–487.

[21] Stachurski, J. and V. L. Martin (2008). Computing the distributions of economic models via simulation, *Econometrica*, 76 (2), 443–450.

[22] Stachurski, J. (2012). A Hilbert space central limit theorem for geometrically ergodic Markov chains, mimeo, Australian National University.

DEPARTMENT OF ECONOMICS, THE UNIVERSITY OF MELBOURNE,
E-MAIL: vance@unimelb.edu.au

INSTITUTE OF ECONOMIC RESEARCH, KYOTO UNIVERSITY,
E-MAIL: nishiyama@kier.kyoto-u.ac.jp

RESEARCH SCHOOL OF ECONOMICS, AUSTRALIAN NATIONAL UNIVERSITY,
E-MAIL: john.stachurski@anu.edu.au