The Annals of Statistics
2011, Vol. 39, No. 2, 990–1011
DOI: 10.1214/10-AOS828

KERNEL ESTIMATORS OF ASYMPTOTIC VARIANCE FOR ADAPTIVE MARKOV CHAIN MONTE CARLO[1]

BY YVES F. ATCHADÉ

University of Michigan

We study the asymptotic behavior of kernel estimators of asymptotic variances (or long-run variances) for a class of adaptive Markov chains. The convergence is studied both in L^p and almost surely. The results also apply to Markov chains and improve on the existing literature by imposing weaker conditions. We illustrate the results with applications to the GARCH(1, 1) Markov model and to an adaptive MCMC algorithm for Bayesian logistic regression.

1. Introduction. Adaptive Markov chain Monte Carlo (adaptive MCMC) provides a flexible framework for optimizing MCMC samplers on the fly (see, e.g., [3, 8, 27] and the reference therein). If π is the probability measure of interest, then these adaptive MCMC samplers generate random processes $\{X_n, n \geq 0\}$ that typically are not Markov, but they nevertheless satisfy a law of large numbers and the empirical average $n^{-1}\sum_{k=1}^{n} h(X_k)$ provides a consistent estimate of the integral $\pi(h) \stackrel{\text{def}}{=} \mathbb{E}(h(X))$, $X \sim \pi$. A measure of uncertainty in approximating $\pi(h)$ by the random variable $n^{-1}\sum_{k=1}^{n} h(X_k)$ is given by the variance $\operatorname{Var}(n^{-1/2}\sum_{k=1}^{n} h(X_k))$. In particular, the asymptotic variance $\sigma^2(h) \stackrel{\text{def}}{=} \lim_{n\to\infty} \operatorname{Var}(n^{-1/2}\sum_{k=1}^{n} h(X_k))$ (also known as the *long-run variance*) plays a fundamental role in assessing the performances of Monte Carlo simulations. But the problem of estimating asymptotic variances for adaptive MCMC samplers has not been addressed in the literature.

We study kernel estimators of asymptotic variances for a general class of adaptive Markov chains. These adaptive Markov chains (the precise definition is given in Section 2 below), which include Markov chains, constitute a theoretical framework for analyzing adaptive MCMC algorithms. More precisely, if $\{X_n, n \geq 0\}$ is an adaptive Markov chain and $h : \mathsf{X} \to \mathbb{R}$ a function of interest, then we consider estimators of the form

$$\Gamma_n^2(h) = \sum_{k=-n}^{n} w(kb)\gamma_n(k),$$

Received November 2009; revised April 2010.
[1]Supported in part by NSF Grant DMS-09-06631.
MSC2010 subject classifications. 60J10, 60C05.
Key words and phrases. Adaptive Markov chain Monte Carlo, kernel estimators of asymptotic variance.

where $\gamma_n(k) = \gamma_n(k; h)$ is the kth order sample autocovariance of $\{h(X_n), n \geq 0\}$, $w : \mathbb{R} \to \mathbb{R}$ is a kernel with support $[-1, 1]$ and $b = b_n$ is the bandwidth. These are well-known methods pioneered by M. S. Bartlett, M. Rosenblatt, E. Parzen and others (see, e.g., [26] for more details). But, with a few notable exceptions in the econometrics literature (see references below), these estimators have mostly been studied with the assumption of stationarity. Thus, more broadly, this paper contributes to the literature on the behavior of kernel estimators of asymptotic variances for ergodic nonstationary processes.

It turns out that, in general, the asymptotic variance $\sigma^2(h)$ does not characterize the limiting distribution of $n^{-1/2}\sum_{k=1}^n (h(X_k) - \pi(h))$ as, for example, with ergodic Markov chains. For adaptive Markov chains, we show that $n^{-1/2}\sum_{k=1}^n (h(X_k) - \pi(h))$ converges weakly to a mixture of normal distributions of the form $\sqrt{\Gamma^2(h)}Z$ for some mixing random variable $\Gamma^2(h)$, where Z is a standard normal random variable independent of $\Gamma^2(h)$. Under a geometric drift stability condition on the adaptive Markov chain and some verifiable conditions on the kernel w and the bandwidth b_n, we prove that the kernel estimator $\Gamma_n^2(h)$ converges to $\Gamma^2(h)$ in L^p-norm, $p > 1$, and almost surely. For Markov chains, $\Gamma^2(h)$ coincides with $\sigma^2(h)$, the asymptotic variance of h. Another important special case where we have $\Gamma^2(h) = \sigma^2(h)$ is the one where the adaptation parameter converges to a deterministic limit as, for instance, with the adaptive Metropolis algorithm of [17]. The general case where $\Gamma^2(h)$ is random poses some new difficulties to Monte Carlo error assessment in adaptive MCMC that we discuss in Section 4.3.

We derive the rate of convergence for $\Gamma_n^2(h)$, which suggests selecting the bandwidth to be $b_n \propto n^{-(2/3)(1-0.5\vee(1/p))}$. When $p = 2$ is admissible, we obtain the bandwidth $b_n \propto n^{-1/3}$, as in [16].

The problem of estimating asymptotic variances is well known in MCMC and Monte Carlo simulation in general. Besides the estimator described above, several other methods have been proposed, including batch means, overlapping batch means and regenerative simulation ([12, 13, 16, 24]). For the asymptotics of kernel estimators, the important work of [16] proves the L^2-consistency and strong consistency of kernel estimators for Markov chains under the assumption of geometric ergodicity and $\mathbb{E}(|h(X)|^{4+\varepsilon}) < \infty$, $X \sim \pi$, for some $\varepsilon > 0$. We weaken these moment conditions to $\mathbb{E}(|h(X)|^{2+\varepsilon}) < \infty$.

Estimating asymptotic variances is also a well-known problem in econometrics and time series modeling. For example, if $\hat{\beta}_n$ is the ordinary least-squares estimator of β in the simple linear model $y_i = \alpha + \beta x_i + u_i, i = 1, \ldots, n$, where $\{u_k, k \geq 1\}$ is a dependent noise process, then, under some mild conditions on the sequence $\{x_i\}$ and on the noise process, $\sqrt{n}(\hat{\beta}_n - \beta)$ converges weakly to a normal distribution $\mathcal{N}(0, \sigma^2/c^2)$, where

$$\sigma^2 = \lim_{n\to\infty} \operatorname{Var}\left(n^{-1/2}\sum_{k=1}^n u_k\right), \qquad c^2 = \lim_{n\to\infty} n^{-1}\sum_{k=1}^n (x_i - \bar{x}_n)^2, \qquad \bar{x}_n = n^{-1}\sum_{k=1}^n x_k.$$

Therefore, a valid inference on β requires the estimation of the asymptotic variance σ^2. The multivariate version of this problem involves estimating the so-called *heteroskedasticity and autocorrelation* (HAC) matrices. Several authors have studied the kernel estimation of HAC matrices and attention has been paid to nonstationarity under various mixing assumptions or mixingale-type assumptions ([1, 14, 15, 19]). But these results require mixing conditions that do not hold in the present setup.

On a more technical note, the proof of our main results (Theorems 4.1–4.3) is based on a martingale approximation approach adapted from [29]. The crux of the argument consists in approximating the periodogram of the adaptive Markov chain by a quadratic form of a martingale difference process which is then treated as a martingale array. As part of the proof, we develop a strong law of large numbers for martingale arrays which may also be of some independent interest. The approach taken here thus differs from the almost sure strong approximation approach taken in [13, 16].

The paper is organized as follows. In Section 2, we define the class of adaptive Markov chains that will be studied. In Section 3, we give a general central limit theorem for adaptive Markov chains that sets the stage to better understand the limiting behavior of the kernel estimator $\Gamma_n^2(h)$. In Section 4, we state the assumptions and the main results of the paper. We also discuss some practical implications of these theoretical results. The proofs are postponed to Section 6 and to the supplementary paper [5]. Section 5 presents applications to generalized autoregressive conditional heteroscedastic (GARCH) processes and to a Bayesian analysis of logistic regression.

We end this introduction with some general notation that will be used throughout the paper. For a Markov kernel Q on a measurable space $(\mathcal{Y}, \mathcal{A})$, say, we denote by Q^n, $n \geq 0$, its nth iterate. Any such Markov kernel Q acts both on bounded measurable functions f and on σ-finite measures μ, as in $Qf(\cdot) \stackrel{\text{def}}{=} \int Q(\cdot, dy) f(y)$ and $\mu Q(\cdot) \stackrel{\text{def}}{=} \int \mu(dx) Q(x, \cdot)$. If $W : \mathcal{Y} \to [1, +\infty)$ is a function, then the W-norm of a function $f : \mathcal{Y} \to \mathbb{R}$ is defined as $|f|_W \stackrel{\text{def}}{=} \sup_{\mathcal{Y}} |f|/W$. The set of measurable functions $f : \mathcal{Y} \to \mathbb{R}$ with finite W-norm is denoted by $\mathcal{L}_W$. Similarly, if μ is a signed measure on $(\mathcal{Y}, \mathcal{A})$, then the W-norm of μ is defined as $\|\mu\|_W \stackrel{\text{def}}{=} \sup_{\{g, |g|_W \leq 1\}} |\mu(g)|$, where $\mu(g) \stackrel{\text{def}}{=} \int g(y) \mu(dy)$. If ν is a σ-finite measure on $(\mathcal{Y}, \mathcal{A})$ and $q \geq 1$, we denote by $L^q(\nu)$ the space of all measurable functions $f : (\mathcal{Y}, \mathcal{A}) \to \mathbb{R}$ such that $\nu(|f|^q) < \infty$. Finally, for $a, b \in \mathbb{R}$, we define $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$.

2. Adaptive Markov chains. Let $(\mathsf{X}, \mathcal{X})$ be a measure state space measure space endowed with a countably generated σ-field $\mathcal{X}$. Let $(\Theta, \mathcal{B}(\Theta))$ be a measure space. In practice, we will take Θ to be a compact subspace of $\mathbb{R}^q$, the q-dimensional Euclidean space. Let $\{P_\theta, \theta \in \Theta\}$ be a family of Markov transition

kernels on $(\mathsf{X}, \mathcal{X})$ such that for any $(x, A) \in \mathsf{X} \times \mathcal{X}$, $\theta \mapsto P_\theta(x, A)$ is measurable. Let π be a probability measure on $(\mathsf{X}, \mathcal{X})$. We assume that for each $\theta \in \Theta$, P_θ admits π as its invariant distribution.

The stochastic processes of interest in this work are defined as follows. Let $\Omega = (\mathsf{X} \times \Theta)^\infty$ be the product space equipped with its product σ-algebra $\mathcal{F}$ and let $\bar{\mu}$ be a probability measure on $(\mathsf{X} \times \Theta, \mathcal{X} \times \mathcal{B}(\Theta))$. Let $\mathbb{P}_{\bar{\mu}}$ be the probability measure on $(\Omega, \mathcal{F})$ with associated expectation operator $\mathbb{E}_{\bar{\mu}}$, associated process $\{(X_n, \theta_n), n \geq 0\}$ and associated natural filtration $\{\mathcal{F}_n, n \geq 0\}$, with the following properties: $(X_0, \theta_0) \sim \bar{\mu}$ and, for each $n \geq 0$ and any nonnegative measurable function $f : \mathsf{X} \to \mathbb{R}$,

$$\mathbb{E}_{\bar{\mu}}(f(X_{n+1})|\mathcal{F}_n) = P_{\theta_n} f(X_n) = \int P_{\theta_n}(X_n, dy) f(y), \qquad \mathbb{P}_{\bar{\mu}}\text{-a.s.} \tag{2.1}$$

We call the X-marginal process $\{X_n, n \geq 0\}$ an *adaptive Markov chain*. In this definition, we have left the adaptation dynamics (i.e., the conditional distribution of θ_{n+1} given $\mathcal{F}_n$ and X_{n+1}) unspecified. This can be done in many different ways (see, e.g., [27]). But it is well known, as we will see later, that the adaptation dynamics needs to be *diminishing* in order for the adaptive Markov chain to maintain π as its limiting distribution.

The simplest example of an adaptive Markov chain is the case where $\theta_n \equiv \bar{\theta} \in \Theta$ for all $n \geq 0$. Then $\{X_n, n \geq 0\}$ is a Markov chain with transition kernel $P_{\bar{\theta}}$. In other words, our analysis also applies to Markov chains and, in particular, to Markov chain Monte Carlo.

EXAMPLE 2.1. To illustrate the definitions and, later, the results, we present a version of the adaptive Metropolis algorithm of [17]. We take $\mathsf{X} = \mathbb{R}^d$ equipped with its Euclidean norm and inner product, denoted by $|\cdot|$ and $\langle \cdot, \cdot \rangle$, respectively. Let π be a positive, possibly unnormalized, density (with respect to the Lebesgue measure). We construct the parameter space Θ as follows. We equip the set $\mathcal{M}_+$ of all d-dimensional symmetric positive semidefinite matrices with the Frobenius norm $|A| \stackrel{\text{def}}{=} \sqrt{\text{Tr}(A^T A)}$ and inner product $\langle A, B \rangle = \text{Tr}(A^T B)$. For $r > 0$, let $\Theta_+(r)$ be the compact subset of elements $A \in \mathcal{M}_+$ such that $|A| \leq r$. Let $\Theta_\mu(r)$ be the ball centered at 0 and with radius r in $\mathbb{R}^d$. We then define $\Theta \stackrel{\text{def}}{=} \Theta_\mu(r_1) \times \Theta_+(r_2)$ for some constants $r_1, r_2 > 0$.

We introduce the functions $\Pi_\mu : \mathbb{R}^d \to \Theta_\mu(r_1)$ and $\Pi_+ : \mathcal{M}_+ \to \Theta_+(r_2)$, defined as follows. For $v \in \Theta_\mu(r_1)$, $\Pi_\mu(v) = v$ and for $v \notin \Theta_\mu(r_1)$, $\Pi_\mu(v) = \frac{M}{|v|} v$. Similarly, for $\Sigma \in \Theta_+(r_2)$, $\Pi_+(\Sigma) = \Sigma$ and for $\Sigma \notin \Theta_+(r_2)$, $\Pi_+(\Sigma) = \frac{M}{|\Sigma|}\Sigma$.

For $\theta = (\mu, \Sigma) \in \Theta$, let P_θ be the transition kernel of the random walk Metropolis (RWM) algorithm with proposal kernel $\mathcal{N}(x, \frac{2.38^2}{d}\Sigma + \varepsilon I_d)$ and target distribution π. The adaptive Metropolis algorithm works as follows.

ALGORITHM 2.1. *Initialization*: Choose $X_0 \in \mathbb{R}^d$, $(\mu_0, \Sigma_0) \in \Theta$. Let $\{\gamma_n\}$ be a sequence of positive numbers (we use $\gamma_n = n^{-0.7}$ in the simulations).

Iteration: Given (X_n, μ_n, Σ_n):

(1) generate $Y_{n+1} \sim \mathcal{N}(X_n, \frac{2.38^2}{d}\Sigma_n + \varepsilon I_d)$; with probability $\alpha_{n+1} = \alpha(X_n, Y_{n+1})$, set $X_{n+1} = Y_{n+1}$ and with probability $1 - \alpha_{n+1}$, set $X_{n+1} = X_n$;

(2) set

$$\mu_{n+1} = \Pi_\mu\big(\mu_n + (n+1)^{-1}(X_{n+1} - \mu_n)\big), \tag{2.2}$$

$$\Sigma_{n+1} = \Pi_+\big(\Sigma_n + (n+1)^{-1}\big((X_{n+1} - \mu_n)(X_{n+1} - \mu_n)^T - \Sigma_n\big)\big). \tag{2.3}$$

Thus, given $\mathcal{F}_n = \sigma\{X_k, \mu_k, \Sigma_k, k \le n\}$, $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$, where P_{θ_n} is the Markov kernel of the random walk Metropolis with target π and proposal $\mathcal{N}(x, \frac{2.38^2}{d}\Sigma_n + \varepsilon I_d)$. So, this algorithm generates a random process $\{(X_n, \theta_n), n \ge 0\}$ that is an adaptive Markov chain, as defined above. Here, the adaptation dynamics is given by (2.2) and (2.3).

Throughout the paper, we fix the initial measure of the process to some arbitrary measure $\bar{\mu}$ and simply write $\mathbb{E}$ and $\mathbb{P}$ for $\mathbb{E}_{\bar{\mu}}$ and $\mathbb{P}_{\bar{\mu}}$, respectively. We impose the following geometric ergodicity assumption.

A1: For each $\theta \in \Theta$, P_θ is phi-irreducible and aperiodic with invariant distribution π. There exists a measurable function $V : \mathsf{X} \to [1, \infty)$ with $\int V(x)\bar{\mu}(dx, d\theta) < \infty$ such that for any $\beta \in (0, 1]$, there exist $\rho \in (0, 1)$, $C \in (0, \infty)$ such that for any $(x, \theta) \in \mathsf{X} \times \Theta$,

$$\|P_\theta^n(x, \cdot) - \pi(\cdot)\|_{V^\beta} \le C\rho^n V^\beta(x), \qquad n \ge 0. \tag{2.4}$$

Furthermore, there exist constants $b \in (0, \infty)$, $\lambda \in (0, 1)$ such that for any $(x, \theta) \in \mathsf{X} \times \Theta$,

$$P_\theta V(x) \le \lambda V(x) + b. \tag{2.5}$$

Condition (2.4) is a standard geometric ergodicity assumption. We impose (2.5) in order to control the moments of the adaptive process. Condition (2.5) is probably redundant since geometric ergodicity intuitively implies a drift behavior of the form (2.5). But this is rarely an issue because both (2.4) and (2.5) are implied by the following minorization and drift conditions.

DR: Uniformly for $\theta \in \Theta$, there exist $\mathcal{C} \in \mathcal{X}$, ν a probability measure on $(\mathsf{X}, \mathcal{X})$, $b, \varepsilon > 0$ and $\lambda \in (0, 1)$ such that $\nu(\mathcal{C}) > 0$, $P_\theta(x, \cdot) \ge \varepsilon\nu(\cdot)\mathbb{1}_{\mathcal{C}}(x)$ and

$$P_\theta V \le \lambda V + b\mathbb{1}_{\mathcal{C}}. \tag{2.6}$$

This assertion follows from Theorem 1.1 of [10]. DR is known to hold for many Markov kernels used in MCMC simulation (see, e.g., [16] for some references). Either drift condition (2.5) or (2.6) implies that $\pi(V) < \infty$ ([22], Theorem 14.3.7). Therefore, under A1, if $f \in \mathcal{L}_{V^\beta}$ for some $\beta \in [0, 1]$, then $f \in L^{1/\beta}(\pi)$. Finally,

we note that under A1, a law of large numbers can be established for the adaptive chain (see, e.g., [7]). A short proof is provided here for completeness.

To state the law of large numbers, we need the following pseudo-metric on Θ. For $\beta \in [0,1]$, $\theta, \theta' \in \Theta$, set

$$D_\beta(\theta,\theta') \stackrel{\text{def}}{=} \sup_{|f|_{V^\beta}\le 1} \sup_{x\in\mathsf{X}} \frac{|P_\theta f(x) - P_{\theta'} f(x)|}{V^\beta(x)}.$$

PROPOSITION 2.1. *Assume* A1. *Let* $\beta \in [0,1)$ *and* $\{h_\theta \in \mathcal{L}_{V^\beta}, \theta\in\Theta\}$ *be a family of functions such that* $\pi(h_\theta)=0$, $(x,\theta)\to h_\theta(x)$ *is measurable and* $\sup_{\theta\in\Theta}|h_\theta|_{V^\beta} < \infty$. *Suppose also that*

$$\sum_{k\ge 1} k^{-1}\big(D_\beta(\theta_k,\theta_{k-1}) + |h_{\theta_k} - h_{\theta_{k-1}}|_{V^\beta}\big)V^\beta(X_k) < \infty, \qquad \mathbb{P}\text{-a.s.} \tag{2.7}$$

Then $n^{-1}\sum_{k=1}^n h_{\theta_{k-1}}(X_k)$ *converges almost surely* ($\mathbb{P}$) *to zero.*

PROOF. See Section 6.1. □

3. A central limit theorem. Central limit theorems are useful in assessing Monte Carlo errors. Several papers have studied central limit theorems for adaptive MCMC ([2, 7, 28]). The next proposition is adapted from [6]. For $h \in \mathcal{L}_V$, we introduce the resolvent functions

$$g_\theta(x) \stackrel{\text{def}}{=} \sum_{j\ge 0} \bar{P}_\theta^j h(x),$$

where $\bar{P}_\theta \stackrel{\text{def}}{=} P_\theta - \pi$. The dependence of g_θ on h is omitted for notational convenience. We also define $G_\theta(x,y) = g_\theta(y) - P_\theta g_\theta(x)$, where $P_\theta g_\theta(x) \stackrel{\text{def}}{=} \int P_\theta(x,dz) g_\theta(z)$. Whenever g_θ is well defined, it satisfies the so-called *Poisson equation*

$$h(x) = g_\theta(x) - \bar{P}_\theta g_\theta(x). \tag{3.1}$$

PROPOSITION 3.1. *Assume* A1. *Let* $\beta \in [0,1/2)$ *and* $h \in \mathcal{L}_{V^\beta}$ *be such that* $\pi(h)=0$. *Suppose that there exists a nonnegative random variable* $\Gamma^2(h)$, *finite* $\mathbb{P}$*-a.s.*, *such that*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^n G^2_{\theta_{k-1}}(X_{k-1},X_k) = \Gamma^2(h) \qquad \text{in } \mathbb{P}\text{-probability}. \tag{3.2}$$

Suppose also that

$$\sum_{k\ge 1} k^{-1/2} D_\beta(\theta_k,\theta_{k-1}) V^\beta(X_k) < \infty, \qquad \mathbb{P}\text{-a.s.} \tag{3.3}$$

Then $n^{-1/2}\sum_{k=1}^n h(X_k)$ *converges weakly to a random variable* $\sqrt{\Gamma^2(h)}Z$, *where* $Z\sim\mathcal{N}(0,1)$ *is a standard normal random variable independent of* $\Gamma^2(h)$.

PROOF. See Section 6.2. □

Condition (3.3), which strengthens (2.7), is a *diminishing adaptation condition* and is not hard to check in general. It follows from the following assumption which is much easier to check in practice.

A2: There exist $\eta \in [0, 1/2)$ and a nonincreasing sequence of positive numbers $\{\gamma_n, n \geq 1\}$, $\gamma_n = O(n^{-\alpha})$, $\alpha > 1/2$, such that for any $\beta \in [0, 1]$, there exists a finite constant C such that

$$D_\beta(\theta_{n-1}, \theta_n) \leq C\gamma_n V^\eta(X_n), \qquad \mathbb{P}\text{-a.s.} \tag{3.4}$$

[2] establishes A2 for the random walk Metropolis and the independence sampler. A similar result is obtained for the Metropolis adjusted Langevin algorithm in [4]. The constant η in A2 reflects the additional fluctuations due to the adaptation. For example, for a Metropolis algorithm with adaptation driven by a stochastic approximation of the form $\theta_{n+1} = \theta_n + \gamma_n H(\theta_n, X_{n+1})$, η is any nonnegative number such that $\sup_{\theta\in\Theta} |H(\theta, \cdot)|_{V^\eta} < \infty$.

PROPOSITION 3.2. *Under* A1–A2, (3.3) *holds*.

PROOF. Under A2, the left-hand side of (3.3) is bounded almost surely by $C \sum_{k\geq 1} k^{-1/2}\gamma_k V^{\eta+\beta}(X_k)$, the expectation of which is bounded by the term $C \sum_{k\geq 1} k^{-1/2}\gamma_k$ according to Lemma A.1(a), assuming A1. Since $\alpha > 1/2$, we conclude that (3.3) holds. □

Equation (3.2) is also a natural assumption. Indeed, in most adaptive MCMC algorithms, we seek to find the "best" Markov kernel from the family $\{P_\theta, \theta \in \Theta\}$ to sample from π. Thus, it is often the case that θ_n converges to some limit $\theta_\star$, say (see, e.g., [2, 3, 6, 9]). In these cases, (3.2) actually holds.

PROPOSITION 3.3. *Assume* A1–A2. *Let* $\beta \in [0, (1 - \eta)/2)$, *where* η *is as in* A2, *and let* $h \in \mathcal{L}_{V^\beta}$ *be such that* $\pi(h) = 0$. *Suppose that there exists a* Θ*-valued random variable* $\theta_\star$ *such that* $D_\beta(\theta_n, \theta_\star) + D_{2\beta}(\theta_n, \theta_\star)$ *converges in probability to zero*. *Then* (3.2) *holds*. *Furthermore*,

$$\Gamma^2(h) = \int_{\mathsf{X}\times\mathsf{X}} \pi(dx) P_{\theta_\star}(x, dy) G^2_{\theta_\star}(x, y).$$

PROOF. See Section 6.3. □

DEFINITION 3.1. We call the random variable $\Gamma^2(h)$ the *asymptotic average squared variation* of h and $\sigma^2(h) \stackrel{\text{def}}{=} \mathbb{E}(\Gamma^2(h))$ the *asymptotic variance* of h.

This definition is justified by the following result.

PROPOSITION 3.4. *Assume* A1–A2. *Let* $\beta \in [0, 1/2)$ *and* $h \in \mathcal{L}_{V^\beta}$ *be such that* $\pi(h) = 0$. *Assume that* (3.2) *holds. Then*

$$\lim_{n\to\infty} \operatorname{Var}\left(n^{-1/2}\sum_{k=1}^{n} h(X_k)\right) = \sigma^2(h).$$

PROOF. See Section 6.4. □

4. Asymptotic variance estimation. Denote by $\pi_n(h) = n^{-1}\sum_{k=1}^{n} h(X_k)$ the sample mean of $h(X_k)$ and denote by $\gamma_n(k)$ the sample autocovariance: $\gamma_n(k) = 0$ for $|k| \geq n$, $\gamma_n(-k) = \gamma_n(k)$ and for $0 \leq k < n$,

$$\gamma_n(k) = \frac{1}{n}\sum_{j=1}^{n-k}\bigl(h(X_j) - \pi_n(h)\bigr)\bigl(h(X_{j+k}) - \pi_n(h)\bigr).$$

Let $w : \mathbb{R} \to \mathbb{R}$ be a function with support $[-1, 1]$ [$w(x) = 0$ for $|x| \geq 1$]. We assume that w satisfies the following.

A3. The function w is even [$w(-x) = w(x)$] and $w(0) = 1$. Moreover, the restriction $w : [0, 1] \to \mathbb{R}$ is twice continuously differentiable.

Typical examples of kernels that satisfy A3 include, among others, the family of kernels

$$w(x) = \begin{cases} 1 - |x|^q, & \text{if } |x| \leq 1, \\ 0, & \text{if } |x| > 1, \end{cases} \tag{4.1}$$

for $q \geq 1$. The case $q = 1$ corresponds to the Bartlett kernel. A3 is also satisfied by the Parzen kernel

$$w(x) = \begin{cases} 1 - 6x^2 + 6|x|^3, & \text{if } |x| \leq \frac{1}{2}, \\ 2(1 - |x|)^3, & \text{if } \frac{1}{2} \leq |x| \leq 1, \\ 0, & \text{if } |x| > 1. \end{cases} \tag{4.2}$$

Our analysis does not cover nontruncated kernels such as the quadratic spectral kernel. But truncated kernels have the advantage of being computationally more efficient.

Let $\{b_n, n \geq 1\}$ be a nonincreasing sequence of positive numbers such that

$$b_n^{-1} = O(n^{1/2}) \quad \text{and} \quad |b_n - b_{n-1}| = O(b_n n^{-1}) \qquad \text{as } n \to \infty. \tag{4.3}$$

We consider the class of kernel estimator of the form

$$\Gamma_n^2(h) = \sum_{k=-n}^{n} w(kb_n)\gamma_n(k) = \sum_{k=-b_n^{-1}+1}^{b_n^{-1}-1} w(kb_n)\gamma_n(k). \tag{4.4}$$

The following is the main L^p-convergence result.

THEOREM 4.1. *Assume* A1–A3. *Let* $\beta \in (0, 1/2 - \eta)$ *and* $h \in \mathcal{L}_{V^\beta}$, *where* η *is as in* A2. *Then*

$$\Gamma_n^2(h) = \frac{1}{n}\sum_{k=1}^{n} G_{\theta_{k-1}}^2(X_{k-1}, X_k) + Q_n + D_n + \varepsilon_n, \qquad n \geq 1. \tag{4.5}$$

The random process $\{(Q_n, D_n, \varepsilon_n), n \geq 1\}$ *is such that for any* $p > 1$ *such that* $2p(\beta + \eta) \leq 1$, *there exists a finite constant* C *such that*

$$\begin{aligned} &\mathbb{E}(|Q_n|^p) \leq C\big(b_n + n^{-\alpha} b_n^{-1+\alpha} + n^{-1+(1/2)\vee(1/p)} b_n^{-1/2}\big)^p, \\ &\mathbb{E}(|D_n|^p) \leq C b_n^p \quad \text{and} \quad \mathbb{E}(|\varepsilon_n|^p) \leq C(n^{-1} b_n^{-1})^p. \end{aligned} \tag{4.6}$$

In particular, if $\lim_{n\to\infty} n^{-1+(1/2)\vee(1/p)} b_n^{-1/2} = 0$, *then*

$$\Gamma_n^2(h) - \frac{1}{n}\sum_{k=1}^{n} G_{\theta_{k-1}}^2(X_{k-1}, X_k)$$

converges to zero in L^p.

PROOF. The proof is given in the supplementary article [5]. □

REMARK 4.1. In Theorem 4.1, we can always take $p = 1/(2(\beta + \eta)) > 1$. In this case, the condition $\lim_{n\to\infty} n^{-1+(1/2)\vee(1/p)} b_n^{-1/2} = 0$ translates to $0.5 \vee (2(\beta + \eta)) + 0.5\delta < 1$. Therefore, if $\beta + \eta$ is close to $1/2$, we need to choose δ small. This remark implies that in applying the above result, one should always try to find the smallest possible β such that $h \in \mathcal{L}_{V^\beta}$.

It can be easily checked that the choice of bandwidth $b_n \propto n^{-\delta}$ with $\delta = \frac{2}{3}(1 - 0.5 \vee (2(\beta + \eta)))$ always satisfies Theorem 4.1. In fact, we will see in Section 4.2 that this choice of b_n is optimal in the L^p-norm, $p = (2(\beta + \eta))^{-1}$.

It is possible to investigate more carefully the rate of convergence of $\Gamma_n^2(h)$ in Theorem 4.1. Indeed, consider the typical case where $p = 2$ is admissible and we have $\alpha = 1$. If we choose b_n such that $b_n = o(n^{-1/3})$ and $n^{-1} = o(b_n)$, then the slowest term in (4.6) is $n^{-1+(1/2)\vee(1/p)} b_n^{-1/2} = (nb_n)^{-1/2}$. By inspecting the proof of Theorem 4.1, the only term whose L^p-norm enjoys such rate $n^{-1+(1/2)\vee(1/p)} b_n^{-1/2}$ is

$$Q_n^{(1)} = 2n^{-1}\sum_{j=2}^{n} Z_{n,j}^{(1)} G_{\theta_{j-1}}(X_{j-1}, X_j),$$

where

$$Z_{n,j}^{(1)} = \sum_{\ell=1}^{j-1} w\big((j-\ell)b_n\big) G_{\theta_{\ell-1}}(X_{\ell-1}, X_\ell).$$

Now, $\{(Q_n^{(1)}, \mathcal{F}_n), n \geq 2\}$ is a martingale array and we conjecture that as $n \to \infty$,

$$(nb_n)^{1/2}\left(\Gamma_n^2(h) - \frac{1}{n}\sum_{k=1}^{n} G_{\theta_{k-1}}^2(X_{k-1}, X_k)\right) \xrightarrow{w} \mathcal{N}(0, \Lambda^2),$$

at least in the special case where θ_n converges to a deterministic limit. But we do not pursue this further since the issue of a central limit theorem for $\Gamma_n^2(h)$ is less relevant for Monte Carlo simulation.

When $\{X_n, n \geq 0\}$ is a Markov chain, Theorem 4.1 improves on [16], as it imposes weaker moment conditions. Almost sure convergence is often more desirable in Monte Carlo settings, but typically requires stronger assumptions. One can impose either more restrictive growth conditions on h (which translates into stronger moment conditions, as in [16]) or one can impose stronger smoothness conditions on the function w. We prove both types of results.

THEOREM 4.2. *Assume* A1–A3 *with* $\eta < 1/4$, *where* η *is as in* A2. *Let* $\beta \in (0, 1/4 - \eta)$ *and* $h \in \mathcal{L}_{V^\beta}$. *Suppose that* $b_n \propto n^{-\delta}$, *where* $\delta \in (2(\beta + \eta), 1/2)$. *Then*

$$\lim_{n\to\infty}\left(\Gamma_n^2(h) - \frac{1}{n}\sum_{k=1}^{n} G_{\theta_{k-1}}^2(X_{k-1}, X_k)\right) = 0$$

almost surely.

PROOF. The proof is given in the supplementary article [5]. □

We can remove the growth condition $h \in \mathcal{L}_{V^\beta}$, $0 < \beta < 0.25 - \eta$, and the constraint on b_n in Theorem 4.2 if we are willing to impose a stronger smoothness condition on w. To do so, we replace A3 with A4.

A4: The function w is even $[w(-x) = w(x)]$ and $w(0) = 1$. Moreover, the restriction $w : [0, 1] \to \mathbb{R}$ is $(r + 1)$-times continuously differentiable for some $r \geq 2$.

THEOREM 4.3. *Assume* A1–A2 *and* A4. *Let* $\beta \in (0, 1/2 - \eta)$ *and* $h \in \mathcal{L}_{V^\beta}$, *where* η *is as in* A2. *Let* $p > 1$ *be such that* $2p(\beta + \eta) \leq 1$. *Suppose, in addition, that*

$$\begin{aligned} &\sum_{n\geq 1}(n^{-1}b_n^{-1})^p < \infty, \qquad \sum_{n\geq 1}(n^{-2}b_n^{-1})^{1\wedge(p/2)} < \infty, \\ &\sum_{n\geq 1} n^{-2+(1/2)\vee(1/p)} b_n^{-1/2} < \infty \quad \text{and} \quad \sum_{n\geq 1} b_n^{(r-1)p} < \infty. \end{aligned} \tag{4.7}$$

The conclusion of Theorem 4.2 then holds.

PROOF. The proof is given in the supplementary article [5]. □

REMARK 4.2. Not all kernels used in practice will satisfy A4. For instance, A4 holds for kernels in the family (4.1) but fails to hold for the Parzen kernel (4.2).

In Theorem 4.3, we can again choose $b_n \propto n^{-\delta}$, where $\delta = \frac{2}{3}(1 - 0.5 \vee (2(\beta + \eta)))$. It is easy to check that if A4 holds with $r > 1 + 2(\beta + \eta)\delta^{-1}$ and we take $p = (2(\beta + \eta))^{-1}$, then this choice of b_n satisfies (4.7).

In the next corollary, we consider the Markov chain case.

COROLLARY 4.1. *Suppose that* $\{X_n, n \geq 0\}$ *is a phi-irreducible, aperiodic Markov chain with transition kernel* P *and invariant distribution* π. *Assume that* P *satisfies* A1. *Let* $\beta \in (0, 1/2)$ *and* $h \in \mathcal{L}_{V^\beta}$. *Then* $\sigma^2(h) := \pi(h^2) + 2\sum_{j\geq 1} \pi(hP^j h)$ *is finite. Assume* A3 *and take* $b_n \propto n^{-\delta}$ *with* $\delta = \frac{2}{3}(1 - 0.5 \vee (2\beta))$. *Then*

$$\lim_{n\to\infty} \Gamma_n^2(h) = \sigma^2(h) \qquad \text{in } L^{(2\beta)^{-1}}.$$

Supposing, in addition, that $\beta \in (0, 1/4)$ *and* $\delta \in (2\beta, 1/2)$, *or that* A4 *holds with* $r > 1 + 2\beta\delta^{-1}$, *then the convergence holds almost surely* $(\mathbb{P})$ *as well.*

4.1. *Application to the adaptive Metropolis algorithm.* We shall now apply the above result to the adaptive Metropolis algorithm described in Example 2.1. We continue to use the notation established in that example. We recall that $\mathsf{X} = \mathbb{R}^d$, $\Theta = \Theta_\mu(r_1) \times \Theta_+(r_2)$, where $\Theta_\mu(r_1)$ is the ball in X with center 0 and radius $r_1 > 0$ and $\Theta_+(r_2)$ is the set of all symmetric positive semidefinite matrices A with $|A| \leq r_2$. Define $\ell(x) = \log \pi(x)$. We assume that:

B1: π is positive and continuously differentiable,

$$\lim_{|x|\to\infty} \left\langle \frac{x}{|x|}, \nabla\ell(x) \right\rangle = -\infty$$

and

$$\lim_{|x|\to\infty} \left\langle \frac{x}{|x|}, \frac{\nabla\ell(x)}{|\nabla\ell(x)|} \right\rangle < 0,$$

where $\nabla\ell$ is the gradient of ℓ.

B1 is known to imply A1 with $V(x) = (\sup_{x\in\mathsf{X}} \pi^\zeta(x))\pi^{-\zeta}(x)$, for any $\zeta \in (0, 1)$ ([2, 20]). We denote by $\mu_\star$ and $\Sigma_\star$ the mean and covariance matrix of π, respectively. We assume that $(\mu_\star, \Sigma_\star) \in \Theta$, which can always be achieved by taking r_1, r_2 large enough.

By Lemma 12 of [2], for any $\beta \in (0, 1]$,

$$D_\beta(\theta_n, \theta_{n-1}) \leq C|\Sigma_n - \Sigma_{n-1}| \leq \gamma_n V^\eta(X_n) \tag{4.8}$$

for any $\eta > 0$. Thus, A2 holds and η can be taken to be arbitrarily small. We can now summarize Proposition 3.1 and Theorems 4.1–4.3 for the random Metropolis algorithm. We focus here on the choice of bandwidth $b_n \propto n^{-\delta}$, where $\delta = \frac{2}{3}(1 - 0.5 \vee (2\beta))$, but similar conclusions can be derived from the theorems for other bandwidths.

PROPOSITION 4.1. *Assume* B1, *let* $V(x) = (\sup_{x\in\mathsf{X}} \pi^\zeta(x))\pi^{-\zeta}(x)$ *for* $\zeta \in (0, 1)$ *and suppose that* $(\mu_\star, \Sigma_\star) \in \Theta$. *Then* $\theta_n = (\mu_n, \Sigma_n)$ *converges in probability*

to $\theta_\star = (\mu_\star, \Sigma_\star)$. *Let* $\beta \in (0, 1/2)$ *and* $h \in \mathcal{L}_{V^\beta}$.

1. $n^{-1/2}\sum_{k=1}^n h(X_k)$ *converges weakly to* $\mathcal{N}(\pi(h), \sigma_\star^2(h))$ *as* $n \to \infty$, *where* $\sigma_\star^2(h) = \pi(h^2) + 2\sum_{j\geq 1}\pi(hP_{\theta_\star}^j h)$ *and* $\theta_\star = \Sigma_\star + \varepsilon I_d$.

2. *Suppose that* A3 *holds and we choose* $b_n \propto n^{-\delta}$, $\delta = \frac{2}{3}(1 - 0.5 \vee (2\beta))$. *Then* $\Gamma_n^2(h)$ *converges to* $\sigma_\star^2(h)$ *in* L^p *for* $p = (2\beta)^{-1}$. *If we additionally suppose that* $\beta \in (0, 1/4)$ *and* $\delta \in (2\beta, 1/2)$, *or that* A4 *holds with* $r > 1 + 2\beta\delta^{-1}$, *then the convergence of* $\Gamma_n^2(h)$ *holds almost surely* $(\mathbb{P})$ *as well.*

4.2. *Choosing the bandwidth* b_n. Consider Theorem 4.1. Suppose that $\alpha \geq 2/3$ and that we take $b_n \propto n^{-\delta}$ for some $\delta \in (0, 1/2]$. Then $n^{-\alpha}b_n^{-1+\alpha} = O(n^{-1/2})$. Similarly, $n^{-1}b_n^{-1} = O(n^{-1/2})$. Thus, the L^p-rate of convergence of $\Gamma_n^2(h)$ is driven by b_n and $n^{-1+(1/2)\vee(1/p)}b_n^{-1/2}$, and we deduce from equating these two terms that the optimal choice of b_n is given by $b_n \propto n^{-\delta}$ for $\delta = \frac{2}{3}(1 - \frac{1}{2} \vee \frac{1}{p})$. Equation (4.6) then gives that

$$\mathbb{E}^{1/p}\left(\left|\Gamma_n^2(h) - \frac{1}{n}\sum_{k=1}^n G_{\theta_{k-1}}^2(X_{k-1}, X_k)\right|^p\right) \leq Cn^{-\delta}.$$

In particular, if $4(\beta + \eta) \leq 1$ (and $\alpha \geq 2/3$), we can take $p = 2$ and then $\delta = 1/3$, which leads to

$$\mathbb{E}^{1/2}\left(\left|\Gamma_n^2(h) - \frac{1}{n}\sum_{k=1}^n G_{\theta_{k-1}}^2(X_{k-1}, X_k)\right|^2\right) \leq Cn^{-1/3}.$$

The same L^2-rate of convergence was also derived in [16].

Even with $b_n = \frac{1}{cn^{1/3}}$, the estimator is still very sensitive to the choice of c. Choosing c is a difficult issue where more research is needed. Here, we follow a data-driven approach adapted from [1] and [25]. In this approach, we take $b_n = \frac{1}{cn^{1/3}}$, where

$$c = c_0\left\{\frac{2\sum_{\ell=1}^m \ell\hat{\rho}_\ell}{1 + 2\sum_{\ell=1}^m \hat{\rho}_\ell}\right\}^{1/3}$$

for some constants c_0 and m, where $\hat{\rho}_\ell$ is the ℓth order sample autocorrelation of $\{h(X_n), n \geq 0\}$. [25] suggests choosing $m = n^{2/9}$. Our simulation results show that small values of c_0 yield small variances but high biases, and inversely for large values of c_0. The value c_0 also depends on how fast the autocorrelation of the process decays. [25] derives some theoretical results on the consistency of this procedure in the stationary case. Whether these results hold in the present nonstationary case is an open question.

4.3. *Discussion*. The above results raise a number of issues. On one hand, we note from Theorems 4.1–4.3 that the kernel estimator $\Gamma_n^2(h)$ does not converge to the asymptotic variance $\sigma^2(h)$, but rather to the asymptotic average squared variation $\Gamma^2(h)$. On the other hand, Proposition 3.1 shows that although the asymptotic variance $\sigma^2(h)$ controls the fluctuations of $n^{-1/2}\sum_{k=1}^k h(X_k)$ as $n \to \infty$, the limiting distribution of $n^{-1/2}\sum_{k=1}^k h(X_k)$ is not the Gaussian $\mathcal{N}(0, \sigma^2(h))$, but instead a mixture of Gaussian distribution of the form $\sqrt{\Gamma^2(h)}Z$. With these conditions, how can one undertake a valid error assessment from adaptive MCMC samplers?

If the adaptation parameter θ_n converges to a deterministic limit $\theta_\star$, then one gets a situation similar to that of Markov chains. This is the ideal case. Indeed, in such cases, $\Gamma^2(h) \equiv \sigma^2(h)$, $n^{-1/2}\sum_{k=1}^n h(X_k)$ converges weakly to a random variable $\mathcal{N}(0, \sigma^2(h))$ and the kernel estimator $\Gamma_n^2(h)$ converges to the asymptotic variance $\sigma^2(h)$, where

$$\sigma^2(h) = \int_{\mathsf{X}\times\mathsf{X}} \pi(dx) P_{\theta_\star}(x, dy) G_{\theta_\star}^2(x, y) = \pi(h^2) + 2\sum_{j\geq 1} \pi(h P_{\theta_\star}^j h).$$

This case includes the adaptive Metropolis algorithm of [17], as discussed in Section 4.1.

However, in some other cases (see, e.g., [2, 7]), what one can actually prove is that $\theta_n \to \theta_\star$, where $\theta_\star$ is a discrete random variable with values in a subset $\{\tau_1, \tau_2, \ldots, \tau_N\}$, say, of Θ. This is typically the case when the adaptation is driven by a stochastic approximation $\theta_{n+1} = \theta_n + \gamma_n H(\theta_n, X_{n+1})$, where the mean field equation $h(\theta) \stackrel{\text{def}}{=} \int_{\mathsf{X}} H(\theta, x)\pi(dx) = 0$ has multiple solutions.

In these cases, $\Gamma_n^2(h)$ clearly provides a poor estimate for $\sigma^2(h)$, even though it is not hard to see that

$$\lim_{n\to\infty} \mathbb{E}(\Gamma_n^2(h)) = \mathbb{E}(\Gamma^2(h)) = \sigma^2(h).$$

Furthermore, a confidence interval for $\pi(h)$ becomes difficult to build. Indeed, the asymptotic distribution $n^{-1/2}\sum_{k=1}^n h(X_k)$ is a mixture

$$\sum_{k\geq 1} p_k \mathcal{N}(0, \sigma_k),$$

where $p_k \stackrel{\text{def}}{=} \mathbb{P}(\theta_\star = \tau_k)$ and $\sigma_k^2(h) = \pi(h^2) + 2\sum_{j\geq 1} \pi(h P_{\tau_k}^j h)$. As a consequence, a valid confidence interval for $\pi(h)$ requires the knowledge of the mixing distribution p_k and the asymptotic variances $\sigma_k^2(h)$, which is much more than one can obtain from $\Gamma_n^2(h)$. It is possible to improve on the estimation of $\sigma^2(h)$ by running multiple chains, but this takes away some of the advantages of the adaptive MCMC framework.

In view of this discussion, when Monte Carlo error assessment is important, it seems that the framework of adaptive MCMC is most useful when the adaptation mechanism is such that there exists a unique, well-defined, optimal kernel $P_{\theta_\star}$ that

the algorithm converges to. This is the case, for example, with the popular adaptive RWM of [17] discussed above and its extension to the MALA (Metropolis adjusted Langevin algorithm; see, e.g., [4]).

5. Examples.

5.1. *The* GARCH(1, 1) *model.* To illustrate the above results in the Markov chain case, we consider the linear GARCH(1, 1) model defined as follows: $h_0 \in (0, \infty)$, $u_0 \sim \mathcal{N}(0, h_0)$ and, for $n \geq 1$,

$$u_n = h_n^{1/2} \varepsilon_n,$$

$$h_n = \omega + \beta h_{n-1} + \alpha u_{n-1}^2,$$

where $\{\varepsilon_n, n \geq 0\}$ is i.i.d. $\mathcal{N}(0, 1)$ and $\omega > 0$, $\alpha \geq 0$, $\beta \geq 0$. We assume that α, β satisfy the following.

E1: There exists $\nu > 0$ such that

$$\mathbb{E}[(\beta + \alpha Z^2)^\nu] < 1, \qquad Z \sim \mathcal{N}(0, 1). \tag{5.1}$$

It is shown by [21], Theorem 2, that under (5.1), the joint process $\{(u_n, h_n), n \geq 0\}$ is a phi-irreducible aperiodic Markov chain that admits an invariant distribution and is geometrically ergodic with a drift function $V(u, h) = 1 + h^\nu + |u|^{2\nu}$. Therefore, A1 holds and we can apply Corollary 4.1. We write $\mathbb{E}_\pi$ to denote expectation taken under the stationary measure. We are interested in the asymptotic variance of the functions $h(u) = u^2$. We can calculate the exact value. Define $\rho_n \stackrel{\text{def}}{=} \operatorname{Corr}_\pi(u_0^2, u_n^2)$. As observed by [11] in introducing the GARCH models, if (5.1) hold with some $\nu \geq 2$, then

$$\rho_1 = \frac{\alpha(1 - \alpha\beta - \beta^2)}{1 - 2\alpha\beta - \beta^2}, \qquad \rho_n = \rho_1(\alpha + \beta)^{n-1}, \qquad n \geq 2.$$

Also,

$$\operatorname{Var}_\pi(u_0^2) = \frac{3\omega^2(1 + \alpha + \beta)}{(1 - \alpha - \beta)(1 - \beta^2 - 2\alpha\beta - 3\alpha^2)} - \left(\frac{\omega}{1 - \alpha - \beta}\right)^2$$

and we obtain

$$\sigma^2(h) = \operatorname{Var}_\pi(u_0^2)\left(1 + 2\frac{\rho_1}{1 - \alpha - \beta}\right).$$

For the simulations, we set $\omega = 1$, $\alpha = 0.1$, $\beta = 0.7$, which gives $\sigma^2(h) = 119.1$. For these values, (5.1) holds with at least $\nu = 4$. We tested the Bartlett and the Parzen kernels for which A3 holds. We choose the bandwidth following the approach outlined in Remark 4.2 with $c_0 = 1.5$. We run the GARCH(1, 1) Markov chain for 250,000 iterations and discard the first 10,000 iterations as burn-in. We compute $\Gamma_n^2(h)$ at every 1000 along the sample path. The results are plotted in Figure 1.

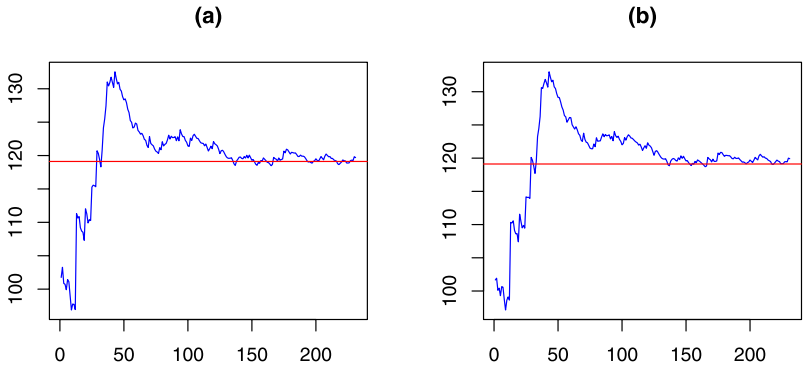

FIG. 1. *Asymptotic variance estimation for* GARCH(1, 1) *with* $\omega = 1$, $\alpha = 0.1$, $\beta = 0.7$ *based on* 250,000 *iterations*. (a) *is Bartlett kernel*, (b) *is Parzen kernel*.

5.2. *Logistic regression*. We also illustrate the results with MCMC and adaptive MCMC. We consider the logistic regression model

$$y_i \sim \mathcal{B}(p_\beta(x_i)), \qquad i = 1, \ldots, n,$$

where $y_i \in \{0, 1\}$ and $p_\beta(x) = e^{x\beta}(1 + e^{x\beta})^{-1}$ for a parameter $\beta \in \mathbb{R}^d$ and a covariate vector $x^T \in \mathbb{R}^d$, where x^T denotes the transpose of x. $\mathcal{B}(p)$ is the Bernoulli distribution with parameter p. The log-likelihood is

$$\ell(\beta|X) = \sum_{i=1}^{n} y_i x_i \beta - \log(1 + e^{x_i\beta}).$$

We assume a Gaussian prior distribution $\pi(\beta) \propto e^{-1/(2s^2)|\beta|^2}$ for some constant $s > 0$ leading to a posterior distribution

$$\pi(\beta|X) \propto e^{\ell(\beta|X)} e^{-1/(2s^2)|\beta|^2}.$$

The RWM algorithm described in Example 2.1 is a possible choice to sample from the posterior distribution. We compare a plain RWM with proposal density $\mathcal{N}(0, e^c I_d)$ with $c = -2.3$ and the adaptive RWM described in Algorithm 2.1 using the family $\{P_\theta, \theta \in \Theta\}$, where $\Theta = \Theta_\mu(r_1) \times \Theta_+(r_2)$, as defined in Example 2.1. It is easy to check that B1 holds. Indeed, we have

$$\langle \beta, \nabla \log \pi(\beta) \rangle = -\frac{|\beta|^2}{s^2} + \sum_{i=1}^{n} \big(y_i - p_\beta(x_i)\big) \langle \beta, x_i^T \rangle$$

and $|\sum_{i=1}^n (y_i - p_\beta(x_i)) \langle \beta, x_i^T \rangle| \le |\beta| \sum_{i=1}^n |x_i|$. We deduce that

$$\left\langle \frac{\beta}{|\beta|}, \nabla \log \pi(\beta) \right\rangle \le -\frac{|\beta|}{s^2} + \sum_{i=1}^{n} |x_i| \to -\infty \qquad \text{as } |\beta| \to \infty.$$

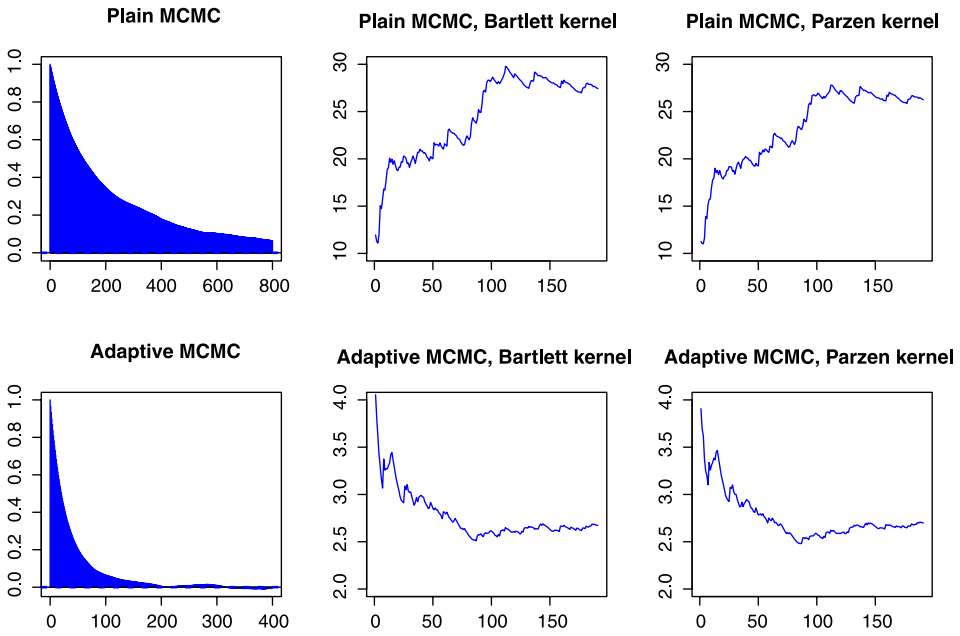

FIG. 2. *Asymptotic variance estimation for logistic regression modeling of the heart data set. Outputs of the coefficient* $\beta^{(2)}$ *are reported, based on* 250,000 *iterations.*

Similarly,

$$\left\langle \frac{\beta}{|\beta|}, \frac{\nabla \log \pi(\beta)}{|\nabla \log \pi(\beta)|} \right\rangle \leq -\frac{1}{s^2} \frac{|\beta|}{|\nabla \log \pi(\beta)|} + \frac{\sum_{i=1}^{n} |x_i|}{|\nabla \log \pi(\beta)|} \to -1 \qquad \text{as } |\beta| \to \infty$$

since $|\nabla \log \pi(\beta)| \sim s^{-2}|\beta|$ as $|\beta| \to \infty$. Therefore, B1 holds. If we choose r_1, r_2 large enough so that $(\mu_\star, \Sigma_\star) \in \Theta$, then Proposition 4.1 holds and applies to any measurable function h such that $|h(\beta)| \leq c\pi^{-t}(\beta|X)$ for some $t \in [0, 1/2)$.

As a simulation example, we test the model with the *Heart data set* which has $n = 217$ cases and $d = 14$ covariates. The dependent variable is the presence or absence of a heart disease and the explanatory variables are relevant covariates. More details can be found in [23]. We use Parzen and Bartlett kernels with $c_0 = 20$ for the Markov chain and $c_0 = 5$ for the adaptive chain. We run both chains for 250,000 iterations and discard the first 50,000 iterations as burn-in. The results are plotted in Figure 2 for the coefficient β_2. We also report in Table 1 below the resulting confidence for the first four coefficients $(\beta_1, \ldots, \beta_4)$.

6. Proofs. This section contains the proofs of the statements from Sections 2–3. The remaining proofs are available in the supplementary paper [5]. Throughout this section, we shall use C to denote a generic constant whose actual value might change from one appearance to the next. On multiple occasions,

TABLE 1
Confidence interval for the first four parameters of the model for the heart data set

Parameters	Plain MCMC	Adaptive RWM
β_1	$[-0.271, -0.239]$	$[-0.272, -0.257]$
β_2	$[-0.203, -0.158]$	$[-0.182, -0.170]$
β_3	$[0.744, 0.785]$	$[0.776, 0.793]$
β_4	$[0.727, 0.756]$	$[0.736, 0.750]$

we make use of the Kronecker lemma and the Toeplitz lemma. We refer the reader to [18], Section 2.6, for a statement and proof of these lemmata.

We shall routinely use the following martingale inequality. Let $\{D_i, \mathcal{F}_i, i \geq 1\}$ be a martingale difference sequence. For any $p > 1$,

$$\mathbb{E}\left(\left|\sum_{i=1}^{n} D_i\right|^p\right) \leq C\left\{\sum_{i=1}^{n} \mathbb{E}^{1\wedge(2/p)}(|D_i|^p)\right\}^{1\vee(p/2)}, \tag{6.1}$$

where C can be taken as $C = (18pq^{1/2})^p$, $p^{-1} + q^{-1} = 1$.

We also notice that for any $q \in [1, \beta^{-1}]$, Lemma A.1(a)–(b) implies that

$$\sup_{k\geq 1} \mathbb{E}(|G_{\theta_{k-1}}(X_{k-1}, X_k)|^q) < \infty. \tag{6.2}$$

6.1. *Proof of Proposition 2.1.* Let $S_n \stackrel{\text{def}}{=} \sum_{k=1}^{n} h_{\theta_{k-1}}(X_k)$. For $\theta \in \Theta$, we define $\tilde{g}_\theta(x) = \sum_{j\geq 0} P_\theta^j h_\theta(x)$. When h_θ does not depend on θ, we obtain $\tilde{g}_\theta = g_\theta$, as defined in Section 3. Similarly, we define $\tilde{G}_\theta(x, y) = \tilde{g}_\theta(y) = P_\theta \tilde{g}_\theta(x)$. Using the Poisson equation $\tilde{g}_\theta - P_\theta \tilde{g}_\theta = h_\theta$, we rewrite S_n as $S_n = M_n + R_n$, where

$$M_n \stackrel{\text{def}}{=} \sum_{k=1}^{n} \tilde{G}_{\theta_{k-1}}(X_{k-1}, X_k)$$

and

$$R_n \stackrel{\text{def}}{=} P_{\theta_0}\tilde{g}_{\theta_0}(X_0) - P_{\theta_n}\tilde{g}_{\theta_n}(X_n) + \sum_{k=1}^{n}\big(\tilde{g}_{\theta_k}(X_k) - \tilde{g}_{\theta_{k-1}}(X_k)\big).$$

Using Lemma A.1 and A1, we easily see that

$$|R_n| \leq C\left(V^\beta(X_0) + V^\beta(X_n) + \sum_{k=1}^{n}\big(D_\beta(\theta_k, \theta_{k-1}) + |h_{\theta_k} - h_{\theta_{k-1}}|_{V^\beta}\big)V^\beta(X_k)\right).$$

For $p > 1$ such that $\beta p \leq 1$, $\sum_{k\geq 1} n^{-p}\mathbb{E}((V^\beta(X_0) + V^\beta(X_n))^p) < \infty$. This is a consequence of Lemma A.1(a) and the Minkowski inequality. Thus, $n^{-1} \times (V^\beta(X_0) + V^\beta(X_n))$ converges almost surely to zero. By (2.7) and the Kronecker lemma, the term $n^{-1}\sum_{k=1}^{n}(D_\beta(\theta_k, \theta_{k-1}) + |h_{\theta_k} - h_{\theta_{k-1}}|_{V^\beta})V^\beta(X_k)$ converges almost surely to zero. We conclude that $n^{-1}R_n$ converges almost surely to zero.

$\{(M_n, \mathcal{F}_n), n \geq 1\}$ is a martingale. Again, let $p > 1$ be such that $\beta p \leq 1$. Equation (6.1) and Lemma A.1(a) together imply that $\mathbb{E}(|M_n|^p) = O(n^{1\vee(p/2)})$, which, combined with Proposition A.1 of [5], implies that $n^{-1}M_n$ converges almost surely to zero.

6.2. *Proof of Proposition 3.1.* This is a continuation of the previous proof. In the present case, $h_\theta \equiv h$, so we write g_θ and G_θ instead of $\tilde{g}_\theta$ and $\tilde{G}_\theta$, respectively. Again, let $S_n \stackrel{\text{def}}{=} \sum_{k=1}^n h(X_k)$. We have $S_n = M_n + R_n$, where $M_n \stackrel{\text{def}}{=} \sum_{k=1}^n G_{\theta_{k-1}}(X_{k-1}, X_k)$ and

$$|R_n| \leq C\left(V^\beta(X_0) + V^\beta(X_n) + \sum_{k=1}^n D_\beta(\theta_k, \theta_{k-1})V^\beta(X_k)\right).$$

$\mathbb{E}(V^\beta(X_0) + V^\beta(X_n))$ is bounded in n, thus $n^{-1/2}(V^\beta(X_0) + V^\beta(X_n))$ converges in probability to zero. By (3.3) and the Kronecker lemma, the term $n^{-1/2}\sum_{k=1}^n D_\beta(\theta_k, \theta_{k-1})V^\beta(X_k)$ converges almost surely to zero. We conclude that $n^{-1/2}R_n$ converges in probability to zero.

$\{(M_n, \mathcal{F}_n), n \geq 1\}$ is a martingale. Since $\beta < 1/2$, (6.2) implies that $\{(M_n, \mathcal{F}_n), n \geq 1\}$ is a square integrable martingale and also that we have

$$\begin{aligned} &\sup_{n\geq 1}\mathbb{E}\Big(\max_{1\leq k\leq n} n^{-1}G^2_{\theta_{k-1}}(X_{k-1}, X_k)\Big) < \infty \quad \text{and} \\ &\lim_{n\to\infty}\max_{1\leq k\leq n} n^{-1/2}G_{\theta_{k-1}}(X_{k-1}, X_k) = 0 \qquad \text{(in probability)}. \end{aligned} \tag{6.3}$$

Equations (3.2) and (6.3) imply, by Theorem 3.2 of [18], that $n^{-1/2}M_n$ converges weakly to a random variable $\sqrt{\Gamma^2(h)}Z$, where $Z \sim \mathcal{N}(0, 1)$, and is independent of $\Gamma^2(h)$.

6.3. *Proof of Proposition 3.3.* We have

$$\begin{aligned} &\frac{1}{n}\sum_{k=1}^n G^2_{\theta_{k-1}}(X_{k-1}, X_k) \\ &\quad= \frac{1}{n}\sum_{k=1}^n \big(G^2_{\theta_{k-1}}(X_{k-1}, X_k) - P_{\theta_{k-1}}G^2_{\theta_{k-1}}(X_{k-1})\big) \\ &\qquad+ \frac{1}{n}\sum_{k=1}^n \left(P_{\theta_{k-1}}G^2_{\theta_{k-1}}(X_{k-1}) - \int_{\mathsf{X}} \pi(dx)P_{\theta_{k-1}}G^2_{\theta_{k-1}}(x)\right) \\ &\qquad+ \frac{1}{n}\sum_{k=1}^n \int_{\mathsf{X}} \pi(dx)\big(P_{\theta_{k-1}}G^2_{\theta_{k-1}}(x) - P_{\theta_\star}G^2_{\theta_\star}(x)\big) + \int_{\mathsf{X}} P_{\theta_\star}G^2_{\theta_\star}(x)\pi(dx) \\ &\quad= T_n^{(1)} + T_n^{(2)} + T_n^{(3)} + \int_{\mathsf{X}} P_{\theta_\star}G^2_{\theta_\star}(x)\pi(dx), \end{aligned}$$

say. The term $T_n^{(1)}$ is an $\mathcal{F}_n$-martingale. Indeed, $\mathbb{E}(G^2_{\theta_{k-1}}(X_{k-1}, X_k)|\mathcal{F}_{k-1}) = P_{\theta_{k-1}} G^2_{\theta_{k-1}}(X_{k-1})$, $\mathbb{P}$-a.s. Furthermore, by (6.2), the martingale differences $G^2_{\theta_{k-1}}(X_{k-1}, X_k) - P_{\theta_{k-1}} G^2_{\theta_{k-1}}(X_{k-1})$ are L^p-bounded for some $p > 1$. By [18], Theorem 2.22, we conclude that $T_n^{(1)}$ converges in L^1 to zero.

The term $T_n^{(2)}$ converges in probability to zero as a consequence of the law of large numbers (Proposition 2.1). Using the definition of D_β and Lemma A.1(a)–(b), we can find a constant C such that

$$\left|\int_{\mathsf{X}} \pi(dx)\big(P_{\theta_n} G^2_{\theta_n}(x) - P_{\theta_\star} G^2_{\theta_\star}(x)\big)\right| \le C\big(D_\beta(\theta_n, \theta_\star) + D_{2\beta}(\theta_n, \theta_\star)\big) \int_{\mathsf{X}} V^{2\beta}(x)\pi(dx),$$

almost surely. It follows that $T_n^{(3)}$ also converges in $\mathbb{P}$-probability to zero.

6.4. *Proof of Proposition 3.4.* From the proof of Proposition 2.1 above, we have seen that $S_n = M_n + R_n$, and it is easy to check that $\mathbb{E}(|R_n|^2) = O(n^{2(1-\alpha)})$ and, by (6.2), $\mathbb{E}(|M_n|^2) = O(n)$. Therefore,

$$\begin{aligned} &|\operatorname{Var}(n^{-1/2} S_n) - n^{-1}\mathbb{E}(M_n^2)| \\ &\quad = |2n^{-1}\mathbb{E}(M_n R_n) + n^{-1}\mathbb{E}(R_n^2) - n^{-1}(\mathbb{E}(R_n))^2| \\ &\quad = O(n^{1/2-\alpha}) \to 0 \qquad \text{as } n \to \infty \end{aligned}$$

since $\alpha > 1/2$. Now,

$$n^{-1}\mathbb{E}(M_n^2) = \mathbb{E}\left(n^{-1}\sum_{k=1}^n G^2_{\theta_{k-1}}(X_{k-1}, X_k)\right).$$

Again, from (6.2), the sequence $n^{-1}\sum_{k=1}^n G^2_{\theta_{k-1}}(X_{k-1}, X_k)$ is uniformly integrable which, combined with (3.2) and Lebesgue's dominated convergence theorem, implies that $n^{-1}\mathbb{E}(M_n^2)$ converges to $\mathbb{E}(\Gamma^2(h))$.

APPENDIX A: SOME USEFUL CONSEQUENCES OF A1

LEMMA A.1. *Assume that* $\{P_\theta, \theta \in \Theta\}$ *satisfies* A1.

(a) *There exists a finite constant* C *such that*

$$\sup_{n\ge 0} \mathbb{E}(V(X_n)) \le C. \tag{A.1}$$

(b) *Let* $\beta \in (0, 1]$ *and* $\{h_\theta \in \mathcal{L}_{V^\beta}, \theta \in \Theta\}$ *be such that* $\pi(h_\theta) = 0$, $\sup_{\theta\in\Theta} |h_\theta|_{V^\beta} < \infty$. *The function* $\tilde{g}_\theta \stackrel{\text{def}}{=} \sum_{j\ge 0} P_\theta^j h_\theta(x)$ *is then well defined,* $|\tilde{g}_\theta|_{V^\beta} \le$

$C|h_\theta|_{V^\beta}$, where the constant C does not depend on $\{h_\theta \in \mathcal{L}_{V^\beta}, \theta \in \Theta\}$. Moreover, we can take C such that for any $\theta, \theta' \in \Theta$,

$$|\tilde{g}_\theta - \tilde{g}_{\theta'}|_{V^\beta} \le C \sup_{\theta\in\Theta} |h_\theta|_{V^\beta} \big(D_\beta(\theta,\theta') + |h_\theta - h_{\theta'}|_{V^\beta}\big). \tag{A.2}$$

(c) *Assume* A2. *Let $\beta \in (0, 1-\eta)$ and $h \in \mathcal{L}_{V^\beta}$ be such that $\pi(h) = 0$. Define $S_n(j) = \sum_{\ell=j+1}^{j+n} h(X_\ell)$. Let $p \in (1, (\beta+\eta)^{-1})$. There then exists a finite constant C that does not depend on n, j, θ or h such that*

$$\mathbb{E}(|S_n(j)|^p) \le C|h|_{V^\beta} n^{1\vee(p/2)}.$$

PROOF. Parts (a) and (b) are standard results (see, e.g., [2]). To prove (c), we use the Poisson equation (3.1) to write

$$\begin{aligned} S_n(j) = {} & \sum_{\ell=j+1}^{j+n} G_{\theta_{\ell-1}}(X_{\ell-1}, X_\ell) + P_{\theta_j} g_{\theta_j}(X_j) - P_{\theta_{j+n}} g_{\theta_{j+n}}(X_{j+n}) \\ & + \sum_{\ell=j+1}^{j+n} \big(g_{\theta_{\ell-1}}(X_\ell) - g_{\theta_\ell}(X_l)\big). \end{aligned}$$

By A1 and part (a), we have

$$\sup_{n\ge1}\sup_{j\ge0} \mathbb{E}[|P_{\theta_j} g_{\theta_j}(X_j) - P_{\theta_{j+n}} g_{\theta_{j+n}}(X_{j+n})|^p] \le C|h|_{V^\beta}.$$

By Burkholder's inequality and some standard inequalities,

$$\begin{aligned} \mathbb{E}\left[\left|\sum_{\ell=j+1}^{j+n} G_{\theta_{\ell-1}}(X_{\ell-1}, X_\ell)\right|^p\right] &\le C\left\{\sum_{\ell=j+1}^{j+n} \mathbb{E}^{1\wedge(2/p)}(|G_{\theta_{\ell-1}}(X_{\ell-1}, X_\ell)|^p)\right\}^{1\vee(p/2)} \\ &\le C|h|_{V^\beta} n^{1\vee(p/2)}. \end{aligned}$$

Part (b) and A2 together give

$$\begin{aligned} &\mathbb{E}\left[\left|\sum_{\ell=j+1}^{j+n} g_{\theta_{\ell-1}}(X_\ell) - g_{\theta_\ell}(X_l)\right|^p\right] \\ &\quad\le C|h|_{V^\beta}\mathbb{E}\left[\left(\sum_{\ell=j+1}^{j+n} D_\beta(\theta_{\ell-1},\theta_\ell) V^\beta(X_\ell)\right)^p\right] \\ &\quad\le C|h|_{V^\beta}\mathbb{E}\left[\left(\sum_{\ell=j+1}^{j+n} \gamma_{k+\ell} V^{\beta+\eta}(X_\ell)\right)^p\right] \le C|h|_{V^\beta}\left(\sum_{\ell=j+1}^{j+n} \gamma_{k+\ell}\right)^p \end{aligned}$$

and, since $\gamma_n = O(n^{-1/2})$, we are done. □

Acknowledgments. The author is grateful to Galin Jones for helpful discussions, and to Prosper Dovonon for pointing out some of the references in the econometrics literature and for helpful comments on an earlier version of this paper.

SUPPLEMENTARY MATERIAL

Supplement to "Kernel estimators of asymptotic variance for adaptive Markov chain Monte Carlo" (DOI: 10.1214/10-AOS828SUPP; .pdf). The proofs of Theorems 4.1–4.3 require some technical and lengthy arguments that we develop in this supplement.

REFERENCES

[1] ANDREWS, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59** 817–858. MR1106513

[2] ANDRIEU, C. and MOULINES, É. (2006). On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Probab.* **16** 1462–1505. MR2260070

[3] ANDRIEU, C. and THOMS, J. (2008). A tutorial on adaptive MCMC. *Statist. Comput.* **18** 343–373.

[4] ATCHADE, Y. F. (2006). An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift. *Methodol. Comput. Appl. Probab.* **8** 235–254. MR2324873

[5] ATCHADÉ, Y. F. (2010). Supplement to "Kernel estimators of asymptotic variance for adaptive Markov chain Monte Carlo." DOI:10.1214/10-AOS828SUPP.

[6] ATCHADE, Y. F. and FORT, G. (2009). Limit theorems for some adaptive MCMC algorithms with subgeometric kernels: Part II. Technical report, Univ. Michigan.

[7] ATCHADE, Y. and FORT, G. (2010). Limit theorems for some adaptive MCMC algorithms with sub-geometric kernels. *Bernoulli* **16** 116–154. MR2648752

[8] ATCHADE, Y. F., FORT, G., MOULINES, E. and PRIOURET, P. (2009). Adaptive Markov chain Monte Carlo: Theory and methods. Technical report, Univ. Michigan.

[9] ATCHADE, Y. F. and ROSENTHAL, J. S. (2005). On adaptive Markov chain Monte Carlo algorithm. *Bernoulli* **11** 815–828. MR2172842

[10] BAXENDALE, P. H. (2005). Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *Ann. Appl. Probab.* **15** 700–738. MR2114987

[11] BOLLERSLEV, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econometrics* **31** 307–327. MR0853051

[12] BRATLEY, P., FOX, B. and SCHRAGE, L. (1987). *A Guide to Simulation*, 2nd ed. Springer, New York.

[13] DAMERDJI, H. (1995). Mean-square consistency of the variance estimator in steady-state simulation output analysis. *Oper. Res.* **43** 282–291. MR1327416

[14] DE JONG, R. M. (2000). A strong consistency proof for heteroskedasticity and autocorrelation consistent covariance matrix estimators. *Econometric Theory* **16** 262–268. MR1763435

[15] DE JONG, R. M. and DAVIDSON, J. (2000). Consistency of kernel estimators of heteroscedastic and autocorrelated covariance matrices. *Econometrica* **68** 407–423. MR1748008

[16] FLEGAL, J. M. and JONES, G. L. (2009). Batch means and spectral variance estimators in Markov chain Monte Carlo. Available at http://www.citebase.org/abstract?id=oai:arXiv.org:0811.1729. MR2604704

[17] HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242. MR1828504

[18] HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic Press, New York. MR0624435

[19] HANSEN, B. E. (1992). Consistent covariance matrix estimation for dependent heterogeneous processes. *Econometrica* **60** 967–972. MR1168743

[20] JARNER, S. F. and HANSEN, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stochahstic Process. Appl.* **85** 341–361. MR1731030

[21] MEITZ, M. and SAIKKONEN, P. (2008). Ergodicity, mixing, and existence of moments of a class of Markov models with applications to GARCH and ACD models. *Econometric Theory* **24** 1291–1320. MR2440741

[22] MEYN, S. P. and TWEEDIE, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer, London. MR1287609

[23] MICHIE, D., SPIEGELHALTER, D. and TAYLOR, C. (1994). *Machine Learning, Neural and Statistical Classification*. Prentice Hall, Upper Saddle River, NJ.

[24] MYKLAND, P., TIERNEY, L. and YU, B. (1995). Regeneration in Markov chain samplers. *J. Amer. Statist. Assoc.* **90** 233–241. MR1325131

[25] NEWEY, W. K. and WEST, K. D. (1994). Automatic lag selection in covariance matrix estimation. *Rev. Econom. Stud.* **61** 631–653. MR1299308

[26] PRIESTLEY, M. B. (1981). *Spectral Analysis and Time Series: Volume 1: Univariate Series*. Academic Press, London. MR0628735

[27] ROBERTS, G. and ROSENTHAL, J. (2009). Examples of adaptive MCMC. *J. Comput. Graph. Statist.* **18** 349–367.

[28] SAKSMAN, E. and VIHOLA, M. (2009). On the ergodicity of the adaptive Metropolis algorithm on unbounded domains. Technical report. Available at arXiv:0806.2933v2.

[29] WU, W. B. and SHAO, X. (2007). A limit theorem for quadratic forms and its applications. *Econometric Theory* **23** 930–951. MR2396738

DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
1085 S. UNIVERSITY AVENUE
ANN ARBOR, MICHIGAN 48109
USA
E-MAIL: yvesa@umich.edu