# Asymptotic variance estimation for Adaptive MCMC

Yves Atchadé

Department of Statistics, University of Michigan
Work partly supported by NSF grant DMS 0906631.

Based in part on collaborative work with Matias Cattaneo, University of Michigan.

January 4, 2011

# Outline of the talk

- Variance estimation is an important aspect of Monte Carlo simulation.

$$\sigma^2(h) = \lim_{n\to\infty} \mathrm{Var}\left(n^{-1/2}\sum_{k=1}^{n} h(X_k)\right) = \mathrm{Var}_\pi\left(h(X_0)\right) + \sum_{j\geq 1} \mathrm{Cov}_\pi\left(h(X_0), h(X_j)\right),$$

  for Markov chains.

- We focus on the so-called lag-window estimator or kernel estimators:

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

- For various class of processes, including Markov chains, It is known that when $nb_n \to \infty$, under regularity conditions, $\Gamma_n^2(h) \to \sigma^2(h)$.

# Outline of the talk

- Variance estimation is an important aspect of Monte Carlo simulation.

$$\sigma^2(h) = \lim_{n\to\infty} \operatorname{Var}\left(n^{-1/2}\sum_{k=1}^{n} h(X_k)\right)$$
$$= \operatorname{Var}_\pi\left(h(X_0)\right) + \sum_{j\geq 1} \operatorname{Cov}_\pi\left(h(X_0), h(X_j)\right),$$

  for Markov chains.

- We focus on the so-called lag-window estimator or kernel estimators:

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

- For various class of processes, including Markov chains, It is known that when $nb_n \to \infty$, under regularity conditions, $\Gamma_n^2(h) \to \sigma^2(h)$.

# Outline of the talk

- Variance estimation is an important aspect of Monte Carlo simulation.

$$\sigma^2(h) = \lim_{n\to\infty} \mathrm{Var}\left(n^{-1/2}\sum_{k=1}^{n} h(X_k)\right)$$
$$= \mathrm{Var}_\pi\left(h(X_0)\right) + \sum_{j\geq 1} \mathrm{Cov}_\pi\left(h(X_0), h(X_j)\right),$$

  for Markov chains.

- We focus on the so-called lag-window estimator or kernel estimators:

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

- For various class of processes, including Markov chains, It is known that when $nb_n \to \infty$, under regularity conditions, $\Gamma_n^2(h) \to \sigma^2(h)$.

# Outline of the talk

- For AMCMC, the behavior of $\Gamma_n^2(h)$ is similar. But we have to be careful when the limiting distribution in the CLT is a mixture of Gaussians.
- The analysis of $\Gamma_n^2(h)$ is related to the asymptotics of quadratic forms of Markov chains

$$\sum_{1 \le \ell \le j \le n} w_n(\ell, j) h_n(X_\ell, X_j),$$

  which has many applications in nonparametric time-series.
- We will also introduce a "small bandwidth" version of $\Gamma_n^2(h)$ where $b_n = 1/n$. $\Gamma_n^2(h)$ is then inconsistent, but as we will see, valid and much improved confidence intervals can still constructed.

# Outline of the talk

- For AMCMC, the behavior of $\Gamma_n^2(h)$ is similar. But we have to be careful when the limiting distribution in the CLT is a mixture of Gaussians.
- The analysis of $\Gamma_n^2(h)$ is related to the asymptotics of quadratic forms of Markov chains

$$\sum_{1\leq\ell\leq j\leq n} w_n(\ell,j)h_n(X_\ell,X_j),$$

which has many applications in nonparametric time-series.
- We will also introduce a "small bandwidth" version of $\Gamma_n^2(h)$ where $b_n = 1/n$. $\Gamma_n^2(h)$ is then inconsistent, but as we will see, valid and much improved confidence intervals can still constructed.

# Outline of the talk

- For AMCMC, the behavior of $\Gamma_n^2(h)$ is similar. But we have to be careful when the limiting distribution in the CLT is a mixture of Gaussians.
- The analysis of $\Gamma_n^2(h)$ is related to the asymptotics of quadratic forms of Markov chains

$$\sum_{1 \leq \ell \leq j \leq n} w_n(\ell, j) h_n(X_\ell, X_j),$$

which has many applications in nonparametric time-series.

- We will also introduce a "small bandwidth" version of $\Gamma_n^2(h)$ where $b_n = 1/n$. $\Gamma_n^2(h)$ is then inconsistent, but as we will see, valid and much improved confidence intervals can still constructed.

# Asymptotic variance estimation for MCMC

- Let $\{X_n,\ n \geq 0\}$ be a Markov chain with invariant distribution $\pi$ and transition kernel $P$.
- If $P^n$ converges to $\pi$ fast enough and under appropriate moment condition on $h$, then

$$\sigma^2(h) = \mathrm{Var}_\pi\left(h(X_0)\right) + 2\sum_{j\geq 1} \mathrm{Cov}_\pi\left(h(X_0), h(X_j)\right) < \infty,$$

  and

$$n^{-1/2}\sum_{i=1}^{n}\left(h(X_i) - \pi(h)\right) \xrightarrow{w} N(0, \sigma^2(h)),$$

- We call $\sigma^2(h)$ the asymptotic variance of $h$.

# Asymptotic variance estimation for MCMC

- Let $\{X_n,\ n \geq 0\}$ be a Markov chain with invariant distribution $\pi$ and transition kernel $P$.
- If $P^n$ converges to $\pi$ fast enough and under appropriate moment condition on $h$, then

$$\sigma^2(h) = \mathrm{Var}_\pi\left(h(X_0)\right) + 2\sum_{j\geq 1}\mathrm{Cov}_\pi\left(h(X_0), h(X_j)\right) < \infty,$$

and

$$n^{-1/2}\sum_{i=1}^{n}\left(h(X_i) - \pi(h)\right) \xrightarrow{w} N(0, \sigma^2(h)),$$

- We call $\sigma^2(h)$ the asymptotic variance of $h$.

# Asymptotic variance estimation for MCMC

- Let $\{X_n,\ n \geq 0\}$ be a Markov chain with invariant distribution $\pi$ and transition kernel $P$.
- If $P^n$ converges to $\pi$ fast enough and under appropriate moment condition on $h$, then

$$\sigma^2(h) = \mathrm{Var}_\pi\left(h(X_0)\right) + 2\sum_{j\geq 1}\mathrm{Cov}_\pi\left(h(X_0), h(X_j)\right) < \infty,$$

  and

$$n^{-1/2}\sum_{i=1}^{n}\left(h(X_i) - \pi(h)\right) \xrightarrow{w} N(0, \sigma^2(h)),$$

- We call $\sigma^2(h)$ the asymptotic variance of $h$.

# Asymptotic variance estimation for MCMC

- There are many Markov chain CLT results.
- See example Meyn & Tweedie (2009), Maxwell & Woodroofe (AP 2000).
- G. Jones (EJP 2006) gathers together many sets of assumptions under which the Markov chain CLT holds.

# Asymptotic variance estimation for MCMC

- The CLT naturally leads to an asymptotically valid confidence interval for $\pi(h)$:

$$\hat{\pi}_n(h) \pm z \frac{\hat{\sigma}_n(h)}{\sqrt{n}},$$

  for appropriate Gaussian quantile $z$, provided we have a consistent estimate $\hat{\sigma}_n(h)$ of $\sigma(h)$.

# Asymptotic variance estimation for MCMC

- There are many such estimators: Batch Means, overlapping Batch Means, regenerative simulations ( G. Jones et al. (2009), A. Tan (2009)). A popular estimator is the lag-window/kernel estimator,

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

- $\gamma_n(k) = n^{-1} \sum_{j=1}^{n-k} (h(X_j) - \pi_n(h)) (h(X_{j+k}) - \pi_n(h))$, $\gamma_n(-k) = \gamma_n(k)$, and $b_n^{-1}$ is the truncation point. I'll refer to $b_n$ as "bandwidth".
- $w : \mathbb{R} \to [0, 1]$ has support $[-1, 1]$, $w(-x) = w(x)$, $w(1) = 0$, and $w(0) = 1$.
- 

$$\Gamma_n^2(h) = \gamma_n(0) + 2 \sum_{k=1}^{b_n^{-1}-1} w(kb_n)\gamma_n(k).$$

# Asymptotic variance estimation for MCMC

- There are many such estimators: Batch Means, overlapping Batch Means, regenerative simulations ( G. Jones et al. (2009), A. Tan (2009)). A popular estimator is the lag-window/kernel estimator,

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

- $\gamma_n(k) = n^{-1} \sum_{j=1}^{n-k} (h(X_j) - \pi_n(h)) (h(X_{j+k}) - \pi_n(h))$, $\gamma_n(-k) = \gamma_n(k)$, and $b_n^{-1}$ is the truncation point. I'll refer to $b_n$ as "bandwidth".

- $w : \mathbb{R} \to [0, 1]$ has support $[-1, 1]$, $w(-x) = w(x)$, $w(1) = 0$, and $w(0) = 1$.

-

$$\Gamma_n^2(h) = \gamma_n(0) + 2 \sum_{k=1}^{b_n^{-1}-1} w(kb_n)\gamma_n(k).$$

# Asymptotic variance estimation for MCMC

- There are many such estimators: Batch Means, overlapping Batch Means, regenerative simulations ( G. Jones et al. (2009), A. Tan (2009)). A popular estimator is the lag-window/kernel estimator,

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

- $\gamma_n(k) = n^{-1}\sum_{j=1}^{n-k}\left(h(X_j) - \pi_n(h)\right)\left(h(X_{j+k}) - \pi_n(h)\right)$, $\gamma_n(-k) = \gamma_n(k)$, and $b_n^{-1}$ is the truncation point. I'll refer to $b_n$ as "bandwidth".
- $w : \mathbb{R} \to [0, 1]$ has support $[-1, 1]$, $w(-x) = w(x)$, $w(1) = 0$, and $w(0) = 1$.
-

$$\Gamma_n^2(h) = \gamma_n(0) + 2\sum_{k=1}^{b_n^{-1}-1} w(kb_n)\gamma_n(k).$$

# Asymptotic variance estimation for MCMC

- There are many such estimators: Batch Means, overlapping Batch Means, regenerative simulations ( G. Jones et al. (2009), A. Tan (2009)). A popular estimator is the lag-window/kernel estimator,

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

- $\gamma_n(k) = n^{-1} \sum_{j=1}^{n-k} (h(X_j) - \pi_n(h)) (h(X_{j+k}) - \pi_n(h))$, $\gamma_n(-k) = \gamma_n(k)$, and $b_n^{-1}$ is the truncation point. I'll refer to $b_n$ as "bandwidth".
- $w : \mathbb{R} \to [0,1]$ has support $[-1,1]$, $w(-x) = w(x)$, $w(1) = 0$, and $w(0) = 1$.
-

$$\Gamma_n^2(h) = \gamma_n(0) + 2 \sum_{k=1}^{b_n^{-1}-1} w(kb_n)\gamma_n(k).$$

# Asymptotic variance estimation for MCMC

- The kernel estimator is well known estimator with a rich literature. Parzen (1960), Priestley (1981), Brockwell & Davis (1991)).
- A large literature in Econometrics as well: Andrews (1991); Hansen (1992); de Jong & Davidson (2000) and their references.
- Operation research: H. Damerdji (1991,1994).
- Most of the work impose some fairly strong mixing conditions.
- Jones et al. (2006, 2009) studies the Markov chain case where $P$ is geometrically ergodic.

# Asymptotic variance estimation for MCMC

- The kernel estimator is well known estimator with a rich literature. Parzen (1960), Priestley (1981), Brockwell & Davis (1991)).
- A large literature in Econometrics as well: Andrews (1991); Hansen (1992); de Jong & Davidson (2000) and their references.
- Operation research: H. Damerdji (1991,1994).
- Most of the work impose some fairly strong mixing conditions.
- Jones et al. (2006, 2009) studies the Markov chain case where $P$ is geometrically ergodic.

# Asymptotic variance estimation for MCMC

- The kernel estimator is well known estimator with a rich literature. Parzen (1960), Priestley (1981), Brockwell & Davis (1991)).
- A large literature in Econometrics as well: Andrews (1991); Hansen (1992); de Jong & Davidson (2000) and their references.
- Operation research: H. Damerdji (1991,1994).
- Most of the work impose some fairly strong mixing conditions.
- Jones et al. (2006, 2009) studies the Markov chain case where $P$ is geometrically ergodic.

# Asymptotic variance estimation for MCMC

- The kernel estimator is well known estimator with a rich literature. Parzen (1960), Priestley (1981), Brockwell & Davis (1991)).
- A large literature in Econometrics as well: Andrews (1991); Hansen (1992); de Jong & Davidson (2000) and their references.
- Operation research: H. Damerdji (1991,1994).
- Most of the work impose some fairly strong mixing conditions.
- Jones et al. (2006, 2009) studies the Markov chain case where $P$ is geometrically ergodic.

# Asymptotic variance estimation for AMCMC

We consider the adaptive MCMC case.

- Let $\{P_\theta, \ \theta \in \Theta\}$ be a family of TK. $P_\theta$ is inv. wrt $\pi$ for any $\theta \in \Theta$.
- Define $\{(X_n, \theta_n), \ n \geq 0\}$ a stoch. process s.t.

$$\theta_n \in \mathcal{F}_n, \quad \mathbb{P}\left(X_n \in A | \mathcal{F}_{n-1}\right) = P_{\theta_{n-1}}(X_{n-1}, A).$$

- $\{X_n, \ n \geq 0\}$ is an adaptive Markov chain (no longer Markov).

# Asymptotic variance estimation for AMCMC

We consider the adaptive MCMC case.

- Let $\{P_\theta, \ \theta \in \Theta\}$ be a family of TK. $P_\theta$ is inv. wrt $\pi$ for any $\theta \in \Theta$.
- Define $\{(X_n, \theta_n), \ n \geq 0\}$ a stoch. process s.t.

$$\theta_n \in \mathcal{F}_n, \quad \mathbb{P}\left(X_n \in A | \mathcal{F}_{n-1}\right) = P_{\theta_{n-1}}(X_{n-1}, A).$$

- $\{X_n, \ n \geq 0\}$ is an adaptive Markov chain (no longer Markov).

# Asymptotic variance estimation for AMCMC

We consider the adaptive MCMC case.

- Let $\{P_\theta, \ \theta \in \Theta\}$ be a family of TK. $P_\theta$ is inv. wrt $\pi$ for any $\theta \in \Theta$.
- Define $\{(X_n, \theta_n), \ n \geq 0\}$ a stoch. process s.t.

$$\theta_n \in \mathcal{F}_n, \quad \mathbb{P}\left(X_n \in A | \mathcal{F}_{n-1}\right) = P_{\theta_{n-1}}(X_{n-1}, A).$$

- $\{X_n, \ n \geq 0\}$ is an adaptive Markov chain (no longer Markov).

# Central limit theorem

- Throughout, we assume diminishing adaptation:

$$D(\theta_n, \theta_{n-1}) \leq \gamma_n, \quad \text{where } \gamma_n \to 0.$$

- What can we say about $\Gamma_n^2(h)$ when $\{X_n,\ n \geq 0\}$ is an adaptive MCMC process?
- We take a very familiar and standard approach to the question.

# Central limit theorem

- Throughout, we assume diminishing adaptation:

$$D(\theta_n, \theta_{n-1}) \leq \gamma_n, \quad \text{where } \gamma_n \to 0.$$

- What can we say about $\Gamma_n^2(h)$ when $\{X_n,\ n \geq 0\}$ is an adaptive MCMC process?
- We take a very familiar and standard approach to the question.

# Central limit theorem

- Throughout, we assume diminishing adaptation:

$$D(\theta_n, \theta_{n-1}) \leq \gamma_n, \quad \text{where } \gamma_n \to 0.$$

- What can we say about $\Gamma_n^2(h)$ when $\{X_n,\ n \geq 0\}$ is an adaptive MCMC process?
- We take a very familiar and standard approach to the question.

# Central limit theorem

- Assume $\pi(h) = 0$ and for simplicity assume that $\gamma_n(k) = n^{-1} \sum_{j=1}^{n-k} h(X_j)h(X_{j+k})$. Define

$$g_\theta(x) := \sum_{j \geq 0} P_\theta^j h(x), \quad G_\theta(x, y) = g_\theta(y) - P_\theta g_\theta(x).$$

- Define $G_k = G_{\theta_{k-}}(X_{k-1}, X_k)$. $\mathbb{E}\left(G_k | \mathcal{F}_{k-1}\right) = 0$. We know from Andrieu & Moulines (06) that:

$$\sum_{k=1}^{n} h(X_k) = \sum_{k=1}^{n} G_k + R_n,$$

where $R_n = o_P(n^{1/2})$.

- The same martingale difference sequence $\{G_k\}$ can be used to decompose $\Gamma_n^2(h)$.

# Central limit theorem

- Assume $\pi(h) = 0$ and for simplicity assume that $\gamma_n(k) = n^{-1} \sum_{j=1}^{n-k} h(X_j)h(X_{j+k})$. Define

$$g_\theta(x) := \sum_{j \geq 0} P_\theta^j h(x), \quad G_\theta(x, y) = g_\theta(y) - P_\theta g_\theta(x).$$

- Define $G_k = G_{\theta_{k-}}(X_{k-1}, X_k)$. $\mathbb{E}\left(G_k | \mathcal{F}_{k-1}\right) = 0$. We know from Andrieu & Moulines (06) that:

$$\sum_{k=1}^{n} h(X_k) = \sum_{k=1}^{n} G_k + R_n,$$

where $R_n = o_P(n^{1/2})$.

- The same martingale difference sequence $\{G_k\}$ can be used to decompose $\Gamma_n^2(h)$.

# Central limit theorem

- Assume $\pi(h) = 0$ and for simplicity assume that $\gamma_n(k) = n^{-1} \sum_{j=1}^{n-k} h(X_j)h(X_{j+k})$. Define

$$g_\theta(x) := \sum_{j \geq 0} P_\theta^j h(x), \quad G_\theta(x, y) = g_\theta(y) - P_\theta g_\theta(x).$$

- Define $G_k = G_{\theta_{k-}}(X_{k-1}, X_k)$. $\mathbb{E}\left(G_k | \mathcal{F}_{k-1}\right) = 0$. We know from Andrieu & Moulines (06) that:

$$\sum_{k=1}^{n} h(X_k) = \sum_{k=1}^{n} G_k + R_n,$$

where $R_n = o_P(n^{1/2})$.

- The same martingale difference sequence $\{G_k\}$ can be used to decompose $\Gamma_n^2(h)$.

# Ingredients of the Proof

## Theorem

*Under appropriate moments and ergodicity conditions,*

$$\sum_{\ell=1}^{n} h(X_\ell) = \sum_{\ell=1}^{n} G_\ell + R_n$$

$$\Gamma_n^2(h) = n^{-1} \sum_{\ell=1}^{n} G_\ell^2 + \frac{2}{n} \sum_{\ell=1}^{n} \sum_{j=1}^{\ell-1} w\left((\ell - j) b_n\right) G_\ell G_j + \zeta_n.$$

*where for* $p > 1$*,*

$$\mathbb{E}\left(|\zeta_n|^p\right) \leq C \left(n^{-1} + b_n + b_n^{1-\frac{1}{2} \vee \frac{1}{p}}\right)^p, \quad \textit{and}$$

$$\mathbb{E}\left(\left|\frac{1}{n} \sum_{\ell=1}^{n} \sum_{j=1}^{\ell-1} w\left((\ell - j) b_n\right) G_\ell G_j\right|^p\right) \leq C \left(n b_n\right)^{-\frac{p}{2}} n^{-\frac{p}{2}+1 \vee \frac{p}{2}}.$$

# Asymptotic variance estimation

## Corollary

$$\lim_{n\to\infty}\left(\Gamma_n^2(h)-\frac{1}{n}\sum_{k=1}^{n}G^2_{\theta_{k-1}}\left(X_{k-1},X_k\right)\right)=0,\quad \textit{in probab.},\qquad (1)$$

*provided* $nb_n \to \infty$*.*

The main condition is:

$$\sup_{\theta\in\Theta}|g_\theta(x)|\le W(x),\quad \text{where}\quad \sup_{n\ge 0}\mathbb{E}\left(W^{2p}(X_n)\right)<\infty.$$

# Few remarks

- The results says nothing about the CLT for the partial sum $\sum_{k=1}^{n} h(X_k)$.
- When $\frac{1}{n}\sum_{k=1}^{n} G^2_{\theta_{k-1}}(X_{k-1}, X_k)$ converges to a deterministic limit $\sigma^2(h)$, say, then $\Gamma^2_n(h)$, also converges to $\sigma^2(h)$.
- Of course, under our moments conditions, this implies that $n^{-1/2}\sum_{k=1}^{n} h(X_k) \xrightarrow{w} \mathcal{N}(0, \sigma^2(h))$.
- Same as with Markov chains.

# Few remarks

- The results says nothing about the CLT for the partial sum $\sum_{k=1}^{n} h(X_k)$.
- When $\frac{1}{n}\sum_{k=1}^{n} G^2_{\theta_{k-1}}(X_{k-1}, X_k)$ converges to a deterministic limit $\sigma^2(h)$, say, then $\Gamma_n^2(h)$, also converges to $\sigma^2(h)$.
- Of course, under our moments conditions, this implies that $n^{-1/2}\sum_{k=1}^{n} h(X_k) \xrightarrow{w} \mathcal{N}(0, \sigma^2(h))$.
- Same as with Markov chains.

# Few remarks

- The results says nothing about the CLT for the partial sum $\sum_{k=1}^{n} h(X_k)$.
- When $\frac{1}{n}\sum_{k=1}^{n} G_{\theta_{k-1}}^2(X_{k-1}, X_k)$ converges to a deterministic limit $\sigma^2(h)$, say, then $\Gamma_n^2(h)$, also converges to $\sigma^2(h)$.
- Of course, under our moments conditions, this implies that $n^{-1/2}\sum_{k=1}^{n} h(X_k) \xrightarrow{w} \mathcal{N}(0, \sigma^2(h))$.
- Same as with Markov chains.

# Few remarks

- But if the average square variation $n^{-1}\sum_{k=1}^{n} G^2_{\theta_{k-1}}(X_{k-1}, X_k)$ converges n probability to a stochastic limit $\Gamma^2(h)$, then of course $\Gamma^2_n(h)$ also converges in probability to $\Gamma^2(h)$ and

$$n^{-1/2}\sum_{k=1}^{n} h(X_k) \xrightarrow{w} \sqrt{\Gamma^2(h)}Z,$$

  where $Z \sim N(0,1)$ independent of $\Gamma^2(h)$ (Hall & Heyde (80), Theorem 3.2).

- $\sqrt{\Gamma^2(h)}Z$ is a mixture of Gaussians.

- But

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

  Thus is this case, $\Gamma^2_n(h)$ is not consistent for the asymptotic variance $\sigma^2(h)$.

# Few remarks

- But if the average square variation $n^{-1}\sum_{k=1}^{n} G^2_{\theta_{k-1}}(X_{k-1}, X_k)$ converges n probability to a stochastic limit $\Gamma^2(h)$, then of course $\Gamma_n^2(h)$ also converges in probability to $\Gamma^2(h)$ and

$$n^{-1/2}\sum_{k=1}^{n} h(X_k) \xrightarrow{w} \sqrt{\Gamma^2(h)}Z,$$

  where $Z \sim N(0,1)$ independent of $\Gamma^2(h)$ (Hall & Heyde (80), Theorem 3.2).
- $\sqrt{\Gamma^2(h)}Z$ is a mixture of Gaussians.
- But

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

  Thus is this case, $\Gamma_n^2(h)$ is not consistent for the asymptotic variance $\sigma^2(h)$.

# Choice of the bandwidth

- One of the main limitations of this estimator is the difficulty in choosing the bandwidth $b_n$. Often $b_n = n^{-\rho}$, $\rho \in \{1/2, 2/3\}$ resulting in a slowly converging $\Gamma_n^2(h)$.
- Neave (Annals Math. Stat. 1970) criticized the assumption $nb_n \to \infty$ as "a mathematically convenient assumption to ensure consistency of the estimates, but which is unrealistic when such estimators are used in practice where the value $nb_n$ cannot be zero".
- Another way to look at it is that: killing both the remainder and the off-diagonal term is a difficult exercise.

# Choice of the bandwidth

- One of the main limitations of this estimator is the difficulty in choosing the bandwidth $b_n$. Often $b_n = n^{-\rho}$, $\rho \in \{1/2, 2/3\}$ resulting in a slowly converging $\Gamma_n^2(h)$.
- Neave (Annals Math. Stat. 1970) criticized the assumption $nb_n \to \infty$ as "a mathematically convenient assumption to ensure consistency of the estimates, but which is unrealistic when such estimators are used in practice where the value $nb_n$ cannot be zero".
- Another way to look at it is that: killing both the remainder and the off-diagonal term is a difficult exercise.

# Small "bandwidth" asymptotics

- Neave (1970) suggested using $b_n = n^{-1}$ which is known to give an inconsistent estimator but still confidence intervals can be constructed.
- Kiefer & Vogelsang (2002,2005,2009) have recently further developed the idea in the Econometrics literature under some restrictive model assumptions.
- We extend the approach in the context of Markov Chains. Thus in the sequel $\{X_n,\ n \geq 0\}$ is a Markov chain with invariant distribution $\pi$ and transition kernel $P$.

# Small "bandwidth" asymptotics

- Neave (1970) suggested using $b_n = n^{-1}$ which is known to give an inconsistent estimator but still confidence intervals can be constructed.
- Kiefer & Vogelsang (2002,2005,2009) have recently further developed the idea in the Econometrics literature under some restrictive model assumptions.
- We extend the approach in the context of Markov Chains. Thus in the sequel $\{X_n,\ n \geq 0\}$ is a Markov chain with invariant distribution $\pi$ and transition kernel $P$.

# Small "bandwidth" asymptotics

- Neave (1970) suggested using $b_n = n^{-1}$ which is known to give an inconsistent estimator but still confidence intervals can be constructed.
- Kiefer & Vogelsang (2002,2005,2009) have recently further developed the idea in the Econometrics literature under some restrictive model assumptions.
- We extend the approach in the context of Markov Chains. Thus in the sequel $\{X_n, \ n \geq 0\}$ is a Markov chain with invariant distribution $\pi$ and transition kernel $P$.

# Small "bandwidth" asymptotics

- We saw above that:

$$\sum_{\ell=1}^{n} h(X_\ell) = \sum_{\ell=1}^{n} G_\ell + R_n$$

$$\Gamma_n^2(h) = \frac{1}{n}\sum_{\ell=1}^{n}\sum_{j=1}^{n} w((\ell-j)b_n)G_\ell G_j + \zeta_n.$$
$$= \frac{1}{n}\sum_{\ell=1}^{n} G_\ell \sum_{j=0}^{n-1} w((\ell-1-j)b_n)G_{j+1} + \zeta_n.$$

- Define $W_{n,\ell} = \frac{G_\ell}{\sqrt{n}\sigma(h)}$, where $\sigma^2(h) = \mathbb{E}_\pi(G^2(X_0, X_1))$, the asymp. variance. And

$$B_n(t) = \sum_{\ell=1}^{\lfloor nt \rfloor} W_{n,\ell}, \quad 0 \le t \le 1$$

# Small "bandwidth" asymptotics

- We saw above that:

$$\sum_{\ell=1}^{n} h(X_\ell) = \sum_{\ell=1}^{n} G_\ell + R_n$$

$$\Gamma_n^2(h) = \frac{1}{n}\sum_{\ell=1}^{n}\sum_{j=1}^{n} w((\ell - j)b_n)G_\ell G_j + \zeta_n.$$

$$= \frac{1}{n}\sum_{\ell=1}^{n} G_\ell \sum_{j=0}^{n-1} w((\ell - 1 - j)b_n)G_{j+1} + \zeta_n.$$

- Define $W_{n,\ell} = \frac{G_\ell}{\sqrt{n}\sigma(h)}$, where $\sigma^2(h) = \mathbb{E}_\pi(G^2(X_0, X_1))$, the asymp. variance. And

$$B_n(t) = \sum_{\ell=1}^{\lfloor nt \rfloor} W_{n,\ell}, \quad 0 \le t \le 1$$

# Small "bandwidth" asymptotics

Then

$$\sum_{\ell=1}^{n} \frac{h(X_\ell)}{\sqrt{n}\sigma(h)} = \sum_{\ell=1}^{n} W_{n,\ell} + \tilde{R}_n = B_n(1) + \tilde{R}_n,$$

and for $b_n = n^{-1}$,

$$\sum_{j=0}^{n-1} w((\ell - 1 - j)b_n) \frac{G_{j+1}}{\sigma(h)\sqrt{n}} = Z_n\left((\ell - 1)b_n\right).$$

where

$$Z_n(t) = \int_0^1 w(t-u) dB_n(u).$$

# Small "bandwidth" asymptotics

Thus with $b_n = n^{-1}$,

$$\sum_{\ell=1}^{n} \frac{h(X_\ell)}{\sqrt{n}\sigma(h)} = B_n(1) + \tilde{R}_n$$

$$\Gamma_n^2(h) = \sigma^2(h)\sigma^2(h) \int_0^1 Z_n(t) dB_n(t) + \zeta_n.$$

- As $n \to \infty$, $\{B_n(t),\ 0 \leq t \leq 1\}$ converges weakly to $\{B(t),\ 0 \leq t \leq 1\}$ the standard Brownian motion.
- Then we use Kurtz & Protter (AP 1992) on the weak convergence of stochastic integrals to get.

# Small "bandwidth" asymptotics

Thus with $b_n = n^{-1}$,

$$\sum_{\ell=1}^{n} \frac{h(X_\ell)}{\sqrt{n}\sigma(h)} = B_n(1) + \tilde{R}_n$$

$$\Gamma_n^2(h) = \sigma^2(h)\sigma^2(h) \int_0^1 Z_n(t) dB_n(t) + \zeta_n.$$

- As $n \to \infty$, $\{B_n(t),\ 0 \leq t \leq 1\}$ converges weakly to $\{B(t),\ 0 \leq t \leq 1\}$ the standard Brownian motion.
- Then we use Kurtz & Protter (AP 1992) on the weak convergence of stochastic integrals to get.

# Small "bandwidth" asymptotics

## Theorem

*Assume* $\{X_n,\ n \geq 0\}$ *is a Markov chain. Let* $\{B(t),\ 0 \leq t \leq 1\}$ *be the standard Brownian motion. If* $b_n = n^{-1}$, *then*

$$\Gamma_n^2(h) \xrightarrow{w} \sigma^2(h) \int_0^1 \int_0^1 w(t-s)dB(s)dB(t).$$

*Furthermore, assuming* $\Gamma_n^2(h) > 0$ *almost surely,*

$$\frac{n^{-1/2}\sum_{j=1}^n h(X_j)}{\sqrt{\Gamma_n^2(h)}} \xrightarrow{w} \frac{B(1)}{\sqrt{\int_0^1 \int_0^1 w(t-s)dB(s)dB(t)}}.$$

# Small "bandwidth" asymptotics

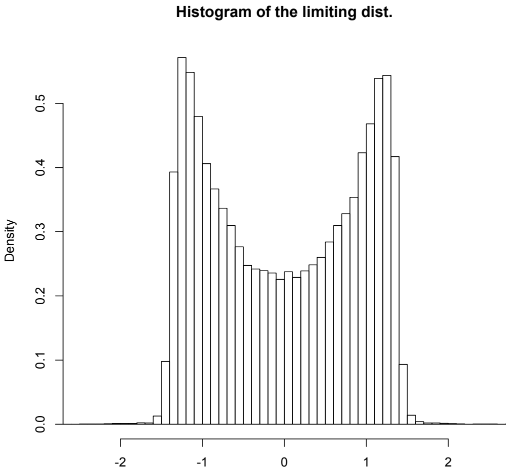
- We can then still construct a valid confidence interval for 0:

$$n^{-1}\sum_{\ell=1}^{n} h(X_\ell) \pm \bar{z}\frac{\sqrt{\Gamma_n^2(h)}}{\sqrt{n}}.$$

- $\bar{z}$ is the appropriate quantile of the distribution of $\frac{B(1)}{\sqrt{\int_0^1\int_0^1 w(t-s)dB(s)dB(t)}}$.
- We approximate this by Monte Carlo.

# Small "bandwidth" asymptotics

For the Parzen kernel, the distribution of $\frac{B(1)}{\sqrt{\int_0^1 \int_0^1 w(t-s)dB(s)dB(t)}}$.



# Small "bandwidth" asymptotics

- The 97.5% quantile is estimated at 1.360.
- When $\pi(h) \neq 0$ and we use $\gamma_{n,k} = n^{-1} \sum_{j=1}^{n-k} (h(X_j) - \hat{\pi}_n(h)) (h(X_{j+k}) - \hat{\pi}_n(h))$ instead of $\gamma_n(k) = n^{-1} \sum_{j=1}^{n-k} h(X_j)h(X_{j+k})$, a similar result holds.
- But the limiting distribution is different from $\frac{B(1)}{\sqrt{\int_0^1 \int_0^1 w(t-s)dB(s)dB(t)}}$.

# $AR(1)$ example

- For $\rho = 0.95$ and $\{\epsilon_n, \ n \geq 1\}$ is an iid $N(0,1)$

$$X_0 = 0, \quad \text{and} \quad X_n = \rho X_{n-1} + \epsilon_n$$

.

- For $|\rho| < 1$, the chain is geometrically ergodic with target distribution $N(0,(1-\rho^2)^{-1})$. We estimate $\mu = \int x\pi(x)dx = 0$.
- We build a confidence interval for $\mu$ comparing the regular bandwidth $b_n = n^{-1/2}$ with the small bandwidth $b_n = n^{-1}$.

# $AR(1)$ example

- For $\rho = 0.95$ and $\{\epsilon_n,\ n \geq 1\}$ is an iid $N(0,1)$

$$X_0 = 0, \quad \text{and} \quad X_n = \rho X_{n-1} + \epsilon_n$$

.

- For $|\rho| < 1$, the chain is geometrically ergodic with target distribution $N(0, (1-\rho^2)^{-1})$. We estimate $\mu = \int x\pi(x)dx = 0$.
- We build a confidence interval for $\mu$ comparing the regular bandwidth $b_n = n^{-1/2}$ with the small bandwidth $b_n = n^{-1}$.

# An example

| | $b_n = n^{-1/2}$ | $\mathrm{b}_n = n^{-1}$ |
|---|---|---|
| $N = 1000$ | 0.83% | 0.96 |
| $N = 5000$ | 0.86% | 0.97 |
| $N = 10,000$ | 0.865% | 0.935 |
| $N = 20,000$ | 0.928% | 0.950 |

# An example

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

We assume a Gaussian prior distribution $\pi(\beta) \propto e^{-\|\beta\|^2/c}$. Posterior distribution:

$$\pi\left(\beta|X\right) \propto e^{\ell(\beta|X)} e^{-\|\beta\|^2/c}.$$

# An example

We build a Random Walk Metropolis to sample fom the posterior distribution:

## Algorithm

- *Given* $\beta_n$*: propose* $\beta' \sim N(\beta_n, \Sigma)$*.*
- *Set* $\beta_{n+1} = \beta'$ *with prob.* $\min\left(1, \frac{\tilde{\pi}(\beta'|X)}{\tilde{\pi}(\beta_n|X)}\right)$*. Otherwise, set* $\beta_{n+1} = \beta_n$*.*

We compare two choices of: $\Sigma = \sigma_\star \Sigma_\pi$ and $\Sigma = 0.0005 \times \mathcal{I}^{-1}$.

# An example

| | $b_n = n^{-1/2}$ | $b_n = n^{-1}$ |
|---|---|---|
| $\Sigma = \sigma_\star \Sigma_\pi$ | 0.940 | 0.942 |
| $\Sigma = 0.0005 \times I_d$ | 0.63 | 0.96 |

$N = 20,000$