Bootstrap with Larger Resample Size for Root-n Consistent Density Estimation with Time Series Data

Christopher C. Chang*, Dimitris N. Politis*

1 February 2011

Abstract

We consider finite-order moving average and nonlinear autoregressive processes with no parametric assumption on the error distribution, and present a kernel density estimator of a bootstrap series that estimates their marginal densities root-n consistently. This is equal to the rate of the best known convolution estimators, and faster than the standard kernel density estimator. We also conduct simulations to check the finite sample properties of our estimator, and the results are generally better than corresponding results for the standard kernel density estimator.

Keywords: Kernel function; convolution estimator; nonparametric density estimation; moving-average process; nonlinear autoregressive process.

1 Introduction

A common statistical problem involves estimating an unknown density function $f(x)$ given a limited number of observations $X_1, X_2, \ldots, X_n$ independently drawn from that density. The standard approach today, first suggested by Rosenblatt (1956) and Parzen (1962), is to use a kernel density estimator

$$f(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{1}$$

where K is a nonnegative kernel function and h_n is a bandwidth. With optimal bandwidth determination, this estimator typically has a $n^{-2/5}$ rate of convergence.

Often, e.g. in a time-series setting, independence does not hold. Roussas (1969) and Rosenblatt (1970) were among the first to study the behavior of the kernel estimator under dependence; many later references can be found in Györfi et al. (198) chapter 4 and Fan & Yao (2003) chapter 5.

Recently, methods have been developed to exploit information about the form of dependence to improve density estimates. Saavedra & Cao (1999) introduced a convolution-kernel estimator for the marginal density of a moving average process of order 1 ($Z_t = a_t - \theta a_{t-1}$ with unknown θ), which they proved to have a $n^{-1/2}$ rate of convergence—surprisingly superior to what is achievable in the independent case. Müller et al. (2005) introduced a similar estimator for the innovation density in nonlinear parametric autoregressive models, Schick & Wefelmeyer (2007) (SW, for short) proved root-n consistency of the convolution density estimator for weakly dependent invertible linear processes, and Støve and Tjøstheim (2007) (ST, for short) proved root-n consistency of a convolution estimator for the density in a nonlinear regression model.

This article is concerned with demonstrating that one can get root-n consistent estimation of the marginal density for MA(p) and nonlinear AR(1) time series with a simple kernel density estimator of a bootstrap series, thus bypassing the need for a convolution. Our bootstrap is the usual model-based (semiparametric) residual bootstrap (see e.g. Efron & Tibshirani (1993) or Davison & Hinkley (1997)). Interestingly, and in contrast to some recent work involving bootstraps with smaller resample sizes (e.g. Bretagnolle (1983), Swanepoel (1986), Politis (1993), Datta (1995), Bickel (1997), Politis (1999)), our proposed bootstrap has resample size larger than n by orders of magnitude.

*Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112 USA. E-mails: chc007@math.ucsd.edu, dpolitis@ucsd.edu

The estimator is presented in section 2, and its root-n consistency is first proved in the MA(1) case and then extended to MA(p). An application of the estimator to the nonlinear AR(1) case is presented and analyzed in section 3; simulation results are described in section 4, and a short conclusion is stated in section 5. Appendix A contains all technical assumptions; all proofs are in Appendix B.

2 MA(p) Density Estimation

2.1 MA(1)

Consider a stationary linear process with MA(1) representation

$$X_t = \varepsilon_t + a\varepsilon_{t-1}, \qquad t \in \mathbb{Z}, a \neq 0, |a| < 1, \varepsilon_t \text{ iid with density } f. \tag{2}$$

The density f is assumed to satisfy smoothness conditions to be specified later.

Our objective is to estimate the stationary density h of the X_t's as accurately as possible. A first step toward this is a good estimate $\hat{a}$ of a. The usual choice is the least squares (LS) estimate regressing $X_2, \ldots, X_n$ on $X_1, \ldots, X_{n-1}$, which minimizes $\sum_{j=2}^{n}(\sum_{k=0}^{j-1}(-\hat{a})^k X_{j-k})^2$; this is adequate for our purposes.

To execute the residual bootstrap that is based on the MA model, it is necessary to use $\hat{a}$ to estimate the sequence of residuals, use the estimated sequence to estimate the underlying residual density, and finally, use the density estimate to construct bootstrap replications of the linear process. We address each of these steps in turn.

If we express ε_j in terms of a and the X_is, we get an infinite geometric sum:

$$\begin{aligned}
\varepsilon_j &= X_j - a\varepsilon_{j-1} \\
&= X_j - aX_{j-1} + a^2\varepsilon_{j-2} \\
&= \ldots \\
&= \sum_{k=0}^{\infty}(-a)^k X_{j-k}
\end{aligned}$$

Thus it is necessary to choose a sequence of cutoff values p_n indicating the number of X_i terms we will use in extracting residuals. We use $p_n := \min(1, \lfloor(\log n)(\log\log n)\rfloor)$. Then our residual estimates are

$$\hat{\varepsilon}_{n,j} = X_j + \sum_{k=1}^{p_n}(-\hat{a}_n)^k X_{j-k},$$

Next, apply a kernel density estimator to this sequence that utilizes the centering assumption and converges at a $o(n^{-1/2})$ rate. Müller et al.'s (2005) weighted kernel density estimator

$$\hat{f}_n(x) := \frac{1}{n - p_n}\sum_{j=p_n+1}^{n} w_{n,j} k_{b_n}(x - \hat{\varepsilon}_{n,j}),$$

where k_{b_n} is a kernel, b_n is a bandwidth, and $w_{n,j} := \frac{1}{1+\lambda\hat{\varepsilon}_j}$ are the weights, suffices for this purpose. We'll use a bandwidth proportional to $n^{-1/4}$.

Then, construct a bootstrap residual sequence ε_j^* for $1 - p_n \le j \le N(n)$ using iid sampling from density $\hat{f}_n$; here the replication length $N(n)$ satisfies $n^{5/2}/N(n) = o(1)$—see the subsection "Determination of necessary bootstrap length" in Appendix B. Finally, calculate bootstrap pseudo-data $X_j^* = \varepsilon_j^* + \hat{a}_n\varepsilon_{j-1}^*$ for $j = 1, \ldots, N(n)$, and estimate h with

$$\hat{h}_n^* := \frac{1}{N}\sum_{j=1}^{N} K_{d_N}(x - X_j^*) \tag{3}$$

where $\{d_n\}$ is a second sequence of bandwidths, and K is another kernel function. We'll use d_n proportional to $n^{-1/5}$.

Our main result is the following:

Theorem 2.1. *Given an MA(1) process of form (2), let $\hat{h}_n^*$ be as defined above, $d_n := Dn^{-1/5}$ for some constant D, N satisfy $n^{5/2}/N = o(1)$, and all the conditions in Section 6.1 hold. Then $\hat{h}_n^* = h + O_P(n^{-1/2})$.*

Note that the notation $\hat{h}_n^* = h + O_P(n^{-1/2})$ is short-hand for $\hat{h}_n^*(x) = h(x) + O_P(n^{-1/2})$, uniformly in x.

2.2 Extending to MA(p)

Now consider the process

$$X_t = \varepsilon_t + \sum_{j=1}^{p} a_j \varepsilon_{t-j}, \qquad a_p \neq 0, \varepsilon_t \text{ iid with density } f, \tag{4}$$

where the a_j's are such that $1 + \sum_{j=1}^{p} a_j z^j$ has no roots on the complex unit disk, and f satisfies (SW-F). Since the process is invertible, the least squares estimators $\hat{a}_{1,n}, \ldots, \hat{a}_{p,n}$ of $a_1, \ldots, a_p$ are root-n consistent and satisfy (SW-R) with $p_n = \min(\lfloor |\log_{|b|} n\rfloor| + 1, \lfloor \frac{n}{2} \rfloor)$, where b is the root of $1 + \sum_{j=1}^{p} a_j z^j$ with magnitude closest to 1. Next, calculate the residuals $\hat{\varepsilon}_{n,j} = X_j - \sum_{s=1}^{p_n} \hat{\varrho}_s X_{j-s}$, where $1 - \sum_{s=1}^{\infty} \hat{\varrho}_s z^s = \frac{1}{1+\sum_{s=p_n}^{\infty} \hat{a}_s z^s}$. Compute the weighted kernel estimator

$$\hat{f}_n(x) := \frac{1}{n - p_n} \sum_{j=p_n+1}^{n} w_{n,j} k_{b_n}(x - \hat{\varepsilon}_j).$$

where $w_{n,j}$ satisfies (MSW-W), k satisfies (SW-K), and b_n satisfies (SW-Q) for some ζ satisfying (SW-B). Construct a bootstrap replication ε_j^* of the residuals (iid $\hat{f}_n$) for $1 - p_n \leq j \leq N$, and calculate $X_j^* = \varepsilon_j^* + \sum_{s=1}^{p_n} \hat{a}_{s,n} \varepsilon_{j-s}^*$. Finally, estimate h with $\hat{h}_n^*(x) := \frac{1}{N} \sum_{j=1}^{n} K_{d_n}(x - X_j^*)$ where K satisfies (ST-K).

Then we have the following result:

Theorem 2.2. *Given a MA(p) process of form (4), let $\hat{h}_n^*$ be as defined above, $d_n := Dn^{-1/5}$ for some constant D, N satisfy $n^{5/2}/N = o(1)$, and all the conditions in Section 6.1 hold. Then $\hat{h}_n^* = h + O_P(n^{1/2})$.*

3 Nonlinear AR(1)

Next, consider a stationary and geometrically ergodic nonlinear process with representation

$$X_{i+1} = g(X_i) + e_i, \qquad e_i \text{ iid with density } f, \tag{5}$$

where f has mean zero and g is differentiable and invertible. Note that the differentiability condition excludes some common nonlinear AR(1) models, such as SETAR.

For clarity of exposition, we will assume S.1 and S.2 in Appendix A are satisfied; this is slightly stronger than stationary and geometrically ergodic.

As before, let h be the stationary density of the X_i's. Since X_i has the same distribution as $g(X_i) + e_i$, following Stove (2008) we have

$$h(x) = \int f(x - g(u))h(u)\, du = E[f(x - g(X))].$$

In light of this, construct an estimator

$$\tilde{h}_n(x) = \hat{E}[\hat{f}_n(x - \tilde{g}_n(X))] \tag{6}$$

where $\hat{f}_n$ is a weighted kernel estimator of the density of the e_i's, $\tilde{g}_n$ is a root-n consistent estimator of g (such as a parametric least squares estimator), and $\hat{E}$ represents an average taken over the observed X_is. (Note that a root-n consistent estimator of g may not always exist.)

More precisely, estimate $\tilde{e}_{n,i} = X_i - \tilde{g}_n(X_{i-1})$ for $2 \leq i \leq n$. Then, for some kernel k satisfying (SW-K) and $\inf_{x\in C} k(x) > 0$ for all compact sets C, and a sequence of bandwidths b_n satisfying (SW-B), set $\hat{f}_n(x) = \frac{1}{n-1} \sum_{j=2}^{n} w_{n,j} k_{b_n}(x - \tilde{e}_{n,j})$ where $w_{n,j}$ satisfies (MSW-W) with $\hat{\varepsilon}$ replaced with $\tilde{e}$. Plugging that into (6) yields $\tilde{h}_n(x) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=2}^{n} w_{n,j} k_{b_n}(x - \tilde{g}_n(X_i) - \tilde{e}_{n,j})$.

Preliminary results by Støve and Tjøstheim (2008) suggest that $\tilde{h}_n^u$ is a root-n consistent estimator of h, i.e.

$$\tilde{h}_n^u = h + O_P(n^{-1/2}). \tag{7}$$

Since $\tilde{h}_n$ performs no worse than $\tilde{h}_n^u$, (7) implies

$$\tilde{h}_n = h + O_P(n^{-1/2}).$$

Now we propose a bootstrap kernel estimator of h that is root-n consistent given (7).

Construct a bootstrap replication $e_{j,n}^*$ of the residuals using $\hat{f}_n$ for $-m_n \le j \le N(n)$ where $m_n := \lceil (\log n)^2 \rceil$ and $N(n)$ is to be determined later. Let $X_{-m_n-1,n}^*$ be randomly drawn from the observed X_i's, and compute $X_{j,n}^* := \tilde{g}_n(X_{j-1,n}) + e_{j,n}^*$ for $-m_n \le j \le N(n)$. Our estimator of h is

$$\hat{h}_n^* := \frac{1}{N} \sum_{j=1}^{N} K_{d_N}(x - X_{j,n}^*)$$

where K and d_N are still defined as in the first section.

Then we have the following result:

Theorem 3.1. *Given a nonlinear AR(1) process of form (5), let $\hat{h}_n^*$ and $\tilde{h}_n$ be as defined above, $d_n := Dn^{-1/5}$ for some constant D, N satisfy $n^{5/2}/N = o(1)$, and all the conditions in Section 6.2 hold. If (7) is true, then $\hat{h}_n^* = h + O_P(n^{-1/2})$.*

3.1 Application: AR(1) Density Estimation

Assume a stationary linear process with AR(1) representation

$$X_t = aX_{t-1} + \varepsilon_t, t \in \mathbb{Z}, a \neq 0, |a| < 1, \varepsilon_t \sim f \forall t,$$

where f has mean zero and $\inf_{x \in C} f(x) > 0$ for all compact sets C. As usual, let h be the true density of the X_t's.

Compute the least squares estimator of a (i.e. minimize $\sum_{j=2}^{n}(X_j - aX_{j-1})^2$); this estimator, which we'll denote as $\hat{a}_n$, is root-n consistent. Then estimate $\tilde{e}_{n,t} = X_t - \hat{a}_n X_{t-1}$ for $2 \le t \le n$, and finish the calculation of $\tilde{h}_n$ as with a nonlinear AR(1) process. If (7) is true for the general nonlinear case, it's true for this $\tilde{h}_n$.

We now propose a bootstrap kernel estimation procedure that's root-n consistent given (7). Draw an iid sample $\varepsilon_{j,n}^*$ from the density $\hat{f}_n$ for $-m_n \le j \le N(n)$ where $m_n = \lceil (\log n)^2 \rceil$ and $N(n) \sim n^{5/2+\epsilon}$. Let $X_{-m_n-1,n}^*$ be randomly drawn from the observed X_i's, and compute $X_{j,n}^* := \hat{a} X_{j-1,n} + \varepsilon_{j,n}^*$ for $-m_n \le j \le N(n)$. Estimate h with

$$\hat{h}_n^* := \frac{1}{N} \sum_{j=1}^{N} K_{d_N}(x - X_{j,n}^*)$$

where K and d_N are defined as in the first section.

Root-n consistency of this estimator, given (7), is shown by Theorem 3.1.

3.2 Application: Nonlinear Parametric AR(1) Density Estimation

Now assume a stationary and geometrically ergodic nonlinear process

$$X_{i+1} = g_\varphi(X_i) + e_i$$

just like the general nonlinear AR(1) case, except that g is known up to a q-dimensional parameter φ, and this provides a framework for estimating g root-n consistently. For instance, we can have a root-n consistent estimator $\hat{\varphi}$ of φ, and have the parametrization of g obey the following condition from Muller (2005):

The function $\tau \mapsto g_\tau(x)$ is differentiable for all x with derivative $\tau \mapsto \dot{g}_\tau(x)$, and for each constant C,

$$\sup_{|\tau - \varphi| \le Cn^{-1/2}} \sum_{i=1}^{n} (g_\tau(X_i) - g_\varphi(X_i) - \dot{g}_\varphi(X_i)(\tau - \varphi))^2 = o_P(1).$$

Also, $E[|\dot{g}_\varphi(X)|^{5/2}] < \infty$.

Then (given (7)) a root-n consistent estimator of h can be constructed as follows:

Estimate $\tilde{e}_{n,t} = X_t - g_{\hat{\varphi}}(X_{t-1})$ for $2 \le t \le n$, and finish the calculation of $\tilde{h}_n$ as with a nonlinear AR(1) process. Draw an iid sample $\varepsilon^*_{j,n}$ from the density $\hat{f}_n$ for $-m_n \le j \le N(n)$ where, as before, $m_n = \lceil (\log n)^2 \rceil$ and $N(n) \sim n^{5/2+\epsilon}$. Let $X^*_{-m_n-1,n}$ be randomly drawn from the observed X_i's, and compute $X^*_{j,n} := \hat{a}X_{j-1,n} + \varepsilon^*_{j,n}$ for $-m_n \le j \le N(n)$. Estimate h with

$$\hat{h}^*_n := \frac{1}{N}\sum_{j=1}^{N} K_{d_N}(x - X^*_{j,n})$$

where K and d_N are defined as in the first section.

4 Simulation study

To evaluate our proposed estimator on finite samples, we compare its (numerically estimated) mean integrated squared error (MISE) to that of the classical kernel estimator (1).

For each entry in the following tables, 200 simulated realizations with fixed sample size (usually $n = 100$ or $n = 400$) of the process $\{X_t\}$ were generated, and then a bootstrap replication of length $n^{5/2}$ was generated off each sample. The first 200 elements of these replications were discarded. (Note that the computation of a single long bootstrap replication of length $\ge 1000n$ is as computer intensive as the usual procedure of generating 1000 or more length-n replications and averaging the results; but using a single replication is slightly advantageous because the initial "break-in" period doesn't have to be repeated. In the $n = 100$ case, $n^{5/2}$ is precisely $1000n$, while $n^{5/2} = 8000n$ when $n = 400$.)

The estimated MISEs (denoted by $\widehat{\text{MISE}}$) of our proposed estimator and the classical kernel estimator were computed by averaging the results of numerically integrating the square of the difference between the density estimates and the true marginal density.

Gaussian kernels were used. Bandwidth selection was left to R 2.9's default behavior, namely $0.9\min(\text{stdev}, \frac{\text{IQR}}{1.34})n^{-1/5}$.

The AR(1) model $X_t = \phi X_{t-1} + e_t$ was investigated first, with the following choices of densities for e_t:

Gaussian: $N(0,1)$

Skewed unimodal: $\frac{1}{5}N(0,1) + \frac{1}{5}N(\frac{1}{2},\frac{2}{3}) + \frac{3}{5}N(\frac{4}{5},\frac{5}{9})$

Kurtotic unimodal: $\frac{2}{3}N(0,1) + \frac{1}{3}N(0,\frac{1}{10})$

Separated bimodal: $\frac{1}{2}N(-\frac{3}{2},\frac{1}{2}) + \frac{1}{2}N(\frac{3}{2},\frac{1}{2})$

Density	Coef.	Sample size	Bootstrap $\hat{MISE}$	Standard kernel $\hat{MISE}$	SE of difference	% advantage
Gaussian	0.8	100	.00286	.01256	.01084	77
		400	.00075	.00440	.00397	83
	0.5	100	.00272	.00859	.00626	68
		400	.00072	.00247	.00130	71
	0.2	100	.00423	.00695	.00383	39
		400	.00132	.00219	.00102	39
	-0.2	100	.00407	.00604	.00255	32
		400	.00134	.00203	.00080	34
Skewed unimodal	0.8	100	.00481	.01867	.01623	74
		400	.00166	.00553	.00432	70
	0.5	100	.00502	.01347	.01017	63
		400	.00157	.00390	.00199	60
	0.2	100	.00698	.01000	.00592	30
		400	.00222	.00359	.00166	38
	-0.2	100	.00680	.00897	.00465	24
		400	.00251	.00338	.00144	26
Kurtotic unimodal	0.8	100	.00338	.01414	.01082	76
		400	.00078	.00414	.00360	83
	0.5	100	.00302	.00880	.00628	66
		400	.00078	.00305	.00186	74
	0.2	100	.00518	.00825	.00441	37
		400	.00195	.00289	.00121	32
	-0.2	100	.00562	.00743	.00303	24
		400	.00192	.00262	.00102	27
Separated bimodal	0.8	100	.00135	.00712	.00698	81
		400	.00035	.00204	.00178	83
	0.5	100	.00242	.00544	.00441	56
		400	.00101	.00173	.00086	41
	0.2	100	.02702	.02047	.00880	-32
		400	.01059	.00876	.00395	-21
	-0.2	100	.02759	.01989	.00868	-39
		400	.01104	.00866	.00453	-28

Table 1: AR(1) simulation results.

It's easily seen from Table 1 that our bootstrap estimator almost always yields better results, though the improvement is smaller when the AR coefficient is low (unsurprising since our theoretical results show the bootstrap estimator would yield no improvement in the $a = 0$ case), and in the separated bimodal subcase the bootstrap estimator exhibits worse performance than the classical kernel estimator. However, even there the superior asymptotic performance of the bootstrap is in evidence, as a 32% to 39% MISE disadvantage when $n = 100$ declines to a roughly 25% disadvantage when n increases to 400; and larger sample sizes are slightly associated with better relative performance of our estimator across the board.

Next, we looked at the MA(1) model $X_t = e_t + ae_{t-1}$, with the same mix of densities.

Density	Coef.	Sample size	Bootstrap $\widehat{MISE}$	Standard kernel $\widehat{MISE}$	SE of difference	% advantage
Gaussian	0.8	100	.00504	.00632	.00600	20
		400	.00112	.00222	.00103	49
	0.5	100	.00462	.00689	.00320	33
		400	.00137	.00241	.00105	43
	0.2	100	.00600	.00670	.00241	11
		400	.00199	.00245	.00075	19
	-0.2	100	.00477	.00575	.00230	17
		400	.00174	.00213	.00063	18
Skewed unimodal	0.8	100	.00856	.01045	.00758	18
		400	.00327	.00464	.00192	29
	0.5	100	.00772	.01024	.00484	25
		400	.00256	.00395	.00182	17
	0.2	100	.00899	.01002	.00389	10
		400	.00315	.00367	.00127	14
	-0.2	100	.00814	.00900	.00436	9
		400	.00257	.00311	.00100	17
Kurtotic unimodal	0.8	100	.02130	.02106	.00975	-1
		400	.00807	.01140	.00336	29
	0.5	100	.01873	.02268	.00933	17
		400	.00792	.01190	.00325	33
	0.2	100	.03822	.03645	.01388	-5
		400	.01373	.01614	.00520	15
	-0.2	100	.03407	.03244	.01490	-5
		400	.01385	.01500	.00631	8
Separated bimodal	0.8	100	.02141	.01560	.00523	-37
		400	.00980	.00789	.00189	-24
	0.5	100	.00706	.00726	.00207	3
		400	.00354	.00336	.00103	-5
	0.2	100	.02554	.02038	.00820	-25
		400	.01075	.00921	.00471	-17
	-0.2	100	.02659	.01990	.00946	-34
		400	.01068	.00884	.00481	-21

Table 2: MA(1) simulation results.

Table 2 exhibits most of the same patterns seen in Table 1. Our estimator outperforms the standard kernel density estimator for all error densities except the separated bimodal, though, as expected, the performance advantage is smaller for low MA(1) coefficients. Larger sample sizes are associated with superior relative performance.

Our third simulation generated data from the MA(3) process $X_t = e_t + a_1 e_{t-1} + a_2 e_{t-2} + a_3 e_{t-3}$.

Density	Coefs.	Sample size	Bootstrap $\widehat{MISE}$	Std. kernel $\widehat{MISE}$	SE of difference	% advantage
Gaussian	1, 0, -0.5	100	.00237	.00554	.00325	57
		400	.00064	.00166	.00087	61
	0.6, 0.3, 0.1	100	.00528	.00757	.00345	30
		400	.00157	.00272	.00115	42
Skewed unimodal	1, 0, -0.5	100	.00437	.00789	.00421	45
		400	.00210	.00372	.00175	44
	0.6, 0.3, 0.1	100	.00869	.01271	.00571	32
		400	.00320	.00466	.00193	31
Kurtotic unimodal	1, 0, -0.5	100	.00519	.00779	.00439	33
		400	.00154	.00323	.00140	52
	0.6, 0.3, 0.1	100	.01194	.01543	.00866	23
		400	.00319	.00508	.00243	37
Separated bimodal	1, 0, -0.5	100	.00212	.00342	.00162	38
		400	.00083	.00119	.00062	30
	0.6, 0.3, 0.1	100	.00418	.00469	.00145	11
		400	.00150	.00172	.00064	13

Table 3: MA(3) simulation results. The MA coefficients are from lowest to highest order.

From Table 3, we can observe that a more complex known dependence structure leads to consistently better relative performance of our estimator even on moderately sized samples.

Finally, we simulated nonlinear AR(1) data from the process $X_t = \phi \tan^{-1} X_{t-1} + e_t$.

Density	Coef.	Sample	Bootstrap $\widehat{\text{MISE}}$	Std. kernel $\widehat{\text{MISE}}$	SE of difference	% advantage
Gaussian	1	36	.14843	.15623	.01682	5
		100	.14344	.14591	.00865	2
		400	.14290	.14302	.00399	0
	0.5	36	.00844	.01851	.01254	54
		100	.00375	.00782	.00522	48
		400	.00141	.00284	.00130	50
	-0.2	36	.00796	.01272	.00783	37
		100	.00421	.00594	.00246	29
		400	.00151	.00218	.00077	31
	-0.8	36	.01584	.02057	.00914	23
		100	.01557	.01798	.00518	13
		400	.01679	.01731	.00240	3
Skewed unimodal	1	36	.21769	.22613	.02888	4
		100	.20533	.20829	.01607	1
		400	.19800	.19827	.00694	0
	0.5	36	.01306	.02533	.01940	48
		100	.00463	.00996	.00639	53
		400	.00200	.00419	.00255	52
	-0.2	36	.01645	.02134	.01067	23
		100	.00675	.00824	.00335	18
		400	.00230	.00300	.00124	23
	-0.8	36	.02332	.02827	.01543	18
		100	.01889	.02216	.00900	15
		400	.01972	.02082	.00400	5
Kurtotic unimodal	1	36	.15627	.16352	.01981	4
		100	.14788	.15114	.00951	2
		400	.14799	.14773	.00419	0
	0.5	36	.00891	.01828	.01273	51
		100	.00324	.00788	.00510	59
		400	.00161	.00311	.00157	48
	-0.2	36	.01104	.01582	.01076	30
		100	.00530	.00652	.00256	19
		400	.00193	.00239	.00080	19
	-0.8	36	.01899	.02219	.00885	14
		100	.01696	.01846	.00539	8
		400	.01736	.01787	.00264	3
Separated bimodal	1	36	.07139	.07211	.00389	1
		100	.07154	.07089	.00209	-1
		400	.07309	.07210	.00102	-1
	0.5	36	.00788	.01126	.00476	30
		100	.00968	.00990	.00472	1
		400	.01540	.01407	.00537	-9
	-0.2	36	.04551	.02861	.01399	-59
		100	.02152	.01424	.00837	-51
		400	.00586	.00411	.00239	-42
	-0.8	36	.01364	.01482	.00404	8
		100	.01500	.01454	.00350	-3
		400	.01584	.01531	.00245	-3

Table 4: Nonlinear AR(1) simulation results.

From Table 4, we can see that, with the exception of the separated bimodal $\phi = -0.2$ case, our estimator continued to outperform (or match, in the nearly nonstationary $\phi = 1$ case) the standard kernel density estimator. It appears that multimodality of the error distribution genuinely lowers effectiveness in the nonlinear AR case as also noted by Støve and Tjøstheim (2008) in the non-bootstrap implementation of the convolution estimator.

However, there was one unexpected pattern: larger sample sizes were no longer associated with better relative performance, and this phenomenon was not due to errors in estimating ϕ. Our limited simulation data does not appear to exhibit a root-n convergence rate. Since our theoretical root-n convergence result is dependent on the validity of eq. (7) as conjectured by Støve and Tjøstheim (2008), one possibility is that the conjecture is false. Further investigation of this case is in order.

5 Conclusions

A bootstrap-based kernel density estimator was presented, and proved to estimate the marginal density of certain finite-order moving average processes and order 1 autoregressive processes root-n consistently. This matches the asymptotic performance of the best known convolution estimators, and is a significant improvement over the $n^{-2/5}$ rate of the usual kernel density estimator.

Simulations indicate that a sample size of 100 is sufficient to realize this performance advantage in most cases, though the advantage is greater across the board given a sample size of 400 (confirming our asymptotic analysis). Small dependence coefficients lower the effectiveness of our estimator, as would be expected from considering the independent case where no improvement is possible. Multimodality of the error distribution also lowers effectiveness, as also noted by Støve and Tjøstheim (2008). When these factors are present, simulation results indicate that our estimator still does not perform much worse than the standard kernel density estimator, but it is unlikely to provide a significant advantage, either.

Our estimator also tends to outperform the usual kernel density estimator for nonlinear autoregressions. However, the picture there is less complete as our simulation does not appear to exhibit a root-n rate, and our theoretical result predicting that convergence rate is dependent on a conjecture.

6 Appendix A: Technical conditions

6.1 MA(1), MA(p)

Conditions on estimation of $\hat{a}$ and initial extraction of residuals:

(SW-R) p_n is a sequence of positive integers where $\frac{p_n}{n} \to 0$ and $np_n c^{2p_n} \to 0$ for all $c \in (-1, 1)$. If $\{X_t\}$ is instead expressed as an autoregression, viz. $\varepsilon_t = X_t - \sum_{s=1}^{\infty} \varrho_s X_{t-s}$, the estimators $\hat{\varrho}_{i,n} = -(-\hat{a}_n)^i$ of the autoregression coefficients $\varrho_i = -(-a)^i$ satisfy

$$\sum_{i=1}^{p_n} (\hat{\varrho}_{i,n} - \hat{\varrho}_i)^2 = O_p(q_n n^{-1})$$

Conditions on the weighted kernel density estimator:

(MSW-W) $w_{n,j} := \frac{1}{1+\lambda\hat{\varepsilon}_j}$ for a choice of λ satisfying $\sum_{j=p_n+1}^{n} w_{n,j} \hat{\varepsilon}_{n,j} = 0$,

(SW-K) $k \geq 0$ integrates to one, and has bounded, continuous, and integrable derivatives up to order two satisfying $\int t^i k(t)\, dt = 0$ for $i = 1, 2$ and $\int |t|^4 |k(t)|\, dt < \infty$,

(SW-Q) $\sum_{s>p_n} |a_s| = O(n^{-1/2-\zeta})$ for some $\zeta > 0$.

(SW-B) The sequences b_n, p_n and q_n and the exponent ζ satisfy $p_n q_n b_n^{-1} \times n^{-1/2} \to 0$, $nb_n^4 = O(1)$, $n^{1/4} s_n \to 0$ and $n^{1/2} b_n s_n = O(1)$, where $s_n = b_n^{-1/2} n^{-1/2} + p_n q_n b_n^{-5/2} n^{-1} + b_n^{-3/2} n^{-\zeta-1/2}$.

Conditions on the kernel used in constructing the final marginal density estimate:

(ST-K) $K \geq 0$ is bounded, two times differentiable, symmetric, integrates to one, $\int K'(z)\,dz = 0$, and $\int z^2 K'(z)\,dz = 0$.

Conditions required to use results in Schick & Wefelmeyer (2007) in the proof of the MA(1) convergence result:

(SW-C) If X_t is expressed as $\varepsilon_t + \sum_{s=1}^{\infty} \varphi_s \varepsilon_{t-s}$, at least one of the moving average coefficients φ_s is nonzero.
(SW-I) The function $\phi(z) = 1 + \sum_{s=1}^{\infty} \varphi_s z^s$ is bounded, and bounded away from zero, on the complex unit disk.
(SW-S) $\sum_{s=1}^{\infty} s|\varphi_s| < \infty$.

6.2 Nonlinear AR(1)

Pair of sufficient conditions for stationarity and geometric ergodicity (Franke (2002a)):

S.1. $\inf_{x \in C} f(x) > 0$ for all compact sets C.
S.2. g is bounded on compact sets and $\limsup_{|x| \to \infty} \frac{E[|g(x)+e_1|]}{|x|} < 1$.

Franke et al.'s (2002b) geometric ergodicity theorem and conditions (used in the final proof):

F.1. There exists a compact set K such that
(i) there exist $\rho > 1$ and $\varepsilon > 0$ with

$$E[|X_t||X_{t-1} = x] \leq \rho^{-1}|x| - \varepsilon \qquad \forall x \notin K$$

(ii) there exists $A < \infty$ with

$$\sup_{x \in K} \{E[|X_t||X_{t-1} = x]\} \leq A.$$

F.2. K is a small set, i.e. there exist $n_0 \in \mathbb{N}, \gamma > 0$ and a probability measure ϕ such that

$$\inf_{x \in K} \{P^{n_0}(x, B)\} \geq \gamma \phi(B)$$

holds for all measurable sets B. $P^n(x, \cdot)$ denotes the n-step transition probability of the Markov chain started in x.
F.3. There exists $\kappa > 0$ such that

$$\inf_{x \in K} \{P(x, K)\} \geq \kappa.$$

Theorem 6.1. *(Franke et al. (2002b)) Given F.1, F.2, and F.3, $\{X_t\}$ is geometrically ergodic with convergence rate ρ_μ only dependent on K, ρ, ε, A, n_0, γ, and κ.*

This is used to establish the existence of a single geometric bound in the proof of Theorem 3.1.

7 Appendix B: Proofs

7.1 Determination of necessary bootstrap length

The bootstrap length $N(n)$ must be chosen such that the pdf $\hat{h}_n^*$ is within $Cn^{-1/2}$ of

$$\hat{h}_n := \hat{f}_n * \hat{f}_{n,\hat{a}_n} \tag{8}$$

everywhere with probability converging to 1. I.e., $P^*(\sup_x |\hat{h}_n^*(x) - \hat{h}_n(x)| > Cn^{-1/2}) \to 0$ as $n \to \infty$, where C is some constant, $\hat{f}_{n,c}(x) := c^{-1}\hat{f}_n(x/c)$, and $*$ indicates convolution. The following lemma tells us how to do this.

Lemma 7.1. *If $\hat{h}_n^*$ is as defined in* (3), *$\hat{h}_n$ is as defined in* (8), *and $d_n := Dn^{-1/5}$ for some constant D, choosing $N(n)$ such that $n^{5/2}/N(n) = o(1)$ guarantees $P^*(\sup_x |\hat{h}_n^*(x) - \hat{h}_n(x)| > Cn^{-1/2}) \to 0$ as $n \to \infty$.*

Proof. $\hat{h}_n^*$ is a convergent kernel density estimator of $\hat{h}_n$ with mean integrated squared error (MISE) of order $N^{-4/5}$ over bootstrap resamples (see e.g. Jones (1995) pg. 22–23). Thus, the L^2 distance between $\hat{h}_n^*$ and $\hat{h}_n$ in a bootstrap resample will, for any fixed probability $p < 1$, be less than a constant multiple of $\frac{N^{-4/5}}{1-p}$ with probability p. Also, the first derivative of $\hat{h}_n^*$ is bounded above by a constant multiple of $N^{1/5}$, because the maximal first derivative of K_{d_N} is of order d_N^{-1}, and similarly, the first derivative of $\hat{h}_n$ is bounded above by a constant multiple of b_n^{-1}. So the first derivative of $|\hat{h}_n^* - \hat{h}_n|$ is bounded above by a constant multiple of $\max(d_N^{-1}, b_n^{-1})$; for $n^{5/2}/N = o(1)$ and $b_n^{-1} = O(n^{1/4})$, d_N^{-1} is asymptotically larger.

Note that, if one is trying to maximize the L^∞ norm of a function with fixed L^2 norm and bounded first derivative, a triangular spike with sides of maximal slope is optimal. To see this, assume toward a contradiction that there exists a function g with identical L^2 norm but greater L^∞ norm γ', and denote the L^∞ norm of the triangular spike by γ. Then, there must exist some x for which $|g(x)| = \frac{\gamma+\gamma'}{2}$. Let the function j be the triangular spike centered at x. $|g(x)| > |j(x)|$, and $|g|$ cannot descend faster than $|j|$ on either side of x since first derivatives are bounded and $|j|$ is defined to attain the extremal values. Thus, $|g| \geq |j|$ everywhere and g must have a larger L^2 norm than j.

We can now use calculus to compute an upper bound on $\max_x |\hat{h}_n^*(x) - \hat{h}_n(x)|$ as a function of N.

$$\begin{aligned}
N^{-4/5} &= 2\int_0^{HN^{-1/5}} (N^{1/5}x)^2\,dx \\
&= \frac{2}{3}N^{2/5}(HN^{-1/5})^3 \\
&= \frac{2}{3}H^3N^{-1/5} \\
\frac{3}{2}N^{-3/5} &= H^3 \\
H &= O(N^{-1/5})
\end{aligned}$$

So choosing N such that $n^{5/2}/N = o(1)$ guarantees $\max_x |\hat{h}_n^*(x) - \hat{h}_n(x)| \leq H = o(n^{-1/2})$ for $d_n = Dn^{-1/5}$ with probability converging to 1. □

7.2 Proof of Theorem 2.1

Proof. First, we verify that conditions (SW-C), (SW-S), and (SW-I) are satisfied. $a \neq 0$ ensures (SW-C) is met. (SW-S) is automatic since there's only one moving average coefficient. $|a| < 1$ guarantees (SW-I).

Next, Lemma 7.1 shows that $\hat{h}_n^* = \hat{h}_n + O_P(n^{-1/2})$, so it remains to prove that $\hat{h}_n = \hat{f}_n * \hat{f}_{n,\hat{a}_n}$ is a root-n consistent estimator of h. Since the true density h satisfies $h = f * f_a$ (where $f_a(x) := a^{-1}f(x/a)$), we can write $\hat{h}_n - h$ as:

$$\hat{h}_n - h = (\hat{f}_n * \hat{f}_{n,\hat{a}} - \hat{f}_n * f_{n,\hat{a}}) + (\hat{f}_n * f_{\hat{a}} - f * f_{\hat{a}}) + (f * f_{\hat{a}} - f * f_a). \tag{9}$$

Now Muller (2005) demonstrates that the weighted estimator $\hat{f}_n$ performs no worse than the corresponding unweighted estimator $\hat{f}_n^u$, so we can use results in SW concerning $\hat{f}_n^u$.

The second and third components of (9) are $o(n^{-1/2})$ under the supremum norm (by Theorem 4 and Theorem 3 in SW, respectively; these theorems apply as long as (SW-C), (SW-I), (SW-S), (SW-F), (SW-R), (SW-K), (SW-Q), and (SW-B) hold, all of which have been verified above). The first component can be rewritten as $\hat{f} * (\hat{f}_{\hat{a}} - f_{\hat{a}_n})$, which has supremum norm equal to $\hat{a}_n^{-1}$ times that of $\hat{f}_{\hat{a}_n^{-1}} * (\hat{f} - f)$. This last convolution is $o(n^{-1/2})$ by SW Theorem 4. □

7.3 Proof of Theorem 2.2

Proof. Lemma 7.1 shows that $\hat{h}_n^*$ is a root-n consistent estimator of $\hat{h}_n$. Since $\hat{h}_n = \hat{f}_n * \hat{f}_{n,\hat{a}_{1,n}} * \cdots * \hat{f}_{n,\hat{a}_{p,n}}$ and $h = f * f_{a_{1,n}} * f_{a_{2,n}} * \cdots * f_{a_{p,n}}$, we have

$$\hat{h}_n - h = (\hat{f}_n * \hat{g}_{1,\hat{a},n} - \hat{f}_n * g_{1,\hat{a},n}) + (\hat{f}_n * g_{1,\hat{a},n} - f * g_{1,\hat{a},n}) + (f * g_{1,\hat{a},n} - f * g_{1,a}) \tag{10}$$

where we define $g_{k,a} := f_{a_k} * f_{a_{k+1}} * \cdots * f_{a_p}$, $g_{k,\hat{a},n} := f_{\hat{a}_{k,n}} * f_{\hat{a}_{k+1,n}} * \cdots * f_{\hat{a}_{p,n}}$, and $\hat{g}_{k,\hat{a},n} := \hat{f}_{n,\hat{a}_{k,n}} * \hat{f}_{n,\hat{a}_{k+1,n}} * \cdots * \hat{f}_{n,\hat{a}_{p,n}}$.

Note that (SW-C) and (SW-S) are satisfied by any nondegenerate MA(p) process, and the statement of (4) ensures (SW-I). Also, as before, we need not concern ourselves with the difference between $\hat{f}_n$ and $\hat{f}_n^u$. Thus, as in the MA(1) case, the second and third components of (10) are shown by SW to be $o(n^{-1/2})$. The first component can be rewritten as $(\hat{f} * (\hat{g}_{1,\hat{a},n} - g_{1,\hat{a},n}))$, which has supremum norm bounded above by that of $\hat{g}_{1,\hat{a},n} - g_{1,\hat{a},n}$ since $||\hat{f}||_1 = 1$. We can rewrite this upper bound as

$$\hat{g}_{1,\hat{a},n} - g_{1,\hat{a},n} = (\hat{f}_{n,\hat{a}_{1,n}} * \hat{g}_{2,\hat{a},n} - \hat{f}_{n,\hat{a}_{1,n}} * g_{2,\hat{a},n}) + (\hat{f}_{n,\hat{a}_{1,n}} * g_{2,\hat{a},n} - f_{n,\hat{a}_{1,n}} * g_{2,\hat{a},n});$$

the second term is $o(n^{-1/2})$ again, and the first term can be bounded and recursively expanded in the same manner. In the end, we have p separate terms, all $o(n^{-1/2})$. □

7.4 Proof of Theorem 3.1

Proof. Define $\hat{h}_{-m_n,n}(x)$ to be the density function of $X^*_{-m_n,n}$, $\hat{h}_{k,n}(x) := \int \hat{f}_n(x - \tilde{g}_n(u))\hat{h}_{k-1,n}(u)\,du$ for $k > -m_n$ (i.e. the density function of $X^*_{k,n}$), and $\hat{h}_{\infty,n}(x) := \lim_{k\to\infty} \hat{h}_{k,n}(x)$ (the existence of this limit will be proved below). Then $\hat{h}_n^* - \tilde{h}_n = (\hat{h}_n^* - \hat{h}_{\infty,n}) + (\hat{h}_{\infty,n} - \tilde{h}_n)$.

Because $\inf_{x\in C} k(x) > 0$ for all compact sets C, and $\tilde{g}_n$ satisfies S.2, the process $\{X^*_{j,n}\}$ (for fixed n) is geometrically ergodic and the associated autoregression has a stationary solution. Furthermore, geometric ergodicity assures us that $\hat{h}_{k,n}$ converges (as $k \to \infty$) at a geometric rate to the density of the autoregression's stationary solution. Thus the latter is $\lim_{k\to\infty} \hat{h}_{k,n}$.

The next question is whether the rate of geometric convergence can be bounded by the same value across different values of n.

For this, F.1, F.2, and F.3 are verified to hold when n is allowed to vary, and then Theorem 6.1 is applied. Because of S.2, there exists $c < 1$ where $\limsup_{|x|\to\infty} \frac{E[|g(x)+e_1|]}{|x|} < c$. It follows that $E[|\tilde{g}_n(X_t)||X_{t-1} = x] \leq \frac{1+c}{2}|x| - e_1$ for all sufficiently large n, so F.1.i holds. Also, S.2 ensures $\tilde{g}_n$ is uniformly bounded on compact sets for sufficiently large n, so F.1.ii also holds. F.2 and F.3 follow from S.1 and the consistency of $\hat{f}_n$ as an estimator of f.

Therefore, since $\frac{\log n}{m_n} \to 0$, and $||\hat{h}_{-m_n,n} - \hat{h}_{\infty,n}||_\infty = O_P(1)$, $||\hat{h}_{1,n} - \hat{h}_{\infty,n}|| = O_P(c^n)$ where $c < 1$ is a positive constant. It follows that $\hat{h}_n^*$ is close to a convergent kernel density estimator of $\hat{h}_{\infty,n}$. If the $X^*_{j,n}$'s were drawn from $\hat{h}_{\infty,n}$, $\hat{h}_n^*$ would have mean integrated squared error of order $N^{-4/5}$ as long as N only grows polynomially in n, and by Lemma 7.1 we can choose $N \sim n^{5/2+\epsilon}$ to ensure $\hat{h}_n^* - \hat{h}_{\infty,n} = O_P(1/\sqrt{n})$. Since the actual $X^*_{j,n}$'s are drawn from distributions differing from $\hat{h}_{\infty,n}$ by a geometrically small (w.r.t. n) amount, the additional bias and variance introduced by nonstationarity is of no consequence.

Finally, since $\tilde{h}_n$ is at least as good an estimator of $E[\hat{f}_n(x - \tilde{g}_n(X))]$ as it is of $E[f(x - g(X))]$ (two sources of error are eliminated, and none are introduced), and the former has density $\hat{h}_{\infty,n}$, we have $\hat{h}_{\infty,n} - \tilde{h}_n = O_P(n^{-1/2})$. Since $\tilde{h}_n - h = O_P(n^{-1/2})$ given (7), it immediately follows that $\hat{h}_n^* = h + O_P(n^{-1/2})$. □

8 References

P. Bickel, F. Götze, W. van Zwet, 1997. Resampling fewer than n observations: Gains, losses, and remedies for losses, *Statistica Sinica* **7**, pp. 1–32.

J. Bretagnolle, 1983. Lois limites du Bootstrap de certaines fonctionnelles, *Ann. Inst. Henri Poincare.* **19**, pp. 281–296.

S. Datta, 1995. On a modified bootstrap for certain asymptotically nonnormal statistics, *Stat. & Probability Letters* **24**, pp. 91–98.

A. Davison, D. Hinkley, 1997. Bootstrap Methods and their Application, Cambridge University Press, Cambridge.

B. Efron, R. Tibshirani, 1993. An Introduction to the Bootstrap, Chapman & Hall, New York.

J. Fan, Q. Yao, 2003. Nonlinear Time Series: Nonparametric and Parametric Methods, Springer-Verlag, New York.

J. Franke, J.-P. Kreiss, E. Mammen, 2002a. Bootstrap of kernel smoothing in nonlinear time series, *Bernoulli* **8**, pp. 1–37.

J. Franke, J.-P. Kreiss, E. Mammen, M. Neumann, 2002b. Properties of the Nonparametric Autoregressive Bootstrap, *J. Time Series Analysis* **23**, pp. 555–585.

L. Györfi, W. Härdle, P. Sarda, P. Vieu, 1989. Nonparametric Curve Estimation from Time Series (Lecture Notes in Statistics), Springer-Verlag, Berlin.

M. Jones, M. Wand, 1995. Kernel Smoothing, CRC Press, London.

U. Müller, A. Schick, W. Wefelmeyer, 2005. Weighted Residual-Based Density Estimators For Nonlinear Autoregressive Models, *Statistica Sinica* **15** (2005), pp. 177–195.

E. Parzen, 1962. On estimation of a probability density function and mode, *Ann. Math. Stat.* **33**, pp. 1065–1076.

D. Politis, J. Romano, 1993. Estimating the distribution of a studentized statistic by subsampling, *Bulletin of the International Statistical Institute* **49**, pp. 315–316.

D. Politis, J. Romano, M. Wolf, 1999. Subsampling, Springer, New York.

M. Rosenblatt, 1956. Remarks on some nonparametric estimates of a density function, *Ann. Math. Stat.* **27**, pp. 832–837.

M. Rosenblatt, 1970. Density estimates and Markov sequences. Nonparametric Techniques in Statistical Inference, Cambridge University Press, Cambridge, pp. 199–213.

G. Roussas, 1969. Nonparametric estimation in Markov processes, *Ann. Math. Stat.* **33**, pp. 73–87.

A. Saavedra, R. Cao, 1999. Rate of convergence of a convolution-type estimator of the marginal density of a MA(1) process, *Stochastic processes and their applications* **80**, pp. 129–155.

A. Schick, W. Wefelmeyer, 2007. Uniformly Root-N Consistent Density Estimators For Weakly Dependent Invertible Linear Processes, *Annals of Statistics* **35**, pp. 815–843.

B. Støve, D. Tjøstheim, 2007. A convolution estimator for the density of nonlinear regression observations, NHH Dept of Finance & Management Science Discussion Paper No. 2007/25.

B. Støve, D. Tjøstheim, 2008. A convolution density estimator for nonlinear time series: Simulations and some preliminary analysis, Preprint, Dept. of Mathematics, Univ. of Bergen, Norway.

J. Swanepoel, 1986. A note on proving that the (modified) bootstrap works, *Comm. Statist. Theory Methods.* **15**, pp. 3193–3203.