Title
Topics in nonparametric statistics

Author
Chang, Christopher

Publication Date
2011

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Topics in Nonparametric Statistics

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Mathematics

by

Christopher Chang

Committee in charge:

Professor Dimitris Politis, Chair
Professor Ian Abramson
Professor Ery Arias-Castro
Professor Anthony Gamst
Professor Karen Messer

2011

The dissertation of Christopher Chang is approved, and it is acceptable in quality and form for publication on microfilm:

Chair

University of California, San Diego

2011

DEDICATION

To mom and dad.

EPIGRAPH

To see what is in front of one's nose
needs a constant struggle.
—George Orwell

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor, Dimitris Politis. His careful guidance and unending patience were invaluable in the completion of this work.

Chapter 1 is essentially a reprint, with minor modifications, of the paper "Bootstrap with Larger Resample Size for Root-n Consistent Density Estimation with Time Series Data" by C. Chang and D.N. Politis, which has been published in *Statistics and Probability Letters*. The dissertation author was the primary investigator and author of this paper.

Chapter 2 is essentially a reprint, with minor modifications, of the paper "Aggregation of Spectral Density Estimators" by C. Chang and D.N. Politis, which has been submitted for publication in *IEEE Transactions on Information Theory*. The dissertation author was the primary investigator and author of this paper.

Chapter 3 is essentially a reprint, with minor modifications, of the paper "Robust Autocorrelation Estimation" by C. Chang and D.N. Politis, which is now in preparation for publication. The dissertation author was the primary investigator and author of this paper.

VITA

1979	Born, Newton, Massachusetts
2000	B. S. in Mathematics, California Institute of Technology
2000-2002	Software Design Engineer, Microsoft Corporation
2004-2009	Graduate Teaching Assistant, University of California, San Diego
2009-2011	Senior Engineer, Counsyl
2011	Ph. D. in Mathematics, University of California, San Diego

PUBLICATIONS

B.S. Srinivasan, C. Chang, et al., "A Universal Carrier Test for the Long Tail of Mendelian Disease", *Reprod. Biomed. Online*, 21, 2010.

C. Chang, D.N. Politis, "Bootstrap with Larger Resample Size for Root-n Consistent Density Estimation with Time Series Data", *Statistics and Probability Letters*, 2011.

ABSTRACT OF THE DISSERTATION

Topics in Nonparametric Statistics

by

Christopher Chang
Doctor of Philosophy in Mathematics

University of California San Diego, 2011

Professor Dimitris Politis, Chair

This thesis is concerned with nonparametric techniques for inferring properties of time series.

First, we consider finite-order moving average and nonlinear autoregressive processes with no parametric assumption on the innovation 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.

Next, given stationary time series data, we study the problem of finding the best linear combination of a set of lag window spectral density estimators with respect to the mean squared risk. We present an aggregation procedure and prove a sharp oracle inequality for its risk. We also provide simulations demonstrating the performance of our aggregation procedure, given Bartlett and other estimators of varying bandwidths as input. This extends work by Rigollet and Tsybakov on aggregation of density estimators.

The last part of this thesis introduces a class of robust autocorrelation estimators based on interpreting the sample autocorrelation function as a linear regression. We investigate the efficiency and robustness properties of the estimators that result from plugging on three common robust regression techniques. Construction of robust autocovariance and positive definite autocorrelation estimates is discussed, as well as application of the estimators to AR model fitting. We finish with simulations, which suggest that the estimators are especially well suited for AR model fitting.

Chapter 1

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

1.1 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

$$f(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \tag{1.1}$$

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. (1989) 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.

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.

1.2 MA(p) Density Estimation

1.2.1 MA(1)

Consider a stationary linear process with MA(1) representation

$$X_t = \varepsilon_t + a\varepsilon_{t-1}, \quad t \in \mathbb{Z}, a \neq 0, |a| < 1, \varepsilon_t \text{ iid with density } f. \tag{1.2}$$

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

$$\hat{h}_n^* := \frac{1}{N} \sum_{j=1}^{N} K_{d_N}(x - X_j^*) \tag{1.3}$$

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 1.2.1. *Given an MA(1) process of form* (1.2)*, 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 1.6.1 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.

1.2.2 Extending to MA(p)

Now consider the process

$$X_t = \varepsilon_t + \sum_{j=1}^{p} a_j \varepsilon_{t-j}, \quad a_p \neq 0, \varepsilon_t \text{ iid with density } f, \tag{1.4}$$

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 1.2.2. *Given a MA(p) process of form* (1.4)*, 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 1.6.1 hold. Then* $\hat{h}_n^* = h + O_P(n^{1/2})$*.*

1.3 Nonlinear AR(1)

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

$$X_{i+1} = g(X_i) + e_i, \quad e_i \text{ iid with density } f, \tag{1.5}$$

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

$$\tilde{h}_n(x) = \hat{E}[\hat{f}_n(x - \tilde{g}_n(X))] \tag{1.6}$$

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 (1.6) 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.

$$\tilde{h}_n^u = h + O_P(n^{-1/2}). \tag{1.7}$$

Since $\tilde{h}_n$ performs no worse than $\tilde{h}_n^u$, (1.7) 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 (1.7).

Construct a bootstrap replication $e^*_{j,n}$ of the residuals using $\hat{f}_n$ for $-m_n \leq j \leq N(n)$ 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 $-m_n \leq j \leq N(n)$. 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 1.3.1. *Given a nonlinear AR(1) process of form* (1.5)*, 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 1.6.2 hold. If* (1.7) *is true, then* $\hat{h}^*_n = h + O_P(n^{-1/2})$.

1.3.1 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 $2 \leq t \leq n$, and finish the calculation of $\tilde{h}_n$ as with a nonlinear AR(1) process. If (1.7) 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 (1.7). Draw an iid sample $\varepsilon^*_{j,n}$ from the density $\hat{f}_n$ for $-m_n \leq j \leq N(n)$ 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 $-m_n \leq j \leq N(n)$. 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 (1.7), is shown by Theorem 1.3.1.

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,

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

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

Then (given (1.7)) 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.

1.4 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 (1.1).

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

It's easily seen from Table 1.1 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.

Table 1.2 exhibits most of the same patterns seen in Table 1.1. 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 $X_t = e_t + a_1e_{t-1} +$

Table 1.1: AR(1) Simulation Results

Density	Coef.	Sample size	Bootstrap $\hat{MISE}$	Std. kernel $\hat{MISE}$	SE of diff.	% 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

$a_2 e_{t-2} + a_3 e_{t-3}$.

From Table 1.3, 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 $X_t = \phi \tan^{-1} X_{t-1} + e_t$.

From Table 1.4, 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 non-

Table 1.2: MA(1) simulation results.

Density	Coef.	Sample size	Bootstrap $\widehat{\text{MISE}}$	Std. kernel $\widehat{\text{MISE}}$	SE of diff.	% 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

linear 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. (1.7) as conjectured by Støve and Tjøstheim (2008), one possibility is that the conjecture is false. Further investigation of this case is in order.

Table 1.3: MA(3) simulation results. (The MA coefficients are from lowest to highest order.)

Density	Coefs.	Sample size	Bootstrap $\widehat{\text{MISE}}$	Std. kernel $\widehat{\text{MISE}}$	SE of diff.	% adv.
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

1.5 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 simu-

Table 1.4: Nonlinear AR(1) simulation results.

Density	Coef.	Sample	Bootstrap $\hat{MISE}$	Std. kernel $\hat{MISE}$	SE of diff.	% adv.
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

lation does not appear to exhibit a root-n rate, and our theoretical result predicting that convergence rate is dependent on a conjecture.

1.6 Appendix A: Technical conditions

1.6.1 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) $\sum_{s=1}^{\infty} s|\varphi_s| < \infty$.

1.6.2 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

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

(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 1.6.1. *(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 1.3.1.

1.7 Appendix B: Proofs

1.7.1 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

$$\hat{h}_n := \hat{f}_n * \hat{f}_{n,\hat{a}_n} \tag{1.8}$$

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 1.7.1. *If $\hat{h}_n^*$ is as defined in* (1.3)*, $\hat{h}_n$ is as defined in* (1.8)*, 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. □

1.7.2 Proof of Theorem 1.2.1

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 1.7.1 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:

$$
\hat{h}_n - h = (\hat{f}_n * \hat{f}_{n,\hat{a}} - \hat{f}_n * f_{n,\hat{a}}) + (\hat{f}_n * f_{\hat{a}} - f * f_{\hat{a}}) + (f * f_{\hat{a}} - f * f_a). \tag{1.9}
$$

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 (1.9) 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. □

1.7.3 Proof of Theorem 1.2.2

Proof. Lemma 1.7.1 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

$$\hat{h}_n - h = (\hat{f}_n * \hat{g}_{1,\hat{a},n} - \hat{f}_n * g_{1,\hat{a},n}) + (\hat{f}_n * g_{1,\hat{a},n} - f * g_{1,\hat{a},n}) + (f * g_{1,\hat{a},n} - f * g_{1,a}) \quad (1.10)$$

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 (1.4) 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 (1.10) 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 $o(n^{-1/2})$. □

1.7.4 Proof of Theorem 1.3.1

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 1.6.1 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 $E[|\tilde{g}_n(X_t)||X_{t-1} = x] \le \frac{1+c}{2}|x| - e_1$ 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 1.7.1 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 (1.7), it immediately follows that $\hat{h}^*_n = h + O_P(n^{-1/2})$. □

1.8 Acknowledgements

Chapter 1 is essentially a reprint, with minor modifications, of the paper "Bootstrap with Larger Resample Size for Root-n Consistent Density Estimation with Time Series Data" by C. Chang and D.N. Politis, which has been published in *Statistics and Probability Letters*. The dissertation author was the primary investigator and author of this paper.

Chapter 2

Aggregation of Spectral Density Estimators

2.1 Introduction

Consider stationary time series data $X_1, \ldots, X_n$ and autocovariances $\{\gamma(k)\}$ where the underlying process has true mean zero and spectral density

$$p(\omega) := \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma(j) e^{-i\omega j} \tag{2.1}$$

defined for all $\omega \in [-\pi, \pi)$. For an estimator $\hat{p}(X_1, \ldots, X_n)$ of p, define the L_2-risk

$$R_n(\hat{p}, p) = E\left[\int_{-\pi}^{\pi} (\hat{p}(x) - p(x))^2 \, dx\right]. \tag{2.2}$$

Let $\hat{p}_1, \ldots, \hat{p}_J$ be a collection of lag window (a.k.a. covariance averaging kernel) spectral density estimators of p. We investigate the construction of a new estimator $\hat{p}_n^L$ which is asymptotically as good, in terms of L_2-risk, as using the best possible linear combination of $\hat{p}_1, \ldots, \hat{p}_J$; more precisely, $\hat{p}_n^L$ satisfies the oracle inequality

$$R_n(\hat{p}_n^L, p) \leq \inf_{\lambda \in \mathbb{R}^J} R_n(\sum_{j=1}^{J} \lambda_j \hat{p}_j, p) + \Delta_{n,J} \tag{2.3}$$

where $\Delta_{n,J}$ is a small remainder term independent of p.

Such an estimator has a variety of applications. For instance, to perform bandwidth or model selection, one can set the $\hat{p}$s to cover a wide spread of possibly reasonable bandwidths/models. Or, when a linear combination of kernels outperforms all the individual inputs (e.g. when the $\hat{p}$s are Bartlett windows; see Politis (2011)), our estimator is capable of discovering it.

Kernel density estimation dates back to Rosenblatt (1956) and Parzen (1962); Priestley (1981) and Brillinger (1981) discuss its application to spectral densities. More recently, Rigollet and Tsybakov (2007) analyzed aggregation of probability density estimators. We extend Rigollet and Tsybakov's work to spectral estimation.

To perform aggregation, we use a sample splitting scheme. The time series data is divided into a training set, a buffer zone, and a validation set; with an exponential mixing rate, the buffer zone need not be more than logarithmic in the size of the other sets to ensure approximate independence between the training and validation sets.

The estimator, and theoretical results concerning its performance, are presented in section 2. Simulation studies are conducted in section 3, and our conclusions are stated in section 4.

2.2 Theoretical Results

2.2.1 Aggregation Procedure

Split the time series into a training set $X_1, \ldots, X_{n_t}$, a buffer zone $X_{n_t+1}, \ldots, X_{n_t+n_b}$, and a validation set $X_{n_t+n_b+1}, \ldots, X_{n_t+n_b+n_v}$, where the first and third sets can be treated as independent. We investigate appropriate choices of n_t, n_b, and n_v at the end of this section.

With the training set, we produce an initial estimate

$$\hat{\gamma}_1(k) := \frac{1}{n_t} \sum_{j=1}^{n_t-k} X_{j+k} X_j \tag{2.4}$$

of the autocovariance function, after centering the data. (In practice, the data will be centered to the sample mean rather than the true mean, but the resulting discrepancy is asymptotically negligible w.r.t. autocovariance and spectral density estimation. So, for simplicity of presentation, we center at the true mean above.)

We then propose the following candidate estimators:

$$p_j(\lambda) := \frac{1}{\sqrt{2\pi}} \sum_{k=-b_j}^{b_j} \hat{\gamma}_1(k) \cdot w_j\left(\frac{k}{b_j}\right) \frac{e^{ik\lambda}}{\sqrt{2\pi}} \tag{2.5}$$

where the b_js ($j = 1, \ldots, J$) are candidate bandwidths arrived at via some selection procedure, and the w_js ($j = 1, \ldots, J$) are lag windows with $w_j(0) = 1$, $w_j(x) \leq 1$ for $x \in (-1, 1)$, and $w_j(x) = 0$ for $|x| \geq 1$ for all j. The p_js have some linear span $\mathcal{L}$ in L_2 whose dimension is denoted by M where $M \leq J$. Now construct an orthonormal basis $\{\phi_j\}$ ($j = 1, \ldots, M$), and note that the ϕ_js are–by necessity–trigonometric polynomials of degree at most $b := \max_j b_j$, i.e.,

$$\phi_j(\lambda) = \sum_{k=-b}^{b} a_{j,k} \frac{e^{ik\lambda}}{\sqrt{2\pi}} \tag{2.6}$$

for some collection of coefficients $a_{j,k}$.

Then, based our validation set, we produce a different estimate of the autocovariance function, namely

$$\hat{\gamma}_2(k) := \frac{1}{n_v} \sum_{j=1}^{n_v-k} X_{n_t+m+j+k} X_{n_t+m+j} \tag{2.7}$$

and compute the coefficients

$$\hat{K}_j := \frac{1}{\sqrt{2\pi}} \sum_{k=-b}^{b} \hat{\gamma}_2(k) a_{j,k} \tag{2.8}$$

(so $a_{j,k}$ is the inner product of ϕ_j and $\frac{e^{ik\lambda}}{\sqrt{2\pi}}$ in L_2).

Finally, our proposed aggregate estimator of the spectral density is given by

$$\hat{p}(\lambda) := \sum_{j}^{M} \hat{K}_j \phi_j(\lambda). \tag{2.9}$$

2.2.2 Performance Bounds

We start with the simplest mixing assumption, m-dependence (i.e. for all positive integers j and k where $k \geq m$, X_j and X_{j+k} are independent).

Theorem 2.2.1. *If $\frac{b}{n} \to 0$, $EX_t^4 < \infty$, and the time series satisfies m-dependence, the L_2 risk is bounded above as follows:*

$$\begin{aligned} R_n(\hat{p}, p) \leq \min_{c_1,\ldots,c_M} ||\sum_{j=1}^{M} c_j p_j - p||^2 + \frac{bp^2(0)M}{n_v \pi} \\ + o(bM/n_v), \end{aligned} \tag{2.10}$$

where p is the true spectral density and $||\cdot||$ denotes the L_2 norm $(\int_{-\pi}^{\pi} (\cdot(x))^2\, dx)^{1/2}$.

Proof: Projecting p onto $\mathcal{L}$, we get $p^*_{\mathcal{L}} := \sum_{j=1}^{M} K^*_j \phi_j$, where K^*_j is the scalar product of p and ϕ_j in L_2. Then, by the Pythagorean theorem, we have

$$||\hat{p} - p||^2 = \sum_{j=1}^{M} (\hat{K}_j - K_j^*)^2 + ||p_{\mathcal{L}}^* - p||^2. \tag{2.11}$$

Next, we have $E[\hat{K}_j] = \frac{1}{\sqrt{2\pi}} \sum_{k=-b}^{b} E[\hat{\gamma}_2(k) a_{j,k}]$. Under m-dependence, the size-n_b buffer zone is sufficient to make all the $\hat{\gamma}_2(k)$s (functions only of the validation set) independent of the $a_{j,k}$s (functions only of the training set), so

$$\begin{aligned} E[\hat{K}_j] =& \frac{1}{\sqrt{2\pi}} \sum_{k=-b}^{b} E[\hat{\gamma}_2(k)] E[a_{j,k}] \\ =& \frac{1}{\sqrt{2\pi}} \sum_{k=-b}^{b} \left(1 - \frac{|k|}{n_v}\right) \gamma(k) a_{j,k} \end{aligned} \tag{2.12}$$

Now, $p(\lambda) = \frac{1}{\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} \gamma(k) \frac{e^{ik\lambda}}{\sqrt{2\pi}}$, so

$$\begin{aligned} E[K_j^*] &= E[\langle p, \phi_j \rangle] \\ &= \frac{1}{\sqrt{2\pi}} \sum_{k=-b}^{b} \gamma(k) a_{j,k} \end{aligned} \tag{2.13}$$

Then,

$$
\begin{aligned}
E[(\hat{K}_j - K_j^*)^2] &= \mathrm{Var}[\hat{K}_j] + (\mathrm{Bias}[\hat{K}_j])^2 \\
&= \mathrm{Var}\left[\frac{1}{\sqrt{2\pi}} \sum_{k=-b}^{b} \hat{\gamma}_2(k) a_{j,k}\right] \\
&\quad + \left(\frac{1}{\sqrt{2\pi}} \sum_{k=-b}^{b} \frac{|k|}{n_v} \gamma(k) a_{j,k}\right)^2 \\
&= \frac{1}{2\pi} \mathrm{Var}\left[\sum_{k=-b}^{b} \hat{\gamma}_2(k) a_{j,k}\right] \\
&\quad + \frac{2}{n_v^2 \pi} \left(\sum_{k=1}^{b} k\gamma(k) a_{j,k}\right)^2
\end{aligned}
\tag{2.14}
$$

$\hat{K}_j$ can be seen as a lag window spectral density estimator at $\lambda = 0$, except the kernel function is allowed to be negative and doesn't necessarily evaluate to 1 at zero. Parzen's (1957) formula for the variance of such an estimator does not require nonnegativity of the kernel function, but does require that it be normalized to $K(0) = 1$; we can fix the latter by replacing $a_{j,k}$ with $\frac{a_{j,k}}{a_{j,0}}$ and then multiplying the resulting formulaic variance by $a_{j,0}^2$. (This just cancels out.) As an asymptotic result, it also requires that the kernel function be continuous rather than discrete, so we interpolate $a_{j,k+x} = (1-x)a_{j,k} + x a_{j,k+1}$ for $0 < x < 1$. Then, applying Parzen's formula,

$$
\begin{aligned}
&\mathrm{Var}\left[\sum_{k=-b}^{b} \hat{\gamma}_2(k) a_{j,k}\right] \\
&= \left[\frac{2a_{j,0}^2 b}{n_v} p^2(0) \int_{-\infty}^{\infty} \frac{a_{j,k}^2}{a_{j,0}^2}\, dk\right] + o(b/n_v)
\end{aligned}
\tag{2.15}
$$

and plugging this into (2.14),

$$
\begin{aligned}
&E[(\hat{K}_j - K_j^*)^2] \\
=&\frac{b}{n_v \pi} p^2(0) \int_{-\infty}^{\infty} a_{j,k}^2 \, dk \, + \frac{2}{n_v^2 \pi} \left(\sum_{k=1}^{b} k\gamma(k) \right)^2 + o(b/n_v).
\end{aligned} \tag{2.16}
$$

$\sum_{k=-b}^{b} a_{j,k}^2 = 1$, so, by convexity of x^2, the integral is bounded above by 1. The square of the bias can be absorbed into the $o(b/n_v)$ term. We conclude that

$$
\begin{aligned}
&E[||\hat{p} - p||^2] \\
\leq & \min_{\hat{K_1}, \ldots, \hat{K_M}} \left|\left| \sum_{j=1}^{M} \hat{K}_j p_j - p \right|\right|^2 + \frac{bp^2(0)M}{n_v \pi} + o(bM/n_v).
\end{aligned} \tag{2.17}
$$

□

Next, we consider the exponential mixing. Defining $\alpha(\cdot)$ as in Definition A.0.1 in Politis (1999),

Theorem 2.2.2. *If $\frac{b}{n} \to 0$, $EX_t^4 < \infty$, the time series satisfies the α-mixing assumption $\alpha(k) \leq c^k$ for some constant $c > 1$ and all $k \geq n_b$, and n_b is chosen such that $n_b \geq (2+\epsilon)\log_c n$ for some $\epsilon > 0$, the L_2 risk of our estimator has the same upper bound as in Theorem 2.2.1.*

Proof: We wish for the dependence between the $\hat{\gamma}_2$'s and the $a_{j,k}$'s to have an impact of order $o(b/n)$ on $||\hat{p} - p||^2 - \min \left|\left| \sum_{j=1}^{M} \hat{K}_j p_j - p \right|\right|^2$.

By Lemma A.0.1 in Politis (1999), with $\xi = \hat{\gamma}_2(k), \zeta = a_{j,k}, p = 2$, and $q = \infty$, we have

$$
\begin{aligned}
&|\operatorname{Cov}(\hat{\gamma_2}(k), a_{j,k})| \\
&\leq 8(E|\hat{\gamma_2}|^2)^{1/2} \cdot 1 \cdot \sqrt{\alpha(n_b)}
\end{aligned} \tag{2.18}
$$

since $|a_{j,k}| \leq 1$ (because, by construction of the orthonormal basis, $\sum_j a_{j,k}^2 = 1$);

$$
\begin{aligned}
&\leq 8\sqrt{\frac{(n_v - k)^2}{n_v^2}\gamma^2(k) + \operatorname{Var}\hat{\gamma}_2(k)}\sqrt{\alpha(n_b)} \\
&= \Omega(8\gamma(k)\sqrt{\alpha(n_b)}) \\
&= \Omega(8\gamma(k)c^{-n_b/2})
\end{aligned} \tag{2.19}
$$

Plugging this back into $E[\hat{K}_j]$, we get an additional term with absolute value bounded by $\Omega\left(\frac{1}{\sqrt{2\pi}}\sum_{k=-b}^{b} 8\gamma(k)c^{-n_b/2}\right)$. Since we chose $n_b \geq (2+\epsilon)\log_c n$, $c^{-n_b/2} \leq n^{-1-(\epsilon/2)}$ so the term's impact on $E[\hat{K}_j]$ is $o(b/n)$. Thus, its impact on $E[(\hat{K}_j - K_j^*)^2]$ is also $o(b/n)$ as desired. □

Theorem 2.2.3. *If $\frac{b}{n} \to 0$, $EX_t^4 < \infty$, the time series satisfies the α-mixing assumption $\alpha(k) = O(k^{-c})$ for all $k \geq n_b$ and some $c > 2$, and n_b is chosen such that $n_b \geq n^{\frac{2}{c}+\epsilon}$ for some $\epsilon > 0$, the L_2 risk of our estimator has the same upper bound as in Theorem 2.2.1.*

Proof: The proof is identical to that of Theorem 2.2.2 up to (2.19). Plugging (2.19) into $E[\hat{K}_j]$, we get an additional term with absolute value bounded by $O\left(\frac{1}{\sqrt{2\pi}}\sum_{k=-b}^{b} 8\gamma(k)n_b^{-c/2}\right)$. Since we chose $n_b \geq n^{\frac{2}{c}+\epsilon}$, the term's impact on $E[\hat{K}_j]$ is $o(b/n)$, and the result follows. □

Remark. *If $\gamma(k)$ decays at only a polynomial rate, Theorem 3.1 from Politis (2011) is only able to bound $\min_{c_1,\ldots,c_M}\left|\left|\sum_{j=1}^{M} c_j p_j - p\right|\right|^2$ by a term of order $n_t^{\frac{1}{2r+1}-1}$, where*

$r \geq 1$ satisfies $\sum_{k=1}^{\infty} k^r \gamma(k) < \infty$. In this case, when the bandwidth candidates are of smaller order than $n_v^{\frac{1}{2r+1}}$, n_v should be larger than n_t.

However, if $\gamma(k)$ decays at least exponentially, the same theorem offers a bound of $O\left(\frac{\log n_t}{n_t}\right)$. In this case, if the bandwidth candidates increase more than logarithmically in n_v, we'll want to choose $n_v > n_t$.

2.3 Simulation Results

2.3.1 Bartlett Aggregation

The Bartlett kernel is defined by

$$w(x) = \begin{cases} 1 - |x| & \text{for } |x| < 1; \\ 0 & \text{elsewhere} \end{cases} \tag{2.20}$$

In the following simulations, we aggregate the estimators

$$p_j(\lambda) = \frac{1}{\sqrt{2\pi}} \sum_{k=-b_j}^{b_j} \hat{\gamma}_1(k) w\left(\frac{k}{b_j}\right) \frac{e^{ik\lambda}}{\sqrt{2\pi}}, \tag{2.21}$$

for various collections of b_js.

Let $\{Z_t\} \sim IID(0, \sigma^2)$. The MA(1) model $X_t = Z_t + \theta Z_{t-1}$ then has autocovariances $\gamma(0) = (1+\theta^2)\sigma^2$, $\gamma(1) = \theta\sigma^2$, and $\gamma(k) = 0$ for $k > 1$. From Politis (2003), the optimal large sample block size is $(6n)^{1/3} \left| \frac{\sum_{k=1}^{\infty} k\gamma(k)}{\sum_{k=-\infty}^{\infty} \gamma(k)} \right|^{2/3}$, which evaluates to $(6n)^{1/3} \left| \frac{\theta}{(1+\theta)^2} \right|^{2/3}$ in the MA(1) case. Most of our simulations use $\theta = 0.5$, for which this reduces to $\frac{2n^{1/3}}{3}$.

In the tables below, "length" denotes the length of the time series, the b_js in the aggregate are listed under "bandwidth", "avg. $\hat{K}$" denotes the average weight assigned by the aggregate to the bandwidth, and "MSE" is the empirical mean square error (MSE) of the kernel spectral density estimate. All values are averages over 200

Table 2.1: MA(1) $\theta = 0.5$ Bartlett aggregation results, optimal bandwidth with single alternative.

Length	Bandwidth	Avg. $\hat{K}$	MSE
100	3	.8391	.015955
	12	.2250	.028312
	agg.		.022932
500	5	.9608	.004253
	20	.0895	.008967
	agg.		.006289
1000	7	.9936	.002756
	28	.0437	.006518
	agg.		.003739
27000	20	.9869	.000274
	80	.0266	.000654
	agg.		.000293
125000	33	.9778	.000099
	133	.0280	.000237
	agg.		.000102
1000	6	-.5530	.002717
	7	1.5914	.002622
	agg.		.003234
1000	7	.9195	.002787
	14	.1186	.003483
	agg.		.003642
1000	7	1.0387	.002985
	50	-.0028	.011457
	agg.		.003721

trials, except for the length 125k time series (for which only 50 trials were averaged).

We first tried aggregations of two bandwidths, with one roughly optimal and the other much larger. Theoretically, we expect the optimal linear combination to basically ignore the second bandwidth, and this is what our aggregates tended towards doing. However, for smaller sample sizes, the lone inefficient alternative raised the MSE by close to 50%. This penalty was reduced to 5-10% once the sample size reached the tens of thousands; see blocks 1–5 of Table 2.1.

We then tried varying the alternative bandwidth; see blocks 6–8. There was no noticeable difference between the 2x optimal and 7x optimal alternatives. However, if the second bandwidth was instead a near-duplicate of the first, the MSE penalty was found lower. Of course, there would be little potential gain from aggregation in that case.

We then tried increasing the number of aggregate components, with geometric spreads of bandwidths. As expected, the MSE penalty was roughly linear in the number of components, and was more acceptable with larger sample sizes; see Table 2.2.

It did not really matter whether the aggregate included a near-optimal component; the (3, 5, 10) aggregate outperformed the (4, 7, 14) aggregate and the (3, 12, 48) aggregate noticeably outperformed the (7, 14, 28) aggregate for length 1k time series, despite the fact that the optimal bandwidth was about 7.

In the theory of kernel spectral estimation, the so-called 'flat-top' lag windows have been shown to have very favorable asymptotic and finite-sample properties, especially when the autocovariance decays quite rapidly. The simplest flat-top lag-window is the trapezoid proposed by Politis and Romano (1995); for the definition and properties of general flat-top lag windows see Politis (2001), Politis (2005) and Politis (2011).

Since the trapezoid can be constructed as a linear combination of two triangular (Bartlett) kernels, we wanted to investigate the conditions under which conditions the aggregate estimator would tend to approximate a trapezoid. Note, however, that the aggregate estimator shoots for minimum MSE, and the flat-top estimators only achieve optimal performance when their bandwidth is chosen to be sufficiently small. Hence, in Table 2.3 we investigate our aggregate's ability to outperform its near-optimal bandwidth component when a very low bandwidth component is also provided.

Indeed, the weight assignments chosen by the aggregate are trapezoid approximations, and the aggregate is able to achieve a MSE advantage of 20% with sample

Table 2.2: MA(1) $\theta = 0.5$ Bartlett aggregation results, geometric bandwidth spreads.

Length	Bandwidth	Avg. $\hat{K}$	MSE
100	2	-.7411	.019814
	3	.9637	.016115
	5	.8571	.017252
	agg.		.026471
100	2	-1.3568	.021141
	3	3.0841	.017913
	5	-1.1890	.019320
	8	.5268	.024443
	agg.		.031413
1000	3	-1.4652	.005982
	5	3.0790	.003430
	10	-.6013	.003141
	agg.		.003696
1000	4	-.6085	.004167
	7	1.7816	.002973
	14	-.1436	.003627
	agg.		.003801
1000	5	.2156	.003218
	10	.9896	.003148
	20	-.1754	.005082
	agg.		.004350
1000	7	.7596	.002926
	14	.5107	.003740
	28	-.2355	.006496
	agg.		.004669
1000	3	.1063	.005605
	12	1.0519	.003285
	48	-.1280	.010856
	agg.		.004144
1000	5	.1666	.003345
	10	.9547	.003253
	20	-.1622	.005148
	40	.0678	.009356
	agg.		.005392
27000	10	-.3527	.000480
	20	1.5431	.000255
	40	-.1843	.000340
	agg.		.000289
27000	10	-.3709	.000510
	20	1.6316	.000289
	40	-.2834	.000377
	80	.3200	.000683
	agg.		.000338

Table 2.3: MA(1) $\theta = 0.5$ Bartlett aggregation results, two-bandwidth trapezoid discovery simulations.

Length	Bandwidth	Avg. $\hat{K}$	MSE
100	1	-.6542	.046679
	3	1.7287	.016935
	agg.		.018653
500	1	-.2461	.041030
	5	1.2560	.004652
	agg.		.003629
1000	1	-.1461	.040431
	7	1.1477	.002836
	agg.		.002472
27000	1	-.0542	.039811
	20	1.0544	.000283
	agg.		.000185
27000	2	-.0848	.009934
	20	1.0842	.000269
	agg.		.000212
27000	3	-.1285	.004485
	20	1.1316	.000293
	agg.		.000228
125000	1	-.0298	.039795
	33	1.0311	.000096
	agg.		.000059
125000	2	-.0528	.009884
	33	1.0503	.000092
	agg.		.000069
125000	3	-.0901	.004471
	33	1.0915	.000101
	agg.		.000073
125000	1	-.0240	.039793
	40	1.0227	.000096
	agg.		.000077
125000	3	-.0516	.004406
	40	1.0527	.000102
	agg.		.000089
125000	5	-.0825	.001601
	40	1.0852	.000098
	agg.		.000088
125000	1	-.0143	.039795
	60	1.0158	.000124
	agg.		.000115
125000	3	-.0261	.004442
	60	1.0283	.000120
	agg.		.000117
125000	5	-.0141	.001631
	60	1.0187	.000119
	agg.		.000119

sizes in the hundreds, which rises to close to 40% in the 125k sample size case. However, the trapezoid's advantage appears to vanish as soon as the primary bandwidth reaches 2x optimal.

The particularly favorable performance of the aggregates including a bandwidth 1 component in the last batch of simulations suggested that geometric bandwidth spreads starting from 1 might significantly outperform the spreads investigated in Table 2.2. This is in fact the case; see Table 2.4. While previously the aggregate did not outperform the best individual component even with a length 27k time series, now we see outperformance at length 4k, and by 27k it is by more than a factor of 2. Note that, in the length 4k case, the two additional bandwidths roughly double the MSE compared to the simple trapezoid aggregate, but the procedure would still be worthwhile if one was not aware of the value of using trapezoidal kernels directly.

In Table 2.5 we tried using our procedure just to select a bandwidth (picking the one assigned the highest weight). Performance was very poor; in fact, the best bandwidth was never selected the most frequently in any test case.

Finally, we tried aggregating Epanechnikov-Priestley kernels, i.e.

$$w(x) = \begin{cases} \frac{3}{4}(1-x^2) & \text{for } |x| < 1; \\ 0 & \text{elsewhere.} \end{cases} \tag{2.22}$$

There is no exact result involving linear combinations of these kernels that is analogous to the relation between trapezoidal and Bartlett kernels. However, for the largest sample sizes our aggregate was able to significantly outperform all the individual components, and across all sample sizes the aggregate never had MSE worse than twice the best individual component.

Table 2.4: Geometric bandwidth spreads starting at 1.

Length	Bandwidth	Avg. $\hat{K}$	MSE
100	1	-.7459	.046781
	2	4.5984	.022110
	3	-11.690	.018185
	4	8.9482	.017763
	agg.		.024602
500	1	-1.0408	.041107
	2	1.8388	.001223
	4	.5200	.005488
	8	-.3026	.005160
	agg.		.006807
4000	1	-.6016	.039950
	3	1.9790	.004779
	7	-.4626	.001348
	15	.0903	.001122
	agg.		.000830
27000	1	-.3707	.039813
	4	1.5363	.002495
	15	-.1786	.000299
	50	.0127	.000437
	agg.		.000135
125000	1	-.2439	.039796
	5	1.2143	.001604
	25	.0287	.000115
	125	.0036	.000230
	agg.		.000027
4000	1	-.4991	.039954
	3	1.4966	.004734
	agg.		.000386

Table 2.5: Model selection.

Length	Bandwidth	Selection freq.	MSE
100	1	.010	.046307
	2	.370	.021443
	3	.335	.017748
	4	.285	.017466
	avg.		.019362
500	1	.000	.040951
	2	.540	.012115
	4	.360	.005291
	8	.100	.004772
	avg.		.008814
4000	1	.000	.039972
	3	.570	.004650
	7	.365	.001285
	15	.065	.001108
	avg.		.003172
27000	1	.000	.039809
	4	.750	.002487
	15	.240	.000294
	50	.010	.000415
	avg.		.001944
125000	1	.00	.039795
	5	.82	.001627
	25	.18	.000116
	125	.00	.000217
	avg.		.001341

Table 2.6: Epanechnikov-Priestley kernels.

Length	Bandwidth	Avg. $\hat{K}$	MSE
100	1	.1070	.047035
	2	-19.102	.014780
	3	66.422	.015464
	4	-46.390	.018206
	agg.		.024315
500	1	-.4764	.040901
	2	1.8497	.004509
	4	-.2803	.003498
	8	-.0877	.006134
	agg.		.005989
4000	1	-.1417	.039942
	3	1.3346	.000836
	7	-.1769	.000726
	15	-.0177	.001411
	agg.		.001004
27000	1	-.0707	.039811
	4	1.1460	.000219
	15	-.0823	.000199
	50	.0074	.000658
	agg.		.000161
125000	1	-.0471	.039794
	5	1.2220	.000084
	25	-.1909	.000067
	125	.0151	.000349
	agg.		.000037

2.4 Conclusions

We presented an aggregation procedure for kernel spectral density estimators with asymptotically optimal performance. Our simulations verified that the aggregate consistently performed within a factor of two (in MSE terms) of its best component, and that it was capable of discovering nontrivial optimal linear combinations such as the trapezoid kernel.

The procedure works best with large sample sizes (> 1000), but reasonable results were obtained with a sample size as small as 500. It is particularly important to minimize the number of aggregate components (preferably to two) in the latter case, since there is a large error term linear in the number of components; however, this term has favorable asymptotics, so very large sample sizes allow diverse aggregates to be employed at minimal cost.

The viability of the first aggregation step as a model selection procedure was also briefly investigated via simulation, and we found that it was unsuitable.

2.5 Acknowledgements

Chapter 2 is essentially a reprint, with minor modifications, of the paper "Aggregation of Spectral Density Estimators" by C. Chang and D.N. Politis, which has been submitted for publication in *IEEE Transactions on Information Theory.* The dissertation author was the primary investigator and author of this paper.

Chapter 3

Robust Autocorrelation Estimation

3.1 Introduction

The estimation of the autocorrelation function plays a crucial role in time series analysis. For example, in the common case where a time series is modeled as an AR process, the model coefficient estimates are straightforward functions of the estimated autocorrelations [4].

Given a stationary time series $X_1, \ldots, X_n$, recall that the autocovariance function (acvf for short) is $\gamma(h) := E[(X_{t+h} - \mu)(X_t - \mu)]$ (where $\mu := E[X_t]$), and the autocorrelation function (acf for short) is $\rho(h) := \gamma(h)/\gamma(0)$. The classical estimator of the acf is the sample acf:

$$\hat{\rho}(h) := \hat{\gamma}(h)/\hat{\gamma}(0)$$

where $\hat{\gamma}$ is the sample acvf:

$$\hat{\gamma}(h) := n^{-1} \sum_{j=1}^{n-h} (X_{j+h} - \bar{X})(X_j - \bar{X}) \quad \text{(where } \bar{X} := n^{-1} \textstyle\sum_{j=1}^{n} X_j\text{)}.$$

Unfortunately, the sample acf is not a robust statistic–contamination of a single point is enough to clobber the rest of the data and drive the estimate, masking the real dependence structure. In practice, it is not uncommon for 10% or more of measured time series values to be outliers [15], so this weakness is highly relevant.

In the past, the computational advantages enjoyed by the classical estimator over robust techniques justified its near-universal usage, sometimes in combination with an outlier identification method to patch its weakness. However, thanks to a massive increase in available computing power, robust estimation is now frequently practical, and it's far from clear that classical estimation plus outlier elimination yields better results than just using an intrinsically robust estimator.

The remainder of this paper is structured as follows: In section 3.2, we introduce a new class of robust autocorrelation estimators, based on interpreting the sample autocorrelation as a linear regression. Next, in section 3.3, we analyze the estimators that result from plugging in three common robust regression techniques, and compare their performance to that of the sample acf. Then, in sections 3.4-3.5, we discuss the derivation of autocovariance and positive definite autocorrelation estimates from our initial estimator. We apply our method to AR model fitting in section 3.6. Finally, we present the results of a simulation study in section 3.7.

3.2 Robust acf estimation

Assume we have time series data $X_1, \ldots, X_n$ generated by a second-order stationary process (except for outliers), i.e. [20]

$$(i)\, E(X_t^2) < \infty \qquad \forall t$$

$$(ii)\, E(X_t) = \mu = \text{constant} \qquad \forall t$$

$$(iii)\ \text{cov}(X_{t+h}, X_t) = \gamma(h) \qquad \forall t, h$$

Fix $h < n$ where $h \in \mathbb{Z}^+$. If the time series is Gaussian, we have $E[X_{t+h} - \mu | X_t] = (X_t - \mu)\rho(h)$ for $t \in \{1, \ldots, n-h\}\}$. This motivates the following idea: create a scatterplot with the points $\{(X_t - \bar{X}, X_{t+h} - \bar{X}), t \in \{1, \ldots, n-h\}\}$ (where the x-coordinate is first); then use the slope of a regression line on the points as an estimate of autocorrelation. It is well known that this regression slope estimate of ρ is valid even if the time series is not Gaussian. [1]

See Figure 3.1 for an example. Indeed, the least-squares estimate of slope is almost identical to the sample acf for $\frac{h}{n}$ small. If the points in the scatterplot are denoted $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, then the ordinary least squares (OLS) estimate of slope is

[1] Since the independent variables are not known precisely–'errors-in-variables' –a technique like orthogonal regression may be appropriate [13]. However, we do not pursue this here, since robust estimation has been more thoroughly studied in the context of linear regression, and some robust linear regression techniques are resistant to outliers in the x-coordinates. See Zamar [46] for a discussion of robust estimation under errors-in-variables.

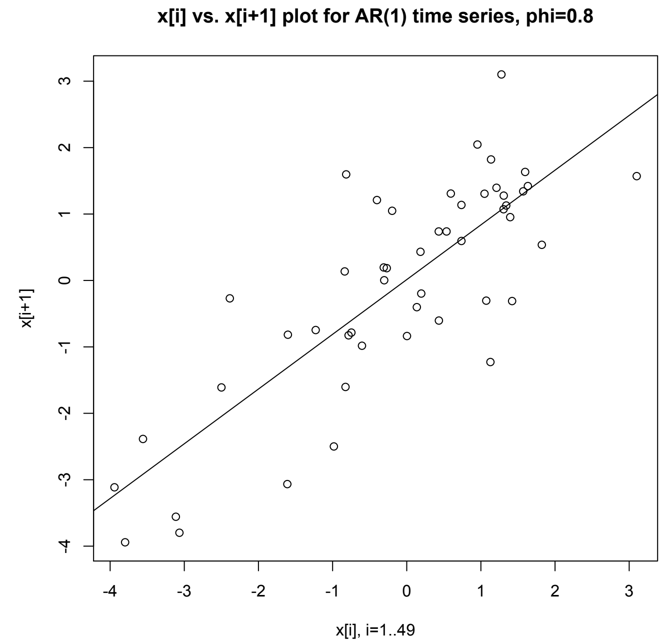

Figure 3.1: Scatterplot of (X_t, X_{t+1}) for a realization of the AR(1) time series $X_t = 0.8X_{t-1} + Z_t$, Z_t iid $N(0,1)$. Regression line is $y = 0.82375x + 0.01289$.

$$\begin{aligned}
\hat{\rho}_{OLS}(h) &= \frac{\sum_{j=1}^{n-h}(x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^{n-h}(x_j - \bar{x})^2} \\
&= \frac{\sum_{j=1}^{n-h}(x_{j+h} - \bar{x}_{(h+1)\dots n})(x_j - \bar{x}_{1\dots(n-h)})}{\sum_{j=1}^{n-h}(x_j - \bar{x}_{1\dots(n-h)})^2} \\
&\approx \frac{\sum_{j=1}^{n-h}(x_{j+h} - \bar{x})(x_j - \bar{x})}{\frac{n-h}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2} \\
&= \frac{n}{n-h}\hat{\rho}(h)
\end{aligned}$$

where $\bar{x}_{a\dots b} := (b - a + 1)^{-1}\sum_{j=a}^{b} x_j$ and $\bar{x} := \bar{x}_{1\dots n}$.

The additional $\frac{n}{n-h}$ factor is expected, since the regression slope is an unbiased estimator while the sample acf is biased low by construction. The only other difference is the inclusion/exclusion of the first and last time series points in computing sample mean and variance; the impact of that is negligible.

The implication is that if we run a robust linear regression on $\{(X_t, X_{t+h})\}$, we should get a robust estimate of autocorrelation. (Since we are only interested in the slope, the $(-\bar{X}, -\bar{X})$ displacement can be dropped.) This is then our proposal for robust acf estimation.

To fix ideas, we investigate in detail three estimators in this class:

1. $\hat{\rho}_{L1}$. Recall that a residual r_i of a linear regression is the vertical distance between the point (x_i, y_i) and the regression line, i.e. $r_i = y_i - (ax_i + b)$ where a is the slope and b the intercept of the regression line. The simplest robust regression technique, L1 regression, minimizes the sum of absolute residuals instead of the sum of squares of those residuals; the effect is to find a "median regression line".

2. $\hat{\rho}_{LTS}$. Least trimmed squares regression, or LTS for short, takes a different approach: instead of changing the pointwise loss function, we use the usual squared residuals but throw the largest values out of the sum. More precisely, define $|r|_{(1)} \leq \ldots \leq |r|_{(n-h)}$ to be the ordered residual absolute values. Then α-trimmed squares minimizes

$$\hat{\sigma} := \left(\sum_{j=1}^{\lceil (1-\alpha)(n-h) \rceil} |r|_{(j)}^2 \right)^{1/2} .$$

We look at α-trimmed squares for $\alpha = \frac{1}{2}$ (so we sum up to the median absolute residual).

3. $\hat{\rho}_{MM}$. An M-estimate [16] minimizes

$$L(\beta) := \sum_{i=1}^{n} \ell \left(\frac{r_i(\beta)}{\hat{\sigma}} \right).$$

for some pointwise loss function ℓ, where $\hat{\sigma}$ is an estimate of the scale of the residuals.

It is efficient, but not resistant to outliers in the x values. A "redescending" M-estimate utilizes a loss function with derivative decreasing to zero at the tails.

In contrast, an S-estimate (S for "scale") minimizes a robust estimate of the scale of the residuals:

$$\hat{\beta} := \operatorname*{argmin}_{\beta} \hat{\sigma}(\mathbf{r}(\beta))$$

where $\mathbf{r}(\beta)$ denotes the vector of residuals and $\hat{\sigma}$ satisfies

$$\frac{1}{n}\sum_{j=1}^{n-h} \ell\left(\frac{r_j}{\hat{\sigma}}\right) = \delta.$$

(δ is usually chosen to be $\frac{1}{2}$.) It has superior robustness, but is inefficient.

MM-estimates, pioneered by Yohai (1987), combine these two techniques in a way intended to retain the robustness of S-estimation while gaining the asymptotic efficiency of M-estimation. Specifically, an initial robust-but-inefficient estimate $\hat{\beta}_0$ is computed, then a scale M-estimate of the residuals, and finally the iteratively reweighted least squares algorithm is used to identify a nearby $\hat{\beta}$ that satisfies the redescending M-estimate equation.

For further discussion of these three robust regression techniques, see Maronna (2006).

3.3 Theoretical Properties

3.3.1 General

We focus our attention on normal efficiency and two measures of robustness (breakdown point and influence function).

Relative normal efficiency is the ratio between the asymptotic variance of the classical estimator and that of another estimator under consideration, assuming Gaussian residuals and no contamination. This is a measure of the price we are paying for any robustness gains.

The breakdown point (BP) is the asymptotic fraction of points that can be contaminated without entirely masking the original relation. Now, in the case of time series and ARMA processes, we distinguish two types of outliers (Denby (1979)):

1. *innovation outliers* that affect all subsequent observations, and can be observed in a pure ARMA process with a heavy-tailed innovation distribution.

2. *additive outliers* or replacement outliers that exist outside the ARMA process and do not affect other observations. For second-order stationary data, the difference between them is minimal (a replacement outlier functions like a slightly variable additive outlier), so for brevity we just concern ourselves with additive outliers.

For additive outliers, the classical autocorrelation estimator has a breakdown point of zero since a single very large outlier is enough to force the estimate to a neighborhood of $\frac{-1}{n-h}$ (see Figure 3.2). Since one additive outlier influences the position of at most two points in the regression, our robust autocorrelation estimators will exhibit BPs at least half that of the robust regression techniques they are built on. (See Ma and Genton (2000) on "temporal breakdown point" for a more exhaustive discussion.)

The impact of an innovation outlier on the regression line varies. For instance, only one point is moved off the regression line in the AR(1) case, but three points are affected in the MA(1) case. So in the former scenario, our robust autocorrelation estimators can be expected to fully inherit the BPs of the robust regressors with respect to innovation outliers, but we cannot expect as much reliability with MA models.

Interestingly, infinite variance symmetric alpha-stable innovation distributions result in a faster sample acf convergence rate than the finite variance innovation case

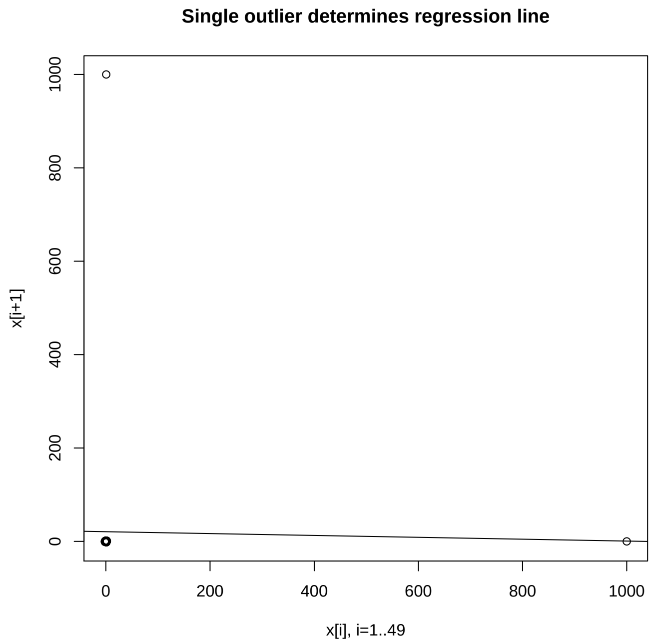

Figure 3.2: Degenerate OLS regression line from 50 N(0,1) points contaminated by one outlier at 1000.

(Davis (2000)); this is possible because the innovation outliers create high leverage points in the scatterplot that are very close to the "correct" regression line. We will investigate whether our robust regression estimates keep up.

Next, the influence function (IF) describes the impact on an autocorrelation estimate $\hat{\rho}$ of adding an infinitesimal probability of an outlier. For additive outliers, it is defined as follows:

$$IF(x, \hat{\rho}, F) := \lim_{\epsilon \to 0^+} \frac{\hat{\rho}((1-\epsilon)F + \epsilon\Delta_x) - \hat{\rho}(F)}{\epsilon}$$

for x such that this limit exists, where F is the time series distribution and Δ_x denotes a probability point mass at x. This is a measure of the asymptotic bias caused by observation contamination (Ma (2000)). We use a similar definition for innovation outliers under an ARMA model: letting G be the innovation distribution and $F(G)$ the resulting time series distribution,

$$IF(x, \hat{\rho} \circ F, G) := \lim_{\epsilon \to 0^+} \frac{\hat{\rho}(F((1-\epsilon)G + \epsilon\Delta_x)) - \hat{\rho}(F(G))}{\epsilon}$$

For the classical estimator, the value of the influence function increases without bound as $|x| \to \infty$ for both additive and innovation outliers, since the numerator in the limit converges to a nonzero constant while the denominator goes to zero.

Finally, we note that our robust autocorrelation estimates are not guaranteed to be in the range [-1, 1]; consider the time series $\{1, 2, 0\}$, which defines a slope -2 regression line for $h = 1$. See section 5 on making our estimate mathematically better-behaved.

3.3.2 L1

Because the x-coordinates are not fixed, $\hat{\rho}_{L1}$ does not inherit all the asymptotic robustness advantages normally enjoyed by L1 regression. Any outlier in the middle of the time series appears as both an x- and a y-coordinate, and while L1 regression shrugs off the y outlier, the x outlier point can have an extreme influence on it. Therefore, the BP is zero in the additive outliers case and the influence function increases without bound again. Since, if the underlying process is AR(1), an additive outlier can have an effect similar to that of two adjacent innovation outliers, the theoretical bounds are no better in the innovation outliers case.

3.3.3 LTS

LTS regression exhibits the highest possible breakdown point ($\frac{1}{2}$). It is robust with respect to both x- and y-outliers, so $\hat{\rho}_{LTS}$ retains the $\frac{1}{2}$ BP in the AR(1) innovation outliers case and has a BP of at least $\frac{1}{4}$ with respect to additive outliers. The influence function flattens at the tails since the probability of mistaking the outlier for a "real" point declines exponentially in n.

It also exhibits the optimal convergence rate, but has a very low normal efficiency of around 7%; cf. Rousseeuw (1987) for details.

3.3.4 MM

MM-estimates also have an asymptotic breakdown point of $\frac{1}{2}$ and are resistant to both x- and y-outliers, so $\hat{\rho}_{MM}$ has a BP of $\frac{1}{2}$ in the innovation outliers case and at least $\frac{1}{4}$ in the additive outliers case. The influence function flattens because a robust estimate of residual scale is used.

The normal efficiency is actually a user-adjustable parameter. In practice, it it is usually chosen to be between 0.7 and 0.95; aiming for an even higher normal efficiency results in too large a region where the MM-estimate tracks the performance of the classical estimator rather than exhibiting the S-estimate's robustness. We use 0.85 in our simulations.

3.4 Robust Autocovariance Estimation

In order to derive an autocovariance estimate from our robust regression slopes, we need to multiply by some estimate of variance. Here, we present a way to obtain this estimate using the robust regression insight.

Our first objective is to obtain a robust estimate of location. Now, from each robust autocorrelation regression we perform, we can derive an estimate of the process mean μ as a function of the estimated slope and intercept:

$$
\begin{aligned}
Y_t &= \beta_0 + \beta_1 Y_{t-h} + \text{error} && (3.1)\\
Y_t - \mu &= \beta_1(Y_{t-h} - \mu) + \text{error, since this line should have zero intercept}\\
Y_t &= \mu + \beta_1 Y_{t-h} - \beta_1\mu + \text{error} && (3.2)\\
\beta_0 &= \mu(1-\beta_1) \qquad \text{(combining (3.1) and (3.2))}\\
\hat{\mu} &:= \frac{\hat{\beta}_0}{1-\hat{\beta}_1}
\end{aligned}
$$

Each value of $h = 1, \ldots, H$ (for some H) yields a distinct $\hat{\mu}$, so we use L1 (i.e. compute the median) or LTS to aggregate these into a single estimate.

Since

$$(Y_t - \mu)^2 = \gamma(0) + \text{error},$$

we can then estimate $\gamma(0)$ by using L1 or LTS on our centered sample values $(Y_t - \hat{\mu})^2$; denote this estimator by $\hat{\gamma}(0)$. Finally, we multiply $\hat{\rho}(h)$ by $\hat{\gamma}(0)$ to get a robust estimate $\hat{\gamma}(h)$ of $\gamma(h)$.

We note that Ma and Genton's (2000) robust autocovariance estimator is an alternative here.

3.5 Robust and positive definite estimation of autocorrelation and autocovariance matrices

The most obvious way to robustly estimate the autocorrelation matrix Σ (where $\Sigma_{i,j} = \rho(|i-j|)$; $i, j = 1, \ldots, q$ for some $q \leq n$) is by plugging our robust correlation estimates directly into the diagonals and subdiagonals; designate this matrix by $\hat{\Sigma}$. (I.e. $\hat{\Sigma}_{i,j} := \hat{\rho}(|i-j|)$.) Unfortunately, this is not guaranteed to be positive definite.

However, following McMurry and Politis (2010), we can define a tapered weight function κ as

$$\kappa(x) = \begin{cases} 1 & \text{if } |x| \leq 1 \\ g(|x|) & \text{if } 1 < |x| \leq c_\kappa \\ 0 & \text{if } |x| > c_\kappa, \end{cases}$$

where $|g(x)| < 1$ and $c_\kappa \geq 1$ is some constant, and let the l-scaled version be denoted as $\kappa_l(x) := \kappa(x/l)$. Also define the tapered estimator

$$\hat{\Sigma}_{\kappa,l} = [\kappa_l(i-j)\hat{\gamma}_{|i-j|}]_{i,j=1}^{q}.$$

Fix κ and l. If TDT^t is the spectral decomposition of $\hat{\Sigma}_{\kappa,l}$ (T is an orthogonal matrix, and $D = \text{diag}(d_1, \ldots, d_n)$ which are the eigenvalues of $\hat{\Sigma}_{\kappa,l}$), define

$$D^\epsilon := \text{diag}(d_1^\epsilon, \ldots, d_n^\epsilon),$$
$$\text{where } d_i^\epsilon := \max(d_i, \epsilon/n^\beta).$$

Then

$$\hat{\Sigma}_{\kappa,l}^\epsilon := TD^\epsilon T^t \tag{3.3}$$

is positive definite for any positive β and ϵ.

McMurry and Politis (2010) have observed that the parameter choice $\beta = 1$, $\epsilon = 1$ with $g(x)$ linear (so κ is trapezoidal) works well in practice. Choosing l is also addressed by McMurry and Politis (2010) in the difficult case where q is large (even the case $q = n$); if q is small w.r.t. n, tapering is not necessary and estimator (3.3) is applicable with $l = n$.

3.6 Application to AR Model Fitting

3.6.1 Direct method

In the context of a pure AR(p) model $X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + Z_t$, autocovariance estimates are often directly used to derive AR coefficient estimates via the Yule-Walker equations:

$$\begin{aligned} \Sigma_p \underline{\phi}_p &= \underline{\gamma}_p \\ \sigma^2 &= \gamma(0) - (\underline{\phi}_p)' \underline{\gamma}_p \end{aligned}$$

where Σ_p is the autocovariance matrix, $\underline{\phi}_p = (\phi_1, \ldots, \phi_p)'$, and $\underline{\gamma}_p = (\gamma_1, \ldots, \gamma_p)'$

However, if the standard autocovariance estimates are used, a single outlier of size B perturbs the coefficient estimates by $O(B/n)$, and a pair of such outliers can perturb $\hat{\phi}_1$ by $O(B^2/n)$.

One way to address this vulnerability is to plug the robust, positive definite autocovariance matrix estimate discussed in the previous section into the linear system. (Note that a positive definite matrix is necessary to ensure the system is solvable.) For p small w.r.t. n, compute $\hat{\Sigma}^{\epsilon}_{\kappa,l}$ from (3.3) with $\kappa(x) = 1$ everywhere, $l = n$, $\epsilon = 1$, and $q = p$; then solve the Yule-Walker equation $\hat{\Sigma}^1_{\kappa,n}\phi = \hat{\gamma}_p$ where $\hat{\gamma}_p$ is the first column of $\hat{\Sigma}^1_{\kappa,n}$. The algorithm is similar for large p, just with different choices of κ and l.

3.6.2 Extended Yule-Walker method

Another technique for increasing robustness, which can be used simultaneously, was explored by Politis (2009). He observed that the 'extended' Yule-Walker equations yield additional valid estimators for the AR coefficients; e.g. for an AR(1), valid estimators for ϕ_1 include $\hat{\gamma}_1/\hat{\gamma}_0$, $\hat{\gamma}_2/\hat{\gamma}_1$, $\hat{\gamma}_3/\hat{\gamma}_2$, etc. Thus, in the AR(1) case, a straight line regression on the $(\hat{\gamma}_k, \hat{\gamma}_{k+1})$ scatterplot (with no intercept term) yields an estimator of ϕ_1 that is somewhat resistant to individual anomalous $\hat{\gamma}_k$s.

Generalizing this idea, fix $p' \geq p$, and let $\underline{\phi}_p := (\phi_1, \ldots, \phi_p)'$, $\underline{\gamma}_k := (\gamma_1, \ldots, \gamma_k)'$, $\underline{\hat{\gamma}}_k := (\hat{\gamma}_1, \ldots, \hat{\gamma}_k)'$. Denote the $p' \times p$ matrix with jth column equal to $(\gamma_{1-j}, \gamma_{2-j}, \ldots, \gamma_{p'-j})$ by $\Sigma_{p',p}$. Then the extended Yule-Walker equations up to $k = p'$ are given by

$$\underline{\gamma}_{p'} = \Sigma_{p',p}\underline{\phi}_p$$

Following Politis (2009), define $\hat{\Sigma}_{p',p}$ to be the $p' \times p$ matrix with jth column $(\hat{\gamma}_{1-j}, \hat{\gamma}_{2-j}, \ldots, \hat{\gamma}_{p'-j})$, and write

$$\underline{\hat{\gamma}}_{p'} = \hat{\Sigma}_{p',p}\underline{\phi}_p + \underline{\epsilon}, \tag{3.4}$$

which defines an error vector $\underline{\epsilon}$.

Equation (3.4) can be viewed as a multivariate linear regression with 'errors-in-variables', and identical x- and y-axis scales; running the regression gives us an estimate of $\underline{\phi}_p$. To ensure uniqueness of the solution, plug the first p columns of $\hat{\Sigma}^{\epsilon}_{\kappa,l}$ from (3.3) (with $q = p'$) rather than the raw autocovariance estimates into equation (3.4).

3.7 Simulation Results

3.7.1 Baseline

First, we generated time series data $X_1, \ldots, X_n$ according to the MA(1) model $X_t = Z_t + \phi Z_{t-1}$ (with no outliers) with $\phi \in \{0.2, 0.5, 0.8\}$, $n \in \{50, 200, 800\}$, and Z_t i.i.d. N(0, 1). We estimated the lag-1 and lag-2 autocorrelations, and compared them to the true values ($\frac{\phi}{1+\phi^2}$ and 0, respectively).

As baselines for comparison, we included OLS regression, which as discussed above is nearly identical to the sample acf, and Ma and Genton's (2000) robust autocorrelation estimator (denoted as MG).

We did the same thing for the AR(1) model $X_t = \phi X_{t-1} + Z_t$. (True autocorrelations are ϕ and ϕ^2 in this case.)

As expected, the OLS (classical) estimator performed best in the no contamination case. (See Tables **??**-3.2.) However, the MM estimator's performance was nearly indistinguishable from OLS's. The L1 and Ma-Genton estimators were somewhat less efficient, with MSEs roughly 1.5x to 2x that of the OLS estimator, and LTS's known terrible normal efficiency was clearly in evidence.

Sample size did not affect the performance of the estimators relative to each other, but a larger sample size reduced the downward bias of them all.

3.7.2 Innovation Outliers

Next, we investigated estimator performance in the face of innovation outliers, modifying Z_t to be distributed according to a Gaussian mixture, 90 or 96 percent N(0, 1) and 10 or 4 percent N(0, 625).

From Table 3.3, we can see that for $\phi = -0.2$, the Ma-Genton, L1, and MM estimators do a substantially better job of handling the innovation outliers than the sample acf. However, for larger values of ϕ and large sample sizes, our robust estimates of $\rho(1)$ cluster toward ϕ instead of $\frac{\phi}{1+\phi^2}$, because any innovation outlier not immediately followed by a second one creates a point of the form $(x + \epsilon_1, \phi x + \epsilon_2)$ (where $|x| >> |\epsilon_i|$)–all of these high-magnitude points trace a single line of slope ϕ which are picked up by the robust estimators as the primary signal, and the other high-magnitude outlier points (which bring the OLS estimate in line) are ignored. The Ma-Genton estimator, not being based on linear regression, is not affected by

Table 3.1: Uncontaminated MA(1) simulation results, averages of 200 trials.

ϕ	n	Estimator	Avg. $\hat{\rho}(1)$	MSE	Avg. $\hat{\rho}(2)$	MSE
0.2	50	OLS	.16815	.01669	-.04035	.02312
		MG	.17428	.02465	-.03364	.03676
		L1	.15741	.02938	-.04618	.03622
		LTS	.12148	.11283	-.06980	.13513
		MM	.16728	.01731	-.04223	.02533
	200	OLS	.18238	.00458	-.01316	.00546
		MG	.18174	.00629	-.01753	.00714
		L1	.18875	.00827	-.01559	.00861
		LTS	.18659	.04622	-.02034	.04300
		MM	.18328	.00489	-.01330	.00574
	800	OLS	.19202	.00120	.00173	.00108
		MG	.19266	.00127	.00152	.00135
		L1	.19457	.00190	.00080	.00213
		LTS	.20289	.01614	.00342	.01447
		MM	.19253	.00122	.00154	.00123
0.5	50	OLS	.35834	.01685	-.03677	.02702
		MG	.36166	.02319	-.02692	.03660
		L1	.35859	.02194	-.01190	.03290
		LTS	.38351	.07726	.00142	.10233
		MM	.35940	.01748	-.02757	.02745
	200	OLS	.39859	.00216	-.00520	.00516
		MG	.39992	.00308	-.00571	.00707
		L1	.39810	.00520	-.00163	.00862
		LTS	.40652	.03394	.01994	.04868
		MM	.39731	.00252	-.00528	.00560
	800	OLS	.39746	.00094	-.00465	.00183
		MG	.39809	.00111	-.00344	.00239
		L1	.39897	.00175	-.00113	.00258
		LTS	.39555	.01439	.00574	.01894
		MM	.39780	.00100	-.00395	.00199
0.8	50	OLS	.45355	.01053	-.05546	.03023
		MG	.45369	.01663	-.06168	.04081
		L1	.44862	.01992	-.06792	.04046
		LTS	.46865	.08112	-.06159	.12601
		MM	.45345	.01106	-.05628	.03074
	200	OLS	.48315	.00242	-.00775	.00667
		MG	.48289	.00322	-.00604	.00877
		L1	.48248	.00470	-.00235	.00847
		LTS	.49077	.02759	.02308	.03534
		MM	.48340	.00256	-.00730	.00663
	800	OLS	.48415	.00055	-.00434	.00166
		MG	.48349	.00067	-.00541	.00186
		L1	.48356	.00121	-.00320	.00202
		LTS	.47204	.01296	.00645	.01402
		MM	.48402	.00059	-.00436	.00166

Table 3.2: Uncontaminated AR(1) simulation results, averages of 200 trials.

ϕ	n	Estimator	Avg. $\hat{\rho}(1)$	MSE	Avg. $\hat{\rho}(2)$	MSE
0.2	50	OLS	.16358	.02592	.02875	.01956
		MG	.15360	.03837	.02700	.03465
		L1	.17565	.03564	.01710	.03553
		LTS	.18526	.11907	-.00201	.12429
		MM	.16702	.02758	.02804	.02197
	200	OLS	.20110	.00439	.02818	.00552
		MG	.20064	.00550	.02512	.00688
		L1	.19851	.00733	.02330	.00762
		LTS	.19576	.04101	.02079	.03917
		MM	.20009	.00459	.02691	.00562
	800	OLS	.19193	.00125	.04054	.00123
		MG	.19286	.00162	.04009	.00146
		L1	.19139	.00206	.04056	.00211
		LTS	.19555	.01551	.05124	.01590
		MM	.19191	.00137	.04066	.00124
0.5	50	OLS	.44600	.01603	.18352	.02630
		MG	.44176	.02597	.18445	.03796
		L1	.45312	.02454	.19821	.03591
		LTS	.46085	.09045	.21308	.11105
		MM	.44471	.01738	.18691	.02687
	200	OLS	.48241	.00417	.23662	.00681
		MG	.47893	.00494	.23194	.00776
		L1	.48157	.00635	.23560	.00937
		LTS	.48630	.03007	.22803	.03912
		MM	.48229	.00429	.23674	.00699
	800	OLS	.49777	.00100	.24495	.00157
		MG	.49708	.00125	.24396	.00202
		L1	.49994	.00147	.24465	.00210
		LTS	.50000	.00983	.24269	.01308
		MM	.49796	.00105	.24512	.00165
0.8	50	OLS	.72894	.01682	.52273	.04186
		MG	.70482	.02413	.48780	.05783
		L1	.72172	.02256	.51311	.05671
		LTS	.69385	.06811	.49295	.15527
		MM	.72896	.01790	.51800	.04563
	200	OLS	.78556	.00191	.61795	.00502
		MG	.78135	.00235	.61327	.00565
		L1	.78586	.00291	.61878	.00646
		LTS	.78713	.01646	.61040	.03228
		MM	.78498	.00193	.61847	.00489
	800	OLS	.79622	.00045	.63450	.00142
		MG	.79563	.00052	.63324	.00166
		L1	.79702	.00076	.63717	.00185
		LTS	.80020	.00573	.64809	.00765
		MM	.79634	.00048	.63522	.00149

Table 3.3: MA(1) simulation results with innovation outliers, averages of 200 trials.

ϕ	Contam. %	n	Estimator	Avg. $\hat{\rho}(1)$	MSE	Avg. $\hat{\rho}(2)$	MSE
-0.2	4	50	OLS	-.19785	.01077	-.02147	.01086
			MG	-.18534	.02753	-.03046	.03823
			L1	-.17678	.00886	-.01231	.00707
			LTS	-.16145	.07133	-.01445	.05180
			MM	-.18117	.00742	-.00452	.00842
		800	OLS	-.19071	.00112	-.00615	.00154
			MG	-.18134	.00169	-.00437	.00191
			L1	-.19446	.00010	.00032	.00011
			LTS	-.18800	.00164	.00205	.00058
			MM	-.19424	.00006	.00028	.00007
	10	50	OLS	-.19866	.01171	-.01778	.01596
			MG	-.18570	.02871	-.06054	.03894
			L1	-.18540	.00228	-.00367	.00309
			LTS	-.16230	.03224	-.00381	.02583
			MM	-.18483	.00187	-.00340	.00314
		800	OLS	-.19148	.00112	-.00318	.00155
			MG	-.17732	.00167	-.00538	.00205
			L1	-.19368	.00004	-.00017	.00006
			LTS	-.19485	.00022	-.00025	.00013
			MM	-.19312	.00002	-.00011	.00004
0.5	4	50	OLS	.34683	.04265	-.05107	.01870
			MG	.36554	.02180	-.05204	.04099
			L1	.42316	.01562	-.02221	.01304
			LTS	.35159	.08751	-.04056	.07510
			MM	.38470	.02550	-.02562	.01032
		800	OLS	.39890	.00067	-.00156	.00148
			MG	.39308	.00119	-.00587	.00252
			L1	.46748	.00475	-.00097	.00014
			LTS	.45444	.01428	-.00032	.00088
			MM	.48818	.00786	-.00121	.00010
	10	50	OLS	.37823	.00939	-.03809	.01739
			MG	.34596	.02506	-.06730	04132
			L1	.43980	.01020	-.00796	.00501
			LTS	.33623	.06072	-.00761	.01634
			MM	.36369	.03569	.00016	.00302
		800	OLS	.39977	.00083	-.00338	.00181
			MG	.39091	.00120	-.00774	.00246
			L1	.47008	.00501	.00072	.00007
			LTS	.49193	.01064	.00257	.00022
			MM	.48947	.00805	.00006	.00004
0.8	4	50	OLS	.46616	.01131	-.04233	.03611
			MG	.46974	.01749	-.05979	.03702
			L1	.55699	.03682	-.00934	.01905
			LTS	.43306	.10956	-.03247	.05134
			MM	.49038	.07561	-.01341	.01176
		800	OLS	.48720	.00054	-.00442	.00168
			MG	.49182	.00087	-.01179	.00261
			L1	.59438	.01447	-.00013	.00013
			LTS	.55985	.02922	.00078	.00066
			MM	.68805	.06670	-.00008	.00010
	10	50	OLS	.45878	.00836	-.04923	.01955
			MG	.48799	.01891	-.06083	.04586
			L1	.61845	.04446	-.00939	.00426
			LTS	.46685	.12663	-.01234	.01295
			MM	.50545	.11626	-.00938	.00443
		800	OLS	.48400	.00063	-.00768	.00170
			MG	.51178	.00147	-.00867	.00247
			L1	.63333	.02528	-.00110	.00005
			LTS	.71091	.08715	-.00049	.00014
			MM	.76902	.09312	-.00084	.00004

this pattern.

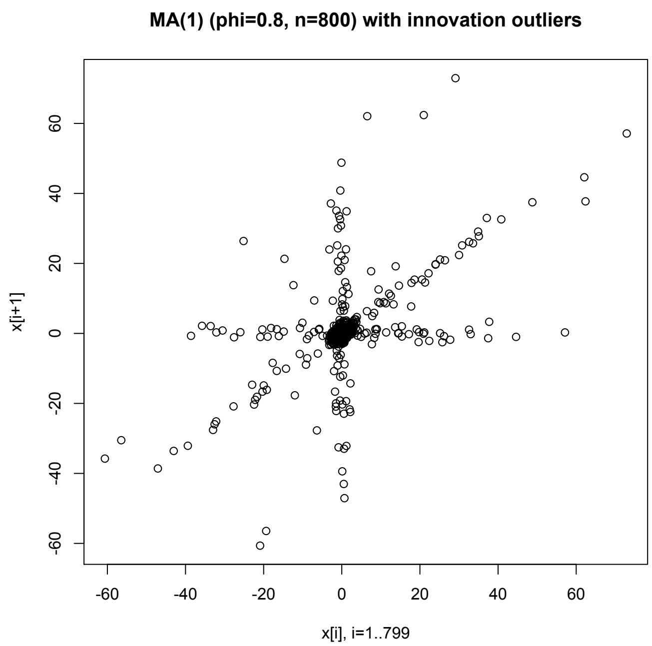

Figure 3.3: X_t vs. X_{t+1} plot for the MA(1) model $X_t = Z_t + 0.8Z_{t-1}$ with innovation outliers. With an innovation outlier at Z_t, (X_{t-1}, X_t) usually lies on the vertical line, (X_t, X_{t+1}) on the diagonal, and (X_{t+1}, X_{t+2}) on the horizontal. The robust estimators tend to fit the diagonal line.

From Table 3.4, we can see that the robust regression estimators all shine in the AR(1) innovation outlier case. This is unsurprising, since an AR(1) innovation outlier only pulls one point off the appropriate regression line, while generating several other high-magnitude points on it (see Figure 3.4). Note that the high-magnitude (and thus high leverage) points are in fact proportionally much closer to the regression line than the rest of the points; this accounts for the fast heavy tail sample acf convergence rate mentioned earlier, which can be seen in the table (the MSEs for $n = 800$ are especially small).

The Ma-Genton estimator does not appear to share the fast convergence rate.

Table 3.4: AR(1) simulation results with innovation outliers, averages of 200 trials.

ϕ	Contam. %	n	Estimator	Avg. $\hat{\rho}(1)$	MSE	Avg. $\hat{\rho}(2)$	MSE
-0.2	4	50	OLS	-.22576	.02227	.02086	.01577
			MG	-.20588	.03829	.00309	.03341
			L1	-.20750	.01018	.02877	.01126
			LTS	-.18120	.06381	.01468	.05601
			MM	-.20344	.00958	.03090	.00692
		800	OLS	-.19775	.00145	.03814	.00091
			MG	-.20515	.00160	.04088	.00161
			L1	-.19949	.00009	.03948	.00010
			LTS	-.19582	.00053	.03674	.00062
			MM	-.19983	.00005	.03962	.00007
	10	50	OLS	-.20993	.01148	.02600	.00906
			MG	-.22251	.03201	.03557	.03274
			L1	-.20522	.00131	.04086	.00157
			LTS	-.19185	.01821	.04927	.01699
			MM	-.20467	.00070	.04440	.00126
		800	OLS	-.19992	.00125	.04020	.00173
			MG	-.21774	.00209	.03777	.00178
			L1	-.20048	.00004	.03959	.00005
			LTS	-.19996	.00016	.03800	.00017
			MM	-.20039	.00002	.03967	.00003
0.5	4	50	OLS	.46520	.00956	.19043	.02019
			MG	.48690	.02355	.19364	.04024
			L1	.49198	.00532	.22763	.00791
			LTS	.48511	.02905	.19989	.04247
			MM	.49183	.00377	.23505	.00649
		800	OLS	.49840	.00097	.24964	.00152
			MG	.53888	.00282	.26705	.00255
			L1	.49969	.00006	.25023	.00009
			LTS	.50038	.00039	.25076	.00039
			MM	.49966	.00004	.24984	.00007
	10	50	OLS	.43619	.04085	.17919	.02531
			MG	.55736	.02541	.23512	.03739
			L1	.48964	.00309	.23227	.00662
			LTS	.48506	.00814	.24911	.01338
			MM	.49613	.00106	.24566	.00249
		800	OLS	.49832	.00086	.24440	.00151
			MG	.59379	.00993	.28994	.00367
			L1	.49924	.00003	.24811	.00007
			LTS	.49902	.00012	.24713	.00018
			MM	.49941	.00002	.24885	.00004
0.8	4	50	OLS	.74099	.01184	.53776	.03219
			MG	.81219	.01316	.61055	.03626
			L1	.77752	.00572	.59431	.01536
			LTS	.76933	.02006	.59469	.03543
			MM	.77987	.00425	.59493	.01619
		800	OLS	.79691	.00037	.63449	.00104
			MG	.88504	.00760	.74106	.01129
			L1	.80011	.00003	.63971	.00008
			LTS	.80013	.00016	.64059	.00040
			MM	.80001	.00002	.64043	.00006
	10	50	OLS	.72992	.01731	.53105	.03459
			MG	.89090	.01489	.70164	.02350
			L1	.79232	.00165	.61721	.00596
			LTS	.79450	.00806	.61635	.01611
			MM	.79677	.00068	.62704	.00465
		800	OLS	.79714	.00046	.63659	.00137
			MG	.93719	.01892	.80715	.02847
			L1	.79990	.00001	.63971	.00004
			LTS	.79943	.00008	.64034	.00012
			MM	.80009	.00001	.64031	.00003

Table 3.5: AR(2) simulation results with innovation outliers, averages of 200 trials. True $(\rho(1), \rho(2))$ is $(\frac{5}{9}, \frac{17}{45})$ in the $(\phi_1, \phi_2) = (0.5, 0.1)$ case, and $(\frac{6}{7}, \frac{57}{70})$ in the $(\phi_1, \phi_2) = (0.6, 0.3)$ case.

ϕ_1, ϕ_2	Contam. %	n	Estimator	Avg. $\hat{\rho}(1)$	MSE	Avg. $\hat{\rho}(2)$	MSE
0.5, 0.1	4	50	OLS	.49805	.01063	.30211	.05545
			MG	.58988	.02912	.36280	.03815
			L1	.53168	.00797	.34237	.01120
			LTS	.54746	.03708	.37782	.04575
			MM	.54553	.00746	.35229	.00849
		800	OLS	.55232	.00110	.37241	.00147
			MG	.64522	.00940	.43532	.00545
			L1	.55813	.00034	.37576	.00018
			LTS	.61018	.00749	.38777	.00166
			MM	.57101	.00151	.37671	.00010
	10	50	OLS	.50255	.01400	.30139	.04159
			MG	.70744	.04561	.44201	.03910
			L1	.53896	.00265	.35993	.00599
			LTS	.59708	.01678	.37594	.01517
			MM	.56102	.00395	.36632	.00395
		800	OLS	.55514	.00106	.37448	.00143
			MG	.74715	.03765	.50552	.01800
			L1	.55795	.00021	.37708	.00011
			LTS	.63746	.00934	.38582	.00082
			MM	.55693	.00030	.37730	.00005
0.6, 0.3	4	50	OLS	.73869	.02945	.66437	.04485
			MG	.85771	.01835	.79481	.03038
			L1	.83786	.01150	.75807	.02043
			LTS	.85036	.02913	.77909	.03403
			MM	.85324	.01226	.77049	.01785
		800	OLS	.85081	.00060	.80749	.00101
			MG	.96729	.01227	.94209	.01663
			L1	.90582	.00242	.83797	.00066
			LTS	.91584	.00372	.84984	.00160
			MM	.91882	.00383	.84796	.00121
	10	50	OLS	.70326	.04351	.62912	.05943
			MG	.91934	.01062	.86392	.01527
			L1	.84315	.00793	.76536	.01284
			LTS	.88486	.01410	.81898	.01243
			MM	.88306	.00714	.78694	.00987
		800	OLS	.84891	.00076	.80323	.00119
			MG	.98441	.01621	.96065	.02146
			L1	.90649	.00248	.83758	.00062
			LTS	.92019	.00409	.85299	.00166
			MM	.92185	.00421	.84226	.00084

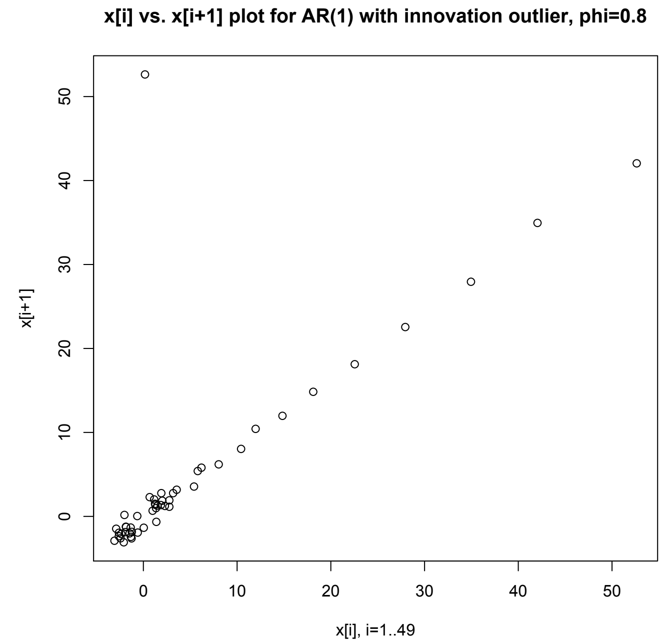

Figure 3.4: (x_t, x_{t+1}) plot for a realization of the AR(1) time series $X_t = 0.8X_{t-1} + Z_t$ with one innovation outlier.

Moving on to the AR(2) case (Table 3.5), we see that with innovation outliers, the L1 and MM robust estimators exhibit much better performance than OLS given a small (50) sample size, but the difference fades with a larger sample size. The Ma-Genton estimator performs relatively poorly across the board.

3.7.3 Additive Outliers

Next, we investigated the performance of our estimators in the additive outlier case by perturbing one or two elements in the middle of the time series by a large number (where, as before, innovations are i.i.d. N(0, 1)).

The Ma-Genton and MM estimators do the best (Table 3.6). The OLS estimator performed especially badly in the $\phi = 0.8$ case, L1 was fairly good but failed the

Table 3.6: AR(1) simulation results with additive outliers, averages of 200 trials. In a length-n time series, an "a, b" contamination pattern means that a was added to the $\frac{n}{2}$th element and b was added to the $(\frac{n}{2}+1)$th element.

ϕ	n	Contam. Pattern	Estimator	Avg. $\hat{\rho}(1)$	MSE	Avg. $\hat{\rho}(2)$	MSE
-0.2	50	25, 25	OLS	.45142	.42544	-.04117	.00796
-0.2	50	25, 25	MG	-.15163	.03272	.01986	.03361
-0.2	50	25, 25	L1	.00949	.04707	-.00093	.00337
-0.2	50	25, 25	LTS	-.16946	.09823	.02005	.06239
-0.2	50	25, 25	MM	-.11206	.03400	.01052	.00573
-0.2	50	25, 0	OLS	-.03535	.02868	-.02372	.00719
-0.2	50	25, 0	MG	-.23079	.03396	.03604	.03247
-0.2	50	25, 0	L1	-.00350	.03885	-.00308	.00348
-0.2	50	25, 0	LTS	-.22194	.11734	.04899	.09225
-0.2	50	25, 0	MM	-.12271	.03509	.01925	.01019
-0.2	50	25, -25	OLS	-.48932	.08450	.00144	.00328
-0.2	50	25, -25	MG	-.27360	.03155	.01328	.03401
-0.2	50	25, -25	L1	-.22172	.02783	.00269	.00268
-0.2	50	25, -25	LTS	-.18993	.09892	.00917	.06258
-0.2	50	25, -25	MM	-.11939	.04269	.00924	.00569
-0.2	800	25, 25	OLS	.21900	.17617	.01230	.00140
-0.2	800	25, 25	MG	-.20102	.00130	.03990	.00146
-0.2	800	25, 25	L1	-.12892	.00658	.01522	.00135
-0.2	800	25, 25	LTS	-.21102	.01470	.04796	.01365
-0.2	800	25, 25	MM	-.18993	.00220	.01481	.00112
-0.2	800	25, 0	OLS	-.11684	.00753	.02222	.00138
-0.2	800	25, 0	MG	-.20590	.00128	.03766	.00138
-0.2	800	25, 0	L1	-.16603	.00271	.01995	.00165
-0.2	800	25, 0	LTS	-.20086	.01327	.02622	.01303
-0.2	800	25, 0	MM	-.19773	.00208	.02053	.00118
-0.2	800	25, -25	OLS	-.38096	.03324	.01549	.00167
-0.2	800	25, -25	MG	-.20358	.00136	.03339	.00181
-0.2	800	25, -25	L1	-.19913	.00164	.01230	.00178
-0.2	800	25, -25	LTS	-.19461	.01404	.02058	.01134
-0.2	800	25, -25	MM	-.18644	.00255	.01274	.00143
0.8	50	25, 25	OLS	.49677	.09497	-.00211	.42308
0.8	50	25, 25	MG	.73375	.01678	.48080	.05831
0.8	50	25, 25	L1	.71784	.02180	.02922	.38034
0.8	50	25, 25	LTS	.76085	.05544	.41964	.17343
0.8	50	25, 25	MM	.80826	.03660	.40465	.13308
0.8	50	25, 0	OLS	.08162	.52154	.04809	.35910
0.8	50	25, 0	MG	.69233	.02827	.46751	.06854
0.8	50	25, 0	L1	.34407	.25730	.13002	.29226
0.8	50	25, 0	LTS	.71216	.07760	.44343	.15611
0.8	50	25, 0	MM	.70241	.02940	.42005	.10891
0.8	50	25, -25	OLS	-.40196	1.44785	.03820	.36352
0.8	50	25, -25	MG	.69580	.02682	.50972	.04958
0.8	50	25, -25	L1	.07515	.55373	.05980	.34270
0.8	50	25, -25	LTS	.73795	.06227	.45480	.13595
0.8	50	25, -25	MM	.73087	.01986	.44154	.11602
0.8	800	25, 25	OLS	.68855	.01329	.40271	.05906
0.8	800	25, 25	MG	.79444	.00052	.62901	.00148
0.8	800	25, 25	L1	.79499	.00071	.59307	.00407
0.8	800	25, 25	LTS	.79494	.00556	.62680	.00936
0.8	800	25, 25	MM	.79541	.00047	.63165	.00126
0.8	800	25, 0	OLS	.61925	.03378	.49346	.02354
0.8	800	25, 0	MG	.79651	.00048	.63460	.00134
0.8	800	25, 0	L1	.78514	.00097	.61991	.00197
0.8	800	25, 0	LTS	.80580	.00493	.63532	.00849
0.8	800	25, 0	MM	.79848	.00046	.63659	.00125
0.8	800	25, -25	OLS	.31915	.23308	.39943	.05927
0.8	800	25, -25	MG	.79172	.00065	.62866	.00160
0.8	800	25, -25	L1	.76533	.00208	.59569	.00394
0.8	800	25, -25	LTS	.78979	.00514	.63881	.00996
0.8	800	25, -25	MM	.79527	.00050	.63117	.00136

$\phi = 0.8$, $n = 50$ case, and LTS generally acted as a much less efficient MM.

3.7.4 Austrian Bank Data

We then applied our estimators to some real-world data, monthly interest rates of an Austrian bank over a 91 month period (see Figure 3.5). This data set has previously been analyzed by Künsch (1983) (1984) and by Ma and Genton (2000).

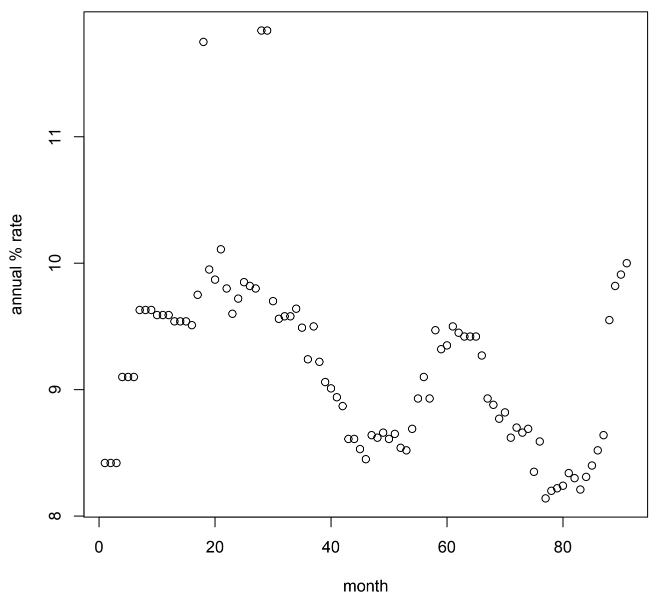

Figure 3.5: 91 consecutive monthly interest rates of an Austrian bank.

Note the three outliers at months 18, 28, and 29. Following Künsch, we run our estimators on both the original data set, and a slightly revised data set where the three outliers are replaced with 9.85.

The L1 and Ma-Genton estimators both gave reasonable numbers and were less affected by the outliers than OLS. However, the LTS estimator was erratic, overestimating the low lag autocorrelation, exhibiting a discontinuity at $\hat{\rho}(6)$ when outliers

Table 3.7: Simulation results with Austrian bank data. ($\hat{\rho}(2)$ was omitted since it was always close to $\frac{\hat{\rho}(1)+\hat{\rho}(3)}{2}$.)

Estimator	Outliers replaced?	$\hat{\rho}(1)$	$\hat{\rho}(3)$	$\hat{\rho}(4)$	$\hat{\rho}(5)$	$\hat{\rho}(6)$	$\hat{\rho}(12)$
OLS	no	.79184	.58923	.51249	.44414	.40440	.08583
	yes	.93920	.77965	.67369	.58264	.50113	.07476
MG	no	.96571	.82703	.73727	.65968	.55046	-.18033
	yes	.96923	.82350	.77709	.69337	.60000	-.15294
L1	no	.97222	.83459	.78351	.72603	.65957	-.02786
	yes	.98361	.89655	.83505	.78169	.76991	-.03361
LTS	no	.99451	.95588	.87975	.85556	.36749	-.94203
	yes	1.00000	.96667	.87603	.86441	.81633	-.94203
MM	no	.97194	.81113	.49292	.40119	.34198	.04550
	yes	.96779	.86493	.79272	.69961	.59654	.07344

Table 3.8: AR(2) simulation results with innovation outliers (10 percent frequency, SD 25x normal), averages of 50 (with $n = 800$) or 200 (with $n = 50$) trials.

ϕ_1, ϕ_2	n	Estimator	Avg. $\hat{\phi}(1)$	MSE	Avg. $\hat{\phi}(2)$	MSE
0.5, 0.1	50	OLS	.56535	.03910	-.01931	.03450
		MG	.78635	.16480	-.15357	.11939
		L1	.52698	.01167	.06052	.01173
		LTS	.53997	.03552	.04648	.02372
		MM	.56058	.01610	.03653	.01423
	800	OLS	.61277	.01600	.03479	.00733
		MG	1.04081	.29954	-.28389	.15202
		L1	.51674	.00072	.08902	.00042
		LTS	.64642	.02789	-.01545	.01772
		MM	.52117	.00120	.08589	.00072

were present, and yielding a bizarre value of -.94203 for the 12-month autocorrelation. MM yielded fine results up to lag 3, but the lag 4-6 numbers were heavily affected by the outliers.

3.7.5 AR Model Fitting

Finally, we combined the direct AR model fitting method described in section 3.6 with our robust autocorrelation estimators.

As we can see in Table 3.8, the robust AR model fitter yields reasonable results even when given the raw sample acf. However, performance was noticeably better with $n = 50$ when combining it with the L1 or MM robust autocorrelation estimators, and with $n = 800$ instead, the performance advantage was overwhelming. Thus, these two methods are not redundant; they complement each other very well.

The Ma-Genton estimator did not estimate the autocorrelations well in Table 3.5, so it is not surprising that the inferred AR coefficients are also far off.

3.8 Conclusions

A procedure for constructing robust autocorrelation estimators out of robust linear regression techniques was proposed, and applied to L1, LTS, and MM regression. A simulation study was then performed, comparing these estimators to the sample acf and a scale-based robust estimator proposed by Ma and Genton. It was found that the Ma-Genton estimator was superior at handling MA(1) models, while our L1- and MM-based estimators shined in the AR case (where Ma-Genton performed poorly). The L1 and MM estimators worked especially well with Politis' suggested procedure for robustly estimating AR coefficients.

3.8.1 Acknowledgements

Thanks to Tim McMurry for helpful R tips, and to Marc Genton for the Austrian bank data.

Chapter 3 is essentially a reprint, with minor modifications, of the paper "Robust Autocorrelation Estimation" by C. Chang and D.N. Politis, which is now in preparation for publication. The dissertation author was the primary investigator and author of this paper.

Bibliography

[1] 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.

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

[3] D. Brillinger, 1981. Time Series: Data Analysis and Theory, Holden Day, San Francisco.

[4] P.J. Brockwell, R.A. Davis, 1991. Time Series: Theory and Methods, Second Edition, Springer, New York.

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

[6] R.A. Davis and T. Mikosch, 2000. The sample autocorrelations of financial time series models, in *Nonlinear and Nonstationary Signal Processing*, Cambridge University Press, Cambridge, pp. 247–274.

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

[8] L. Denby and R.D. Martin, 1979. Robust estimation of the first-order autoregressive parameter, *Journal of the American Statistical Association* **74**, pp. 140–146.

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

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

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

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

[13] W.A. Fuller, 1987. Measurement error models, John Wiley & Sons, New York.

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

[15] F.R. Hampel, 1973. Robust estimation, a condensed partial survey, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **27**, pp. 87–104.

[16] P.J. Huber, 1973. Robust regression: Asymptotics, conjectures and Monte Carlo, *The Annals of Statistics* **1**, pp. 799–821.

[17] M. Jones, M. Wand, 1995. Kernel Smoothing, CRC Press, London.

[18] H. Künsch, 1983. Robust estimation for autoregressive processes, *Proc. Institute for Statist. Math* **31**, pp. 51–64.

[19] H. Künsch, 1984. Infinitesimal robustness for autoregressive processes. *The Annals of Statistics* **12**, pp. 843–863.

[20] Y. Ma and M.G. Genton, 2000. Highly robust estimation of the autocovariance function, *Journal of Time Series Analysis* **21**, pp. 663–684.

[21] R.A. Maronna, R.D. Martin, and V.J. Yohai, 2006. Robust Statistics: Theory and Methods, Wiley, New York.

[22] T. McMurry and D.N. Politis, 2010. Banded and tapered estimates of autocovariance matrices and the linear process bootstrap, submitted to *Journal of Time Series Analysis.*

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

[24] E. Parzen, 1957. On consistent estimates of the spectrum of a stationary time series, *Annals of Mathematical Statistics* **28**, pp. 329–348.

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

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

[27] D.N. Politis, J.P. Romano, 1995. Bias-Corrected Nonparametric Spectral Estimation, *J. Time Ser. Anal.*, **16**, pp. 67–104.

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

[29] D.N. Politis, 2001. On nonparametric function estimation with infinite-order flat-top kernels, in *Probability and Statistical Models with applications*, Ch. Charalambides et al. (Eds.), Chapman and Hall/CRC, Boca Raton, pp. 469–483.

[30] D.N. Politis, 2003. Adaptive bandwidth choice, *J. Nonparam. Statist.* **15**, pp. 517–533.

[31] D.N. Politis, 2005. Complex-valued tapers, *IEEE Signal Processing Letters* **12**, pp. 512–515.

[32] D.N. Politis, 2009. An algorithm for robust fitting of autoregressive models, *Economics Letters* **102**, pp. 128–131.

[33] D.N. Politis, 2011. Higher-order accurate, positive semi-definite estimation of large-sample covariance and spectral density matrices, to appear in *Econometric Theory*.

[34] M. Priestley, 1981. Spectral Analysis and Time Series, Academic Press, London.

[35] P. Rigollet, A. Tsybakov, 2007. Linear and convex aggregation of density estimators, *Mathematical Methods of Statistics* **16**, pp. 260–280.

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

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

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

[39] 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.

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

[41] 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.

[42] 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.

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

[44] P.J. Rousseeuw and A.M. Leroy. Robust Regression and Outlier Detection, John Wiley & Sons, New York.

[45] V.J. Yohai, 1987. High breakdown-point and high efficiency estimates for regression, *The Annals of Statistics* **15**, pp. 642–656.

[46] R.H. Zamar, 1989. Robust estimation in the errors-in-variables model, *Biometrika* **76**, pp. 149–160.