

Semiparametric Estimation for Non-Gaussian Non-minimum Phase ARMA models

Richard A. Davis and Jing Zhang
Columbia University

Abstract

We consider inference for the parameters of general autoregressive moving average (ARMA) models which are possibly noncausal/noninvertible (also referred to as non-minimum phase) and driven by a non-Gaussian distribution. For non-minimum phase models, the observations can depend on both the past and future shocks in the system. The non-Gaussianity constraint is necessary to distinguish between causal-invertible and non-causal/noninvertible models. Many of the existing estimation procedures adopt quasi-likelihood methods by assuming a non-Gaussian density function for the noise distribution that is fully known up to a scalar parameter. To relax such distributional restrictions, we borrow ideas from nonparametric density estimation and propose a semiparametric maximum likelihood estimation procedure, in which the noise distribution is projected onto the space of log-concave measures. We show the maximum likelihood estimators in this semiparametric setting are consistent. In fact, the MLE is robust to the misspecification of log-concavity in cases where the true distribution of the noise is close to its log-concave projection (Cule and Samworth, 2010; Dümbgen et al., 2011). We derive a lower bound for the best asymptotic variance of regular estimators at rate $n^{-\frac{1}{2}}$ for AR models and construct a semiparametric efficient one-step estimator. The estimation procedure is illustrated via a simulation study and an empirical example illustrating the methodology is also provided.

Keywords: semiparametric; ARMA; non-causal/noninvertible; log-concave

1 Introduction

The ARMA model is perhaps the most successful and well studied class of models for the analysis of univariate time series. In the cases where the noise distribution is Gaussian, they are easy to fit by maximizing the resulting Gaussian likelihood. One often resorts to maximizing the Gaussian likelihood even if the noise is non-Gaussian. The parameter estimated in this fashion have the same asymptotic behavior as in the “Gaussian” case. A univariate time series X_t is called an $\text{ARMA}(p, q)$ process if it is stationary and satisfies the difference equations

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \quad (1)$$

where $\{Z_t\}$ is a sequence of independent and identically distributed (i.i.d) random variables with zero mean. Define the AR and MA polynomials by $\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p$ and $\theta(z) = 1 + \theta_1 z + \dots + \theta_q z^q$, respectively. Then (1) can be written in the compact form

$$\phi(B)X_t = \theta(B)Z_t,$$

where B is the backward-shift operator. We assume $\phi(z)$ and $\theta(z)$ have no common roots and satisfy the constraint that $\phi(z)\theta(z) \neq 0$ for any $|z| = 1$. Thus there exists a unique strictly stationary solution $\{X_t\}$ to (1). An ARMA (p, q) process X_t is said to be causal-invertible if

$$\phi(z)\theta(z) \neq 0 \text{ for all } z \in \mathbb{C} \text{ with } |z| \leq 1,$$

that is, both the AR and MA polynomials have no roots inside the unit circle, for which X_t (Z_t) can be expressed as a function of only the present and past noise $\{Z_s : s \leq t\}$ (observations $\{X_s, s \leq t\}$). It turns out that any noncausal/noninvertible ARMA process admits an equivalent causal-invertible representation for which

$$\phi^*(B)X_t = \theta^*(B)Z_t^*.$$

However, the sequence $\{Z_t^*\}$ is i.i.d if and only if Z_t is Gaussian, otherwise, $\{Z_t^*\}$ is only uncorrelated, see Breidt and Davis (1991). Thus in the Gaussian case, one requires causality and invertibility to ensure identifiability of the model parameters. In the non-Gaussian case, ARMA models are allowed to be noncausal/noninvertible. As a result, the Gaussian noise must be excluded in order to study noncausal/noninvertible models. On the other hand, it is common to observe non-Gaussian sequences in the real world. The non-minimum phase ARMA models are useful in a variety of applications. The Wal-Mart stock volume data in (Andrews et al., 2009), the U.S. inflation data in (Lanne and Saikkonen, 2008) and the Microsoft stock volume data in (Breidt et al., 2001) are all such examples. Allowing noncausality/noninvertibility can enlarge the pool of ARMA models, eliminate more of the serial dependence and enhance our understanding of the data.

Any second-order based estimation procedure, including maximizing the Gaussian likelihood, are not efficient unless the noise distribution is Gaussian. Moreover, since second order estimation procedures cannot distinguish between causal-invertible and noncausal/noninvertible models, they are not applicable for estimating non-minimum phase models. In fact, statistical inference for non-minimum phase models are comparatively limited due to the complicated dependence structure of the process itself (Breidt et al., 2001; Lii and Rosenblatt, 1996; Wu and Davis, 2010). Approximations of the likelihood functions for the non-causal/noninvertible ARMA process are derived in (Breidt et al., 2001; Lii and Rosenblatt, 1996). Many of the existing estimation procedures are then based on the idea of maximum likelihood estimation by assuming a common pre-specified noise distribution. In (Breidt et al., 2001; Wu and Davis, 2010), a least absolute deviation (LAD) criterion is proposed, which is frequently used for modeling time series in the non-Gaussian setting. While the LAD method is derived by using the Laplace distribution, it can produce consistent estimators even when the noise distribution is not Laplace under mild conditions. To relax the strong parametric

assumption about the innovation distribution, we have to extend the maximum likelihood principle to a nonparametric framework and consider semiparametric models. (Kreiss, 1987) proposes one-step adaptive estimators for ARMA models where the error density is estimated by a kernel density estimator. (Chen and Samworth, 2015) studies semiparametric time series models including causal-invertible ARMA processes, in which the distribution of the noise satisfies minor conditions and it has been shown that the semiparametric estimation procedure produces consistent estimators of the ARMA parameters. In addition, the estimate of the noise distribution consistently estimates the log-concave projection of the true density. In particular, if the noise density is log concave, then the density estimator is consistent. Inspired by (Chen and Samworth, 2015), we apply the log-concave projection method to the non-causal/noninvertible ARMA models. We show the consistency of the estimators for both the coefficients and the density under mild conditions. We also obtain a lower bound for the asymptotic variances of regular estimators at rate $n^{-\frac{1}{2}}$ for the semiparametric AR models.

The rest of the paper is organized as follows. Section 2 provides a quick review of the definitions and basic properties of log-concave densities and log-concave projection. Section 3 applies log-concave projection to general ARMA models and derives the objective function. Section 4 shows the consistency of the estimators and derives a lower bound for the asymptotic variances of regular estimators at rate $n^{-\frac{1}{2}}$ for general AR models. Section 5 presents a simulation study and a real data application to further illustrate the results in Section 4. Technical details are given in the Appendix.

2 The Log-Concave Projection

A probability density function f is said to be log-concave if $\log f$ is a concave function. The family of log-concave densities has some attractive properties and behaves to some extent as a parametric family, see (Bagnoli and Bergstrom, 2006; Walther, 2009). It has been shown that for a given probability measure P on \mathbb{R} , there exists a unique log-concave density f that maximizes the log-likelihood type functional

$$\int_{\mathbb{R}} \log f dP,$$

when the maximum is with respect to log-concave densities under mild conditions (Cule and Samworth, 2010; Dümbgen et al., 2011). The log-concave maximum likelihood estimator of P based on iid observations from P can be viewed as a projection of the empirical measure onto the space of distributions with log-concave densities. This estimation procedure possesses good properties and sheds light on the area of nonparametric density estimation. To apply this nonparametric estimation procedure to ARMA models, it is helpful to review the properties of such projections first (see Cule and Samworth (2010); Dümbgen et al. (2011); Walther (2009) for more details).

Let \mathcal{P} denote the class of all non-degenerate probability measures P on \mathbb{R} with finite first moment. Let \mathcal{F} be the set of log concave densities on \mathbb{R} . Then the functional mapping $\Pi : \mathcal{P} \rightarrow \mathcal{F}$

$$\Pi(P) = \arg \max_{f \in \mathcal{F}} \int_{\mathbb{R}} \log f dP$$

is well-defined if and only if $P \in \mathcal{P}$. The quantity $\Pi(P)$ is referred to as the log-concave projection of P onto \mathcal{F} . The maximal function $L : \mathcal{P} \rightarrow \mathbb{R}$ is defined as

$$L(P) = \max_{f \in \mathcal{F}} \int_{\mathbb{R}} \log f dP,$$

and is finite if and only if $P \in \mathcal{P}$ (If the first moment of P does not exist, $L(P) = -\infty$, while if P is a dirac measure, $L(P) = \infty$). For convenience, we also use $L(X)$ and $\Pi(X)$ to denote $L(P)$ and $\Pi(P)$ respectively when X is some random variable distributed as P . The key properties of $L(\cdot)$ and $\Pi(\cdot)$ are summarized below.

1. Affine equivariance:

$$L(a + CX) = L(X) - \log|C| \text{ for any } a \in \mathbb{R} \text{ and nonzero constant } C$$

2. Non-increasing under convolution:

$$L(X + Y) \leq L(X) \tag{2}$$

if X is independent of Y and $X \in \mathcal{P}$. The equal sign holds if and only if Y is a constant.

3. Mean preservation:

$$\int_{\mathbb{R}} x dP = \int_{\mathbb{R}} x \Pi(P)(x) dx.$$

Further interesting properties of $\Pi(\cdot)$ and $L(\cdot)$ have been well presented by (Dümbgen et al., 2011). Here we state the main results in (Dümbgen et al., 2011) for completeness.

First we introduce two useful measures of distance between probability distributions: the first moment Mallows distance (Mallows, 1972) and the bounded Lipschitz metric. Suppose P and Q are any two probability measures in \mathcal{P} , the first moment Mallows distance between P and Q is defined by

$$M_1(P, Q) := \inf_F (\mathbb{E}|X - Y| : (X, Y) \sim F, X \sim P, Y \sim Q),$$

where X and Y are any integrable random variables distributed as P and Q respectively, and F is the joint probability distribution of (X, Y) satisfying the marginal distribution constraint. $M_1(\cdot, \cdot)$ is also known as the Wasserstein, Monge-Kantorovich or Earth Mover's distance (Levina and Bickel, 2001). Kantorovich and Rubinstein (1958) established a useful duality formula for Mallows's distance:

$$M_1(P, Q) := \sup_{\|g\|_L \leq 1} \left| \int g d(P - Q) \right| \tag{3}$$

with $\|g\|_L = \sup_{x \neq y} |g(x) - g(y)| / |x - y|$, and the supremum is over all Lipschitz functions with Lipschitz constant bounded by one. It's also known that for any sequence of probability measures Q_n and Q ,

$$M_1(Q_n, Q) \longrightarrow 0 \text{ if and only if } Q_n \xrightarrow{w} Q \text{ and } \int |x| dQ_n \longrightarrow \int |x| dQ. \tag{4}$$

More detailed information about the first moment Mallows distance can be found in Villani (2009).

The bounded Lipschitz distance metrizes the weak convergence of probability measures

$$D_{BL}(P, Q) := \sup_{\|g\|_\infty \leq 1, \|g\|_L \leq 1} \left| \int g d(P - Q) \right|$$

with $\|g\|_\infty := \sup_x |g(x)|$. It's obvious that the first moment Mallows distance is stronger than the bounded Lipschitz metric:

$$D_{BL}(P, Q) \leq M_1(P, Q).$$

The following continuity properties of $L(\cdot)$ and $\Pi(\cdot)$ with respect to $M_1(\cdot, \cdot)$ and $D_{BL}(\cdot, \cdot)$ are adapted from Theorem 2.15 in (Dümbgen et al., 2011).

Lemma 2.1. *Let the sequence $\{P_n\}$ and P be distributions on \mathbb{R} with finite first moment. Then,*

- (i) *If $\lim_{n \rightarrow \infty} D_{BL}(P_n, P) = 0$, then $\limsup_{n \rightarrow \infty} L(P_n) \leq L(P)$*
- (ii) *If $\lim_{n \rightarrow \infty} M_1(P_n, P) = 0$, then $\lim_{n \rightarrow \infty} L(P_n) = L(P)$.*
- (ii) *If $\lim_{n \rightarrow \infty} M_1(P_n, P) = 0$, then $\Pi(P_n)$ converges to $\Pi(P)$ in L^1 .*

Remark 2.2. For our results, Lemma 2.1 will be applied by taking P_n to be the empirical distribution of observations coming from a stationary ergodic time series. Suppose $\{X_t\}$ is a stationary ergodic time series with marginal distribution P in \mathcal{P} . Then it follows from (4) that the empirical distribution $\mathbb{P}_n := \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ converges to P in the first moment Mallows distance almost surely. As a result, the log-concave maximum likelihood estimator \hat{f}_n

$$\hat{f}_n := \Pi(\mathbb{P}_n) = \arg \max_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \log f(X_i)$$

is well defined for large n with probability one and

$$L(\mathbb{P}_n) \xrightarrow{a.s.} L(P), \quad \int |\hat{f}_n - \Pi(P)| \xrightarrow{a.s.} 0.$$

Lemma 2.3 summarizes some convergence results of the log-concave density sequences shown in (Cule and Samworth, 2010), which play an important role in the application of the log-concave density estimator to ARMA processes.

Lemma 2.3. *Let f_n be a sequence of log-concave densities on \mathbb{R} and f be some density function on \mathbb{R} such that $F_n \xrightarrow{d} F$ where (F_n, F) are the associated cdfs of (f_n, f) . Then,*

- (i) *f is log-concave*
- (ii) *f_n converges to f almost everywhere*
- (iii) *Let $a_0 > 0$ and $b_0 \in \mathbb{R}$ such that $f(x) \leq e^{(-a_0|x|+b_0)}$. Then for every $a < a_0$, we have $\int_{\mathbb{R}} e^{a|x|} |f_n(x) - f(x)| dx \rightarrow 0$. Furthermore, if f is continuous,*

$$\sup_{x \in \mathbb{R}} e^{a|x|} |f_n(x) - f(x)| \rightarrow 0.$$

Lemma 2.3 further implies that \hat{f} converges to the log-concave projection $\Pi(P)$ in a stronger exponential weighting norm. Moreover, the log-concave maximum likelihood density estimator $\hat{\varphi} := \log \hat{f}$ is shown to be a piecewise linear function with knots at the observations $\{X_i\}_{i=1}^n$ and is zero outside the interval $\left[\min_{i=1, \dots, n} X_i, \max_{i=1, \dots, n} X_i \right]$. It is not differentiable at the sample points $\{X_i\}_{i=1}^n$. As a substitute for \hat{f} , a smoothed log concave density estimator \hat{f}_σ , the convolution of \hat{f} with a zero mean, σ^2 variance normal density, is proposed in (Chen and Samworth, 2013). Detailed construction of \hat{f}_σ can be found in (Dümbgen and Rufibach, 2010).

3 Model specification

Denote ϕ and θ as the AR and MA parameter vectors $(\phi_1, \dots, \phi_p) \in \mathbb{R}^p$ and $(\theta_1, \dots, \theta_q) \in \mathbb{R}^q$ respectively. Let the parameter space $\Theta := \{\beta = (\phi, \theta)^T\}$ be a compact subset of \mathbb{R}^{p+q} such that the AR and MA polynomials $\phi(z)$ and $\theta(z)$ have no common zeros and no zeros on the unit circle. Let $\beta_0 = (\phi_0, \theta_0)^T$ denote the true parameter vector and P_0 denote the true distribution of Z_t . Since the polynomials $\phi(z)$ and $\theta(z)$ have no zeros of absolute value one, then $\beta(z) := \theta^{-1}(z)\phi(z)$ admits a two sided power expansion

$$\beta(z) = \sum_{i=-\infty}^{\infty} a_i(\beta) z^i$$

in some annulus $\{z : 0 < r(\beta) < |z| < R(\beta)\}$ where $r(\beta) < 1, R(\beta) > 1$ (Brockwell and Davis, 2009). The coefficients $a_i(\beta)$ decay geometrically fast to zero as $|i| \rightarrow \infty$. Although Z_t is unobserved, it's expressible in terms of β_0 and $\{X_t\}$. Rearranging (1), we obtain the linear representation of Z_t in terms of $\{X_t\}$:

$$Z_t(\beta_0) = \beta_0(B)X_t = \sum_{i=-\infty}^{\infty} a_i(\beta_0)X_{t-i} = Z_t.$$

Analogously, for any $\beta \in \Theta$, define the process

$$Z_t(\beta) := \beta(B)X_t = \theta^{-1}(B)\phi(B)X_t = \sum_{i=-\infty}^{\infty} a_i(\beta)X_{t-i}.$$

It's easy to see that $Z_t(\beta)$ is stationary and ergodic. We define a convergent representation of $Z_t(\beta)$ as introduced in Lii and Rosenblatt (1996):

$$Z_{t,m(n)}(\beta) = \sum_{|i| \leq m(n)} a_i(\beta)X_t,$$

where $m(n) \rightarrow \infty$ as $n \rightarrow \infty$ with $m(n) = o(n)$. By such truncation, $Z_{t,m(n)}(\beta)$ is completely computable from the observed sequence $\{X_1, \dots, X_n\}$ for $t = m(n) + 1, \dots, n - m(n)$. Let

$$\mathbb{P}_{\beta,n} := \frac{1}{n - 2m(n)} \sum_{t=m(n)+1}^{n-m(n)} \delta_{Z_{t,m(n)}(\beta)}$$

and

$$\tilde{\mathbb{P}}_{\beta,n} := \frac{1}{n - 2m(n)} \sum_{t=m(n)+1}^{n-m(n)} \delta_{Z_t(\beta)}$$

be the empirical measures of the truncated residuals $\{Z_{t,m(n)}(\beta)\}_{t=m(n)+1}^{n-m(n)}$ and the untruncated residuals $\{Z_t(\beta)\}_{t=m(n)+1}^{n-m(n)}$, respectively. And let P_β denote the stationary distribution of $Z_t(\beta)$. We have the following convergence results for $\mathbb{P}_{\beta,n}$ and $\tilde{\mathbb{P}}_{\beta,n}$.

Lemma 3.1. *Suppose that β_0 is an interior point in the compact parameter space Θ and $P_0 \in \mathcal{P}$. Then,*

$$\sup_{\beta \in \Theta} M_1(\mathbb{P}_{\beta,n}, \tilde{\mathbb{P}}_{\beta,n}) \xrightarrow{a.s.} 0 \quad \text{and} \quad \sup_{\beta \in \Theta} M_1(\tilde{\mathbb{P}}_{\beta,n}, P_\beta) \xrightarrow{a.s.} 0 \quad \text{as } n \rightarrow \infty.$$

It follows that

$$\sup_{\beta \in \Theta} M_1(\mathbb{P}_{\beta,n}, P_\beta) \xrightarrow{a.s.} 0 \quad \text{as } n \rightarrow \infty. \quad (5)$$

Lemma 3.1 indicates that the truncated residuals are asymptotically equivalent to the untruncated version in the first moment Mallow's distance. In (Lii and Rosenblatt, 1996), the following approximations to the log-likelihood function of the sequence $\{X_t\}_{t=1}^n$ are derived:

$$h_{\beta,f}^n := \frac{1}{n - 2m(n)} \sum_{i=m(n)+1}^{n-m(n)} l_{\beta,f}(\mathbf{z}_{i,m(n)}(\beta)) = \int l_{\beta,f} d\mathbb{P}_{\beta,n}, \quad (6)$$

where

$$l_{\beta,f}(u) := \log f(u) + \log \kappa(\beta),$$

and f is the assumed pdf of Z_t . The deterministic piece $\kappa(\beta)$ is the Jacobian of the transformation introduced in deriving $h_{\beta,f}^n$, which equals the reciprocal of the products of $\theta(z)$'s noninvertible roots multiplied by the product of $\phi(z)$'s noncausal roots (Lii and Rosenblatt, 1996). The generic notation f used here refers to a certain candidate density of Z_t . Since in reality it is unlikely to know the true distribution of Z_t , the error distribution is usually assumed to belong to a fairly general class of elliptical distributions (Breidt et al., 2001; Huang and Pawitan, 2000; Lii and Rosenblatt, 1996; Wu and Davis, 2010) to facilitate parameter estimation. The LAD methods (Breidt et al., 2001; Wu and Davis, 2010) maximize variants of (6) by using a Laplace error distribution and the objective functions generate consistent estimators under regularity conditions.

In order to relax the distributional assumptions, we consider a semiparametric model and take the noise distribution itself as a parameter. The model consists of two parts: the finite dimensional parameter β and the infinite dimensional nuisance parameter P . Both β and P are unknown. We adopt the classic semiparametric estimation procedures, which consist of estimating P first, followed by maximizing the resultant profile likelihood with respect to β .

In our framework, we consider the log-concave density projection method to estimate P in step one, that is, projecting the empirical measure of the residuals $\mathbb{P}_{\beta,n}$

onto the space of log concave distributions on \mathbb{R} to obtain a log concave maximum likelihood estimator of P (Cule and Samworth, 2010; Dümbgen et al., 2011). The profile log likelihood can be expressed as:

$$h_n(\beta) = \max_{f \in \mathcal{F}} h_{\beta, f}^n = L(\mathbb{P}_{\beta, n}) + \log \kappa(\beta) \text{ for } \beta \in \Theta. \quad (7)$$

Theorem 3.2. *Under the assumption that $P_0 \in \mathcal{P}$ and β_0 is a interior point of the compact parameter space Θ , there exists $(\hat{\beta}, \hat{f})$ that maximize $h_{\beta, f}^n$ over $\Theta \times \mathcal{F}$*

Proof. Note that $\beta \rightarrow \mathbb{P}_{\beta, n}$ defines a continuous mapping from Θ to the probability space \mathcal{P} equipped with the first moment Mallow's distance. On the other hand, the functional mapping $L(\cdot)$ is continuous on \mathcal{P} with respect to Mallow's distance. Therefore, $h_n(\beta)$ is a continuous function on Θ and attains its maximum on Θ at some $\hat{\beta} \in \Theta$. Then it follows that $(\hat{\beta}, \hat{f} := \Pi(\mathbb{P}_{\hat{\beta}}))$ maximizes $h_n(\beta, f)$ over $\Theta \times \mathcal{F}$. \square

The joint maximizer $(\hat{\beta}, \hat{f})$ is referred to as the maximum log-concave likelihood estimator (MLCLE). In Section 4, we will show $\hat{\beta}$ is strongly consistent.

4 Asymptotic results

4.1 Consistency

For causal-invertible ARMA models, $\kappa(\beta)$ is identically equal to one. Thus (6) reduces to the conditional log-likelihood of the sequence $\{X_i\}_{i=1}^n$. The maximizer $(\hat{\beta}, \hat{f})$ of (6) is exactly the estimator proposed in (Chen and Samworth, 2015), where consistency results were established. We now turn to the general case of noncausal/noninvertible models. The main result is contained in the following theorem.

Theorem 4.1. *In (1), suppose Z_t satisfies the following condition,*

$$L\left(\sum_{k=-\infty}^{\infty} d_k Z_{t-k}\right) \leq L(Z_t), \quad (8)$$

for any geometrically decaying sequence d_k with $\sum_{k=-\infty}^{\infty} d_k^2 \geq 1$ and the equality holding if and only if only one d_k is non-zero. Then,

$$\hat{\beta} \xrightarrow{a.s.} \beta_0 \text{ and } \int |\hat{f} - \Pi(P_0)| dx \xrightarrow{a.s.} 0 \text{ as } n \rightarrow \infty.$$

Remark 4.2. In the causal-invertible case, (Chen and Samworth, 2015) did not require condition (8). This is due to the fact that noncausal/noninvertible models were not allowed. So if one expands the family of models to be noncausal/noninvertible, then a condition like (8) is required even if the true model is causal-invertible.

We state the relevant consistency result shown in (Chen and Samworth, 2015) for comparison.

Proposition 4.3. *For causal-invertible ARMA models, assume that $P_0 \in \mathcal{P}$ and the parameter space Θ is compact, then*

$$\hat{\beta} \xrightarrow{a.s.} \beta_0 \quad \text{and} \quad \int |\hat{f} - \Pi(P_0)| \xrightarrow{a.s.} 0 \quad \text{as} \quad n \rightarrow \infty.$$

The consistency of $\hat{\beta}$ even when the true density is not log concave is a somewhat surprising and interesting result. The proof takes advantage of the property (2) of the $L(\cdot)$ function. In short, under causality and invertibility, Z_t is independent of $Z_t(\beta) - Z_t$. Therefore,

$$L(Z_t(\beta)) = L(Z_t + Z_t(\beta) - Z_t) \leq L(Z_t),$$

implying that β_0 is a global maximizer of $L(Z_t(\beta))$ over $\beta \in \Theta$. Furthermore, it can be shown that β_0 is actually the unique global maximizer, which is a key ingredient in verifying the consistency of maximum likelihood estimators. However, for non-causal/noninvertible models, the same argument does not apply since X_t may depend on future errors and Z_t is not independent of $Z_t(\beta) - Z_t$. We will show the strong consistency of the MLCLE for general ARMA processes from a different perspective. Recall that $Z_t(\beta)$ is a stationary ARMA process with AR polynomial $\phi_0(z)\theta(z)$ and MA polynomial $\phi(z)\theta_0(z)$, which is possibly noncausal or noninvertible. Since $\phi_0(z)\theta(z)$ and $\phi(z)\theta_0(z)$ have no roots on the unit circle, the Laurent expansion

$$\beta(z)\beta_0^{-1}(z) = \sum_{k=-\infty}^{\infty} a_k z^k$$

is valid on some annulus containing the unit circle. Correspondingly, $Z_t(\beta)$ can be represented as

$$Z_t(\beta) = \sum_{k=-\infty}^{\infty} a_k Z_{t-k}.$$

Proof of Theorem 4.1: From Remark 2.2, we have $L(\mathbb{P}_{\beta,n}) \xrightarrow{a.s.} L(P_\beta)$ for each $\beta \in \Theta$, that is, the profile log-likelihood function $h_n(\beta) = L(\mathbb{P}_{\beta,n}) + \log \kappa(\beta)$ converges almost surely to $h(\beta) = L(P_\beta) + \log \kappa(\beta)$. The proof of the theorem consists of two steps. First we show that the sequence of functions $\{L(\mathbb{P}_{\beta,n}) + \log \kappa(\beta)\}_n$ converges not only pointwise but uniformly to $L(P_\beta) + \log \kappa(\beta)$. Second we show that the limiting function $L(P_\beta) + \log \kappa(\beta)$ is uniquely maximized at β_0 .

- (i) *Uniform convergence of the sequence $\{L(\mathbb{P}_{\beta,n}) + \log \kappa(\beta)\}_n$*

Similar to the argument of the continuity of $h_n(\beta)$ in the proof of Theorem 3.2, the limiting function is continuous in β . Define

$$\Omega := \{\omega : \lim_{n \rightarrow \infty} \sup_{\beta \in \Theta} M_1(\mathbb{P}_{\beta,n}, P_\beta) = 0\}.$$

Then for fixed $\omega \in \Omega$, and for any convergent sequence $\{\beta^n\} \in \Theta$ with limit β^* , we have

$$\begin{aligned} M_1(\mathbb{P}_{\beta^n,n}, P_{\beta^n}) &\leq \sup_{\beta \in \Theta} M_1(\mathbb{P}_{\beta,n}, P_\beta) \\ \limsup_{n \rightarrow \infty} M_1(\mathbb{P}_{\beta^n,n}, P_{\beta^n}) &\leq \lim_{n \rightarrow \infty} \sup_{\beta \in \Theta} M_1(\mathbb{P}_{\beta,n}, P_\beta) = 0. \end{aligned}$$

Furthermore,

$$\limsup_{n \rightarrow \infty} M_1(\mathbb{P}_{\beta^n, n}, P_{\beta^*}) \leq \limsup_{n \rightarrow \infty} [M_1(\mathbb{P}_{\beta^n, n}, P_{\beta^n}) + M_1(P_{\beta^n}, P_{\beta^*})] = 0,$$

since the distribution of $Z_t(\beta_n)$ converges in the first moment Mallows distance to the distribution of $Z_t(\beta^*)$. Then according to Lemma 2.1,

$$|L(\mathbb{P}_{\beta^n, n}) - L(P_{\beta^*})| \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

As a result,

$$|L(\mathbb{P}_{\beta^n, n}) - L(P_{\beta^n})| \leq |L(\mathbb{P}_{\beta^n, n}) - L(P_{\beta^*})| + |L(P_{\beta^n}) - L(P_{\beta^*})| \rightarrow 0 \text{ for the fixed } \omega \in \Omega.$$

Since $\{\beta^n\}$ is arbitrary and Θ is compact, we have

$$\sup_{\beta \in \Theta} |L(\mathbb{P}_{\beta, n}) - L(P_{\beta})| \rightarrow 0 \text{ on } \Omega.$$

Now since the function $\kappa(\beta)$ is continuous and deterministic on Θ and the set Ω has probability one, this establish the uniform convergence of $\{L(\mathbb{P}_{\beta, n}) + \log \kappa(\beta)\}_n$ a.s..

(ii) *Unique maximizer of $L(P_{\beta}) + \log \kappa(\beta)$*

Denote the difference $L(P_{\beta_0}) + \log \kappa(\beta_0) - L(P_{\beta}) - \log \kappa(\beta)$ as $d(\beta)$:

$$\begin{aligned} d(\beta) &= L(Z_t) - L(Z_t(\beta)) + \log \frac{\kappa(\beta_0)}{\kappa(\beta)} \\ &= L(Z_t) - L\left(\sum_{k=-\infty}^{\infty} a_k Z_{t-k}\right) + \log \frac{\kappa(\beta_0)}{\kappa(\beta)} \\ &= L(Z_t) - L\left(\frac{\kappa(\beta_0)}{\kappa(\beta)} \sum_{k=-\infty}^{\infty} a_k Z_{t-k}\right). \end{aligned}$$

According to Proposition 6.2, $\left(\frac{\kappa(\beta_0)}{\kappa(\beta)}\right)^2 \sum_{k=-\infty}^{\infty} a_k^2 \geq 1$. Then by condition (8), $d(\beta) \geq 0$ for all $\beta \in \Theta$, or equivalently, β_0 is a global maximizer of $L(P_{\beta}) + \log \kappa(\beta)$. If there exists another $\beta \neq \beta_0 \in \Theta$ such that $d(\beta) = 0$, where the equal sign in (8) holds, then we know there is only one a_k being non-zero and the coefficients must satisfies

$$\left(\frac{\kappa(\beta_0)}{\kappa(\beta)}\right)^2 \sum_{k=-\infty}^{\infty} a_k^2 = 1.$$

The Laurent expansion of $\beta(z)\beta_0^{-1}(z)$ only has one non-zero coefficient. It then follows $\beta(z)\beta_0^{-1}(z) \equiv 1$ and $\beta = \beta_0$. Therefore, β_0 is the unique global maximizer of the limiting function $L(P_{\beta}) + \log \kappa(\beta)$.

Since the parameter space Θ is assumed to be compact, it follows from the continuous mapping theorem that the MLCLE $\hat{\beta}$ maximizing $L(\mathbb{P}_{\beta,n}) + \kappa(\beta)$ converges almost surely to β_0 . In addition,

$$M_1(\mathbb{P}_{n,\hat{\beta}}, P_{\beta_0}) \leq M_1(P_{\hat{\beta}}, P_{\beta_0}) + M_1(\mathbb{P}_{n,\hat{\beta}}, P_{\hat{\beta}}) \xrightarrow{a.s.} 0,$$

from which we conclude that $\int |\hat{f}_n - \Pi(P_0)| dx \xrightarrow{a.s.} 0$. \square

Verification of (8) has to be checked on a case-by-case basis. We show that (8) is true for log-concave distributions and symmetric α stable distributions with $\alpha \in (1, 2)$.

Corollary 4.4. *If Z_t is non-Gaussian and follows a log concave distribution, then the MLCLE $\hat{\beta}$ is strongly consistent for β_0 and $\int |\hat{f} - \Pi(P_0)| dx \xrightarrow{a.s.} 0$.*

Proof. We will use the celebrated Entropy Power Inequality from information theory, due to Shannon (Shannon, 2001), to show (8) is true for any non-Gaussian log-concave distribution. For completeness, this inequality is stated in Lemma 6.1.

For any random variable X that has a log-concave distribution, the entropy of X is well-defined. Let $H(X)$ denote the differential entropy of X . In this case, the log concave projection $\Pi(X)$ is exactly the true density of X itself, implying $L(X) = -H(X)$. For any geometrically decaying sequence $\{d_k\}_{k=-\infty}^{\infty}$ with $\sum_k d_k^2 \geq 1$, let $Y_j = \sum_{|k| \leq j} d_k Z_{t-k}$. Since the log concave measures are closed under convolution, Y_j also has log concave distribution under the assumption that Z_t is log concave. And hence, $L(Y_j) = -H(Y_j)$. Applying the Entropy-Power Inequality repeatedly, we obtain

$$\exp(2H(Y_j)) > \sum_{|k| \leq j} \exp(2H(d_k Z_{t-k})).$$

The strict inequality follows from the fact that Z_t is assumed to be non-Gaussian. Since $H(d_k Z_{t-k}) = H(Z_t) + \log|d_k|$ if $d_k \neq 0$,

$$\sum_{|k| \leq j} \exp(2H(d_k Z_{t-k})) = \exp(2H(Z_t)) \sum_{|k| \leq j} d_k^2,$$

and hence

$$H(Y_j) > H(Z_t) + \frac{1}{2} \log \left(\sum_{|k| \leq j} d_k^2 \right).$$

Then,

$$L(Y_j) < L(Z_t) - \frac{1}{2} \log \left(\sum_{|k| \leq j} d_k^2 \right). \quad (9)$$

It's straightforward to see that Y_j converges to $\sum_{k=-\infty}^{\infty} d_k Z_{t-k}$ in the first moment Mallow's distance as $j \rightarrow \infty$. Thus we can let j goes to infinity in (9) and obtain

$$L(Z_t) \geq L \left(\sum_{k=-\infty}^{\infty} d_k Z_{t-k} \right) + \frac{1}{2} \log \left(\sum_{k=-\infty}^{\infty} d_k^2 \right).$$

Since $\sum_{k=-\infty}^{\infty} d_k^2 \geq 1$ by assumption, we have

$$L(Z_t) \geq L\left(\sum_{k=-\infty}^{\infty} d_k Z_{t-k}\right).$$

When the equality holds, it's easy to see $\sum_{k=-\infty}^{\infty} d_k^2 = 1$. If there exists at least two non zero terms of these d_k s, $Y := \sum_{k=-\infty}^{\infty} d_k Z_{t-k}$ can be written as a sum of two non-degenerate independent random variables $Y^1 + Y^2$, where $Y^i = \sum_{k \in J_i} d_k Z_{t-k}$ for $i = 1, 2$ and J_1, J_2 is a partition of the integers. As linear combinations of independent non-Gaussian random variables, Y^1 and Y^2 are also non-Gaussian. By the Entropy Power Inequality,

$$\begin{aligned} \exp(2H(Y)) &> \exp\left(2H(Y^1)\right) + \exp\left(2H(Y^2)\right) \\ &\geq \sum_{k \in J_1} \exp(2H(d_k Z_{t-k})) + \sum_{k \in J_2} \exp(2H(d_k Z_{t-k})) \\ &\geq \exp(2H(Z_t)) \sum_{k \in J_1} d_k^2 + \exp(2H(Z_t)) \sum_{k \in J_2} d_k^2 \\ &\geq \exp(2H(Z_t)) \sum_k d_k^2. \end{aligned}$$

The first strict inequality is due to the non-Gaussianity of Y^1 and Y^2 . Although the Entropy Power Inequality only works for a finite sum, the second inequality follows from the fact that $\sum_{k \in J_i} d_k^2 \exp(2H(Z_{t-k}))$ is finite for $i = 1, 2$. By some simple algebra,

$$\begin{aligned} H(Y) &> H(Z_t) + \frac{1}{2} \log \sum_k d_k^2 \\ H(Y) &> H(Z_t). \end{aligned}$$

As Y is the weak limit of the log-concave distributed sequence $Y_j = \sum_{|k| \leq j} d_k Z_{t-k}$, Y has a log-concave distribution, indicating $L(Y) = -H(Y)$. It follows that

$$L\left(\sum_{k=-\infty}^{\infty} d_k Z_{t-k}\right) = L(Y) < L(Z_t)$$

strictly, which is a contradiction. Therefore, there is at most one nonzero d_k if

$$L\left(\sum_{k=-\infty}^{\infty} d_k Z_{t-k}\right) = L(Z_t).$$

And this nonzero term has absolute value one, which indicates that log-concave random variable satisfies (8). \square

Remark 4.5. Even under misspecification of log-concavity, the MLCLE may still be consistent in cases the true distribution P_0 is close to its log-concave projection $\Pi(P_0)$ and preserves the property (8). Simulation results suggests that $\hat{\beta}$ is still consistent given Z_t follows a student- t distribution which is not log concave, although not yet proved.

Corollary 4.6. *If Z_t is symmetric- α -stable with exponent $\alpha \in (1, 2)$, then*

$$\hat{\beta} \xrightarrow{a.s.} 0 \text{ and } \int |\hat{f}_n - \Pi(P_0)| dx \xrightarrow{a.s.} 0. \quad \text{as } n \rightarrow \infty.$$

Proof. For any geometrically decaying sequence $\{d_k\}_{k=-\infty}^{\infty}$ with $\sum_{k=-\infty}^{\infty} d_k^2 \geq 1$, $\sum_{k=-\infty}^{\infty} d_k Z_{t-k}$ is equal in distribution to $\left(\sum_{k=-\infty}^{\infty} |d_k|^\alpha\right)^{\frac{1}{\alpha}} Z_t$. Now for $\alpha \in (1, 2)$,

$$\left(\sum_{k=-\infty}^{\infty} |d_k|^\alpha\right)^{\frac{1}{\alpha}} \geq \left(\sum_{k=-\infty}^{\infty} d_k^2\right)^{\frac{1}{2}} \geq 1$$

Therefore,

$$\begin{aligned} L\left(\sum_{k=-\infty}^{\infty} d_k Z_{t-k}\right) &= L\left(\left(\sum_{k=-\infty}^{\infty} |d_k|^\alpha\right)^{\frac{1}{\alpha}} Z_t\right) \\ &= L(Z_t) - \log\left(\sum_{k=-\infty}^{\infty} |d_k|^\alpha\right)^{\frac{1}{\alpha}} \\ &\leq L(Z_t) \end{aligned}$$

When equality holds,

$$\sum_{k=-\infty}^{\infty} |d_k|^\alpha = \sum_{k=-\infty}^{\infty} d_k^2 = 1,$$

implying that there exists only one non-zero d_k with absolute value one and all other d_k s being zero. This completes the proof and hence (8) is satisfied. \square

4.2 Asymptotic properties

The asymptotic distribution of semiparametric M-estimators has been studied extensively in the literature Andrews (1994); Ichimura and Lee (2010); van der Vaart (1996). Unfortunately there is no general approach that is applicable to a wide range of problems. Rather, each modeling framework, which often involves the interaction of a nuisance parameter with the main parameter of interest, has to be considered on a case-by-case basis. Specifically, unlike the classical Taylor expansion of the maximum likelihood equations, the score function depends on an estimated and hence random nuisance parameter. Therefore, extra effort is needed to quantify the smoothness of the model with respect to the nonparametric component. We make the the following assumptions on f_0 , the true density for Z_t :

- A_1 $f_0(x) > 0$ for all x
- A_2 f_0 is continuously differentiable and φ_0'' is bounded
- A_3 $\int z \dot{f}_0 dz = z f_0(z)|_{-\infty}^{\infty} - \int f_0(z) dz = -1$ and $\int z f_0(z) dz = 0$
- A_4 f_0 is log-concave and non-Gaussian

Following the ideas in Chapter 7 of (van der Vaart, 2002), we construct a semi-parametric efficient estimator by using the efficient score function. For notational consistency, β is again used to denote the parameter vector, where $\beta = \phi$ is the autoregressive polynomial coefficients of the $AR(p)$ process. Define an augmented process \mathbf{X}_t as $(X_t, X_{t-1}, \dots, X_{t-p})^T$. Then the residuals $Z_t(\beta) = \phi(B)X_t = (1, -\beta^T)\mathbf{X}_t$ is a function of \mathbf{X}_t , and hence can be completely recovered from the data for $t = p+1, \dots, n$. So there is no need for truncation. The derivative of $Z_t(\beta)$ with respect to the vector β : $\dot{Z}_t(\beta)$, has a nice form in terms of \mathbf{X}_t , which is

$$\dot{Z}_t(\beta) = (-X_{t-1}, -X_{t-2}, \dots, -X_{t-p})^T = (\mathbf{0}_{p \times 1}, -I_{p \times p})\mathbf{X}_t.$$

To simplify notation, we ignore the index t and use Z_β and \dot{Z}_β to denote $Z_t(\beta)$ and $\dot{Z}_t(\beta)$, respectively, for a general t . Recall that the pseudo log-likelihood function is

$$l_{\beta, f}(Z_\beta) = \log f(Z_\beta) + \log \kappa(\beta) \quad (\beta, f) \in \Theta \times \mathcal{F}. \quad (10)$$

Since $f \in \mathcal{F}$ is a log concave function, it is differentiable at all but at most countably many points. If f is not differentiable at some point, use the left derivative instead. Then we can differentiate $l_{\beta, f}$ with respect to β and obtain the ordinary parametric score for β when f is fixed:

$$\dot{l}_{\beta, f} = \frac{\dot{f}(Z_\beta)}{f(Z_\beta)} \dot{Z}_\beta + \frac{\dot{\kappa}(\beta)}{\kappa(\beta)} \quad (11)$$

It has been shown in Davis and Song (2012) that the parametric score $\dot{l}_{\beta, f}$ is unbiased, that is,

$$\mathbb{E}_{\beta, f} \dot{l}_{\beta, f} = \mathbb{E}_{\beta, f} \left(\frac{\dot{f}(Z_\beta)}{f(Z_\beta)} \dot{Z}_\beta + \frac{\dot{\kappa}(\beta)}{\kappa(\beta)} \right) = 0 \quad (12)$$

given f satisfies $A_1 - A_3$. The efficient score function for β is defined to be the parametric score function $\dot{l}_{\beta, f}$ minus its orthogonal projection onto the closed linear span of the score functions for the nuisance parameter f (van der Vaart, 2002; Kosorok, 2007). By looking at the efficient score function, we can obtain a lower bound on the asymptotic variance of regular estimators at rate $n^{-\frac{1}{2}}$. See (Kreiss, 1987; Drost et al., 1997; Koul and Schick, 1997) for nice introductions to semiparametric estimation for time series models.

Nw we consider the efficient score function. For fixed $(\beta, f) \in \Theta \times \mathcal{F}$, define a path $s \rightarrow (\beta_s, f_s)$ given by:

$$\beta_s = \beta + sa, \quad f_s(\cdot) = (1 + sg(\cdot))f(\cdot), \quad (13)$$

where $a \in \mathbb{R}^p$ and $g(\cdot)$ is a bounded continuous function satisfying the constraint $\int_{\mathbb{R}} g(x)f(x)dx = 0$. The functions f_s are valid densities for s small enough, since

g is bounded. Differentiating the log-likelihood function $l_{\beta_s, f_s}(Z_{\beta_s}) = \log f_s(Z_{\beta_s}) + \log \kappa(\beta_s)$ with respect to s , we obtain the score function at (β, f) along the one-dimensional parametric submodel (13)

$$\begin{aligned}
S_{a,g} &:= \frac{\partial}{\partial s} l_{\beta_s, f_s} \big|_{s=0} \\
&= \frac{a^T \dot{Z}_{\beta_s} \dot{f}(Z_{\beta_s}) + g(Z_{\beta_s}) f(Z_{\beta_s}) + s \frac{\partial}{\partial s} [g(Z_{\beta_s}) f(Z_{\beta_s})]}{f(Z_{\beta_s}) + s g(Z_{\beta_s}) f(Z_{\beta_s})} + a^T \frac{\dot{\kappa}(\beta_s)}{\kappa(\beta_s)} \big|_{s=0} \\
&= a^T \left(\frac{\dot{f}(Z_\beta)}{f(Z_\beta)} \dot{Z}_\beta + \frac{\dot{\kappa}(\beta)}{\kappa(\beta)} \right) + g(Z_\beta) \\
&= a^T \dot{l}_{\beta, f} + g.
\end{aligned}$$

The information of this submodel is defined as

$$\mathcal{I}_{a,g} := \mathbb{E}_{\beta, f} (S_{a,g})^2.$$

For a fixed vector $a \in \mathbb{R}^p$, $\mathcal{I}_{a,g}$ is minimized over $g \in L^2(P_f)$ when g equals $g^*(u) := -a^T E_{\beta, f} [\dot{l}_{\beta, f} \mid Z_\beta = u]$, where P_f is the probability measure associated with density f . The minimal information over all paths is referred to as the efficient information. If the minimum is attained, the score of the submodel that has the minimal information (least favorable submodel) is the efficient score function. Thus, we take a candidate for the efficient score function to be of the form

$$\tilde{l}_{\beta, f} = \frac{\dot{f}(Z_\beta)}{f(Z_\beta)} \left(\dot{Z}_\beta - E_{\beta, f} [\dot{Z}_\beta \mid Z_\beta] \right) \quad (14)$$

since $\mathbb{E}_{\beta, f} (a^T \tilde{l}_{\beta, f})^2 = \inf_{g \in L^2(P_f)} \mathcal{I}_{a,g}$ for any $a \in \mathbb{R}$.

Proposition 4.7. *Replacing f with f_0 , we have $\mathbb{E}_{\beta, f_0} [\dot{Z}_t(\beta) \mid Z_t(\beta)] = \frac{\dot{\kappa}(\beta)}{\kappa(\beta)} Z_t(\beta)$. And hence,*

$$\tilde{l}_{\beta, f_0} = \frac{\dot{f}_0(Z_t(\beta))}{f_0(Z_t(\beta))} \left[\dot{Z}_t(\beta) - \frac{\dot{\kappa}(\beta)}{\kappa(\beta)} Z_t(\beta) \right]. \quad (15)$$

Proof. Each coordinate of the vector $\dot{Z}_t(\beta)$ admits a unique linear representation in terms of the sequence $\{Z_t(\beta)\}$, so $\dot{Z}_t(\beta)$ can be expressed as $\sum_{i=-\infty}^{+\infty} a_{\beta, i} Z_{t-i}(\beta)$, where $a_{\beta, i} \in \mathbb{R}^p$ is uniquely determined by β . Then we have

$$\begin{aligned}
\mathbb{E}_{\beta, f_0} \left(\frac{\dot{f}_0(Z_t(\beta))}{f_0(Z_t(\beta))} \dot{Z}_t(\beta) \right) &= \mathbb{E}_{\beta, f_0} \left(\frac{\dot{f}_0(Z_t(\beta))}{f_0(Z_t(\beta))} \sum_{i=-\infty}^{+\infty} a_{\beta, i} Z_{t-i}(\beta) \right) \\
&= \sum_{i=-\infty}^{+\infty} a_{\beta, i} \mathbb{E}_{\beta, f_0} \left(\frac{\dot{f}_0(Z_t(\beta))}{f_0(Z_t(\beta))} Z_{t-i}(\beta) \right) \\
&= a_{\beta, 0} \mathbb{E}_{\beta, f_0} \left(\frac{\dot{f}_0(Z_t(\beta))}{f_0(Z_t(\beta))} Z_t(\beta) \right) \\
&= -a_{\beta, 0},
\end{aligned}$$

which together with (12) implies that $a_{\beta,0} = \frac{\dot{\kappa}(\beta)}{\kappa(\beta)}$. Therefore,

$$\mathbb{E}_{\beta,f_0} [\dot{Z}(\beta) \mid Z_t(\beta)] = \mathbb{E}_{\beta,f_0} \left[\sum_{i=-\infty}^{+\infty} a_{\beta,i} Z_{t-i}(\beta) \mid Z_t(\beta) \right] = \frac{\dot{\kappa}(\beta)}{\kappa(\beta)} Z_t(\beta).$$

□

Remark 4.8. Here and after, let $\tilde{l}_{\beta,f} = \frac{\dot{f}(Z_t(\beta))}{f(Z_t(\beta))} \left[\dot{Z}_t(\beta) - \frac{\dot{\kappa}(\beta)}{\kappa(\beta)} Z_t(\beta) \right]$. Note that by such modification, $\tilde{l}_{\beta,f}$ may not be the efficient score function at points (β, f) other than (β, f_0) . The function $\tilde{l}_{\beta,f}$ is unbiased in the sense that

$$\mathbb{E}_{\beta,f_0} \tilde{l}_{\beta,f} = \mathbb{E}_{\beta,f_0} \left(\varphi'(Z(\beta)) \left(\dot{Z}(\beta) - \frac{\dot{\kappa}(\beta)}{\kappa(\beta)} Z(\beta) \right) \right) = 0, \quad (16)$$

where $\varphi = \log f$. Hence $\mathbb{E}_{\beta,f_0} \tilde{l}_{\beta,\hat{f}} = 0$.

Unfortunately, we are not able to show the asymptotic efficiency of the MLCLE $\hat{\beta}$. Alternatively, we follow the ideas of the one-step estimators constructed in Chapter 7 of (van der Vaart, 2002) and design a semiparametric efficient estimator. Set $\hat{\varphi}_{\sigma_n} = \log \hat{f}_{\sigma_n}$, where \hat{f}_{σ_n} is the smoothed log-concave density estimator. Write $\tilde{l}_{\beta,f}$ as a function of the augmented process $\{\mathbf{X}_t\}$:

$$\tilde{l}_{\beta,f}(\mathbf{X}_t) = \varphi' \left((1, -\beta^T) \mathbf{X}_t \right) \left[(\mathbf{0}_{p \times 1}, -I_{p \times p}) \mathbf{X}_t - \frac{\dot{\kappa}(\beta)}{\kappa(\beta)} \left((1, -\beta^T) \mathbf{X}_t \right) \right].$$

Suppose that an initial \sqrt{n} consistent estimator $\tilde{\beta}$ (LAD estimator as an example) for β_0 is available, and define the one-step estimator as

$$\check{\beta} := \tilde{\beta} - \left(\sum_{i=p+1}^n \tilde{l}_{\tilde{\beta}, \hat{f}_{\sigma_n}}(\mathbf{X}_i) \tilde{l}_{\tilde{\beta}, \hat{f}_{\sigma_n}}^T(\mathbf{X}_i) \right)^{-1} \sum_{i=p+1}^n \tilde{l}_{\tilde{\beta}, \hat{f}_{\sigma_n}}(\mathbf{X}_i). \quad (17)$$

Theorem 4.9. *Suppose that f_0 satisfies the conditions $A_1 - A_4$, and the efficient information matrix $\tilde{I}_{\beta_0, f_0} := E \left(\tilde{l}_{\beta_0, f_0} \tilde{l}_{\beta_0, f_0}^T \right)$ is nonsingular. Then, $\check{\beta}$ is asymptotic efficient at (β_0, f_0) in the sense that*

$$\sqrt{n}(\check{\beta} - \beta_0) \xrightarrow{\mathcal{D}} N(0, \tilde{I}_{\beta_0, f_0}^{-1}).$$

Proof. The function $\tilde{l}_{\beta, \hat{f}_{\sigma_n}}$ is unbiased according to (16) and satisfies the integrability conditions stated in Proposition 6.3. Then the conclusion follows from Theorem 7.2 in (van der Vaart, 2002). □

Remark 4.10. In practice, we can iterate by replacing $\tilde{\beta}$ with the last update $\check{\beta}$ in equation (17). We suspect that the MLCLE $\hat{\beta}$ is semiparametric efficient, though this is not yet proved. Simulation results for investigating its asymptotic behavior are included in Section 5.

5 Examples

5.1 Simulation study

A simulation study was conducted to evaluate the finite performance of the MLCLE and to compare with LAD and MLE methods, when the pdf of Z_t is known. We considered a mixed AR(2) process and a ARMA(1,1) process from a symmetric α -stable ($S\alpha S$) distribution, respectively, i.e.,

1. $X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} = Z_t$
2. $X_t - \phi X_{t-1} = Z_t - \theta Z_{t-1}$,

where $\{Z_t\}$ is a sequence of iid $S\alpha S$ random variables. Three values of α are considered: 1.1, 1.5, 1.9. For each case, a time series of length 500 is simulated and the parameters of interest are estimated by MLCLE, LAD and MLE methods. This procedure is replicated 5,000 times, and the results of this experiment are summarized in the following tables.

For the mixed AR(2) model, we set the true value (ϕ_1, ϕ_2) to be (1.2, 0.6), with a noncausal root in the AR polynomial. As shown in Table 1, for smaller α , the MLCLE is comparable to the LAD estimation. As α gets larger, the MLCLE outperforms the LAD estimation. In addition, as α decreases, both MLCLE and LAD estimation have improved performance. For the ARMA(1,1) model, we set the (ϕ, θ) to be (0.5, 1.5) and (1.5, 0.5). Similar conclusions as for Table 1 are seen in Table 2.

$Z_t \sim S\alpha S$				
True value	α	MLE	MLCLE	LAD
$\phi_1 = 1.2$ $\phi_2 = 0.6$	1.1	1.2002 (0.0141)	1.2011 (0.0159)	1.2005 (0.0157)
		0.6001 (0.011)	0.6005 (0.0129)	0.6002 (0.0125)
	1.5	1.2020 (0.0438)	1.2059 (0.0567)	1.2033 (0.0587)
		0.6003 (0.0327)	0.6014 (0.0365)	0.601 (0.0373)
	1.9	1.2059 (0.0709)	1.2034 (0.1124)	1.2045 (0.1449)
		0.5981 (0.0593)	0.6044 (0.0620)	0.6011 (0.0709)

Table 1: Mean and root-mean-squared error (\cdot) for MLE, MLCLE and LAD estimates for AR(2)

In regard to the asymptotic behavior, we consider an AR(1) process driven by the following log concave distributions: Laplace distribution with λ equal to one, logistic distribution with mean zero and scale parameter equal to one. Time series of lengths 100, 500, 5000 were simulated and for each realization, an AR(1) model was fitted via the MLCLE, LAD and MLE methods, respectively. For each sample size, this procedure was replicated 1000 times.

Tables 3 and 4 reports the mean, the root-mean-squared error and the asymptotic standard deviation of each method given different noise distributions. Note that the LAD coincides with MLE for the Laplace distribution. The MLCLE and MLE estimates are comparable for the two log-concave distributions. As the sample sizes grows, the normalized empirical variance $\hat{\sigma}^2$ by the MLCLE approaches the inverse efficient

$$Z_t \sim S\alpha S$$

True value	α	MLE	MLCLE	LAD
$\phi = 0.5$ $\theta = 1.5$	1.1	0.5000 (0.0059)	0.4998 (0.0071)	0.5000 (0.0070)
		1.5000 (0.0107)	1.5006 (0.0161)	1.5007 (0.0160)
	1.5	0.4999 (0.0182)	0.4994 (0.0205)	0.5002 (0.0210)
		1.4998 (0.0311)	1.5017 (0.0402)	1.5027 (0.0439)
	1.9	0.4994 (0.0364)	0.4977 (0.0422)	0.5000 (0.0479)
		1.5009 (0.0445)	1.5040 (0.0831)	1.5089 (0.1001)
$\phi = 1.5$ $\theta = 0.5$	1.1	1.5001 (0.0109)	1.5008 (0.0126)	1.5004 (0.0121)
		0.4999 (0.0059)	0.4998 (0.0100)	0.4999 (0.0105)
	1.5	1.5010 (0.0316)	1.5038 (0.0406)	1.5023 (0.0414)
		0.4997 (0.0188)	0.5000 (0.0211)	0.5001 (0.0219)
	1.9	1.5019 (0.0448)	1.5149 (0.0846)	1.5101 (0.0993)
		0.5001 (0.0364)	0.5014 (0.0423)	0.5000 (0.0499)

Table 2: Mean and root-mean-squared error (\cdot) for MLE, MLCLE and LAD estimates for ARMA(1,1)

information. For logistic distributions, the MLCLE outperforms the LAD estimates, suggesting the efficiency of the MLCLE.

$$Z_t \sim \text{Logistic}(0, 1), \phi = 2$$

n	MLE	MLCLE	LAD
100	2.1032 (0.4373) [4.3516]	2.1507 (0.4402) [4.3803]	2.1129 (0.5003) [4.9783]
500	2.0178 (0.1548) [3.4570]	2.0303 (0.1593) [3.5583]	2.0198 (0.1804) [4.0295]
5000	2.0025 (0.0473) [3.3410]	2.0032 (0.0473) [3.3473]	2.0023 (0.0545) [3.8557]

Table 3: Mean, root-mean-squared error (\cdot) and normalized empirical standard error [\cdot] of MLE, MLCLE and LAD estimates for non-causal AR(1) model

$$Z_t \sim \text{Laplace}(1), \phi = 2$$

	MLE	MLCLE
100	2.0694 (0.3681) [3.6628]	2.1267 (0.3849) [3.8298]
500	2.0115 (0.1208) [2.6994]	2.0196 (0.1238) [2.7663]
5000	2.0012 (0.0352) [2.4900]	2.0019 (0.0356) [2.5194]

Table 4: Mean, root-mean-squared error (\cdot) and normalized empirical standard error [\cdot] of MLE and MLCLE estimates for non-causal AR(1) model

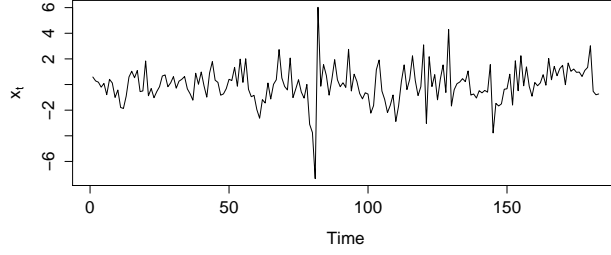


Figure 1: The demeaned differences of U.S. Total Government Revenue

5.2 An empirical example

Figure 1 contains the time series plot of the quarterly data of the demeaned differences of U.S. total government revenue from 1955:1 to 2000:4 (184 observations). The Jarque-Bera test for normality gives a p-value smaller than e^{-12} and the Shapiro-Wilk test gives a p-value smaller than e^{-8} . Both tests are significant and show strong evidence of rejecting normality of the data. The sample ACF and PACF plots of x_t in Figure 2 suggest fitting an AR(2) model to this data.

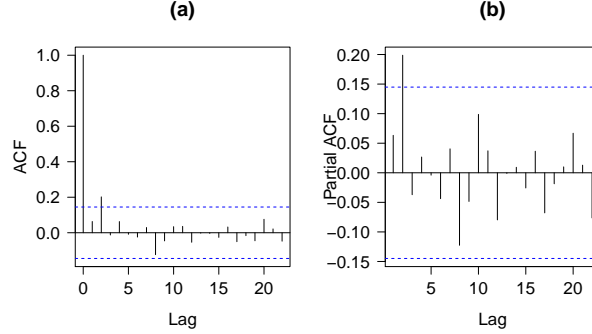


Figure 2: (a) Sample ACF of x_t , and (b) sample PACF of x_t

The best fitting causal Gaussian AR(2) model is given by

$$X_t - 0.0507X_{t-1} - 0.1995X_{t-2} = W_t.$$

While the sample ACF of the residuals $\{\hat{W}_t\}$ in Figure 3 indicate that \hat{W}_t is white noise, the ACF of the squared residuals $\{\hat{W}_t^2\}$ show significant lag one correlation. And hence, $\{\hat{W}_t\}$ is uncorrelated but not independent. In contrast, the best fitting mixed AR(2) model, by applying the MLCLE method, is given by

$$X_t - 1.3042X_{t-1} - 0.7606X_{t-2} = Z_t.$$

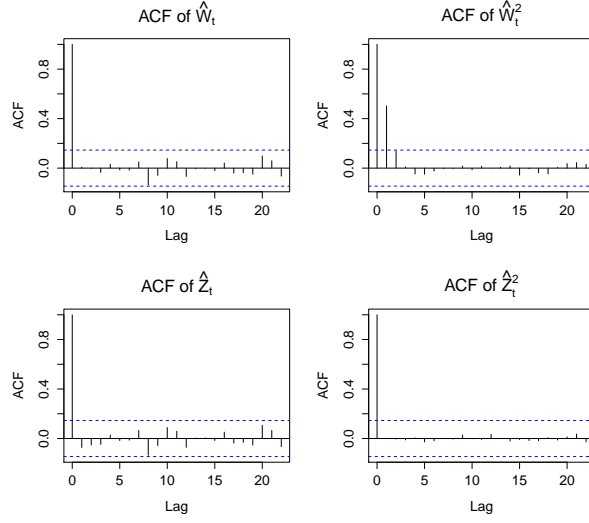


Figure 3: sample of ACF of the residuals \hat{W}_t and \hat{Z}_t

The AR polynomial $1 - \phi_1 z - \phi_2 z^2$ has one root inside the unit circle and one root outside the unit circle. The bottom panel of Figure 3 shows the plot of the ACF of $\{\hat{Z}_t\}$ and the ACF of $\{\hat{Z}_t^2\}$ from the mixed model. The ACF of $\{\hat{Z}_t\}$ looks very similar to those of $\{\hat{W}_t\}$, indicating both of them effectively remove the serial correlation structure in the data. Moreover, $\{\hat{Z}_t^2\}$ is also uncorrelated by looking at the ACF of $\{\hat{Z}_t^2\}$. Therefore, the noncausal model produces residuals that look more independent at least in terms of the squares of the residuals.

Acknowledgement

This research was supported in part by ARO MURI grant W911NF-12-1-0385.

6 Appendix: Auxiliary results and proof

Lemma 6.1. (*Entropy Power Inequality*)

$$\exp(2H(X + Y)) \geq \exp(2H(X)) + \exp(2H(Y))$$

where X and Y are independent real-valued random variables and $H(X)$ is the differential entropy of the probability density function f_X

$$H(X) = - \int_{\mathbb{R}} f_X(x) \log f_X(x) dx.$$

The equality holds if and only if X and Y are normal random variables.

Proof of Lemma 3.1: Since the parameter set Θ is assumed to be compact and all $\beta(z)$ have no zeros of absolute value one, there exists some $0 < \rho < 1$ and $K > 0$ such that $|a_j(\beta)| \leq K\rho^{|j|}$ for all j (see (Brockwell and Davis, 2009)).

$$\begin{aligned}
M_1(\mathbb{P}_{\beta,n}, \tilde{\mathbb{P}}_{\beta,n}) &= \sup_{\|g\|_L \leq 1} \left(\left| \int g d\mathbb{P}_{\beta,n} - \int g d\tilde{\mathbb{P}}_{\beta,n} \right| \right) \\
&\leq \frac{1}{n-2m(n)} \sum_{i=m(n)+1}^{n-m(n)} |Z_{i,m(n)}(\beta) - Z_i(\beta)| \\
&\leq \frac{1}{n-2m(n)} \sum_{i=m(n)+1}^{n-m(n)} \sum_{|j| > m(n)} K\rho^{|j|} |Z_{i-j}| \\
&\leq \frac{1}{n-2m(n)} \sum_{i=m(n)+1}^{n-m(n)} Y_{i,m(n)},
\end{aligned}$$

where $Y_{i,m(n)} = \sum_{|j| \geq m(n)} K\rho^{|j|} |Z_{i-j}|$. Denote the right-hand side of the last inequality above as W_n . Then, $\sum_{n=1}^{\infty} E(W_n)$ is finite since

$$EW_n = E \left(\frac{1}{n-2m(n)} \sum_{i=m(n)+1}^{n-m(n)} Y_{i,m(n)} \right) = \sum_{|j| \geq m(n)} K\rho^{|j|} E|Z_i| = 2 * KE(|Z_i|) \frac{\rho^{m(n)}}{1-\rho},$$

indicating that W_n converges to 0 almost surely by Borel-Cantelli lemma. Thus,

$$\sup_{\beta \in \Theta} M_1(\mathbb{P}_{\beta,n}, \tilde{\mathbb{P}}_{\beta,n}) \xrightarrow{a.s.} 0.$$

For any $\beta, \beta' \in \Theta$,

$$\begin{aligned}
M_1(\tilde{\mathbb{P}}_{\beta,n}, \tilde{\mathbb{P}}_{\beta',n}) &= \sup_{\|g\|_L \leq 1} \left(\left| \int g d\tilde{\mathbb{P}}_{\beta,n} - \int g d\tilde{\mathbb{P}}_{\beta',n} \right| \right) \\
&\leq \frac{1}{n-2m(n)} \sum_{i=m(n)+1}^{n-m(n)} |Z_i(\beta) - Z_i(\beta')| \\
&\leq \frac{1}{n-2m(n)} \sum_{i=m(n)+1}^{n-m(n)} \sum_{j=-\infty}^{\infty} |a_j(\beta) - a_j(\beta')| |Z_{i-j}| \\
&\leq \frac{1}{n-2m(n)} \sum_{i=m(n)+1}^{n-m(n)} \sum_{|j| \leq M} |a_j(\beta) - a_j(\beta')| |Z_{i-j}| + \frac{1}{n-2m(n)} \sum_{i=m(n)+1}^{n-m(n)} \sum_{|j| \geq M} 2K\rho^{|j|} |Z_{i-j}| \\
&\leq \frac{\max_{|j| \leq M} |a_j(\beta) - a_j(\beta')|}{n-2m(n)} \sum_{i=m(n)+1}^{n-m(n)} \sum_{|j| \leq M} |Z_{i-j}| + \frac{1}{n-2m(n)} \sum_{i=m(n)+1}^{n-m(n)} \sum_{|j| \geq M} 2K\rho^{|j|} |Z_{i-j}|
\end{aligned}$$

The second term converges almost surely to $4KE(|Z_i|)^{\frac{\rho^M}{1-\rho}}$. Therefore, it can be arbitrarily small by choosing M large, and for this large M ,

$$\frac{1}{n-2m(n)} \sum_{i=m(n)+1}^{n-m(n)} \sum_{|j| \leq M} |Z_{i-j}|$$

converges almost surely to some constant and one can show that

$$\max_{|j| \leq M} |a_j(\beta) - a_j(\beta')| \leq C \|\beta - \beta'\|$$

for some constant C not depends on β, β' . Therefore,

$$\lim_{\substack{n \rightarrow \infty \\ \|\beta - \beta'\| \rightarrow 0}} M_1(\tilde{\mathbb{P}}_{\beta,n}, \tilde{\mathbb{P}}_{\beta',n}) \xrightarrow{a.s.} 0.$$

On the other hand, notice that $M_1(\tilde{\mathbb{P}}_{\beta,n}, P_\beta) \xrightarrow{a.s.} 0$ since $Z_t(\beta)$ is stationary and ergodic, and hence $M_1(P_{\beta'}, P_\beta)$ is uniformly continuous on $\Theta \times \Theta$. This implies that $M_1(\tilde{\mathbb{P}}_{\beta,n}, P_\beta)$ is stochastically equicontinuous since

$$\left| M_1(\tilde{\mathbb{P}}_{\beta,n}, P_\beta) - M_1(\tilde{\mathbb{P}}_{\beta',n}, P_{\beta'}) \right| \leq M_1(\tilde{\mathbb{P}}_{\beta,n}, \tilde{\mathbb{P}}_{\beta',n}) + M_1(P_{\beta'}, P_\beta).$$

It follows that

$$\sup_{\beta \in \Theta} M_1(\tilde{\mathbb{P}}_\beta, P_\beta) \xrightarrow{a.s.} 0.$$

□

Proposition 6.2. *The coefficients a_k of the Laurent expansion of $\beta(z)\beta_0^{-1}(z)$ satisfies the inequality*

$$\left(\frac{\kappa(\beta_0)}{\kappa(\beta)} \right)^2 \sum_{k=-\infty}^{\infty} a_k^2 \geq 1. \quad (18)$$

Proof. Let $V_t = \sum_{k=-\infty}^{\infty} a_k W_{t-k}$ where $W_t \stackrel{iid}{\sim} N(0, 1)$. There exists a causal-invertible version of $V_t \equiv \sum_{k=0}^{\infty} a_k^* W_{t-k}^*$ with $a_0^* = 1$ and $\text{var}(W_t^*) = \left(\frac{\kappa(\beta)}{\kappa(\beta_0)} \right)^2$ (Brockwell and Davis, 2009)). Then we know

$$\sum_{k=-\infty}^{\infty} a_k^2 = \text{var}(V_t) = \sum_{k=0}^{\infty} a_k^{*2} \text{var}(W_{t-k}^*) \geq \text{var}(W_t^*) = \left(\frac{\kappa(\beta)}{\kappa(\beta_0)} \right)^2,$$

which implies that

$$\left(\frac{\kappa(\beta_0)}{\kappa(\beta)} \right)^2 \sum_{k=-\infty}^{\infty} a_k^2 \geq 1.$$

$\left(\frac{\kappa(\beta_0)}{\kappa(\beta)} \right)^2 \sum_{k=-\infty}^{\infty} a_k^2 = 1$ implies $a_k^* = 0$ for all $k \neq 0$. □

Proposition 6.3. *For every deterministic sequence β_n converges to β_0 , the sequence $\tilde{l}_{\beta_n, \hat{f}_{\sigma_n}}$ satisfies the following integrability conditions.*

$$\mathbb{E}_{\beta_n, f_0} \left[\left\| \tilde{l}_{\beta_n, f} - \tilde{l}_{\beta_0, f_0} \right\|^2 \right] \Big|_{f=\hat{f}_{\sigma_n}} = o_P(1).$$

Proof of Proposition 6.3: Let μ_n be the mode of \hat{f}_{σ_n} , then $\hat{\varphi}'_{\sigma_n} \geq 0$ for $x \leq \mu_n$ and $\hat{\varphi}'_{\sigma_n} \leq 0$ for $x \geq \mu_n$. It follows that

$$\int |\hat{\varphi}'_{\sigma_n}| \hat{f}_{\sigma_n}^{\frac{1}{3}} dx = \int_{-\infty}^{\mu_n} \hat{\varphi}'_{\sigma_n} \hat{f}_{\sigma_n}^{\frac{1}{3}} dx - \int_{\mu_n}^{\infty} \hat{\varphi}'_{\sigma_n} \hat{f}_{\sigma_n}^{\frac{1}{3}} dx = 6\hat{f}_{\sigma_n}^{\frac{1}{3}}(\mu_n) \xrightarrow{a.s.} 6f_0^{\frac{1}{3}}(\mu),$$

where μ is the mode of f_0 . Then, by following the same argument as Lemma 3 in (Cule and Samworth, 2010), $|\hat{\varphi}'_{\sigma_n}| \hat{f}_{\sigma_n}^{\frac{1}{3}}$ is uniformly bounded with probability one. Besides, there exists some $c > 0$ such that $\hat{f}_{\sigma_n} \geq cf_0$ with probability one according to the proof of Theorem 4.1 of (Cule et al., 2010). Thus, $|\hat{\varphi}'_{\sigma_n}|$ can be bounded by $f_0^{-\frac{1}{3}}$ up to some constant. Proposition 6.4 implies $\tilde{l}_{\beta_n, \hat{f}_{\sigma_n}}$ converges to \tilde{l}_{β_0, f_0} almost surely. Then the results follow from the dominated convergence theorem. \square

Proposition 6.4. *Assume that f_0 is continuously differentiable. Then for any compact set $S \subseteq \mathbb{R}$,*

$$\lim_{n \rightarrow \infty} \sup_{x \in S} |\hat{\varphi}_{\sigma_n}(x) - \varphi_0(x)| \xrightarrow{a.s.} 0 \text{ and } \lim_{n \rightarrow \infty} \sup_{x \in S} |\hat{\varphi}'_{\sigma_n}(x) - \varphi'_0(x)| \xrightarrow{a.s.} 0.$$

Proof. \hat{f}_{σ_n} is not only log-concave but also infinitely differentiable. Particularly, the first derivative of $\hat{\varphi}_{\sigma_n} := \log \hat{f}_{\sigma_n}$ exist. Since f_0 is assumed to be continuous, then according to Theorem 2 in (Chen and Samworth, 2013), we have

$$\lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} |\hat{f}_{\sigma_n}(x) - f_0(x)| \xrightarrow{a.s.} 0.$$

And accordingly, let S be any compact set in \mathbb{R} , we obtain

$$\lim_{n \rightarrow \infty} \sup_{x \in S} |\hat{\varphi}_{\sigma_n}(x) - \varphi_0(x)| \xrightarrow{a.s.} 0,$$

since \hat{f}_{σ_n} and f_0 are supported on the real line. Note that $\hat{\varphi}_{\sigma_n}$ and φ_0 are continuous concave functions. Thus, $\hat{\varphi}'_{\sigma_n}(x)$ converges pointwise to $\varphi'_0(x)$ as n goes to infinity. Further, since both $\hat{\varphi}'_{\sigma_n}$ and φ'_0 are continuous non-increasing functions, this pointwise convergence actually can be strengthened to be uniform, that is,

$$\lim_{n \rightarrow \infty} \sup_{\{x \in S\}} |\hat{\varphi}'_{\sigma_n}(x) - \varphi'_0(x)| \xrightarrow{a.s.} 0.$$

\square

References

- Andrews, B., M. Calder, and R. A. Davis (2009). Maximum likelihood estimation for α -stable autoregressive processes. *Annals of Statistics* 37(4), 1946–1982.
- Andrews, D. W. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, 43–72.
- Bagnoli, M. and T. Bergstrom (2006). Log-concave probability and its applications. pp. 217–241.
- Breidt, F. J., R. A. Davis, and A. A. Trindade (2001). Least absolute deviation estimation for all-pass time series models. *Annals of statistics*, 919–946.
- Brockwell, P. J. and R. A. Davis (2009). *Time Series: Theory and Methods*. Springer Science & Business Media.
- Chen, Y. and R. J. Samworth (2013). Smoothed log-concave maximum likelihood estimation with applications. *Statistica Sinica*, 1373–1398.
- Chen, Y. and R. J. Samworth (2015). Semiparametric time series models with log-concave innovations: maximum likelihood estimation and its consistency. *Scandinavian Journal of Statistics*, 42, 1–31.
- Cule, M. and R. Samworth (2010). Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electronic Journal of Statistics* 4, 254–270.
- Cule, M., R. Samworth, and M. Stewart (2010). Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72(5), 545–607.
- Davis, R. A. and L. Song (2012). Noncausal vector ar processes with application to economic time series. *Working paper, Columbia University*.
- Drost, F. C., C. A. Klaassen, and B. J. Werker (1997). Adaptive estimation in time-series models. *Annals of Statistics* 25(2), 786–817.
- Dümbgen, L. and K. Rufibach (2010). logcondens: Computations related to univariate log-concave density estimation. *Journal of Statistical Software*, to appear.
- Dümbgen, L., R. Samworth, and D. Schuhmacher (2011). Approximation by log-concave distributions, with applications to regression. *Annals of Statistics* 39(2), 702–730.
- Huang, J. and Y. Pawitan (2000). Quasi-likelihood estimation of non-invertible moving average processes. *Scandinavian Journal of Statistics*, 689–702.
- Ichimura, H. and S. Lee (2010). Characterization of the asymptotic distribution of semiparametric m-estimators. *Journal of Econometrics* 159(2), 252–266.

- Kosorok, M. R. (2007). *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media.
- Koul, H. L. and A. Schick (1997). Efficient estimation in nonlinear autoregressive time-series models. *Bernoulli* 3(3), 247–277.
- Kreiss, J.-P. (1987). On adaptive estimation in stationary arma processes. *Annals of Statistics*, 112–133.
- Lanne, M. and P. Saikkonen (2008). Modeling expectations with noncausal autoregressions. *Available at SSRN 1210122*.
- Levina, E. and P. Bickel (2001). The earth mover’s distance is the mallows distance: Some insights from statistics. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, Volume 2, pp. 251–256. IEEE.
- Lii, K.-S. and M. Rosenblatt (1996). Maximum likelihood estimation for nongaussian nonminimum phase arma sequences. *Statistica Sinica* 6(1), 1–22.
- Mallows, C. (1972). A note on asymptotic joint normality. *Annals of Mathematical Statistics*, 508–515.
- Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* 5(1), 3–55.
- van der Vaart, A. (1996). Efficient maximum likelihood estimation in semiparametric mixture models. *Annals of Statistics* 24(2), 862–878.
- van der Vaart, A. (2002). Part iii: Semiparameric statistics. *Lectures on Probability Theory and Statistics*, 331–457.
- Walther, G. (2009). Inference and modeling with log-concave distributions. *Statistical Science*, 319–327.
- Wu, R. and R. A. Davis (2010). Least absolute deviation estimation for general autoregressive moving average time-series models. *Journal of Time Series Analysis* 31(2), 98–112.