Time Series Modeling with Shape Constraints

Jing Zhang

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2017

ABSTRACT

Time Series Modeling with Shape Constraints

Jing Zhang

This thesis focuses on the development of semiparametric estimation methods for a class of time series models using shape constraints. Many of the existing time series models assume the noise follows some known parametric distributions. Typical examples are the Gaussian and t distributions. Then the model parameters are estimated by maximizing the resultant likelihood function.

As an example, the autoregressive moving average (ARMA) models (Brockwell and Davis, 2009) assume Gaussian noise sequence and are estimated under the causal-invertible constraint by maximizing the Gaussian likelihood. Although the same estimates can also be used in the causal-invertible non-Gaussian case, they are not asymptotically optimal (Rosenblatt, 2012). Moreover, for the noncausal/noninvertible cases, the Gaussian likelihood estimation procedure is not applicable, since any second-order based methods cannot distinguish between causal-invertible and noncausal/noninvertible models (Brockwell and Davis, 2009). As a result, many estimation methods for noncausal/noninvertible ARMA models assume the noise follows a known non-Gaussian distribution, like a Laplace distribution or a t distribution. To relax this distributional assumption and allow noncausal/noninvertible models, we borrow ideas from nonparametric shape-constraint density estimation and propose a semiparametric estimation procedure for general ARMA models by projecting the

underlying noise distribution onto the space of log-concave measures (Cule and Samworth, 2010; Dümbgen et al., 2011). 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. 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 estimator.

We also consider modeling time series of counts with shape constraints. Many of the formulated models for count time series are expressed via a pair of generalized state-space equations. In this set-up, the observation equation specifies the conditional distribution of the observation Y_t at time t given a *state-variable* X_t. For count time series, this conditional distribution is usually specified as coming from a known parametric family such as the Poisson or the Negative Binomial distribution. To relax this formal parametric framework, we introduce a concave shape constraint into the one-parameter exponential family. This essentially amounts to assuming that the *reference measure* is log-concave. In this fashion, we are able to extend the class of observation-driven models studied in Davis and Liu (2016). Under this formulation, there exists a stationary and ergodic solution to the state-space model. In this new modeling framework, we consider the inference problem of estimating both the parameters of the mean model and the log-concave function, corresponding to the reference measure. We then compute and maximize the likelihood function over both the parameters associated with the mean function and the reference measure subject to a concavity constraint. The estimator of the mean function and the conditional distribution are shown to be consistent and perform well compared to a full parametric model specification. The finite-sample behavior of the estimators is studied via simulation and two empirical examples are provided to illustrate the methodology.

Contents

List of Tables

List of Figures

Acknowledgments

I am extremely grateful to my Ph.D advisor, Richard A. Davis, for his constant encouragement and guidance. He is a great mentor and one of the smartest people I know. I hope that I could be as enthusiastic and energetic as Richard.

I would also like to thank Professor Zhiliang Ying, Professor Bodhisattva Sen, Professor Joel Cohen and Professor Chris Wiggins for agreeing to serve on my committee.

I am thankful to my friends at the Department of Statistics in Columbia University for always being so helpful and friendly.

Finally, I would like to thank my parents for all their patience, understanding and love.

To my parents, grandparents and those who educate me

Chapter 1

Introduction

1.1 ARMA models

The ARMA models are perhaps the most successful, well studied and easy to use models for the analysis of univariate time series (Brockwell and Davis, 2009; Rosenblatt, 2012; Box et al., 2015). These models form an important part of the classical literature in time series analysis. Probabilistic and statistical aspects of ARMA models related to model identification, estimation, model checking, and forecasting have been thoroughly investigated. A univariate stochastic process $\{X_t : t = 0, \pm 1, \pm 2, \dots\}$ is called an ARMA(p, q) process if it is stationary and satisfies the difference equations

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \quad \text{for all } t = 0, \pm 1, \pm 2, \dots \tag{1.1}$$

where the noise sequence $\{Z_t\}$ is assumed to be independently and identically distributed (iid) random variables with zero mean and variance σ^2. In many applications, the independence assumption can be replaced by the weaker condition that $\{Z_t\}$ is white noise. Throughout this thesis, we consider the iid setting. Moreover, for some applications, the noise sequence $\{Z_t\}$ can be allowed to have infinite variance, for example, assuming symmetric $\alpha-$stable noise (Cline and Brockwell, 1985).

Define the autoregressive (AR) polynomial of degree p by $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ and the moving average (MA) polynomial of degree q by $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$. Then (1.1) can be written in a compact form

$$\phi(B)X_t = \theta(B)Z_t, \quad t = 0, \pm 1, \pm 2, \ldots,$$

where B is the backward-shift operator defined by $B^j X_t = X_{t-j}$ for $j = 0, \pm 1, \pm 2, \ldots$. The polynomials $\phi(z)$ and $\theta(z)$ are assumed to have no common roots. Then, the recursive equations (1.1) admits a unique stationary solution if and only if the AR polynomial $\phi(z)$ has no roots on the unit circle, that is, $\phi(z) \neq 0$ for any $|z| = 1$. The solution is given by

$$X_t = \frac{\theta(B)}{\phi(B)} Z_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \tag{1.2}$$

where $\sum_{j=-\infty}^{\infty} \psi_j z^j$ is the Laurent expansion of $\frac{\theta(z)}{\phi(z)}$ in an annulus $\{z : R < |z| < \frac{1}{R}\}$ with $0 < R < 1$. When $\theta(z)$ has no zeros on the unit circle, $\{Z_t\}$ also has a two sided representation in terms of $\{X_t\}$:

$$Z_t = \frac{\phi(B)}{\theta(B)} X_t = \sum_{j=-\infty}^{\infty} \pi_j X_{t-j}, \tag{1.3}$$

where $\frac{\phi(z)}{\theta(z)} = \sum_j \pi_j z^j$ in an annulus $\{z : r < |z| < \frac{1}{r}\}$ with $0 < r < 1$. Throughout this thesis, we assume that the polynomial $\phi(z)\theta(z)$ has no zeros on the unit circle such that the equations (1.2) and (1.3) are well-defined.

1.1.1 Minimum and nonminimum phase ARMA models

An ARMA(p, q) process is said to be causal-invertible (minimum phase) if

$$\phi(z)\theta(z) \neq 0 \text{ for any } z \in \mathbb{C} \text{ with } |z| \leqslant 1.$$

That is, both the AR and MA polynomials have no zeros inside the unit circle. In such cases, X_t can be expressed as a function of only the present and the past noise $\{Z_s : s \leqslant t\}$, i.e., $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ as in (1.2). If $\phi(z)$ has any roots inside the unit circle, there are negative power terms in (1.2) and X_t depends on the future noise variables and we say X_t is noncausal. Correspondingly, invertibility means that Z_t can be written as a causal function of $\{X_t\}$: $Z_t = \sum_{j=0} \pi_j X_{t-j}$ and it only depends on the present and past observations $\{X_s : s \leqslant t\}$. If the MA polynomial $\theta(z)$ has no roots inside the unit circle, then the ARMA(p, q) process is invertible; otherwise, the process is said to be noninvertible.

It turns out that a general ARMA(p, q) process, $\phi(B)X_t = \theta(B)Z_t$, which is possibly noncausal/noninvertible (nonminimum phase), admits an equivalent causal-invertible representation (Brockwell and Davis, 2009). More specifically, we can find polynomials $\phi^*(z)$, $\theta^*(z)$ and a white noise sequence $\{Z_t^*\}$ that satisfy the difference equations

$$\phi^*(B)X_t = \theta^*(B)Z_t^* \quad t = 0, \pm 1, \pm 2, \ldots,$$

where $\phi^*(z)$ and $\theta^*(z)$ have no zeros inside the unit circle. However, $\{Z_t^*\}$ is not independent in general. In fact, $\{Z_t^*\}$ is iid if and only if $\{Z_t\}$ is Gaussian, otherwise, $\{Z_t^*\}$ is only uncorrelated (Breidt et al., 2001). Thus, the Gaussian likelihood cannot distinguish between causal-invertible and noncausal/noninvertible ARMA models. The assumptions of causality and invertibility are necessary to ensure identifiability of the model parameters when using the Gaussian likelihood or any second-order based estimation method.

Let $\phi = (\phi_1, \ldots, \phi_q)^T$ and $\theta = (\theta_1, \ldots, \theta_q)^T$ denote the AR and MA coefficients, respectively. If $\{Z_t\}$ is Gaussian, the observed vector $\mathbf{X}_n = (X_1, \ldots, X_n)'$ is also Gaussian with zero mean and covariance matrix denoted as $\Gamma_n(\phi, \theta, \sigma^2)$. The likelihood of $\mathbf{X}_n$ is

$$L_n(\phi, \theta, \sigma^2) = (2\pi \det \Gamma_n(\phi, \theta, \sigma^2))^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\mathbf{X}_n' \Gamma_n^{-1}(\phi, \theta, \sigma^2)\mathbf{X}_n\right). \qquad (1.4)$$

The maximizer of the likelihood function above, $(\hat{\phi}, \hat{\theta})$, is asymptotically efficient for estimating the causal-invertible ARMA models driven by Gaussian noise. If $\{Z_t\}$ is non-Gaussian, $L_n(\phi, \theta, \sigma^2)$ is referred to as the quasi-Gaussian likelihood function and $(\hat{\phi}, \hat{\theta})$ that maximizes (1.4) is still consistent and asymptotically normal for the true causal-invertible parameters, but is no longer efficient. One can also derive the Gaussian likelihood function by conditioning on the previous observations for causal-invertible models, by which we have

$$L_n(\phi, \theta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} (r_1 r_2 \cdots r_n)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{(X_i - \hat{X}_j)^2}{r_j}\right), \qquad (1.5)$$

where $\hat{X}_j$ is the one-step linear predictor of X_j given $\{X_1, \ldots, X_{j-1}\}$ and $r_j = \frac{E\left(X_j - \hat{X}_j\right)^2}{\sigma^2}$. See Brockwell and Davis (2009) for innovation algorithms to calculate $\hat{X}_j$ and r_j.

The conditional expectation $E\left[X_t \mid X_s, s < t\right]$ is known to be a linear combination of $X_s, s < t$ when $\{Z_t\}$ is Gaussian. In fact, in the minimum phase case, $E\left[X_t \mid X_s, s < t\right]$ is the same as the Gaussian case for any iid sequence $\{Z_t\}$ (Rosenblatt, 2012). When $\{Z_t\}$ is non-Gaussian, we can remove the causal-invertible constraint and allow noncausal/noninvertible ARMA models. In the nonminimum phase case, the conditional expectation $E\left[X_t \mid X_s, s < t\right]$ is no longer a linear function of $X_s, s < t$ since X_t depends on future noises. Nonminimum phase ARMA models driven by non-Gaussian noise sequences 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 examples where noncausal models fit better than causal ones. Allowing noncausality/noninvertibility can enlarge the pool of ARMA models, eliminate more of the serial dependence of the residuals and enhance our understanding of the data.

Statistical inference for nonminimum phase models is comparatively limited due to the complicated dependence structure of the process. The standard least squares methods developed under the causal-invertible constraint are only Gaussian efficient and are not applicable for nonminimum phase models. Many of the existing estimation procedures are based on the idea of maximum likelihood estimation by assuming a common pre-specified noise distribution, such as a Laplace or a t distribution. Breidt et al. (1991) considered inference for the parameters of possibly noncausal AR models by factoring the AR polynomial $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p = \phi^\dagger(z)\phi^*(z)$ with

$$\phi^\dagger(z) = 1 - \theta_1 z - \cdots - \theta_r z^r \neq 0 \text{ for } |z| \leqslant 1$$

$$\phi^*(z) = 1 - \theta_{r+1} z - \cdots - \theta_p z^s \neq 0 \text{ for } |z| \geqslant 1$$

$r, s \geqslant 0, r + s = p$. They derived approximations of the likelihood function and showed the consistency and asymptotic efficiency of the MLE of the parameters $(\theta_1, \ldots, \theta_p)$. Lii and Rosenblatt (1992) established similar results for noninvertible MA models. Lii and Rosenblatt (1996) further studied inference for nonminimum phase ARMA models driven by non-Gaussian noises. They proposed an approximate maximum likelihood estimation procedure and established the consistency and asymptotic normality of the MLE.

The least absolute deviation (LAD) criterion based on quasi-Laplace likelihood function is frequently used for modeling time series in the non-Gaussian setting. While the absolute-deviation-type estimators are obtained by assuming a Laplace distribution for the noise, they are still consistent even when the noise distribution is not Laplace under mild conditions.

Huang and Pawitan (2000) established consistency of LAD estimators for noninvertible MA processes driven by standard heavy tailed errors. Breidt et al. (2001) studied LAD estimation for all-pass time series models. All-pass models are a special class of ARMA models where all of the roots of the AR polynomial are reciprocals of the roots of the MA polynomials and vice versa. They generate uncorrelated time series, but these series are not independent in the non-Gaussian case. Wu and Davis (2010) proposed a LAD estimation procedure for nonminimum phase ARMA models and established the consistency and asymptotic normality of the LAD estimators.

Moreover, there is a large literature on the nonminimum phase models estimation based on rank (Andrews et al., 2007) or cumulants of order greater than two (Nikias and Petropulu, 1993). In this thesis, we limit our interest to likelihood based estimation methods.

1.1.2 Semiparametric inference for ARMA mdoels

Maximum likelihood based estimation procedures usually require full knowledge of the underlying noise distribution. However, the error distribution is rarely known in practice. It may be more realistic to learn the distribution of the noise from the data using a nonparametric estimation approach. The ARMA(p, q) model (1.1) has two parameters: the AR/MA coefficients (ϕ, θ) and the noise distribution P. It is then very natural to study ARMA models from the semiparametric perspective, in which we have a finite dimensional parameter (ϕ, θ) and an infinite dimensional parameter P.

Semiparametric inference forms a very important part of classical statistical modeling. It enjoys the flexibility of nonparametric modeling and has various important applications. Kosorok (2007) presents an overview of semiparametric inference techniques and provides full treatments of several useful examples. See also van der Vaart (2002) and Tsiatis (2007). Estimation of a semiparametric model is more difficult than estimation of any parametric submodel. A regular estimator is said to be semiparametric efficient if its information is equal to the minimum of the information over all efficient estimators for all parametric submodels. If there exists a parametric submodel that attains this minimum, then it is called a least favorable submodel. For a semiparameric model $P_{\beta,f}$, where β is the finite dimensional parameter and f is the infinite dimensional parameter, the semiparametric estimators are obtained by jointly maximizing the likelihood function over the parameter space of β and f. The semiparametric MLE of β, $\hat{\beta}$, depends on a random element $\hat{f}$ and so is the score function of $\hat{\beta}$. Thus, the classical Taylor expansion of the maximum likelihood equations is not applicable and the semiparametric efficiency of $\hat{\beta}$ is not guaranteed. Extra effort is needed to quantify the smoothness of the model with respect to the nonparametric component. See a general approach for asymptotic efficiency of semiparametric estimators via computing efficient score function and constructing a least favorable submodel in van der Vaart (2002); Kosorok (2007).

Kreiss (1987) considered the problem of estimating the parameters of minimum phase ARMA models when the noise distribution was unknown. He constructed adaptive estimates based on the kernel density estimator of the noise distribution. This methodology was able to establish local asymptotic normality (LAN) of minimum phase ARMA processes. Gassiat (1993) showed LAN properties and obtained LAM estimators for noncausal AR processes provided the noise distribution was known. She showed that adaptive efficient estimation was impossible for the parameters ϕ of noncausal AR models when the noise distribution was unknown. Drost et al. (1997) and Koul and Schick (1997) studied adaptive estimation for more general time series models. We aim to develop a semiparametric estimation procedure for nonminimum phase ARMA processes using the theory from nonparametric log-concave density estimation. The asymptotic properties of the resulting semiparametric estimators are also studied. This topic is the subject of Chapter 2.

1.2 Time series of counts models

Time series of counts arises naturally from counting the number of discrete events over some period of time. There are two main frameworks that are typically used for time series of counts data (Cox et al., 1981): parameter-driven and observation-driven. Parameter-driven models assume the conditional mean process depends solely on a latent process while observation-driven models formulate the conditional mean process explicitly as a function of the lagged observations. Estimation for parameter-driven models is difficult since they depend on a latent process and it is not easy to evaluate the likelihood function. Simulation-based numerical methods are used to obtain parameter estimates. In contrast, since the conditional mean process of observation-driven models is a function of past observations, it is relatively easy to obtain parameter estimates via maximum likelihood method. However, stability properties, such as stationarity and ergodicity, are difficult to derive.

We consider observation-driven time series of counts models in this thesis. The classical ARMA models driven by noise with a continuous distribution are not applicable for modeling count data. Many time series of counts models then fall into the generalized linear model (GLM) framework where the conditional distribution of the response is assumed to belong to an exponential family. One typically assumes a Poisson distribution (Davis et al., 2003; Heinen, 2003; Ferland et al., 2006; Fokianos et al., 2009) or a Negative Binomial distribution (Davis and Wu, 2009; Christou and Fokianos, 2014). The observations are generated as

$$Y_t \mid \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t),$$

where $\mathcal{F}_t$ is the filtration generated by observations up to time t; $\lambda_t = E\left[Y_t \mid \mathcal{F}_{t-1}\right]$ is the conditional mean process. For the integer-valued generalized autoregressive conditional heteroscedasticity (INGARCH) (p, q) process, the conditional mean process is modeled as

$$\lambda_t = \gamma_0 + \sum_{i=1}^{p} \gamma_i Y_{t-i} + \sum_{j=1}^{q} \delta_j \lambda_{t-j},$$

where $\gamma_0 > 0, \gamma_j \geqslant 0, i = 1, \ldots, p, \delta_j \geqslant 0, j = 1, \ldots, q$. For Poisson INGARCH$(p, q)$ models, the conditional mean happens to also be the conditional variance. The INGARCH(p, q) process is an integer-valued analogue of a GARCH(p, q) process introduced in Bollerslev (1986). Ferland et al. (2006) considered the Poisson INGARCH(p, q) model and showed the second-order stationarity of the process. Fokianos et al. (2009) studied the consistency and the asymptotic normality of the MLE for Poisson autoregressive models, where more general conditional mean models were considered. Davis and Liu (2016) focused on observation-driven models and studied the theory and inference for a relatively comprehensive class of count time series models, where the observations were assumed to follow a one-parameter exponential family given the conditional mean process that was modeled as a function of lagged observations.

The models considered in Davis and Liu (2016) can be viewed as an extension of the GLM framework, although no covariates are involved. They showed the stationarity and ergodicity of the underlying processes under fairly general conditions and established the asymptotic normality of the maximum likelihood estimators. Another GLM type model is the generalized linear autoregressive moving average (GLARMA) model (Shephard, 1995; Davis et al., 1999, 2003; Davis and Wu, 2009), where the observations are assumed to be generated from a one-parameter exponential family conditional on a latent process and covariates. See an overview of the likelihood-based estimation methods for analysis and modeling of count time series from the GLM perspective in Liboschik et al. (2015); Fokianos (2015).

We exploit the exponential family distribution assumption in GLM and propose a semiparametric estimation framework for modeling time series of counts data. Assume Y_t given the past history $\mathcal{F}_{t-1}$ is generated as

$$Y_t \mid \mathcal{F}_{t-1} \sim p(y \mid \eta_t, \varphi),$$

where $p(y \mid \eta, \varphi) = \exp\left(\varphi(y) + \eta y - A_\varphi(\eta)\right)$ is from an exponential family. Here $A_\varphi(\eta) = \log \sum_{y=0}^{\infty} \exp\left(\varphi(y) + \eta y\right)$. Let $B_\varphi(\eta) = A'_\varphi(\eta)$. Then the conditional mean process $X_t = E\left[Y_t \mid \mathcal{F}_{t-1}\right]$ enters through the link function $X_t = B_\varphi(\eta_t)$. Allowing the baseline function $\varphi(y)$ to vary as a parameter leads to a rich and flexible class of models. Many time series of counts models naturally fall into this formulation. We impose a concave shape constraint on the function $\varphi(y)$ and illustrate the semiparametric estimation procedure for the observation-driven models considered in Davis and Liu (2016). The MLEs are shown to be strongly consistent. This work is described in Chapter 4.

1.3 Shape constraint function estimation

Shape constraint function estimation has been receiving increasing interest in nonparametric inference. Instead of making smoothness assumptions on the underlying nonparametric function and using local averaging methods, e.g., kernel smoothing, it assumes the function of interest satisfies certain qualitative constraints, such as monotonicity, convexity or concavity. The corresponding estimation procedure is usually fully automatic and is free of choosing tuning parameters like bandwidth or kernel functions (Dümbgen and Rufibach, 2009; Seijo and Sen, 2011).

Grenander (1956) was the first to study the non-parametric maximum likelihood estimation of a monotone density. Since then, various nonparametric shape restricted estimators for different density estimation/regression problems have been developed. For example, the nonparametric least squares estimator of a convex (concave) regression function was considered in Birke and Dette (2007), Shively et al. (2009), Seijo and Sen (2011) and Chen and Samworth (2015a), and nonparametric least squares estimation of a monotone regression function was given in Brunk (1955) and Mammen (1991). Related consistency and asymptotic properties have been studied under suitable smoothness conditions. A general framework for isotonic optimization (optimizing the likelihood and computing function estimators) is described in Mair et al. (2009). Dümbgen and Rufibach (2009), Cule and Samworth (2010), and Cule et al. (2010) have given comprehensive characterizations of nonparametric MLE of log-concave densities in regards to existence, consistency, and other theoretical properties. Kim and Samworth (2016) showed that the minimax risk of log-concave density estimation with respect to the squared Hellinger loss is lower bounded by $n^{-\frac{4}{5}}$ for $d = 1$ and $n^{-\frac{2}{d+1}}$ for $d \geqslant 2$, where d is the dimension of the observations. Computational details of univariate and multivariate log-concave density estimators are presented in Dümbgen and Rufibach (2010) and Cule et al. (2009), respectively.

A probability density function f is said to be log-concave if $\log f$ is a concave function. The family of log-concave measures is a very useful nonparametric class of distributions and behaves almost like a parametric class (Walther, 2009; Schuhmacher et al., 2011). It contains many of the commonly used parametric distributions such as the Gaussian density, logistic density, Laplace density and Gamma densities parameter greater than or equal to one (Bagnoli and Bergstrom, 2006). Dümbgen et al. (2011) showed that there exists a unique log-concave density that maximizes the log likelihood type functional $\int \log f \, dP$ over the class of log-concave densities if and only if the probability measure P is non-degenerate and has finite first moment. The maximizer of $\int \log f \, dP$ is referred to as the log-concave projection of P onto the space of log-concave densities, which was successfully applied to regression problems in Dümbgen et al. (2011). Chen and Samworth (2015b) developed a semiparametric estimation framework for a class of time series models including causal-invertible ARMA models by projecting the noise distributions onto the class of log-concave densities. Inspired by Chen and Samworth (2015b), we study general ARMA processes without the causal-invertible constraint using log-concave projection in Chapter 2. We also consider estimating causal vector autoregressive (VAR) models using log-concave projection in Chapter 3. Chen and Samworth (2015a) considered inference for generalized additive models with shape restrictions (monotonicity, convexity, and concavity) on each additive regression function where the response given the covariates follows an exponential family distribution. Time series of counts models usually assume that the observations follow an exponential family distribution given the conditional mean process. We exploit this assumption and propose a semiparametric GLM framework with concave shape constraints for modeling time series of counts data in Chapter 4.

1.4 Organization of the thesis

Chapter 2 considers inference for nonminimum phase ARMA models driven by non-Gaussian noise and presents a semiparametric estimation procedure using the log-concave density estimator. Chapter 3 generalizes this method to causal VAR models. Chapter 4 derives an extension of the natural one-parameter exponential family by imposing a concave shape constraint on the baseline function and develops a semiparametric estimation procedure for the observation-driven time series of counts models. Chapter 5 summarizes our conclusions and discusses open research questions.

Chapter 2

Semiparametric Estimation for Nonminimum Phase ARMA Models

2.1 Introduction

This chapter focuses on the inference for nonminimum phase ARMA models given in (1.1) driven by non-Gaussian noise. In the case of minimum phase models, one often resorts to maximizing the Gaussian likelihood even if the noise is non-Gaussian. The parameters estimated in this fashion have the same asymptotic behavior as in the "Gaussian" case. However, Gaussian likelihood is blind to minimum and nonminimum phase models. 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 and it is very natural and useful to consider nonminimum phase models.

Many of the existing estimation methods postulate a known noise distribution. To relax the parametric distributional assumption, we extend the maximum likelihood principle to a nonparametric framework and consider semiparametric models. Chen and Samworth (2015b) 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 (2015b), we apply the log-concave projection method to the noncausal/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. We conjecture that the semiparametric estimators are asymptotically normal, although not proved yet.

The rest of the chapter is organized as follows. Section 2.2 provides a quick review of the definitions and basic properties of log-concave densities and log-concave projection. Section 2.3 applies log-concave projection to general ARMA models and derives the objective function. Section 2.4 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 2.5 presents a simulation study and a real data application to further illustrate the results in Section 2.4. Appendix 2.6.1 contains the proofs of the propositions. Appendix 2.6.2 presents the current progress in studying the asymptotic properties of the semiparametric MLE for AR models.

2.2 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) and Walther (2009). It has been shown that for a given probability measure P on $\mathbb{R}^d$, there exists a unique log-concave density

f that maximizes the log-likelihood type functional (i.e., the Kullback-Leibler discrepancy)

$$D(f, P) := \int_{\mathbb{R}^d} \log f dP,$$

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), and Walther (2009) for more details.

Let $\mathcal{P}$ denote the class of all probability measures P on $\mathbb{R}^d$ such that $\int \|x\| \, dP < \infty$ and $P(H) < 1$ for any hyperplane $H \subset \mathbb{R}^d$. When $d = 1$, this means we rule out the Dirac measure. Let $\mathcal{F}$ be the set of log concave densities on $\mathbb{R}^d$. Then the functional mapping $\Pi : \mathcal{P} \to \mathcal{F}$

$$\Pi(P) = \arg\max_{f \in \mathcal{F}} D(f, P)$$

is well-defined if and only if $P \in \mathcal{P}$ (Dümbgen et al., 2011). The quantity $\Pi(P)$ is referred to as the log-concave projection of P onto $\mathcal{F}$. The maximal function $L : \mathcal{P} \to \mathbb{R}^d$ is defined as

$$L(P) = \max_{f \in \mathcal{F}} D(f, P),$$

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 supported on some hyperplane of $\mathbb{R}^d$, $L(P) = \infty$). In particular, if $P \in \mathcal{F}$, and hence has all moments, then $\Pi(P) = P$. 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:

   $$L(a+CX) = L(X) - \log|\det C| \text{ for any } a \in \mathbb{R}^d \text{ and nonsingular } d \times d \text{ real matrix } C. \tag{2.1}$$

2. Non-increasing under convolution:

   $$L(X+Y) \leqslant L(X) \tag{2.2}$$

   if X is independent of Y and $X \in \mathcal{P}$. The equal sign holds if and only if $Y = \delta_a$ for some vector $a \in \mathbb{R}^d$.

3. Mean preservation:

   $$\int_{\mathbb{R}^d} x\, d\,P(x) = \int_{\mathbb{R}^d} x\, \Pi(P)(x)\, d\,x.$$

Further interesting properties of Π and L have been 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 measures: the first moment Mallows distance 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

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

where X and Y are any integrable random variables distributed as P and Q respectively, and F is a 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). Kantorović and Rubinśteín (1958) established a useful duality formula for Mallow's distance:

$$M_1(P, Q) := \sup_{\|g\|_L \leqslant 1} \left| \int g \, d\,(P - Q) \right| \tag{2.3}$$

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 (Mallows, 1972) for any sequence of probability measures Q_n and Q,

$$M_1(Q_n, Q) \longrightarrow 0 \text{ if and only if } Q_n \xrightarrow{w} Q \text{ and } \int \|x\| \, d\,Q_n \longrightarrow \int \|x\| \, d\,Q, \tag{2.4}$$

where $\xrightarrow{w}$ denotes weak convergence. More detailed information about the first moment Mallows distance can be found in Villani (2008).

The bounded Lipschitz distance metrizes the weak convergence of probability measures

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

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

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

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 2.2.1. *Let the sequence $\{P_n\}$ and P be distributions on $\mathbb{R}^d$ with finite first moment. Then* *(a) If* $\lim_{n\to\infty} D_{BL}(P_n, P) = 0$*, then* $\limsup_{n\to\infty} L(P_n) \leqslant L(P)$*.*

(b) If $\lim_{n\to\infty} M_1(P_n, P) = 0$*, then* $\lim_{n\to\infty} L(P_n) = L(P)$*.*

(c) If $\lim_{n\to\infty} M_1(P_n, P) = 0$*, then* $\Pi(P_n)$ *converges to* $\Pi(P)$ *in* L^1*.*

Remark 1. For our results, Lemma 2.2.1 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 $\mathcal{P}$. Let $\{X_i\}_{i=1}^n$ be an observed sequence of $\{X_t\}$. Then it follows from (2.4) that the empirical distribution $\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ converges to P in the first moment Mallow's distance almost surely. As a result, the nonparametric log-concave maximum likelihood estimator $\hat{f}_n$

$$\hat{f}_n := \Pi(\mathbb{P}_n) = \arg\max_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \log f(X_i) \tag{2.5}$$

is well defined for large n with probability one and

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

Lemma 2.2.2 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 2.2.2. *Let* f_n *be a sequence of log-concave densities on* $\mathbb{R}^d$ *and* f *be some density function on* $\mathbb{R}^d$ *such that* $F_n \overset{D}{\Rightarrow} F$ *where* (F_n, F) *are the associated distributions 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 $f(x) \leqslant e^{-a_0\|x\|+b_0}$. Then for every $a < a_0$, we have $\int_{\mathbb{R}^d} e^{a\|x\|}|f_n(x) - f(x)|\,dx \to 0$. Furthermore, if f is continuous,

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

Lemma 2.2.2 further implies that the nonparametric log-concave density estimator $\hat{f}_n$ defined in (2.5) converges to the log-concave projection $\Pi(P)$ in a stronger exponential weighting norm. In particular, for the univariate case, i.e., $d = 1$, the log level maximum likelihood density estimator $\hat{\varphi}_n := \log \hat{f}_n$ is shown to be a piecewise linear function with knots at the observations $\{X_i\}_{i=1}^n$ and is zero outside the interval $\left[\min_{i=1,\cdots,n} X_i, \max_{i=1,\cdots,n} X_i\right]$. It is not differentiable at the sample points $\{X_i\}_{i=1}^n$. As a substitute for $\hat{f}_n$, a smoothed log concave density estimator $\hat{f}_{\sigma_n}$, the convolution of $\hat{f}_n$ with a zero mean, σ_n^2 variance normal density, is proposed in Chen and Samworth (2013). Detailed construction of $\hat{f}_{\sigma_n}$ including the choice of σ_n can be found in Chen and Samworth (2013). See also Dümbgen and Rufibach (2010) for a detailed description of the usage of the R package **logcondens** to compute $\hat{f}_n$ and the smoothed version $\hat{f}_{\sigma_n}$ in the univariate case.

2.3 Model specification

Denote ϕ and θ as the AR and MA parameter vectors $(\phi_1, \cdots, \phi_p) \in \mathbb{R}^p$ and $(\theta_1, \cdots, \theta_q) \in \mathbb{R}^q$ 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 $\beta(z) \coloneqq \theta^{-1}(z)\phi(z)$ 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 $|i| \to \infty$. Although Z_t is unobserved, it's expressible in terms of β_0 and $\{X_t\}$. Rearranging (1.1), 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

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

Since the filter weights $a_i(\beta)$ are absolutely summable, it's easy to see that $Z_t(\beta)$ is stationary and ergodic; see Brockwell and Davis (2009). We define a convergent representation of $Z_t(\beta)$ as introduced in Lii and Rosenblatt (1996):

$$Z_{t,m(n)}(\beta) = \sum_{|i| \leqslant m(n)} a_i(\beta)X_{t-i},$$

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

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

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. Let P_β denote the stationary distribution of $Z_t(\beta)$. Recall β_0 denote the true parameters. So P_{β_0} is the true disribution P_0. We have the following convergence results for $\mathbb{P}_{\beta,n}$ and $\tilde{\mathbb{P}}_{\beta,n}$.

Proposition 2.3.1. *Suppose that β_0 is an interior point in the compact parameter space Θ and $P_0 \in \mathcal{P}$. Then,*

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

It follows that

$$\sup_{\beta\in\Theta} M_1(\mathbb{P}_{\beta,n}, P_\beta) \xrightarrow{a.s.} 0 \quad \textit{as} \quad n \to \infty. \tag{2.6}$$

Proposition 2.3.1 indicates that the truncated residuals are asymptotically equivalent to the untruncated version in the first moment Mallow's distance. Lii and Rosenblatt (1996) derived the following approximations to the log-likelihood function of β, f given the observations $\{X_i\}_{i=1}^n$:

$$h_{\beta,f}^n := \frac{1}{n-2m(n)} \sum_{i=m(n)+1}^{n-m(n)} l_{\beta,f}\left(Z_{i,m(n)}(\beta)\right) = \int l_{\beta,f} d\mathbb{P}_{\beta,n}, \tag{2.7}$$

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 noninvertiable 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 (2.7) 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 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:

$$h_n(\beta) = \max_{f \in \mathcal{F}} h^n_{\beta,f} = L(\mathbb{P}_{\beta,n}) + \log \kappa(\beta) \text{ for } \beta \in \Theta. \tag{2.8}$$

Theorem 2.3.2. *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 $h^n_{\beta,f}$ over $\Theta \times \mathcal{F}$*

Proof. Note that $\beta \to \mathbb{P}_{\beta,n}$ defines a continuous mapping from Θ to the space of probability measures $\mathcal{P}$ 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 $\left(\hat{\beta}, \hat{f} := \Pi(\mathbb{P}_{\hat{\beta}})\right)$ maximizes $h_n(\beta, f)$ over $\Theta \times \mathcal{F}$. □

The joint maximizer $(\hat{\beta}, \hat{f})$ is referred to as the maximum log-concave likelihood estimator (MLCLE). In Section 2.4, we will show $\hat{\beta}$ is strongly consistent.

2.4 Asymptotic results

2.4.1 Consistency

For causal-invertible ARMA models, $\kappa(\beta)$ is identically equal to one. Thus (2.7) reduces to the conditional log-likelihood of the sequence $\{X_i\}_{i=1}^n$. The maximizer $(\hat{\beta}, \hat{f})$ of (2.7) is exactly the estimator proposed in Chen and Samworth (2015b), where consistency results were established. We now turn to the general case of noncausal/noninvertible models. The main result is:

Theorem 2.4.1. *In (1.1), suppose Z_t satisfies the following condition,*

$$L\left(\sum_{k=-\infty}^{\infty} d_k Z_{t-k}\right) \leqslant L(Z_t), \tag{2.9}$$

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

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

Remark 2. In the causal-invertible case, Chen and Samworth (2015b) did not require condition (2.9) because they excluded noncausal/noninvertible models. So if one expands the family of models to be noncausal/noninvertible, then a condition like (2.9) is required even if the true model is causal-invertible.

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

Proposition 2.4.2. *For causal-invertible ARMA models, assume that $P_0 \in \mathcal{P}$ and the parameter space Θ is compact, then*

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

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 (2.2) of the $L(\cdot)$ function. In short, under causality and invertibility, Z_t is independent of $Z_t(\beta) - Z_t$. Therefore,

$$L\left(Z_t(\beta)\right) = L\left(Z_t + Z_t(\beta) - Z_t\right) \leqslant L(Z_t),$$

implying that β_0 is a global maximizer of $L\left(Z_t(\beta)\right)$ 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 noncausal/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 2.4.1: From Remark 1, 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 2.3.2, the limiting function is continuous in β. Define

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

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

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

Furthermore,

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

since the distribution of $Z_t(\beta_n)$ converges in the first moment Mallows distance to the distribution of $Z_t(\beta^\star)$. Then according to Lemma 2.2.1,

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

As a result,

$$|L(\mathbb{P}_{\beta^n,n}) - L(P_{\beta^n})| \leqslant |L(\mathbb{P}_{\beta^n,n}) - L(P_{\beta^*})| + |L(P_{\beta^n}) - L(P_{\beta^*})| \to 0,$$

for the fixed $\omega \in \Omega$. Since $\{\beta^n\}$ is arbitrary and Θ is compact, we have

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

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

(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 $d(\beta)$:

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

The last equality is due to affine equivariance property; see equation (2.1). According to Proposition 2.6.2, $\left(\frac{\kappa(\beta_0)}{\kappa(\beta)}\right)^2 \sum_{k=-\infty}^{\infty} a_k^2 \geqslant 1$. Then by condition (2.9), $d(\beta) \geqslant 0$ 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 (2.9) holds, then we know there is only one a_k being non-zero and the coefficients must satisfy

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

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 $L(P_\beta) + \kappa(\beta)$.

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,

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

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

Verification of (2.9) has to be checked on a case-by-case basis. We show that (2.9) is true for log-concave distributions and symmetric α stable distributions with $\alpha \in (1, 2)$.

Corollary 2.4.3. *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 (2.9) is true for any non-Gaussian log-concave distribution. For completeness, this inequality is stated in Lemma 2.6.1.

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 $\sum_k d_k^2 \geqslant 1$, let $Y_j = \sum_{|k| \leqslant j} d_k Z_{t-k}$. 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

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

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 $d_k \neq 0$,

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

and hence

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

Then,

$$L(Y_j) < L(Z_t) - \frac{1}{2}\log\left(\sum_{|k|\leqslant j} d_k^2\right). \tag{2.10}$$

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 $j \to \infty$. Thus we can let j goes to infinity in (2.10) and obtain

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

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

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

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,

$$\begin{aligned}
\exp\{2H(Y)\} &> \exp\{2H(Y^1)\} + \exp\{2H(Y^2)\} \\
&\geqslant \sum_{k \in J_1, |k| \leqslant N} \exp\{2H\left(d_k Z_{t-k}\right)\} + \sum_{k \in J_2, |k| \leqslant N} \exp\{2H\left(d_k Z_{t-k}\right)\} \\
&= \exp\{2H(Z_t)\} \sum_{k \in J_1, |k| \leqslant N} d_k^2 + \exp\{2H(Z_t)\} \sum_{k \in J_2, |k| \leqslant N} d_k^2,
\end{aligned}$$

where N is some large integer. The first strict inequality is due to the non-Gaussianity of Y^1 and Y^2. Now by letting $N \to \infty$, we obtain

$$\exp\{2H(Y)\} > \exp\{2H(Z_t)\} \sum_k d_k^2.$$

It follows that

$$H(Y) \quad > \quad H(Z_t) + \frac{1}{2} \log \sum_k d_k^2 = H(Z_t),$$

since $\sum_k d_k^2 = 1$. As Y is the weak limit of the log-concave distributed sequence $Y_j = \sum_{|k| \leqslant j} d_k Z_{t-k}$, Y has a log-concave distribution, indicating $L(Y) = -H(Y)$. We deduce that

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

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 (2.9). □

Remark 3. Even under misspecification of log-concavity, the MLCLE may still be consistent in cases the true distribution P_0 is close to it's log-concave projection $\Pi(P_0)$ and preserves the property (2.9). Simulation results suggests that $\hat{\beta}$ is still consistent given Z_t follows the non log-concave student-t distribution; however, this has not been proved.

Corollary 2.4.4. *If Z_t is symmetric-α-stable with exponent $\alpha \in (1,2)$, then*

$$\hat{\beta} \xrightarrow{a.s.} 0 \textit{ and } \int |\hat{f}_n - \Pi(P_0)|\, d\, x \xrightarrow{a.s.} 0. \quad as \quad n \to \infty.$$

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

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

Therefore,

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

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 (2.9) is satisfied. □

2.4.2 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 $(\log f_0)''$ is bounded

A_3 $\int z\dot{f}_0 dz = zf_0(z)|_{-\infty}^{\infty} - \int f_0(z)dz = -1$ and $\int zf_0(z)dz = 0$

A_4 f_0 is log-concave and non-Gaussian

Following the ideas in Chapter 7 of van der Vaart (2002), we construct a semiparametric 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 $(X_t, X_{t-1}, \cdots, X_{t-p})^T$. 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 $t = p+1, \cdots, n$. 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

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

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

$$l_{\beta,f}(Z_\beta) = \log f(Z_\beta) + \log \kappa(\beta) \quad (\beta, f) \in \Theta \times \mathcal{F}. \tag{2.11}$$

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:

$$\dot{l}_{\beta,f} = \frac{\dot{f}(Z_\beta)}{f(Z_\beta)}\dot{Z}_\beta + \frac{\dot{\kappa}(\beta)}{\kappa(\beta)} \tag{2.12}$$

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

$$\mathbb{E}_{\beta,f}\dot{l}_{\beta,f} = \mathbb{E}_{\beta,f}\left(\frac{\dot{f}(Z_\beta)}{f(Z_\beta)}\dot{Z}_\beta + \frac{\dot{\kappa}(\beta)}{\kappa(\beta)}\right) = 0 \tag{2.13}$$

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.

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

$$\beta_s = \beta + sa, \quad f_s = (1 + sg)f, \tag{2.14}$$

where $a \in \mathbb{R}^p$ and g is a bounded continuous function which satisfies the constraint $\int_{\mathbb{R}} g(x)f(x)\,d\,x = 0$. 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 (2.14)

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

The information of this submodel is defined as

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

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 $g^*(u) := -a^T E_{\beta,f}\left[\dot{l}_{\beta,f} \mid Z_\beta = u\right]$, 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

$$\tilde{l}_{\beta,f} = \frac{\dot{f}(Z_\beta)}{f(Z_\beta)}\left(\dot{Z}_\beta - E_{\beta,f}\left[\dot{Z}_\beta \mid Z_\beta\right]\right) \tag{2.15}$$

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

Proposition 2.4.5. *Replacing f with f_0, we have $\mathbb{E}_{\beta,f_0}\left[\dot{Z}_t(\beta) \mid Z_t(\beta)\right] = \frac{\dot{\kappa}(\beta)}{\kappa(\beta)}Z_t(\beta)$. And* *hence,*

$$\tilde{l}_{\beta,f_0} = \frac{\dot{f}_0\left(Z_t(\beta)\right)}{f_0\left(Z_t(\beta)\right)}\left[\dot{Z}_t(\beta) - \frac{\dot{\kappa}(\beta)}{\kappa(\beta)}Z_t(\beta)\right]. \tag{2.16}$$

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

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

which together with (2.13) implies that $a_{\beta,0} = \frac{\dot{\kappa}(\beta)}{\kappa(\beta)}$. Therefore,

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

□

Remark 4. 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

$$\mathbb{E}_{\beta,f_0}\tilde{l}_{\beta,f} = \mathbb{E}_{\beta,f_0}\left(\varphi'\left(Z(\beta)\right)\left(\dot{Z}(\beta) - \frac{\dot{\kappa}(\beta)}{\kappa(\beta)}Z(\beta)\right)\right) = 0, \tag{2.17}$$

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 $\hat{f}_{\sigma_n}$ is the smoothed log-concave density estimator based on convolving $\hat{f}_n$ with a normal density with mean zero and variance σ_n^2. 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

$$\check{\beta} := \tilde{\beta} - \left(\sum_{i=p+1}^{n} \tilde{l}_{\tilde{\beta},\hat{f}_{\sigma_n}}(\mathbf{X}_i)\tilde{l}^T_{\tilde{\beta},\hat{f}_{\sigma_n}}(\mathbf{X}_i)\right)^{-1} \sum_{i=p+1}^{n} \tilde{l}_{\tilde{\beta},\hat{f}_{\sigma_n}}(\mathbf{X}_i). \tag{2.18}$$

Theorem 2.4.6. *Suppose that f_0 satisfies the conditions $A_1 - A_4$, and the efficient information matrix $\tilde{I}_{\beta_0,f_0} = E\left(\tilde{l}_{\beta_0,f_0}\tilde{l}^T_{\beta_0,f_0}\right)$ is nonsingular. Then, $\check{\beta}$ is asymptotic efficient at (β_0, f_0) in the sense that*

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

Proof. The function $\tilde{l}_{\beta,\hat{f}_{\sigma_n}}$ is unbiased according to (2.17) and satisfies the integrablibility conditions stated in Proposition 2.6.4. Then the conclusion follows from Theorem 7.2 in van der Vaart (2002). □

Remark 5. In practice, we can iterate by replacing $\tilde{\beta}$ with the last update $\check{\beta}$ in equation (2.18). We suspect that the MLCLE $\hat{\beta}$ is semiparametric efficient, although this is not yet proved. Simulation results for investigating its asymptotic behavior are included in Section 2.5.

For illustration, we compute $\tilde{I}_{\beta_0,f_0}$ explicitly for the noncausal AR(1) models. In this case, $\dot{Z}_t(\beta) = -X_{t-1}$. Since $E\left[\dot{Z}_t(\beta) \mid Z_t\right] = \frac{1}{\beta} Z_t$, we have

$$\begin{aligned}\tilde{l}_{\beta_0,f_0} &= \varphi'(Z_t)\left(\dot{Z}_t(\beta) - E\left[\dot{Z}_t(\beta) \mid Z_t\right]\right) = \varphi'(Z_t)\left(-X_{t-1} - \frac{1}{\beta} Z_t\right) \\ &= -\frac{1}{\beta} X_t \varphi'(Z_t).\end{aligned}$$

Therefore, $\tilde{I}_{\beta_0,f_0} = E(\tilde{l}^2_{\beta_0,f_0}) = \frac{1}{\beta^2} E X_t^2 E\varphi'(Z_t)^2 = \frac{\sigma^2}{\beta^2(\beta^2-1)} E\varphi'(Z_t)^2$, where σ^2 is the variance of Z_t. See examples below for calculating the inverse efficient information for the noncausal AR(1) model: $X_t - 2X_{t-1} = Z_t$ driven by Laplace and logistic distributed noise, respectively.

1. Laplace distribution $f(x) = \frac{1}{2}\exp(-|x|)$: $E\varphi'(Z_t)^2 = 1$, $\sigma^2 = 2$

   $\tilde{I}_{\beta_0,f_0} = \frac{1}{6}$, $\frac{1}{\sqrt{\tilde{I}_{\beta_0,f_0}}} = \sqrt{6} \approx 2.45$

2. Logistic distribution $f(x) = \frac{e^{-x}}{(1+e^{-x})^2}$: $E\varphi'(Z_t)^2 = \frac{1}{3}$, $\sigma^2 = \frac{\pi^2}{3}$

   $\tilde{I}_{\beta_0,f_0} = \frac{\pi^2}{108}$, $\frac{1}{\sqrt{\tilde{I}_{\beta_0,f_0}}} = \frac{6}{\pi}\sqrt{3} \approx 3.31$

2.5 Examples

2.5.1 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. The R package **logcondens** (Dümbgen and Rufibach, 2010) is used to compute the log-concave density MLE. We considered a mixed AR(2) process and a ARMA(1,1) process from a symmetric $\alpha-$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 $(1.2, 0.6)$ so that the AR roots are 0.63 and -2.63. As shown in Table 2.1, 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 2.1 are seen in Table 2.2.

$Z_t \sim S\alpha S$

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 (0.0110)	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)	0.6010 (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 2.1: Mean and root-mean-squared error $(\cdot)$ of 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, 1000, 5000, 10000 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 2.3, 2.4 and 2.5 reports the mean, the root-mean-squared error and the normalized empirical standard error of each method given different noise distribu-

$Z_t \sim S\alpha S$

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

tions. Note that the LAD coincides with MLE for the Laplace distribution. The conjectured asymptotic variance of the MLCLE, σ^2 (the inverse efficient fisher information), under each setting is contained in Table 2.6. The MLCLE and MLE estimates are comparable for the three log-concave distributions. As the sample sizes grows, the normalized empirical variance $\hat{\sigma}^2$ by the MLCLE approaches the inverse efficient information. For normal and 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]
1000	2.0095 (0.1070) [3.3824]	2.0134 (0.1097) [3.4666]	2.0102 (0.1234) [3.8989]
5000	2.0025 (0.0473) [3.3410]	2.0032 (0.0473) [3.3473]	2.0023 (0.0545) [3.8557]
10000	2.0013 (0.0330) [3.2997]	2.0012 (0.0334) [3.3399]	2.0014 (0.0383) [3.8324]

Table 2.3: Mean, root-mean-squared error $(\cdot)$ and normalized empirical standard error $[\cdot]$ of MLE, MLCLE and LAD estimates for non-causal AR(1) model

$Z_t \sim N(0,1)$, $\phi = 0.5$

	MLE	MLCLE	LAD
100	0.4908 (0.0875) [0.8710]	0.4849 (0.0988) [0.9826]	0.4893 (0.1100) [1.0940]
500	0.4980 (0.0386) [0.8615]	0.4948 (0.0426) [0.9519]	0.4979 (0.0480) [1.0765]
1000	0.4990 (0.0271) [0.8562]	0.4970 (0.0291) [0.9186]	0.4991 (0.0340) [1.0757]
5000	0.4999 (0.0123) [0.8696]	0.4995 (0.0123) [0.8706]	0.4999 (0.0152) [1.0733]
10000	0.4999 (0.0087) [0.8713]	0.4997 (0.0087) [0.8743]	0.4998 (0.0108) [1.0805]

Table 2.4: Mean, root-mean-squared error $(\cdot)$ and normalized root-mean-squared error $[\cdot]$ of MLE, MLCLE and LAD estimates for 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]
1000	2.0057 (0.0819) [2.5874]	2.0097 (0.0851) [2.6908]
5000	2.0012 (0.0352) [2.4900]	2.0019 (0.0356) [2.5194]
10000	2.0000 (0.0248) [2.4825]	2.0005 (0.0249) [2.4912]

Table 2.5: Mean, root-mean-squared error $(\cdot)$ and normalized empirical standard error $[\cdot]$ of MLE and MLCLE estimates for non-causal AR(1) model

	Logistic(0,1)	N(0,1)	Laplace(1)
$\tilde{I}^{-1}_{\beta_0,f_0}$	$\frac{108}{\pi^2} \approx 10.94$	$\frac{3}{4} = 0.75$	6

Table 2.6: Inverse efficient information of AR(1) process with AR coefficient equal to two

2.5.2 An empirical example

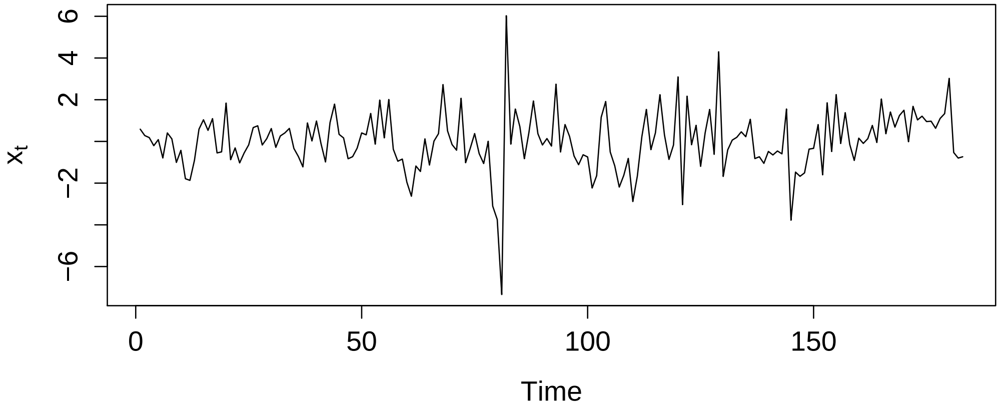

Figure 2.1: The demeaned differences of U.S. Total Government Revenue

Figure 2.1 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 2.2 suggest fitting an AR(2) model to this data. Table 2.7 compare the log-likelihood function values of the best fit of causal Gaussian (CG) AR(2), causal non-Gaussian (CN) AR(2) and the mixed (MX) AR(2).

Model assumptions	CG	CN	MX
Log-likelihood	-318.6452	-300.4979	-296.4701

Table 2.7: Comparison of log-likelihood

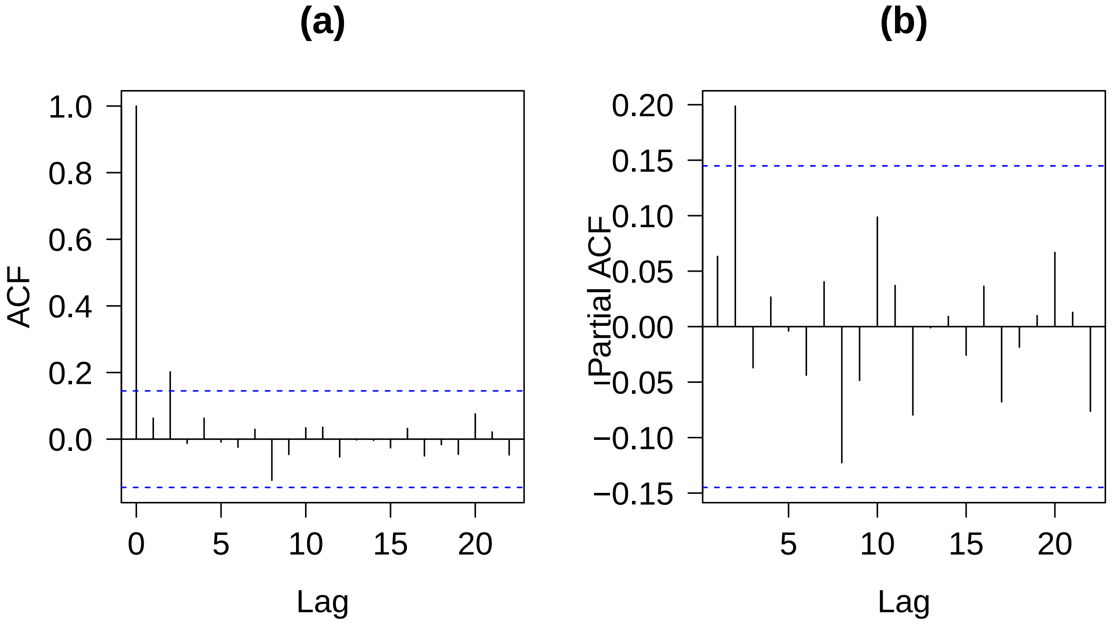

Figure 2.2: (a) Sample ACF of x_t, and (b) Sample PACF of x_t

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 2.5.2 indicate that $\hat{W}_t$ is white noise, the ACF of the absolute values of the residuals $\{|\hat{W}_t|\}$ and those 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

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

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. Figure 2.5.2 plots the residuals $\{\hat{Z}_t\}$, the ACF of $\{\hat{Z}_t\}$, the ACF of $\{|\hat{Z}_t|\}$ and those 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 $\{|\hat{Z}_t|\}$ and those of $\{\hat{Z}_t^2\}$. Therefore, the noncausal model products residuals that look more independent at least in terms of the squares of the residuals.

Another benefit of the MLCLE estimation is that we can estimate the noise distribution as well. Figure 2.5 contains the estimated density $\hat{f}_n$ based on $\{\hat{Z}_t\}$ which is skewed to the right. Also plotted is the estimated density based on the residuals from fitting a Gaussian AR(2) model. From an interpretation perspective, noncausal models may be difficult to accept since such models imply that *shocks* depend on the future. However, an alternative explanation may be that the past, only defined in terms of the time series, is not a rich enough information set for modeling X_t. That is, the information set should perhaps include exogenous variables, such as news or auxiliary time series in order to produce a *causal* model.

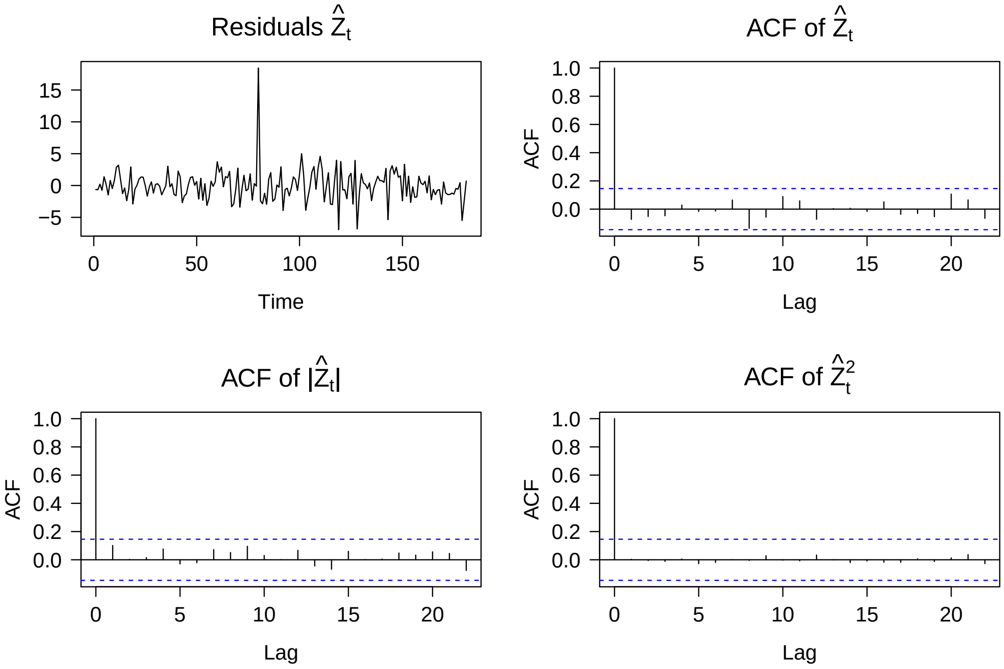

Figure 2.3: AR model-fitting using the MLCLE method

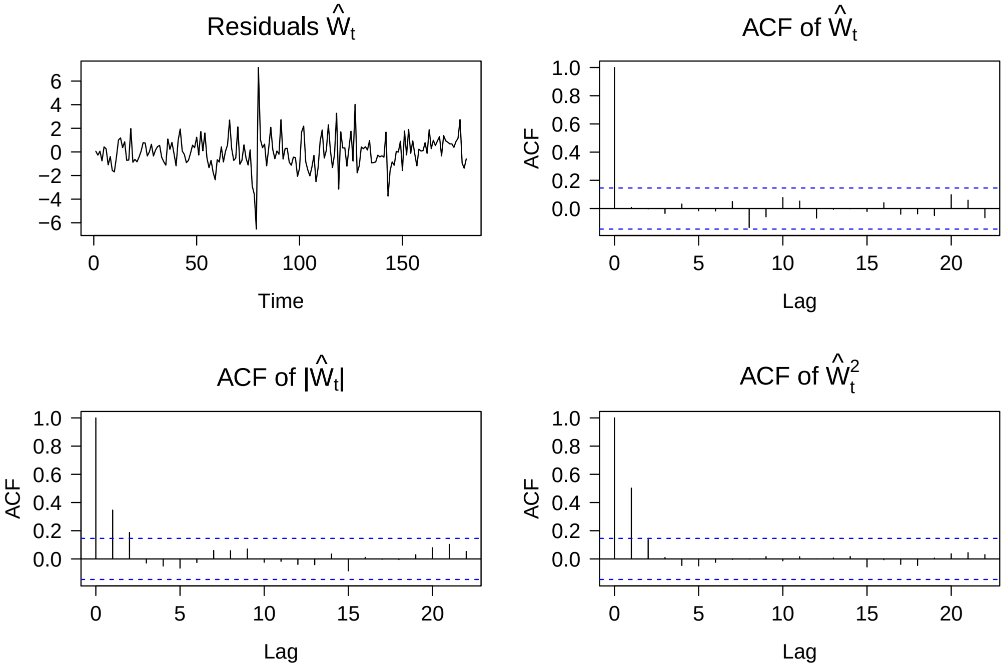

Figure 2.4: Causal AR model using Least Square

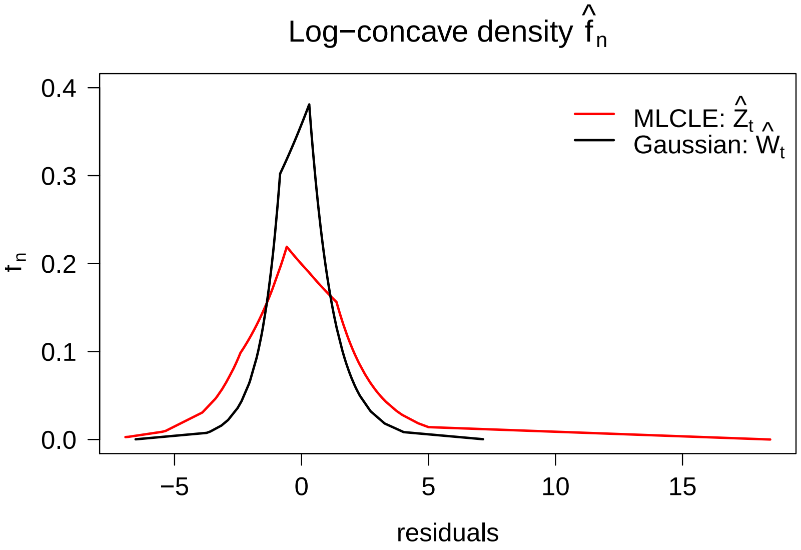

Figure 2.5: Estimated log-concave densities from MLCLE and GL residuals

2.6 Appendix

2.6.1 Auxiliary results and proof

Lemma 2.6.1. *(Entropy Power Inequality)*

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

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 only X and Y are normal random variables.

Proof of Proposition 2.3.1: 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 $|a_j(\beta)| \leqslant K\rho^{|j|}$ for all j (see Brockwell and Davis (2009)).

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

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

$$EW_n == \sum_{|j|>m(n)} K\rho^{|j|}E|Z_i| = 2KE(|Z_i|)\frac{\rho^{m(n)}}{1-\rho},$$

indicating that W_n converges to 0 almost surely by the 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 $\beta, \beta' \in \Theta$,

$$M_1(\tilde{\mathbb{P}}_{\beta,n}, \tilde{\mathbb{P}}_{\beta',n}) = \sup_{\|g\|_L \leqslant 1} \left(|\int g\, d\tilde{\mathbb{P}}_{\beta,n} - \int g\, d\tilde{\mathbb{P}}_{\beta',n}| \right)$$

$$\begin{aligned}
&\leqslant \frac{1}{n-2m(n)} \sum_{i=m(n)+1}^{n-m(n)} |Z_i(\beta) - Z_i(\beta')| \\
&\leqslant \frac{1}{n-2m(n)} \sum_{i=m(n)+1}^{n-m(n)} \sum_{j=-\infty}^{\infty} |a_j(\beta) - a_j(\beta'))||Z_{i-j}| \\
&\leqslant \frac{1}{n-2m(n)} \sum_{i=m(n)+1}^{n-m(n)} \sum_{|j|\leqslant M} |a_j(\beta) - a_j(\beta')||Z_{i-j}| + \frac{1}{n-2m(n)} \sum_{i=m(n)+1}^{n-m(n)} \sum_{|j|>M} 2K\rho^{|j|}|Z_{i-j}| \\
&\leqslant \frac{\max_{|j|\leqslant M}|a_j(\beta) - a_j(\beta')|}{n-2m(n)} \sum_{i=m(n)+1}^{n-m(n)} \sum_{|j|\leqslant M} |Z_{i-j}| + \frac{1}{n-2m(n)} \sum_{i=m(n)+1}^{n-m(n)} \sum_{|j|>M} 2K\rho^{|j|}|Z_{i-j}|.
\end{aligned}$$

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,

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

converges almost surely to some constant and one can show that

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

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

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

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

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

It follows that

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

□

Proposition 2.6.2. *The coefficients a_k of the Laurent expansion of $\beta(z)\beta_0^{-1}(z)$ satisfies the inequality*

$$\left(\frac{\kappa(\beta_0)}{\kappa(\beta)}\right)^2\sum_{k=-\infty}^{\infty}a_k^2\geqslant 1. \tag{2.19}$$

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

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

which implies that

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

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

Proposition 2.6.3. *Assume that f_0 is continuously differentiable.Then for any compact set $S \subseteq \mathbb{R}$,*

$$\lim_{n\to\infty} \sup_{x\in S} |\hat{\varphi}_{\sigma_n}(x) - \varphi_0(x)| = 0 \ \textit{a.s. and} \ \lim_{n\to\infty} \sup_{x:\in S} |\hat{\varphi}'_{\sigma_n}(x) - \varphi'_0(x)| = 0 \, a.s..$$

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

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

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

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

since $\hat{f}_{\sigma_n}$ and f_0 are supported on the real line. Note that $\hat{\varphi}_{\sigma_n}$ and φ_0 are continuous concave functions. Thus, $\hat{\varphi}'_{\sigma}(x)$ 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,

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

□

Proposition 2.6.4. *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.*

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

Proof of Proposition 2.6.4: Let μ_n be the mode of $\hat{f}_{\sigma_n}$, then $\hat{\varphi}'_{\sigma_n} \geqslant 0$ for $x \leqslant \mu_n$ and $\hat{\varphi}'_{\sigma_n} \leqslant 0$ for $x \geqslant \mu_n$. It follows that

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

where μ is the mode of f_0. Then, by following the same argument as Lemma 3 in Cule and Samworth (2010), $|\hat{\varphi}'_{\sigma_n}|\hat{f}^{\frac{1}{3}}_{\sigma_n}$ is uniformly bounded with probability one. Besides, there exists some $c > 0$ such that $\hat{f}_{\sigma_n} \geqslant cf_0$ 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 2.6.3 implies $\tilde{l}_{\beta_n,\hat{f}_{\sigma_n}}$ converges to $\tilde{l}_{\beta_0,f_0}$ almost surely. Then the results follow from the dominated convergence theorem. □

2.6.2 Comments on asymptotic properties of MLCLE $\hat{\beta}$ for AR models

In this subsection, we present the current progress in studying the asymptotic properties of MLCLE $\hat{\beta}$ for AR models. We consider a path that is similar to the one used in Section 6.2 of Murphy et al. (1999) to find the least favorable submodel with score being a good approximation of the efficient score function. Define $\Psi := \{\varphi \mid \varphi \text{ is concave and } \int e^{\varphi(u)}\, d\,u < \infty\}$ and rewrite the objective function (2.7) in terms of log density $\varphi = \log f$, i.e.,

$$\begin{aligned} h^n_{\beta,\varphi} &= \frac{1}{n-p}\sum_{i=p+1}^{n} \varphi\left(Z_i(\beta)\right) - \int e^{\varphi(u)}\, d\,u + \log \kappa(\beta) \\ &= \int \varphi\, d\,\mathbb{P}_{\beta,n} - \int e^{\varphi(u)}\, d\,u + \log \kappa(\beta) \quad (\beta, \phi) \in \Theta \times \Psi. \end{aligned} \tag{2.20}$$

In fact, maximizing (2.7) over $\Theta \times \mathcal{F}$ is equivalent to maximizing (2.20) over $\Theta \times \Psi$ by reparametrizing f as e^{φ}. Although a general concave function $\varphi \in \Psi$ is not necessarily differentiable, there exists a right continuous (or left continuous) non-increasing function φ' that satisfies

$$\varphi(b) = \varphi(a) + \int_a^b \varphi'(u)\, d\,u, \text{ when } b > a.$$

For any $\varphi \in \Psi$ with $\varphi(0)$ well-defined and $\gamma \in \Theta$ such that $\|\beta - \gamma\|$ is sufficiently close to zero, we define a path $\{\gamma, \xi_{\beta,\varphi}(\gamma)\}$ such that $\xi_{\beta,\varphi}(\beta) = \varphi$:

$$\xi_{\beta,\varphi}(\gamma)(u) := \int_0^u \varphi'\left(y + (\beta-\gamma)^T \frac{\dot\kappa(\beta)}{\kappa(\beta)} y\right)\, d\,y + (\beta-\gamma)^T \left[\frac{\dot\kappa(\beta)}{\kappa(\beta)} - \frac{\dot\kappa(\beta)}{\kappa(\beta)}\varphi(0)\right] + \varphi(0).$$

Since φ' is non-increasing and $1 + (\beta-\gamma)^T \frac{\dot\kappa(\beta)}{\kappa(\beta)}$ is positive when $\|\beta - \gamma\|$ is small enough, $\xi_{\beta,\varphi}(\gamma)$ is a concave function in u. In particular, $\{\gamma, \xi_{\beta,\varphi}(\gamma)\}$ is a well-defined concave function at the MLCLE $(\hat\beta, \hat\varphi := \hat f)$ since the support of the density of Z_t is assumed to be $\mathbb{R}$, thus $\hat\varphi(0)$ is finite for n large. Similar to the efficient score function $\tilde{l}_{\beta,f_0}$, for fixed $(\beta, \varphi) \in \Theta \times \Phi$, we define

$$\psi_{\beta,\varphi}\left(Z(\beta)\right) := \varphi'\left(Z(\beta)\right)\left[\dot Z(\beta) - \frac{\dot\kappa(\beta)}{\kappa(\beta)} Z(\beta)\right].$$

Remark 6. Taking $\varphi_0 = \log f_0$, we have $\psi_{\beta,\varphi_0} = \tilde{l}_{\beta,f_0}$. By equation (2.17), the function $\psi_{\beta,\varphi}$

satisfies

$$\mathbb{E}_{\beta,f_0}\psi_{\beta,\varphi} = \mathbb{E}_{\beta,f_0}\left(\varphi'\left(Z(\beta)\right)\left(\dot{Z}(\beta) - \frac{\dot{\kappa}(\beta)}{\kappa(\beta)}Z(\beta)\right)\right) = 0. \tag{2.21}$$

Define $\mathbb{P}_{\beta,n}\psi_{\beta,\varphi} = \frac{1}{n-p}\sum_{i=p+1}^{n}\varphi'\left(Z_i(\beta)\right)\left[\dot{Z}_i(\beta) - -\frac{\dot{\kappa}(\beta)}{\kappa(\beta)}Z(\beta)\right]$. Recall that the smoothed log-concave density estimator $\hat{f}_{\sigma_n}$ is the convolution of $\hat{f}$ with $N(0,\sigma_n^2)$. Let $\hat{\varphi}_{\sigma_n} := \log \hat{f}_{\sigma_n}$. We show $\hat{\beta}$ satisfies the equation (2.22)

$$\sqrt{n}\mathbb{P}_{\hat{\beta},n}\psi_{\hat{\beta},\hat{\varphi}_{\sigma_n}} = o_p(1) \tag{2.22}$$

in Proposition 2.6.5. The function $\psi_{\beta,\varphi}\left(Z_t(\beta)\right)$ can be expressed in terms of the augmented process $\{\mathbf{X}_t\}$ as

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

and $\psi_{\beta,\varphi}$ can be viewed as a function from $\mathbb{R}^{p+1}$ to $\mathbb{R}^p$ indexed by (β,φ). Let $\mathcal{C}$ be the collection of the $\psi_{\beta,\varphi}$ functions given by

$$\mathcal{C} := \{\psi_{\beta,\varphi} : \|\beta - \beta_0\| \leqslant \epsilon_1, \mathbb{E}_{\beta_0,f_0}\|\psi_{\beta,\varphi} - \tilde{l}_{\beta_0,f_0}\|^2 \leqslant \epsilon_2, \varphi \in \Phi\}.$$

Proposition 2.6.4 implies $\psi_{\hat{\beta},\hat{\varphi}_{\sigma_n}} \in \mathcal{C}$ for n large. Although the function φ' is non-increasing and β is a finite dimensional vector, we are not able to show that the class $\mathcal{C}$ is a V-C subgraph class at this time. We save it for a future work and proceed as if we have this condition. See Remark 7 for the remaining proof of the asymptotic normality of $\hat{\beta}$.

Remark 7. Suppose that f_0 satisfies the conditions $A_1 - A_4$, the AR(p) process is β-mixing, the class $\mathcal{C}$ is a V-C subgraph class 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, $\sqrt{n}(\hat{\beta} - \beta_0)$ is asymptotically normal with mean 0 and covariance matrix given by the inverse of the efficient information matrix; that is

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

Proof. For any function $\psi_{\beta,\varphi} \in \mathbb{C}$, define $\mathbb{G}_n\psi_{\beta,\varphi} := \sqrt{n}\left(\mathbb{P}_{\beta,n} - P_{\beta_0,f_0}\right)\psi_{\beta,\varphi}$, where $P_{\beta_0,f_0}\psi_{\beta,\varphi} = \mathbb{E}_{\beta_0,f_0}\left(\varphi'(Z_\beta)\left(\dot{Z}_\beta - \frac{\dot{\kappa}(\beta)}{\kappa(\beta)}Z_\beta\right)\right)$. Since $\sqrt{n}\ \mathbb{P}_{\hat{\beta},n}\psi_{\hat{\beta},\hat{\varphi}_{\sigma_n}} = o_p(1)$ as shown in Proposition 2.6.5, we have

$$\mathbb{G}_n\psi_{\hat{\beta},\hat{\varphi}_{\sigma_n}} = -\sqrt{n}P_{\beta_0,f_0}\psi_{\hat{\beta},\hat{\varphi}_{\sigma_n}} + o_p(1). \tag{2.23}$$

Note that X_t is assumed to be β-mixing and the class $\mathcal{C}$ is a V-C subgraph class. By applying Theorem 2.1 in Arcones and Yu (1994), the empirical process

$$\{\mathbb{G}_n\psi_{\beta,\varphi} := \sqrt{n}\left(\mathbb{P}_{\beta,n} - P_{\beta_0,f_0}\right)\psi_{\beta,\varphi}, \psi_{\beta,\varphi} \in \mathcal{C}\}$$

converges in law to a centered, tight Gaussian process indexed by the class $\mathcal{C}$. So the empirical process $\{\mathbb{G}_n\psi_{\beta,\varphi}\}$ is asymptotically equicontinuous, and we have

$$\mathbb{G}_n\left(\psi_{\hat{\beta},\hat{\varphi}_{\sigma_n}} - \tilde{l}_{\beta_0,f_0}\right) = o_p(1). \tag{2.24}$$

Then, by equations (2.21) and (2.23), (2.24) implies

$$\sqrt{n}\left(P_{\hat{\beta},f_0} - P_{\beta_0,f_0}\right)\psi_{\hat{\beta},\hat{\varphi}_{\sigma_n}} = \mathbb{G}_n\tilde{l}_{\beta_0,f_0} + o_p(1). \tag{2.25}$$

To complete the theorem, it remains to show that

$$\sqrt{n}\left(P_{\hat{\beta},f_0} - P_{\beta_0,f_0}\right)\psi_{\hat{\beta},\hat{\varphi}_{\sigma_n}} = \tilde{I}_{\beta_0,f_0}\sqrt{n}(\hat{\beta} - \beta_0) + o_p\left(\sqrt{n}(\hat{\beta} - \beta_0)\right). \tag{2.26}$$

Equations (2.25) and (2.26) imply

$$\tilde{I}_{\beta_0,f_0}\sqrt{n}(\hat{\beta}-\beta_0)=\mathbb{G}_n\tilde{l}_{\beta_0,f_0}+o_p\left(1+\sqrt{n}(\hat{\beta}-\beta_0)\right).$$

The result now follows

$$\sqrt{n}(\hat{\beta}-\beta_0)=\tilde{I}^{-1}_{\beta_0,f_0}\mathbb{G}_n\tilde{l}_{\beta_0,f_0}+o_p(1)\xrightarrow{\mathcal{D}}N(0,\tilde{I}^{-1}_{\beta_0,f_0}).$$

Due to the square integrability of $\psi_{\hat{\beta},\hat{\varphi}_\sigma}$ shown in Proposition 2.6.4 and the fact that the likelihood function of $\mathbf{X}_t$ is differentiable in quadratic mean, the equation (2.26) can be established in the same way as in the proof of Theorem 6.20 in van der Vaart (2002). □

Proposition 2.6.5. *There exists a random sequence σ_n such that $\hat{\beta}$ satisfies the equation*

$$\mathbb{P}_{\hat{\beta},n}\psi_{\hat{\beta},\hat{\varphi}_{\sigma_n}}=o_p(\frac{1}{\sqrt{n}}).$$

Proof. First we calculate the partial derivative of $\xi_{\beta,\varphi}(\gamma)$ at β. Define $\tau_\gamma:=1+(\beta-\gamma)^T\frac{\dot{\kappa}(\beta)}{\kappa(\beta)}$, then

$$\begin{aligned}
\xi_{\beta,\varphi}(\gamma)(u) &= \frac{1}{\tau_\gamma}\int_0^{\tau_\gamma u}\varphi'(y)\,dy+(\beta-\gamma)^T\left[\frac{\dot{\kappa}(\beta)}{\kappa(\beta)}-\frac{\dot{\kappa}(\beta)}{\kappa(\beta)}\varphi(0)\right]+\varphi(0)\\
\frac{\partial\xi_{\beta,\varphi}(\gamma)(u)}{\partial\gamma}\,|_{\gamma=\beta} &= -\frac{\dot{\tau}_\gamma}{\tau_\gamma^2}\int_0^{\tau_\gamma u}\varphi'(y)\,dy+\frac{1}{\tau_\gamma}\varphi'(\tau_\gamma u)\dot{\tau}_\gamma u-\left[\frac{\dot{\kappa}(\beta)}{\kappa(\beta)}-\frac{\dot{\kappa}(\beta)}{\kappa(\beta)}\varphi(0)\right]|_{\gamma=\beta}\\
&= \frac{\dot{\kappa}(\beta)}{\kappa(\beta)}\int_0^{u}\varphi'(y)\,dy-\frac{\dot{\kappa}(\beta)}{\kappa(\beta)}\varphi'(u)u-\left[\frac{\dot{\kappa}(\beta)}{\kappa(\beta)}-\frac{\dot{\kappa}(\beta)}{\kappa(\beta)}\varphi(0)\right]\\
&= \frac{\dot{\kappa}(\beta)}{\kappa(\beta)}\varphi(u)-\frac{\dot{\kappa}(\beta)}{\kappa(\beta)}\varphi'(u)u-\frac{\dot{\kappa}(\beta)}{\kappa(\beta)}
\end{aligned}$$

The partial derivative of $\xi_{\beta,\varphi}(\gamma)\,(Z(\gamma))$:

$$\begin{aligned}
\frac{\partial \xi_{\beta,\varphi}(\gamma)\,(Z(\gamma))}{\partial \gamma}\,|_{\gamma=\beta} &= -\frac{\dot{\tau}_\gamma}{\tau_\gamma^2}\int_0^{\tau_\gamma Z(\gamma)} \varphi'(y)dy + \frac{1}{\tau_\gamma}\varphi'\left(\tau_\gamma Z(\gamma)\right)\left[\dot{\tau}_\gamma Z(\gamma) + \dot{Z}(\gamma)\tau_\gamma\right] - \frac{\dot{\kappa}(\beta)}{\kappa(\beta)}\left[1-\varphi(0)\right]|_{\gamma=\beta} \\
&= \frac{\dot{\kappa}(\beta)}{\kappa(\beta)}\int_0^{Z(\beta)} \varphi'(y)\,d\,y + \varphi'\left(Z(\beta)\right)\left[-\frac{\dot{\kappa}(\beta)}{\kappa(\beta)}Z(\beta) + \dot{Z}(\beta)\right] - \frac{\dot{\kappa}(\beta)}{\kappa(\beta)}\left[1-\varphi(0)\right] \\
&= \frac{\dot{\kappa}(\beta)}{\kappa(\beta)}\varphi\left(Z(\beta)\right) + \psi_{\beta,\varphi}\left(Z(\beta)\right) - \frac{\dot{\kappa}(\beta)}{\kappa(\beta)}
\end{aligned}$$

The partial derivatives of $\xi_{\beta,\varphi}\left(Z(\gamma)\right)$ and $\xi_{\beta,\varphi}(u)$ can be calculated in the same way as above at the point $(\hat{\beta}, \hat{\varphi} = \log \hat{f})$. As $\hat{\varphi}$ maximizes the function

$$\int \varphi\, d\,\mathbb{P}_{\hat{\beta},n} - \int e^{\varphi(y)}\, d\,y \text{ over } \Psi,$$

the derivative of the function $v\int \hat{\varphi}\, d\,\mathbb{P}_{\hat{\beta},n} - \int e^{v\hat{\varphi}(y)}\, d\,y$ is zero at $v = 1$, which implies

$$\int \hat{\varphi}\, d\,\mathbb{P}_{\hat{\beta},n} - \int \hat{\varphi}(y) e^{\hat{\varphi}(y)}\, d\,y = 0. \tag{2.27}$$

The function

$$h_n(\gamma) := \int \xi_{\hat{\beta},\hat{\varphi}}(\gamma)\, d\,\mathbb{P}_{\gamma,n} - \int \exp \xi_{\hat{\beta},\hat{\varphi}}(\gamma)(u)\, d\,u + \log \kappa(\gamma)$$

attains its maximum at $\hat{\beta}$. To get around the nondiffiability issue of $\hat{\varphi}$ on a most countable set, we consider the smoothed version of the log-concave density estimator $\hat{f}_\sigma$ defined in Proposition 2.6.3. Define $h_{n,\sigma}(\gamma)$ as a substitute of $h_n(\gamma)$

$$h_{n,\sigma}(\gamma) := \int \xi_{\hat{\beta},\hat{\varphi}_\sigma}(\gamma)\, d\,\mathbb{P}_{\gamma,n} - \int \exp \xi_{\hat{\beta},\hat{\varphi}_\sigma}(\gamma)(u)\, d\,u + \log \kappa(\gamma),$$

where $\hat{\varphi}$ is replaced by $\hat{\varphi}_\sigma$. It's easy to see that $h_{n,\sigma}(\gamma)$ converges uniformly to $h_n(\gamma)$ in

a neighborhood of $\hat{\beta}$ as σ approaches zero. Therefore, $\hat{\gamma}$, the maximizer of $h_{n,\sigma}(\gamma)$, can be arbitrarily close to $\hat{\beta}$ as σ goes to zero. And $\|\hat{\gamma} - \hat{\beta}\|$ only depends on σ for fixed n.
On the other hand, Proposition 2.6.3 further implies that the functions $\frac{\partial h_{n,\sigma}(\gamma)}{\partial \gamma}$ indexed by n and σ are uniformly Lipschitz in γ for large enough n and small enough σ. Therefore, there exists a sequence of σ_n approaches 0 such that the following conditions are satisfied:

$$\|\frac{\partial h_n(\gamma)}{\partial \gamma}|_{\gamma=\hat{\beta}} - \frac{\partial h_{n,\sigma_n}(\gamma)}{\partial \gamma}|_{\gamma=\hat{\gamma}}\| = o_p(\frac{1}{\sqrt{n}}) \tag{2.28}$$

$$\|\int \hat{\varphi}_{\sigma_n}\, d\mathbb{P}_{\hat{\beta},n} - \int \hat{\varphi}_{\sigma_n}(y)e^{\hat{\varphi}_{\sigma_n}(y)}\, dy\| = o_p(\frac{1}{\sqrt{n}}) \tag{2.29}$$

Equation (2.29) follows from (2.27) and the fact that

$$|\left[\int \hat{\varphi}_{\sigma}\, d\mathbb{P}_{\hat{\beta},n} - \int \hat{\varphi}_{\sigma}(y)e^{\hat{\varphi}_{\sigma}(y)}\, dy\right] - \left[\int \hat{\varphi}\, d\mathbb{P}_{\hat{\beta},n} - \int \hat{\varphi}(y)e^{\hat{\varphi}(y)}\, dy\right]| \to 0$$

as σ approaches 0 for fixed n.

Since $h_{n,\sigma_n}(\gamma)$ is locally maximized at $\hat{\gamma}$, (2.28) actually implies

$$\frac{\partial h_{n,\sigma_n}(\gamma)}{\partial \gamma}\,|_{\gamma=\hat{\beta}} \quad = \quad o(\frac{1}{\sqrt{n}}),$$

and

$$\frac{\partial h_{n,\sigma_n}(\gamma)}{\partial \gamma}\,|_{\gamma=\hat{\beta}} = \int \frac{\partial \xi_{\hat{\beta},\hat{\varphi}_{\sigma_n}}(\gamma)(Z(\gamma))}{\partial \gamma}\,|_{\gamma=\hat{\beta}}\, d\mathbb{P}_{\hat{\beta},n} - \int \exp \hat{\varphi}_{\sigma_n} \frac{\partial \xi_{\hat{\beta},\hat{\varphi}_{\sigma_n}}(\gamma)(u)}{\partial \gamma}\,|_{\gamma=\hat{\beta}}\, du + \frac{\dot{\kappa}(\hat{\beta})}{\kappa(\hat{\beta})}.$$

Substituting the partial derivatives of $\xi_{\hat{\beta},\hat{\varphi}_{\sigma_n}}$ into the above equation:

$$\begin{aligned}
\int \frac{\partial \xi_{\beta,\hat{\varphi}_{\sigma_n}}(\gamma)\,(Z(\gamma))}{\partial \gamma} \,|_{\gamma=\beta} \; d\,\mathbb{P}_{\hat{\beta},n} &= \frac{\dot{\kappa}(\hat{\beta})}{\kappa(\hat{\beta})} \int \hat{\varphi}_{\sigma_n} d\mathbb{P}_{\hat{\beta},n} + \int \psi_{\hat{\beta},\hat{\varphi}_{\sigma_n}} d\mathbb{P}_{\hat{\beta},n} - \frac{\dot{\kappa}(\hat{\beta})}{\kappa(\hat{\beta})} \\
\int \exp \hat{\varphi}_{\sigma_n} \frac{\partial \xi_{\hat{\beta},\hat{\varphi}_{\sigma_n}}(\gamma)(u)}{\partial \gamma} \,|_{\gamma=\hat{\beta}} \; d\,u &= \frac{\dot{\kappa}(\hat{\beta})}{\kappa(\hat{\beta})} \left(\int \hat{\varphi}_{\sigma_n}(u) e^{\hat{\varphi}_{\sigma_n}(u)}\, d\,u - \int u \hat{\varphi}'_{\sigma_n}(u) e^{\hat{\varphi}_{\sigma_n}(u)}\, d\,u - 1 \right) \\
&= \frac{\dot{\kappa}(\hat{\beta})}{\kappa(\hat{\beta})} \int \hat{\varphi}_{\sigma_n}(u) e^{\hat{\varphi}_{\sigma_n}(u)}\, d\,y
\end{aligned}$$

The last equality follows from the fact

$$\int u \hat{\varphi}'_{\sigma_n}(y) e^{\hat{\varphi}_{\sigma_n}(u)}\, d\,u \quad = \quad u e^{\hat{\varphi}_{\sigma_n}(u)} \,|_{-\infty}^{\infty} - \int e^{\hat{\varphi}_{\sigma_n}(u)}\, d\,u = -1$$

Thus,

$$\begin{aligned}
\frac{\partial h_{n,\sigma_n}(\gamma)}{\partial \gamma} \,|_{\gamma=\hat{\beta}} &= \int \psi_{\hat{\beta},\hat{\varphi}_{\sigma_n}} d\mathbb{P}_{\hat{\beta},n} + \frac{\dot{\kappa}(\hat{\beta})}{\kappa(\hat{\beta})} \left[\int \hat{\varphi}_{\sigma_n} d\mathbb{P}_{\hat{\beta},n} - \int \hat{\varphi}_{\sigma_n}(y) e^{\hat{\varphi}_{\sigma_n}(y)}\, d\,y \right] \\
&= o_p(\frac{1}{\sqrt{n}})
\end{aligned}$$

By (2.29), we have

$$\mathbb{P}_{\hat{\beta},n} \psi_{\hat{\beta},\hat{\varphi}_{\sigma_n}} = \int \psi_{\hat{\beta},\hat{\varphi}_{\sigma_n}} d\,\mathbb{P}_{\hat{\beta},n} = o_p(\frac{1}{\sqrt{n}}).$$

□

Chapter 3

Causal Vector Autoregression with Log-Concave Projection

3.1 Introduction

A multivariate time series $\mathbf{X}_t = (X_{t,1}, \cdots, X_{t,m})^T$ is a mean zero VAR(p) process if it is stationary and satisfies the difference equations,

$$\mathbf{X}_t - \phi_1 \mathbf{X}_{t-1} - \cdots - \phi_p \mathbf{X}_{t-p} \quad = \quad \mathbf{Z}_t, \tag{3.1}$$

where $\phi_1, \ldots, \phi_p$ are real-valued $m \times m$ matrices of AR coefficients; $\mathbf{Z}_t = (Z_{t,1}, \cdots, Z_{t,m})^T$ is an iid sequence of random vectors with mean $\mathbf{0}$ and distributed as P. Throughout this chapter, we assume $P \in \mathcal{P}$, where $\mathcal{P}$ is the set of probability measures on $\mathbb{R}^m$ with finite first moment and not supported on any sub-hyperplane of $\mathbb{R}^m$. Note that P need not to be continuous and discrete-valued noise is allowed. Define the AR matrix polynomial by $\mathbf{\Phi}(z) = \mathbb{I} - \phi_1 z - \cdots - \phi_p z^p$. Then (3.1) can be written in the compact form $\mathbf{\Phi}(B)\mathbf{X}_t = \mathbf{Z}_t$. The matrix polynomial $\mathbf{\Phi}(\mathbf{z}) = \mathbb{I} - \phi_1 z - \cdots - \phi_p z^p$ satisfies the condition $\det \mathbf{\Phi}(z) \neq 0$ for

$|z| \leqslant 1$. That is, $\det \mathbf{\Phi}(z)$ has no zeros inside the unit circle. Then, the process $\mathbf{X}_t$ is said to be causal and $\mathbf{X}_t$ only depends on the past and present shocks $\mathbf{Z}_s, s \leqslant t$ (Brockwell and Davis, 2009). Otherwise, the process $\mathbf{X}_t$ is noncausal and depends on the future noises. Causal VAR models usually assume Gaussian innovations. Even if $\mathbf{Z}_t$ is non-Gaussian, the quasi Gaussian likelihood estimator is still consistent. But we can get a more efficient estimator asymptotically by maximizing the actual likelihood function, assuming a known distribution for $\mathbf{Z}_t$. The noncausal VAR models are more complicated than the univariate case. We only consider estimating causal VAR models using the log-concave density estimator.

Cule et al. (2010) studied maximum likelihood density estimation for multi-dimensional log-concave measures. See Section 2.2 in Chapter 2 for a quick review of the properties of the log-concave density estimator. We stick to the notation used in Section 2.2 for consistency. The rest of the chapter is organized as follows. Section 3.2 describes the semiparametric estimation framework for causal VAR(p) models and shows the consistency of the estimators. Section 3.3 contains simulation studies to evaluate the finite sample performance of the semiparametric estimator.

3.2 Model formulation

Given the $m \times m$ matrix A, let $vec(A)$ denote the vectorization of the $m \times m$ matrix A, which is the $m^2 \times 1$ column vector obtained by stacking the columns of the matrix A. Denote $\mathbf{\Phi}$ as the VAR parameter vectors $\left(vec(\phi_1)^T, vec(\phi_2)^T, \ldots, vec(\phi_p)^T\right)^T \in \mathbb{R}^{m^2 p}$. Assume that the parameter space Θ is a compact subset of $\{\mathbf{\Phi} \in \mathbb{R}^{m^2 p} : \det \mathbf{\Phi}(z) \neq 0 \,\text{for}\, |z| \leqslant 1\}$. Let $\mathbf{\Phi}_0$ denote the true parameter and P_0 be the true distribution of $\mathbf{Z}_t$. Define $\mathbf{Z}_t(\mathbf{\Phi}) := \mathbf{\Phi}(B)\mathbf{X}_t$. Then $\mathbf{Z}_t(\mathbf{\Phi})$ is stationary and ergodic.

Define $\mathbf{Y}_t$ as

$$\mathbf{Y}_t = \begin{bmatrix} \mathbf{X}_t \\ \mathbf{X}_{t-1} \\ \vdots \\ \mathbf{X}_{t-p} \end{bmatrix}_{m(p+1)\times 1} .$$

Then $\mathbf{Z}_t(\mathbf{\Phi}) = \mathbf{\Phi}_{\mathbf{Y}}\mathbf{Y}_t$, where $\mathbf{\Phi}_Y = \begin{pmatrix} \mathbb{I}_m & -\phi_1 & -\phi_2 & \cdots & \cdots & -\phi_p \end{pmatrix}_{m\times m(p+1)}$. Let $P_{\mathbf{\Phi}}$ denote the stationary distribution of $\mathbf{Z}_t(\mathbf{\Phi})$, so that $P_{\mathbf{\Phi_0}}$ is the true distribution P_0. Given the observed sequence $\mathbf{X}_1, \ldots, \mathbf{X}_n$, let $\mathbb{P}_{\mathbf{\Phi},n} := \frac{1}{n-p}\sum_{i=p+1}^{n} \delta_{\mathbf{Z}_i(\mathbf{\Phi})}$ be the empirical measure of the residuals $\{\mathbf{Z}_i(\mathbf{\Phi})\}_{i=p+1}^{n}$. Recall that the bounded Lipschitz distance D_{BL} (see the definitions in Section 2.2) metrizes the weak convergence of probability measures; that is, a sequence of probability measures Q_n converges weakly to some probability measure Q if and only if $\lim_{n\to\infty} D_{BL}(Q_n, Q) = 0$. By Theorem 2.2 in Berti et al. (2006), the sequence of random measures $\mathbb{P}_{\mathbf{\Phi},n}$ converges weakly to $P_{\mathbf{\Phi}}$ almost surely for each $\mathbf{\Phi} \in \Theta$, i.e.,

$$D_{BL}\left(\mathbb{P}_{\mathbf{\Phi},n}, P_{\mathbf{\Phi}}\right) \xrightarrow{a.s.} 0 \text{ as } n \to \infty.$$

In fact, as shown in Proposition 3.2.1, $\mathbb{P}_{\mathbf{\Phi},n}$ admits a stronger convergence result.

Proposition 3.2.1.

$$\sup_{\mathbf{\Phi}\in\Theta} D_{BL}\left(\mathbb{P}_{\mathbf{\Phi},n}, P_{\mathbf{\Phi}}\right) \xrightarrow{a.s.} 0 \ \textit{as}\ n \to \infty.$$

Proof. Let f be a bounded Lipschitz function on $\mathbb{R}^m$ with $\|f\|_\infty \leqslant 1$ and $\|f\|_L \leqslant 1$. For any $\mathbf{\Phi} \in \Theta$ and $x, x' \in \mathbb{R}^{m(p+1)}$ we have

$$\begin{aligned}\left|f\left(\mathbf{\Phi}_Y x\right) - f\left(\mathbf{\Phi}_Y x'\right)\right| &\leqslant \|\mathbf{\Phi}_Y x - \mathbf{\Phi}_Y x'\| \\ &\leqslant \|\mathbf{\Phi}_Y\|\|x - x'\| \\ &\leqslant \sup_{\mathbf{\Phi}\in\Theta}\|\mathbf{\Phi}_Y\|\|x - x'\|.\end{aligned}$$

As a result, the function $f_{\mathbf{\Phi}}(x) := f\left(\mathbf{\Phi}_{\mathbf{Y}} x\right)$ is a bounded Lipschitz function on $\mathbb{R}^{m(p+1)}$ with Lipschitz constant $\|f_{\mathbf{\Phi}}\|_L \leqslant \|\mathbf{\Phi}_Y\|$. Note that

$$f\left(\mathbf{Z}_t(\mathbf{\Phi})\right) = f_{\mathbf{\Phi}}(\mathbf{Y}_t).$$

And hence,

$$\int f\, d\left(\mathbb{P}_{\mathbf{\Phi},n} - P_{\mathbf{\Phi}}\right) = \int f_{\mathbf{\Phi}}\, d\left(Q_n - Q\right),$$

where Q_n and Q denote the empirical measure $\frac{1}{n-p}\sum_{i=p+1}^{n}\delta_{\mathbf{Y}_i}$ and the stationary measure of the vector $\mathbf{Y}_t$, respectively. It then follows that

$$\begin{aligned}\sup_{\mathbf{\Phi}\in\Theta, \|f\|_\infty\leqslant 1, \|f\|_L\leqslant 1}\left|\int f\, d\left(\mathbb{P}_{\mathbf{\Phi},n} - P_{\mathbf{\Phi}}\right)\right| &= \sup_{\mathbf{\Phi}\in\Theta}\sup_{\substack{\|f_{\mathbf{\Phi}}\|_\infty\leqslant 1 \\ \|f_{\mathbf{\Phi}}\|_L\leqslant\|\mathbf{\Phi}_Y\|}}\left|\int f_{\mathbf{\Phi}} d\left(Q_n - Q\right)\right| \\ &\leqslant \sup_{\mathbf{\Phi}\in\Theta}\|\mathbf{\Phi}_Y\| D_{BL}(Q_n, Q). \qquad (3.2)\end{aligned}$$

Since Θ is assumed to be compact, the quantity $\sup_{\mathbf{\Phi}\in\Theta}\|\mathbf{\Phi}_Y\|$ is finite. And note that Q_n converges weakly to Q almost surely. Therefore, inequality (3.2) implies

$$\sup_{\mathbf{\Phi}\in\Theta} D_{BL}\left(\mathbb{P}_{\mathbf{\Phi},n}, P_{\mathbf{\Phi}}\right) \xrightarrow{a.s.} 0 \text{ as } n \to \infty.$$

□

Let $\Psi := \{\varphi \mid \varphi \text{ is concave and } \int e^{\varphi(u)}\, du = 1\}$ denote the set of concave functions φ such that e^{φ} is a pdf. A log-concave density e^{φ} is viewed as a generic candidate for the noise distribution. Then the likelihood of $\mathbf{\Phi}$ based on the sequence $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ and conditioning on $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p$ is given by

$$l_n(\mathbf{\Phi}, \varphi) = \int \varphi \, d\, \mathbb{P}_{\mathbf{\Phi},\mathbf{n}},$$

where $\varphi \in \Psi$. We consider estimating $(\mathbf{\Phi}_0, P_0)$ by maximizing $l_n(\mathbf{\Phi}, \varphi)$ over $\Theta \times \Psi$. We first project the empirical measure of the residuals $\mathbb{P}_{\mathbf{\Phi},n}$ on to the space of log-concave densities $\mathcal{F}$ on $\mathbb{R}^m$ to obtain the profile log-likelihood function

$$L(\mathbb{P}_{\mathbf{\Phi},n}) = \max_{\varphi \in \Psi} l_n(\mathbf{\Phi}, \varphi).$$

Recall that $\Pi(\mathbb{P}_{\mathbf{\Phi},n}) = \arg\max_{\varphi \in \Psi} l_n(\mathbf{\Phi}, \varphi)$ is the log-concave density estimator of $P_{\mathbf{\Phi}}$ based on the residuals $\{\mathbf{Z}_i(\mathbf{\Phi})\}_{i=p+1}^{n}$.

Proposition 3.2.2. *Under the assumption that $P_0 \in \mathcal{P}$ and Θ is compact, there exists a $\hat{\mathbf{\Phi}} \in \Theta$ maximizes $L(\mathbb{P}_{\mathbf{\Phi},n})$ over Θ.*

Proof. It is easy to see that $\mathbf{\Phi} \to \mathbb{P}_{\mathbf{\Phi},n}$ defines a continuous mapping from Θ to the space of probability measures $\mathcal{P}$ on $\mathbb{R}^m$ equipped with the first moment Mallow's distance. According to Lemma 2.2.1, the function $L(\mathbb{P}_{\mathbf{\Phi},n})$ of $\mathbf{\Phi}$ is continuous. Thus, $L(\mathbb{P}_{\mathbf{\Phi},n})$ attains its maximum on Θ at some $\hat{\mathbf{\Phi}} \in \Theta$. Then we have $\left(\hat{\mathbf{\Phi}}, \Pi(\mathbb{P}_{\hat{\mathbf{\Phi}},n})\right)$ maximizes $l_n(\mathbf{\Phi}, \varphi)$ over $\Theta \times \Psi$. □

We call $\hat{\mathbf{\Phi}}$ the maximum log-concave likelihood estimator (MLCLE). Theorem 3.2.3 shows the consistency of $\left(\hat{\mathbf{\Phi}},\, \Pi(\mathbb{P}_{\hat{\mathbf{\Phi}},n})\right)$.

Theorem 3.2.3. *Under the assumption that* $P_0 \in \mathcal{P}$ *and* Θ *is compact,* $\hat{\mathbf{\Phi}} = \arg\max_{\mathbf{\Phi}\in\Theta} L(\mathbb{P}_{\mathbf{\Phi},n})$ *is strongly consistent for* $\mathbf{\Phi}_0$ *and*

$$\sup_{z\in\mathbb{R}^m} \left|\Pi(\mathbb{P}_{\hat{\mathbf{\Phi}},n})(z) - \Pi\left(P_0\right)(z)\right| \xrightarrow{a.s.} 0.$$

Proof. Rewrite $\mathbf{Z}_t(\mathbf{\Phi})$ as

$$\mathbf{Z}_t(\mathbf{\Phi}) = \mathbf{Z}_t(\mathbf{\Phi}_0) + \mathbf{Z}_t(\mathbf{\Phi}) - \mathbf{Z}_t(\mathbf{\Phi}_0).$$

The causality of the process $\mathbf{X}_t$ implies that $\mathbf{Z}_t(\mathbf{\Phi}_0)$ is independent of $\mathbf{Z}_t(\mathbf{\Phi}) - \mathbf{Z}_t(\mathbf{\Phi}_0)$. Due to the *non-increasing under convolution* property of log-concave projection (see equation (2.2)), we have

$$\begin{aligned} L\left(\mathbf{Z}_t(\mathbf{\Phi})\right) &= L\left(\mathbf{Z}_t(\mathbf{\Phi}_0) + \mathbf{Z}_t(\mathbf{\Phi}) - \mathbf{Z}_t(\mathbf{\Phi}_0)\right) \\ &\leqslant L\left(\mathbf{Z}_t(\mathbf{\Phi}_0)\right). \end{aligned} \tag{3.3}$$

The inequality holds if and only if $\mathbf{Z}_t(\mathbf{\Phi}) - \mathbf{Z}_t(\mathbf{\Phi}_0) = \left(\mathbf{\Phi}_Y - \mathbf{\Phi}_{\mathbf{0}Y}\right)\mathbf{Y}_t = \delta_a$ for some vector $a \in \mathbb{R}^m$, which implies that $\mathbf{\Phi} = \mathbf{\Phi}_0$. Therefore, $\mathbf{\Phi}_0$ is the unique maximizer of $L\left(\mathbf{Z}_t(\mathbf{\Phi})\right)$ over Θ. In addition, by virtue of Lemma 2.2.1 (a), for any $\mathbf{\Phi} \in \Theta$, we have

$$\limsup_{n\to\infty} L(\mathbb{P}_{\mathbf{\Phi},n}) \leqslant L(P_{\mathbf{\Phi}}) \quad a.s.. \tag{3.4}$$

Inequalities (3.3) and (3.4) imply

$$\limsup_{n\to\infty} L(\mathbb{P}_{\mathbf{\Phi},n}) \leqslant L(P_{\mathbf{\Phi}_0}) \quad a.s.. \tag{3.5}$$

Let $\omega \in \Omega$ such that (3.5) holds and $\lim_{n\to\infty} L(\mathbb{P}_{\mathbf{\Phi}_0,n}) = L(P_{\mathbf{\Phi}_0})$. For such ω, let $\hat{\mathbf{\Phi}}_{n(k)}$ be any convergent subsequence of $\hat{\boldsymbol{\Phi}}$ with limit denoted as $\boldsymbol{\Phi'}$, we have

$$\begin{aligned} D_{BL}\left(\mathbb{P}_{\hat{\boldsymbol{\Phi}}_{n(k)},n(k)}, P_{\boldsymbol{\Phi'}}\right) &\leqslant D_{BL}\left(\mathbb{P}_{\hat{\boldsymbol{\Phi}}_{n(k)},n(k)}, P_{\hat{\boldsymbol{\Phi}}_{\mathbf{n(k)}}}\right) + D_{BL}\left(P_{\hat{\boldsymbol{\Phi}}_{n(k)}}, P_{\boldsymbol{\Phi'}}\right). \\ &\leqslant \sup_{\boldsymbol{\Phi}\in\Theta} D_{BL}\left(\mathbb{P}_{\boldsymbol{\Phi},n(k)}, P_{\boldsymbol{\Phi}}\right) + D_{BL}\left(P_{\hat{\boldsymbol{\Phi}}_{n(k)}}, P_{\boldsymbol{\Phi'}}\right). \end{aligned}$$

Since $\hat{\boldsymbol{\Phi}}_{n(k)}$ converges to $\boldsymbol{\Phi'}$, we have $D_{BL}\left(P_{\hat{\boldsymbol{\Phi}}_{n(k)}}, P_{\boldsymbol{\Phi'}}\right) \longrightarrow 0$. It then follows from Proposition 3.2.1 that

$$D_{BL}\left(\mathbb{P}_{\hat{\boldsymbol{\Phi}}_{n(k)},n(k)}, P_{\boldsymbol{\Phi'}}\right) \leqslant \sup_{\boldsymbol{\Phi}\in\Theta} D_{BL}\left(\mathbb{P}_{\boldsymbol{\Phi},n}, P_{\boldsymbol{\Phi}}\right) + D_{BL}\left(P_{\hat{\boldsymbol{\Phi}}_{n(k)}}, P_{\boldsymbol{\Phi'}}\right) \longrightarrow 0.$$

Therefore, according to Lemma 2.2.1 (a), we have $\limsup_{k\to\infty} L\left(\mathbb{P}_{\hat{\boldsymbol{\Phi}}_{n(k)},n(k)}\right) \leqslant L\left(P_{\boldsymbol{\Phi'}}\right)$. Notice that $\lim_{k\to\infty} L\left(\mathbb{P}_{\boldsymbol{\Phi}_0,n(k)}\right) \leqslant \liminf_{k\to\infty} L\left(\mathbb{P}_{\hat{\boldsymbol{\Phi}}_{n(k)},n(k)}\right)$. As a result,

$$L(P_{\boldsymbol{\Phi}_0}) \leqslant \liminf_{k\to\infty} L\left(\mathbb{P}_{\hat{\boldsymbol{\Phi}}_{n(k)},n(k)}\right) \leqslant \limsup_{k\to\infty} L\left(\mathbb{P}_{\hat{\boldsymbol{\Phi}}_{n(k)},n(k)}\right) \leqslant L\left(P_{\boldsymbol{\Phi'}}\right).$$

Inequality (3.3) further implies that $L\left(P_{\boldsymbol{\Phi}_0}\right) = L\left(P_{\boldsymbol{\Phi'}}\right)$. And hence, $\boldsymbol{\Phi}_0 = \boldsymbol{\Phi'}$. Therefore, we conclude that $\hat{\boldsymbol{\Phi}}_n$ converges to $\boldsymbol{\Phi}_0$ almost surely.

Note that $\mathbf{Z}_t(\hat{\boldsymbol{\Phi}}_n) = \hat{\boldsymbol{\Phi}}_{\mathbf{n},\mathbf{Y}}\mathbf{Y}_t$. Since $\mathbf{Y}_t$ is an ergodic process and has finite first moment, by triangular inequality, we have

$$\frac{1}{n-p}\sum_{t=p+1}^{n} \|\mathbf{Z}_t(\hat{\boldsymbol{\Phi}}_n)\| = \frac{1}{n-p}\sum_{t=p+1}^{n} \|\hat{\boldsymbol{\Phi}}_{\mathbf{n},\mathbf{Y}}\mathbf{Y}_t\| \xrightarrow{a.s.} E\|\boldsymbol{\Phi}_{\mathbf{0},\mathbf{Y}}\mathbf{Y}_t\| = E\|\mathbf{Z}_t\|.$$

In addition, $D_{BL}\left(\mathbb{P}_{\hat{\boldsymbol{\Phi}}_n}, P_{\boldsymbol{\Phi}_0}\right) \xrightarrow{a.s.} 0$. Therefore, we have $M_1\left(\mathbb{P}_{\hat{\boldsymbol{\Phi}}_n}, P_{\boldsymbol{\Phi}_0}\right) \xrightarrow{a.s.} 0$ by property (2.4). It then follows from Lemma 2.2.1 (c) and Lemma 2.2.2 that

$$\sup_{z\in\mathbb{R}^m} \left|\Pi(\mathbb{P}_{\hat{\boldsymbol{\Phi}},n})(z) - \Pi(P_0)(z)\right| \xrightarrow{a.s.} 0.$$

□

3.3 Numerical results

In order to evaluate the finite sample behavior of $\hat{\mathbf{\Phi}}$ (MLCLE) and to compare its performance with quasi Gaussian Likelihood (GL) and MLE (when distribution of Z_t is known) estimation methods, we consider a bivariate VAR(1) process: $\mathbf{X}_t - \phi\mathbf{X}_{t-1} = \mathbf{Z}_t$. Set $\phi = \begin{bmatrix} 0.1 & 0.3 \\ 0.9 & -0.5 \end{bmatrix}$. The R package **LogConcDEAD** (Cule et al., 2009) is used to compute the log-concave density MLE. We consider three innovation distributions: bivariate Normal distribution with mean zero and identity covariance matrix ($N(0, I_2)$); bivariate Laplace distribution with identity covariance matrix and scale parameter 1 ($L_{1,\Sigma=I_2}$); bivariate t distribution with identity covariance matrix and degree of freedom 6 ($t_{6,\Sigma=I_2}$). Note that $N(0, I_2)$ and $L_{1,\Sigma=I_2}$ are log-concave densities while $t_{6,\Sigma=I_2}$ is not log-concave. For each case, time series of lengths $100, 200, 500, 1000$ were simulated and for each realization, a VAR(1) model was fitted via the MLCLE, GL and MLE methods, respectively. For each sample size, this procedure was replicated 1000 times.

The results are reported in the following tables. Tables 3.1, 3.2, 3.3, 3.4 correspond to sample size 1000, 500, 200, 100, respectively. For $N(0, I_2)$ and $L_{1,\Sigma=I_2}$ which are log-concave, the root-mean-squared errors of MLCLE given in $(\cdot)$ are comparable to those of MLE across all sample sizes, suggesting the asymptotic efficiency of MLCLE when the true distribution is log-concave. For $L_{1,\Sigma=I_2}$, MLCLE has smaller root-mean-squared errors than those of GL. For $t_{6,\Sigma=I_2}$ which is not log-concave, MLCLE is slightly better than GL for moderate sample sizes.

$\mathbf{Z}_t$	Method	ϕ_{11}	ϕ_{21}	ϕ_{12}	ϕ_{22}
		0.1	0.9	0.3	-0.5
$N(0, I_2)$	GL	0.1004 (0.0290)	0.8991 (0.0288)	0.2986 (0.0178)	-0.4991 (0.0182)
	MLCLE	0.1004 (0.0294)	0.9001 (0.0290)	0.2983 (0.0194)	-0.4999 (0.0193)
$L_{1,\Sigma=I_2}$	GL	0.0996 (0.0289)	0.8994 (0.0290)	0.2991 (0.0182)	-0.4989 (0.0179)
	MLCLE	0.0992 (0.0212)	0.9014 (0.0214)	0.2992 (0.0135)	-0.4990 (0.0137)
	MLE	0.1005 (0.0212)	0.9000 (0.0209)	0.3001 (0.0132)	-0.4990 (0.0133)
$t_{6,\Sigma=I_2}$	GL	0.1003 (0.0285)	0.8992 (0.0283)	0.2987 (0.0182)	-0.4993 (0.0178)
	MLCLE	0.1004 (0.0256)	0.9002 (0.0266)	0.2983 (0.0176)	-0.4999 (0.0171)
	MLE	0.1003 (0.0261)	0.8996 (0.0258)	0.2989 (0.0164)	-0.4993 (0.0163)

Table 3.1: $n = 1000$: Mean and root-mean-squared error $(\cdot)$ of GL, MLCLE and MLE for VAR(1)

$\mathbf{Z}_t$	Method	ϕ_{11}	ϕ_{21}	ϕ_{12}	ϕ_{22}
		0.1	0.9	0.3	-0.5
$N(0, I_2)$	GL	0.0998 (0.0406)	0.8980 (0.0414)	0.2970 (0.0256))	-0.4988 (0.0255)
	MLCLE	0.0987 (0.0412)	0.8991 (0.0419)	0.2975 (0.0255)	-0.4990 (0.0266)
$L_{1,\Sigma=I_2}$	GL	0.0999 (0.0407)	0.8989 (0.0417)	0.2979 (0.0256)	-0.4980 (0.0257)
	MLCLE	0.0986 (0.0321)	0.9000 (0.0317)	0.2987 (0.0202)	-0.4990 (0.0196)
	MLE	0.1007 (0.0304)	0.9000 (0.0310)	0.2995 (0.0189)	-0.4988 (0.0192)
$t_{6,\Sigma=I_2}$	GL	0.1010 (0.0410)	0.8988 (0.0411)	0.2981 (0.0258)	-0.4977 (0.0261)
	MLCLE	0.0987 (0.0398)	0.8991 (0.0389)	0.2975 (0.0255)	-0.4990 (0.0247)
	MLE	0.1010 (0.0373)	0.8991 (0.0374)	0.2985 (0.0233)	-0.4979 (0.0237)

Table 3.2: $n = 500$: Mean and root-mean-squared error $(\cdot)$ of GL, MLCLE and MLE for VAR(1)

$\mathbf{Z}_t$	Method	ϕ_{11}	ϕ_{21}	ϕ_{12}	ϕ_{22}
		0.1	0.9	0.3	-0.5
$N(0, I_2)$	GL	0.0997 (0.0651)	0.8984 (0.0640)	0.2942 (0.0414)	-0.4961 (0.0409)
	MLCLE	0.0974 (0.0661)	0.8989 (0.0646)	0.2930 (0.0420)	-0.4974 (0.0425)
$L_{1,\Sigma=I_2}$	GL	0.1017 (0.0634)	0.8977 (0.0649)	0.2955 (0.0411)	-0.4962 (0.0404)
	MLCLE	0.0967 (0.0553)	0.8948 (0.0578)	0.2968 (0.0338)	-0.4980 (0.0344)
	MLE	0.1028 (0.0502)	0.9003 (0.0508)	0.2986 (0.0315)	-0.4970 (0.0318)
$t_{6,\Sigma=I_2}$	GL	0.0995 (0.0646)	0.8959 (0.0653)	0.2948 (0.0409)	-0.4964 (0.0415)
	MLCLE	0.0974 (0.0649)	0.8989 (0.0646)	0.2930 (0.0411)	-0.4974 (0.0425)
	MLE	0.0992 (0.0594)	0.8967 (0.0599)	0.2956 (0.0375)	-0.4971 (0.0379)

Table 3.3: $n = 200$: Mean and root-mean-squared error $(\cdot)$ of GL, MLCLE and MLE for VAR(1)

$\mathbf{Z}_t$	Method	ϕ_{11}	ϕ_{21}	ϕ_{12}	ϕ_{22}
		0.1	0.9	0.3	-0.5
$N(0, I_2)$	GL	0.1014 (0.0904)	0.8947 (0.0922)	0.2900 (0.0595)	-0.4911 (0.0595)
	MLCLE	0.0978 (0.1004)	0.8978 (0.1029)	0.2879 (0.0679)	-0.4953 (0.0635)
$L_{1,\Sigma=I_2}$	GL	0.1026 (0.0907)	0.8946 (0.0944)	0.2911 (0.0589)	-0.4921 (0.0590)
	MLCLE	0.0927 (0.0878)	0.8963 (0.0916)	0.2897 (0.0570)	-0.4965 (0.0581)
	MLE	0.1021 (0.0759)	0.8988 (0.0761)	0.2955 (0.0479)	-0.4940 (0.0480)
$t_{6,\Sigma=I_2}$	GL	0.1009 (0.0922)	0.8952 (0.0921)	0.2914 (0.0590)	-0.4914 (0.0584)
	MLCLE	0.0978 (0.1004)	0.8978 (0.1029)	0.2879 (0.0679)	-0.4953 (0.0635)
	MLE	0.1016 (0.0861)	0.8960 (0.0854)	0.2923 (0.0553)	-0.4924 (0.0546)

Table 3.4: $n = 100$: Mean and root-mean-squared error $(\cdot)$ of GL, MLCLE and MLE for VAR(1)

Chapter 4

Modeling Time Series of Counts with Shape Constraint

4.1 Introduction

In recent years there has been an increasing interest in analysis and modelling of time series of counts. Many time series of counts models assume the observations follow a Poisson distribution given the conditional mean process that characterizes the serial dynamics of the observed process, e.g., Davis et al. (2003); Heinen (2003); Ferland et al. (2006); Fokianos et al. (2009). Davis and Liu (2016) assumed that the observations follow a one-parameter exponential family distribution given the conditional mean process. They showed the stationarity and ergodicity of the underlying processes under fairly general conditions and established the asymptotic normality of the maximum likelihood estimators. The generalized linear autoregressive moving average (GLARMA) models (Shephard, 1995; Davis et al., 1999, 2003; Davis and Wu, 2009) also assumed that the observations are generated from a one-parameter exponential family distribution conditional on a latent process and covariates. Liboschik et al. (2015); Fokianos (2015) have given a review of the likelihood-based estimation methods for analysis and modeling of time series of counts data from a GLM perspective.

To relax the distributional assumption, we consider nonparametric shape constraints which are becoming increasingly popular and have various important applications in statistical inference, e.g., Dümbgen et al. (2011); Chen and Samworth (2015a,b). We take advantage of the one-parameter exponential family assumption and propose a semiparametric estimation procedure to the observation-driven models studied in Davis and Liu (2016). The underlying pmf p from the one-parameter exponential family is assumed to be log-concave, that is, $\log p$ is a concave function. For discrete distributions, "log-concavity" is defined in the sense that the linear interpolation of the logarithm of the probability mass function is concave. A more detailed description about the concave shape constraint is given in Section 4.2.1. This semiparametric estimation framework is rather generic and can be naturally applied to other time series of counts models which are based on an exponential family distribution assumption besides the conditional mean models considered in Davis and Liu (2016).

The rest of the chapter is organized as follows. Section 4.2.1 introduces the extended natural one-parameter exponential family which incorporates a concave component. Section 4.2.2 applies the extended natural one-parameter exponential family to the observation-driven models considered in Davis and Liu (2016). Section 4.3 explains some computational details. Section 4.4 shows the consistency of the MLE for the mean model parameters and the baseline function. Section 4.5 presents a simulation study and real data applications to further illustrate the results in Section 4.4. Technical details can be found in the Appendix.

4.2 Model formulation

4.2.1 Extended one-parameter exponential family

A random variable Y is said to follow a distribution of the one parameter exponential family if its pdf (or pmf in discrete case) with respect to some σ-finite measure μ can be written in the form

$$p(y \mid \eta) = h(y) \exp\left(\eta y - A(\eta)\right), \, y \geqslant 0, \tag{4.1}$$

where $h(\cdot)$ is a non-negative function and is supported on $\mathcal{X} \subset \mathbb{R}^+$; η is referred to as the natural parameter and $A(\eta)$ is the cumulant function given by

$$A(\eta) = \log \int_{\mathcal{X}} h(y) \exp\left(\eta y\right) \, d\,\mu(y).$$

The σ-finite measure μ is either Lebesgue measure or counting measure, which depends on whether $p(y \mid \eta)$ is a continuous distribution or not. Since we focus on time series of counts data, we express $A(\eta)$ explicitly as $A(\eta) = \log \sum_{y=0}^{\infty} h(y) \exp\left(\eta y\right)$, where μ is chosen to be counting measure and $\mathcal{X}$ is the discrete set $\mathcal{N}_0 = \{0, 1, \cdots\}$. The natural parameter η belongs to a subset of $\mathbb{R}$ such that $A(\eta)$ is well-defined, that is, $\eta \in \{\eta \in \mathbb{R} : A(\eta) < \infty\}$. It's known that $EY = A'(\eta)$. Since Y is assumed to be non-negative, the function $A(\cdot)$ is non-decreasing. The exponential family enjoys many useful properties and contains many well-known parametric classes. Distribution families indexed by a single parameter that belong to the one-parameter exponential family include the Poisson distribution, the exponential distribution, the Bernoulli distribution. Some of the two-parameter exponential families can be written in the form of (4.1) by fixing one of the parameters, for example, the normal distribution and the Gamma distribution. Classic expositions of the exponential family can be found in Lehmann and Casella (2006); Bickel and Doksum (2006).

We consider extending the natural one-parameter exponential family (4.1) by allowing the function $h(\cdot)$ to change. Let $\varphi(y) = \log h(y)$ and considering $\varphi(y)$ as a baseline function, we rewrite $p(y \mid \eta)$ explicitly as a function of η and φ:

$$p(y \mid \eta, \varphi) = \frac{\exp\left(\varphi(y) + \eta y\right)}{\sum_{y=0}^{\infty} \exp\left(\varphi(y) + \eta y\right)},\ y = 0, 1, 2, \ldots. \tag{4.2}$$

For any function φ on the discrete set $\mathcal{N}_0$, $p(y \mid \eta, \varphi)$ defines a natural one-parameter exponential family in regards to η as long as $\sum_{y=0}^{\infty} \exp\left(\varphi(y) + \eta y\right) < \infty$. For example, $\varphi(y) = -\log y!$ corresponds to the Poisson distribution. By introducing the infinite dimensional parameter φ, we will have more flexibility and obtain a richer class of models since the associated distribution family $p(y \mid \eta, \varphi)$ not only retains the good properties of the exponential family but also avoids being constrained to a particular parametric class. Note that $p(y \mid \eta, \varphi)$ is invariant under shift of φ, that is,

$$p(y \mid \eta, \varphi) = p(y \mid \eta, \varphi + \mathit{constant})$$

and choices of baseline functions φ for the exponential family (4.1) are not unique since

$$p(y \mid \eta, \varphi) = p(y \mid \eta - \tilde{\eta}, \varphi + \tilde{\eta} y).$$

Thus extra constraints on φ are required for identifiability. Denote $A_{\varphi}(\eta)$ as the log integral

$$\log \sum_{y=0}^{\infty} \exp\left(\varphi(y) + \eta y\right)$$

and let $B_{\varphi}(\eta)$ be the partial derivative of $A_{\varphi}(\eta)$ with respect to η. It is known that $E(Y) = B_{\varphi}(\eta)$ and $Var(Y) = \frac{\partial B_{\varphi}(\eta)}{\eta}$. Thus the function $B_{\varphi}(\eta)$ is strictly increasing provided that $p(y \mid \eta, \varphi)$ is non-degenerate. As a result, we can establish a one-to-one correspondence between η and $B_\varphi(\eta)$ such that the inverse function B_φ^{-1} is well-defined. In addition, since the observations Y considered in this paper are assumed to be non-negative, the function $A_\varphi(\eta)$ is not only convex but also increasing in η.

A concave shape constraint is imposed on the baseline function φ to control the complexity of the exponential families indexed by η and φ. For the exponential family $p(y \mid \eta, \varphi)$ with respect to counting measure, its support $\mathcal{X}$ is a collection of integers and the baseline function φ is only defined on a countable set. In this case, φ is said to be concave if the linear interpolation of the points $\{(y, \varphi(y)) : y \in \mathcal{X}\}$ is concave. More specifically, let $\{y_k\}_{k=1}^{\infty}$ be the elements in $\mathcal{X}$ which is in an increasing order. Define a piecewise linear function through

$$\bar{\varphi}(y) = \left(1 - \frac{y - y_k}{y_{k+1} - y_k}\right)\varphi(y_k) + \frac{y - y_k}{y_{k+1} - y_k}\varphi(y_{k+1}) \quad \text{for} \quad y \in [y_k, y_{k+1}], \tag{4.3}$$

where $\bar{\varphi}(y) = -\infty$ on $\mathbb{R} \setminus [\min_k y_k, \sup_k y_k]$. Then $p(y \mid \eta, \varphi)$ is said to satisfy the concave shape constraint if $\bar{\varphi}$ is a concave function. To simplify notation, $\bar{\varphi}$ is also used to denote the baseline function φ when the distribution (4.2) is continuous, where we actually have $\bar{\varphi} = \varphi$. Although choices of φ for (4.2) are not unique, all associated functions $\bar{\varphi}$ are concave if one of them is concave. Let $\mathcal{H}$ be the collection of concave functions $\psi : \mathbb{R}^+ \to \mathbb{R}$ such that $\psi(y) \to -\infty$ as $y \to \infty$ and $\int_0^\infty e^{\psi(y)} dy = 1$. Then a probability density function f on $\mathbb{R}^+$ is said to be log-concave if $\log f \in \mathcal{H}$. We assume that $\bar{\varphi} \in \mathcal{H}$, that is, $e^{\bar{\varphi}(y)}$ is a log-concave density with

$$\int_0^\infty e^{\bar{\varphi}(y)} dy = 1. \tag{4.4}$$

Moreover, in order to ensure identifiability, $\bar{\varphi}$ is also assumed to satisfy

$$\int_0^\infty y e^{\bar{\varphi}(y)} dy = 1. \tag{4.5}$$

Remark 8. There exists a unique function $\bar{\varphi}$ satisfying equations (4.4) and (4.5) for the exponential family (4.2). Let φ and φ^1 be any two functions such that $p(y \mid \eta, \varphi) = p(y \mid \eta^1, \varphi^1)$ on $\mathcal{X}$, where $\eta, \eta^1 \in \mathbb{R}$. Then, for any $y \in \mathcal{X}$,

$$\varphi(y) + \eta y - A_{\varphi}(\eta) = \varphi^1(y) + \eta^1 y - A_{\varphi^1}(\eta^1),$$

which implies

$$\bar{\varphi}^1(y) = \bar{\varphi}(y) + (\eta - \eta^1)y + A_{\varphi^1}(\eta^1) - A_{\varphi}(\eta).$$

Suppose that both the linear interpolations $\bar{\varphi}$ and $\bar{\varphi}^1$ satisfy equations (4.4) and (4.5):

$$\int_0^{\infty} e^{\bar{\varphi}^1(y)} dy = \int_0^{\infty} e^{\bar{\varphi}(y)+(\eta-\eta^1)y+A_{\varphi^1}(\eta^1)-A_{\varphi}(\eta)} dy = \int_0^{\infty} e^{\bar{\varphi}(y)} dy = 1$$
$$\int_0^{\infty} y e^{\bar{\varphi}^1(y)} dy = \int_0^{\infty} y e^{\bar{\varphi}(y)+(\eta-\eta^1)y+A_{\varphi^1}(\eta^1)-A_{\varphi}(\eta)} dy = \int_0^{\infty} y e^{\bar{\varphi}(y)} dy = 1.$$

Then the exponential family $\frac{\exp(\bar{\varphi}(y)+\zeta y)}{\int_0^{\infty} \exp(\bar{\varphi}(y)+\zeta y)\, d\, y}$ supported on $\mathbb{R}^+$ has mean value one when $\zeta = 0$ and $\zeta = \eta - \eta^1$, which implies $\eta - \eta^1 = 0$ since there is a one-to-one mapping from the natural parameter to the mean of an exponential family. As a result, $\bar{\varphi} = \bar{\varphi}^1$ and $\varphi = \varphi^1$. It should be noted that equation (4.5) is only to ensure identifiability of the baseline function φ and can be modified by $\int_0^{\infty} y e^{\bar{\varphi}(y)}\, d\, y = c$ for any $c \in \mathbb{R}^+$ with $\min_{y \in \mathcal{X}} y < c < \sup_{y \in \mathcal{X}} y$.

Let $\mathcal{H}_1 := \{\varphi \in \mathcal{H} : \int_0^{\infty} y e^{\varphi(y)} dy = 1\}$. Then the extended exponential family (4.2) indexed by the natural parameter η and function $\varphi \in \mathcal{H}_1$ is well-defined. For example, the exponential distribution has a unique representation as in (4.2) by $\eta < 1$ and $\varphi = -y \in \mathcal{H}_1$, where $\mathcal{X} = \mathbb{R}^+$ and μ is Lebesgue measure. Many of the standard exponential family distributions satisfy the concave shape constraint, for example, the Poisson distribution, the Bernoulli distribution and the exponential distribution. The negative binomial (NB) distribution $NB(r, p)$, the COM-Poisson distribution (λ, ν), and the Gamma distribution

$\Gamma(\alpha, \beta)$ have concave baseline functions when fixing the parameters r, ν and α, respectively.

Remark 9. Let $\Gamma(z)$ be the gamma function defined on $z \geqslant 0$. The second derivative of $\log \Gamma(z)$ is given by

$$\frac{d^2 \log \Gamma(z)}{dz^2} = \sum_{m=0}^{\infty} \frac{1}{(z+m)^2} \text{ for } z \geqslant 0; \tag{4.6}$$

see e.g., Medina and Moll (2007). Thus $\log \Gamma(z)$ is a strictly convex function. Taking log of the Poisson distribution $p(y \mid \lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$, we have $\log p(y \mid \lambda) = -\lambda + y \log \lambda - \log y!$. One of the baseline functions for Poisson distributions, $\varphi(y) = -\log y!$, is concave. And hence, the piecewise linear interpolation of the points $(k, -\log k!)\,, k = 0, 1, \ldots$ is also a concave function. As a result, the Poisson distribution satisfies the concave shape constraint. Figure 4.1 plots the baseline function $\varphi_0 \in \mathcal{H}_1$ of the Poisson distribution, where the black dots correspond to $-\log y!$, $y = 0, 1, 2, \ldots$.

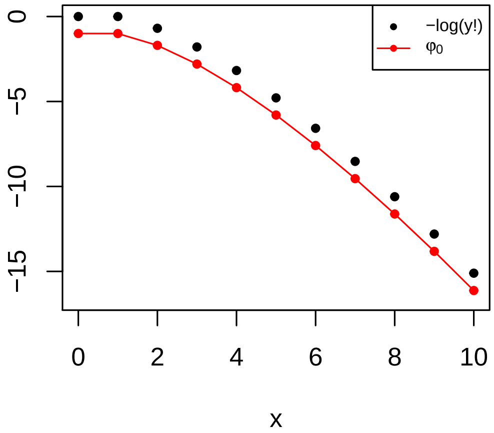

Figure 4.1: Baseline function φ_0 of the Poisson distribution

Moreover, it follows directly that COM-Poisson distribution $p(y \mid \lambda, \nu) = \frac{\lambda^y}{(y!)^\nu} \frac{1}{\sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}}$ also has a concave baseline function when the parameter ν is fixed. The logarithm of the

NB distribution $p(y \mid r, p) = \binom{y+r-1}{y} p^r (1-p)^y$ is given by

$$\log p(y \mid r, p) = \log(y+r-1)! - \log y! - \log(r-1)! + y \log(1-p) + r \log p.$$

In this case, consider $\varphi(y) = \log(y+r-1)! - \log y!$. According to equation (4.6), we have

$$\varphi''(y) = \sum_{m=0}^{\infty} \frac{1}{(y+r)^2} - \sum_{m=0}^{\infty} \frac{1}{(y+1)^2} \leqslant 0 \text{ for } r \geqslant 1,$$

and hence the function $\varphi(y)$ is concave so that the NB distribution also satisfies the concave shape constraint.

Based on the extended exponential family (4.2) indexed by the natural parameter η and function $\varphi \in \mathcal{H}_1$, we develop a semiparametric estimation framework for modeling time series of counts and show its flexibility compared with traditional parametric inference methods.

4.2.2 Semiparametric time series of counts models

We consider the observation-driven models studied by Davis and Liu (2016) where the ergodicity and stationarity properties of the underlying process were established. We state the relevant results in Davis and Liu (2016) for completeness. Let $\mathcal{F}_0 = \sigma(\eta_1)$, where η_1 is a natural parameter of (4.2) and is assumed to be fixed. The time series Y_t is generated as follows

$$Y_t \mid \mathcal{F}_{t-1} \sim p(y \mid \eta_t, \varphi), \quad X_t = g_\theta(X_{t-1}, Y_{t-1}), \tag{4.7}$$

where $\mathcal{F}_t = \sigma = (\eta_1, Y_1, \ldots, Y_t)$; $p(y \mid \eta_t, \varphi)$ is defined in (4.2) with $\varphi \in \mathcal{H}_1$ and $X_t := E\left[Y_t \mid \mathcal{F}_{t-1}\right]$ is the conditional mean process. The non-negative bivariate function $g_\theta(x, y)$ is defined on $[0, \infty) \times \mathcal{X}$. Davis and Liu (2016) showed that the process $\{X_t, Y_t\}$ generated by (4.7) is strictly stationary and ergodic provided that the function $g_\theta(x, y)$ satisfies a contraction condition: there exist two non-negative constants a and b with $a + b < 1$ such that for any $x, x' \geqslant 0$ and $y, y' \in \mathcal{X}$

$$|g_\theta(x, y) - g_\theta(x', y')| \leqslant a|x - x'| + b|y - y'|. \tag{4.8}$$

For integer valued Y_t, there exists a unique stationary and ergodic solution to the recursive equation $X_t = g_\theta(X_{t-1}, Y_{t-1})$ in terms of the process $\{Y_t\}$ if $g_\theta(\cdot, \cdot)$ satisfies condition (4.8), that is, X_t can be expressed as

$$X_t = g_\infty^\theta(Y_{t-1}, Y_{t-2}, \dots), \tag{4.9}$$

where g_∞^θ is a measurable function from $\mathcal{N}_0^\infty = \{(n_1, n_2, \dots), n_i \in \mathcal{N}_0, i = 1, 2, \dots\}$ to $[0, \infty)$.

Davis and Liu (2016) provided many examples with completely specified distribution function $p(y \mid \eta, \varphi)$ and $g_\theta(\cdot, \cdot)$, for which the expressions of mean, variance and autocorrelation functions were derived, and conditions for the maximum likelihood estimator of θ to be asymptotically normal were presented. Here we briefly cite some of the examples studied in Davis and Liu (2016).

Integer-valued generalized autoregressive conditional heteroscedasticity (INGARCH)$(1, 1)$ models: set $g_\theta(x, y) = \delta + \alpha x + \beta y$, where $\theta = (\delta, \alpha, \beta)^T \in \mathbb{R}^3$ with $\delta, \alpha, \beta > 0$ and $\alpha + \beta < 1$.

1. Poisson INGARCH(1,1) model:

$$Y_t \mid \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t),\ \lambda_t = \delta + \alpha\lambda_{t-1} + \beta Y_{t-1}.$$

2. NB INGARCH(1,1) model:

$$Y_t \mid \mathcal{F}_{t-1} \sim NB(r, p_t),\ X_t = \delta + \alpha X_{t-1} + \beta Y_{t-1},\ p_t = \frac{r}{X_t + r}.$$

3. Gamma INGARCH(1,1) model:

$$Y_t \mid \mathcal{F}_{t-1} \sim \Gamma(\kappa, s_t), \quad s_t = \frac{\delta}{\kappa} + \alpha s_{t-1} + \frac{\beta}{\kappa} Y_{t-1}.$$

Besides the linear dynamic models, Davis and Liu (2016) used spline basis functions to incorporate nonlinear dynamic models, where $g_\theta(x, y) = \delta + \alpha x + \beta y + \sum_{k=1}^{K} \beta_k (y - \epsilon_k)^+$. Here $\{\epsilon_k\}_{k=1}^{K}$ are the knots and $\theta = (\delta, \alpha, \beta, \beta_1, \ldots, \beta_K)^T \in \mathbb{R}^{K+3}$. The parameters satisfy the constraints $\alpha + \beta < 1, \beta + \sum_{k=1}^{K} \beta_k \geqslant 0$ and $\alpha + \beta + \sum_{k=1}^{s} \beta_k < 1$ for $s = 1, \ldots, K$ to ensure $g_\theta(\cdot, \cdot)$ is contractive in the sense of inequality (4.8). Distributions like Poisson, negative binomial and Gamma are also applicable for the nonlinear conditional mean model $g_\theta(\cdot, \cdot)$.

The dynamic behavior of the conditional mean process X_t is completely characterized by the recursive equation $g_\theta(X_{t-1}, Y_{t-1})$, which is indexed by a finite dimensional parameter $\theta \in \mathbb{R}^d$. The function B_φ is strictly increasing and its inverse function is well defined. The corresponding natural parameter η_t is determined as follows

$$X_t = B_\varphi(\eta_t), \quad \eta_t = B_\varphi^{-1}(X_t).$$

It serves as a connection between the parametric part θ and the nonparametric part φ. Here the function B_φ^{-1} plays the role of link function as in GLM. Davis and Liu (2016) showed the consistency and asymptotic normality of the maximum likelihood estimator of θ under some regularity conditions for fixed φ; that is, assuming a known parametric distribution class. We relax the distributional restriction by introducing the infinite dimensional parameter φ

and propose a semiparametric estimation procedure for model (4.7). Let θ_0 and $\varphi_0 \in \mathcal{H}_1$ denote the true parameters. Assumptions (A1-A6) are adapted from Davis and Liu (2016).

(A1) θ_0 is an interior point in the compact parameter space $\Theta \subset \mathbb{R}^d$

(A2) For any $\theta \neq \theta_0$ and all t, $P_{\theta_0}(\{X_t(\theta) \neq X_t(\theta_0)\}) > 0$, that is, $X_t(\theta)$ can distinguish θ_0 from $\theta \neq \theta_0$.

(A3) For any $\mathbf{y} \in [0, \infty)^\infty$ or $\mathcal{N}_0^\infty$, the mapping $\theta \to g_\infty^\theta$ is continuous.

(A4) The mapping $\theta \to g_\infty^\theta$ is twice continuously differentiable.

(A5) Denote $\mathcal{R}(B_{\varphi_0})$ as the range of $B_{\varphi_0}(\eta)$. For any $\theta \in \Theta$, $g_\infty^\theta \geqslant x_\theta^* > 0 \in \mathcal{R}(B_{\varphi_0})$. Moreover $x_\theta^* \geqslant x^* > 0 \in \mathcal{R}(B_{\varphi_0})$ for all θ.

(A6) $g(x, y)$ is increasing in (x, y) if Y_t given $\mathcal{F}_{t-1}$ has a continuous distribution.

(A7) There exists some $c > 0$ such that $A''_{\varphi_0}(\eta) \geqslant c > 0$ for all $\eta \in [B_{\varphi_0}^{-1}(x^*), \infty) \cap \{\eta : A_{\varphi_0}(\eta) < \infty\}$.

(A8) $E(X_t^2) < \infty$.

Remark 10. Assumptions (A7-A8) are sufficient conditions for the semiparametric MLE to be consistent. Time series models (4.7) based on Poisson, negative binomial and Gamma distributions all satisfy these two conditions.

4.3 The likelihood function

The likelihood function of model (4.7) based on the sequence $\{Y_1, \cdots, Y_n\}$ and conditioning on η_1 is given by

$$\begin{aligned} L(\theta, \varphi \mid Y_1, \cdots, Y_n, \eta_1) &= \prod_{t=1}^{n} p(Y_t \mid \eta_t(\theta), \varphi) \\ &= \prod_{t=1}^{n} \exp\left(\varphi(Y_t) + \eta_t(\theta) Y_t - A_\varphi(\eta_t(\theta))\right). \end{aligned}$$

The sequence $\eta_t(\theta)$ is calculated through the function $B_\varphi^{-1}(X_t(\theta))$. We can choose $X_1(\theta) = 1$ and $\eta_1 = 0$. The corresponding log likelihood function is

$$\begin{aligned} l_n(\theta, \varphi) &:= \frac{1}{n} \log L(\theta, \varphi \mid Y_1, \cdots, Y_n, \eta_1) \\ &= \frac{1}{n} \sum_{t=1}^{n} \left[\varphi(Y_t) + \eta_t(\theta) Y_t\right] - \frac{1}{n} \sum_{t=1}^{n} A_\varphi\left(\eta_t(\theta)\right). \end{aligned} \tag{4.10}$$

For integer-valued Y_t, there may be repeated values of the data. Let K be the number of unique values of $\{Y_1, \ldots, Y_n\}$ and $Y_{(1)}, \ldots, Y_{(K)}$ be the corresponding ordered unique values. Given the order statistics $Y_{(1)}, \ldots, Y_{(K)}$, let $\mathcal{G}_n$ be the collection of vectors $(\varphi_1, \ldots, \varphi_K) \in \mathbb{R}^K$ such that the linear interpolation $\varphi^n(y)$ through the points $\left(Y_{(k)}, \varphi_k\right), k = 1, \ldots, K$, belongs to $\mathcal{H}_1$. That is,

$$\varphi^n(y) = \left(1 - \frac{y - Y_{(k)}}{\Delta_k}\right) \varphi_k + \frac{y - Y_{(k)}}{\Delta_k} \varphi_{k+1} \;\text{ for }\; y \in \left[Y_{(k)}, Y_{(k+1)}\right] \tag{4.11}$$

is concave with $\int_0^\infty e^{\varphi^n(y)} dy = 1$ and $\int_0^\infty y e^{\varphi^n(y)} dy = 1$, where $\Delta_k = Y_{(k+1)} - Y_{(k)}$ and $\varphi^n(y) = -\infty$ on $\mathbb{R} \setminus \left[Y_{(1)}, Y_{(K)}\right]$. The index set of $k's$ consists of the points in which the slopes of φ^n change. The function φ^n defined in equation (4.11) is fully specified by the vector $(\varphi_1, \ldots, \varphi_K)$. The elements in $\mathcal{G}_n$ can be viewed as piecewise linear functions and $\mathcal{G}_n$ is essentially a subset of $\mathcal{H}_1$. We assume $\mathcal{G}_n$ satisfies conditions (G1-G3). For any $\varphi^n = (\varphi_1, \dots, \varphi_K) \in \mathcal{G}_n$,

(G1) $\max_k \varphi_k \leqslant M_1$,

(G2) $\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\varphi_k \geqslant -M_2$,

(G3) $B_{\varphi^n}^{-1}\left(X_t(\theta)\right) \leqslant B_{\varphi^n}^{-1}(x^*) + M_3\left(X_t(\theta) - x^*\right)$ for $t = 1, \dots, n$ and $\theta \in \Theta$,

where $M_1, M_2, M_3 > 0$ are constants such that $\max_{y\in\mathbb{R}} \varphi_0(y) \leqslant M_1$, $E\left(\varphi_0(Y_t)\right) \geqslant -M_2$ and $\frac{1}{c} \leqslant M_3$ for the constant c in assumption (A7). In practice, we can choose M_1, M_2, M_3 large. Regularity conditions G1-G3 imposed on $\mathcal{G}_n$ play an important role in establishing the consistency of the estimators. See details in the Appendix.

Let $\tilde{\varphi}_0^n$ be the function as defined in (4.11) based on the vector $\left(\varphi_0(Y_{(1)}), \dots, \varphi_0(Y_{(K)})\right)$. Note that $\tilde{\varphi}_0^n$ is concave but does not necessarily belong to $\mathcal{G}_n$. To ensure it belongs to $\mathcal{G}_n$, we can construct a modified version of $\tilde{\varphi}_0^n$ as

$$\varphi_0^n(y) = \tilde{\varphi}_0^n(y) + \eta_n^* y - \log \int_0^\infty e^{\tilde{\varphi}_0^n(y) + \eta_n^* y} dy,$$

where η_n^* satisfies the equation $\frac{\int_0^\infty y e^{\tilde{\varphi}_0^n(y)+\eta y} dy}{\int_0^\infty e^{\tilde{\varphi}_0^n(y)+\eta y} dy} = 1$. Proposition 4.6.10 shows that $\varphi_0^n(y) \in \mathcal{G}_n$ for large n *a.s.* and $l(\theta_0, \varphi_0) \leqslant \liminf_{n\to\infty} l_n(\theta_0, \varphi_0^n)$ *a.s.*, where

$$l(\theta_0, \varphi_0) := E\left[\varphi_0(Y_t) + \eta_t(\theta_0)Y_t - A_{\varphi_0}\left(\eta_t(\theta_0)\right)\right].$$

The expectation is with respect to the stationary measure of $(X_t, Y_t)^T$. Given the observations $\{Y_1, \dots, Y_n\}$, the set $\mathcal{G}_n$ is a subset of $\mathbb{R}^K$ by construction. Hence, maximizing the log likelihood function $l_n(\theta, \varphi)$ over $\Theta \times \mathcal{G}_n$ reduces to a finite-dimensional optimization problem. For any $\varphi = (\varphi_1, \dots, \varphi_K) \in \mathcal{G}_n$ and integer-valued Y_t, the log integral $A_\varphi(\eta) = \log \sum_{k=1}^K \exp\left(\varphi_k + \eta Y_{(k)}\right)$. It is straightforward to compute $\eta_t(\theta) = B_\varphi^{-1}\left(X_t(\theta)\right)$

using a binary search algorithm, where $B_\varphi(\eta) = A'_\varphi(\eta)$. Since A_φ and B_φ are both continuous functions, the log likelihood function $l_n(\theta,\varphi)$ is continuous on $\Theta \times \mathcal{G}_n$.

Proposition 4.3.1. *Under assumption A1, there exists $(\hat\theta_n, \hat\varphi_n) \in \Theta \times \mathcal{G}_n$ that maximizes $l_n(\theta,\varphi)$ over $\Theta \times \mathcal{G}_n$.*

Remark 11. We alternate between maximizing $l_n(\theta,\varphi)$ in θ and φ. Note that the maximization in $\varphi = (\varphi_1, \ldots, \varphi_K)$ is subject to the following $K-2$ linear constraints:

$$-\frac{1}{\Delta_{k-1}}\varphi_{k-1} + (\frac{1}{\Delta_{k-1}} + \frac{1}{\Delta_k})\varphi_k - \frac{1}{\Delta_k}\varphi_{k+1} \geqslant 0 \text{ for } k = 2, \ldots, K-1,$$

which adapt from the inequalities

$$\frac{\varphi_{k+1} - \varphi_k}{\Delta_k} \leqslant \frac{\varphi_k - \varphi_{k-1}}{\Delta_{k-1}} \text{ for } k = 2, \ldots, K-1,$$

to ensure concavity of φ. We call $(\hat\theta_n, \hat\varphi_n)$ the CMLE (concave MLE) of (θ_0, φ_0) and we show its consistency in Section 4.4.

4.4 Consistency

Let $\mathcal{H}_{\varphi_0} \coloneqq \{\varphi \in \mathcal{H}_1 : \varphi(y) = -\infty \text{ implies } \varphi_0(y) = -\infty \text{ for } y \in \mathbb{R}\}$. For fixed $(\theta,\varphi) \in \Theta \times \mathcal{H}_{\varphi_0}$, by virtue of the mean ergodic theorem, we have $l_n(\theta,\varphi) \xrightarrow{a.s.} l(\theta,\varphi)$, where $l(\theta,\varphi)$ is the limiting function of (4.10) given by

$$\begin{aligned} l(\theta,\varphi) &= E\left[\log p\left(Y_t \mid \eta_t(\theta), \varphi\right)\right] \qquad (4.12) \\ &= E\left[\varphi(Y_t) + \eta_t(\theta)Y_t - A_\varphi\left(\eta_t(\theta)\right)\right]. \end{aligned}$$

Proposition 4.4.1. *For any $\varphi \in \mathcal{H}_{\varphi_0}$, θ_0 is the unique maximizer of $l(\theta,\varphi)$; that is, for* *any* $\theta \in \Theta \setminus \theta_0$*,* $l(\theta, \varphi) - l(\theta_0, \varphi) < 0$*.*

Remark 12. Following the same argument as in Proposition 4.4.1, we know θ_0 is the unique maximizer of the function $l(\theta) := E\left[-\frac{Y_t}{X_t(\theta)} - \log X_t(\theta)\right]$ over Θ. In fact, $l(\theta)$ is obtained by substituting $p(y \mid \eta, \varphi)$ with the exponential distribution in equation (4.12). Let

$$\tilde{\theta} = \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{t=1}^{n} \left[-\frac{Y_t}{X_t(\theta)} - \log X_t(\theta)\right].$$

By assumption (A5), there exists $x^* > 0$ such that $X_t(\theta) \geqslant x^*$ for all t and $\theta \in \Theta$. Thus,

$$-\frac{Y_t}{X_t(\theta)} - \log X_t(\theta) \leqslant -\log x^* \text{ for all } t \text{ and } \theta \in \Theta.$$

Let $\dot{X}_t = \frac{\partial X_t(\theta)}{\partial \theta} \mid_{\theta=\theta_0}$ and suppose that $\Omega := E\left((\frac{Y_t}{X_t^2} - \frac{1}{X_t})^2 \dot{X}_t \dot{X}_t^T\right)$ exists. Then strong consistency and asymptotic normality of $\tilde{\theta}$ directly follow from the proof of Theorem 1 and Theorem 2 in Davis and Liu (2016), respectively. In particular,

$$\sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{D} N(0, \Omega^{-1}).$$

Since $\tilde{\theta}$ is obtained by maximizing the quasi-log likelihood function, it is referred to as QMLE (quasi MLE) and can be used as an initial point when maximizing $l_n(\theta, \varphi)$. Theorem 4.4.2 shows the consistency of the CMLE $(\hat{\theta}_n, \hat{\varphi}_n)$.

Theorem 4.4.2. *Assume (A1-A8). Then,* $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$ *and* $\hat{\varphi}_n \xrightarrow{a.s.} \varphi_0$*.*

Proof. Proposition 4.6.10 implies that

$$l(\theta_0, \varphi_0) \leqslant \liminf_{n \to \infty} l_n(\theta_0, \varphi_0^n) \leqslant \liminf_{n \to \infty} l_n(\hat{\theta}_n, \hat{\varphi}_n) \; a.s.. \tag{4.13}$$

On the other hand, by Proposition 4.6.8,

$$\limsup_{n\to\infty} l_n(\hat{\theta}_n, \hat{\varphi}_n) \leqslant \sup_{\theta\in\Theta,\varphi\in\mathcal{H}_{\varphi_0}} l(\theta,\varphi)\, a.s.. \tag{4.14}$$

According to Corollary 4.6.9, for any subsequence of $(\hat{\theta}_n, \hat{\varphi}_n)_n$, we can find a subsequence $(\hat{\theta}_{n_k}, \hat{\varphi}_{n_k})_k$ that converges pointwise to some $(\theta^*, \psi) \in \Theta \times \mathcal{H}_{\varphi_0}$ and

$$\limsup_{k\to\infty} l_{n_k}(\hat{\theta}_{n_k}, \hat{\varphi}_{n_k}) \leqslant l(\theta^*, \psi).$$

Together with inequalities (4.13) and (4.14), we have

$$l(\theta_0, \varphi_0) \leqslant l(\theta^*, \psi) \leqslant \sup_{\theta\in\Theta,\varphi\in\mathcal{H}_{\varphi_0}} l(\theta,\varphi).$$

Proposition 4.6.5 further implies that

$$\theta^* = \theta_0 \text{ and } \bar{\psi} = \bar{\varphi}_0 \text{ almost everywhere.} \tag{4.15}$$

(4.15) is true for any convergent subsequence of $(\hat{\theta}_n, \hat{\varphi}_n)$. Thus $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$ and $\hat{\varphi}_n \xrightarrow{a.s.} \varphi_0$. □

4.5 Examples

We illustrate the finite-sample performance of the semiparametric estimation procedure and compare its performance with MLE obtained by maximizing the actual log likelihood function and QMLE via simulation studies for the linear Poisson model. Empirical examples are also provided.

4.5.1 Poisson INGARCH(1,1) model

We consider the Poisson INGARCH(1,1) model:

$$Y_t \mid \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t), \quad \lambda_t = 0.5 + 0.5\lambda_{t-1} + 0.4Y_{t-1}.$$

Time series of lengths 500, 1000 are simulated and the parameter $\theta = (\delta, \alpha, \beta)$ is estimated with MLE $\check{\theta}$, CMLE $\hat{\theta}$ and QMLE $\tilde{\theta}$. This procedure is replicated 1,000 times and the resulting means and root-mean-squared errors of the estimates are summarized in Table 4.1. The performance of CMLE $\hat{\theta}$ is comparable to MLE $\check{\theta}$ in terms of root-mean-squared error. It is not surprising that the MLE $\check{\theta}$ has relatively smaller root-mean-squared-errors than the QMLE $\tilde{\theta}$. Histograms of the estimators $\hat{\theta}$ and $\tilde{\theta}$ are given in Figures 4.3 and 4.4, which are bell-shaped. The quasi-likelihood estimates $\tilde{\theta}$ are asymptotically normal according to Remark 12. Figure 4.2 shows the estimated conditional distribution $p\left(y \mid B_{\hat{\varphi}}^{-1}(x), \hat{\varphi}\right)$ with mean x based on one realization of time series Y_t of length 1000. It is very close to the corresponding Poisson distribution with the same mean.

n	Estimates	$\delta = 0.5$	$\alpha = 0.5$	$\beta = 0.4$
500	$\check{\theta}$	0.5954 (0.1771)	0.4702 (0.0612)	0.4097 (0.0412)
	$\hat{\theta}$	0.5645 (0.1690)	0.4934 (0.0591)	0.3933 (0.0429)
	$\tilde{\theta}$	0.619 (0.2169)	0.4594 (0.0719)	0.416 (0.0457)
1000	$\check{\theta}$	0.5445 (0.1066)	0.4858 (0.0398)	0.4060 (0.0290)
	$\hat{\theta}$	0.5357 (0.1085)	0.4947 (0.0404)	0.3977 (0.0300)
	$\tilde{\theta}$	0.5619 (0.1305)	0.4767 (0.0484)	0.4111 (0.0338)

Table 4.1: Estimates of Poisson INGARCH(1,1); Root mean squared error is given in $(\cdot)$.

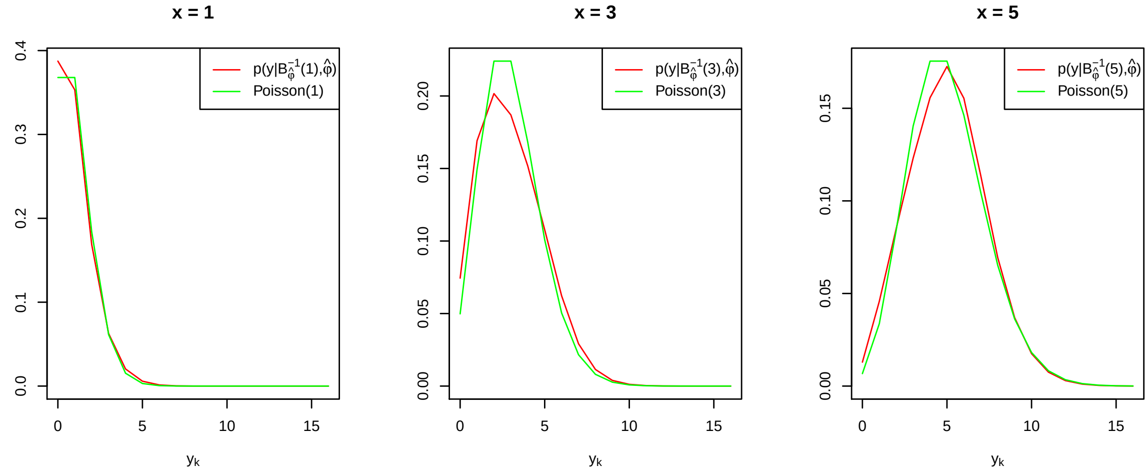

Figure 4.2: Red: Estimated conditional distribution $p\left(y \mid B_{\hat{\varphi}}^{-1}(x), \hat{\varphi}\right)$ based on one realization of time series Y_t of length 1000; Green: pmf of Poisson(x); x denotes the mean of the exponential family; Left: $x = 1$; Middle: $x = 3$; Right: $x = 5$.

4.5.2 Empirical examples

1. Number of transactions of Ericsson stock

We first look at the number of transactions per minute for the stock Ericsson B during July 2nd, 2002. This dataset is a typical example of time series of counts data and has been considered by Guikema and Goffelt (2008); Brännäs and Quoreshi (2010); Davis and Liu (2016), etc. According to the analysis in Davis and Liu (2016), negative binomial based models are more appropriate than Poisson based models, which is reasonable since the mean number of transactions per minute is 9.91 with a sample variance of 32.84, indicating strong overdispersion in the data. Figure 4.5 contains the plot of the time series and the sample ACF of the data, which shows positive serial correlation. Davis and Liu (2016) set knots at sample quantiles and selected the number of knots via AIC for the mean function. They suggest that the NB INGARCH(1,1) and 1-knot NB models fit relatively better than the Poisson based models or those with higher number of knots. Therefore, we fit a semiparametric (SP) INGARCH(1,1) model and a 1-knot SP model by maximizing the objective function (4.10)

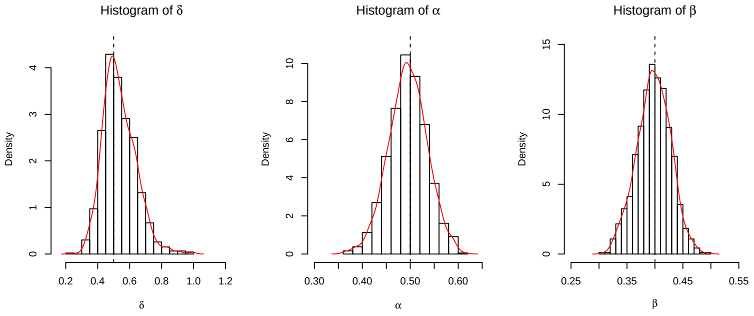

Figure 4.3: Histogram of CMLE $\hat{\theta}$ of Poisson INGARCH(1,1) of sample size 1000; The dashed vertical lines stand for the true values of the parameters.

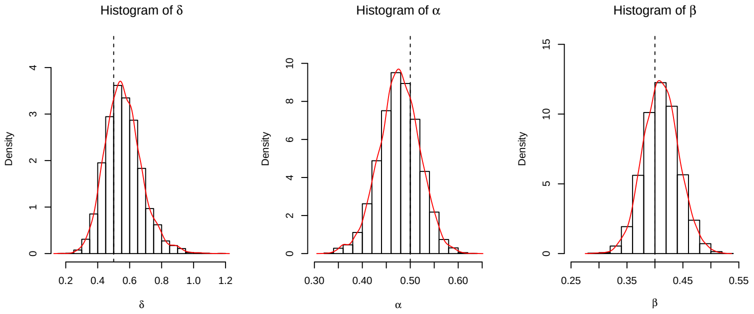

Figure 4.4: Histogram of QMLE $\tilde{\theta}$ of Poisson INGARCH(1,1) of sample size 1000; The dashed vertical lines stand for the true values of the parameters.

over $\Theta \times \mathcal{G}_n$ for comparison.

The fitted SP INGARCH(1, 1) is $\hat{X}_t = 0.2776 + 0.8392\hat{X}_{t-1} + 0.1328Y_{t-1}$. The fitted 1-knot SP model is

$$\hat{X}_t = 1.1014 + 0.7331\hat{X}_{t-1} + 0.1275Y_{t-1} + 0.1085(Y_{t-1} - 9)^+.$$

According to the analysis of Davis and Liu (2016), the estimated conditional distribution of the 1-knot NB model was given by NB$(8, \frac{8}{8+\hat{X}_t})$. Figure 4.6 plots the estimated conditional distribution $p(y \mid B_{\hat{\varphi}}^{-1}(x), \hat{\varphi})$ of the 1-knot SP model for the Ericsson stock data, the negative binomial distribution NB$(8, \frac{8}{x+8})$ and the Poisson(x) distribution, where x denotes the mean of the distribution. For small mean value, the estimated conditional distribution $p(y \mid$

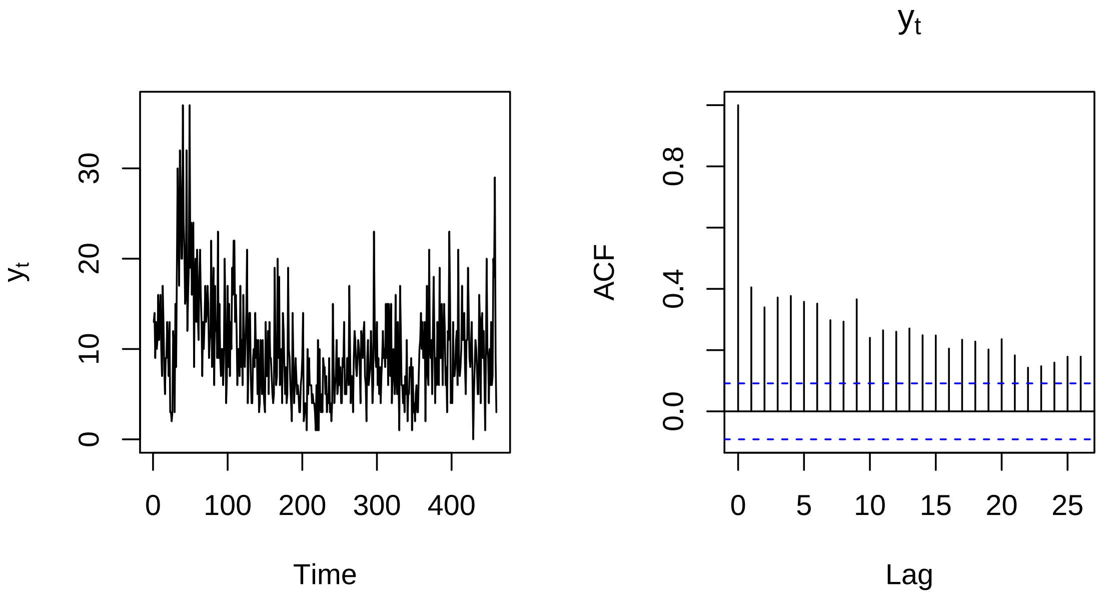

Figure 4.5: Left: Number of transactions per minute of the stock Ericsson during July 2nd 2002; Right: ACF of the data.

$B_{\hat{\varphi}}^{-1}(x), \hat{\varphi})$ is closer to the negative binomial distribution than to the Poisson distribution with the same mean. As the mean increases, $p(y \mid B_{\hat{\varphi}}^{-1}(x), \hat{\varphi})$ is data-driven and is skewed more to the right. We adopt some of the graphical and quantitative diagnostic tools used in Davis and Liu (2016) to measure the goodness of the fitted models. See Davis et al. (2003) and Jung and Tremayne (2011) for a detailed exposition. The standardized Pearson residuals are calculated according to the formula

$$e_t = \frac{Y_t - E(Y_t \mid \mathcal{F}_{t-1})}{\sqrt{Var(Y_t \mid \mathcal{F}_{t-1})}}.$$

The residuals should follow a white noise sequence if the model assumptions are correct. For the SP model, the conditional mean $E(Y_t \mid \mathcal{F}_{t-1})$ is equal to $\hat{X}_t = g_{\hat{\theta}}(\hat{X}_{t-1}, Y_{t-1})$ and the conditional variance $Var(Y_t \mid \mathcal{F}_{t-1})$ can be obtained by computing the second moment of the estimated conditional distribution $p\left(y \mid B_{\hat{\varphi}}^{-1}(\hat{X}_t), \hat{\varphi}\right)$. As shown in the top of Figure

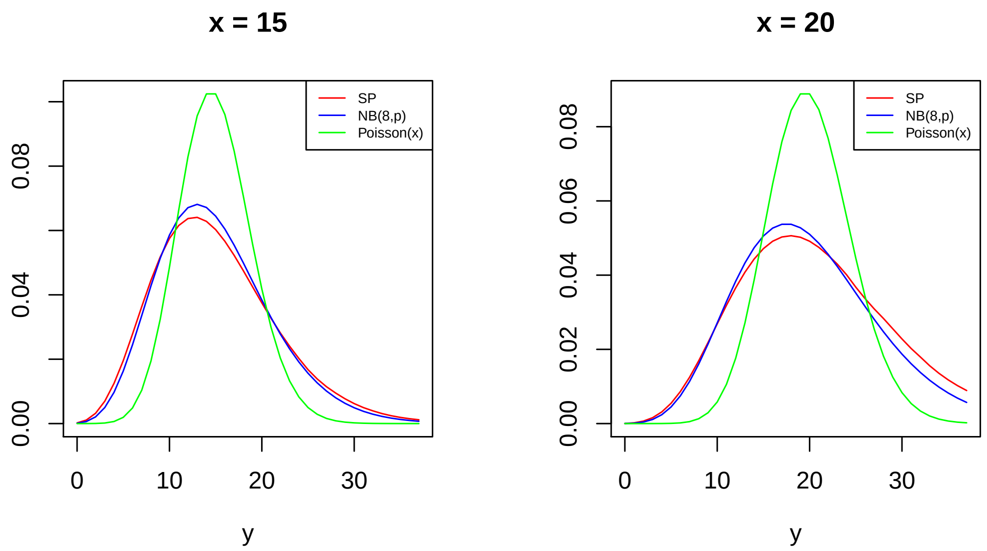

Figure 4.6: The red line is the estimated conditional distribution $p\left(y \mid B_{\hat{\varphi}}^{-1}(x), \hat{\varphi}\right)$ for the transactions data, the blue line is the pmf of NB $(8, \frac{8}{8+x})$ and the green line is the pmf of Poisson(x); x denote the mean of the exponential family; Left: $x = 15$; Right: $x = 20$.

.

4.7, the fitted conditional mean process $X_t(\hat{\theta})$ by the 1-knot SP model follows the count time series very well. The sample ACF of the residuals by SP INGARCH(1,1) and 1-knot SP models plotted in the bottom of Figure 4.7 do not exhibit any serial correlation, indicating that the residuals are compatible with white noise.

The probability integral transform (PIT) is another useful tool for testing the distributional assumptions of models, see Fokianos (2001). For a random variable X with a continuous distribution F, the PIT, $F(X)$, is known to have a standard uniform distribution. When the underlying distributions are discrete, adjustments are required. We consider the randomized PIT, which is obtained by perturbing the step-function nature of CDF for discrete random variables, see Czado et al. (2009). Given time series of counts data Y_t, the

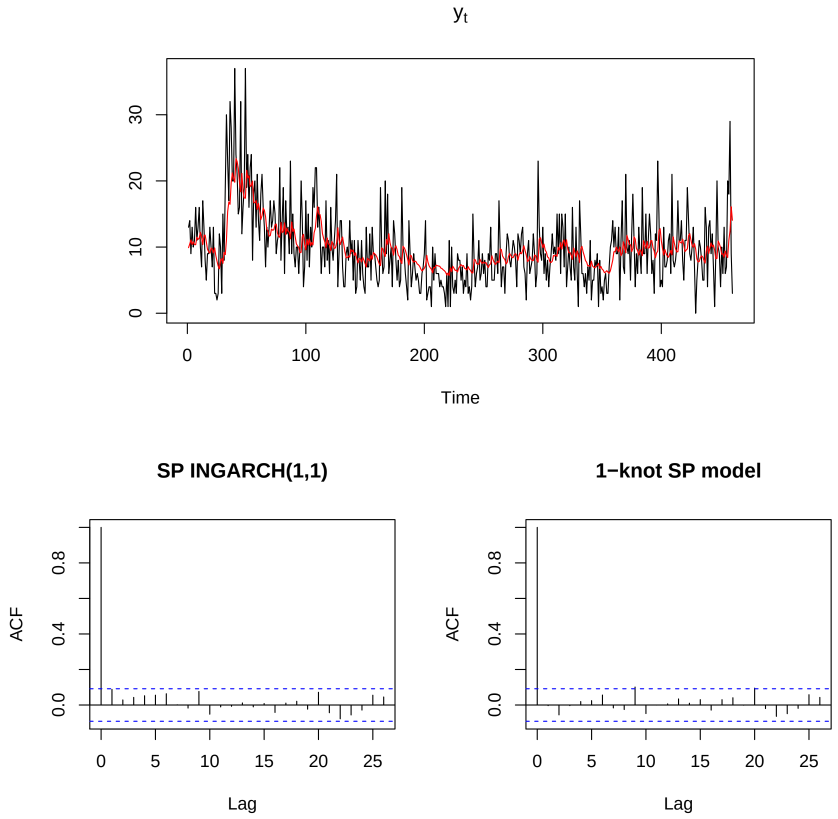

Figure 4.7: Top: The black line is the number of transactions of Ericsson stock, and the red line is the fitted conditional mean process by 1-knot SP model; Bottom: ACF of the standardized Pearson residuals of SP INGRACH$(1,1)$ (left) and 1-knot SP model (right) for the transactions data.

conditional randomized PIT is defined by

$$U_t := F_t(Y_{t-1}) + \nu_t \left[F_t(Y_t) - F_t(Y_t - 1) \right],$$

where $\{\nu_t\}$ is an iid sequence drawn from the standard uniform distribution and $F_t(\cdot)$ is the predictive cumulative distribution of Y_t. If the model is correctly specified, $\{U_t\}$ should be an iid sequence distributed as Uniform $(0,1)$. The histograms and QQ-plots of the randomized PIT are given in Figure 4.8 to test if U_t follows the standard uniform distribution. The p-values of Kolmogorov-Smirnov test are reported in Table 4.3. The randomized PIT of the two fitted SP models are both close to the uniform distribution. The predictive power of models

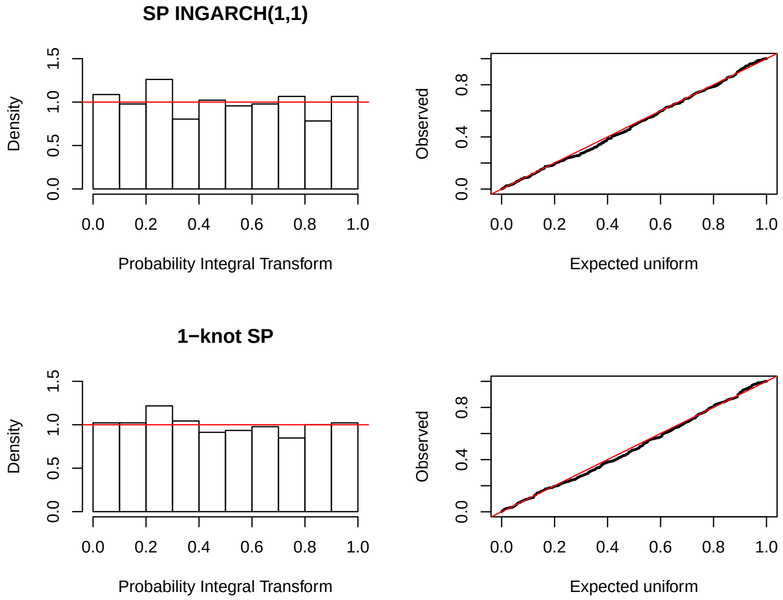

Figure 4.8: Left: Histograms of the randomized PIT for SP INGARCH(1,1) model and 1-knot SP model fitted to the transactions data, respectively; Right: QQ-plots of corresponding randomized PIT against the uniform distribution.

can be assessed by proper scoring rules (Czado et al., 2009; Fokianos, 2015). Denote a scoring rule for the predictive distribution p_t and the observation Y_t by $S(p_t, Y_t)$. The average score for each model is computed as $\frac{1}{n}\sum_{t=1}^{n} S(p_t, Y_t)$. Generally, models with the lowest mean score are preferable. We consider three scoring rules: logarithmic score (LS), quadratic score (QS), and ranked probability score (RPS). See Table 4.2 for exact formulas. Each of them measures different distances between the predictive distribution and the observed data. Table 4.3 reports the scores for the two fitted semiparametric models. The results for NB INGARCH(1,1) and 1-knot NB models are based on the analysis of Davis and Liu (2016). As seen from Table 4.3, there are no significant differences between the scores, which indicates that the negative binomial based models provide good fits to the Ericsson stock data.

Scoring rule	Definition $S(p_t, Y_t)$
logarithmic score	$-\log(p_t(Y_t))$
quadratic score	$-2p_t(Y_t) + \|p_t\|^2$
ranked probability score	$\sum_{y=0}^{\infty}\left(F_t(y) - \delta_{\{Y_t \leqslant y\}}\right)$

Table 4.2: Definitions of some of the scoring rules (Czado et al., 2009); F_t is CDF of p_t and $\|p_t\|^2 = \sum_{y=0}^{\infty} p_t(y)^2$.

Model	log likelihood	$p-$value of PIT	LS	QS	RPS
NB INGARCH(1,1)	-1332.02	0.7386	2.8958	-0.0671	2.6063
SP INGARCH(1,1)	-1330.67	0.6955	2.8732	-0.0676	2.6038
1-knot NB	-1331.34	0.8494	2.8942	-0.0671	2.6021
1-knot SP	-1330.56	0.7204	2.8797	-0.0665	2.6175

Table 4.3: Quantitative model checking for Ericsson B stock data.

2. Campylobacter infections in Canada

For the second example, we study the number of campylobacterosis cases (reported every 28 days) in the North of Québec in Canada from January 1990 to the end of October 2000. Campylobacterosis is an acute bacterial infectious disease and can be caused by eating unhealthy food. Infected people may suffer from abdominal pain, cramps and fever within 2 or 3 days of exposure to the organism. This dataset is available in R package **tscount**. A detailed description of the dataset can be found in Ferland et al. (2006). Figure 4.9 contains the plot of the data. The empirical ACF shows strong positive serial correlation. Ferland et al. (2006) considered a Poisson INGARCH$(13, 1)$ model:

$$Y_t \mid \mathcal{F}_t \sim \text{Poisson}(\lambda_t), \quad \lambda_t = 2.3135 + 0.2752\lambda_{t-13} + 0.5484Y_{t-1}.$$

Following the analysis of Ferland et al. (2006), we consider an INGARCH(1,1) model based on the Negative Binomial distribution and the SP approach, respectively. Table 4.4 reports the estimates of NB INGARCH$(1, 1)$ and the SP INGARCH$(1, 1)$ models. As seen from the table, the estimates obtained by different fitting methods are very close. The estimated

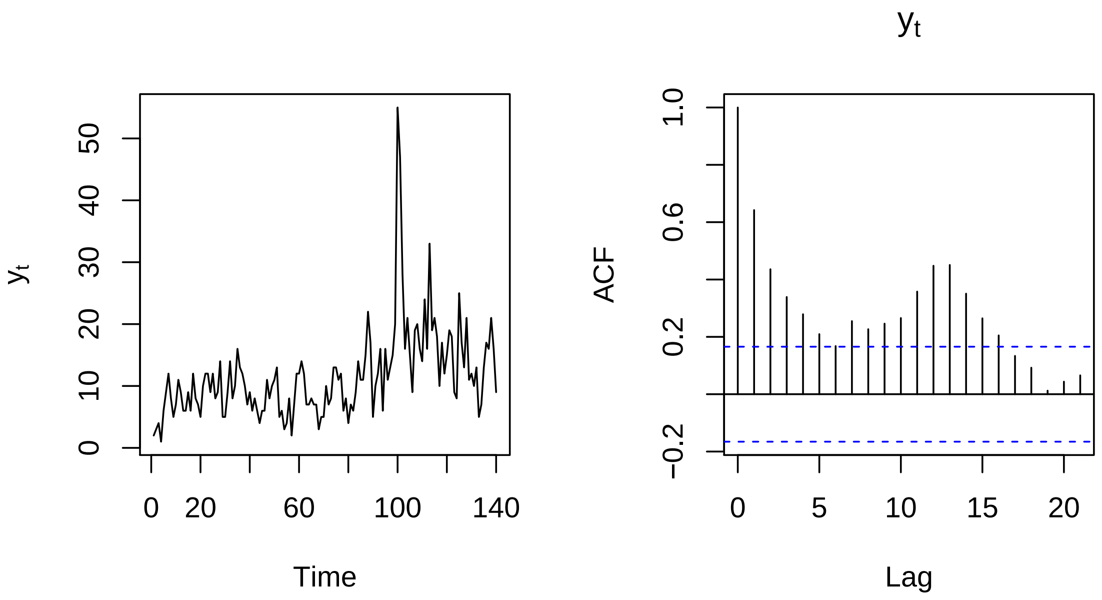

Figure 4.9: Left: Number of campylobacterosis cases (reported every 28 days) in the North of Québec in Canada between January 1990 and October 2000; Right: Sample ACF of the data.

conditional distribution of the NB INGARCH(1, 1) model is NB$(10, \frac{10}{10+\hat{X}_t})$. The estimated conditional distribution $p(y \mid B_{\hat{\varphi}}^{-1}(x), \hat{\varphi})$ of the SP INGARCH(1, 1) model for the campylobacter infections data is given in Figure 4.10. As the mean increases, $p(y \mid B_{\hat{\varphi}}^{-1}(x), \hat{\varphi})$ is more skewed to the right, while the negative binomial distribution NB$(10, \frac{10}{10+x})$ and the Poisson distribution with the same mean change only slightly in shape. The top of Figure 4.11 plots the fitted conditional mean process $X_t(\hat{\theta})$ of the SP INGRACH(1, 1) model, which moves along with the count time series. The standardized Pearson residuals shown in the bottom of Figure 4.11 appear to be white noise. The randomized PIT histogram given in Figure 4.12 corresponding to the SP INGARCH(1,1) model appears to be the closest to the uniform distribution. The scores given in Table 4.5 are also in favor of the SP INGARCH(1,1) model.

	δ	α	β
NB INGARCH$(1,1)$	2.2132	0.2691	0.5463
SP INGARCH$(1,1)$	2.1873	0.2640	0.5465
QMLE $\tilde{\theta}$	1.9248	0.3626	0.4781

Table 4.4: Estimates of the INGARCH$(1,1)$ models for the campylobacter infections data.

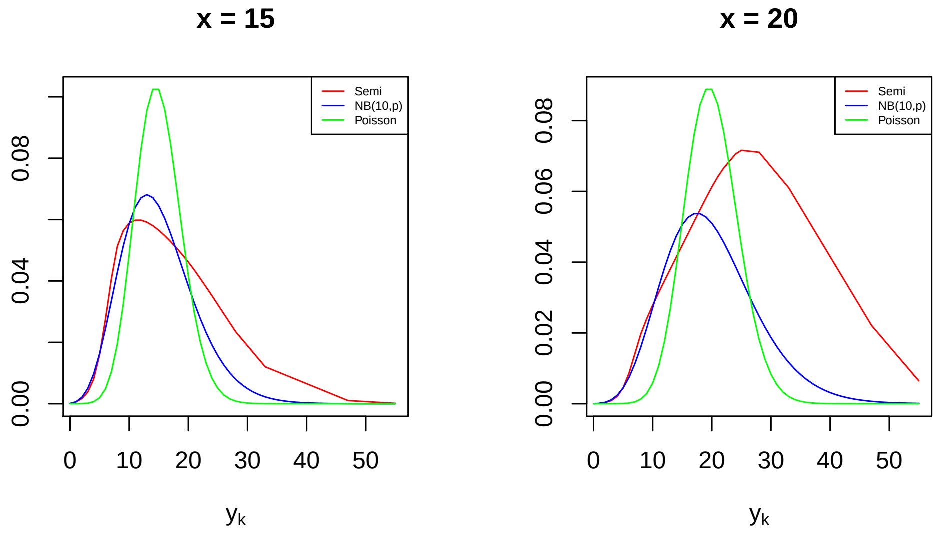

Figure 4.10: The red line is the estimated conditional distribution $p\left(y \mid B_{\hat{\varphi}}^{-1}(x), \hat{\varphi}\right)$ for the campylobacter infections data, the blue line is the pmf of NB $(10, \frac{10}{10+x})$ and the green line is the pmf of Poisson(x); x denote the mean of the exponential family; Left: $x = 15$; Right: $x = 20$.

Model	log likelihood	$p-$value of PIT	LS	QS	RPS
Poisson INGARCH$(13,1)$	-435.4042	0.3846	3.1100	-0.0689	2.6870
NB INGARCH$(1,1)$	-404.1755	0.7409	2.9046	-0.0705	2.6709
SP INGARCH$(1,1)$	-391.6108	0.8387	2.8141	-0.0717	2.6667

Table 4.5: Quantitative model checking for campylobacter infections data.

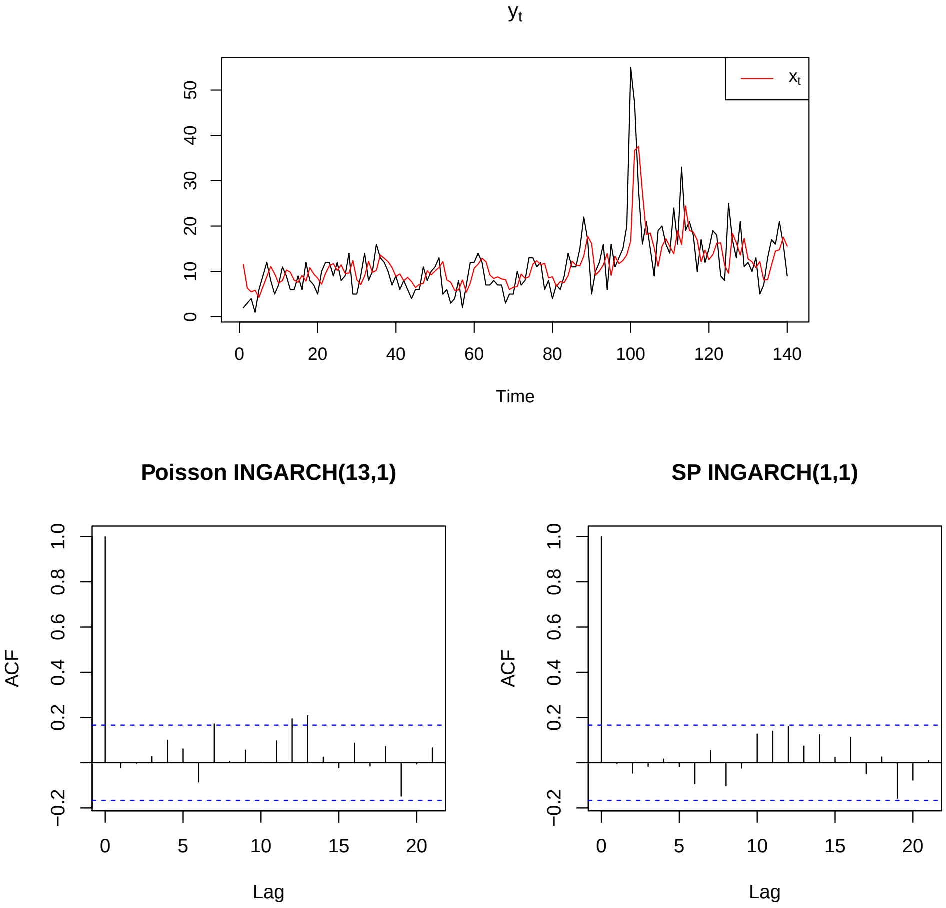

Figure 4.11: Top: The black line is the number of campylobacter cases, and the red line is the fitted conditional mean process by SP INGARCH$(1, 1)$ model; Bottom: ACF of the standardized Pearson residuals of Poisson INGARCH$(13, 1)$ (left) and SP INGARCH$(1, 1)$ (right) for the campylobacter infectious data.

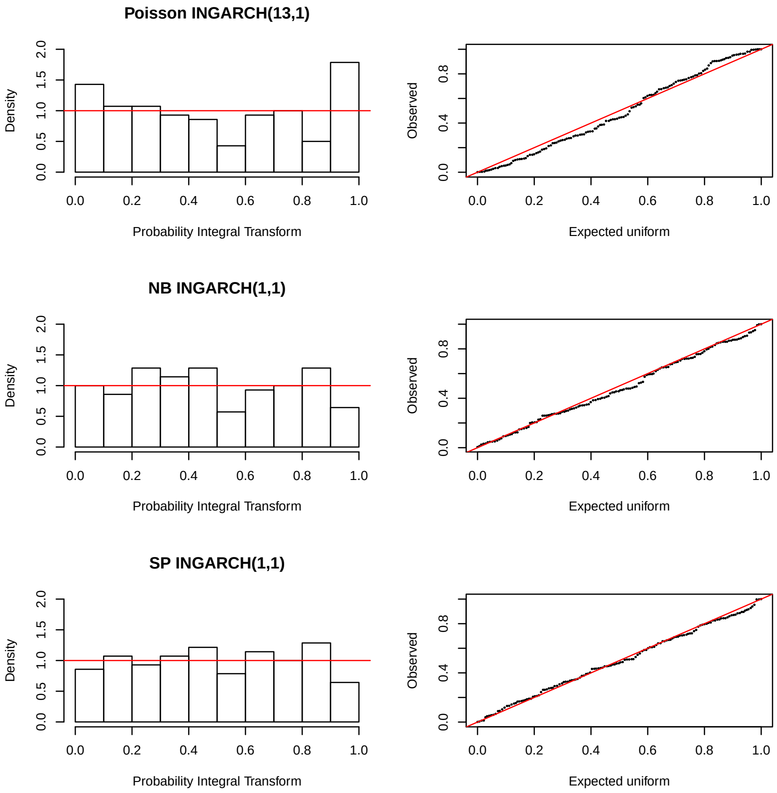

Figure 4.12: Randomized PIT histograms applied to the campylobacter infections data; Left: histograms of randomized PIT; Right: QQ-plots of corresponding randomized PIT against the uniform distribution.

3. Homicides data

Another example is the weekly number of deaths recorded at the Salt River state mortuary, Cape Town, for the period time $1986 - 1991$. This dataset can be downloaded from the website http://www.hmms-for-time-series.de/second/data. The left plot in Figure 4.13 shows this count time series. There are 312 observations in total. The sample mean is 2.63 and the sample variance is 6.59, indicating overdispersion in the data. The sample ACF of the time series, as shown in the right plot of Figure 4.13, shows strong positive serial correlation. We consider an INGARCH(1,1) model based on the SP approach. Table 4.6 shows the MLEs of the mean model based on the Poisson distribution, the negative binomial distribution and the SP approach, respectively. The estimators under the different distributional assumptions are very close. Note that the three fitted models all give small estimated intercept, which is reasonable since the sample mean of the homicides data is relatively small. The

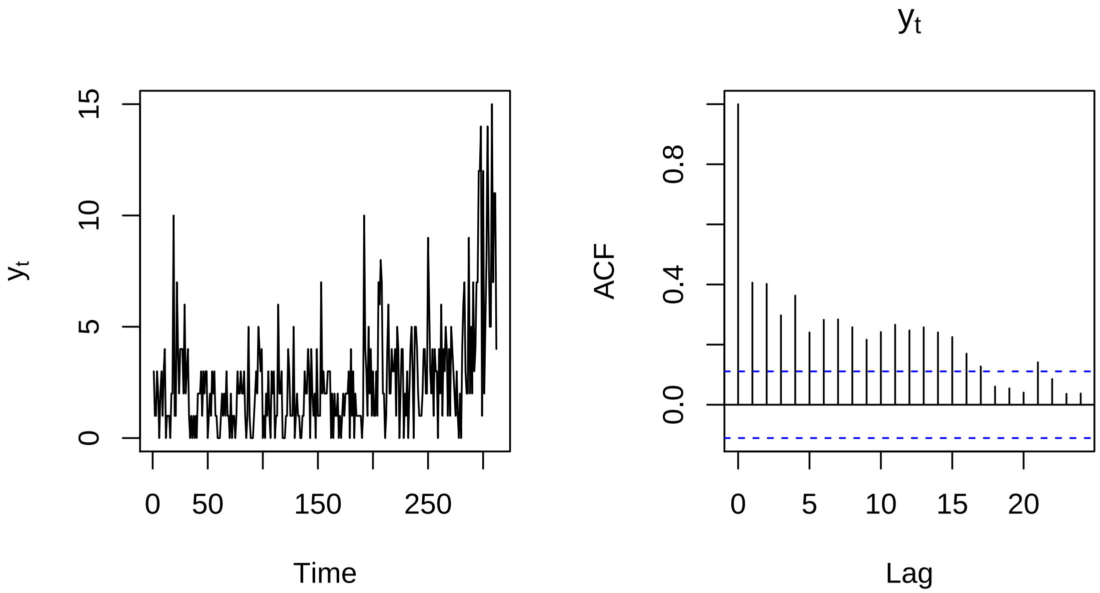

Figure 4.13: Left: Weekly number of deaths recorded at the Salt River state mortuary, Cape Town, between $1986 - 1991$; Right: Sample ACF of the data.

distribution coefficient r of the NB INGARCH$(1,1)$ model is 4. The estimated conditional distribution $p(y \mid B_{\hat{\varphi}}^{-1}(x), \hat{\varphi})$ of the SP INGARCH$(1,1)$ model for the homicides data shown in Figure 4.14 is skewed to the left when the mean is small and has fatter tails than those of the negative binomial distribution and the Poisson distribution with the same mean. When the mean increases, $p(y \mid B_{\hat{\varphi}}^{-1}(x), \hat{\varphi})$ is skewed to the right. The fitted conditional mean process $X_t(\hat{\theta})$ shown in the top of Figure 4.15 follows the count time series well and there is no apparent serial correlation in the standardized Pearson residuals by looking at the corresponding sample ACF plot in the bottom of Figure 4.15. The randomized PIT of the SP INGARCH(1,1) model, shown in Figure 4.16, appears to be the closest to the uniform distribution. The scores given in Table 4.7 are also in favor of the SP INGARCH(1,1) model.

	δ	α	β
Poisson INGARCH$(1,1)$	0.0436	0.8641	0.1259
NB INGARCH$(1,1)$	0.0474	0.8716	0.1159
SP INGARCH$(1,1)$	0.0826	0.8379	0.1307

Table 4.6: Estimates of the INGARCH$(1,1)$ models for homicides data.

Model	log likelihood	$p-$value of PIT	LS	QS	RPS
Poisson INGARCH	-631.6997	0.0752	2.0231	-0.1694	1.1026
NB INGARCH(1,1)	-610.6602	0.4755	1.9565	-0.1753	1.0857
SP INGARCH(1,1)	-607.6660	0.6339	1.9476	-0.1758	1.0884

Table 4.7: Quantitative model checking for homicides data.

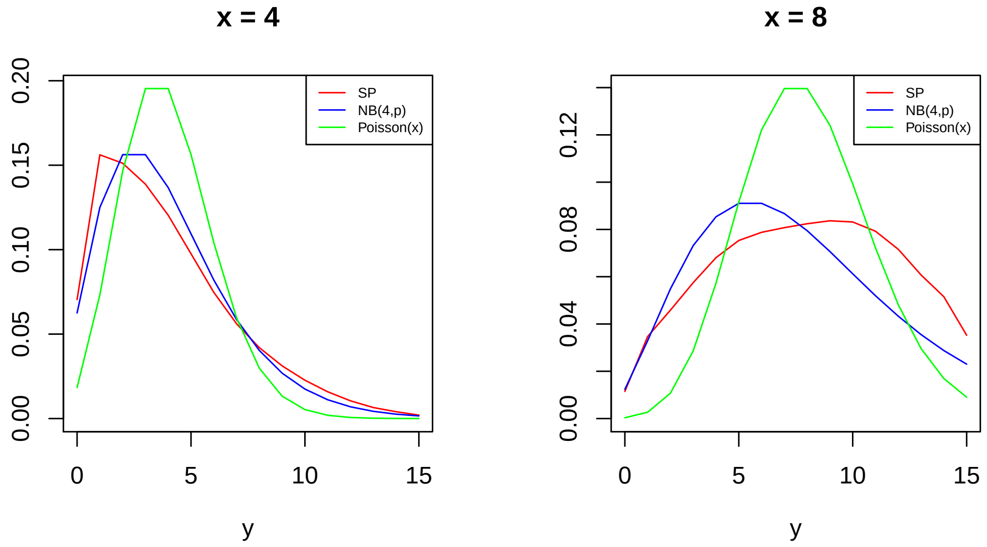

Figure 4.14: The red line is the estimated conditional distribution $p\left(y \mid B_{\hat{\varphi}}^{-1}(x), \hat{\varphi}\right)$ for the homicides data, the blue line is the pmf of NB $(4, \frac{4}{4+x})$ and the green line is the pmf of Poisson(x); x denote the mean of the exponential family; Left: $x = 4$; Right: $x = 8$.

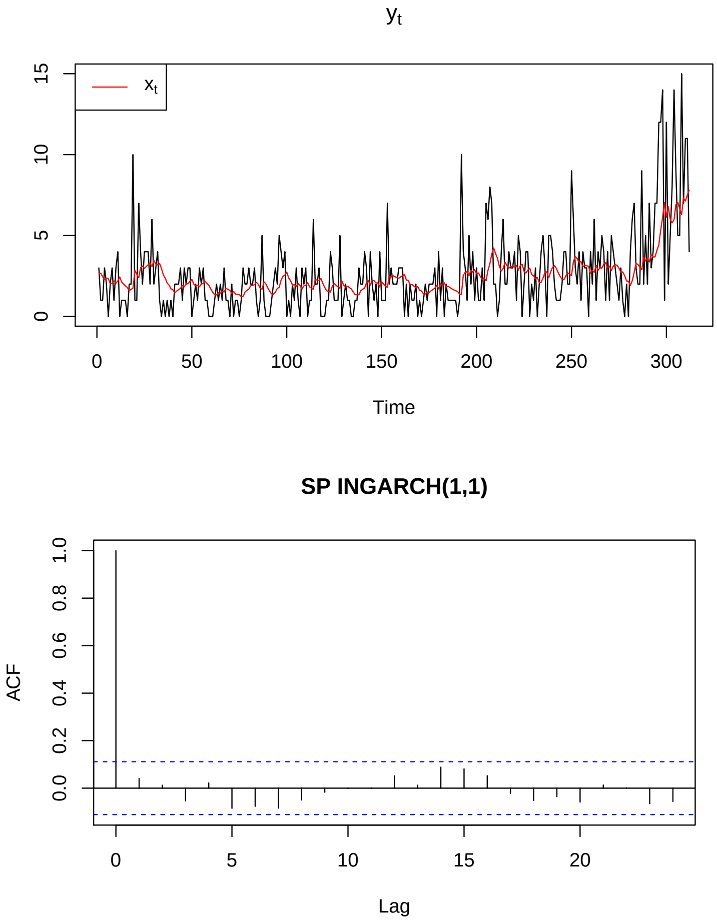

Figure 4.15: Top: The black line is the weekly number of homicides, and the red line is the fitted conditional mean process by SP INGARCH$(1,1)$ model; Bottom: ACF of the standardized Pearson residuals of the SP INGARCH$(1,1)$ for homicides data.

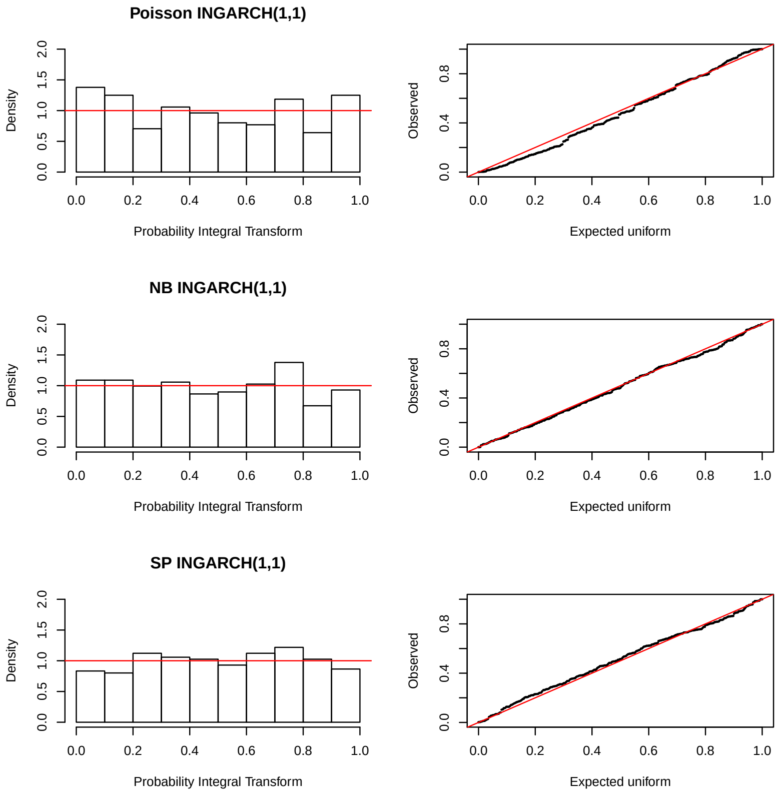

Figure 4.16: Randomized PIT histograms applied to the homicides data; Left: histograms of randomized PIT; Right: QQ-plots of corresponding randomized PIT against the uniform distribution.

4. Breech births data

The left plot of Figure 4.17 shows the number of monthly breech births in Edendale hospital of Pietermaritzburg in South Africa from February 1977 to January 1986. This count time series is of length 108 and is reported by Zucchini and MacDonald (2009). It can be downloaded from the website http://www.hmms-for-time-series.de/second/data. Although the sample ACF of the data (see the right plot of Figure 4.17) shows a reduced degree of autocorrelation between successive observations, we can apply the same analysis as in the previous examples. The data are overdispersed since the sample mean is 18.18 and the sample variance is 62.24. INGARCH(1,1) models based on the Poisson distribution, the negative binomial distribution and the semiparametric approach are considered. The estimates of the three fitted models are reported in Table 4.8. The data analysis in this case shows again that the semiparametric model is superior to the Poisson and negative binomial

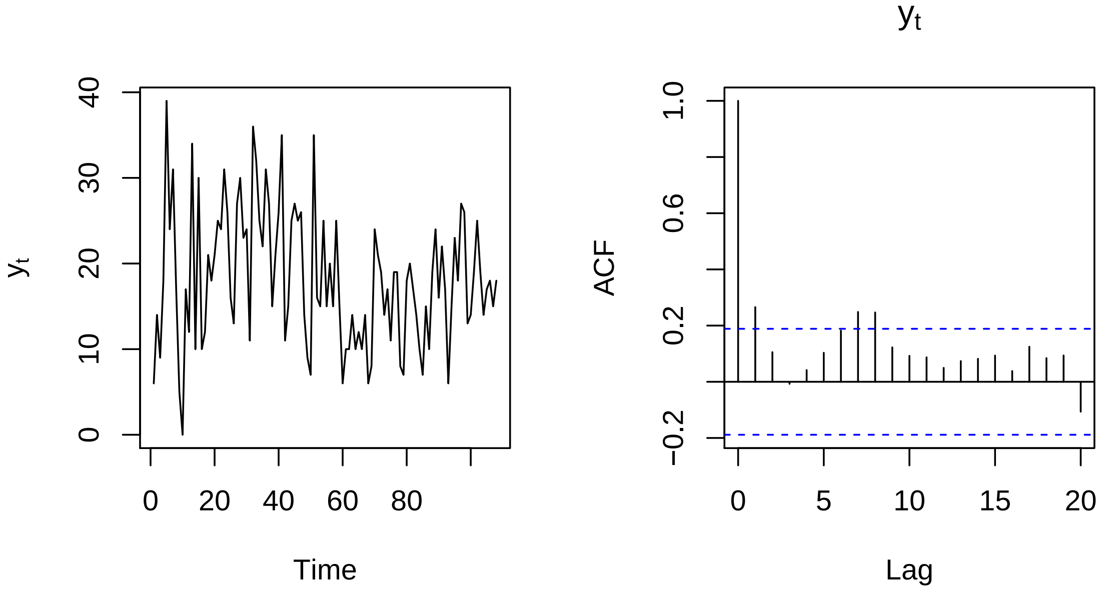

Figure 4.17: Left: Number of monthly breech births in Edendale hospital of Pietermaritzburg in South Africa from February 1977 to January 1986; Right: Sample ACF of the data

based models.

	δ	α	β
Poisson INGARCH$(1,1)$	12.0498	0.0754	0.2675
NB INGARCH$(1,1)$	12.3370	0.0537	0.2736
SP INGARCH$(1,1)$	12.0999	0.0706	0.2636

Table 4.8: Estimates of the INGARCH$(1,1)$ models for breech births data.

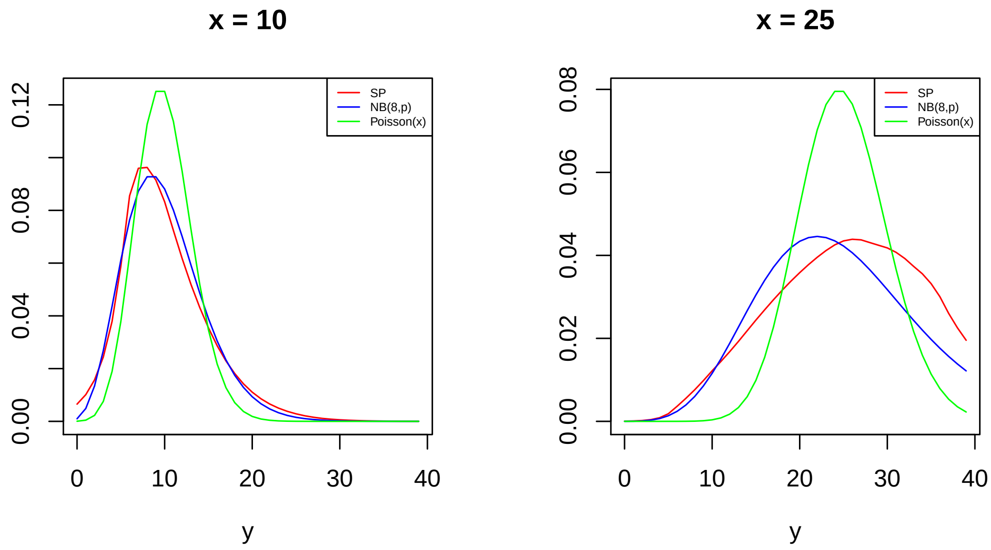

Figure 4.18: The red line is the estimated conditional distribution $p\left(y \mid B_{\hat{\varphi}}^{-1}(x), \hat{\varphi}\right)$ for the breech births data, the blue line is the pmf of NB $(8, \frac{8}{8+x})$ and the green line is the pmf of Poisson(x); x denote the mean of the exponential family; Left: $x = 10$; Right: $x = 25$.

Model	log likelihood	$p-$value of PIT	LS	QS	RPS
Poisson INGARCH	-425.2814	0.0006	3.9694	-0.0236	4.5321
NB INGARCH(1,1)	-372.1538	0.8901	3.4554	-0.0372	4.2706
SP INGARCH(1,1)	-358.5313	0.9893	3.3266	-0.0396	4.2563

Table 4.9: Quantitative model checking for breech births data

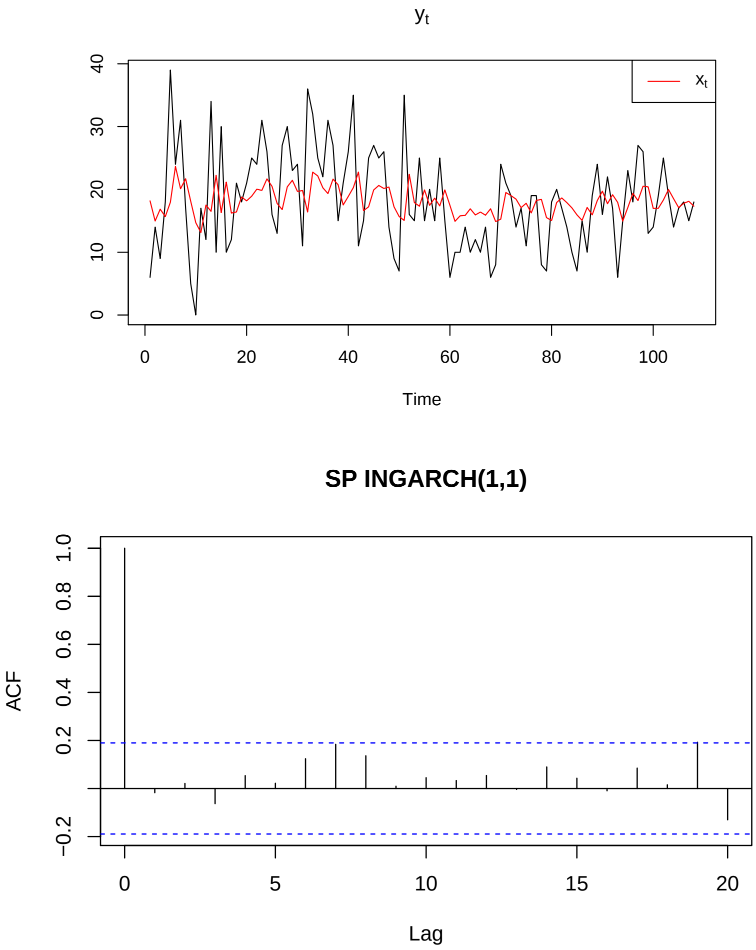

Figure 4.19: Top: The black line is the number of monthly breech births, and the red line is the fitted conditional mean process by SP INGARCH$(1, 1)$ model; Bottom: ACF of the standardized Pearson residuals of SP INGARCH$(1, 1)$ model for breech births data.

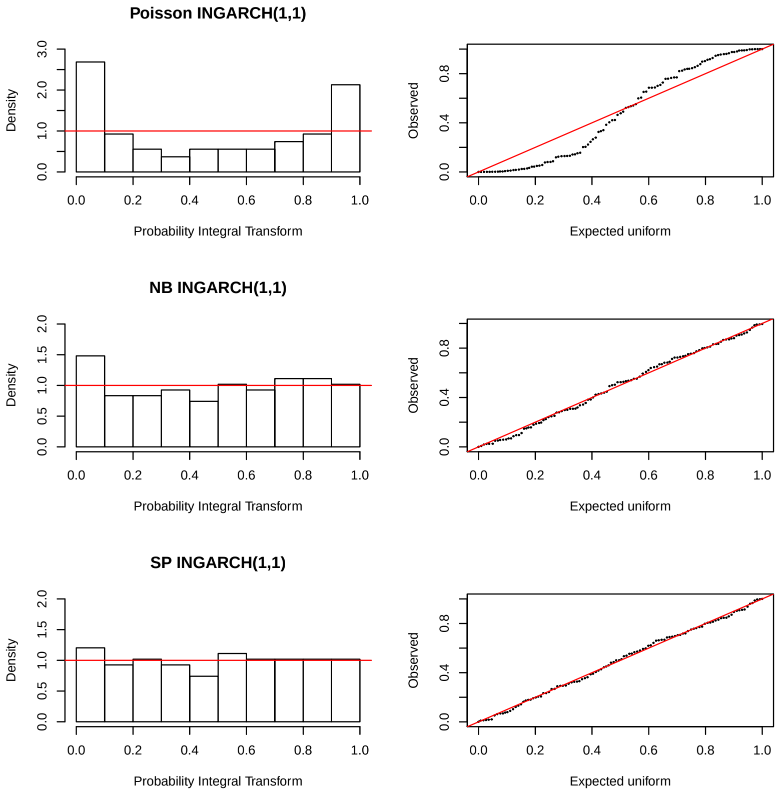

Figure 4.20: Randomized PIT histograms applied to the breech births data; Left: histograms of randomized PIT; Right: QQ-plots of corresponding randomized PIT against uniform distribution.

4.6 Appendix

First we state some relevant results in Dümbgen et al. (2011) that have important applications in our semiparametric estimation framework. Let $\mathcal{P}$ denote the collection of non-degenerate probability measures on the real line with finite first moment. The convex support of the distribution $Q \in \mathcal{P}$ is defined as

$$\operatorname{csupp}(Q) \coloneqq \cap\{C : C \subset \mathbb{R} \text{ closed and convex}, Q(C) = 1\}.$$

Define

$$h(Q, x) \coloneqq \sup\{Q(C) : C \subset \mathbb{R} \text{ closed and convex, } x \notin \text{ interior}(C)\}.$$

Dümbgen et al. (2011) have given a nice introduction to the log-concave density class $\mathcal{H}$ and characterized some interesting properties of $\operatorname{csupp}(Q)$ and $h(Q, x)$ for a general distribution function $Q \in \mathcal{P}$; see Lemmas 4.6.1, 4.6.2, and 4.6.3.

Lemma 4.6.1. *A point $x \in \mathbb{R}$ is an interior point of the convex support of $Q \in \mathcal{P}$ if and only if $h(Q, x) < 1$.*

Lemma 4.6.2. *Suppose that a sequence $Q_n \in \mathcal{P}$ converges weakly to $Q \in \mathcal{P}$. Then*

$$\limsup_{n\to\infty} h(Q_n, x) \leqslant h(Q, x) \textit{ for every } x \in \mathbb{R}.$$

Lemma 4.6.3. *For function $\varphi \in \mathcal{H}$, we have*

$$\varphi(x) \geqslant -\frac{\max(N, 0) - \int \varphi \, d\,Q}{1 - h(Q, x)},$$

where $N \coloneqq \sup_{x\in\mathbb{R}} \varphi(x)$.

Proof. It follows from the proof of Theorem 2.2 in Dümbgen et al. (2011). □

Proposition 4.6.4. *Suppose that a concave sequence $\{\varphi_n \in \mathcal{H}\}$ converges pointwise to some function φ. Then $A_{\varphi_n}(\eta)$ and $B_{\varphi_n}(\eta)$ converge to $A_\varphi(\eta)$ and $B_\varphi(\eta)$ on any compact subset of $\{\eta : A_\varphi(\eta) < \infty\}$, respectively. Moreover, the inverse function of $B_{\varphi_n}(\eta)$, $B^{-1}_{\varphi_n}(x)$, converges to $B^{-1}_\varphi(x)$ uniformly on any compact set of $\mathcal{R}(B_\varphi)$.*

Proof. Due to the concavity, $\{\varphi_n \in \mathcal{H}\}$ converges uniformly to φ on any compact subset of $\{y : \varphi(y) > -\infty\}$. For any $\eta \in \{\eta : A_\varphi(\eta) < \infty\}$, the sequence $\{\varphi_n(y) + \eta y\}$ converges uniformly to $\varphi(y)+\eta y$ on any compact subset of $\{y : \varphi(y) > -\infty\}$ as well. Since $A_\varphi(\eta) < \infty$, $\sup_{y\in\mathcal{R}} (\varphi(y) + \eta y)$ is bounded. As a result, there exist two constants a and $b > 0$ such that

$$\varphi_n(y) + \eta y \leqslant a - by \text{ for all } n \text{ and } y \in \mathbb{R}^+.$$

Recall

$$A_{\varphi_n}(\eta) = \log \sum_{y=0}^{\infty} \exp\left(\varphi_n(y) + \eta y\right)$$

and

$$B_{\varphi_n}(\eta) = A'_{\varphi_n}(\eta) = \frac{\sum_{y=0}^{\infty} y \exp\left(\varphi_n(y) + \eta y\right)}{\exp\left(A_{\varphi_n}(\eta)\right)}.$$

By the dominated convergence theorem, we have $A_{\varphi_n}(\eta) \to A_\varphi(\eta)$ and $B_{\varphi_n}(\eta) \to B_\varphi(\eta)$ as $n \to \infty$ for any $\eta \in \{\eta : A_\varphi(\eta) < \infty\}$. Note that $A_{\varphi_n}(\eta)$ and $B_{\varphi_n}(\eta)$ are strictly increasing functions for n large and their pointwise limits, $A_\varphi(\eta)$ and $B_\varphi(\eta)$, are continuous functions. Thus the sequences $A_{\varphi_n}(\eta)$ and $B_{\varphi_n}(\eta)$ converge to $A_\varphi(\eta)$ and $B_\varphi(\eta)$ uniformly on any compact subset of $\{\eta : A_\varphi(\eta) < \infty\}$, respectively. Furthermore, it's easy to see that the inverse function $B^{-1}_{\varphi_n}(x)$ converges to $B^{-1}_\varphi(x)$ pointwise on $\mathcal{R}(B_\varphi)$. The uniform convergence of $B^{-1}_{\varphi_n}(x)$ on any compact set of $\mathcal{R}(B_\varphi)$ follows from the fact that $B^{-1}_{\varphi_n}(x)$ are increasing functions and the limit $B^{-1}_\varphi(x)$ is continuous. □

Proof of Proposition 4.3.1. Let Q_n be the empirical distribution $\frac{1}{n}\sum_{t=1}^{n}\delta_{Y_t}$. For any point $y \in$ interior(cuspp(Q_n)) and $\varphi = (\varphi_1, \ldots, \varphi_K) \in \mathcal{G}_n$, by virtue of Lemma 4.6.3, we have

$$\varphi(y) \geqslant -\frac{\max(M_1, 0) - \int \varphi \, d\, Q_n}{1 - h(Q_n, y)}.$$

By construction, $\int \varphi \, d\, Q_n \leqslant -M_2$ and the empirical distribution $Q_n \in \mathcal{P}$ with probability one as $n \to \infty$. Therefore, the vector $\varphi = (\varphi_1, \ldots, \varphi_K)$ belongs to a compact set

$$\left[-\max_{y \in \{Y_1, \ldots, Y_n\}} \frac{\max(M_1, 0) + M_2}{1 - h(Q_n, y)}, M_1\right]^K.$$

The pointwise limit of a sequence of concave functions preserves concavity and the limit also satisfies assumptions (G1-G3) by Proposition 4.6.4. As a result, $\mathcal{G}_n$ is a compact subset of $\mathbb{R}^K$. Then the continuous function $l_n(\theta, \varphi)$ attains its maximum on $\Theta \times \mathcal{G}_n$ at some point $(\hat{\theta}_n, \hat{\varphi}_n) \in \Theta \times \mathcal{G}_n$. □

Proof of Proposition 4.4.1. For fixed $\varphi \in \mathcal{H}_{\varphi_0}$, $p(y \mid \eta, \varphi)$ belongs to the natural exponential family. Recall $\eta_t(\theta)$ satisfies the equation

$$X_t(\theta) = B_\varphi\left(\eta_t(\theta)\right).$$

Therefore, we have $\eta_t(\theta) = B_\varphi^{-1}\left(X_t(\theta)\right)$ and

$$\begin{aligned} l(\theta, \varphi) &= E\left[\varphi(Y_t) + \eta_t(\theta) Y_t - A_\varphi\left(\eta_t(\theta)\right)\right] \\ &= E\left[\varphi(Y_t) + B_\varphi^{-1}\left(X_t(\theta)\right) Y_t - A_\varphi\left(B_\varphi^{-1}\left(X_t(\theta)\right)\right)\right]. \end{aligned}$$

The expectation is taken with respect to the stationary measure of the process $\{X_t, Y_t\}$.

Also,

$$\begin{aligned}
l(\theta,\varphi)-l(\theta_0,\varphi) &= E\left[Y_t\left(B_\varphi^{-1}\left(X_t(\theta)\right)-B_\varphi^{-1}\left(X_t(\theta_0)\right)\right)\right.\\
&\quad \left.-\left(A_\varphi\left(B_\varphi^{-1}\left(X_t(\theta)\right)\right)-A_\varphi\left(B_\varphi^{-1}\left(X_t(\theta_0)\right)\right)\right)\right]\\
&= E\left[X_t(\theta_0)\left(B_\varphi^{-1}\left(X_t(\theta)\right)-B_\varphi^{-1}\left(X_t(\theta_0)\right)\right)\right.\\
&\quad \left.-\left(A_\varphi\left(B_\varphi^{-1}\left(X_t(\theta)\right)\right)-A_\varphi\left(B_\varphi^{-1}\left(X_t(\theta_0)\right)\right)\right)\right]\\
&= \int_{\{X_t(\theta)\neq X_t(\theta_0)\}} X_t(\theta_0)\left(B_\varphi^{-1}\left(X_t(\theta)\right)-B_\varphi^{-1}\left(X_t(\theta_0)\right)\right)\\
&\quad -\left(A_\varphi\left(B_\varphi^{-1}\left(X_t(\theta)\right)\right)-A_\varphi\left(B_\varphi^{-1}\left(X_t(\theta_0)\right)\right)\right)dP_{\theta_0}.
\end{aligned}$$

The second equality is obtained by conditioning on $\mathcal{F}_{t-1}$. On the set $\{X_t(\theta)\neq X_t(\theta_0)\}$, it follows from the mean value theorem that there exists some $c_t(\theta)\in\mathbb{R}$ between $B_\varphi^{-1}\left(X_t(\theta)\right)$ and $B_\varphi^{-1}\left(X_t(\theta_0)\right)$ such that

$$A_\varphi\left(B_\varphi^{-1}\left(X_t(\theta)\right)\right)-A_\varphi\left(B_\varphi^{-1}\left(X_t(\theta_0)\right)\right)=B_\varphi\left(c_t(\theta)\right)\left(B_\varphi^{-1}\left(X_t(\theta)\right)-B_\varphi^{-1}\left(X_t(\theta_0)\right)\right).$$

Therefore,

$$l(\theta,\varphi)-l(\theta_0,\varphi)=\int_{\{X_t(\theta)\neq X_t(\theta_0)\}}\left(X_t(\theta_0)-B(c_t(\theta)\right)\left(B_\varphi^{-1}\left(X_t(\theta)\right)-B_\varphi^{-1}\left(X_t(\theta_0)\right)\right)dP_{\theta_0}.$$

Note that $c_t(\theta)$ is a random element that depends on $X_t(\theta)$ and $X_t(\theta_0)$. Since the function $B_\varphi(\eta)$ is strictly increasing, in the case that $X_t(\theta)>X_t(\theta_0)$, we have

$$\left(X_t(\theta_0)-B(c_t(\theta)\right)\left(B_\varphi^{-1}\left(X_t(\theta)\right)-B_\varphi^{-1}\left(X_t(\theta_0)\right)\right)<0.$$

The same result holds as well when $X_t(\theta)<X_t(\theta_0)$. Together with assumption (A2), we have $l(\theta,\varphi)-l(\theta_0,\varphi)<0$ for any $\theta\neq\theta_0$. □

Proposition 4.6.5. *For any $(\theta, \varphi) \in \Theta \times \mathcal{H}_{\varphi_0} \setminus (\theta_0, \varphi_0)$, $l(\theta, \varphi) < l(\theta_0, \varphi_0)$.*

Proof. Since $l(\theta, \varphi) < l(\theta_0, \varphi)$ for any $\varphi \in \mathcal{H}_{\varphi_0}$, we have

$$\begin{aligned}
\sup_{\theta \in \Theta, \varphi \in \mathcal{H}_{\varphi_0}} l(\theta, \varphi) &= \sup_{\varphi \in \mathcal{H}_{\varphi_0}} l(\theta_0, \varphi) \\
&= \sup_{\varphi \in \mathcal{H}_{\varphi_0}} E\left[\log p\left(Y_t \mid B_{\varphi}^{-1}(X_t), \varphi\right)\right] \\
&= \sup_{\varphi \in \mathcal{H}_{\varphi_0}} E\left[E\left(\log p\left(Y_t \mid B_{\varphi}^{-1}(X_t), \varphi\right) \mid X_t\right)\right] \\
&\leqslant E\left[E\left(\log p\left(Y_t \mid B_{\varphi_0}^{-1}(X_t), \varphi_0\right) \mid X_t\right)\right] \\
&= E\left[\log p\left(Y_t \mid B_{\varphi_0}^{-1}(X_t), \varphi_0\right)\right]. \\
&= l(\theta_0, \varphi_0),
\end{aligned}$$

where the inequality follows from the non-negativity of Kullback-Leibler divergence. Equality holds if and only if $p\left(y \mid B_{\varphi}^{-1}(x), \varphi\right) = p\left(y \mid B_{\varphi_0}^{-1}(x), \varphi_0\right)$ on $\mathcal{X}$, which implies $\varphi = \varphi_0$ almost everywhere. Therefore, (θ_0, φ_0) is the unique maximizer of $l(\theta, \varphi)$ over $\Theta \times \mathcal{H}_{\varphi_0}$. □

Proposition 4.6.6. *Assume $\{\varphi_n \in \mathcal{G}_n\}$ converges pointwise to some $\varphi \in \mathcal{H}$. Suppose that η is bounded below by some $\eta^* \in \{A_\varphi(\eta) < \infty\}$ and $A_\varphi''(\eta)$ is bounded away from 0 on $[\eta^*, \infty) \cap \{\eta : A_\varphi(\eta) < \infty\}$. Then there exists some constant $c \in \mathbb{R}^+$ such that*

$$A_{\varphi_n}''(\eta) \geqslant c > 0 \textit{ for all } n \textit{ large and } \eta \in [\eta^*, \infty) \cap \{\eta : A_\varphi(\eta) < \infty\}.$$

Moreover,

$$B_{\varphi_n}^{-1}(x) \leqslant B_{\varphi_n}^{-1}(x^*) + \frac{1}{c}(x - x^*).$$

Proof. Since $A_\varphi(\eta)$ is convex, the set $[\eta^*, \infty) \cap \{\eta : A_\varphi(\eta) < \infty\}$ is an interval $[\eta^*, \eta_1)$, where $\eta_1 = \sup\{\eta : A(\eta) < \infty\}$ and η_1 can be $+\infty$. Let $U_{n,\eta}$ be a random variable distributed as

$p(\cdot \mid \eta, \varphi_n)$. Then we have $A''_{\varphi_n}(\eta) = Var(U_{n,\eta})$. If there exists a sequence $\{\eta_n\}$ such that $Var(U_{n,\eta_n}) \to 0$ as $n \to \infty$, we consider the two cases:

Case I: There exists a subsequence $\{\eta_{n_k}\}$ of $\{\eta_n\}$ that converges to some $\tilde{\eta} \in [\eta^*, \eta_1)$. Since $p(\cdot \mid \eta_{n_k}, \varphi_{n_k})$ converges to $p(\cdot \mid \tilde{\eta}, \varphi)$ pointwise, we can find $a, b \in \mathbb{R}$ with $b > 0$ such that the concave sequence $\log p(y \mid \eta_{n_k}, \varphi_{n_k})$ is uniformly bounded above by $a - by$. Hence, by the dominated convergence theorem, we have $Var(U_{n_k,\eta_{n_k}}) \to Var(U) = 0$, where U is a random variable distributed as $p(\cdot \mid \tilde{\eta}, \varphi)$. However, the distribution $p(y \mid \tilde{\eta}, \varphi)$ is non-degenerate and $Var(U) = A''_{\varphi}(\tilde{\eta}) > 0$, which is a contradiction.

Case II: There exists a subsequence $\{\eta_{n_k}\}$ of $\{\eta_n\}$ that converges to η_1. For notation convenience, let $\{\eta_n\}$ itself denote the subsequence $\{\eta_{n_k}\}$. Since $B^{(3)}_{\varphi_n}(\eta) = E\left(U_{n,\eta} - EU_{n,\eta}\right)^4 \geqslant 0$, $B''_{\varphi_n}(\eta) = E\left(U_{n,\eta} - EU_{n,\eta}\right)^3$ is a convex function and it converges pointwise to $B''_{\varphi}(\eta)$ on $\{\eta : A_{\varphi}(\eta) < \infty\}$ by the dominated convergence theorem. Thus there exists $\eta_2 \in \{\eta : A_{\varphi}(\eta) < \infty\} < \eta_1$ such that $B''_{\varphi_n}(\eta)$ and $B''_{\varphi}(\eta)$ do not change sign on $[\eta_2, \eta_1)$ for all n large. Then we can see that $B'_{\varphi_n}(\eta)$ and $B'_{\varphi}(\eta)$ are decreasing functions on $[\eta_2, \eta_1)$ for all n large since $\lim_{n\to\infty} B'_{\varphi_n}(\eta_n) = Var(U_{n,\eta_n}) = 0$, which implies

$$\lim_{n\to\infty} B'_{\varphi}(\eta_n) = \lim_{n\to\infty} B'_{\varphi_n}(\eta_n) = 0.$$

This contradicts the assumption that $A''_{\varphi}(\eta)$ is bounded away from 0 on $[\eta^*, \eta_1)$.

From Cases I and II, we conclude that there exists $c > 0$ such that $A''_{\varphi_n}(\eta) \geqslant c > 0$ for all n large and $\eta \in [\eta^*, \eta_1)$. In addition,

$$\left(B^{-1}_{\varphi_n}\right)'(x) = \frac{1}{B'_{\varphi_n}\left(B^{-1}_{\varphi_n}(x)\right)} \leqslant \frac{1}{c}, \text{ for all } n \text{ and } x \in \mathcal{R}(B_{\varphi}).$$

It then follows that

$$\begin{aligned} B_{\varphi_n}^{-1}(x) &= B_{\varphi_n}^{-1}(x^*) + \int_{x^*}^{x} \left(B_{\varphi_n}^{-1}\right)'(v)dv \\ &\leqslant B_{\varphi_n}^{-1}(x^*) + \frac{1}{c}(x - x^*). \end{aligned}$$

□

The bounded Lipschitz distance between two probability measures P and Q is defined as

$$D_{BL}(P,Q) := \sup_{\|f\|_\infty \leqslant 1, \|f\|_L \leqslant 1} |\int f d(P-Q)|,$$

with $\|f\|_\infty := \sup_x |f(x)|$ and $\|f\|_L = \sup_{x \neq y} |f(x) - f(y)|/|x-y|$. It is well-known that the bounded Lipschitz distance metrizes the weak convergence of probability measures (Pollard, 1984), that is, a sequence of probability measure Q_n converges weakly to some probability measure Q if and only if $\lim_{n\to\infty} D_{BL}(Q_n, Q) = 0$. Let $P_n(\theta)$ and $P(\theta)$ denote the empirical measure $\frac{1}{n}\sum_{t=1}^{n} \delta_{(X_t(\theta), Y_t)^T}$ and the stationary measure of $(X_t(\theta), Y_t)^T$, respectively. By Theorem 2.2 in Berti et al. (2006), the sequence of random measures $P_n(\theta)$ converges weakly to $P(\theta)$ almost surely for each $\theta \in \Theta$, i.e.,

$$D_{BL}\left(P_n(\theta), P(\theta)\right) \xrightarrow{a.s.} 0 \text{ as } n \to \infty.$$

In fact, $\{P_n(\theta)\}_\theta$ admits a stronger convergence result as shown in Proposition 4.6.7.

Proposition 4.6.7.

$$\sup_{\theta \in \Theta} D_{BL}\left(P_n(\theta), P(\theta)\right) \xrightarrow{a.s.} 0 \ \textit{as}\ n \to \infty.$$

Proof. For any $\theta \in \Theta$ and any bounded Lipschitz function f on $\mathbb{R}^2$ with $\|f\|_\infty \leqslant 1$ and

$\|f\|_L \leqslant 1$, by inequality (4.8), we have

$$\begin{aligned}|f\left(g_\theta(x,y),z\right)-f\left(g_\theta(x',y'),z'\right)| &\leqslant |g_\theta(x,y)-g_\theta(x',y')|+|z-z'| \\ &\leqslant a|x-x'|+b|y-y'|+|z-z'| \\ &\leqslant 2^{\frac{1}{2}}\sqrt{(x-x')^2+(y-y')^2+(z-z')^2},\end{aligned}$$

for any $x, x' \geqslant 0$ and $y, y', z, z' \in \mathcal{X}$, which implies the function $f_\theta(x,y,z) \coloneqq f\left(g_\theta(x,y),z\right)$ is a bounded Lipschitz function on $\mathbb{R}^+ \times \mathcal{X} \times \mathcal{X}$ with $\|f_\theta\|_L \leqslant \sqrt{2}$. Note that

$$f\left(X_t(\theta),Y_t\right)=f_\theta(X_{t-1},Y_{t-1},Y_t).$$

Hence,

$$\int f\, d\left(P_n(\theta)-P(\theta)\right)=\int f_\theta\, d\left(Q_n-Q\right),$$

where Q_n and Q denote the empirical measure $\frac{1}{n}\sum_{t=1}^n \delta_{(X_{t-1},Y_{t-1},Y_t)^T}$ and the stationary measure of the vector $(X_{t-1},Y_{t-1},Y_t)^T$, respectively. Then

$$\begin{aligned}\sup_{\theta\in\Theta,\|f\|_\infty\leqslant 1,\|f\|_L\leqslant 1}\left|\int f\, d\left(P_n(\theta)-P(\theta)\right)\right| &= \sup_{\theta\in\Theta,\|f_\theta\|_\infty\leqslant 1,\|f_\theta\|_L\leqslant\sqrt{2}}\left|\int f_\theta\, d\left(Q_n-Q\right)\right| \\ &\leqslant \sqrt{2}D_{BL}(Q_n,Q),\end{aligned}$$

which implies

$$\sup_{\theta\in\Theta} D_{BL}\left(P_n(\theta),P(\theta)\right)\xrightarrow{a.s.} 0 \text{ as } n\to\infty.$$

□

Proposition 4.6.8.

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta,\varphi\in\mathcal{G}_n} l_n(\theta,\varphi) \leqslant \sup_{\theta\in\Theta,\varphi\in\mathcal{H}_{\varphi_0}} l(\theta,\varphi)\, a.s.. \tag{4.16}$$

Proof. First we follow the argument of proof of Theorem 4.5 in Dümbgen et al. (2011) to show a compactness property of log-concave density sequence. Denote Ω' as the set $\{\omega\in\Omega : \lim_{n\to\infty}\sup_{\theta\in\Theta} D_{BL}\left(P_n(\theta), P(\theta)\right) = 0\}$. For each $\omega\in\Omega'$, the empirical measure $Q_n := \frac{1}{n}\sum_{t=1}^{n}\delta_{Y_t}$ converges weakly to the stationary distribution of Y_t, which is denoted as Q. Thus, by Lemma 4.6.2, we have

$$\limsup_{n\to\infty} h\left(Q_n, y\right) \leqslant h\left(Q, y\right) < 1$$

for any point $y\in$ interior(cuspp$\,(Q)$). According to Lemma 4.6.3, for any sequence $\{\varphi_n \in \mathcal{G}_n\}_n$, we have

$$\varphi_n(y) \geqslant -\frac{\max(N_n, 0) - \int \varphi_n \, d\, Q_n}{1 - h(Q_n, y)},$$

where $N_n := \sup_{y\in\mathbb{R}}\varphi_n(y)$. The sequence N_n is uniformly bounded above by M_1 and the integral $\int\varphi_n\, d\, Q_n$ is uniformly bounded below by $-M_2$. As a result,

$$\inf_{n\geqslant 1}\varphi_n(y) \geqslant -\frac{M_1 + M_2}{1 - h(Q, y)} > -\infty$$

for any $y\in$ interior(cuspp(Q)). In addition, since the sequence N_n is uniformly bounded, we deduce from the proof of Theorem 2.2 in Dümbgen et al. (2011) that there exist two constants a and b with $b > 0$ such that

$$\varphi_n(y) \leqslant a - by \text{ for all } n \text{ and } y\in\mathbb{R}.$$

Then according to Lemma 4.2 in Dümbgen et al. (2011), there exists a subsequence $\{\varphi_{n_k}\}_k$ of $\{\varphi_n\}_n$ which converges to some concave function ψ pointwise on interior(cuspp(Q)). In particular, interior(cuspp(Q)) $\subset \{y : \psi(y) > -\infty\}$, $\int e^{\psi(y)}\, dy = 1$ and $\int ye^{\psi(y)}\, d\, y = 1$ by the dominated convergence theorem, which implies $\psi \in \mathcal{H}_{\varphi_0}$. For notation convenience, let $\{\varphi_n\}_n$ be the subsequence that converges to the function ψ. Due to the concavity, the sequence $\{\varphi_n\}_n$ converges to ψ uniformly on any compact subset of $\{y : \psi(y) > -\infty\}$. By virtue of Proposition 4.6.4, the sequences $A_{\varphi_n}(\eta)$ and $B_{\varphi_n}(\eta)$ converge to $A_\psi(\eta)$ and $B_\psi(\eta)$ uniformly on any compact subset of $\{\eta \in \mathbb{R} : A_\psi(\eta) < \infty\}$, respectively. The inverse function $B^{-1}_{\varphi_n}(x)$ converges to $B^{-1}_\psi(x)$ uniformly on any compact set of $\mathcal{R}(B_\psi)$ as well.

Let $\{\theta_n \in \Theta\}_n$ be any sequence that converges to some $\theta^* \in \Theta$. Then $P_n(\theta_n)$ converges weakly to $P(\theta^*)$ for the fixed $\omega \in \Omega'$. We follow the same argument of the proof for Theorem 4.5 in Dümbgen et al. (2011). By Skorohod's representation theorem (Billingsley, 2013), there exists a probability space $(\Omega'', \mathcal{A}, \mathbb{P})$ with bivariate random vectors $Z_n = (Z_{n,1}, Z_{n,2})^T \sim P_n(\theta_n)$ and $Z = (Z_1, Z_2) \sim P(\theta^*)$ such that Z_n converges to Z almost surely. Since $A_{\varphi_n}\left(B^{-1}_{\varphi_n}(x)\right)$ is an increasing function, we have

$$A_{\varphi_n}(B^{-1}_{\varphi_n}(Z_{n,1})) - A_{\varphi_n}\left(B^{-1}_{\varphi_n}(x^*)\right) \geqslant 0.$$

By Fatou's lemma,

$$\liminf_{n\to\infty} E\left[A_{\varphi_n}(B^{-1}_{\varphi_n}(Z_{n,1})) - A_{\varphi_n}\left(B^{-1}_{\varphi_n}(x^*)\right)\right] \geqslant E\left[A_\psi\left(B^{-1}_\psi(Z_1)\right) - A_\psi\left(B^{-1}_\psi(x^*)\right)\right],$$

which implies

$$\limsup_{n\to\infty} E\left[-A_{\varphi_n}(B^{-1}_{\varphi_n}(Z_{n,1}))\right] \leqslant -E\left[A_\psi\left(B^{-1}_\psi(Z_1)\right)\right]. \tag{4.17}$$

Note that $\varphi_n \in \mathcal{G}_n$ and hence by construction, we have

$$B_{\varphi_n}^{-1}(Z_{n,1}) \leqslant B_{\varphi_n}^{-1}(x^*) + M_3(Z_{n,1} - x^*).$$

Applying Fatou's lemma again,

$$\begin{aligned} \liminf_{n\to\infty} E\left[B_{\varphi_n}^{-1}(x^*)Z_{n,2} + M_3(Z_{n,1} - x^*)Z_{n,2} - B_{\varphi_n}^{-1}(Z_{n,1})Z_{n,2}\right] \\ \geqslant E\left[B_{\psi}^{-1}(x^*)Z_2 + M_3(Z_1 - x^*)Z_2 - B_{\psi}^{-1}(Z_1)Z_2\right]. \end{aligned}$$

Hence,

$$\limsup_{n\to\infty} E\left[B_{\varphi_n}^{-1}(Z_{n,1})Z_{n,2}\right] \leqslant E\left[B_{\psi}^{-1}(Z_1)Z_2\right]. \tag{4.18}$$

Inequalities (4.17) and (4.18) imply that

$$\limsup_{n\to\infty} E\left[\varphi_n(Z_2) + B_{\varphi_n}^{-1}(Z_{n,1})Z_{n,2} - A_{\varphi_n}\left(B_{\varphi_n}^{-1}(Z_{n,1})\right)\right] \leqslant l(\theta^*, \psi).$$

Therefore, for the fixed $\omega \in \Omega'$,

$$\begin{aligned} \limsup_{n\to\infty} l_n(\theta_n, \varphi_n) &\leqslant l(\theta^*, \psi) \\ &\leqslant \sup_{\theta\in\Theta, \varphi\in\mathcal{H}_{\varphi_0}} l(\theta, \varphi). \end{aligned}$$

Note that (θ_n, φ_n) are arbitrary sequences in $\Theta \times \mathcal{G}_n$ and $\{\omega : \omega \notin \Omega'\}$ is a null-set. Then the inequality (4.16) follows. □

Corollary 4.6.9. *On the set Ω', $\{\mathcal{G}_n\}_n$ is compact in the sense that any subsequence of $\{\varphi_n \in \mathcal{G}_n\}_n$ has a further subsequence with pointwise limit ψ belonging to $\mathcal{H}_{\varphi_0}$. In addition, let $\{\varphi_n \in \mathcal{G}_n\}_n$ be the convergent subsequence with limit ψ and $\{\theta_n \in \Theta\}_n$ be any sequence* *that converges to some $\theta^* \in \Theta$. Then*

$$\limsup_{n\to\infty} l_n(\theta_n, \varphi_n) \leqslant l(\theta^*, \psi).$$

Proof. See the proof of Proposition 4.6.8. □

Proposition 4.6.10. *$\varphi_0^n \in \mathcal{G}_n$ for large n a.s. and $l(\theta_0, \varphi_0) \leqslant \liminf_{n\to\infty} l_n(\theta_0, \varphi_0^n)$ a.s..*

Proof. It's easy to see that $\tilde{\varphi}_0^n$ converges pointwise to φ_0 *a.s.* and $\tilde{\varphi}_0^n \leqslant \varphi_0$. By the dominated convergence theorem, the sequences $A_{\tilde{\varphi}_0^n}(\eta)$ and $B_{\tilde{\varphi}_0^n}(\eta)$ converge pointwise to $A_{\varphi_0}(\eta)$ and $B_{\varphi_0}(\eta)$ *a.s.*, respectively. Since both $A_{\tilde{\varphi}_0^n}(\eta)$ and $B_{\tilde{\varphi}_0^n}(\eta)$ are increasing functions, we have $A_{\tilde{\varphi}_0^n}(\eta)$ and $B_{\tilde{\varphi}_0^n}(\eta)$ converge almost surely to $A_{\varphi_0}(\eta)$ and $B_{\varphi_0}(\eta)$ uniformly on any compact subset of $\{\eta \in \mathbb{R} : A_{\varphi_0}(\eta) < \infty\}$, respectively. The inverse functions $B^{-1}_{\tilde{\varphi}_0^n}(x)$ converge to $B^{-1}_{\varphi_0}(x)$ uniformly on any compact set of $\mathcal{R}(B_{\varphi_0})$ as well. Note that $\frac{\int_0^\infty y e^{\varphi_0(y)}\, dy}{\int_0^\infty e^{\varphi_0(y)}\, dy} = 1$. Therefore, the sequence η_n^*, which satisfies the equation $\frac{\int_0^\infty y e^{\tilde{\varphi}_0^n(y)+\eta y}\, dy}{\int_0^\infty e^{\tilde{\varphi}_0^n(y)+\eta y}\, dy} = 1$, converges to zero almost surely and $\int_0^\infty e^{\tilde{\varphi}_0^n(y)+\eta_n^* y}\, dy \xrightarrow{a.s.} 1$. Hence φ_0^n converges pointwise to φ_0 almost surely. According to assumption (A5), $X_t \geqslant x^* \in \mathcal{R}(B_{\varphi_0})$. Therefore $B^{-1}_{\varphi_0^n}(X_t) \geqslant B^{-1}_{\varphi_0^n}(x^*)$. It follows from Assumption (A7) and Proposition 4.6.6 that $\varphi_0^n \in \mathcal{G}_n$ for large n *a.s.*.

Recall that the empirical distribution $P_n(\theta_0) = \frac{1}{n}\sum_{t=1}^n \delta_{(X_t,Y_t)^T}$ converges to the stationary distribution of $(X_t, Y_t)^T$, $P(\theta_0)$, almost surely. We now apply the idea of the proof for Theorem 4.5 in Dümbgen et al. (2011) again. By Skorohod's representation theorem (Billingsley, 2013), there exists a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with bivariate random vectors $Z_n = (Z_{n,1}, Z_{n,2})^T \sim P_n(\theta_0)$ and $Z = (Z_1, Z_2)^T \sim P(\theta_0)$ such that Z_n converges to Z almost surely. Note that $B^{-1}_{\varphi_0^n}(X_t)Y_t - B^{-1}_{\varphi_0^n}(x^*)Y_t \geqslant 0$, and hence by Fatou's lemma,

$$\begin{aligned}
\liminf_{n\to\infty} \frac{1}{n}\sum_{t=1}^n \left[B^{-1}_{\varphi_0^n}(X_t)Y_t - B^{-1}_{\varphi_0^n}(x^*)Y_t\right] &= \liminf_{n\to\infty} E\left[B^{-1}_{\varphi_0^n}(Z_{n,1})Z_{n,2} - B^{-1}_{\varphi_0^n}(x^*)Z_{n,2}\right] \\
&\geqslant E\left[B^{-1}_{\varphi_0}(Z_1)Z_2 - B^{-1}_{\varphi_0}(x^*)Z_2\right],
\end{aligned}$$

which implies

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} B_{\varphi_0^n}^{-1}(X_t) Y_t \geqslant E\left[B_{\varphi_0}^{-1}(Z_1) Z_2\right]. \tag{4.19}$$

For $x \in \mathcal{R}(B_{\varphi_0})$, the derivative of the function $A_{\varphi_0^n}\left(B_{\varphi_0^n}^{-1}(x)\right)$ with respect to x is given by

$$\begin{aligned} \frac{dA_{\varphi_0^n}\left(B_{\varphi_0^n}^{-1}(x)\right)}{dx} &= B_{\varphi_0^n}\left(B_{\varphi_0^n}^{-1}(x)\right)\left(B_{\varphi_0^n}^{-1}\right)'(x) \\ &= \frac{x}{B'_{\varphi_0^n}\left(\left(B_{\varphi_0^n}^{-1}\right)(x)\right)} \end{aligned}$$

According to assumption (A7) and Proposition 4.6.6, there exists some $c > 0$ such that $B'_{\varphi_0^n}(\eta) \geqslant c$ for n large and $\eta \in [B_{\varphi_0}^{-1}(x^*), \infty) \cap \{A_{\varphi_0}(\eta) < \infty\}$. Therefore

$$\frac{dA_{\varphi_0^n}\left(B_{\varphi_0^n}^{-1}(x)\right)}{dx} \leqslant \frac{x}{c} \text{ for all } n \text{ and } x \in \mathcal{R}(B_{\varphi_0}).$$

It then follows that

$$\begin{aligned} A_{\varphi_0^n}\left(B_{\varphi_0^n}^{-1}(x)\right) &= A_{\varphi_0^n}^{-1}\left(B_{\varphi_0^n}^{-1}(x^*)\right) + \int_{x^*}^{x} \frac{dA_{\varphi_0^n}\left(B_{\varphi_0^n}^{-1}(v)\right)}{dv} dv \\ &\leqslant A_{\varphi_0^n}^{-1}\left(B_{\varphi_0^n}^{-1}(x^*)\right) + \frac{1}{2c}\left(x^2 - (x^*)^2\right). \end{aligned}$$

Now applying Fatou's lemma again,

$$\begin{aligned} &\liminf_{n\to\infty} E\left[A_{\varphi_0^n}^{-1}\left(B_{\varphi_0^n}^{-1}(x^*)\right) + \frac{1}{2c}\left(Z_{n,1}^2 - (x^*)^2\right) - A_{\varphi_0^n}\left(B_{\varphi_0^n}^{-1}(Z_{n,1})\right)\right] \\ &\qquad \geqslant E\left[A_{\varphi_0}^{-1}\left(B_{\varphi_0}^{-1}(x^*)\right) + \frac{1}{2c}\left(Z_1^2 - (x^*)^2\right) - A_{\varphi_0}\left(B_{\varphi_0}^{-1}(Z_1)\right)\right]. \end{aligned} \tag{4.20}$$

Note that $EX_t^2 < \infty$ by assumption (A8), and hence

$$\begin{aligned} \liminf_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} -A_{\varphi_0^n}\left(B_{\varphi_0^n}^{-1}(X_t)\right) &= \liminf_{n\to\infty} E\left[-A_{\varphi_0^n}\left(B_{\varphi_0^n}^{-1}(Z_{n,1})\right)\right] \\ &\geqslant E\left[-A_{\varphi_0}\left(B_{\varphi_0}^{-1}(Z_1)\right)\right]. \qquad (4.21) \end{aligned}$$

Combining (4.19) and (4.21), we have

$$l(\theta_0, \varphi_0) \leqslant \liminf_{n\to\infty} l_n(\theta_0, \varphi_0^n)\, a.s..$$

□

Chapter 5

Conclusions and Future Work

This thesis focuses on developing applications of shape constraint estimation for time series models. First, we propose a semiparametric estimation procedure for non-Gaussian non-minimum phase ARMA models using log-concave projection, where the underlying noise distribution can be learned simultaneously. The consistency of the semiparametric MLE follows from the properties of log-concave projection and the Entropy Power Inequality. We obtain a lower bound for the best asymptotic variance of regular estimators at rate $\sqrt{n}$ for AR models and construct a semiparametric efficient estimator. Extension of this estimation procedure to causal VAR models is also considered. Second, we take advantage of the exponential distribution family assumption made in many time series of counts models and propose a semiparametric GLM framework for modeling count time series by incorporating an infinite dimensional function parameter to the exponential distribution family.

Future directions of this research:

1. As stated in Remark 5, we conjecture that the MLCLE $\hat{\beta}$ is semiparametric efficient with asymptotic variance given by the inverse efficient information matrix. Appendix 2.6.2 contains the current work in studying the asymptotic properties of the MLCLE

$\hat{\beta}$. We need to control the complexity of the function class $\{\tilde{l}_{\beta,f} : \beta \in \Theta, f \in \mathcal{F}\}$ and show it is of finite VC dimension in order to apply empirical process theory for dependent data and to complete the proof of the asymptotic efficiency of the MLCLE.

2. Chapter 3 generalizes the semiparametric estimation method to causal VAR models and only shows the consistency of the MLCLE. We hope to extend the estimation procedure to noncausal VAR models, for which we are only able to show that the true VAR coefficients are a local maximum of the limiting function.

3. Chapter 4 applies the one-parameter exponential family with a concave baseline function to the conditional mean models $X_t(\theta) = g_\theta(X_{t-1}, Y_{t-1})$ studied in Davis and Liu (2016). It may be worthwhile to generalize the extended one-parameter exponential family to other conditional mean models, such as INGARCH models with covariates.

4. In modeling time series of counts, we only show the consistency of the CMLE. It is interesting to study the asymptotic properties of the CMLE as well.

5. In Chapter 4, the built-in optimizing function in MATLAB is used to find the MLE of the concave baseline function. This step is very time consuming. The optimizing function available in MATLAB is very likely to find a suboptimal point. It would be worthwhile optimizing the implementation to run more efficiently.

Bibliography

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, B., R. A. Davis, and F. J. Breidt (2007). Rank-based estimation for all-pass time series models. *The Annals of Statistics*, 844–869.

Andrews, D. W. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, 43–72.

Arcones, M. A. and B. Yu (1994). Central limit theorems for empirical and U-processes of stationary mixing sequences. *Journal of Theoretical Probability 7*(1), 47–71.

Bagnoli, M. and T. Bergstrom (2006). Log-concave probability and its applications. In *Rationality and Equilibrium*, pp. 217–241. Springer.

Berti, P., L. Pratelli, and P. Rigo (2006). Almost sure weak convergence of random probability measures. *Stochastics and Stochastics Reports 78*(2), 91–97.

Bickel, P. J. and K. Doksum (2006). *Mathematical Statistics, Basic Ideas and Selected Topics.* Vol I, Prentice Hall, Saddle River, NJ.

Billingsley, P. (2013). *Convergence of probability measures.* John Wiley & Sons.

Birke, M. and H. Dette (2007). Estimating a convex function in nonparametric regression. *Scandinavian Journal of Statistics 34*(2), 384–404.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics 31*(3), 307–327.

Box, G. E., G. M. Jenkins, G. C. Reinsel, and G. M. Ljung (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Brännäs, K. and A. Quoreshi (2010). Integer-valued moving average modelling of the number of transactions in stocks. *Applied Financial Economics 20*(18), 1429–1440.

Breidt, F. J., R. A. Davis, K.-S. Lh, and M. Rosenblatt (1991). Maximum likelihood estimation for noncausal autoregressive processes. *Journal of Multivariate Analysis 36*(2), 175–198.

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.

Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics*, 607–616.

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 (2015a). Generalized additive and index models with shape constraints. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Chen, Y. and R. J. Samworth (2015b). Semiparametric time series models with log-concave innovations: maximum likelihood estimation and its consistency. *Scandinavian Journal of Statistics, 42, 1-31*.

Christou, V. and K. Fokianos (2014). Quasi-Likelihood Inference for Negative Binomial Time Series Models. *Journal of Time Series Analysis 35*(1), 55–78.

Cline, D. B. and P. J. Brockwell (1985). Linear prediction of ARMA processes with infinite variance. *Stochastic Processes and their Applications 19*(2), 281–296.

Cox, D. R., G. Gudmundsson, G. Lindgren, L. Bondesson, E. Harsaae, P. Laake, K. Juselius, and S. L. Lauritzen (1981). Statistical analysis of time series: Some recent developments [with discussion and reply]. *Scandinavian Journal of Statistics*, 93–115.

Cule, M., R. Gramacy, and R. Samworth (2009). LogConcDEAD: An R package for maximum likelihood estimation of a multivariate log-concave density. *Journal of Statistical Software 29.2*, 1–20.

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 multidimensional log-concave density. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(5), 545–607.

Czado, C., T. Gneiting, and L. Held (2009). Predictive model assessment for count data. *Biometrics 65*(4), 1254–1261.

Davis, R. A., W. T. Dunsmuir, and S. B. Streett (2003). Observation-driven models for Poisson counts. *Biometrika*, 777–790.

Davis, R. A., W. T. Dunsmuir, and W. Ying (1999). Modeling time series of count data. *Statistics Textbooks and Monographs 158*, 63–114.

Davis, R. A. and H. Liu (2016). Theory and inference for a class of observation-driven models with application to time series of counts. *Statistica Sinica 26*, 1673–1707.

Davis, R. A. and L. Song (2012). Noncausal Vector AR Processes with Application to Economic Time Series. *Working paper, Columbia University*.

Davis, R. A. and R. Wu (2009). A negative binomial model for time series of counts. *Biometrika*, 735–749.

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 (2009). Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli 15*(1), 40–68.

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.

Ferland, R., A. Latour, and D. Oraichi (2006). Integer-Valued GARCH Process. *Journal of Time Series Analysis 27*(6), 923–942.

Fokianos, K. (2001). Truncated Poisson regression for time series of counts. *Scandinavian journal of statistics 28*(4), 645–659.

Fokianos, K. (2015). Statistical Analysis of Count Time Series Models: A GLM Perspective. *Handbook of Discrete-Valued Time Series, Handbooks of Modern Statistical Methods*, 3–28.

Fokianos, K., A. Rahbek, and D. Tjøstheim (2009). Poisson autoregression. *Journal of the American Statistical Association 104*(488), 1430–1439.

Gassiat, E. (1993). Adaptive estimation in noncausal stationary AR processes. *The Annals of Statistics*, 2022–2042.

Grenander, U. (1956). On the theory of mortality measurement: part ii. *Scandinavian Actuarial Journal 1956*(2), 125–153.

Guikema, S. D. and J. P. Goffelt (2008). A flexible count data regression model for risk analysis. *Risk analysis 28*(1), 213–223.

Heinen, A. (2003). Modelling time series count data: an autoregressive conditional Poisson model.

Huang, J. and Y. Pawitan (2000). Quasi-likelihood Estimation of Non-invertible Moving Average Processes. *Scandinavian Journal of Statistics 27*(4), 689–702.

Ichimura, H. and S. Lee (2010). Characterization of the asymptotic distribution of semiparametric M-estimators. *Journal of Econometrics 159*(2), 252–266.

Jung, R. C. and A. Tremayne (2011). Useful models for time series of counts or simply wrong ones? *AStA Advances in Statistical Analysis 95*(1), 59–91.

Kantorović, L. V. and G. Ś. Rubinśteín (1958). On a space of completely additive functions. *Vestnik Leningrad. Univ. 13*, 52–59.

Kim, A. K. and R. J. Samworth (2016). Global rates of convergence in log-concave density estimation. *The Annals of Statistics 44*(6), 2756–2779.

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*.

Lehmann, E. L. and G. Casella (2006). *Theory of point estimation.* Springer Science & Business Media.

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.

Liboschik, T., K. Fokianos, and R. Fried (2015). tscount: An R package for analysis of count time series following generalized linear models.

Lii, K.-S. and M. Rosenblatt (1992). An approximate maximum likelihood estimation for non-gaussian non-minimum phase moving average processes. *Journal of Multivariate Analysis 43*(2), 272–299.

Lii, K.-S. and M. Rosenblatt (1996). Maximum likelihood estimation for nonGaussian non-minimum phase ARMA sequences. *Statistica Sinica 6*(1), 1–22.

Mair, P., K. Hornik, and J. de Leeuw (2009). Isotone optimization in R: pool-adjacent-violators algorithm (pava) and active set methods. *Journal of statistical software 32*(5), 1–24.

Mallows, C. (1972). A note on asymptotic joint normality. *Annals of Mathematical Statistics*, 508–515.

Mammen, E. (1991). Estimating a smooth monotone regression function. *The Annals of Statistics*, 724–740.

Medina, L. A. and V. H. Moll (2007). The integrals in Gradshteyn and Ryzhik. Part 10: The digamma function. *arXiv preprint arXiv:0709.3446*.

Murphy, S. A., A. van der Vaart, and J. Wellner (1999). Current status regression. *Mathematical methods of statistics 8*(3), 407–425.

Nikias, C. L. and A. P. Petropulu (1993). *Higher-order spectra analysis : a nonlinear signal processing framework*. PTR Prentice Hall Englewood Cliffs, N.J.

Pollard, D. (1984). *Convergence of stochastic processes*. David Pollard.

Rosenblatt, M. (2012). *Gaussian and non-Gaussian linear time series and random fields*. Springer Science & Business Media.

Schuhmacher, D., A. Hüsler, and L. Dümbgen (2011). Multivariate log-concave distributions as a nearly parametric model. *Statistics & Risk Modeling with Applications in Finance and Insurance 28*(3), 277–295.

Seijo, E. and B. Sen (2011). Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics*, 1633–1657.

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review 5*(1), 3–55.

Shephard, N. (1995). Generalized linear autoregressions. Technical report, Nuffield College, Oxford University.

Shively, T. S., T. W. Sager, and S. G. Walker (2009). A Bayesian approach to non-parametric monotone function estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(1), 159–175.

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

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.

Villani, C. (2008). *Optimal transport: old and new*, Volume 338. Springer Science & Business Media.

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.

Zucchini, W. and I. L. MacDonald (2009). *Hidden Markov Models for time series: an introduction using R*, Volume 22. Boca Raton: CRC press.