Quasi-likelihood Estimation of a Censored Autoregressive Model With Exogenous Variables

Chao Wang Kung-Sik Chan [*]

July 14, 2016

Abstract

Maximum likelihood estimation of a censored autoregressive model with exogenous variables (CARX) requires computing the conditional likelihood of blocks of data of variable dimensions. As the random block dimension generally increases with the censoring rate, maximum likelihood estimation becomes quickly numerically intractable with increasing censoring. We introduce a new estimation approach using the complete-incomplete data framework with the complete data comprising the observations were there no censoring. We introduce a system of unbiased estimating equations motivated by the complete-data score vector, for estimating a CARX model. The proposed quasi-likelihood method reduces to maximum likelihood estimation when there is no censoring, and it is computationally efficient. We derive the consistency and asymptotic normality of the quasi-likelihood estimator, under mild regularity conditions. We illustrate the efficacy of the proposed method by simulations and a real application on phosphorus concentration in river water.

Keywords: Maximum likelihood estimation; Estimating equation; Regression model; Time series.

[*] Chao Wang (email:chao-wang@uiowa.edu) is Post-doctoral Researcher and Kung-Sik Chan (email: kung-sik-chan@uiowa.edu) is Professor, both in the Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, IA 52242, USA. The authors are grateful to Keith Schilling for sharing with us his scientific insights regarding the real application, and thank the Iowa Department of Natural Resources for partial funding support. We also thank an Associate Editor and two referees for very helpful comments leading to a streamlined version of the paper.

1 Introduction

Censored time series data are frequently encountered in diverse fields including environmental monitoring, medicine, economics and social sciences. Censoring may arise when a measuring device is subject to some detection limit beyond which the device cannot yield a reliable measurement. For instance, the total phosphorus concentration in river water is an important indicator about water quality, and its fluctuations over time are often monitored in environmental studies. However, the phosphorus concentration cannot be measured exactly if it falls below certain detection limit.

There is an extensive literature on regression analysis with censored responses, since the pioneering work of Buckley & James (1979). However, the case of regression with both the response and covariates subject to censoring is relatively under-explored. Censored time-series regression analysis, for instance, the Tobit model with auto-correlated regression errors (Tobin 1958; Robinson 1982*a*), falls in the latter framework. Robinson (1980) studied maximum likelihood (ML) estimation of Gaussian time series models with left (right) censored data, and showed that with sufficiently sparse censoring, the likelihood of a censored autoregressive (AR) model only requires 1-dimensional integration. Robinson (1982*a*) showed that Gaussian likelihood estimation of a Tobit model that assumes independent errors is still strongly consistent and asymptotically normal, even if the Gaussian errors are auto-correlated. Moreover, Robinson (1982*b*) proposed two methods, namely, conditional least squares and method of moments, for estimating the residual autocorrelations, and established their strong consistency and asymptotic normality, under the normal error assumption. The consistent autocorrelation estimates can then be used to provide consistent estimation of any finite-order autoregressive-moving-average (ARMA) model specified for the regression errors.

Zeger & Brookmeyer (1986) studied maximum likelihood estimation of a regression model with the errors driven by an autoregressive (AR) model of known order $p \geq 0$. We first outline their approach in the simple case of no covariates and AR(1) errors. Without censoring, the Markov property of the AR(1) errors imply that the log-likelihood can be decomposed as the sum of the conditional log-likelihood of Y_t given Y_{t-1}. The presence of censoring invalidates the Markov property, but Zeger & Brookmeyer (1986) pointed out a generalized Markov property: the conditional log-likelihood of Y_t given all past Y's is equal to that of Y_t given $Y_{t-j}, j = 1, \ldots, m$, if Y_m is uncensored. Thus, the data can be partitioned into blocks that are preceded by an uncensored observation, except possibly the initial block, and the log-likelihood is the sum of the conditional log-likelihood of each block of data given the uncensored observation preceding it, plus that of the initial block. For AR(p) errors, the blocks are delineated by p consecutive uncensored observations, and the log-likelihood is then decomposed as the sum of the conditional log-likelihood of each block given the p consecutive uncensored observations preceding it, plus the log-likelihood of the initial block; if exogenous variables are available, the (conditional) likelihoods are further conditional on the exogenous variables. (See Robinson (1980) for a similar result in the normal case.) However the block dimension increases quickly with increasing censoring and the AR order; see the Appendix. Thus maximum likelihood estimation becomes quickly numerically intractable with increasing censoring even for moderately high AR order. Zeger & Brookmeyer (1986) also briefly discussed a pseudo-likelihood approach, but it was not fully developed. Park, Genton & Ghosh (2007) introduced an imputation method to estimate a censored time series model assuming the complete data is an ARMA process. They proposed to impute the censored values by some random values simulated from their conditional distribution given the observed data and the censoring information, and treat the imputed time series as the complete data with which any estimation procedure for complete time-series data can be used. However, they focused on the AR(1) model and relied on simulation studies to demonstrate their method, with no derivation of theoretical properties.

Our aim is to derive a computationally efficient estimation method via solving a system of fixed number of unbiased estimating equations for estimating a censored regression model with autoregressive errors. The basic idea of our approach assumes that the score of the complete-data conditional log-likelihood of Y_t^* given $Y_{t-j}^*, j = 1, \ldots, p$ and the covariates has a closed-form expression and so does its expectation given the (possibly) censored time series $Y_t, t = 0, \ldots, p$, evaluated at the same set of model parameters. Setting the preceding conditional mean score to zero then provides an unbiased estimating equation for estimating the model. Our proposed quasi-likelihood method becomes maximum likelihood estimation in the absence of censoring. Furthermore, we derive the consistency and asymptotic normality of the proposed estimator under some mild regularity conditions. Implementation of the proposed method for the important special case of normal innovations is discussed in detail.

In Section 2 we elaborate the model, the proposed estimation procedure and its theoretical properties. We report the empirical performance of the proposed method in Section 3, and apply in Section 3.6 the proposed method in a real application with a series of censored phosphorus concentrations in river water. We briefly conclude in Section 4. All technical details are postponed to the Appendix.

2 THE MODEL AND ESTIMATION PROCEDURE

2.1 The CARX model

Let $\{Y_t^*\}$ denote the real-valued time series of interest. Let $C \subset \mathbb{R}$ be the censoring region such that Y_t^* is not observed if $Y_t^* \in C$. The censoring region is frequently an interval of the form $(-\infty, c)$ or (c, ∞), or a finite interval (c_l, c_u), which are referred to as left, right, and interval censoring, respectively (Park et al. 2007). In practice, for data subject to left censoring with censoring limit c, the reported value $Y_t = \max\{c, Y_t^*\}$, $Y_t = \min\{c, Y_t^*\}$ for right censoring, and $Y_t = Y_t^* \times I(Y_t^* \notin C) + c \times I(Y_t^* \in C)$, where $c = (c_l + c_u)/2$, for interval censoring. Below, the censoring pattern is not restricted, but assumed to be predetermined and independent of the underlying process. The proposed method is applicable to left censoring, right censoring, interval censoring, or more general censoring patterns.

Additionally, let X_t be a vector covariate which bears a linear regression relationship with Y_t, with the regression errors assumed to follow an autoregressive model of order p where p is some known non-negative integer. (The proposed method can be readily extended to a nonlinear regression model.) $\{X_t\}$ is assumed to be always observable.

Below, let $v^\intercal$ denote the transpose of a vector or matrix v. For a time series $\{s_t\}$, let $s_{i:j} = (s_i, s_{i-1}, \ldots, s_j)$ if $i > j$, and $s_{i:j} = (s_i, s_{i+1}, \ldots, s_j)$ otherwise. We now state the model with the latent process given by

$$Y_t^* = X_t^\intercal \beta + \eta_t, \tag{1}$$

$$\eta_t = \sum_{i=1}^{p} \psi_i \eta_{t-i} + \varepsilon_t, \tag{2}$$

and their linkage to the observations given by

$$Y_t = \begin{cases} c, & \text{if } Y_t^* \in C, \\ Y_t^*, & \text{otherwise,} \end{cases} \tag{3}$$

where $\{\varepsilon_t\}$ is an independent and identically distributed process with mean 0 and variance σ^2. For the general case of variable censoring region, C and c in Eqn (3) are replaced by C_t and c_t. The preceding model is referred to as the Censored Auto-Regressive model with eXgenous variables (CARX). Let B denote the backshift operator so that for any time series $\{s_t\}$, $B^k s_t = s_{t-k}, k = 1, 2, \ldots,$ and $\Psi(B) = \sum_{i=1}^{p} \psi_i B^i$. Eqns. (1) and (2) can be rewritten as

$$(1 - \Psi(B))(Y_t^* - X_t^\intercal \beta) = \varepsilon_t. \tag{4}$$

Let $\psi = (\psi_1, \cdots, \psi_p)^\intercal$. Throughout, $\theta = (\beta^\intercal, \psi^\intercal, \sigma)^\intercal$ denotes a generic parameter vector, while θ_0 denotes the true parameter vector. Throughout, it is assumed that $\sigma_0 > 0$.

2.2 Estimation

We consider the problem of estimating a CARX model with data $\{(Y_t, X_t)\}_{t=1}^n$ generated from the CARX model with unknown parameter θ_0. Further notations are required. Define the following σ-algebras

$$\begin{aligned} \mathcal{F}_t^* &= \sigma\left\{X_{t:t-p}, Y_{t-1:t-p}^*\right\}, \\ \mathcal{F}_t &= \sigma\left\{X_{t:t-p}, Y_{t-1:t-p}\right\}, \\ \mathcal{G}_t &= \sigma\left\{Y_t, \mathcal{F}_t\right\}. \end{aligned}$$

Our proposed method is motivated by maximum likelihood estimation. Were $\{Y_t^*\}$ observable and supposing $\{\varepsilon_t\}$ admits a marginal probability density function (pdf) $f_\theta(\cdot)$, then the joint log-likelihood function for $Y^*_{n:1}$ conditional on $X_{n:1}$ is given by

$$\ell(Y^*_{n:1}|X_{n:1};\theta) = \sum_{t=p+1}^{n} \ell(Y^*_t|\mathcal{F}^*_t;\theta) + \ell(Y^*_{p:1}|X_{p:1};\theta),$$

where $\ell(Y^*_t|\mathcal{F}^*_t;\theta) = \log f_\theta((1-\Psi(B))(Y^*_t - X_t^\intercal\beta))$, due to the $AR(p)$ representation of the regression errors $\{\eta_t\}$. Suppressing contributions from the initial values yields a simpler conditional log-likelihood

$$\ell_*(\theta) = \sum_{t=p+1}^{n} \ell(Y^*_t|\mathcal{F}^*_t;\theta), \tag{5}$$

the maximization of which requires the first-order optimality condition:

$$0 = \nabla\ell_*(\theta) = \sum_{t=p+1}^{n} \nabla\ell(Y^*_t|\mathcal{F}^*_t;\theta),$$

where ∇ denotes taking the partial derivative with respect to θ.

However, censoring in the observed time series $\{Y_t\}$ entails that its joint log-likelihood cannot be reduced to a simple form similar to the one for $\{Y^*_t\}$ (Zeger & Brookmeyer (1986)). Let $S(Y^*_t|\mathcal{F}^*_t,\theta) = \nabla\ell(Y^*_t|\mathcal{F}^*_t;\theta)$. The proposed method of estimation is motivated by the observation that for all t, $E_\theta(S(Y^*_t|\mathcal{F}^*_t,\theta)|\mathcal{G}_t) = 0$ is an unbiased estimating equation, and so is

$$\sum_{t=p+1}^{n} \mathrm{E}_\theta\left[S(Y^*_t|\mathcal{F}^*_t,\theta)|\,\mathcal{G}_t\right] = 0, \tag{6}$$

which combines information from all data. In principle, the σ-algebra $\mathcal{G}_t$ can be chosen to include more or less information, for instance, $\mathcal{G}_t$ may be enlarged to the σ-algebra generated by all data, namely, $Y_1, Y_2, \ldots, Y_n$, in which case the imputed score defined by the left side of (6) is the observed-data score, and the associated Z-estimation method corresponds to (conditional) maximum likelihood estimation. The current choice of $\mathcal{G}_t$ is motivated by the ease of computation and the fact that in the absence of censoring, solving the preceding estimating equation reduces to maximum likelihood estimation. Below, we state an iterative algorithm for solving (6).

Step(1) Initialize the parameter estimate by some estimate, denoted by $\theta^{(0)}$.

Step(2) For each $k = 1, \ldots,$ obtain an update of estimate $\theta^{(k)}$ by

$$\theta^{(k)} = \mathrm{argmax}_{\theta}\, Q(\theta|\theta^{(k-1)}), \tag{7}$$

where for any current estimate $\theta^{(c)}$, $Q(\theta|\theta^{(c)}) = \sum_{t=p+1}^{n} Q_t(\theta|\theta^{(c)})$, and

$$Q_t(\theta|\theta^{(c)}) = \mathrm{E}_{\theta^{(c)}}\left[\ell_t^*(Y_t^*|\mathcal{F}_t^*;\theta)|\mathcal{G}_t\right]. \tag{8}$$

Step(3) Iterate Step(2) until $\|\theta^{(k)} - \theta^{(k-1)}\|_2/\|\theta^{(k-1)}\|_2 < \epsilon$ for some positive tolerance $\epsilon \approx 0$. Let $\hat{\theta}$ be the estimate obtained from the last iteration.

We now justify the proposed iterative algorithm. In many cases $Q(\theta|\theta^{(c)})$ is differentiable with respect to θ, so solving Eq (7) is equivalent to solving the equation

$$\frac{\partial Q(\theta|\theta^{(k-1)})}{\partial \theta} = 0. \tag{9}$$

Under suitable regularity conditions, differentiation and expectation can be interchanged. Let $f(\cdot,\theta)$ be the density function of some distribution, $S(\cdot,\theta) = \nabla f(x,\theta) = \frac{\partial f(x,\theta)}{\partial \theta}$ be its first derivative. Then,

$$S(Y_t^*|\mathcal{F}_t^*,\theta) = \nabla \log f(Y_t^*|\mathcal{F}_t^*,\theta),$$
$$S(Y_{t:t-p}^*|\mathcal{G}_t,\theta) = \nabla \log f(Y_{t:t-p}^*|\mathcal{G}_t,\theta).$$

Throughout, we assume that

$$\frac{\partial Q_t(\theta|\theta^{(k-1)})}{\partial \theta} = \mathrm{E}_{\theta^{(k-1)}}\left[S(Y_t^*|\mathcal{F}_t^*,\theta)|\mathcal{G}_t\right].$$

Thus, $\theta^{(k)}$ solves the following equation

$$\sum_{t=p+1}^{n} \mathrm{E}_{\theta^{(k-1)}}\left[S(Y_t^*|\mathcal{F}_t^*,\theta)|\mathcal{G}_t\right] = 0. \tag{10}$$

Hence, if the iteration converges to a limit denoted by $\hat{\theta}$, then it holds that

$$\sum_{t=p+1}^{n} \mathrm{E}_{\hat{\theta}}\left[S(Y_t^*|\mathcal{F}_t^*,\hat{\theta})\middle|\,\mathcal{G}_t\right] = 0, \tag{11}$$

so that $\hat{\theta}$ solves the estimating equation (6).

Remark Since Eq. (6) may have multiple roots, it is desirable to initialize the preceding algorithm by some estimate close to the true value. As mentioned in Section 1, in the case of normal innovations and left (right) censoring, the estimators introduced by Robinson (1982*a*) and Robinson (1982*b*) can be used to provide consistent estimates that can serve as initial values for the proposed algorithm. However, these estimators are generally inefficient, for instance, the initial AR estimates utilize information from a subset of the data whose lagged responses from lags 1 to p are uncensored. Consequently, for small samples, the initial AR estimates could be non-stationary. Similarly, the innovation variance estimator could be non-positive (Amemiya 1973). We note that the methods introduced by Robinson (1982*a*) and Robinson (1982*b*) may be lifted to the case of non-normal innovation distributions which admit explicit formulas for their truncated moments; see Jawitz (2004) for a survey of such formulas. An alternative initialization scheme consists of replacing the censored observations by their censoring limits for left or right censoring or the mean of the censoring limits in the case of interval censoring, and fitting model (1) with the modified data as if they were uncensored. While the initial estimates so obtained are biased (Park et al. 2007), our limited experience suggests that the algorithm so initialized generally converges without problems.

We present another characterization of the proposed estimator $\hat{\theta}$. The conditional pdf of $Y^*_{t:t-p}$ given $X_{t:t-p} = x_{t:t-p}$ is $f(y^*_{t:t-p}|x_{t:t-p}, \theta) = f(y^*_t|\mathcal{F}^*_t; \theta)f(y^*_{t-1:t-p}|x_{t-1:t-p}, \theta) = \exp\{\ell(y^*_t|\mathcal{F}^*_t; \theta)\}f(y^*_{t-1:t-p}|x_{t-1:t-p}, \theta)$. Let $R = R(y_{t:t-p})$ denote the collection of $y^*_{t:t-p}$ such that $y_{t:t-p}$ is observed if and only if $y^*_{t:t-p} \in R$. For instance, if none of $y_{t:t-p}$ are censored, then $R = \{y_{t:t-p}\}$ whereas if all of $y_{t:t-p}$ are censored, then $R = \prod_{j=0}^{p} C_{t-j}$. The likelihood of θ based on $Y_{t:t-p}$ given $X_{t:t-p}$ is $f(Y_{t:t-p}|X_{t:t-p}, \theta) = \int_{R(Y_{t:t-p})} f(y^*_{t:t-p}|X_{t:t-p}, \theta)dy^*_{t:t-p}$, where $dy^*_{t:t-p}$ signifies the product measure of the Lebesgue measure and the counting measure induced on R, for instance, the counting measure if R is a singleton. Hence, we can estimate the parameter by maximizing the composite likelihood $\tilde{L}_n(\theta) = \sum_{t=p+1}^{n} \log f(Y_{t:t-p}|X_{t:t-p}, \theta)$. Indeed, this will provide a consistent estimator as, under some general regularity conditions including ergodicity and stationarity, $\tilde{L}_n(\theta)/n \to \tilde{L}(\theta) = E_{\theta_0}\{\log f(Y_{t:t-p}|X_{t:t-p}, \theta)\}$ which is uniquely maximized at θ_0, the true parameter. A major difficulty of this approach is that the maximization problem is generally intractable. Thus, we extend the definition of $f(y_{t:t-p}|x_{t:t-p},\theta)$ by decoupling the parameter indexing the conditional density of Y_t^* given $\mathcal{F}_t^*$ from that of $Y_{t-1:t-p}^*$ given $X_{t-1:t-p}$, as follows: $\pi_{\delta,\theta}(y_{t:t-p}|x_{t:t-p}) = \int_{R(y_{t:t-p})} \exp\{\ell(y_t^*|\mathcal{F}_t^*;\delta)\} f(y_{t-1:t-p}^*|x_{t-1:t-p},\theta) dy_{t:t-p}^*$ and introduce the objective function $\tilde{L}_n(\delta,\theta) = \sum_{t=p+1}^n \log \pi_{\delta,\theta}(Y_{t:t-p}|X_{t:t-p})$. Similarly, it holds under suitable conditions that $\tilde{L}_n(\delta,\theta)/n \to \tilde{L}(\delta,\theta) = E_{\theta_0}\{\log \pi_{\delta,\theta}(Y_{t:t-p}|X_{t:t-p})\}$ which is uniquely maximized at (θ_0,θ_0). In particular, $\theta_0 = \arg\max_\delta \tilde{L}(\delta,\theta_0)$.

The new objective function suggests an iterative estimation scheme by first updating δ by some value that increases $\tilde{L}_n(\cdot,\theta)$ over its current value, with θ fixed at its current estimate, followed by updating θ as the just updated δ, with the procedure repeated until convergence. This is achieved by the proposed estimator $\hat{\theta}$, because

$$\begin{aligned}
& \tilde{L}_n(\delta,\theta) - \tilde{L}_n(\theta,\theta) \\
= & \sum_{t=p+1}^n \log\left[\frac{\int_{R(Y_{t:t-p})} \exp\{\ell(y_t^*|\mathcal{F}_t^*;\delta)\} f(y_{t-1:t-p}^*|X_{t-1:t-p},\theta) dy_{t:t-p}^*}{\int_{R(Y_{t:t-p})} \exp\{\ell(y_t^*|\mathcal{F}_t^*;\theta)\} f(y_{t-1:t-p}^*|X_{t-1:t-p},\theta) dy_{t:t-p}^*}\right] \\
= & \sum_{t=p+1}^n \log\left[\frac{\int_{R(Y_{t:t-p})} \exp\{\ell(y_t^*|\mathcal{F}_t^*;\delta) - \ell(y_t^*|\mathcal{F}_t^*;\theta)\} f(y_{t:t-p}^*|X_{t:t-p},\theta) dy_{t:t-p}^*}{\int_{R(Y_{t:t-p})} f(y_{t:t-p}^*|X_{t:t-p},\theta) dy_{t:t-p}^*}\right] \\
\geq & Q(\delta|\theta) - Q(\theta|\theta),
\end{aligned}$$

where the last inequality follows from Jensen's inequality. The consistency of $\hat{\theta}$ derived in the next section shows that it maximizes the objective function $\tilde{L}_n(\theta)$ asymptotically. Henceforth, the proposed method will be referred to as quasi-likelihood estimation.

2.3 Asymptotic properties of the estimator

In general it is difficult to establish the global consistency of an estimator, which is also the case for our setting. Fortunately, as remarked earlier, a consistent estimator is generally available, so it suffices to establish local consistency for the proposed estimator. Henceforth, the initial estimate $\theta^{(0)}$ for the iterative algorithm is assumed to be consistent. Therefore, we can and shall restrict the parameter space to Θ, a neighborhood of θ_0 which, without loss of generality, is furthermore assumed to be compact.

The covariate process $\{X_t\}$ is assumed to be stationary and β-mixing with exponentially decaying mixing coefficients, which is a mild assumption. So shall we assume the regression error process $\{\eta_t\}$, which holds if (i) all roots of the characteristic polynomial $1-\sum_{j=1}^{p}\psi_j z^j$ lie outside the unit circle (Cryer & Chan 2008) and (ii) ε_t admits a pdf (Pham & Tran 1985).

The asymptotic property of the estimator depends on the following Z functions,

$$\begin{aligned} Z_t(\theta) &= \frac{\partial Q_t(\theta|\theta^{(c)})}{\partial\theta}|_{\theta^{(c)}=\theta} = \mathrm{E}_\theta\left[S(Y_t^*|\mathcal{F}_t^*,\theta)|\mathcal{G}_t\right], \\ Z^{(n)}(\theta) &= \frac{1}{n-p}\sum_{t=p+1}^{n} Z_t(\theta), \\ Z(\theta) &= \mathrm{E}_{\theta_0}\left[Z_t(\theta)\right]. \end{aligned}$$

The estimating equation (6) is equivalent to

$$Z^{(n)}(\theta) = 0. \tag{12}$$

To establish the desired consistency and asymptotic normality for the proposed estimator, we make heavy use of empirical process theories for Vapnick-Cervonenkis (V-C) classes of functions (Arcones & Yu 1994). See Van der Vaart (2000) for a review of V-C class. We shall require the process $\{Z_t(\theta);\theta\in\Theta\}$ to be a V-C class satisfying some moment condition.

In summary, the following assumptions are imposed below. Let $q\in(2,\infty)$ be some fixed real number.

A1. The initial estimate $\theta^{(0)}$ is a consistent estimator and the parameter space Θ is a compact neighborhood of the true parameter θ_0.

A2. The covariate process $\{X_t\}$ is β-mixing with exponentially decaying mixing coefficients.

A3. The censoring region may vary with t but it is assumed to be an interval $C_t=(c_{t,l},c_{t,u})$ with possibly infinite censoring limits. Alternatively, the censoring region is the complement of an interval $C_t=(c_{t,l},c_{t,u})^c$, in order to accommodate simultaneous left and right censoring. The process $\{(Y_t^*,X_t,c_{t,l},c_{t,u})\}$ is stationary. The censoring limits are β-mixing processes with exponentially decaying mixing coefficients, independent of $\{X_t\}$ and $\{Y_t^*\}$.

A4. For all $|z|\le 1$, the polynomial $1-\sum_{j=1}^{p}\psi_j z^j \neq 0$.

A5. The distribution of ε_t has a twice differentiable pdf.

A6. The class of functions $\{Z_t(\theta) : \theta \in \Theta\}$ is a V-C subgraph class and there exists an envelope function F such that $\sup_{\theta\in\Theta} |Z_t(\theta)| \le F$ with $F \in L^q$.

A7. The matrix $\nabla Z(\theta) = \mathrm{E}_{\theta_0}\left[\nabla Z_t(\theta)\right]$ is continuous in θ and it is nonsingular at the true parameter θ_0.

The asymptotic properties of the proposed estimator are established in the following theorem.

Theorem 2.1. *Under A1–A6, the quasi-likelihood estimator $\hat{\theta}$ is locally consistent, i.e., $\hat{\theta} \xrightarrow{P} \theta_0$, over all sufficiently small compact neighborhood Θ of θ_0.*

Furthermore, if A7 holds, it is also asymptotically normal, i.e.,

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{\mathcal{L}} N(0, \Sigma),$$

where $\Sigma = M_1^{-1} M_0 (M_1^{-1})^\intercal$, $M_0 = \sum_{i=-\infty}^{\infty} E_{\theta_0}\left[Z_{t+|i|}(\theta_0) Z_t^\intercal(\theta_0)\right]$ and $M_1 = E_{\theta_0}\left[(\nabla Z_t)(\theta_0)\right]$.

Remark Note that the asymptotic covariance matrix has a sandwich form involving two matrices. The first one M_1 is assumed to be non-singular, and the second matrix M_0 is an infinite sum of auto-covariance matrices. Both matrices are not easy to compute. In this regard, a parametric bootstrap procedure is proposed to estimate the covariance matrix Σ and construct confidence intervals of the unknown parameters. Specifically, given the estimator $\hat{\theta}$ and using the observed $\{X_t\}$, uncensored observations of the same sample size can be readily simulated. Then the uncensored observations can be subject to the observed censoring scheme to yield the simulated censored time-series responses with which a bootstrap estimate $\tilde{\theta}$ can be obtained using the iterative algorithm. The bootstrap procedure can be replicated for, say, B times to yield a sample of bootstrap parameter estimates with which the sample covariance matrix of the bootstrap estimates provides an estimate of the covariance matrix of $\hat{\theta}$. Also, confidence intervals can be constructed directly from the sample quantiles of the bootstrap parameter estimates.

2.4 Normal innovations

We now specialize to the important case that the innovations $\{\varepsilon_t\}$ are normally distributed with zero mean and common variance σ^2. Then the log-likelihood of Y_t^* conditional on $\mathcal{F}_t^*$ is given by the following expression apart from an additive constant,

$$\ell(Y_t^*|\mathcal{F}_t^*, \theta) = -\frac{1}{2}\log(\sigma^2) - \left\{(Y_t^* - X_t^\intercal\beta) - \sum_{j=1}^{p}\psi_j\left(Y_{t-j}^* - X_{t-j}^\intercal\beta\right)\right\}^2 /2\sigma^2.$$

For any given parameter vector θ, let

$$Z_{t_1,t_2}(\theta) = \mathrm{E}_\theta[Y_{t_1}^*|\mathcal{G}_{t_2}],$$

$$\Sigma_t(\theta) = \mathrm{cov}(Y_{t:t-p}^*|\mathcal{G}_t;\theta),$$

which can be readily computed based on the explicit formulas for the moments of a truncated multivariate normal variable (Tallis 1961); see also the R (R Core Team 2015) package `mvtnorm` (Genz & Bretz 2009; Genz, Bretz, Miwa, Mi, Leisch, Scheipl & Hothorn 2014). Then

$$\begin{aligned} Q_t(\theta|\theta^{(c)}) &= \mathrm{E}_{\theta^{(c)}}\left[\ell_t^*(\theta)|\mathcal{G}_t\right] \\ =& -\log(\sigma^2)/2 - \Big([Z_{t,t}(\theta^{(c)}) - X_t^\intercal\beta - \sum_{j=1}^{p}\psi_j\left\{Z_{t-j,t}(\theta^{(c)}) - X_{t-j}^\intercal\beta\right\}]^2 \\ &+(1,-\psi^\intercal)\Sigma_t(\theta^{(c)})(1,-\psi^\intercal)^\intercal\Big)/(2\sigma^2). \end{aligned}$$

The maximization required in Step (2) can be carried out via block co-ordinate descent that sequentially updates β, ψ's and σ^2 block by block as follows:

1 For given $\theta^{(k)}$, ψ, and σ, the regression coefficient β is updated by regressing $Z_{t,t}(\theta^{(k)}) - \sum_{j=1}^{p}\psi_j Z_{t-j,t}(\theta^{(k)})$ on $X_t - \sum_{j=1}^{p}\psi_j X_{t-j}$ for $t = p+1,\ldots,n$.

2 For given $\theta^{(k)}$, β, and σ, update ψ by maximizing the Q function which is a quadratic function of ψ so the optimization can be readily done. Alternatively, it suffices to update ψ to another feasible vector which increases $Q(\cdot|\theta^{(k)})$.

3 For given $\theta^{(k)}$, β, and ψ, update σ by the formula

$$\begin{aligned}\sigma^2 =& \frac{1}{n-p}\sum_{t=p+1}^{n}\Big([Z_{t,t}(\theta^{(k)}) - X_t^\intercal\beta - \sum_{j=1}^{p}\psi_j\{Z_{t-j,t}(\theta^{(k)}) - X_{t-j}^\intercal\beta\}]^2 \\ &+ (1,-\psi^\intercal)\Sigma_t(\theta^{(k)})(1,-\psi^\intercal)^\intercal\Big).\end{aligned}$$

Furthermore, it is clear that assumption A5 is true. A6 can also be verified as in Proposition 5.1 with some mild assumptions about X_t. In addition, A7 can be verified for the case of left (right) censoring with a constant censoring limit. The proof techniques used there can be extended to other continuous innovation distributions, under certain regularity conditions.

2.5 Model prediction

Consider the problem of predicting the future values Y^*_{n+h}, where $h = 1, 2, \ldots, H$, given the observations $\{(Y_t, X_t)\}_{t=1}^n$ and supposing the availability of future covariate values $\{X_{t+h}\}_{h=1}^H$. To simplify the derivation of the predictive distribution, we assume the normality of ε_t and known θ_0, although the following method can be readily extended to non-normal innovations. Due to the autoregressive nature of the regression errors $\eta_t = Y_t^* - X_t^\intercal \beta$, the conditional distribution

$$\begin{aligned}\mathfrak{L}_{n,h} &= \mathfrak{L}(Y^*_{n+h} | \{X_{n+i}\}_{i=1}^h, \{(Y_t, X_t)\}_{t=1}^n) \\ &= \mathfrak{L}(Y^*_{n+h} | \{X_{n+i}\}_{i=1}^h, \{(Y_t, X_t)\}_{t=\tau}^n),\end{aligned}$$

where $\tau = \max\big(\{1\} \cup \big\{1 \le u \le n - p + 1 : \text{none of } \{Y_t\}_{t=u}^{u+p-1} \text{ is censored}\big\}\big)$; see Zeger & Brookmeyer (1986).

If $\tau = n - p + 1$, i.e., the most recent p Y's are uncensored, the prediction problem is the same as that for an ordinary time series regression model. In particular, for any $h = 1, \ldots, H$, $\mathfrak{L}_{n,h}$ is a normal distribution. Specifically, the predicted value $\hat{Y}^*_{n+h}$ is given recursively by $\hat{Y}^*_{n+h} = X^\intercal_{n+h}\beta + \hat{\eta}_{n+h}$, with $\hat{\eta}_{n+h} = \sum_{l=1}^p \hat{\psi}_l \hat{\eta}_{n+h-l}$, and $\hat{\eta}_t = Y_t - X_t^\intercal \beta$ for $t \le n$. The prediction error can be written as $\epsilon_{n+i} = \varepsilon_{n+i} + \sum_{l=1}^p \psi_l \epsilon_{n+i-l} = \sum_{j=0}^i \omega_{i,j} \varepsilon_{n+i-j}$, where the coefficients $\omega_{i,j}$ can be calculated recursively through the preceding identity and the initial condition $\omega_{i,0} = 1$, which results in a formula for the prediction variance, namely, $\text{var}(\hat{Y}^*_{n+h}) = \hat{\sigma}^2 \sum_{j=0}^i \omega_{i,j}^2$.

If $\tau < n - p + 1$, then $\mathfrak{L}_{n,h}$ is generally a truncated multivariate normal distribution. Although its first and second moments can be computed analytically, they are not as useful in constructing predictive intervals. Here a Monte Carlo method is proposed to estimate any interesting characteristic of the predictive distribution of Y^*_{n+h}. Note that the regression errors $\{\eta_t = Y_t^* - X_t^\intercal \beta\}_{t=\tau}^n$ follows a multivariate normal distribution, unconditionally. Let η_c and η_o be the sub-vectors of $\eta_{\tau:n}$ such that the corresponding elements of $Y_{\tau:n}$ are censored and observed, respectively. Then given $Y_{\tau:n}$, η_c follows a truncated multivariate normal distribution, whose realizations can be readily simulated so we can draw realizations from the conditional distribution of $Y_{n-p+1:n}^*$ and thence those of $Y_{n+h}^*, h = 1, \ldots, H$, given $\{(Y_t, X_t)\}_{t=1}^n$ and $\{X_{t+h}\}_{h=1}^H$. We can then construct predictive intervals of Y_{n+h}^* from a random sample from the predictive distribution of Y_{n+h}^*, using the percentile method.

2.6 Simulated residuals

In the presence of censoring, there are several ways to define residuals, for instance, generalized residuals and simulated residuals (Gourieroux, Monfort, Renault & Trognon 1987; Hillis 1995). See also Cox & Snell (1968). If Y_t^* is observed, the corresponding residual is universally defined as $Y_t^* - \hat{Y}_{t|t-1}^*$, where $\hat{Y}_{t|t-1}$ is the mean of $\mathfrak{L}_{t-1,1}$, evaluated at the parameter estimate. In the presence of censoring so that some Y_t^*s are unobserved, we compute the simulated residuals as follows: First, impute each unobserved Y_t^* by a realization from the conditional distribution $\mathfrak{L}(Y_t^* | \{(Y_s, X_s)\}_{s=1}^t)$, evaluated at the parameter estimate. Then, refit the model with $\{(Y_t^*, X_t)\}$ so obtained, via conditional maximum likelihood; the residuals from the latter model are the simulated residuals $\hat{\varepsilon}_t$. Let the corresponding parameter estimate of θ be $\tilde{\theta}$. The corresponding (simulated) partial residuals for the X's, i.e., $X_t^\intercal \tilde{\beta} + \hat{\varepsilon}_t$, can be used to assess the relationship between Y and X, after adjusting for the autoregressive errors. Gourieroux et al. (1987) showed that under some regularity conditions, model diagnostic tests using the simulated residuals have the same asymptotic null distributions as the uncensored case. Limited simulation study reported below shows that the asymptotic null distribution of the Ljung-Box test statistic based on the simulated residuals is the same as that for uncensored data, hence it provides a useful tool for model diagnostics.

3 DATA EXAMPLES

3.1 Empirical performance of the proposed estimation method

We study the empirical performance of the proposed method by simulations. Data were simulated from the CARX model subject to left censoring with a constant censoring limit c, with the covariate X_t's being independent two-dimensional random vectors comprising independent standard normally distributed components. The regression errors follow an $AR(3)$ process with normally distributed innovations of zero mean and variance σ^2. Left censoring is enforced with the censoring limit being -1.5, -0.7, and -0.2 to make approximate censoring rates of 5%, 20%, and 40%, respectively. Several sample sizes including 100, 200, 500 and 1000 were tried. Unless stated otherwise, all experiments were replicated 1000 times. As comparison, we contrast the proposed method (method 1 in Table 1) with the incorrect but convenient method of ignoring censoring and applying conditional maximum likelihood estimation with the censored data simply replaced by their censoring limits (method 0), whose estimates served as the initial values for the proposed method.

Table 1 reports the sample mean for each parameter estimate and their sample standard deviations enclosed in round brackets, for both methods. In addition, the empirical coverage rates of the 95% confidence interval obtained by parametric bootstrap are listed in square brackets, but only for the proposed method because conditional maximum likelihood estimation ignoring censoring is quite biased for the case of moderately high censoring rate. It can be seen that the proposed method performed well in all cases, while ignoring censoring led to increasing bias with the censoring rate. For both methods, the variance of the estimator increased with the censoring rate and decreased with the sample size. Parametric bootstrap seems to work well as the empirical coverage rates of the bootstrap confidence intervals were quite close to the nominal 95%.

3.2 Comparison with maximum likelihood estimation

We examine the potential loss of efficiency of the proposed method as compared with ML estimation, by simulation for the case of AR(1) errors. Specifically, data of size 100, 200, or 400, respectively, were simulated and censored whenever their magnitude exceeded certain

n	c	method	β_1	β_2	ψ_1	ψ_2	ψ_3	σ
-	-	-	0.2	0.4	0.1	0.3	-0.2	0.707
100	-1.5	0	0.187 (0.07)	0.380 (0.07)	0.0936 (0.10)	0.282 (0.10)	-0.183 (0.10)	0.666 (0.05)
		1	0.195 (0.07) [93.4%]	0.398 (0.06) [93.6%]	0.0969 (0.10) [95.0%]	0.286 (0.10) [94.3%]	-0.187 (0.10) [95.2%]	0.690 (0.05) [77.5%]
	-0.7	0	0.154 (0.063)	0.314 (0.059)	0.0930 (0.10)	0.273 (0.10)	-0.148 (0.10)	0.587 (0.047)
		1	0.196 (0.077) [92.8%]	0.399 (0.074) [95.4%]	0.0934 (0.11) [94.2%]	0.285 (0.11) [93.7%]	-0.187 (0.11) [95.9%]	0.689 (0.060) [87.4%]
	-0.2	0	0.116 (0.054)	0.235 (0.054)	0.162 (0.097)	0.339 (0.11)	-0.0269 (0.095)	0.509 (0.050)
		1	0.195 (0.083) [93.5%]	0.399 (0.083) [93.4%]	0.0909 (0.12) [94.9%]	0.278 (0.12) [93.7%]	-0.189 (0.13) [96.4%]	0.682 (0.068) [85.5%]
200	-1.5	0	0.190 (0.045)	0.381 (0.045)	0.0956 (0.069)	0.290 (0.069)	-0.188 (0.068)	0.675 (0.034)
		1	0.199 (0.047) [94.6%]	0.399 (0.049) [93.9%]	0.0980 (0.070) [95.6%]	0.294 (0.070) [94.1%]	-0.193 (0.070) [96.1%]	0.699 (0.038) [90.1%]
	-0.7	0	0.157 (0.040)	0.315 (0.041)	0.0975 (0.070)	0.282 (0.074)	-0.149 (0.066)	0.596 (0.034)
		1	0.200 (0.049) [94.7%]	0.400 (0.052) [93.9%]	0.0979 (0.075) [95.1%]	0.293 (0.073) [93.9%]	-0.193 (0.074) [95.1%]	0.698 (0.042) [89.6%]
	-0.2	0	0.117 (0.036)	0.235 (0.038)	0.165 (0.066)	0.346 (0.077)	-0.0287 (0.063)	0.517 (0.036)
		1	0.199 (0.054) [95.1%]	0.400 (0.059)[93.8%]	0.0974 (0.084) [95.0%]	0.290 (0.083) [93.3%]	-0.193 (0.086) [95.5%]	0.695 (0.048) [90.1%]
500	-1.5	0	0.192 (0.028)	0.382 (0.029)	0.0965 (0.045)	0.294 (0.042)	-0.193 (0.043)	0.678 (0.021)
		1	0.201 (0.029) [94.5%]	0.399 (0.031) [94.6%]	0.0987 (0.045) [94.7%]	0.298 (0.042) [95.5%]	-0.197 (0.044) [95.7%]	0.704 (0.024) [93.7%]
	-0.7	0	0.158 (0.025)	0.314 (0.026)	0.0992 (0.044)	0.287 (0.045)	-0.154 (0.043)	0.600 (0.021)
		1	0.201 (0.031) [94.2%]	0.400 (0.033) [94.1%]	0.0996 (0.048) [94.2%]	0.298 (0.045) [94.8%]	-0.198 (0.046) [95.9%]	0.703 (0.027) [92.9%]
	-0.2	0	0.118 (0.023)	0.235 (0.024)	0.167 (0.041)	0.350 (0.049)	-0.033 (0.042)	0.522 (0.023)
		1	0.200 (0.033) [94.7%]	0.399 (0.036) [94.2%]	0.0985 (0.053) [93.7%]	0.296 (0.050) [94.8%]	-0.196 (0.053) [95.9%]	0.702 (0.030) [93.3%]
1000	-1.5	0	0.190 (0.020)	0.381 (0.020)	0.0969 (0.031)	0.294 (0.030)	-0.196 (0.031)	0.681 (0.014)
		1	0.199 (0.021) [95.1%]	0.399 (0.022) [94.8%]	0.0991 (0.031) [94.2%]	0.299 (0.030) [94.5%]	-0.200 (0.031) [94.9%]	0.705 (0.016) [95.2%]
	-0.7	0	0.156 (0.018)	0.314 (0.019)	0.0990 (0.031)	0.288 (0.032)	-0.156 (0.030)	0.603 (0.014)
		1	0.199 (0.022) [95.3%]	0.399 (0.023) [94.2%]	0.0987 (0.033) [95.5%]	0.298 (0.032) [94.4%]	-0.200 (0.033) [95.0%]	0.705 (0.018) [95.1%]
	-0.2	0	0.117 (0.016)	0.236 (0.017)	0.166 (0.029)	0.352 (0.035)	-0.0351 (0.030)	0.524 (0.016)
		1	0.199 (0.024) [95.1%]	0.399 (0.026) [94.2%]	0.0991 (0.036) [94.9%]	0.297 (0.035) [94.4%]	-0.201 (0.038) [94.8%]	0.705 (0.021) [93.9%]

Table 1: Simulation study. The true parameters are displayed in the first row. For each sample size 100, 200, 500 , or 1000 and left censoring limit -1.5, -0.7, or -0.2, we contrast the proposed method (labeled as 1) with conditional maximum likelihood estimation ignoring censoring, i.e., with the censored data simply replaced by the censoring limit c. For each parameter, the reported values are the sample mean and sample standard deviation (in round brackets). In addition, the empirical coverage rates of the 95% bootstrap confidence intervals based on the proposed method are enclosed in square brackets.

threshold in order to make censoring rate of 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, respectively, from the following model: $(1 - \psi B)(y_t^* - X_t^\intercal (0.2, 0.4)^\intercal) = \varepsilon_t$, where $\psi = -0.5, 0.5, 0.9$, $\{\varepsilon_t\} \sim_{iid} N(0, 0.6^2)$, and $\{X_t\} \sim_{iid} N(0, I)$ are independent of ε_t.

The MLE was computed by implementing the EM algorithm in Section 3.2 of Zeger & Brookmeyer (1986), except that the AR parameter update was done by directly maximizing the imputed log-likelihood. Here we only report the results for the case of sample size equal to 400, as they are representative of results of other sample sizes; results for $\psi = -0.5$ are similar to those for $\psi = .5$, and hence omitted. Figure 1 plots the ratio of the mean squared error (MSE) of the proposed estimator to that of the MLE against the censoring rate, for each parameter of the model, which shows that for $\psi = .5$, there is little loss of efficiency as measured by the MSE, being at most 6% loss at 50% censoring rate. Moreover, the AR parameter seemed to have been slightly more efficiently estimated by the proposed method than ML estimation at low censoring rates, perhaps because the numerical (possibly high-dimensional) integration required by ML estimation more than offsets its theoretical efficiency in such cases. For $\psi = 0.9$, the loss of efficiency is greater but lower than 8%, except for the AR(1) coefficient estimate whose loss elevates to 14% at 50% censoring rate. Mean computation time is another important metric for comparing the two methods. Figure 2 plots against the censoring rate the ratio of the mean computation time of ML estimation to that of the proposed method, which demonstrates that the proposed method is computationally much more efficient than ML estimation as it was almost 50 (800) times faster than ML estimation at 50% censoring rate, for $\psi = 0.5$ (0.9).

3.3 The method applied to missing data

As noted by Zeger & Brookmeyer (1986), missing responses can be regarded as resulting from left censoring with an infinite censoring limit. In particular, a CARX model with the response missing at random provides another setting for comparing the proposed method with Gaussian likelihood estimation which can be readily carried out via Kalman filtering as implemented by the `arima` function in the `stats` package of (R Core Team 2015; Ripley 2002). Data of size 100, 200, 400, respectively, were simulated from the model in Section 3.2 except that the regression errors now follow an AR(2) model. Data cases were subsequently

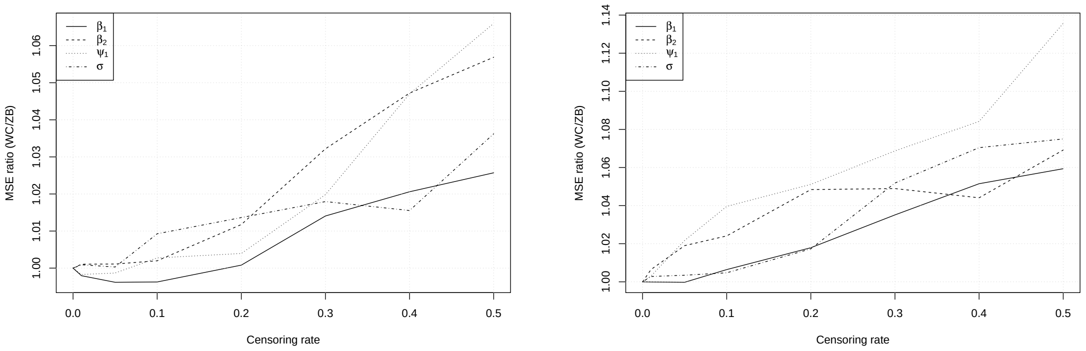

Figure 1: Ratios of the mean squared errors (MSE) of the proposed method to those of maximum likelihood estimation. Left diagram: the true AR(1) coefficient $\psi = 0.5$. Right diagram: $\psi = 0.9$.

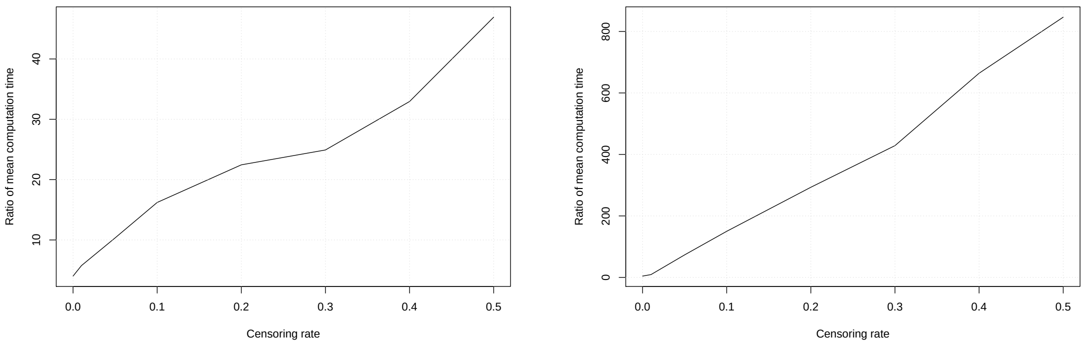

Figure 2: Plot of the ratio of the mean computation time of maximum likelihood estimation to that of the proposed method. Left diagram: $\psi = 0.5$. Right diagram: $\psi = 0.9$.

discarded randomly to make a missing rate of 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, respectively. With each simulated series, the CARX model was estimated once by ML estimation via the `arima` function and once by the proposed method, again with the true parameter values as the initial values. The results were similar across different sample sizes, so we only report the case for sample size 400. Figure 3 plots against the missing rate the ratio of the MSE of the proposed method to that of the ML estimation, for each parameter, which shows that the MSE of the proposed method is less than 5% higher than ML estimation for missing rate up to 20%, and is generally not more than 10% higher even at 50% missing rate, except for the parameter σ. The simulation results in this and last sub-sections indicate that the proposed method generally incurs relatively little loss of efficiency compared with ML estimation.

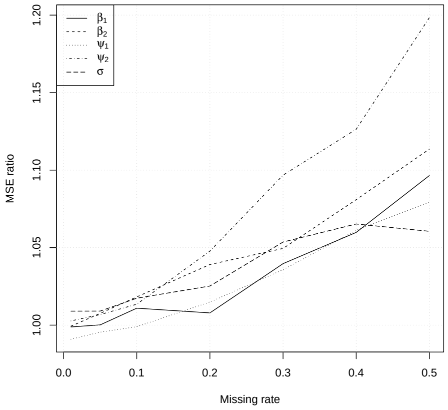

Figure 3: Ratios of the mean squared errors (MSE) of the proposed method to those of maximum likelihood estimation with missing data.

3.4 The robustness of the proposed method

As the proposed method may also be regarded as a generalization of Gaussian likelihood estimation or conditional least squares, it is of interest to assess its robustness against departure from the normal innovation assumption. We did this with a simulation study where the innovations were t-distributed, while the model estimation was done based on normal innovations.

n	distribution	β_1	β_2	ψ_1	ψ_2	ψ_3	σ
-	-	0.2	0.4	0.1	-0.3	-0.2	0.707
100	t (df =5)	0.200 (0.064)[95.0%]	0.396 (0.069)[94.2%]	0.107 (0.108)[94.7%]	0.303 (0.103)[95.9%]	-0.207 (0.109)[96.3%]	0.625 (0.062)[53.5%]
	t (df=10)	0.200 (0.073)[92.6%]	0.400 (0.073)[94.0%]	0.103 (0.104)[95.4%]	0.294 (0.099)[95.3%]	-0.201 (0.106)[96.0%]	0.665 (0.060)[78.7%]
	t (df=20)	0.200 (0.073)[93.1%]	0.405 (0.079)[92.1%]	0.102 (0.105)[95.2%]	0.292 (0.101)[95.7%]	-0.201 (0.107)[96.3%]	0.677 (0.061)[84.5%]
	normal	0.197 (0.079)[92.6%]	0.401 (0.075)[93.4%]	0.097 (0.106)[95.0%]	0.289 (0.102)[95.7%]	-0.191 (0.108)[95.3%]	0.691 (0.062)[88.6%]
200	t (df=5)	0.199 (0.045)[95.4%]	0.395 (0.048)[93.2%]	0.103 (0.076)[93.5%]	0.302 (0.071)[95.4%]	-0.210 (0.074)[94.6%]	0.638 (0.043)[50.1%]
	t (df=10)	0.198 (0.048)[93.9%]	0.396 (0.050)[94.0%]	0.100 (0.074)[94.2%]	0.298 (0.068)[95.6%]	-0.205 (0.072)[95.9%]	0.675 (0.044)[78.3%]
	t (df=20)	0.200 (0.050)[94.1%]	0.399 (0.051)[93.6%]	0.100 (0.073)[95.4%]	0.296 (0.071)[95.1%]	-0.200 (0.074)[95.7%]	0.688 (0.044)[86.4%]
	normal	0.200 (0.049)[94.2%]	0.402 (0.052)[93.7%]	0.098 (0.073)[95.8%]	0.296 (0.070)[95.7%]	-0.196 (0.074)[95.0%]	0.700 (0.043)[91.5%]
400	t (df=5)	0.197 (0.033)[94.1%]	0.395 (0.033)[93.5%]	0.105 (0.053)[93.5%]	0.306 (0.051)[94.0%]	-0.207 (0.053)[94.0%]	0.641 (0.031)[30.9%]
	t (df=10)	0.198 (0.033)[94.6%]	0.397 (0.034)[95.3%]	0.101 (0.051)[96.0%]	0.301 (0.048)[95.8%]	-0.202 (0.052)[94.9%]	0.678 (0.030)[74.6%]
	t (df=20)	0.200 (0.035)[93.9%]	0.400 (0.036)[94.3%]	0.099 (0.051)[95.4%]	0.299 (0.046)[96.4%]	-0.200 (0.053)[94.5%]	0.692 (0.030)[87.5%]
	normal	0.200 (0.036)[94.9%]	0.400 (0.037)[92.7%]	0.098 (0.052)[93.6%]	0.298 (0.049)[95.4%]	-0.197 (0.052)[94.5%]	0.703 (0.030)[93.5%]

Table 2: Summary of simulation results. The true parameter values are displayed in the second row. For each sample size and innovation degree of freedom, the average estimates are reported, together with their sample standard deviations (enclosed in round brackets) and the empirical coverage rates of their nominal 95% confidence intervals (in square brackets).

Data were simulated from a CARX model with independent 2-dimensional standard normal covariates and AR(3) errors with t-distributed innovations of degrees of freedom equal to 5, 10, 20, or ∞ (i.e. normal distribution), and sample size equal to 100, 200 or 400. See Table 2 for the true parameter values. The responses were censored whenever their magnitude exceeds some threshold that makes a censoring rate of approximately 20%. The innovation distributions are scaled to ensure that they have identical unit standard deviation. Computation of the parametric confidence intervals were based on 500 bootstraps, and each experiment was replicated 1000 times. The results summarized in Table 2 shows that the proposed method is robust to heavier tails in the innovation distribution than the normal distribution, as the estimates are comparable in terms of bias and standard deviation, across the range of degrees of freedom, and the empirical coverage rates are all close to the nominal 95%, except for the parameter σ. That the estimator of σ is non-robust can be expected as this is the case even if there is no censoring, which highlights a future problem for developing more robust estimation of σ.

3.5 The Ljung-Box test statistic based on simulated residuals

Next, we report some simulation results on the empirical performance of the Ljung-Box test statistic, based on the simulated residuals, that is used as a tool for model diagnostics. The null model is an $AR(2)$ model with a two-dimensional covariate comprising two independent standard normal variables, with $\beta^\intercal = (0.2, 0.4)$, $\psi^\intercal = (0.28, 0.25)$, and $\sigma = 0.6$. We computed the Ljung-Box statistic using the first 10 or 20 lags of the sample autocorrelation function (ACF) of the simulated residuals. For assessing the power of the test, the AR part of the model is embedded in an AR(3) model with the AR operator given by $(1 - \delta B)(1 - \Psi(B))$, where $\delta = 0.0, 0.1, 0.2, \cdots, 0.8$. The data are left-censored so as to achieve a long-run censoring rate of 15%, or 30%, plus the case of no censoring as a benchmark, with sample size 100, or 200. The empirical rejection rates based on 1000 replications are shown in Figure 4. The sizes of the test under all settings are quite close to the nominal 5% level. The power generally increases with greater deviation of δ from zero, lesser censoring, larger sample size and fewer lags used in computing the Ljung-Box statistic. That using more lags in the test resulted in lower power is expected owing to the geometric decay of the ACF so that its higher lags quickly become non-informative.

3.6 An application to the total phosphorus concentration in river water

Phosphorus is one of the two nutrients of main concern in Iowa river water, as excessive phosphorus in river water can result in eutrophication. Phosphorus concentration in river water has been closely monitored under the ambient water quality program conducted by the Iowa Department of Natural Resources (Libra, Wolter & Langel 2004). Here, we analyze a series of 120 monthly phosphorus concentration (P) in mg/l, in river water collected at an ambient site located at Whitebreast Creek near Knoxville, Iowa, USA, from October 1998 to March 2010. There is a gap of missing P data from September 2008 to March 2009, when data collection was suspended owing to lack of funding. The data were censored when P fell below certain detection limits c_t (red line in Figure 5) that varied over time, resulting in about 10% censoring. It is known that P is generally correlated with the water discharge (Q) (Schilling, Chan, Liu & Zhang 2010). The main interest is to explore the relationship between P and Q with censored P data. The Q data were obtained from the website of the U.S. Geological Survey. See Figure 5 for the time plots of P, Q, and the historical censoring limits.

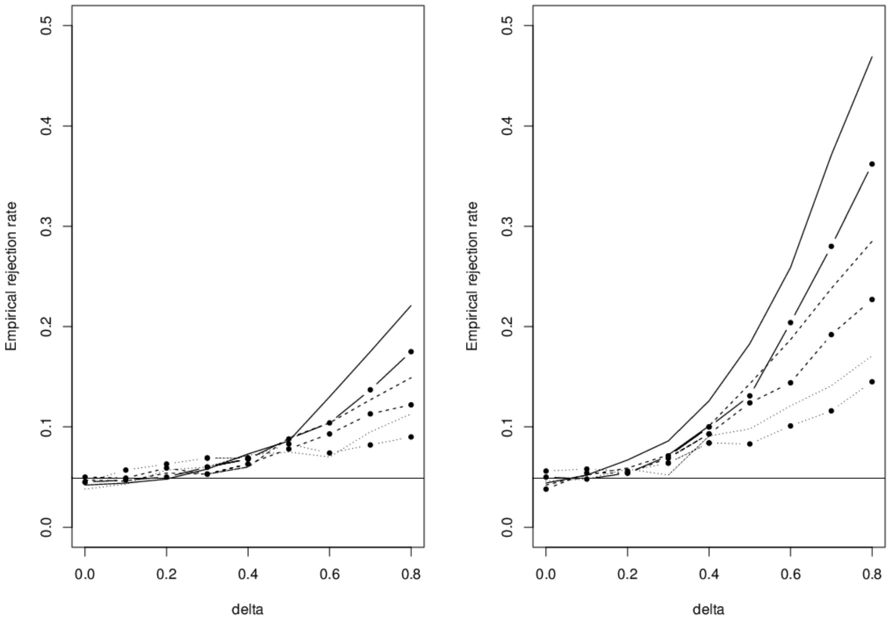

Figure 4: Empirical rejection rates of the Ljung-Box test with (simulated) residuals. Sample size is 100 (200) in the Left (right) diagram. Empirical rejection rates are connected with a solid line for the test using the first 10 lags of residual ACF and no censoring, while those from experiments with left-censored data at long-run censoring rate of 15% (30%) are connected with a dashed (dotted) line, with a horizontal solid line indicating the nominal 5% size. Results based on first 20 lags of residual ACF are similarly plotted except that the rates are superimposed as solid circles.

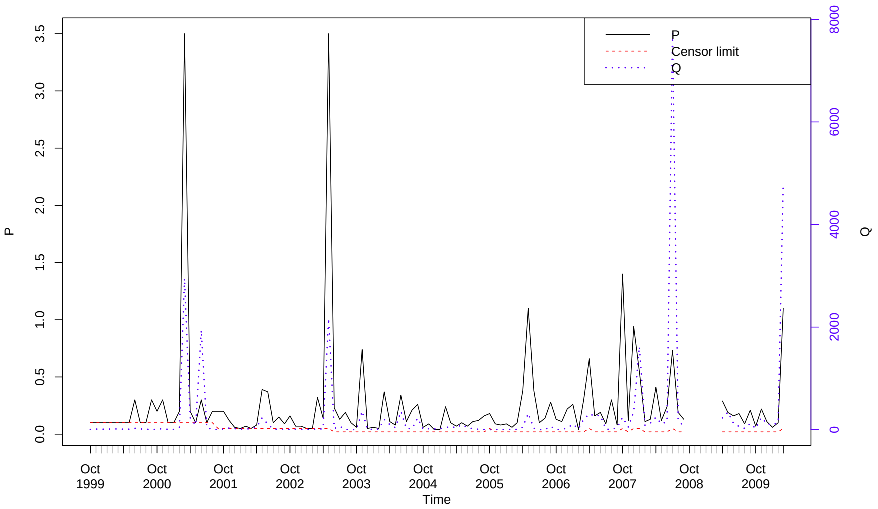

Figure 5: Time series plots of P (black solid line, scale shown on the left vertical axis), Q (blue dotted line, scale shown on the right vertical axis) and the censoring limits c_t (red dashed line, in the same scale as that of P).

Preliminary analysis (unreported) shows that taking the logarithmic transformation for both P and Q renders their relationship more linear. Model diagnostics with a preliminary linear regression model $\log(P_t) = \beta_0 + \beta_1 \log(Q_t) + \eta_t$ indicates (i) the presence of serial residual autocorrelation, hence we model η_t as an autoregressive process, and (ii) that the P-Q relationship is seasonal. We model the seasonal relationship by introducing the dummy seasonal dummy variables $S_j, j = 1, 2, 3, 4$ for quarters 1 to 4 and their interactions with discharge $\log(Q_t)$, where the first quarter comprises January to March, the second quarter from April to June, etc. In summary, the model is

$$\log(P_t) = \sum_{j=1}^{4} \{\beta_{0,j} S_{j,t} + \beta_{1,j} S_{j,t} \times \log(Q_t)\} + \eta_t,$$

where the regression errors $\{\eta_t\}$ follow an autoregressive model as defined in Eq (2) with ε_t independent and identically distributed as $N(0, \sigma^2)$. The coefficients $\beta_{0,1}$ and $\beta_{1,1}$ are the intercept and slope for quarter 1, $\beta_{0,2}$ and $\beta_{1,2}$ those for quarter 2, etc.

Since the AR order is unknown and the seasonal P-Q relationship is uncertain, we fitted a number of models with or without seasonal P-Q relationship, altogether 8 models, and the AR order from 1 to 3. Model selection was then carried out by using an information criterion similar to the AIC with the log-likelihood replaced by the conditional expectation defined in Eq (8). A seasonal regression model with $AR(2)$ regression errors was selected. The final model fit is summarized in Table 3; the parametric bootstrap 95% confidence intervals are based on 1000 replicates. The P-Q relationships in quarters 2 and 4 are quite similar. Indeed, constraining the regression coefficients to be identical for these two quarters slightly reduces the AIC from 17.027 to 17.013. The rate of change in $\log(P)$ per unit change in $\log(Q)$ is highest in quarter 1 and lowest in quarter 3; see Figure 6 which shows the simulated partial residual plot of $\log(Q)$ (c.f. Subsection 2.6), with the fitted quarterly linear relationships superimposed in the diagram. These found relationships are consistent with the fact that discharge is generally lowest in quarter 1 and highest in quarter 3. The AR estimates are moderate in values, suggesting rather short memory in the data. Figure 7 plots the ACF of the simulated residuals, which suggests no residual autocorrelation. The Ljung-Box test statistic using the first 10 lags of the residual ACF is 5.28 with p-value 0.27, suggesting no serial autocorrelation in the residuals and that the model provides a good fit to the data. Finally, Figure 8 plots the time plot of P and the exponentiation of the fitted values, showing that the fitted values generally track the data well, but less so for the larger P peaks.

As an illustration of prediction with a censored time series, we re-fitted the selected model with the data excluding the last 6 observations that are withheld for assessing the real prediction performance. The point predictors and their 95% prediction intervals, computed using the procedure elaborated in Section 2.5 and assuming normal innovations, are shown in Figure 9, with the actual data superimposed on the diagram. Overall, the prediction tracks the data movement reasonably well although the last three data points are somewhat close to the lower prediction limits. We note that the prediction results (unreported) are similar with the innovations bootstrapped by drawing randomly with replacement from the simulated residuals.

Parameter	Estimate (Confidence Interval)
$\beta_{0,1}$	-4.26 (-4.8, -3.7)
$\beta_{0,2}$	-3.83 (-4.7, -3.0)
$\beta_{0,3}$	-2.54 (-3.0, -2.1)
$\beta_{0,4}$	-3.36 (-3.9, -2.8)
$\beta_{1,1}$	0.605 (0.47, 0.72)
$\beta_{1,2}$	0.489 (0.32, 0.67)
$\beta_{1,3}$	0.192 (0.07, 0.32)
$\beta_{1,4}$	0.478 (0.33, 0.64)
ψ_1	0.281 (0.04, 0.45)
ψ_2	0.254 (0.01, 0.44)
σ	0.593 (0.48, 0.64)

Table 3: Estimated parameters and their 95% bootstrap confidence intervals.

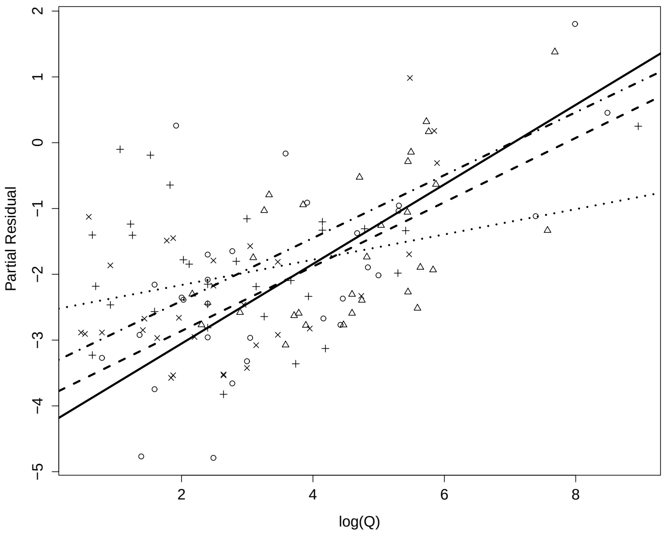

Figure 6: Partial residual plot for log(Q), with a partial residual drawn as an open circle (triangle, plus, cross), if it belongs to quarter 1 (2,3,4). The four lines display the quarterly linear relationship, with solid, dashed, dotted, dot-dashed lines for quarters 1,2,3,4, respectively.

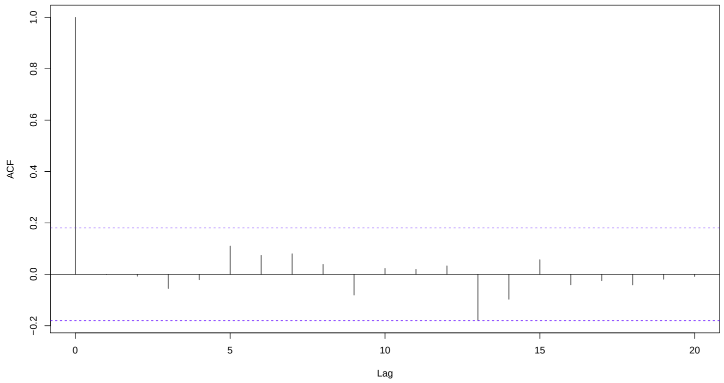

Figure 7: The ACF plot of the simulated residuals.

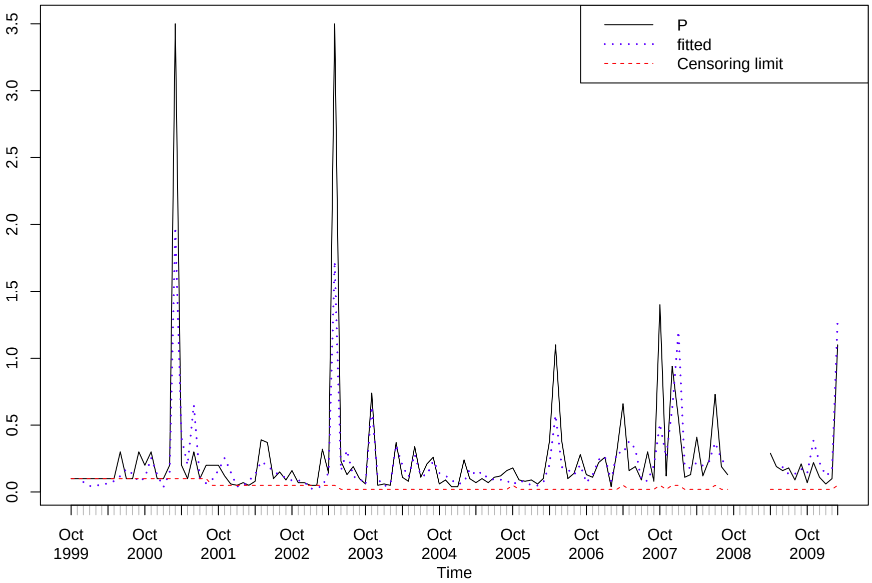

Figure 8: Time plot of P and the exponentiation of the fitted values.

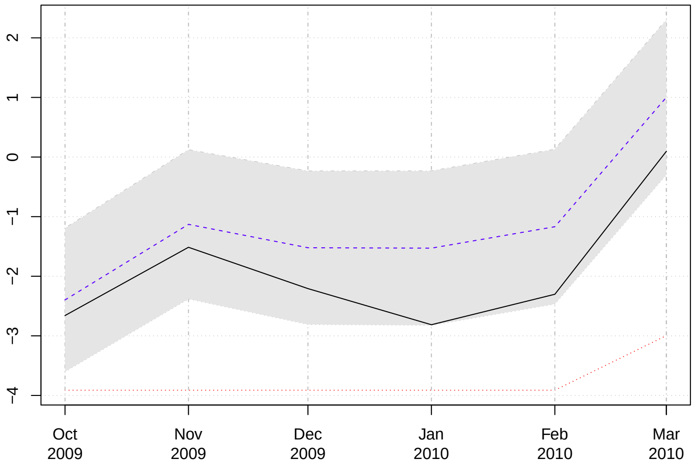

Figure 9: Time series plot of $\log(P)$ and their predicted values. The observed values (predicted values, censoring limits) are connected by a black solid (blue dashed, red dashed) line. The 95% prediction band is shaded in grey.

4 DISCUSSION AND CONCLUSION

We have proposed a new method to estimate a censored regression model with autoregressive errors. The consistency and asymptotic properties are established under some general conditions. The proposed method can be readily implemented for the case of normal innovations. Simulation studies indicate that the proposed method enjoys excellent sampling properties, whereas ignoring censoring results in substantial bias when the censoring rate is moderately high. We illustrate the efficacy of the proposed method by analyzing the seasonal phosphorus-discharge relationship with a phosphorus concentration data. Some interesting future work include extension of the proposed method to estimating a censored regression model with autoregressive moving average regression errors and studying the theoretical properties of the proposed estimator when the covariate process X_t is non-stationary, for instance, in a trend analysis. It is also of practical importance to study the estimation method in the case of non-Gaussian innovation distributions. Also of interest is to study the theoretical properties of the simulated residuals in the current setting.

5 APPENDIX

5.1 Numerical intractability of maximum likelihood estimation

Recall the method proposed by Zeger & Brookmeyer (1986) makes use of the fact that for a censored autoregressive process $\{Y_t\}$ of order p, the conditional distribution of Y_t given all past Y's is the same as that given the past Y's up to p consecutive, uncensored observations. Hence, the likelihood function for the data $Y_i, i = 1, \ldots, n$ can be simplified by first dividing the data into blocks using the following recursive scheme. Let $t_0 = n+1$. Given $t_i, 0 \leq i \leq k$. If $t_k = 1$, the recursive definition of the t_i's is done. Otherwise, set t_{k+1} to be the largest $t < t_k - 1$ such that the p consecutive Y's before time t, i.e. $\{Y_{t-j}, 1 \leq j \leq p\}$, are uncensored and if no such p consecutive Y's exist, set $t_{k+1} = 1$. Since t_k is strictly decreasing, only finitely many t_i's are defined. Suppose $t_K = 1$. Then the data can be blocked into K blocks, namely, $B_i = \{Y_t, t_i \leq t < t_{i-1}\}, i = 1, 2, \ldots, K$. The likelihood function is then equal to the likelihood of the last block B_K times the product of the conditional likelihood of B_i given $Y_{t_i-j}, 1 \leq j \leq p$, for $1 \leq i < K$. The derivation of the latter conditional likelihood generally involves integration, either analytically or numerically, which becomes more complex with the block dimension. Specifically, suppose the ith block B_i comprises censored observations $\{Y_t, t \in C_i\}$ and uncensored observations $\{Y_t, t \in O_i\}$. Then, the conditional likelihood of B_i given $Y_{t_i-j}, 1 \leq j \leq p$ is obtained by integrating the conditional density of $\{Y_t, t \in C_i\}$ given $\{Y_t, t \in O_i\} \cup \{Y_{t_i-j}, 1 \leq j \leq p\}$, over the censoring regions corresponding to $\{Y_t, t \in C_i\}$, and multiplying the integral with the conditional density of $\{Y_t, t \in O_i\}$ given $\{Y_{t_i-j}, 1 \leq j \leq p\}$. The conditional likelihood is complex for high block dimension because the conditional density may involve the parameters nonlinearly and a high-dimensional integration may be required. In the ideal case that none of the Y's are censored, the blocks are of unit dimension, except for the last block. In the presence of censoring, the blocks are of random dimension. The distribution of the block dimension is generally quite complex. As a benchmark, assuming independent censoring and a constant censoring rate, say π (which holds if the uncensored process is independent and identically distributed, and censoring occurs when the underlying process falls within some fixed interval), the mean block dimension, excluding the last block B_K, can be shown to be $\{1-(1-\pi)^p\}/\{\pi(1-\pi)^p\}-p+1$, with variance equal to $\{1-(2p+1)\pi(1-\pi)^p-(1-\pi)^{2p+1}\}/\{\pi^2(1-\pi)^{2p}\}$; see Philippou, Georghiou & Philippou (1983). For instance, for $p=5$ and a 25% censoring rate, the mean block dimension is ≈ 8.9 with standard deviation about 9.3. So, relatively high dimensional integration may be required even for $p=5$ to carry out maximum likelihood estimation, rendering the method increasingly computationally intensive with the censoring rate.

5.2 Proof of Theorem 2.1

Proof. First, note that the function $Z_t(\theta)$ is continuously differentiable with respect to θ. By A1, a consistent initial estimate for the parameter is assumed, so the parameter space is restricted to some compact neighbourhood Θ of θ_0, which is assumed to satisfy some conditions specified below.

The functional central limit theorem of Arcones & Yu (1994) will be used to prove

$$\left\{\sqrt{n}\left(Z^{(n)}(\theta)-\mathrm{E}_{\theta_0}\left[Z_t(\theta)\right]\right):\theta\in\Theta\right\}\xrightarrow{\mathcal{L}}\{G(\theta):\theta\in\Theta\},$$

where $\{G(\theta)\}$ is a Gaussian process which has a version with uniformly bounded and uniformly continuous sample paths with respect to the L^2 norm. See also Chan & Tsay (1998). In order to apply the alluded functional central limit theorem to $\{Z_t\}$, we need to verify that

1 the class of functions $\{Z_t(\theta),\theta\in\Theta\}$ is a Vapnick-Cervonenkis subgraph class of measurable functions,

2 there exists an envelope function $F\in L^p$ for some $p>2$ such that $|Z_t(\theta)|\le F$ for all $\theta\in\Theta$, and

3 The process $\{(Y_t^*,X_t,c_{t,l},c_{t,u})\}$ is stationary, β-mixing with a geometrically decaying β-mixing rate.

The first two conditions are essentially stated by A6. The third condition is ensured by A2–A5.

We verify the consistency result by using Theorem 5.9 of Van der Vaart (2000), with the key steps established below. Firstly, we show that θ_0 is a zero of $Z(\theta)$:

$$\begin{aligned}
&\mathrm{E}_{\theta_0}\left[Z_t(\theta_0)\right]\\
=&\mathrm{E}_{\theta_0}\left[\mathrm{E}_{\theta_0}\left[\frac{\partial \ell(Y_t^*|\mathcal{F}_t^*,\theta)}{\partial\theta}|_{\theta=\theta_0}|\mathcal{G}_t\right]\right]\\
=&\mathrm{E}_{\theta_0}\left[\frac{\partial \ell(Y_t^*|\mathcal{F}_t^*,\theta)}{\partial\theta}|_{\theta=\theta_0}\right]\\
=&\mathrm{E}_{\theta_0}\left[\mathrm{E}_{\theta_0}\left[\frac{\partial \ell(Y_t^*|\mathcal{F}_t^*,\theta)}{\partial\theta}|_{\theta=\theta_0}|\mathcal{F}_t^*\right]\right]\\
=&\mathrm{E}_{\theta_0}\left[0\right]\\
=&0.
\end{aligned}$$

As $Z(\theta) = Z(\theta_0) + \nabla Z(\theta_0)^\intercal(\theta-\theta_0) + o(\|\theta-\theta_0\|)$ around θ_0, the non-singularity of ∇Z at θ_0, owing to A7, implies that, by shrinking the compact neighborhood Θ of θ_0 if necessary, $Z(\theta) \neq 0$, for all $\theta \in \Theta$ and $\theta \neq \theta_0$. Hence, it holds that for all sufficiently small $\epsilon > 0$, $\inf_{\theta\in\Theta,\|\theta-\theta_0\|>\varepsilon} \|\mathrm{E}_{\theta_0}\left[Z_t(\theta)\right]\| > 0$. The uniform convergence

$$\sup_{\theta\in\Theta}\|Z^{(n)}(\theta) - Z(\theta)\| \xrightarrow{P} 0$$

follows from the functional limit theorem for $Z_t(\theta)$. The consistency property then follows from Theorem 5.9 of Van der Vaart (2000).

The asymptotic normality result can be verified by using similar techniques in the proof of Theorem 5.21 of Van der Vaart (2000) and Arcones & Yu (1994).

□

5.3 Verification of Assumptions A6 and A7 for the case of normal innovations

Proposition 5.1. *Assumption A6 holds if the innovations are normally distributed,* $\exp(\epsilon\|X_t\|^2) \in L^1$ *for some small* $\epsilon > 0$*, and the parameter space is a sufficiently small neighborhood of* θ_0*. Furthermore, under the additional conditions that (a)* $E_{\theta_0}[(X_t - \sum_{i=1}^p \psi_{0,i}X_{t-i})(X_t - \sum_{i=1}^p \psi_{0,i}X_{t-i})^\intercal]$ *is non-singular, (b) the pdf of* $X_{t:t-p}$ *is an analytic function and (c) assuming one-sided censoring with a constant censoring limit c, Assumption A7 holds for almost all real-valued c, with the set of exceptional c's being at most countably infinite.*

The additional condition (a) requires that there exists no non-trivial linear combination of $X_t - \sum_{i=1}^{p} \psi_{0,i} X_{t-i}$ that is identically zero, which holds if the the components of $X_{t-i}, 0 \leq i \leq p$ are not exactly collinear. A real-valued function with a multi-dimensional argument is an analytic function if it coincides with its Taylor series expansion locally (Krantz & Parks 2012). Condition (b) is a mild condition which holds for many distributions including normal and t-distributions (Golubev, Levit & Tsybakov 1996).

Proof. First note that the score function has the following form,

$$S(Y_t^*|\mathcal{F}_t^*, \theta) = \frac{1}{\sigma^2} \begin{bmatrix} \varepsilon_t (X_t - \sum_{j=1}^{p} \psi_j X_{t-j}) \\ \varepsilon_t \eta_{t-1:t-p} \\ \frac{1}{2}\left(\frac{\varepsilon_t^2}{\sigma^2} - 1\right) \end{bmatrix} \tag{13}$$

where $\eta_t = Y_t^* - X_t^\intercal \beta$ and $\varepsilon_t = (1 - \Psi(B))\eta_t$.

Expanding each term in the score function, it is seen that each element in the vector $Z_t(\theta)$ can be written as some linear combination of the following terms

$$\mathrm{E}_\theta \left[(Y_{t-i}^*)^j | \mathcal{G}_t \right] \|X_{t-l}\|^m, \tag{14}$$

where (i) $i, l = 0, \ldots, p$; (ii) $j, m = 0, 1, 2$; and (iii) $j + m \leq 2$.

To show the V-C property for $Z_t(\theta)$, it suffices to show that $\mathrm{E}_\theta \left[(Y_{t-i}^*)^j | \mathcal{G}_t \right]$ is a V-C class, then so is $Z_t(\theta)$. For truncated normal distributions, Tallis (1961) gives the closed-form expressions for the moments of first and second orders, from which the V-C property can be deduced. (Note that the V-C property is preserved by forming finite products and sums of elements of a V-C class.)

It remains to find an envelope function for the score functions $Z_t(\theta)$. It suffices to first find some envelope functions for the expression defined by Eq (14).

As the main difficulty lies in bounding $\mathrm{E}_\theta \left[(Y_{t-i}^*)^j | \mathcal{G}_t \right]$, we first present a general result about conditional expectation and change of measure. Suppose we have a conditional expectation $\mathrm{E}_\theta \left[W | \mathcal{G} \right]$ where W is some random variable and $\mathcal{G}$ is a relevant σ-algebra, and θ is the parameter indexing the underlying probability measure to be denoted as P_θ, where θ lies in N_{θ_0}, a neighborhood of a known parameter θ_0. We want to find simple conditions sufficient for the existence of an envelope function for $\mathrm{E}_\theta \left[W | \mathcal{G} \right]$ for $\theta \in N_{\theta_0}$ that has finite absolute q-th moment under P_{θ_0}.

Suppose $P_\theta, \theta \in N_{\theta_0}$ are pairwise mutually absolutely continuous so that $\frac{dP_{\theta_1}}{dP_{\theta_2}}$ is well-defined for all $\theta_1, \theta_2 \in N_{\theta_0}$. The following lemma is instrumental, whose proof is deferred to later.

Lemma 5.2.

$$E_\theta(W|\mathcal{G}) = E_{\theta_0}\left(W\frac{dP_\theta}{dP_{\theta_0}}|\mathcal{G}\right) E_\theta\left(\frac{dP_{\theta_0}}{dP_\theta}|\mathcal{G}\right). \tag{15}$$

Setting $W \equiv 1$ in (15) yields the identity

$$E_\theta\left(\frac{dP_{\theta_0}}{dP_\theta}|\mathcal{G}\right) = E_{\theta_0}^{-1}\left(\frac{dP_\theta}{dP_{\theta_0}}|\mathcal{G}\right).$$

Hence,

$$E_\theta(W|\mathcal{G}) = E_{\theta_0}\left(W\frac{dP_\theta}{dP_{\theta_0}}|\mathcal{G}\right) E_{\theta_0}^{-1}\left(\frac{dP_\theta}{dP_{\theta_0}}|\mathcal{G}\right).$$

Jensen's inequality then implies that

$$|E_\theta(W|\mathcal{G})| \le E_\theta(|W||\mathcal{G}) = E_{\theta_0}\left(|W|\frac{dP_\theta}{dP_{\theta_0}}\bigg|\mathcal{G}\right) E_{\theta_0}^{-1}\left(\frac{dP_\theta}{dP_{\theta_0}}\bigg|\mathcal{G}\right).$$

As the function $f(x) = 1/x, x > 0$ is convex, Jensen's inequality entails that

$$|E_\theta(W|\mathcal{G})| \le E_{\theta_0}\left(|W|\frac{dP_\theta}{dP_{\theta_0}}|\mathcal{G}\right) E_{\theta_0}\left(\frac{dP_{\theta_0}}{dP_\theta}|\mathcal{G}\right). \tag{16}$$

We shall assume that the neighborhood N_{θ_0} is chosen such that there exists a random variable H of finite absolute r-th moment under P_{θ_0} and such that $\frac{dP_{\theta_1}}{dP_{\theta_2}} \le H$ for all $\theta_1, \theta_2 \in N_{\theta_0}$. It then follows from (16) that

$$\sup_{\theta\in N_{\theta_0}} |E_\theta(W|\mathcal{G})| \le E_{\theta_0}\left(|W|H|\mathcal{G}\right) E_{\theta_0}\left(H|\mathcal{G}\right). \tag{17}$$

The right side of (17) is then an envelope function of $\{E_\theta(W|\mathcal{G}), \theta \in N_{\theta_0}\}$.

We now consider the case of normally distributed innovations in our model, in which case the conditional distribution of $Y^*_{t:t-p}$ given the X's is multivariate normal $N(x^\intercal_{t:t-p}\beta, \Sigma)$, where Σ is determined by ψ and σ. Let $y = y_{t:t-p}$ and $\mu_i = x^\intercal_{t:t-p}\beta_i, i = 0, 1, 2$. Direct calculation yields

$$\begin{aligned}\frac{dP_{\theta_1}}{dP_{\theta_2}}(y) =& \exp\left\{y^\intercal\Sigma_1^{-1}(\mu_1-\mu_2) - \frac{1}{2}(\mu_1+\mu_2)^\intercal\Sigma_1^{-1}(\mu_1-\mu_2)\right\}\\ &\times\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right)^{(p+1)/2}\exp\left\{\frac{1}{2}(y-\mu_2)^\intercal\Sigma_1^{-1}(\Sigma_1-\Sigma_2)\Sigma_2^{-1}(y-\mu_2)\right\}\\ \le& c\exp\left\{\epsilon(\|y-\mu_0\|^2 + \|x_{t:t-p}\|^2)\right\}\\ :=& H,\end{aligned}$$

where c is some constant independent of θ_1, θ_2, and ϵ can be arbitrarily small as long as the neighborhood N_{θ_0} is sufficiently small. Then or any $r > 0$,

$$\begin{aligned}\mathrm{E}_{\theta_0}\left[H^r\right] =& \mathrm{E}_{\theta_0}\left[\mathrm{E}_{\theta_0}\left[H^r \middle| x_{t:t-p}\right]\right] \\ =& c\mathrm{E}_{\theta_0}\left[\exp(r\epsilon\|x_{t:t-p}\|^2)\right]\mathrm{E}_{\theta_0}\left[\exp(r\epsilon\|y-\mu_0\|^2)\middle| x_{t:t-p}\right] \\ =& c\mathrm{E}_{\theta_0}\left[\exp(r\epsilon\|x_{t:t-p}\|^2)\right]\mathrm{E}_{\theta_0}\left[\exp(r\epsilon\|y-\mu_0\|^2)\right],\end{aligned}$$

which is finite when ϵ is small enough by the assumption that $\exp(r\epsilon\|X_t\|^2) \in L^1$.

It follows from $\mathrm{E}\left[\exp(\epsilon\|X_t\|^2)\right] < \infty$ for some $\epsilon > 0$ that X_t has finite moments of any finite order. An envelope function for the expression in Eq (14) is $\mathrm{E}_{\theta_0}\left(|Y^*_{t-i}|^j H|\mathcal{G}_t\right)\mathrm{E}_{\theta_0}\left(H|\mathcal{G}_t\right)\|X_{t-l}\|^m$. Because Y^*_{t-j} given the X's is normal, it admits finite moments of any order. It follows from the generalized Hölder inequality and by making the neighborhood N_{θ_0} sufficiently small that $|Y_{t-i}|^j H\mathrm{E}_{\theta_0}[H|\mathcal{G}_t]\|X_{t-l}\|^m$ is L^q and so is the envelope function.

Next we prove that with one-sided censoring and constant censoring limit, $\mathrm{E}_{\theta_0}\left[\nabla Z_t(\theta_0)\right]$ can only be singular for at most countably infinitely many censoring limit values. Without loss of generality, consider the case of left censoring with l being the left censoring limit. First we claim that

$$\mathrm{E}_\theta\left[\nabla Z_t(\theta)\right] = -\mathrm{E}_\theta\left[\mathrm{E}_\theta[S(Y^*_t|\mathcal{F}^*_t,\theta)|\mathcal{G}_t]\mathrm{E}_\theta[S(Y^*_{t:t-p}|X_{t:t-p},\theta)|\mathcal{G}_t]^\intercal\right], \tag{18}$$

the proof of which we outline for the case with AR(1) errors and no covariates, as the proof for the general case is similar. Let $S_{1,t} = S(Y^*_t|\mathcal{F}^*_t,\theta) = S(Y^*_t|Y^*_{t-1},\theta)$, $S_{2,t} = S(Y^*_t, Y^*_{t-1},\theta) = \nabla\log f(Y^*_t, Y^*_{t-1},\theta)$ where we abuse the notation $f(Y^*_t, Y^*_{t-1},\theta)$ to stand for the joint pdf of Y^*_t, Y^*_{t-1} assuming θ is the parameter. Let $S_{3,t} = S(Y^*_{t-1};\theta) = \nabla\log f(Y^*_{t-1},\theta)$, so $S_{2,t} = S_{1,t} + S_{3,t}$. For any pair $\mathbf{y}_t = (y_t, y_{t-1})^\intercal$, there exists a unique region $R = R(\mathbf{y}_t)$ such that $(Y^*_t, Y^*_{t-1})^\intercal \in R$ if and only if the corresponding censored observation is $\mathbf{y}_t$. For instance if both y_t, y_{t-1} are left censored, $R = (-\infty, l] \times (-\infty, l]$ whereas if y_t is censored but y_{t-1} is not, then $R = (-\infty, l] \times \{y_{t-1}\}$. If both of them are not censored, $R = \{y_t\} \times \{y_{t-1}\}$, a singleton. Thus, the sigma algebra $\mathcal{G}_t$ is equivalent to that generated by $(Y^*_t, Y^*_{t-1})^\intercal \in R(\mathbf{y}_t)$.

Then, for $(Y^*_t, Y^*_{t-1}) \in R$ with $R = (-\infty, l] \times (-\infty, l]$, i.e., both y_t and y_{t-1} are censored, and omitting $dy_t^* dy_{t-1}^*$ in the following integrals for simplicity, we have

$$\begin{aligned}\nabla Z_t =& \nabla \left(\int_R \nabla \log f(y_t^*|y_{t-1}^*;\theta) \frac{f(y_t^*, y_{t-1}^*;\theta)}{\int_R f(y_t^*, y_{t-1}^*;\theta)} \right) \\ =& \int_R \nabla^2 \log f(y_t^*|y_{t-1}^*;\theta) \frac{f(y_t^*, y_{t-1}^*;\theta)}{\int_R f(y_t^*, y_{t-1}^*;\theta)} \\ &+ \int_R \nabla \log f(y_t^*|y_{t-1}^*;\theta)(\nabla \log f(y_t^*, y_{t-1}^*;\theta))^\intercal \frac{f(y_t^*, y_{t-1}^*;\theta)}{\int_R f(y_t^*, y_{t-1}^*;\theta)} \\ &- \int_R \nabla \log f(y_t^*|y_{t-1}^*;\theta) f(y_t^*, y_{t-1}^*;\theta) \frac{\int_R (\nabla f(y_t^*, y_{t-1}^*;\theta))^\intercal}{\{\int_R f(y_t^*, y_{t-1}^*;\theta)\}^2} \\ =& \mathrm{E}_\theta \left[\nabla S_{1,t}|\mathcal{G}_t\right] + \mathrm{E}_\theta \left[S_{1,t} S_{2,t}^\intercal|\mathcal{G}_t\right] - \mathrm{E}_\theta \left[S_{1,t}|\mathcal{G}_t\right] \mathrm{E}_\theta \left[S_{2,t}^\intercal|\mathcal{G}_t\right]. \end{aligned} \tag{19}$$

It can be readily verified that the last equality in the preceding display continues to hold even if R is 1-dimensional or a singleton, in which case the integral becomes a 1-dimensional integral with respect to the Lebesgue measure, or the counting measure on a singleton, respectively. Taking expectation on both sides of (19) yields

$$\begin{aligned}\mathrm{E}_\theta \nabla Z_t(\theta) &= \mathrm{E}_\theta \left[\nabla S_{1,t}\right] + \mathrm{E}_\theta \left[S_{1,t} S_{2,t}^\intercal\right] + \mathrm{E}_\theta \left[\mathrm{E}_\theta \left[S_{1,t}|\mathcal{G}_t\right] \mathrm{E}_\theta \left[S_{2,t}^\intercal|\mathcal{G}_t\right]\right] \\ &= \mathrm{E}_\theta \left[\nabla S_{1,t}\right] + \mathrm{E}_\theta \left[S_{1,t}(S_{1,t} + S_{3,t})^\intercal\right] - \mathrm{E}_\theta \left[\mathrm{E}_\theta \left[S_{1,t}|\mathcal{G}_t\right] \mathrm{E}_\theta \left[S_{2,t}^\intercal|\mathcal{G}_t\right]\right],\end{aligned}$$

where $\nabla S_{1,t}$ is evaluated at θ, etc., hence (18), because $\mathrm{E}_\theta \left[\nabla S_{1,t}\right] + \mathrm{E}_\theta \left[S_{1,t} S_{1,t}^\intercal\right] = 0$ and $\mathrm{E}_\theta \left[S_{1,t} S_{3,t}^\intercal\right] = 0$. This completes the proof of the claim in the special case.

Even for fixed θ, all expectations are implicit functions of the censoring limit l. Below, we write $E_l(\cdot) = E_{\theta,l}(\cdot)$ to emphasize its dependence on l, but l will be suppressed whenever the expectation does not depend on l. We now verify the result concerning the non-singularity of $\mathrm{E}_{\theta_0,l}\left[\nabla Z_t(\theta_0)\right]$, except for countably many l. We prove this by showing that for fixed θ, (i) the components of $\mathrm{E}_l\left[\nabla Z_t(\theta)\right]$ are real analytic functions and hence so is the determinant $d_\theta(l) = \det\left\{\mathrm{E}_{\theta,l}\left[\nabla Z_t(\theta)\right]\right\}$ and (ii) for fixed θ_0, $\lim_{l \to -\infty} d_{\theta_0}(l)$ exists and is non-zero, under the condition that $\mathrm{E}_\theta[(X_t - \sum_{i=1}^p \psi_i X_{t-i})(X_t - \sum_{i=1}^p \psi_i X_{t-i})^\intercal]$ is non-singular at $\theta = \theta_0$. Note the latter expectation does not depend on l. Results (i) and (ii) then imply that $\mathrm{E}_{\theta_0,l}\left[\nabla Z_t(\theta_0)\right]$ is non-singular, except for at most countably many l, otherwise the roots of $d_{\theta_0}(l) = 0$ admit an accumulation point prompting $d_{\theta_0}(l) \equiv 0$, hence $\lim_{l \to -\infty} d_{\theta_0}(l) = 0$, resulting in a contradiction. Here, we have relied on the fact that the roots of a non-constant analytic function cannot have an accumulation point.

We first verify (ii). Let $R_l = \{y_{t:t-p} : y_{t-i} > l, i = 0, 1, \ldots, p\}$ and I_{R_l} ($I_{R_l^c}$) the indicator function of R_l (its complement). Note $y_{t:t-p} \in R_l$ if and only if none of y_{t-i} are censored, for $i = 0, 1, \ldots, p$. Hence, on R_l, $\mathrm{E}_{\theta,l}[S(Y_t^*|\mathcal{F}_t^*, \theta)|\mathcal{G}_t] = S(Y_t^*|\mathcal{F}_t^*, \theta)$ and $\mathrm{E}_{\theta,l}[S(Y_{t:t-p}^*|X_{t:t-p}, \theta)|\mathcal{G}_t] = S(Y_{t:t-p}^*|X_{t:t-p}, \theta)$. It follows from Eqn. (18) that

$$\begin{aligned}
&\mathrm{E}_{\theta,l}\left[\nabla Z_t(\theta)\right] \\
&= -\mathrm{E}_{\theta,l}\left[S(Y_t^*|\mathcal{F}_t^*, \theta)S(Y_{t:t-p}^*|X_{t:t-p}, \theta)^\intercal I_{R_l}\right] \\
&\quad - \mathrm{E}_{\theta,l}\left[\mathrm{E}_{\theta,l}[S(Y_t^*|\mathcal{F}_t^*, \theta)|\mathcal{G}_t]\mathrm{E}_{\theta,l}[S(Y_{t:t-p}^*|X_{t:t-p}, \theta)|\mathcal{G}_t]^\intercal I_{R_l^c}\right]. \qquad (20)
\end{aligned}$$

As $l \to -\infty$, it follows from the Lebesgue dominated convergence theorem, with $|S(Y_{t:t-p}^*|X_{t:t-p}, \theta)^\intercal S(Y_t^*|\mathcal{F}_t^*, \theta)|$ as the component-wise dominating function, that the first term on the right side of the preceding display converges to

$$\begin{aligned}
&- \mathrm{E}_\theta\left[S(Y_t^*|\mathcal{F}_t^*, \theta)S(Y_{t:t-p}^*|X_{t:t-p}, \theta)^\intercal\right] \\
&= -\mathrm{E}_\theta\left[S(Y_t^*|\mathcal{F}_t^*, \theta)S(Y_t^*|\mathcal{F}_t^*, \theta)^\intercal\right],
\end{aligned}$$

which equals the block diagonal matrix consisting of the following diagonal blocks $-\mathrm{E}_\theta\left[(X_t - \sum_{i=1}^p \psi_i X_{t-i})(X_t - \sum_{i=1}^p \psi_i X_{t-i})^\intercal\right]$, $-E_\theta\left[\eta_{t-1:t-p}\eta_{t-1:t-p}^\intercal\right]$ and $-1/(2\sigma^4)$; c.f. (13). Note that A3 and A4 imply that $-E_\theta\left[\eta_{t-1:t-p}\eta_{t-1:t-p}^\intercal\right]$ is always strictly negative definite. On the other hand the second term tends to zero as $l \to -\infty$, which can be shown as follows. For any vector or matrix v, let $\|v\|^2$ its squared Euclidean norm and $|v|$ the vector obtained from taking the absolute value component-wise. Upon noting that $I_{R_l^c}$ is $\mathcal{G}_t$-measurable, Cauchy-Schwartz and Jensen inequalities entail that

$$\begin{aligned}
&\mathrm{trace}\left(\left|\mathrm{E}_{\theta,l}\left\{\mathrm{E}_{\theta,l}[S(Y_t^*|\mathcal{F}_t^*, \theta)|\mathcal{G}_t]\mathrm{E}_{\theta,l}[S(Y_{t:t-p}^*|X_{t:t-p}, \theta)|\mathcal{G}_t]^\intercal I_{R_l^c}\right\}\right|\right) \\
&\le \mathrm{E}_\theta^{1/2}[\|S(Y_t^*|\mathcal{F}_t^*, \theta)\|^2 I_{R_l^c}] \times \mathrm{E}_\theta^{1/2}[\|S(Y_{t:t-p}^*|X_{t:t-p}, \theta)\|^2 I_{R_l^c}],
\end{aligned}$$

which converges to 0 as $l \to -\infty$, verifying the claim regarding the second term in the decomposition, and thence (ii).

We sketch the proof for (i) only in the simple case $p = 1$ and no covariates, as the proof for the general case is similar. Letting $R_{l,1} = \{y_t^* > l, y_{t-1} > l\}$, $R_{l,2} = \{y_t^* \le l, y_{t-1} > l\}$,

$R_{l,3} = \{y_t^* > l, y_{t-1} \le l\}$, and $R_{l,4} = \{y_t^* \le l, y_{t-1} \le l\}$, Eqn (18) becomes

$$\begin{aligned}
&\mathrm{E}_{\theta,l}\left[\nabla Z_t(\theta)\right] \\
&= -\mathrm{E}_{\theta,l}\left[\mathrm{E}_{\theta,l}[S(Y_t^*|\mathcal{F}_t^*,\theta)|\mathcal{G}_t]S(Y_{t:t-1}^*,\theta)]^\intercal\right] \\
&= -\sum_{i=1}^{4}\mathrm{E}_{\theta,l}\left[\mathrm{E}_{\theta,l}[S(Y_t^*|\mathcal{F}_t^*,\theta)|\mathcal{G}_t]S(Y_{t:t-p}^*,\theta)^\intercal I_{R_{l,i}}\right]. \qquad (21)
\end{aligned}$$

That $\mathrm{E}_{\theta,l}\left[\nabla Z_t(\theta)\right]$ is an (component-wise) analytic function of l follows from verifying the said property for each summand in the preceding decomposition. For instance, after some algebra, the second summand equals

$$\int_l^\infty \frac{\int_{-\infty}^l S(y_t^*|y_{t-1}^*,\theta)f(y_t^*,y_{t-1}^*,\theta)dy_t^* \times \int_{-\infty}^l S(y_t^*,y_{t-1}^*,\theta)f(y_t^*,y_{t-1}^*,\theta)dy_t^*}{\int_{-\infty}^l f(y_t^*,y_{t-1}^*,\theta)dy_t^*}dy_{t-1}^*, \qquad (22)$$

which is an analytic function of l. This can be verified by noticing (1) $S(y_t^*|\mathcal{F}_t^*,\theta)$, $S(Y_{t:t-1}^*,\theta)$ are polynomials, and thus analytic functions of y_t^*, y_{t-1}^*, (2) $f(y_{t:t-1}^*,\theta)$, being the jointly normal density of y_t^*, y_{t-1}^*, is analytic with respect to y_t^*, y_{t-1}^*, (3) sum, product and composition of analytic functions are analytic and so is the ratio of an analytic function to a positive analytic function, and (4) the indefinite integral of an analytic function with respect to any variable is analytic, see Propositions 2.2.2 and 2.2.3 in Krantz & Parks (2012). Other summands can be similarly shown to be analytic functions of l, which completes the proof. □

Proof of Lemma 5.2. Note that the conditional expectation $E_\theta(W|\mathcal{G})$ is characterized by (i) $E_\theta(W|\mathcal{G})$ is $\mathcal{G}$-measurable function and (ii) for all $A \in \mathcal{G}$, the following equality holds:

$$\int_A W dP_\theta = \int_A E_\theta(W|\mathcal{G})dP_\theta.$$

Consider the following display, where A is an arbitrary element in $\mathcal{G}$:

$$\begin{aligned}
\int_A W dP_\theta &= \int_A W\frac{dP_\theta}{dP_{\theta_0}}dP_{\theta_0} \\
&= \int_A E_{\theta_0}\left(W\frac{dP_\theta}{dP_{\theta_0}}|\mathcal{G}\right)dP_{\theta_0} \\
&= \int_A E_{\theta_0}\left(W\frac{dP_\theta}{dP_{\theta_0}}|\mathcal{G}\right)\frac{dP_{\theta_0}}{dP_\theta}dP_\theta \\
&= \int_A E_{\theta_0}\left(W\frac{dP_\theta}{dP_{\theta_0}}|\mathcal{G}\right)E_\theta\left(\frac{dP_{\theta_0}}{dP_\theta}|\mathcal{G}\right)dP_\theta.
\end{aligned}$$

The claim follows from the preceding equality and that the right side of (15) is $\mathcal{G}$-measurable.

□

References

Amemiya, T. (1973), "Regression analysis when the dependent variable is truncated normal," *Econometrica: Journal of the Econometric Society*, pp. 997–1016.

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

Buckley, J., & James, I. (1979), "Linear regression with censored data," *Biometrika*, 66(3), 429–436.

Chan, K.-S., & Tsay, R. S. (1998), "Limiting properties of the least squares estimator of a continuous threshold autoregressive model," *Biometrika*, 85(2), 413–426.

Cox, D. R., & Snell, E. J. (1968), "A general definition of residuals," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 248–275.

Cryer, J. D., & Chan, K. S. (2008), *Time series analysis: with applications in R*, New York: Springer.

Genz, A., & Bretz, F. (2009), *Computation of Multivariate Normal and t Probabilities*, Lecture Notes in Statistics, Heidelberg: Springer-Verlag.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2014), *mvtnorm: Multivariate Normal and t Distributions.* R package version 1.0-0.

Golubev, Y. K., Levit, B. Y., & Tsybakov, A. B. (1996), "Asymptotically efficient estimation of analytic functions in Gaussian noise," *Bernoulli*, pp. 167–181.

Gourieroux, C., Monfort, A., Renault, E., & Trognon, A. (1987), "Simulated residuals," *Journal of Econometrics*, 34(1), 201–252.

Hillis, S. L. (1995), "Residual plots for the censored data linear regression model," *Statistics in medicine*, 14(18), 2023–2036.

Jawitz, J. W. (2004), "Moments of truncated continuous univariate distributions," *Advances in water resources*, 27(3), 269–281.

Krantz, S., & Parks, H. (2012), *A Primer of Real Analytic Functions* Birkhäuser: Boston.

Libra, R. D., Wolter, C. F., & Langel, R. J. (2004), *Nitrogen and phosphorus budgets for Iowa and Iowa watersheds* Iowa Department of Natural Resources, Geological Survey.

Park, J. W., Genton, M. G., & Ghosh, S. K. (2007), "Censored time series analysis with autoregressive moving average models," *Canadian Journal of Statistics*, 35(1), 151–168.

Pham, T. D., & Tran, L. T. (1985), "Some mixing properties of time series models," *Stochastic processes and their applications*, 19(2), 297–303.

Philippou, A. N., Georghiou, C., & Philippou, G. N. (1983), "A generalized geometric distribution and some of its properties," *Statistics & Probability Letters*, 1(4), 171–175.

R Core Team (2015), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing.

Ripley, B. D. (2002), "Time series in R 1.5. 0," *R News*, 2(2), 2–7.

Robinson, P. M. (1980), Estimation and forecasting for time series containing censored or missing observations,, in *Time Series: Proceedings of the International Conference, Held at Nottingham University, March, 1979*, ed. O. D. Anderson, North Holland, pp. 167–182.

Robinson, P. M. (1982*a*), "On the asymptotic properties of estimators of models containing limited dependent variables," *Econometrica: Journal of the Econometric Society*, pp. 27–41.

Robinson, P. M. (1982*b*), "Analysis of time series from mixed distributions," *The Annals of Statistics*, pp. 915–925.

Schilling, K. E., Chan, K. S., Liu, H., & Zhang, Y. K. (2010), "Quantifying the effect of land use land cover change on increasing discharge in the Upper Mississippi River," *Journal of Hydrology*, 387(3), 343–345.

Tallis, G. M. (1961), "The Moment Generating Function of the Truncated Multi-normal Distribution," *Journal of the Royal Statistical Society. Series B (Methodological)*, 23(1), 223–229.

Tobin, J. (1958), "Estimation of relationships for limited dependent variables," *Econometrica: journal of the Econometric Society*, pp. 24–36.

Van der Vaart, A. W. (2000), *Asymptotic statistics*, Vol. 3, Cambridge: Cambridge university press.

Zeger, S. L., & Brookmeyer, R. (1986), "Regression Analsis with Censored Autocorrelated Data," *Journal of the American Statistical Association*, 81(395), 722–729.