UNICAMP

UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Matemática, Estatística
e Computação Científica

FERNANDA LANG SCHUMACHER

CENSORED REGRESSION MODELS WITH AUTOREGRESSIVE ERRORS

MODELOS DE REGRESSÃO CENSURADOS COM ERROS AUTORREGRESSIVOS

CAMPINAS
2016

Fernanda Lang Schumacher

Censored regression models with autoregressive errors

Modelos de regressão censurados com erros autorregressivos

Dissertation presented to the Institute of Mathematics, Statistics and Scientific Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Statistics.

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestra em Estatística.

Orientador: Prof. Dr. Víctor Hugo Lachos Dávila

Coorientador: Prof. Dr. Filidor Edilfonso Vilca Labra

Este exemplar corresponde à versão final da dissertação defendida pela aluna Fernanda Lang Schumacher, e orientada pelo Prof. Dr. Víctor Hugo Lachos Dávila.

Campinas
2016

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Maria Fabiana Bezerra Muller - CRB 8/6162

Informações para Biblioteca Digital

Título em outro idioma: Modelos de regressão censurados com erros autorregressivos
Palavras-chave em inglês:
Censored observations (Statistics)
Correlated errors (Statistics)
Linear models (Statistics)
Influential observations (Statistics)
Local influence (Regression models)
Área de concentração: Estatística
Titulação: Mestra em Estatística
Banca examinadora:
Filidor Edilfonso Vilca Labra [Coorientador]
Celso Rômulo Barbosa Cabral
Airlane Pereira Alencar
Data de defesa: 04-11-2016
Programa de Pós-Graduação: Estatística

Dissertação de Mestrado defendida em 04 de novembro de 2016 e aprovada

Pela Banca Examinadora composta pelos Profs. Drs.

Prof(a). Dr(a). FILIDOR EDILFONSO VILCA LABRA

Prof(a). Dr(a). CELSO RÔMULO BARBOSA CABRAL

Prof(a). Dr(a). AIRLANE PEREIRA ALENCAR

A Ata da defesa com as respectivas assinaturas dos membros
encontra-se no processo de vida acadêmica do aluno.

Acknowledgement

Resumo

Dados de séries temporais são facilmente encontrados em diversas áreas, incluindo monitoramento do meio ambiente, medicina, economia e ciências sociais, e muitas vezes apresentam autocorrelação. Uma complicação adicional surge quando as medições de séries temporais apresentam irregularidades, tais como observações submetidas a limites de detecção superiores ou inferiores, acima e abaixo dos quais as medições não são quantificáveis, e observações faltantes. Pesquisadores comumente desconsideram os casos censurados ou substituem estas observações por alguma função do limite de detecção, o que muitas vezes resulta em estimativas tendenciosas. Neste trabalho, nós estudamos alguns aspectos de estimação e análise de influência local em modelos de regressão censurados com erros autorregressivos de ordem p (modelos AR(p)-CR). As estimativas de máxima verossimilhança (ML) dos parâmetros são obtidas usando uma aproximação estocástica do algoritmo EM (SAEM). Esta abordagem nos permite estimar os parâmetros de interesse de forma eficiente. Como um subproduto, o algoritmo SAEM permite previsões dos valores não observáveis da variável resposta. A matriz de informação observada é derivada analiticamente para a obtenção dos erros padrões. Além disso, técnicas de diagnóstico de influência local são derivadas para modelos AR(p)-CR com base na função Q sob três esquemas de perturbação. O desempenho dos métodos em amostras finitas é avaliado por meio da análise de vários estudos de simulação e de aplicações em dois conjuntos de dados reais. O algoritmo e métodos propostos são implementados no novo pacote do R *ARCensReg*.

Palavras-chave: Algoritmo SAEM; Dados censurados; Influência local; Limite de detecção; Modelos autorregressivos AR(p); Observações influentes.

Abstract

Time series data are frequently encountered in diverse fields, including environmental monitoring, medicine, economics and social science, and they are often autocorrelated rather than independent. An additional complication arises when time series measurements are observed with data irregularities, such as observations subjected to upper or lower detection limits, below and above which they are not quantifiable, and missing observations. Practitioners commonly disregard censored data cases or replace these observations with some function of the limit of detection, which often results in biased estimates. In this work, we study some aspects of estimation and local influence analysis in censored regression models with autoregressive errors of order p (hereafter, AR(p)-CR models). The estimates of maximum likelihood (ML) of the parameters are obtained using a stochastic approximation of the EM (SAEM) algorithm. This approach allows for easy and fast estimation of the parameters of autoregressive models when censoring is present. As a byproduct, the SAEM algorithm enables predictions of unobservable values of the response variable. The observed information matrix is derived analytically to account for standard errors. Furthermore, local influence diagnostic measures are derived for the AR(p)-CR model on the basis of the Q-function under three usual perturbation schemes. The finite sample performance of the methods is evaluated through the analysis of several simulation studies and its applications to two real datasets. The proposed algorithm and methods are implemented in the new R package *ARCensReg*.

Keywords: Autoregressive AR(p) models; Censored data; Influential observations; Limit of detection; Local influence; SAEM algorithm.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Observations collected over time are often autocorrelated rather than independent, so time series data analysis must deal with temporally collected observations by modeling their autocorrelations. Autoregressive (AR) models for time series data developed by Box et al. (1994) have been widely used as a basic approach. However, modeling AR data can present an additional challenge, from the statistical point of view, in which the observations can be subject to upper or lower detection limits, below and above which they are not quantifiable. For example, environmental monitoring of different variables often involves left-censored observations falling below the minimum limit of detection (LOD) of the instruments used to quantify them. The concentration of a certain mineral in river water can be an important indicator about water quality, and its fluctuations over time are often monitored in environmental studies. However, the mineral concentration cannot be measured exactly if it falls below certain detection limits. In these studies, the proportion of censored data may not be small, so the use of crude/ad hoc methods, such as discarding the censored observations or substituting a threshold value or some arbitrary point like a midpoint between zero and cutoff for detection (LOD/2), might lead to biased estimates of fixed effects related with exogenous variables and autoregressive components. Thus the key issue in estimating the parameters of time series models with exogenous variables based on censored data is to obtain estimates that are at least (asymptotically) unbiased and more efficient than some of the two ad-hoc methods described above.

As an alternative to crude imputation methods, Robinson (1980) suggested imputing the censored part with its conditional expectation, given the completely observed part. Since the conditional expectation has the form of multiple incomplete integrals, his method groups the data vector so that each subgroup includes one censored observation, and thus requires a single integral. However, this method may not be feasible for many consecutive censored observations. Zeger and Brookmeyer (1986) suggested a full likelihood estimation and approximate method for an autoregressive time series model. However, the authors pointed out that the method may not be feasible when the censoring rate is too high. Hopke et al. (2001) used multiple imputation

based on a Bayesian approach. However, little explanation was provided about the theoretical properties of the estimators, such as unbiasedness and efficiency. More recently, Park et al. (2007) presented an alternative method for handling censored data in the setting of AR models. In this method, time series data are regarded as a realization from a multivariate normal distribution, and the censored values are then imputed using the conditional multivariate normal distribution given the observed part. Nevertheless, this method is not exactly a likelihood-based method since the censored observations are first imputed and then any estimation procedure for complete time-series data is used.

Even though some solutions have been proposed in the literature to deal with the problem of censored responses in AR models, there are no studies conducting exact inferences for censored AR models from a likelihood-based perspective. In this work, we aim to derive a computationally efficient estimation method via the stochastic version of the expectation-maximization (SAEM) algorithm in censored regression models with autoregressive errors of order p (hereafter, AR(p)-CR model). The SAEM algorithm was initially proposed by Delyon et al. (1999) using maximum likelihood (ML) techniques as a powerful alternative to the expectation-maximization (EM) when the E-step is intractable. The SAEM algorithm has been proved to be more computationally efficient than the classic Monte Carlo EM (MCEM) algorithm due to the recycling of simulations from one iteration to the next in the smoothing phase of the algorithm. Moreover, as pointed out in Meza et al. (2012), the SAEM algorithm, unlike the MCEM, converges even in a typically small simulation size.

Furthermore, in the interest of evaluating the influence of small deviations on the data, we develop local influence diagnostic techniques for the AR(p)-CR model, following the method proposed by Zhu and Lee (2001) to assess the local influence in a minor perturbation of a statistical model with incomplete data, utilizing Cook's approach to the conditional expectation of the complete-data log-likelihood function in the EM algorithm. Although there are some works in the literature discussing influence diagnostic techniques for regression models with autoregressive errors, to the best of our knowledge, there are no previous studies of diagnostic analysis for censored linear regression models with autoregressive errors.

1.1 Motivation

Given the importance of correctly analysing autocorrelated censored data, our motivation is the analysis of two real datasets described next.

1.1.1 Cloud ceiling height dataset

The first dataset is a meteorological time series of cloud ceiling height previously analyzed by Park et al. (2007). The cloud ceiling height is defined as the distance from the ground to the bottom of a cloud and is measured in hundreds of feet. According to Park et al. (2007), an accurate determination of the cloud ceiling height is important mainly because it is one of the major factors contributing to weather-related accidents and one of the major causes of flight delays. The recording device has a detection limit of $12,000$ feet, so the observed data can be considered a right-censored time serie.

The data were originally collected by the National Center for Atmospheric Research (NCAR) based on hourly observations in San Francisco, recorded during the month of March 1989, consisting of 716 observations. The censoring rate is 40.5%. Figure 1.1 shows the log-transformed data whose original scale is in hundreds of feet. Considering this transformation the limit of detection is $\log(120) = 4.787$.
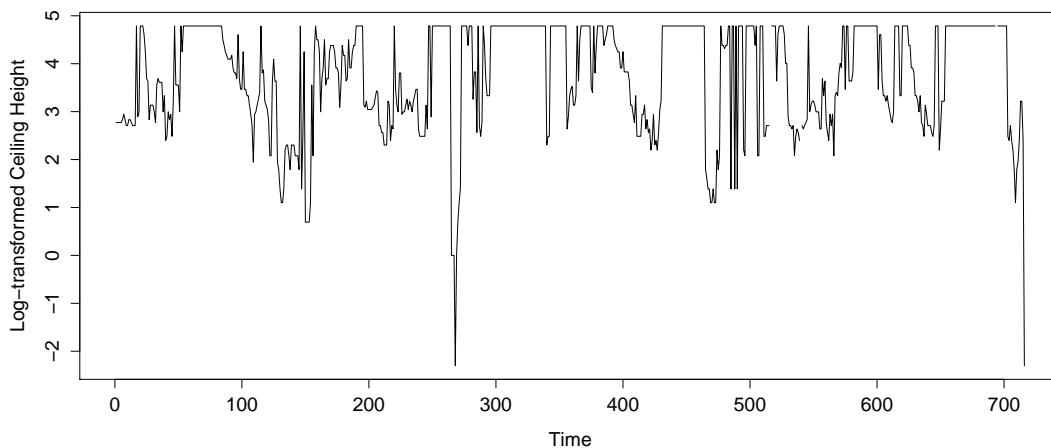
Figure 1.1: Censored time series of log-transformed hourly cloud ceiling height in San Francisco during March 1989.

Park et al. (2007) analysed this dataset using four different approaches, the two first being naive approaches (to discard the censored observation *(a)* and to treat the censored values as observed *(b)*), the third being a Bayesian approach *(c)* and the last being the method proposed in his paper *(d)*: imputing the censored values by considering the time series data as a realization from a multivariate normal distribution and then, once the augmented dataset is obtained, estimating the parameters from autoregressive moving average (ARMA) models by any suitable method. Table 1.1 presents the estimates for the selected ARMA model obtained by Park et al. (2007) for the four approaches. The ARMA(p,q) model is defined as

$$Y_t = \beta_0 + \phi_1(Y_{t-1} - \beta_0) + \ldots + \phi_p(Y_{t-p} - \beta_0) + \eta_t - \psi_1\eta_{t-1} - \ldots - \psi_q\eta_{t-q},$$

where β_0 is the mean parameter, $\phi_1, \ldots, \phi_p$ are the autoregressive parameters and $\psi_1, \ldots, \psi_q$ are the moving average parameters. The error process $\{\eta_t\}$ is assumed to be white noise with mean 0 and variance σ^2.

Table 1.1: Parameter estimates and standard errors obtained by Park et al. (2007) from the selected ARMA model for the cloud ceiling height dataset.

	Approach			
	(a)	*(b)*	*(c)*	*(d)*
Selected Model	ARMA(1,1)	AR(1)	AR(2)	AR(2)
$\widehat{\beta_0}$	3.704 (0.159)	3.004 (0.147)	4.194 (0.407)	4.129 (0.236)
$\widehat{\phi_1}$	0.872 (0.024)	0.841 (0.027)	0.740 (0.049)	0.689 (0.038)
$\widehat{\phi_2}$	-	-	0.159 (0.049)	0.173 (0.038)
$\widehat{\psi_1}$	0.243 (0.046)	-	-	-
$\widehat{\sigma}$	0.723	0.562	1.035	0.877

Source: Park et al. (2007)

1.1.2 Total phosphorus concentration dataset

According to Wang and Chan (2016a) phosphorus is one of the two nutrients of main concern in Iowa river water, as excessive phosphorus in river water can result in eutrophication. It is known that phosphorus concentration (P) is generally correlated with the water discharge (Q) (Schilling et al., 2010).

Phosphorus concentration data of West Fork Cedar River at Finchford, Iowa, USA, that were collected under the ambient water quality program conducted by the Iowa Department of Natural Resources (Iowa DNR) are available in the *R* package *carx* (R Development Core Team, 2016; Wang and Chan, 2016b). The dataset has 15.5% of left censored observations and it was recorded monthly from 10/1998 to 10/2013, being measured in mg/l. A gap from 09/2008 to 03/2009 in the data is due to program suspension owing to lack of funding.

Water discharge data can be obtained from the website of U.S. Geological Survey, measured in cubic feet per second. A similar dataset was studied by Wang and Chan (2016a), who proposed a quasi-likelihood estimation method for censored autoregressive models with exogenous variables.

As a preliminary study, we fitted a censored regression model considering independent errors under the normal distribution using the *R* package *SMNCensReg* (R Development Core Team, 2016). Under this assumption we can calculate the transformation of the martingale residuals (Garay et al., 2015), defined by

$$r_{MT_t} = \text{sign}(r_{M_t})\sqrt{-2\left[r_{M_t} + \delta_t \log(\delta_t - r_{M_t})\right]}, \ t = 1, \ldots, n,$$
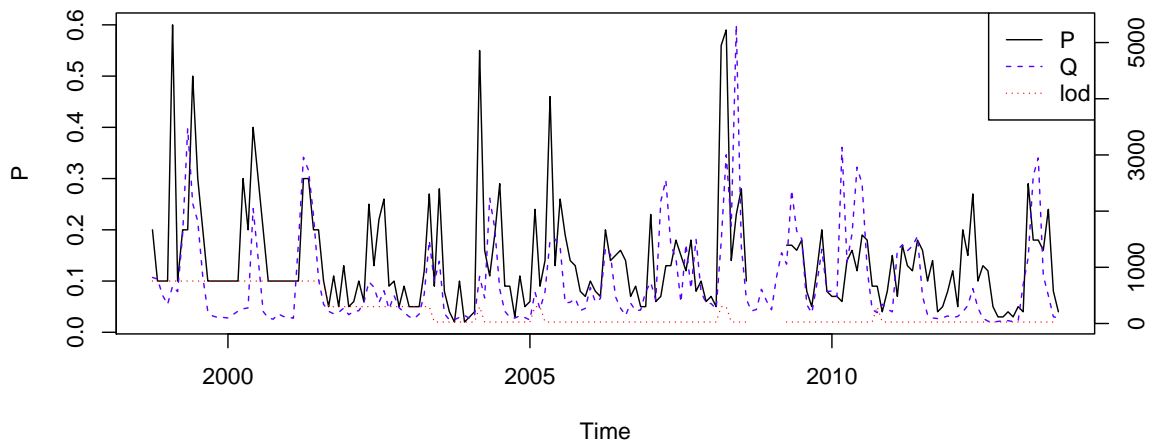
Figure 1.2: Censored time series of phosphorus concentration (black solid line, scale shown on the left vertical axis), water discharge (blue dashed line, scale shown on the right vertical axis) and limit of detection (red dotted line, in the same scale as P).

where $r_{M_t} = \delta_t + \log(S(y_t, \widehat{\boldsymbol{\theta}}))$ is the martingale residual proposed by Ortega et al. (2003), $\delta_t = 0, 1$ indicating whether the tth observation is censored or not, respectively, $\text{sign}(r_{M_t})$ denotes the sign of r_{M_t} and $S(y_t, \widehat{\boldsymbol{\theta}}) = P(Y_t > y_t \mid \widehat{\boldsymbol{\theta}})$ representing the survival function evaluated at y_t. Figure 1.3 presents the partial autocorrelations of transformation of the martingale residuals. It indicates that the assumption of independent errors is not valid, and hence a model with autocorrelated errors should be considered.
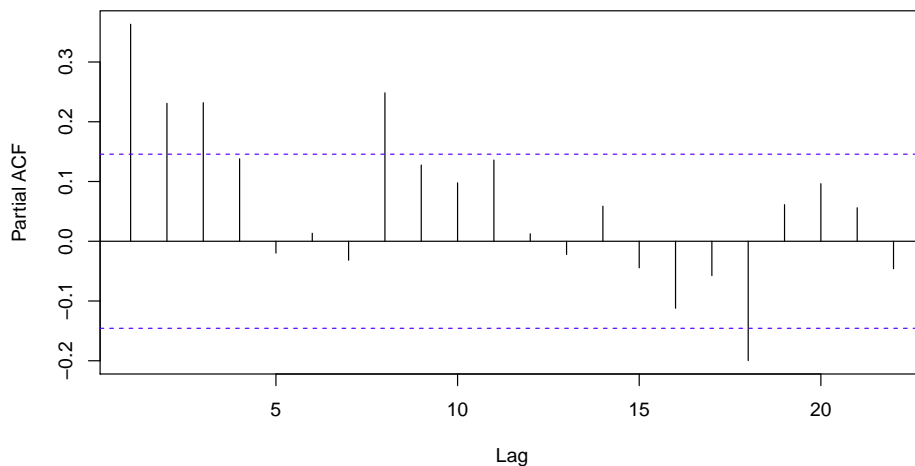
Figure 1.3: Partial autocorrelations of the transformation of the martingale residuals for the model with independent errors for the total phosphorus concentration dataset.

1.2 Censoring

In many practical situations we may not be able to observe the variable of interest $(Y_t, t = 1, \ldots, n)$ directly. Instead, we may observe only an interval in which the realization of this variable is contained. That is called censoring. Let V_t be the value that we observe at time t. In this work, we will deal with two types of censoring, such that

$$V_t = \begin{cases} \max(Y_t, c_t), & \text{in case of left censoring,} \\ \min(Y_t, c_t), & \text{in case of right censoring,} \end{cases} \tag{1.2.1}$$

where c_t is the censoring level or the detection limit at time t.

Notice that a left censoring structure causes a right truncation of the distribution, since we only know that the true observation y_t is less than or equal to the observed quantity V_t. Moreover, the right censored problem can be represented by a left censored problem by simultaneously transforming the response Y_t and censoring level V_t to $-Y_t$ and $-V_t$. Therefore, the methods will be presented considering left censoring.

1.3 The multivariate normal and truncated normal distributions

A random vector $\mathbf{Y}$ is said to have an n-variate normal distribution with mean vector $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ (positive definite), denoted by $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if the probability density function (pdf) of $\mathbf{Y}$ is given by

$$\phi_n(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right\}, \tag{1.3.1}$$

where

$$\boldsymbol{\mu} = \begin{pmatrix} E(Y_1) \\ \vdots \\ E(Y_n) \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \left[Cov(Y_i, Y_j) \right], i = 1, \ldots, n; j = 1 \ldots, n.$$

Let $\mathbf{Y}$ be partitioned as $\mathbf{Y} = \left(\mathbf{Y}_1^\top, \mathbf{Y}_2^\top \right)^\top$, such that $dim(\mathbf{Y}_1) = n_1$ and $dim(\mathbf{Y}_2) = n_2 = n - n_1$ and let

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

be the corresponding partitions. It can be shown that

(i) $\mathbf{Y}_1 \sim N_{n_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$,

(ii) $\mathbf{Y}_2 \mid \mathbf{Y}_1 = \mathbf{y}_1 \sim N_{n_2}\left(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\, \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1),\ \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\, \boldsymbol{\Sigma}_{11}^{-1}\, \boldsymbol{\Sigma}_{12}\right).$

Now, let $\Phi_n(\cdot \mid \boldsymbol{\mu},\, \boldsymbol{\Sigma})$ denote the corresponding cumulative distribution function (cdf) and let $TN_n(\boldsymbol{\mu},\, \boldsymbol{\Sigma}; \mathbb{A})$ denote the n-variate truncated normal distribution on the interval $\mathbb{A}$, where $\mathbb{A} = A_1 \times \ldots \times A_n$ and for the right truncation case $A_t = (-\infty, c_t]$, $t = 1, \ldots, n$. A random vector $\mathbf{Y}$ is said to follow an n-variate truncated normal distribution, denoted by $\mathbf{Y} \sim TN_n(\boldsymbol{\mu},\, \boldsymbol{\Sigma}; \mathbb{A})$, if its pdf is given by

$$f(\mathbf{y} \mid \boldsymbol{\mu},\, \boldsymbol{\Sigma}; \mathbb{A}) = \frac{\phi_n(\mathbf{y} \mid \boldsymbol{\mu},\, \boldsymbol{\Sigma})}{\Phi_n(\mathbf{c} \mid \boldsymbol{\mu},\, \boldsymbol{\Sigma})} \mathbb{1}_{\mathbb{A}}(\mathbf{y}), \tag{1.3.2}$$

where $\mathbf{c} = (c_1, \ldots, c_n)^\top$ and $\mathbb{1}_{\mathbb{A}}(\mathbf{y})$ is the indicator function such that

$$\mathbb{1}_{\mathbb{A}}(\mathbf{y}) = \left\{ \begin{array}{ll} 1, & \text{if } \mathbf{y} \in \mathbb{A} \\ 0, & \text{otherwise} \end{array} \right..$$

The multivariate normal pdf and cdf can be evaluated without much computational burden through the routine mvtnorm() available in R (see Genz et al., 2008; R Development Core Team, 2016). In addition, the truncated multivariate normal pdf and cdf, along with its first and second moments, can be computed through the routine tmvtnorm(), also available in R (see Wilhelm and G., 2015; R Development Core Team, 2016).

1.4 The autoregressive regression model of order p

Ignoring censoring for the moment, we consider the classic linear regression model by introducing autocorrelated errors, defined as a discrete time autoregressive process. The discrete time representation of this model for the observed response at time t is given by

$$\begin{array}{lll} Y_t & = & \mathbf{x}_t^\top \boldsymbol{\beta} + \varepsilon_t, \tag{1.4.1} \\ \varepsilon_t & = & \phi_1 \varepsilon_{t-1} + \ldots + \phi_p \varepsilon_{t-p} + \eta_t, \quad \eta_t \sim N(0, \sigma^2), \quad t = 1, \ldots, n, \tag{1.4.2} \end{array}$$

where Y_t is the response variable, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_l)^\top$ is a vector of regression parameters of dimension l and $\mathbf{x}_t^\top = (x_{t1}, \ldots, x_{tl})$ is a vector of non-stochastic regressor variable values, ε_t is the autoregressive error with Gaussian disturbance η_t and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^\top$ is the vector of autoregressive coefficients. The model defined in (1.4.1)-(1.4.2) will be denoted the AR(p)-LR model.

To ensure stationarity of the AR(p) model given in (1.4.2), the roots of

$$1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p = 0$$

must lie outside the unit circle, where B is the backshift operator, such that $B^j \varepsilon_t = \varepsilon_{t-j}$ for $j = 0, \ldots, p$. The regions of ϕ where the process is stationary will be called the admissible region.

Equivalently, in matrix notation, the AR(p)-LR model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{y} = (Y_1, \ldots, Y_n)^\top$, $\mathbf{X}$ is an $n \times l$ matrix of rows $\mathbf{x}_t^\top$ and $\boldsymbol{\epsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \sim N_n(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \sigma^2 \mathbf{M}_n(\phi)$, such that

$$\mathbf{M}_n(\phi) = \frac{1}{\sigma^2} \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{n-2} \\ \vdots & \vdots & & \vdots \\ \gamma_{n-1} & \gamma_{n-2} & \cdots & \gamma_0 \end{bmatrix}, \tag{1.4.3}$$

where $\gamma_0, \ldots, \gamma_{n-1}$ are the theoretical autocovariances of the process, and, for $k = 1, 2, \ldots$, $\rho_k = \gamma_k / \gamma_0$ are the theoretical autocorrelations of the process. Moreover, the coefficients $\phi_1, \ldots, \phi_p$ satisfy the Yule-Walker equations (Barndorff-Nielsen and Schou, 1973), that can be written as

$$\begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{pmatrix} = \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \cdots & \rho_{p-2} \\ \vdots & \vdots & & \vdots \\ \rho_{p-1} & \rho_{p-2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix}.$$

Besides, when $k = 0$ we have $\gamma_0 = \phi_1 \gamma_1 + \ldots + \phi_p \gamma_p + \sigma^2$. Using the fact that $\gamma_k = \gamma_0 \rho_k$, for $k = 1, 2, \ldots$, γ_0 can be written as

$$\gamma_0 = \frac{\sigma^2}{1 - \phi_1 \rho_1 - \ldots - \phi_p \rho_p}.$$

To ensure the admissibility of ϕ and stabilize the estimating procedure, we follow Barndorff-Nielsen and Schou (1973) to reparameterize ϕ as

$$\begin{aligned} \phi_p^{(p)} &= \pi_p, \\ \phi_v^{(p)} &= \phi_v^{(p-1)} - \pi_p \phi_{p-v}^{(p-1)}, \end{aligned} \tag{1.4.4}$$

where $\phi_v^{(p)}$ is the vth AR parameter under the AR(p) model given in (1.4.2), and $\pi_v = \phi_v^{(v)}$ is the partial autocorrelation at lag v, for $v = 1, \ldots, p-1$.

This recursion can be used to define a transformation

$$\mathscr{B} : (\pi_1, \ldots, \pi_p) \to (\phi_1, \ldots, \phi_p) \qquad (1.4.5)$$

that is one-to-one, continuous and differentiable inside the admissible region. This parameterization has the advantage that in the π-space the admissible region is simply the p-dimensional cube with boundary surfaces corresponding to ± 1, while in the ϕ-space it is very complicated (see for instance McLeod and Zhang, 2006). As an illustration, for $p = 2$, the transformation is simply $\phi_1 = \pi_1(1 - \pi_2)$ and $\phi_2 = \pi_2$. For $p = 3$, it can be written as $\phi_1 = \pi_1(1 - \pi_2) - \pi_2\pi_3$, $\phi_2 = \pi_2(1 + \pi_1\pi_3) - \pi_1\pi_3$ and $\phi_3 = \pi_3$.

Following Box et al. (1994), for the model presented in (1.4.1) and (1.4.2) the exact log-likelihood function is given by

$$\ell(\boldsymbol{\theta} \mid \mathbf{y}) = -\frac{1}{2}\left[n \log \sigma^2 + \frac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^\top \mathbf{M}_n^{-1}(\phi)(\mathbf{y} - \mathbf{X}\beta) + \log |\mathbf{M}_n(\phi)|\right] + C, \qquad (1.4.6)$$

where C is a constant independent of the parameter vector $\boldsymbol{\theta}$. Considering the reparameterization given in (1.4.4) and dropping constant terms (McLeod and Zhang, 2006), the log-likelihood function can be written as

$$\ell(\boldsymbol{\theta}|\mathbf{y}) = -\frac{1}{2}\left[n \log \sigma^2 + \frac{1}{\sigma^2}S(\boldsymbol{\pi}, \beta) + \log g_p\right], \qquad (1.4.7)$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)^\top$, $g_p = |\mathbf{M}_n(\phi)| = |\mathbf{M}_p(\phi)| = \prod_{j=1}^{p}(1 - \pi_j^2)^{-j}$ and

$$S(\boldsymbol{\pi}, \beta) = \boldsymbol{\lambda}^\top D(\mathbf{y}, \beta)\boldsymbol{\lambda},$$

with $D(\mathbf{y}, \beta)$ being the $(p+1) \times (p+1)$ matrix with the (i, j)-entry being the sum of $n - (i - 1) - (j - 1)$ squares and lagged products, defined by

$$D_{i,j} = D_{j,i} = (Y_i - \mathbf{x}_i^\top \beta)(Y_j - \mathbf{x}_j^\top \beta) + \ldots + (Y_{n+1-j} - \mathbf{x}_{n+1-j}^\top \beta)(Y_{n+1-i} - \mathbf{x}_{n+1-i}^\top \beta) \quad (1.4.8)$$

and $\boldsymbol{\lambda}^\top = (-1, \phi^\top) = (-1, \mathscr{B}(\boldsymbol{\pi})^\top)$.

The unknown model parameters can be estimated by maximizing the corresponding log-likelihood function. Since $\partial \ell(\boldsymbol{\theta}|\mathbf{y})/\partial \boldsymbol{\pi} = 0$ does not have an analytic solution for $\boldsymbol{\pi}$, the procedure to obtain the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ requires numerical approximations. Following De Bastiani et al. (2014), given $\boldsymbol{\pi}$, the log-likelihood function (1.4.7)

is maximized at

$$\widehat{\beta}(\boldsymbol{\pi}) = \left(\mathbf{X}^\top \mathbf{M}_n^{-1}\left(\mathscr{B}(\boldsymbol{\pi})\right)\mathbf{X}\right)^{-1}\mathbf{X}^\top \mathbf{M}_n^{-1}\left(\mathscr{B}(\boldsymbol{\pi})\right)\mathbf{y}, \qquad (1.4.9)$$

$$\widehat{\sigma^2}(\boldsymbol{\pi}) = \frac{1}{n}S\left(\boldsymbol{\pi},\widehat{\beta}(\boldsymbol{\pi})\right). \qquad (1.4.10)$$

Thus, substituting expressions (1.4.9)–(1.4.10) into the log-likelihood and dropping constant terms, we obtain a concentrated log-likelihood function:

$$\ell_c(\boldsymbol{\pi}|\mathbf{y}) = -\frac{n}{2}\log S\left(\boldsymbol{\pi},\widehat{\beta}(\boldsymbol{\pi})\right) - \frac{1}{2}\log g_p, \qquad (1.4.11)$$

which must be maximized numerically with respect to $\boldsymbol{\pi} \in (-1,1)^p$, the p-dimensional cube with boundary surfaces corresponding to ± 1, obtaining $\widehat{\boldsymbol{\pi}}$. Finally, the MLE of β and σ^2 are calculated as $\widehat{\beta} = \widehat{\beta}(\widehat{\boldsymbol{\pi}})$ and $\widehat{\sigma^2} = \widehat{\sigma^2}(\widehat{\boldsymbol{\pi}})$. We can also obtain estimates of the original parameter vector ϕ by setting $\widehat{\phi} = \mathscr{B}(\widehat{\boldsymbol{\pi}})$.

These estimates can be obtained through the R function *arima()* (R Development Core Team, 2016), using the argument *xreg* to declare the regressor variables.

1.5 Algorithms for ML estimation

In models with missing or incomplete data, the EM algorithm (Dempster et al., 1977) is a classical approach to obtain the maximum likelihood estimates. The main idea behind this algorithm is that, in some situations, the estimation of the parameters would be easy if the complete data were available, while it is difficult based on the incomplete data only. Moreover, it has many attractive features, such as numerical stability and simplicity of implementation (Couvreur, 1996). Letting $\mathbf{y}_c = (\mathbf{y}^c, \mathbf{y}^o)$ be the complete-data vector, where $\mathbf{y}^c$ represents the missing data and $\mathbf{y}^o$ the observed data respectively and, $\ell(\boldsymbol{\theta}|\mathbf{y}_c)$ be the complete-data log-likelihood function, then the EM-algorithm proceeds in two steps:

- **E-step:** Let $\widehat{\boldsymbol{\theta}}^{(k)}$ be the current kth step estimate of $\boldsymbol{\theta}$. By using the property of conditional expectation, we can compute the $\hat{Q}_k(\boldsymbol{\theta})$ function by

$$\hat{Q}_k(\boldsymbol{\theta}) = E\left[\ell(\boldsymbol{\theta}|\mathbf{y}_c)|\mathbf{y}^o, \widehat{\boldsymbol{\theta}}^{(k)}\right]. \qquad (1.5.1)$$

- **M-step:** Maximize $\hat{Q}_k(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, obtaining $\widehat{\boldsymbol{\theta}}^{(k+1)}$.

Each iteration of the EM algorithm increases the likelihood function $\ell(\boldsymbol{\theta}|\mathbf{y}^o)$, and the EM sequence $\{\widehat{\boldsymbol{\theta}}^{(k)}\}$ converges to a stationary point of the observed likelihood under mild regularity conditions (for more details see Wu (1983) and Vaida (2005)).

However, in some situations, the E-step has no analytic form or involves calculations computationally too expensive and has to be calculated using simulations. Wei and Tanner (1990) proposed the Monte Carlo EM (MCEM) algorithm, in which the E-step is replaced by a Monte Carlo approximation based on a large number of independent simulations of the missing data. Another alternative is to consider a stochastic approximation of the expectations, as proposed by Delyon et al. (1999) with the so-called SAEM algorithm, which seems to be more efficient than the MCEM algorithm because the number of required simulations is considerably smaller. The SAEM algorithm consists, at each iteration, of successively simulating the missing data with the conditional distribution, and updating the unknown parameters of the model. Thus, at iteration k, the SAEM method is performed as follows:

- **E-Step:**

    ◇ *Simulation:* Draw m samples of the missing data $\mathbf{y}^{c(k)}$ with the conditional distribution

    $$f(\mathbf{y}^c|\mathbf{y}^o, \widehat{\boldsymbol{\theta}}^{(k-1)}).$$

    ◇ *Stochastic Approximation:* Update $\hat{Q}_k(\boldsymbol{\theta})$ according to

    $$\hat{Q}_k(\boldsymbol{\theta}) = \hat{Q}_{k-1}(\boldsymbol{\theta}) + \delta_k \left(\frac{1}{m} \sum_{\ell=1}^{m} \ell(\boldsymbol{\theta}|\mathbf{y}^o, \mathbf{y}^{c(k)}) - \hat{Q}_{k-1}(\boldsymbol{\theta}) \right), \qquad (1.5.2)$$

    where δ_k is a decreasing sequence of positive numbers such that

    $$\sum_{k=1}^{\infty} \delta_k = \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \delta_k^2 < \infty, \qquad (1.5.3)$$

    as presented by Kuhn and Lavielle (2004).

- **M-Step:**

    ◇ *Maximization:* Update the estimate $\widehat{\boldsymbol{\theta}}^{(k)}$ according to

    $$\widehat{\boldsymbol{\theta}}^{k+1} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ \hat{Q}_k(\boldsymbol{\theta}).$$

This process is iterated until some distance between two successive parameter estimations, such as

$$\sqrt{\left(\widehat{\boldsymbol{\theta}}^{(k+1)} - \widehat{\boldsymbol{\theta}}^{(k)} \right)^{\top} \left(\widehat{\boldsymbol{\theta}}^{(k+1)} - \widehat{\boldsymbol{\theta}}^{(k)} \right)},$$

becomes small enough.

If the smoothing parameter δ_k is equal to 1 for all k, the SAEM algorithm will have "no memory", and will be equivalent to the MCEM algorithm. The SAEM with no memory will

converge quickly (convergence in distribution) to a neighborhood solution, but the algorithm with memory will converge slowly (almost sure convergence) to the ML solution. As proposed by Galarza et al. (2015), we will consider

$$\delta_k = \begin{cases} 1, & \text{if } 1 \leq k \leq cW; \\ \frac{1}{k-cW}, & \text{if } cW+1 \leq k \leq W, \end{cases} \qquad (1.5.4)$$

where W is the maximum number of iterations and c is a cutoff point ($0 \leq c \leq 1$) which determines the percentage of the initial iterations. Kuhn and Lavielle (2005) recommended choosing the number of initial iterations between 50 and 100.

Note that the SAEM algorithm performs a Monte Carlo E-step like MCEM, but with a small and fixed Monte Carlo sample size ($m \leq 10$), which is then combined with the previous simulations in a "smooth" way. According to Delyon et al. (1999), when the maximization step is much faster than the simulation step, one may set the number of simulations at $m = 1$.

1.5.1 Curved exponential family

Following Kuhn and Lavielle (2005), when the complete-data likelihood function $\ell(\boldsymbol{\theta}|\mathbf{y}_c)$ belongs to the curved exponential family the implementation of the SAEM algorithm is more straightforward. In this case, the complete-data likelihood function can be written as

$$\ell(\boldsymbol{\theta}|\mathbf{y}_c) = \exp\left\{-\Psi(\boldsymbol{\theta}) + \langle \tilde{S}(\mathbf{y}), \Phi(\boldsymbol{\theta}) \rangle\right\}, \qquad (1.5.5)$$

where Ψ and Φ denote two functions of the unknown parameter $\boldsymbol{\theta}$, $\langle \cdot, \cdot \rangle$ denotes the scalar product and $\tilde{S}(\mathbf{y})$ is known as the minimal sufficient statistics of the complete model. For this situation, Delyon et al. (1999) pointed out that the *Stochastic Approximation* step reduces to the approximation of the minimal sufficient statistics.

1.6 Diagnostic analysis

The statistical models are important tools to extract and understand essential features of a dataset. However, the parameter estimation can be greatly affected by the presence of influential observations in the data. In the case of correlated data, a main approach for the detection of influential observations is local influence techniques (Cook, 1986), used to assess the stability of the estimation outputs with respect to the model inputs.

Following the pioneering work of Cook (1986), this area of research has received considerable attention in the recent statistical literature; see, for example Zhu and Lee (2001), Lee and Xu (2004) and Osorio et al. (2007), amongst others. However, in many models, the

marginal log-likelihood function is complex and direct application of Cook's approach may be very difficult, since these measures involve the first and second partial derivatives of this function. The work of Zhu and Lee (2001) proposes a method to assess the local infuence in a minor perturbation of a statistical model with incomplete data, utilizing Cook's approach to the conditional expectation of the complete-data log-likelihood function in the EM algorithm.

Schall and Dunne (1991) discussed diagnostics for regression-ARMA time series, considering local influence for two types of outliers: the observation and innovation outlier. Kim and Huggins (1998) discussed the local influence approach to linear regression model with AR(1) errors. Nevertheless, to the best of our knowledge, there are no previous studies of diagnostic analysis for censored linear regression models with autoregressive errors.

1.7   Work objectives

This work aims to present inferences and diagnostic analysis in censored regression models with autoregressive errors of order p. The specific objectives are:

1. To derive a computationally efficient estimation method via the stochastic version of the expectation-maximization (SAEM) algorithm in AR(p)-CR models;

2. To derive the standard errors of the SAEM estimates in AR(p)-CR models based on a stochastic approximation of the observed Fisher information matrix;

3. To develop influence diagnostic techniques in AR(p)-CR models, based on local influence approach (Cook, 1986; Zhu and Lee, 2001);

4. To implement the proposed methods in the R package *ARCensReg* (Schumacher et al., 2016; R Development Core Team, 2016).

1.8   Organization of the dissertation

The results contained in this dissertation are organized into four chapters. In Chapter 2, we describe the AR(p)-CR model and the ML estimation procedure based in the SAEM algorithm and we discuss how to obtain the standard errors and prediction. To conclude this chapter we examine the performance of the proposed methods through simulation studies as well as the analysis of two real datasets.

In Chapter 3, we develop influence diagnostic techniques, based on local influence approach. The methodology is illustrated using a real dataset and we present a simulation study evaluating the efficiency of our method in detecting influential observations under various scenarios.

Finally, Chapter 4 presents some concluding remarks, the technical production that resulted from this dissertation and some possible directions for future research.

Chapter 2

The censored autoregressive regression model of order p

Time series usually present autocorrelation, which refers to the correlation of a time series with its own past and future values. In this situation, regression models with independent errors assumption are not appropriate. Instead, we should consider modeling their autocorrelation. An additional complication arises when time series measurements are observed with data irregularities, such as observations subjected to upper or lower detection limits and missing observations, which commonly occurs in environmental monitoring, for example. In this situation, the proportion of censored data may not be small, and so the use of crude/ad hoc methods, such as discarding the censored observations or substituting them by a threshold value or by some arbitrary point, like a midpoint between zero and cutoff for detection (LOD/2), might lead to biased estimates of the parameters.

Our motivating datasets presented in Subsection 1.1.1 and 1.1.2 have 40.5% and 15.5% of censored observations, respectively. As an alternative to crude imputation methods, Zeger and Brookmeyer (1986) suggested a full likelihood estimation and approximate method for an autoregressive time series model. However, the authors pointed out that the method may not be feasible when the censoring rate is very high. Some other authors, such as Robinson (1980) and Park et al. (2007), proposed imputation methods for handling censored data in the setting of AR models.

In this chapter we present the AR(p)-CR model and we develop a likelihood-based approach to obtain the ML estimation based on the SAEM algorithm (Delyon et al., 1999), that consists of replacing the E-step from the expectation-maximization (EM) algorithm (Dempster et al., 1977) by a stochastic approximation obtained using simulated data, while the M-step remains unchanged. The SAEM algorithm is a powerful alternative when the E-step is intractable or computationally too expensive, and it has been proved to be more computationally efficient than the classic Monte Carlo EM (MCEM) algorithm (Meza et al., 2012). We also discuss how

to obtain standard errors and prediction. The proposed SAEM algorithm is implemented in the R package *ARCensReg* (Schumacher et al., 2016).

2.1 Proposed model

The autoregressive model with exogenous variables is defined by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.1.1}$$

where $\mathbf{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ are as defined in Section 1.4. Moreover, following Vaida and Liu (2009), we assume that the response Y_t is not fully observed for all t. Thus, assuming left censoring, let the observed data at the tth time be (V_t, C_t), where V_t represents the uncensored reading or censoring level and C_t is the censoring indicator such that

$$Y_t \leq V_t \ \text{if} \ C_t = 1 \ \text{and} \ Y_t = V_t \ \text{if} \ C_t = 0, \tag{2.1.2}$$

so that (2.1.2), along with the model given in (2.1.1), define AR(p)-CR model.

2.2 The log-likelihood function

Following Vaida and Liu (2009), classical inference on the parameter vector $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \sigma^2, \boldsymbol{\phi}^\top\right)^\top$ is based on the marginal distribution of $\mathbf{y}$. For complete-data, we have marginally that $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$. On the other hand, for responses with censoring pattern as in (2.1.2), we have $\mathbf{y} \mid \mathbf{V}, \mathbf{C} \sim TN_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}; \mathbb{A})$, where $TN_n(\cdot; \mathbb{A})$ denotes the truncated normal distribution on the interval $\mathbb{A}$, where $\mathbb{A} = A_1 \times \ldots \times A_n$, with A_t being the interval $(-\infty, \infty)$ if $C_t = 0$, and the interval $(-\infty, V_t]$ if $C_t = 1$. To compute the likelihood function associated with model (2.1.1)-(2.1.2), the first step is to treat separately the observed and censored components of $\mathbf{y}$.

Let $\mathbf{y}^o$ be the n^o-vector of observed outcomes and $\mathbf{y}^c$ be the n^c-vector of censored observations, with $n = n^o + n^c$, such that $C_t = 0$ for all elements in $\mathbf{y}^o$ and $C_t = 1$ for all elements in $\mathbf{y}^c$. After reordering, $\mathbf{y}$, $\mathbf{V}$, $\mathbf{X}$ and $\boldsymbol{\Sigma}$ can be partitioned as follows:

$$\mathbf{y} = vec(\mathbf{y}^o, \mathbf{y}^c), \ \mathbf{V} = vec(\mathbf{V}^o, \mathbf{V}^c), \ \mathbf{X} = vec(\mathbf{X}^o, \mathbf{X}^c) \ \text{and} \ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}^{oo} & \boldsymbol{\Sigma}^{oc} \\ \boldsymbol{\Sigma}^{co} & \boldsymbol{\Sigma}^{cc} \end{pmatrix},$$

where $vec(.)$ denotes the function which stacks vectors or matrices of the same number of columns. Then we have

$$\mathbf{y}^o \sim N_{n^o}(\mathbf{X}^o\boldsymbol{\beta}, \boldsymbol{\Sigma}^{oo}) \ \text{and} \ \mathbf{y}^c|\mathbf{y}^o \sim N_{n^c}(\boldsymbol{\mu}, \mathbf{S}),$$

where $\boldsymbol{\mu} = \mathbf{X}^c \boldsymbol{\beta} + \Sigma^{co} (\Sigma^{oo})^{-1} (\mathbf{y}^o - \mathbf{X}^o \boldsymbol{\beta})$ and $\mathbf{S} = \Sigma^{cc} - \Sigma^{co} (\Sigma^{oo})^{-1} \Sigma^{oc}$. Now, let $\Phi_n(\mathbf{u}; \mathbf{a}, \mathbf{A})$ and $\phi_n(\mathbf{u}; \mathbf{a}, \mathbf{A})$ be the cdf (left tail) and pdf, respectively, of $N_n(\mathbf{a}, \mathbf{A})$ computed at vector $\mathbf{u}$. From Vaida and Liu (2009) and Matos et al. (2013), the likelihood function for the observed data is thus given by (using conditional probability arguments)

$$
\begin{aligned}
L(\boldsymbol{\theta}|\mathbf{y}) &= P(\mathbf{y}^c \le \mathbf{V}^c | \mathbf{y}^o = \mathbf{V}^o, \boldsymbol{\theta}) P(\mathbf{y}^o = \mathbf{V}^o | \boldsymbol{\theta}) \\
&= P(\mathbf{y}^c \le \mathbf{V}^c | \mathbf{y}^o, \boldsymbol{\theta}) f(\mathbf{y}^o | \boldsymbol{\theta}) \\
&= \phi_{n^o}(\mathbf{y}^o; \mathbf{X}^o \boldsymbol{\beta}, \Sigma^{oo}) \Phi_{n^c}(\mathbf{V}^c; \boldsymbol{\mu}, \mathbf{S}),
\end{aligned} \tag{2.2.1}
$$

which can be evaluated without much computational burden through the routine *mvtnorm()* available in R (see Genz et al., 2008; R Development Core Team, 2016). The estimates obtained by maximizing the log-likelihood function $\ell(\boldsymbol{\theta}|\mathbf{y}) = \log(L(\boldsymbol{\theta}|\mathbf{y}))$ are the maximum likelihood (ML) estimates.

The log-likelihood function $\ell(\boldsymbol{\theta}|\mathbf{y})$ is used to compute different model selection criteria, such as

$$
\text{AIC} = 2m - 2\ell_{max} \text{ and BIC} = m\log n - 2\ell_{max},
$$

where $m = p + l + 1$ is the number of model parameters and ℓ_{max} is the maximized log-likelihood value.

2.3 ML estimation via the SAEM algorithm

As the observed log-likelihood function involves complex expressions, such as multidimensional integrals, it is difficult to work directly with $\ell(\boldsymbol{\theta}|\mathbf{y})$, either for the ML estimation or to carry out the influence analysis. Besides, time series data commonly present missing observations and, if the response Y_t were fully observed for all t, then maximum likelihood estimation would be easy. Thus, we will discuss how to obtain the ML estimates using the SAEM algorithm presented in Section 1.5. It is worth emphasizing that this approach enables approximation of the standard errors and direct imputation of the censored and missing observations.

Let $\mathbf{y} = (Y_1, \ldots, Y_n)^\top$, $\mathbf{V} = (V_1, \ldots, V_n)^\top$ and $\mathbf{C} = (C_1, \ldots, C_n)^\top$, and consider that we observe (V_t, C_t) at time t. Let $\mathbf{y}_c = (\mathbf{C}^\top, \mathbf{V}^\top, \mathbf{y}^\top)^\top$ be the complete dataset obtained by augmenting the observed dataset $(\mathbf{C}^\top, \mathbf{V}^\top, \mathbf{y}^{o\top})^\top$ with the censored data $\mathbf{y}^c$. Hence, the EM-type algorithm (Dempster et al., 1977) is applied to the complete-data log-likelihood $\ell(\boldsymbol{\theta}|\mathbf{y}_c)$ which, dropping constant terms, is given by

$$
\ell(\boldsymbol{\theta}|\mathbf{y}_c) = -\frac{1}{2} \left[n\log \sigma^2 + \frac{1}{\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{M}_n^{-1}(\boldsymbol{\phi})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \log |\mathbf{M}_p(\boldsymbol{\phi})| \right], \tag{2.3.1}
$$

where $\mathbf{M}_n(\boldsymbol{\phi})$ is as defined in (1.4.3) and $\mathbf{M}_p(\boldsymbol{\phi})$ is a $p \times p$ matrix containing the first p lines

and p columns of $\mathbf{M}_n(\phi)$.

Given the current estimate $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(k)}$, the E-step calculates the conditional expectation of the complete-data log-likelihood function given by

$$\hat{Q}_k(\boldsymbol{\theta}) = E[\ell(\boldsymbol{\theta}|\mathbf{y}_c)|\mathbf{V},\mathbf{C},\widehat{\boldsymbol{\theta}}^{(k)}] = -\frac{1}{2}\left[n\log\sigma^2 + \log|\mathbf{M}_p(\phi)| + \frac{1}{\sigma^2}\gamma^{(k)}\right], \qquad (2.3.2)$$

where

$$\begin{aligned}
\gamma^{(k)} = \gamma(\mathbf{y}^{(k)},\phi,\boldsymbol{\beta}) &= \mathrm{tr}(\widehat{\mathbf{y}^{2(k)}}\mathbf{M}_n^{-1}(\phi)) - 2\widehat{\mathbf{y}^{\top(k)}}\mathbf{M}_n^{-1}(\phi)\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^{\top}\mathbf{X}^{\top}\mathbf{M}_n^{-1}(\phi)\mathbf{X}\boldsymbol{\beta} \\
&= (\widehat{\mathbf{y}^{(k)}} - \mathbf{X}\boldsymbol{\beta})^{\top}\mathbf{M}_n^{-1}(\phi)(\widehat{\mathbf{y}^{(k)}} - \mathbf{X}\boldsymbol{\beta}) + \mathrm{tr}(Var\{\mathbf{y}|\mathbf{V},\mathbf{C},\widehat{\boldsymbol{\theta}}^{(k)}\}\mathbf{M}_n^{-1}(\phi)).
\end{aligned}$$

It is clear that the E-step reduces only to the computation of

$$\widehat{\mathbf{y}^{(k)}} = E\{\mathbf{y}|\mathbf{V},\mathbf{C},\widehat{\boldsymbol{\theta}}^{(k)}\} \ \text{ and } \ \widehat{\mathbf{y}^{2(k)}} = E\{\mathbf{y}\mathbf{y}^{\top}|\mathbf{V},\mathbf{C},\widehat{\boldsymbol{\theta}}^{(k)}\}, \qquad (2.3.3)$$

that is, the first and second moments of a truncated multivariate normal distribution. Although these can be determined in closed form as a function of multivariate normal probabilities (for more details on the computation of these moments, one may refer to Vaida and Liu, 2009; Arismendi, 2013), this calculation is computationally expensive since it requires high-dimensional numerical integration, resulting in an impracticable time of convergence when the number of censored observations is not small.

For this reason, we will consider the SAEM algorithm as described in Section 1.5 which, for the model defined in (2.1.1), at iteration k, is performed as follows:

- **E-Step:**

  ◇ *Simulation:* sample $\mathbf{y}_{k,l}^c$ ($l = 1,\ldots,m$) from the truncated multivariate normal distribution $TN_{n^c}(\boldsymbol{\mu},\mathbf{S};\mathbb{A})$, where $\boldsymbol{\mu} = \mathbf{X}^c\boldsymbol{\beta} + \Sigma^{co}(\Sigma^{oo})^{-1}(\mathbf{y}^o - \mathbf{X}^o\boldsymbol{\beta})$, $\mathbf{S} = \Sigma^{cc} - \Sigma^{co}(\Sigma^{oo})^{-1}\Sigma^{oc}$ and $\mathbb{A} = \{\mathbf{y}^c = (y_1^c,\ldots,y_{n^c}^c)^{\top} \mid y_1^c \leq V_1,\ldots,y_{n^c}^c \leq V_{n^c}\}$;

  ◇ *Stochastic Approximation:* update $\hat{Q}_k(\boldsymbol{\theta})$ according to

  $$\hat{Q}_k(\boldsymbol{\theta}) = \hat{Q}_{k-1}(\boldsymbol{\theta}) + \delta_k\left(\frac{1}{m}\sum_{l=1}^m \ell(\boldsymbol{\theta} \mid \mathbf{y}_{k,l}) - \hat{Q}_{k-1}(\boldsymbol{\theta})\right), \qquad (2.3.4)$$

  where $\mathbf{y}_{k,l} = vec(\mathbf{y}^o,\mathbf{y}_{k,l}^c)$, $l = 1,\ldots,m$ and δ_k is a smoothing parameter, i.e., a decreasing sequence of positive numbers such that $\sum_{k=1}^{\infty}\delta_k = \infty$ and $\sum_{k=1}^{\infty}\delta_k^2 \leq \infty$.

- **M-Step:**

  ◇ *Maximization*: update the estimate $\widehat{\boldsymbol{\theta}}^{(k)}$ according to

  $$\widehat{\boldsymbol{\theta}}^{k+1} = \underset{\boldsymbol{\theta}}{\arg\max}\ \hat{Q}_k(\boldsymbol{\theta}).$$

Since the complete-data likelihood function given in 2.3.1 belongs to the exponential family, the E-step reduces to the approximation of the expressions given in (2.3.3), and the *Stochastic Approximation* step, defined in (2.3.4), reduces to

$$\widehat{\mathbf{y}^{2(k)}} = \widehat{\mathbf{y}^{2(k-1)}} + \delta_k\left(\frac{1}{m}\sum_{l=1}^{m}\mathbf{y}_{(k,l)}^{c}\mathbf{y}_{(k,l)}^{c\top} - \widehat{\mathbf{y}^{2(k-1)}}\right), \tag{2.3.5}$$

$$\widehat{\mathbf{y}^{(k)}} = \widehat{\mathbf{y}^{(k-1)}} + \delta_k\left(\frac{1}{m}\sum_{l=1}^{m}\mathbf{y}_{(k,l)}^{c} - \widehat{\mathbf{y}^{(k-1)}}\right),\ k = 1,2,3,\dots. \tag{2.3.6}$$

Moreover, for the model given in (2.1.1) and (2.1.2), the conditional maximization step becomes:

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \left(\mathbf{X}^{\top}\mathbf{M}_n^{-1}(\widehat{\boldsymbol{\phi}}^{(k)})\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{M}_n^{-1}(\widehat{\boldsymbol{\phi}}^{(k)})\widehat{\mathbf{y}^{(k)}}, \tag{2.3.7}$$

$$\widehat{\sigma^2}^{(k+1)} = \frac{1}{n}\left(\mathrm{tr}\left(\widehat{\mathbf{y}^{2(k)}}\mathbf{M}_n^{-1}(\widehat{\boldsymbol{\phi}}^{(k)})\right) - 2\widehat{\boldsymbol{\beta}}^{(k)\top}\mathbf{X}^{\top}\mathbf{M}_n^{-1}(\widehat{\boldsymbol{\phi}}^{(k)})\widehat{\mathbf{y}^{(k)}}\right. \tag{2.3.8}$$

$$+ \left.\widehat{\boldsymbol{\beta}}^{(k)\top}\mathbf{X}^{\top}\mathbf{M}_n^{-1}(\widehat{\boldsymbol{\phi}}^{(k)})\mathbf{X}\widehat{\boldsymbol{\beta}}^{(k)}\right),$$

$$\widehat{\boldsymbol{\pi}}^{(k+1)} = \underset{\boldsymbol{\pi}\in(-1,1)^p}{\arg\max}\left\{-\frac{n}{2}\log\left[\left(-1,\mathscr{B}(\boldsymbol{\pi})^{\top}\right)D\left(\widehat{\mathbf{y}^{(k)}},\widehat{\boldsymbol{\beta}}^{(k)}\right)\begin{pmatrix}-1\\\mathscr{B}(\boldsymbol{\pi})\end{pmatrix}\right]\right. \tag{2.3.9}$$

$$\left. - \frac{1}{2}\log\left[\prod_{j=1}^{p}\left(1-\pi_j^2\right)^{-j}\right]\right\},$$

$$\widehat{\boldsymbol{\phi}}^{(k+1)} = \mathscr{B}(\widehat{\boldsymbol{\pi}}^{(k+1)}). \tag{2.3.10}$$

The initial values were calculated by considering the censored values as observed ones and proceeding as discussed in Section 1.4.

In order to make the proposed algorithm easier to understand, we summarized all the steps needed to implement the SAEM algorithm in a flow diagram, presented in Figure 2.1.

To illustrate the gain in considering the stochastic approximations of the expectations instead of the theorical calculations, we generated one sample of size 300 with 40% of left censoring from an AR(1)-CR model, as defined in (2.1.1) and (2.1.2), and estimated the model parameters using the algorithms EM and SAEM ($m = 10$, $W = 400$ and $c = 0.18$), with
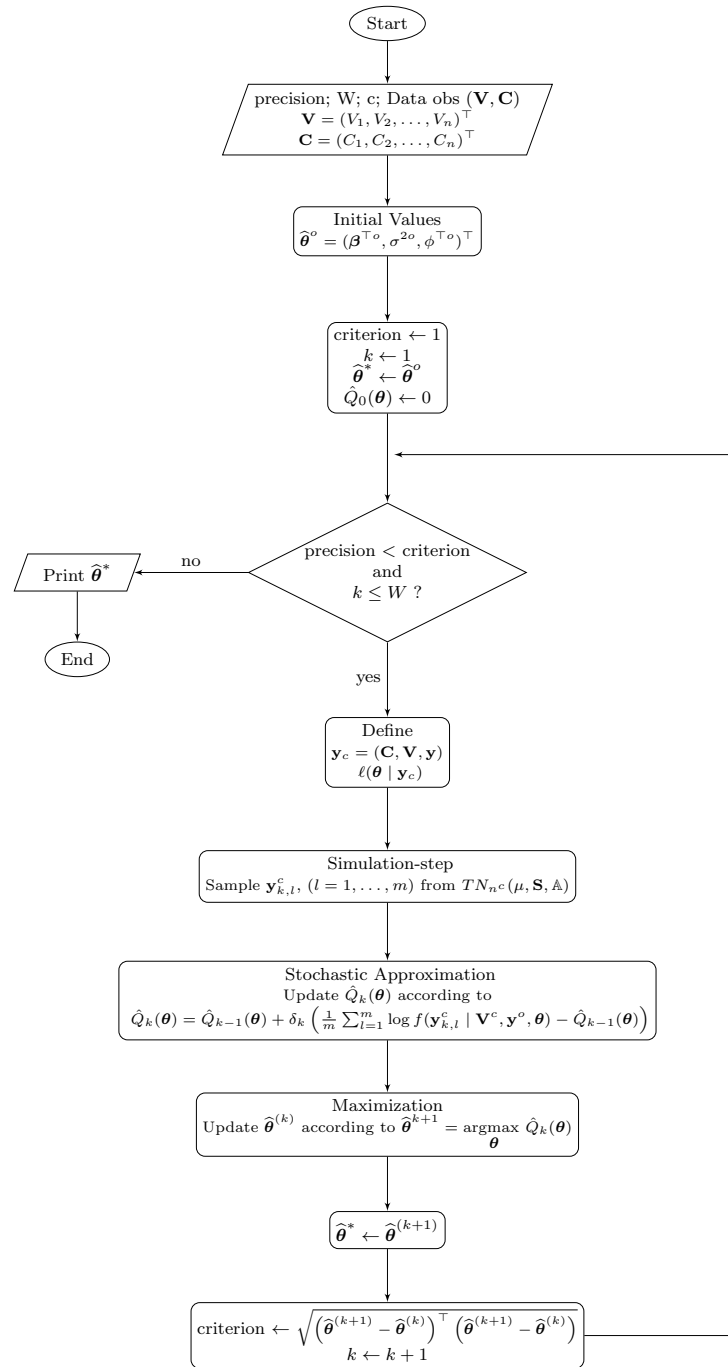
Figure 2.1: Flow diagram of the SAEM algorithm for the AR(p)-CR models.

the same initial values. The SAEM algorithm took approximately 9 minutes to converge, while the EM algorithm took approximately 28 hours.

2.4 Standard error and prediction

2.4.1 The observed Fisher information matrix

Let $\mathbf{y}_{k,l} = vec(\mathbf{y}^o, \mathbf{y}^c_{k,l})$, $(l = 1, \dots, m)$,

$$\partial_\theta \ell(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\partial \ell(\boldsymbol{\theta} \mid \mathbf{y})}{\partial \boldsymbol{\theta}} \quad \text{and} \quad \partial_\theta^2 \ell(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\partial^2 \ell(\boldsymbol{\theta} \mid \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}.$$

Following Meza et al. (2012), the Fisher information matrix can be estimated using the fact that the gradient and the Hessian of the log-likelihood can be obtained almost directly from the simulated missing data $\mathbf{y}^c$. Thus, $\partial_\theta^2 \ell(\boldsymbol{\theta} \mid \mathbf{y})$ can be approximated following the stochastic approximation procedure:

$$
\begin{aligned}
\Delta_k &= \Delta_{k-1} + \delta_k \left(\frac{1}{m} \sum_{l=1}^{m} \partial_\theta \ell(\widehat{\boldsymbol{\theta}}^{(k)} \mid \mathbf{y}_{k,l}) - \Delta_{k-1} \right), \\
G_k &= G_{k-1} + \delta_k \left(\frac{1}{m} \sum_{l=1}^{m} \left(-\partial_\theta^2 \ell(\widehat{\boldsymbol{\theta}}^{(k)} \mid \mathbf{y}_{k,l}) - \partial_\theta \ell(\widehat{\boldsymbol{\theta}}^{(k)} \mid \mathbf{y}_{k,l}) \partial_\theta \ell(\widehat{\boldsymbol{\theta}}^{(k)} \mid \mathbf{y}_{k,l})^\top \right) - G_{k-1} \right), \\
H_k &= G_k - \Delta_k \Delta_k^\top, \quad k = 1, 2, 3, \dots.
\end{aligned}
$$

Provided the SAEM algorithm converges to a limiting value $\boldsymbol{\theta}^*$ and that $\ell(\boldsymbol{\theta} \mid \mathbf{y})$ is regular enough, H_k converges to the observed Fisher information matrix $\mathbf{I}_o(\boldsymbol{\theta}^*) = -\partial_\theta^2 \log L(\boldsymbol{\theta}^* \mid \mathbf{y})$, where $L(\boldsymbol{\theta} \mid \mathbf{y})$ is as defined in (2.2.1). When $\log L(\boldsymbol{\theta} \mid \mathbf{y})$ is a sufficiently smooth incomplete-data log-likelihood function, the maximum likelihood estimator is asymptotically normal and $\mathbf{I}_o(\boldsymbol{\theta}^*)^{-1}$ converges to the asymptotic covariance of the estimators (Delyon et al., 1999).

In this sense, let $\ell(\boldsymbol{\theta}|\mathbf{y}_c)$ be as defined in (1.4.6) or, equivalently, as defined in (1.4.7). Let $D = D(\mathbf{y}_c, \boldsymbol{\beta})$ be partitioned as

$$D = \begin{bmatrix} D_{11} & D_{\phi 1}^\top \\ D_{\phi 1} & D_{\phi\phi} \end{bmatrix}, \tag{2.4.1}$$

such that D_{11} is 1×1, $D_{\phi 1}$ is $p \times 1$ and $D_{\phi\phi}$ is $p \times p$. Then, the sum of squares can be written as

$$\boldsymbol{\lambda}^\top D \boldsymbol{\lambda} = \begin{bmatrix} -1, \boldsymbol{\phi}^\top \end{bmatrix} \begin{bmatrix} D_{11} & D_{1\phi} \\ D_{1\phi}^\top & D_{\phi\phi} \end{bmatrix} \begin{bmatrix} -1 \\ \boldsymbol{\phi} \end{bmatrix} = D_{11} - 2\boldsymbol{\phi}^\top D_{\phi 1} + 2\boldsymbol{\phi}^\top D_{\phi\phi} \boldsymbol{\phi}.$$

Therefore, we have

$$\frac{\partial \boldsymbol{\lambda}^\top D \boldsymbol{\lambda}}{\partial \boldsymbol{\phi}} = -2D_{\phi 1} + 2D_{\phi\phi} \boldsymbol{\phi}.$$

Thus, after some algebraic manipulation and letting $\mathbf{M}_n = \mathbf{M}_n(\boldsymbol{\phi})$ be as given in

(1.4.3), the elements of the gradient vector are

$$\frac{\partial \ell(\boldsymbol{\theta}|\mathbf{y}_c)}{\partial \beta} = \frac{1}{\sigma^2} \left(\mathbf{X}^\top \mathbf{M}_n^{-1} \mathbf{y} - \mathbf{X}^\top \mathbf{M}_n^{-1} \mathbf{X} \beta \right),$$

$$\frac{\partial \ell(\boldsymbol{\theta}|\mathbf{y}_c)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (\mathbf{y} - \mathbf{X}\beta)^\top \mathbf{M}_n^{-1} (\mathbf{y} - \mathbf{X}\beta),$$

$$\frac{\partial \ell(\boldsymbol{\theta}|\mathbf{y}_c)}{\partial \phi} = -\frac{1}{2} \mathrm{tr} \left\{ \mathbf{M}_p^{-1} \frac{\partial \mathbf{M}_p}{\partial \phi} \right\} - \frac{1}{\sigma^2} \left(-D_{\phi 1} + D_{\phi \phi} \phi \right),$$

and the elements of the Hessian matrix are

$$\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{y}_c)}{\partial \beta \partial \beta^\top} = -\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{M}_n^{-1} \mathbf{X},$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{y}_c)}{\partial \beta \partial \sigma^2} = -\frac{1}{\sigma^4} \left(\mathbf{X}^\top \mathbf{M}_n^{-1} \mathbf{y} - \mathbf{X}^\top \mathbf{M}_n^{-1} \mathbf{X} \beta \right),$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{y}_c)}{\partial \beta \partial \phi^\top} = -\frac{1}{\sigma^2} \frac{\partial (D_{\phi \phi} \phi - D_{\phi 1})}{\partial \beta},$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{y}_c)}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} (\mathbf{y} - \mathbf{X}\beta)^\top \mathbf{M}_n^{-1} (\mathbf{y} - \mathbf{X}\beta),$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{y}_c)}{\partial \phi \partial \sigma^2} = \frac{1}{\sigma^4} \left(D_{\phi \phi} \phi - D_{\phi 1} \right),$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{y}_c)}{\partial \phi \partial \phi^\top} = -\frac{1}{\sigma^2} D_{\phi \phi} - \frac{1}{2} \mathrm{tr} \left\{ \frac{\partial}{\partial \phi} \left(\mathbf{M}_p^{-1} \frac{\partial \mathbf{M}_p}{\partial \phi^\top} \right) \right\}.$$

2.4.2 Prediction

The problem related to the prediction of future values has a great impact in many practical applications. Rao et al. (1987) pointed out that the predictive accuracy of future observations can be taken as an alternative measure of "goodness of fit". In order to propose a strategy for generating predicted values from our AR(p)-CR model, we used the plug-in approach proposed by Wang (2013). Thus, let $\mathbf{y}_{obs}$ be the observed response vector of dimension $n_{obs} \times 1$ and $\mathbf{y}_{pred}$ the $n_{pred} \times 1$ response vector n_{pred}-step-ahead. Let $\widetilde{\mathbf{X}} = \left(\mathbf{X}_{obs}, \mathbf{X}_{pred} \right)$ be the $(n_{obs} + n_{pred}) \times p$ design matrix corresponding to

$$\widetilde{\mathbf{y}} = \left(\begin{array}{c} \mathbf{y}_{obs} \\ \mathbf{y}_{pred} \end{array} \right).$$

Replacing the censored values existing in $\mathbf{y}_{obs}$ by $\widehat{\mathbf{y}} = E\{\mathbf{y}|\mathbf{V}, \mathbf{C}, \widehat{\boldsymbol{\theta}}\}$ obtained from the SAEM algorithm, a complete dataset, $\mathbf{y}_{obs}^*$, is obtained. Then, we have that

$$\widetilde{\mathbf{y}}^* = \left(\mathbf{y}_{obs}^{*\top}, \mathbf{y}_{pred}^\top \right)^\top \sim N_{n_{obs} + n_{pred}} \left(\widetilde{\mathbf{X}} \beta, \, \boldsymbol{\Sigma} \right), \tag{2.4.2}$$

where

$$\Sigma = \left(\begin{array}{cc} \Sigma^{obs^*,obs^*} & \Sigma^{obs^*,pred} \\ \Sigma^{pred,obs^*} & \Sigma^{pred,pred} \end{array} \right).$$

Following Wang (2013), the best predictor of $\mathbf{y}_{pred}$, with respect to the minimum mean squared error (MSE) criterion, is the conditional expectation of $\mathbf{y}_{pred}$ given $\mathbf{y}_{obs}^*$, given by

$$\widehat{\mathbf{y}}_{pred}(\boldsymbol{\theta}) = \mathbf{X}_{pred}\beta + \Sigma^{pred,obs^*} \left(\Sigma^{obs^*,obs^*} \right)^{-1} \left(\mathbf{y}_{obs} - \mathbf{X}_{obs}\beta \right). \quad (2.4.3)$$

Therefore, the predictor of $\mathbf{y}_{pred}$ can be calculated by substituting $\widehat{\boldsymbol{\theta}}$ in (2.4.3), obtaining

$$\widehat{\mathbf{y}}_{pred} = \widehat{\mathbf{y}}_{pred}(\widehat{\boldsymbol{\theta}}).$$

2.5 Simulation studies

Two simulation studies were conducted to examine the performance of the proposed method, by analyzing the asymptotic properties of the SAEM estimates and prediction accuracy based in the plug-in method proposed in Subsection 2.4.2. In all the simulation studies we consider a left-censored AR(2)-CR model with one explanatory variable. Thus, the data generating process is as follows:

Step 1: Generate the correlated errors ε_t from the model defined in (1.4.2), with $p = 2$ and x_t from a uniform distribution $(U(0,1))$. Then set $Z_t = \beta_0 + \beta_1 x_t + \varepsilon_t, t = 1, \ldots, n$;

Step 2: Construct the censored time series $Y_t = V \, \mathbb{I}_{(Z_t < V)} + Z_t \, \mathbb{I}_{(Z_t \geq V)}$, where V is the appropriate sample quantile.

The initial estimates were chosen by fitting a linear regression considering the censored values (LOD) as real observed values and by calculating the sample partial autocorrelations of the residuals. For the SAEM algorithm, we fixed the maximum number of iterations at $W = 600$ and the cutoff point at $c = 0.12$ (see Kuhn and Lavielle, 2005).

2.5.1 First study

The main goal of this simulation study is to provide empirical evidence of the consistency of the ML estimates, for different censoring proportions. The simulated data follow an AR(2)-CR model, as defined in (2.1.1)–(2.1.2), with parameters set at $\beta_0 = 2$, $\beta_1 = 1$, $\sigma^2 = 2$, $\phi_1 = 0.48$ and $\phi_2 = -0.2$. In order to investigate the asymptotic properties of the estimates, different censoring proportions (5%, 20% and 40%) and different samples sizes ($n = 50, 100, 200, 300$ and 500) were considered.

For each simulation setting, we considered 100 simulated Monte Carlo datasets.

The ML estimates and their associate standard errors were recorded. We analyzed the mean square error (MSE) and the mean absolute error (MAE) of the coefficient estimates obtained. These measures for the parameter θ_i are defined as

$$\text{MSE}_i = \frac{1}{100} \sum_{j=1}^{100} \left(\widehat{\theta}_i^{(j)} - \theta_i \right)^2 \text{ and MAE}_i = \frac{1}{100} \sum_{j=1}^{100} \left| \widehat{\theta}_i^{(j)} - \theta_i \right|, \qquad (2.5.1)$$

where $\widehat{\theta}_i^{(j)}$ is the ML estimate of the parameter θ_i for the jth sample, $j = 1, \ldots, 100$.

Figures 2.2 and 2.3 show that the MSE and the MAE tend to zero as the sample size increases. The respective tables are presented in the Appendix A. As a general rule, the results indicate that the SAEM estimates of the proposed model do provide good asymptotic properties. Note also that for simulations with higher censoring rates ($p = 40\%$), the convergence still behaved well. Table 2.1 presents the summary statistics for parameter estimation.

Here we also examine the consistency of the approximation method, suggested in Subsection 2.4.1, to get the standard errors (SE) of the SAEM estimates. Considering all the ML estimates obtained (across 100 samples), we computed:

- The Monte Carlo standard deviation of $\widehat{\theta}_i$, defined by

$$\text{MC-SD} = \sqrt{\frac{1}{99} \left[\sum_{j=1}^{100} \left(\widehat{\theta}_i^{(j)} \right)^2 - 100 \left(\overline{\overline{\theta}}_i \right)^2 \right]},$$

where $\overline{\overline{\theta}}_i = \frac{1}{100} \Sigma_{j=1}^{100} \widehat{\theta}_i^{(j)}$; and

- The average values of the approximate standard errors of the SAEM estimates obtained through the method described in Subsection 2.4.1 using the empirical information matrix, denoted by IBM-SE.

Table 2.1 reveals that the estimation method of the standard errors provides relatively close results to the empirical ones, indicating that the proposed approximate method to get the standard errors is reliable. Moreover, the closeness improves as the sample size increases.

2.5.2 Second study

The aim of this simulation study is to compare the estimation and prediction accuracy of the proposed method (denoted by *Cens*) with two naive *(ad-hoc)* methods commonly used in the literature: (a) to treat the censored values as observed (denoted by *Unc*), and (b) to impute the censored values by calculating the censored values divided by two (denoted by *LOD*). For the naive methods estimation, we used the routine *arima()* available in the R software (R Development Core Team, 2016).
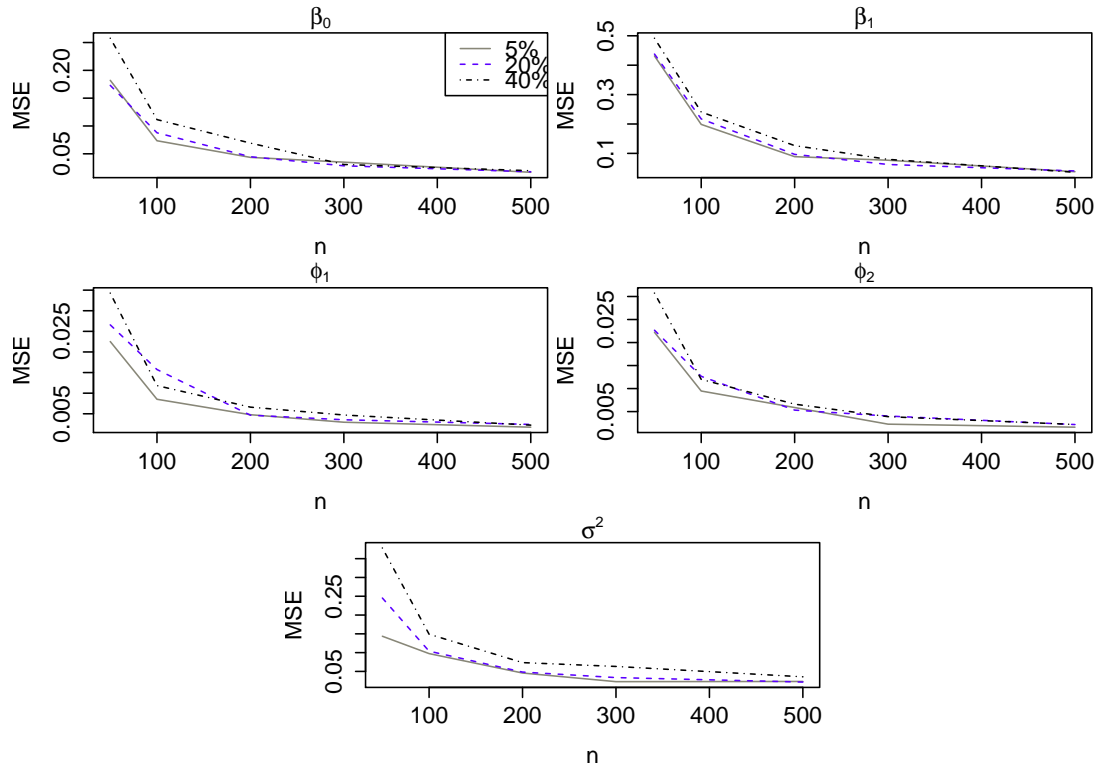
Figure 2.2: Mean square error of the parameter estimates under 5%, 20% and 40% censoring levels and different samples sizes. The solid line (gray) represents 5% censoring, the dashed line (blue) represents 20% and the dot-dashed line (black) represents 40%.

The simulated data follow an AR(2)-CR model, as defined in (2.1.1)–(2.1.2), with parameters set at $\beta_0 = 10$, $\beta_1 = 5$, $\sigma^2 = 2$, $\phi_1 = 0.48$ and $\phi_2 = -0.2$. We considered different censoring proportions (5%, 20% and 40%), with sample size set at $n = 500$. The 3 latest values were preserved with the purpose of comparing the prediction accuracy, and the remaining 497 values were used for estimation and 3-step-ahead prediction.

For each simulation setting we considered 100 simulated Monte Carlo datasets, resulting in 300 predicted values for each method. For each method and censoring rate we analyzed the mean square prediction error (MSPE) and the mean absolute prediction error (MAPE), defined as

$$\text{MSPE} = \frac{1}{300} \sum_{l=1}^{100} \sum_{j=498}^{500} \left(y_j^{(l)} - \widehat{y}_j^{(l)} \right)^2 \text{ and MAPE} = \frac{1}{300} \sum_{l=1}^{100} \sum_{j=498}^{500} \left| y_j^{(l)} - \widehat{y}_j^{(l)} \right|, \qquad (2.5.2)$$

where $y_j^{(l)}$ is jth value of the lth sample and $\widehat{y}_j^{(l)}$ is its predicted value.

Table 2.2 presents the prediction accuracy measures obtained through all the simulated Monte Carlos samples. It can be seen that the SAEM approach provided more accurate predictions than the naive methods for all censoring rates considered and this difference became greater as the censoring rate increased.
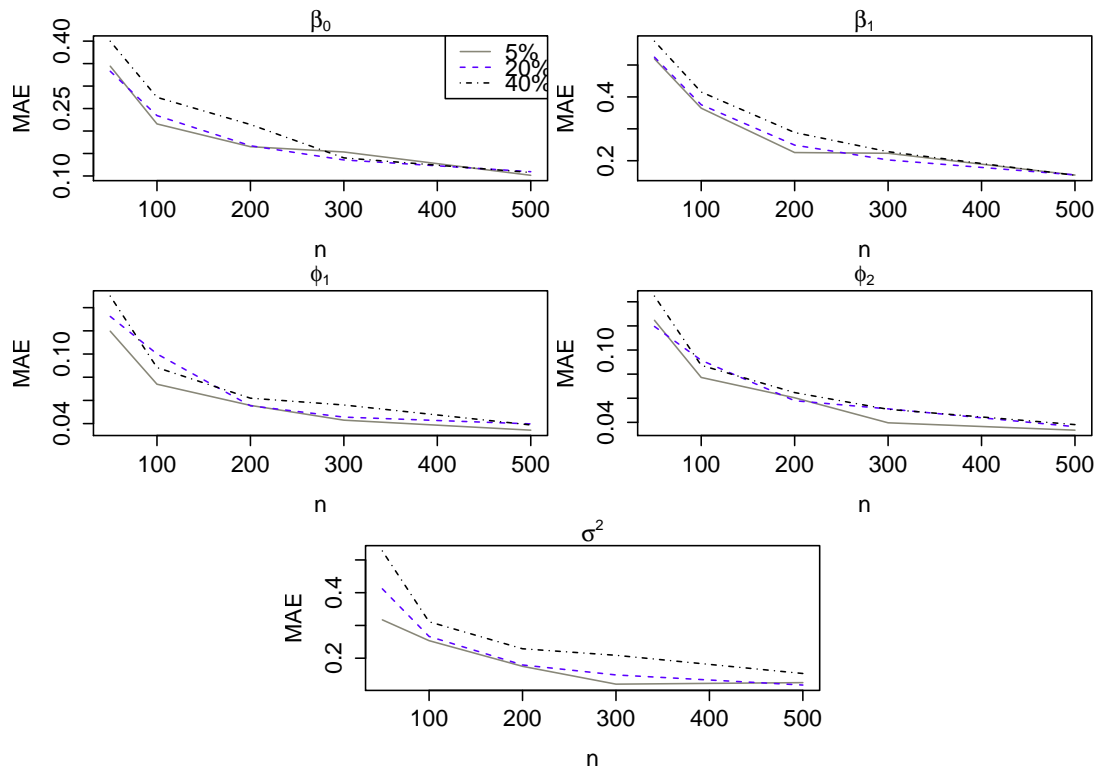
Figure 2.3: Mean absolute error of the parameter estimates under different censoring levels and different samples sizes. The solid line (gray) represents 5% censoring, the dashed line (blue) represents 20% and the dot-dashed line (black) represents 40%.

Figure 2.4 presents the boxplots of the parameter estimates for the three methods and it shows that, on average, the SAEM algorithm produced the closest estimates to the true values for all parameters and all censoring rates.

Table 2.1: Results based on 100 simulated Monte Carlo samples with different sample sizes (n) and different censoring proportions (CP). MC mean and MC SD are the mean and standard deviations of the estimates, respectively. IBM SE is the average value of the approximate standard error obtained through the information-based method as described in Subsection 2.4.1.

n	CP		$\beta_0 = 2$	$\beta_1 = 1$	$\sigma^2 = 2$	$\phi_1 = 0.48$	$\phi_2 = -0.2$
		MC Mean	1.981	0.997	1.880	0.454	-0.233
	5%	IBM SE	0.409	0.627	0.394	0.141	0.142
		MC SD	0.428	0.660	0.361	0.149	0.146
		MC Mean	2.038	0.984	1.858	0.458	-0.223
50	20%	IBM SE	0.425	0.646	0.430	0.149	0.148
		MC SD	0.416	0.665	0.477	0.162	0.150
		MC Mean	2.027	0.978	1.814	0.449	-0.253
	40%	IBM SE	0.453	0.700	0.495	0.163	0.163
		MC SD	0.509	0.704	0.589	0.183	0.168
		MC Mean	1.988	1.005	1.906	0.460	-0.210
	5%	IBM SE	0.289	0.437	0.280	0.100	0.100
		MC SD	0.272	0.448	0.299	0.090	0.097
		MC Mean	1.991	1.033	1.897	0.447	-0.206
100	20%	IBM SE	0.299	0.455	0.311	0.105	0.104
		MC SD	0.298	0.467	0.307	0.122	0.113
		MC Mean	2.007	1.049	1.821	0.480	-0.220
	40%	IBM SE	0.325	0.482	0.350	0.116	0.114
		MC SD	0.335	0.491	0.344	0.109	0.108
		MC Mean	1.989	1.004	1.955	0.478	-0.196
	5%	IBM SE	0.208	0.308	0.203	0.070	0.070
		MC SD	0.209	0.300	0.210	0.069	0.077
		MC Mean	1.999	0.991	1.956	0.488	-0.212
200	20%	IBM SE	0.215	0.320	0.226	0.074	0.074
		MC SD	0.213	0.312	0.216	0.068	0.072
		MC Mean	2.038	0.966	1.883	0.474	-0.211
	40%	IBM SE	0.229	0.339	0.257	0.082	0.081
		MC SD	0.262	0.356	0.246	0.081	0.081
		MC Mean	2.022	0.955	1.980	0.475	-0.205
	5%	IBM SE	0.170	0.254	0.168	0.057	0.057
		MC SD	0.187	0.275	0.150	0.054	0.048
		MC Mean	2.009	0.974	1.997	0.477	-0.214
300	20%	IBM SE	0.175	0.264	0.189	0.060	0.060
		MC SD	0.169	0.249	0.184	0.060	0.062
		MC Mean	2.023	0.999	1.908	0.463	-0.198
	40%	IBM SE	0.188	0.279	0.213	0.067	0.066
		MC SD	0.175	0.284	0.235	0.067	0.063
		MC Mean	2.027	0.969	1.999	0.481	-0.211
	5%	IBM SE	0.131	0.196	0.131	0.044	0.044
		MC SD	0.127	0.193	0.153	0.042	0.039
		MC Mean	1.997	1.015	1.974	0.473	-0.200
500	20%	IBM SE	0.136	0.203	0.145	0.047	0.046
		MC SD	0.134	0.201	0.146	0.049	0.047
		MC Mean	2.005	1.017	1.916	0.474	-0.200
	40%	IBM SE	0.146	0.216	0.166	0.052	0.051
		MC SD	0.141	0.189	0.170	0.047	0.047

Table 2.2: MSPE and MAPE for the 3-step-ahead prediction. The results are based on 300 predicted observations.

MSPE	5%	20%	40%
Cens	1.888	2.248	2.277
LOD	2.003	3.684	6.254
Unc	1.900	2.369	3.266
MAPE	5%	20%	40%
Cens	1.113	1.202	1.209
LOD	1.134	1.552	2.046
Unc	1.118	1.248	1.437
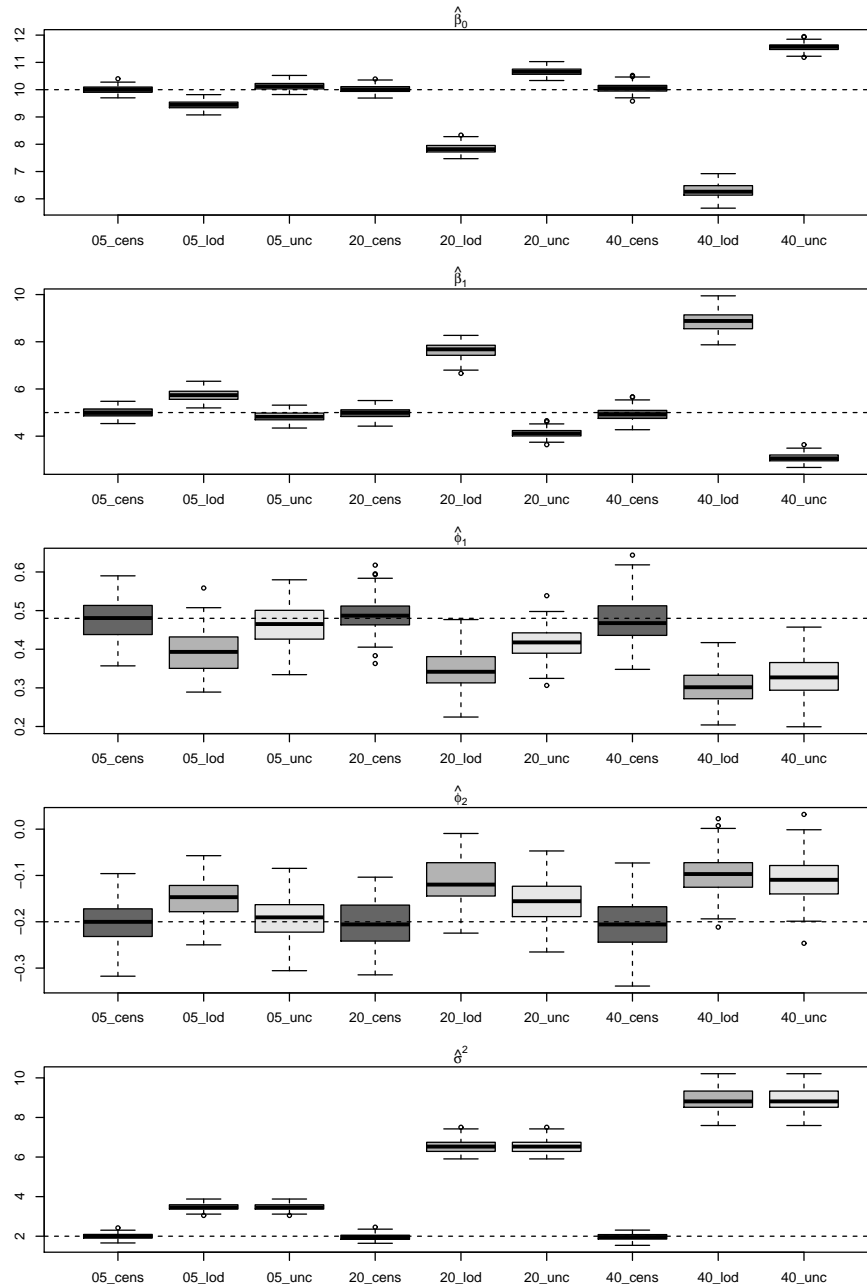
Figure 2.4: Boxplot of the parameter estimates for $n = 497$. The dotted line indicates the true value of the parameter. *05_cens* indicates the result for the censored case with censoring level of 5%, *05_lod* indicates the result for the LOD procedure with censoring level of 5% and so on.

2.6 Application on real datasets

2.6.1 Cloud ceiling height dataset

In this subsection we consider the analysis of the dataset described in Subsection 1.1.1. There were three missing observations and for those observations we considered the censoring interval $A_j = (-\infty, \infty)$. For model selection we considered $p = 1, 2$ and 3, fixing the maximum number of iterations at $W = 600$, and the cutoff point at $c = 0.12$ (see Kuhn and Lavielle, 2005). The AR(2)-CR model presented the smallest AIC value and the smallest BIC value, and therefore it was selected.

Table 2.3: Parameter estimates of AR(p)-CR model for the log-transformed cloud ceiling height data. Bold entries represent the best model.

p	loglik	AIC	BIC	$\hat{\beta}_0$	$\hat{\sigma}^2$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$
1	-470.7	947	961	4.069 (0.182)	0.872 (0.057)	0.808 (0.024)		
2	**-466.5**	**941**	**959**	**4.059 (0.215)**	**0.869 (0.056)**	**0.665 (0.044)**	**0.174 (0.045)**	
3	-466.2	942	965	4.054 (0.232)	0.874 (0.057)	0.656 (0.045)	0.108 (0.056)	0.086 (0.046)

It is worth noting that the model select by our approach is the same that was select by the two non-naive approaches considered by Park et al. (2007) (see Table 1.1), although the estimates obtained differ slightly. Figure 2.5 presents the ceiling height data with the imputed values, estimated using $\widehat{\mathbf{y}} = E\{\mathbf{y}|\mathbf{V}, \mathbf{C}, \widehat{\boldsymbol{\theta}}\}$ obtained from the SAEM algorithm.

Figure 2.5: Censored time series of log-transformed hourly cloud ceiling height in San Francisco during March 1989. The dashed line represents the augmented serie based on the fitted AR(2)-CR model.

For the selected model (AR(2)-CR) , Figure 2.6 shows the convergence of the estimates obtained through the SAEM algorithm. The dashed line indicates the iteration where the simulations start being smoothed. It is important to note that convergence is attained quickly.
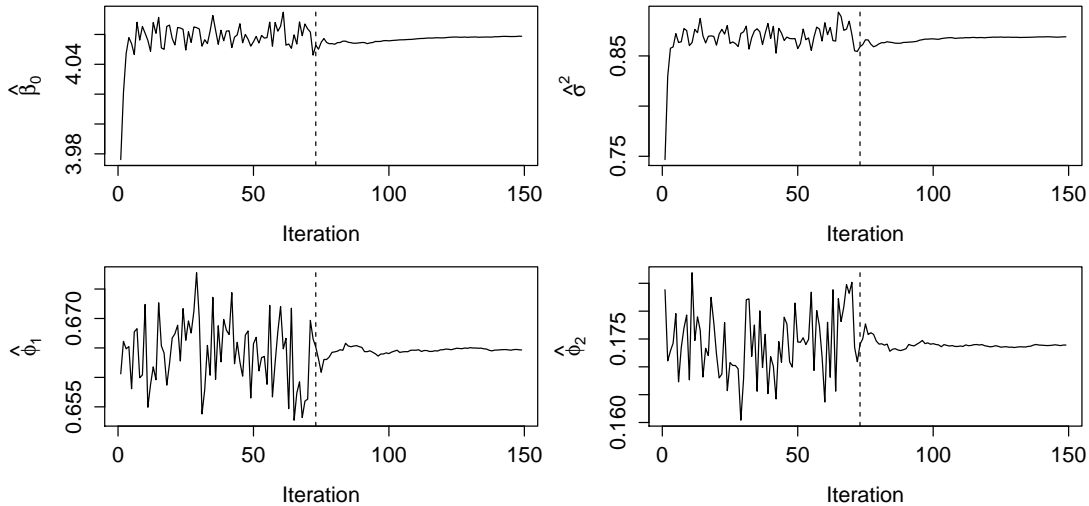
Figure 2.6: Convergence of the SAEM parameter estimates for the AR(2)-CR model.

2.6.2 Total phosphorus concentration dataset

In this subsection we consider the analysis of the dataset described in Subsection 1.1.2. Following Wang and Chan (2016a), since the data present seasonality, we will consider the model

$$\log(P_t) = \sum_{j=1}^{4} \left[\beta_{1,j} S_{j,t} + \beta_{2,j} S_{j,t} \log(Q_t) \right] + \varepsilon_t,$$

where S_j is a quarter indicator variable, $j = 1, 2, 3, 4$, such that $S_{j,t} = 1$ if the tth observation belongs to the jth quarter and $S_{j,t} = 0$ otherwise, with the first quarter comprising January to March, the second quarter April to June, etc; and ε_t is as defined in (1.4.2).

For model selection we fitted an AR(p)-CR model, as defined in (2.1.1)–(2.1.2), with $p = 1, 2, 3$ and 4. The initial estimates were chosen by considering the censored values as real observed values. For the SAEM algorithm, we fixed the maximum number of iterations at $W = 600$ and the cutoff point at $c = 0.12$ (see Kuhn and Lavielle, 2005). Missing observations were handled by replacing them with their conditional expectations at each step of the SAEM algorithm.

Table 2.4: Criteria for model selection for the log-transformed phosphorus concentration data.

p	AIC	BIC	loglik
1	**269.263**	**301.248**	-124.632
2	270.359	305.543	-124.180
3	272.023	310.405	-124.012
4	269.618	311.198	-121.809

Since the AIC criterion is close for $p = 1$ and $p = 4$, we fitted these two models preserving the 12 latest observations for prediction comparison purposes and calculated the mean square prediction error (MSPE) defined as

$$\text{MSPE}_p = \frac{1}{12} \sum_{t=170}^{181} \left(\log(P_t) - \widehat{\log(P_t)}_p \right)^2,$$

where $\widehat{\log(P_t)}_p$ denotes the prediction of the tth value under the AR(p)-CR model, $t = 170, \dots, 181$. Once the obtained results are $\text{MSPE}_1 = 0.1366$ and $\text{MSPE}_4 = 0.1473$, we selected the AR(1)-CR model.

Table 2.5: Estimated parameters and their standard errors for the AR(1)-CR model.

Parameter	Estimate	SE
$\beta_{1,1}$	-4.580	0.429
$\beta_{1,2}$	-2.890	0.708
$\beta_{1,3}$	-4.172	0.421
$\beta_{1,4}$	-4.998	0.463
$\beta_{2,1}$	0.354	0.072
$\beta_{2,2}$	0.167	0.099
$\beta_{2,3}$	0.365	0.071
$\beta_{2,4}$	0.409	0.084
σ^2	0.245	0.028
ϕ_1	-0.103	0.084

The estimated parameters of the selected model (AR(1)-CR) are presented in Table 2.5 and Figure 2.8 shows their convergence. Figure 2.7 presents the log-transformed phosphorus concentration data with the imputed values, estimated using $\widehat{\log(\mathbf{P})} = E\{\log(\mathbf{P})|\mathbf{V}, \mathbf{C}, \widehat{\theta}\}$ obtained from the SAEM algorithm.
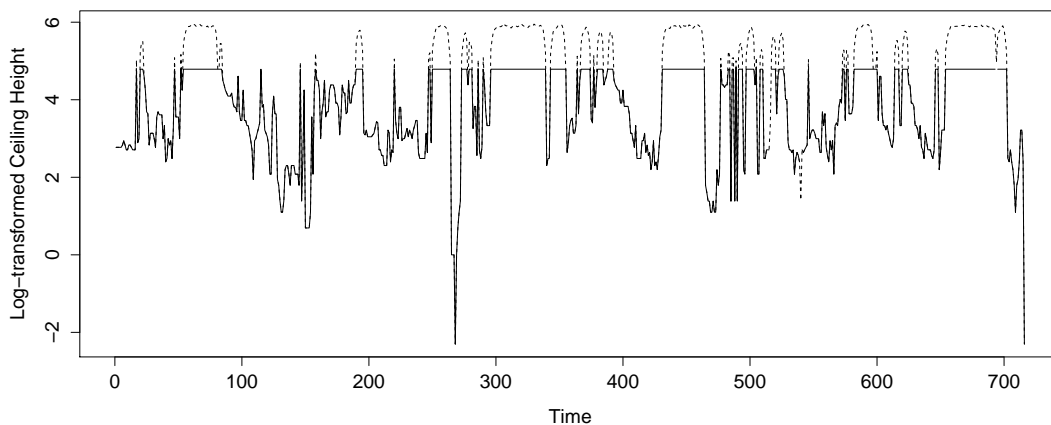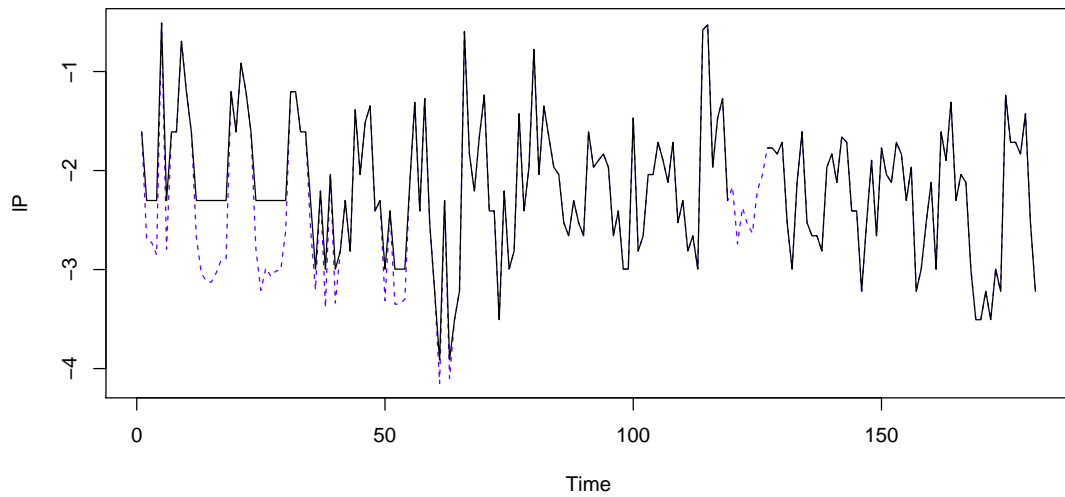
Figure 2.7: Log-transformed phosphorus concentration (lP) with the augmented series based on the fitted AR(1)-CR model (dashed blue line).
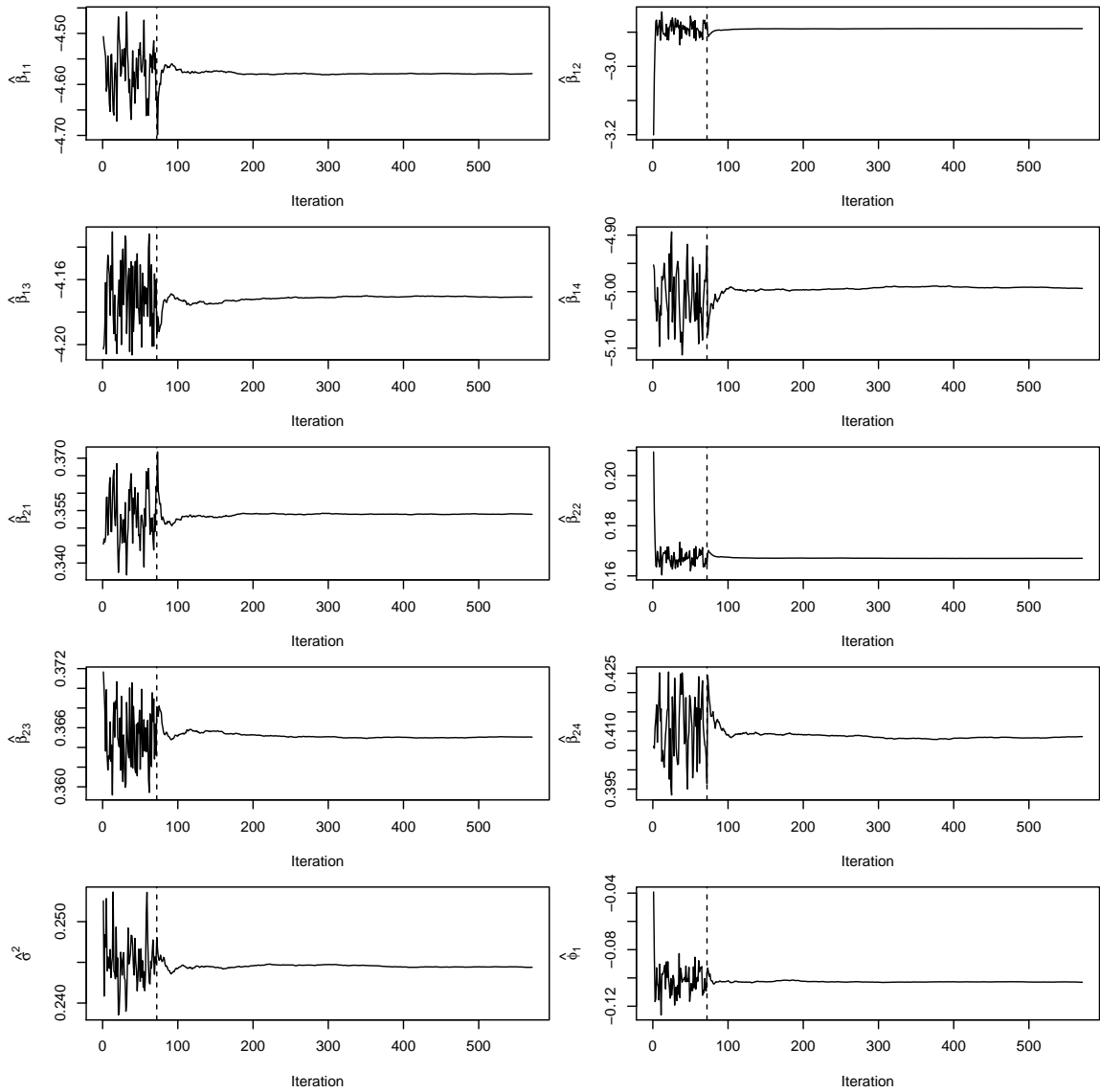
Figure 2.8: Convergence of the SAEM parameter estimates for the AR(1)-CR model.

2.7   Conclusions

This chapter describes a likelihood-based approach to perform inference and prediction in autoregressive censored linear models. We developed a stochastic approximation of the EM algorithm, called the SAEM algorithm, to obtain the maximum likelihood estimates of model parameters.

For practical demonstration, the method was applied to two datasets, measured subject to detection limits of the recording device. We also used simulation to investigate the properties of predictions and parameter estimates, and the robustness of the SAEM algorithm. In this simulation study, comparisons were made between inferences based on the censored data and inferences based on complete-data obtained by a crude/ad hoc imputation method. We showed that the differences in inference between the two approaches can be substantial. Moreover, the SAEM algorithm leads to an improvement in the computation speed of the ML estimates, as opposed to the Monte Carlo EM (MCEM) algorithms, especially when the censoring level increases.

Chapter 3

Diagnostic Analysis

In statistical modeling robustness is desired, meaning that the estimates obtained from the proposed model are not influenceable by small deviations on the data. The interest in diagnostic analysis has grown steadily in recent years. Influence diagnostics are widely used in statistical modeling to identify and evaluate aberrant and influential points which may cause unwanted effects on estimation and goodness of fit.

Of course, when dealing with autocorrelated data, omission approaches are inappropriate. As an alternative, Cook (1986) developed a general method for assessing the local influence of a model perturbation that uses the log-likelihood contours rather than omission approaches. The key idea of the local influence approach is to utilize the concept of normal curvature in differential geometry in assessing the local behavior of the likelihood displacement function (Zhu et al., 2007).

Zhu and Lee (2001) proposed a method to assess the local infuence in a minor perturbation of a statistical model with incomplete data, utilizing Cook's approach to the conditional expectation of the complete-data log-likelihood function in the EM algorithm. Recently, Zhu et al. (2007) developed a perturbation manifold to select an appropriate perturbation for statistical models without missing data. By using the results of Zhu and Lee (2001) and Zhu et al. (2007), in this chapter we derive local influence diagnostics for the AR(p)-CR model on the basis of the Q-function under three perturbation schemes.

3.1   Local influence

In this section, we derive the normal curvature of the local influence (Zhu and Lee, 2001; Cook, 1986) for some common perturbation schemes either in the model or the data. We will consider the response perturbation scheme, the explanatory variable perturbation and the matrix scale perturbation for this purpose.

Consider a perturbation vector $\boldsymbol{\omega} = (\omega_1, ..., \omega_g)^\top$, varying in an open set $\Omega \subset \mathbb{R}^g$. Let $\ell(\boldsymbol{\theta}, \boldsymbol{\omega}|\mathbf{y}_c)$ be the complete-data log-likelihood function of the model perturbed with $\boldsymbol{\omega}$, called the perturbed model. It is assumed that there exists $\boldsymbol{\omega}_0 \in \Omega$, a $g \times 1$ vector of no perturbation, such that $\ell(\boldsymbol{\theta}, \boldsymbol{\omega}_0|\mathbf{y}_c) = \ell(\boldsymbol{\theta}|\mathbf{y}_c)$ for all $\boldsymbol{\theta}$. Let $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ denote the estimate of $\boldsymbol{\theta}$ which maximizes the function

$$\hat{Q}(\boldsymbol{\theta}, \boldsymbol{\omega}) = \mathrm{E}\left[\ell(\boldsymbol{\theta}, \boldsymbol{\omega}|\mathbf{y}_c)|\mathbf{V}, \mathbf{C}, \widehat{\boldsymbol{\theta}}\right].$$

For $\boldsymbol{\omega} = \boldsymbol{\omega}_0$, the conditional expectation of the complete-data log-likelihood function for the model defined in (2.1.1)–(2.1.2) is given by

$$\hat{Q}(\boldsymbol{\theta}) = -\frac{1}{2}\left[n\log\sigma^2 + \log|\mathbf{M}_p(\boldsymbol{\phi})| + \frac{1}{\sigma^2}\left(\mathrm{tr}(\widehat{\mathbf{yy}^\top}\mathbf{M}_n^{-1}(\boldsymbol{\phi}))\right. \right.$$
$$\left.\left. -2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{M}_n^{-1}(\boldsymbol{\phi})\widehat{\mathbf{y}} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{M}_n^{-1}(\boldsymbol{\phi})\mathbf{X}\boldsymbol{\beta}\right)\right], \tag{3.1.1}$$

where

$$\widehat{\mathbf{y}} = E\{\mathbf{y}|\mathbf{V}, \mathbf{C}, \widehat{\boldsymbol{\theta}}\} \text{ and } \widehat{\mathbf{yy}^\top} = E\{\mathbf{yy}^\top|\mathbf{V}, \mathbf{C}, \widehat{\boldsymbol{\theta}}\}$$

are stochastically approximated by the SAEM algorithm.

To assess the influence of minor perturbations on the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ for incomplete data problems, Zhu and Lee (2001) proposed to consider the Q-displacement function

$$f_Q(\boldsymbol{\omega}) = 2\left[\hat{Q}\left(\widehat{\boldsymbol{\theta}}\right) - \hat{Q}\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}\right)\right]$$

and the associated influence graph

$$\alpha(\boldsymbol{\omega}) = \begin{pmatrix} \boldsymbol{\omega} \\ f_Q(\boldsymbol{\omega}) \end{pmatrix}.$$

Following the approach of Cook (1986) and Zhu and Lee (2001), the normal curvature $C_{f_Q, \mathbf{d}}$ of $\alpha(\boldsymbol{\omega})$ at $\boldsymbol{\omega}_0$ in the direction of some unit vector $\mathbf{d}$ can be used to summarize the local behavior of $f_Q(\boldsymbol{\omega})$. It can be shown that

$$C_{f_Q, \mathbf{d}} = -2\mathbf{d}^\top\ddot{Q}_{\boldsymbol{\omega}_o}\mathbf{d} \text{ and } -\ddot{Q}_{\boldsymbol{\omega}_0} = \boldsymbol{\Delta}_{\boldsymbol{\omega}_0}^\top\left\{-\ddot{Q}(\widehat{\boldsymbol{\theta}})\right\}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\omega}_0},$$

where

$$\ddot{Q}(\widehat{\boldsymbol{\theta}}) = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\{\hat{Q}(\boldsymbol{\theta})\}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \text{ and } \boldsymbol{\Delta}_{\boldsymbol{\omega}} = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\omega}^\top}\{\hat{Q}(\boldsymbol{\theta}, \boldsymbol{\omega})\}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}}.$$

Following the same procedure as in Cook (1986), the quantity $-\ddot{Q}_{\boldsymbol{\omega}_0}$ is useful for

detecting influential observations. From the spectral decomposition of a symmetric matrix

$$-2\ddot{Q}\boldsymbol{\omega}_0 = \sum_{k=1}^{g} \lambda_k \mathbf{e}_k \mathbf{e}_k^\top, \tag{3.1.2}$$

where $\{(\lambda_k, \mathbf{e}_k), k = 1, \ldots, g\}$ are eigenvalue–eigenvector pairs of $-2\ddot{Q}\boldsymbol{\omega}_0$ with $\lambda_1 \geq \ldots \geq \lambda_r > \lambda_{r+1} = \ldots = 0$ and orthonormal eigenvectors $\{\mathbf{e}_k, k = 1, \ldots, g\}$, Zhu and Lee (2001) propose to inspect all eigenvectors corresponding to nonzero eigenvalues for capturing more information.

Following the work of Zhu and Lee (2001), we consider the following aggregated contribution vector of all eigenvectors that correspond to nonzero eigenvalues. Let $\tilde{\lambda}_k = \lambda_k/(\lambda_1 + \ldots + \lambda_r)$, $\mathbf{e}_k^2 = (\mathbf{e}_{k1}^2, \ldots, \mathbf{e}_{kg}^2)^\top$ and

$$M(0) = \sum_{k=1}^{r} \tilde{\lambda}_k \mathbf{e}_k^2. \tag{3.1.3}$$

The lth component of $M(0)$, $M(0)_l$, is equal to $\sum_{k=1}^{r} \tilde{\lambda}_k \mathbf{e}_{kl}^2$. The assessment of influential cases is based on visual inspection of $\{M(0)_l, l = 1, \ldots, g\}$ plotted against the index l.

The lth case may be regarded as influential if $M(0)_l$ is larger than some benchmark value. So far, there is no general rule to choose the benchmark value. Let $\overline{M}(0)$ and $SM(0)$ denote, respectively, the mean and standard error of $\{M(0)_l : l = 1, \ldots, g\}$, where $\overline{M}(0) = 1/g$ (Zhu and Lee, 2001). Poon and Poon (1999) propose using $2\overline{M}(0)$ as a benchmark for $M(0)$. An alternative that takes into account the variation of $M(0)$ is to take $\overline{M}(0) + 2SM(0)$ as a benchmark (Zhu and Lee, 2001). According to Lee and Xu (2004), the exact choice of the function of $\overline{M}(0)$ as the benchmark is subjective. Lee and Xu (2004) also propose using

$$\overline{M}(0) + c^* SM(0), \tag{3.1.4}$$

where c^* is a selected constant, and depending on the specific application, c^* may be chosen suitably.

In particular, one may be interested in assessing the influence on a subset $\boldsymbol{\theta}_1$ of $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top\right)^\top$. For such situations, Cook (1986) showed that the curvature in the direction $\mathbf{d}$ is given by

$$C_{f_Q, \mathbf{d}} = 2\mathbf{d}^\top \Delta_{\boldsymbol{\omega}_0}^\top \left(\left\{ -\ddot{Q}(\widehat{\boldsymbol{\theta}}) \right\}^{-1} - B_{22} \right) \Delta_{\boldsymbol{\omega}_0} \mathbf{d}, \tag{3.1.5}$$

where

$$B_{22} = \begin{pmatrix} 0 & 0 \\ 0 & \left\{ -\ddot{Q}_{22}(\widehat{\boldsymbol{\theta}}) \right\}^{-1} \end{pmatrix}$$

and $\ddot{Q}_{22}(\widehat{\boldsymbol{\theta}})$ is obtained from the partition of $\ddot{Q}(\widehat{\boldsymbol{\theta}})$ according to the partition of $\boldsymbol{\theta}$. The influential observations can be identified similarly to the previous case, and the respective aggregated

contribution vector of all eigenvectors of

$$-2\Delta_{\boldsymbol{\omega}_0}^\top \left(\left\{ -\ddot{\hat{Q}}(\widehat{\boldsymbol{\theta}}) \right\}^{-1} - B_{22} \right) \Delta_{\boldsymbol{\omega}_0}$$

will be denoted by $M_{\boldsymbol{\theta}_1}(0)$.

3.1.1 The Hessian matrix

In order to obtain the diagnostic measures for local influence, we need to compute the Hessian matrix

$$\ddot{\hat{Q}}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \left\{ \hat{Q}(\boldsymbol{\theta}) \right\}$$

evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$. Let $\mathbf{M}_n = \mathbf{M}_n(\boldsymbol{\phi})$ be as given in (1.4.3). Taking second derivatives of the conditional expectation of the complete-data log-likelihood function given in (3.1.1) with respect to $\boldsymbol{\theta}$, we obtain the elements of $\ddot{\hat{Q}}(\boldsymbol{\theta})$:

$$
\begin{aligned}
\frac{\partial^2 \hat{Q}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} &= -\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{M}_n^{-1} \mathbf{X}, \\
\frac{\partial^2 \hat{Q}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \sigma^2} &= -\frac{1}{\sigma^4} \left(\mathbf{X}^\top \mathbf{M}_n^{-1} \widehat{\mathbf{y}} - \mathbf{X}^\top \mathbf{M}_n^{-1} \mathbf{X} \boldsymbol{\beta} \right), \\
\frac{\partial^2 \hat{Q}(\boldsymbol{\theta})}{\partial \phi_i \partial \boldsymbol{\beta}^\top} &= \frac{1}{\sigma^2} (\widehat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^\top \frac{\partial \mathbf{M}_n^{-1}}{\partial \phi_i} \mathbf{X}, \quad i = 1, \dots, p, \\
\frac{\partial^2 \hat{Q}(\boldsymbol{\theta})}{\partial (\sigma^2)^2} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \left(\text{tr} \left[\widehat{\mathbf{y}^2} \mathbf{M}_n^{-1} \right) - 2\widehat{\mathbf{y}}^\top \mathbf{M}_n^{-1} \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{M}_n^{-1} \mathbf{X}\boldsymbol{\beta} \right], \\
\frac{\partial^2 \hat{Q}(\boldsymbol{\theta})}{\partial \phi_i \partial \sigma^2} &= \frac{1}{2\sigma^4} \left[\text{tr} \left(\widehat{\mathbf{y}^2} \frac{\partial \mathbf{M}_n^{-1}}{\partial \phi_i} \right) - 2\widehat{\mathbf{y}}^\top \frac{\partial \mathbf{M}_n^{-1}}{\partial \phi_i} \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{X}^\top \frac{\partial \mathbf{M}_n^{-1}}{\partial \phi_i} \mathbf{X}\boldsymbol{\beta} \right], \quad i = 1, \dots, p, \\
\frac{\partial^2 \hat{Q}(\boldsymbol{\theta})}{\partial \phi_i \partial \phi_j} &= -\frac{1}{2} \text{tr} \left\{ \frac{\partial}{\partial \phi_j} \left(\mathbf{M}_p^{-1} \frac{\partial \mathbf{M}_p}{\partial \phi_i} \right) \right\} - \frac{1}{2\sigma^2} \left[\text{tr} \left(\widehat{\mathbf{y}^2} \frac{\partial^2 \mathbf{M}_n^{-1}}{\partial \phi_i \partial \phi_j} \right) - 2\widehat{\mathbf{y}}^\top \frac{\partial^2 \mathbf{M}_n^{-1}}{\partial \phi_i \partial \phi_j} \mathbf{X}\boldsymbol{\beta} \right. \\
&\quad \left. + \boldsymbol{\beta}^\top \mathbf{X}^\top \frac{\partial^2 \mathbf{M}_n^{-1}}{\partial \phi_i \partial \phi_j} \mathbf{X}\boldsymbol{\beta} \right], \quad i = 1, \dots, p; j = 1, \dots, p.
\end{aligned}
$$

3.2 Perturbation schemes

In this section, we will evaluate the matrix $\Delta_{\boldsymbol{\omega}_0}$ for AR(p)-CR models under the following perturbation schemes: *perturbation of the response variable* carried out on the response values, which may indicate observations with large influence on their own predicted values (in our case, the response variables are $\mathbf{V}'s$); *scale perturbation* performed on the scale matrix $\boldsymbol{\Sigma} = \sigma^2 \mathbf{M}_n(\boldsymbol{\phi})$, which may reveal individuals that are most influential on the scale structure; and finally *perturbation of explanatory variables*.

3.2.1 Response perturbation

Suppose that the response vector $\mathbf{V}$ is perturbed according to $\mathbf{V}(\boldsymbol{\omega}) = \mathbf{V} + \boldsymbol{\omega}$, where $\boldsymbol{\omega} = (\omega_1, ..., \omega_n)^\top$. Now, substituting $\mathbf{V}$ for $\mathbf{V}(\boldsymbol{\omega})$ in (2.1.2), we can write the perturbed model as

$$Y_t(\boldsymbol{\omega}) \leq V_t \text{ if } C_t = 1 \text{ and } Y_t(\boldsymbol{\omega}) = V_t \text{ if } C_t = 0, \tag{3.2.1}$$

where $\mathbf{y}(\boldsymbol{\omega}) = \mathbf{y} - \boldsymbol{\omega}$. Hence, the perturbed Q-function, $\hat{Q}(\boldsymbol{\theta}, \boldsymbol{\omega})$, is as given in (3.1.1), with $\widehat{\mathbf{y}}$ and $\widehat{\mathbf{yy}^\top}$ being replaced by $\widehat{\mathbf{y}(\boldsymbol{\omega})} = \widehat{\mathbf{y}} - \boldsymbol{\omega}$ and $\widehat{\mathbf{y}(\boldsymbol{\omega})\mathbf{y}(\boldsymbol{\omega})^\top} = \widehat{\mathbf{yy}^\top} - \widehat{\mathbf{y}}\boldsymbol{\omega}^\top - \boldsymbol{\omega}\widehat{\mathbf{y}}^\top + \boldsymbol{\omega}\boldsymbol{\omega}^\top$, respectively.

Under this perturbation scheme, the vector of no perturbation is given by $\boldsymbol{\omega}_0 = (0, \ldots, 0)^\top$ and

$$\boldsymbol{\Delta}_{\boldsymbol{\omega}_0} = \begin{pmatrix} \boldsymbol{\Delta}_\beta \\ \boldsymbol{\Delta}_{\sigma^2} \\ \boldsymbol{\Delta}_\phi \end{pmatrix}$$

is an $(l + p + 1) \times n$ matrix and has components given by

$$\begin{aligned}
\boldsymbol{\Delta}_\beta &= -\frac{1}{\widehat{\sigma}^2} \mathbf{X}^\top \mathbf{M}^{-1}(\widehat{\boldsymbol{\phi}}), \\
\boldsymbol{\Delta}_{\sigma^2} &= -\frac{1}{\widehat{\sigma}^4} \left(\widehat{\mathbf{y}} - \mathbf{X}\widehat{\boldsymbol{\beta}} \right)^\top \mathbf{M}^{-1}(\widehat{\boldsymbol{\phi}}), \\
\boldsymbol{\Delta}_{\phi_i} &= \frac{1}{\widehat{\sigma}^2} \left(\widehat{\mathbf{y}} - \mathbf{X}\widehat{\boldsymbol{\beta}} \right)^\top \frac{\partial \mathbf{M}^{-1}}{\partial \phi_i}(\widehat{\boldsymbol{\phi}}), \ i = 1, \ldots, p.
\end{aligned} \tag{3.2.2}$$

3.2.2 Scale matrix perturbation

In order to study the effects of perturbation of the scale matrix, we consider the perturbation scheme of the form

$$\Sigma(\boldsymbol{\omega}) = \mathbf{D}(\boldsymbol{\omega}) \, \Sigma = \sigma^2 \mathbf{D}(\boldsymbol{\omega}) \mathbf{M}(\boldsymbol{\phi}),$$

where $\mathbf{D}(\boldsymbol{\omega})$ is a $n \times n$ diagonal matrix with value ω_t in tth diagonal element. Under this scheme, the non-perturbed model is obtained when $\boldsymbol{\omega}_0 = (1, \ldots, 1)^\top$. Thus, considering this perturbation scheme, $\boldsymbol{\Delta}_{\boldsymbol{\omega}_0}$ is an $(l + p + 1) \times n$ matrix and has components given by

$$\begin{aligned}
\boldsymbol{\Delta}_{\beta_t} &= \frac{1}{2\widehat{\sigma}^2} \left(-2\mathbf{X}^\top \mathbf{d}(t)\mathbf{M}^{-1}(\widehat{\boldsymbol{\phi}})\widehat{\mathbf{y}} + \mathbf{X}^\top \left(\mathbf{M}^{-1}(\widehat{\boldsymbol{\phi}})\mathbf{d}(t) + \mathbf{d}(t)\mathbf{M}^{-1}(\widehat{\boldsymbol{\phi}}) \right) \mathbf{X}\widehat{\boldsymbol{\beta}} \right), \\
\boldsymbol{\Delta}_{\sigma^2 t} &= -\frac{1}{2\widehat{\sigma}^4} \left[\text{tr}\left(\widehat{\mathbf{y}^2}\mathbf{M}^{-1}(\widehat{\boldsymbol{\phi}})\mathbf{d}(t) \right) - 2\widehat{\mathbf{y}}^\top \mathbf{M}^{-1}(\widehat{\boldsymbol{\phi}})\mathbf{d}(t)\mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{M}^{-1}(\widehat{\boldsymbol{\phi}})\mathbf{d}(t)\mathbf{X}\widehat{\boldsymbol{\beta}} \right], \\
\boldsymbol{\Delta}_{\phi_i t} &= \frac{1}{2\widehat{\sigma}^2} \left[\text{tr}\left(\widehat{\mathbf{y}^2}\frac{\partial \mathbf{M}^{-1}}{\partial \phi_i}(\widehat{\boldsymbol{\phi}})\mathbf{d}(t) \right) - 2\widehat{\mathbf{y}}^\top \frac{\partial \mathbf{M}^{-1}}{\partial \phi_i}(\widehat{\boldsymbol{\phi}})\mathbf{d}(t)\mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \frac{\partial \mathbf{M}^{-1}}{\partial \phi_i}(\widehat{\boldsymbol{\phi}})\mathbf{d}(t)\mathbf{X}\widehat{\boldsymbol{\beta}} \right],
\end{aligned}$$

for $i = 1, \ldots, p$, $t = 1, \ldots, n$, and where $\mathbf{d}(t)$ is an $n \times n$ matrix with the tth diagonal element equal to one and the others equal to zero.

3.2.3 Explanatory variable perturbation

In the interest of studying the influence that perturbation in the explanatory variables may produce on the parameter estimates, and taking into account the interest of perturbing a subset of the explanatory variables (commonly the subset of continuous explanatory variables), we partitionate the matrix $\mathbf{X}$ as $\begin{bmatrix} \mathbf{X}^p & \mathbf{X}^n \end{bmatrix}$, such that $\mathbf{X}^p$ has dimension $n \times l_p$ and contains the columns of $\mathbf{X}$ that we are interested in perturbing, and $\mathbf{X}^n$ has dimension $n \times l_n$ and contains the remaining columns, where $l_p + l_n = l$.

Therefore, we replace $\mathbf{X}$ in the perturbed Q-function by

$$\mathbf{X}(\omega) = \begin{bmatrix} \mathbf{X}^p + \mathbf{W} & \mathbf{X}^n \end{bmatrix},$$

with $\mathbf{W} = \boldsymbol{\omega} \mathbf{1}^\top$, where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$ and $\mathbf{1}$ is a $l_p \times 1$ vector of ones, thus $\mathbf{W}$ is an $n \times l_p$ matrix. Let $\mathbf{1}^\mathbf{p}$ be an $l \times 1$ vector, such that $\mathbf{1}^\mathbf{p}{}_k = 1$ if the kth column of $\mathbf{X}$ is contained in $\mathbf{X}^p$ and $\mathbf{1}^\mathbf{p}{}_k = 0$ otherwise, $k = 1, \ldots, l$. Thus, considering the non-perturbed vector $\boldsymbol{\omega}_0 = (0, \ldots, 0)^\top$, $\Delta_{\boldsymbol{\omega}_0}$ has the following elements:

$$
\begin{aligned}
\Delta_\beta &= \frac{1}{\widehat{\sigma}^2} \left[\mathbf{1}^\mathbf{p} \left(\widehat{\mathbf{y}} - \mathbf{X}\widehat{\beta} \right)^\top - \left(\mathbf{1}^{\mathbf{p}\top} \beta \right) \mathbf{X}^\top \right] \mathbf{M}^{-1}(\widehat{\phi}), \\
\Delta_{\sigma^2} &= -\frac{1}{\widehat{\sigma}^4} \left(\mathbf{1}^{\mathbf{p}\top} \beta \right) \left(\widehat{\mathbf{y}} - \mathbf{X}\widehat{\beta} \right)^\top \mathbf{M}^{-1}(\widehat{\phi}), \\
\Delta_{\phi_i} &= \frac{1}{\widehat{\sigma}^2} \left(\mathbf{1}^{\mathbf{p}\top} \beta \right) \left(\widehat{\mathbf{y}} - \mathbf{X}\widehat{\beta} \right)^\top \frac{\partial \mathbf{M}^{-1}}{\partial \phi_i}(\widehat{\phi}), \ i = 1, \ldots, p.
\end{aligned}
$$

3.3 Simulation studies

Two simulation studies were conducted to examine the performance of the proposed model. In all the simulation studies we consider a left-censored AR(2)-CR model, as defined in (2.1.1)–(2.1.2), with one explanatory variable, and parameters set at $\beta_0 = 5$, $\beta_1 = 2$, $\sigma^2 = 2$, $\phi_1 = 0.48$ and $\phi_2 = -0.2$. Besides, the data generating process is as described in Section 2.5.

The initial estimates were chosen by considering the censored values as real observed values. For the SAEM algorithm, we fixed the maximum number of iterations at $W = 400$ and the cutoff point at $c = 0.18$ (see Kuhn and Lavielle, 2005).

3.3.1 First simulation study

This study illustrates the proposed diagnostic measures by generating only one sample of size $n = 200$, censoring it in three different levels (5%, 20% and 40%) and replacing the maximum value of each sample by $y_{\max} = y_{\max} + 3\text{sd}(\mathbf{y})$. The atypical point corresponds to the point #54 and Figure 3.1 presents the generated dataset, the censoring limit and the replaced value for each censoring level.
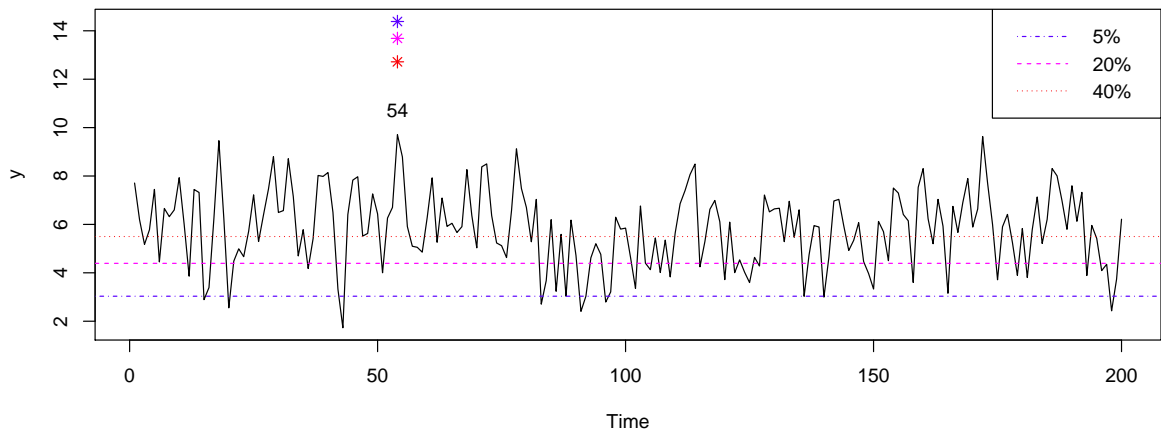
Figure 3.1: Generated time serie with one outlier (asterisk) and censoring limits for 5% (dash-dot blue line), 20% (dashed light violet line) and 40% (dotted red line) of censored data.

Following the approach described in Section 3.1, Figure 3.2 depicts the index plots of $M(0)$ for the response perturbation, explanatory variable perturbation and scale matrix perturbation, respectively, along with the Lee and Xu (2004) benchmark, given in (3.1.4), computed for $c^* = 3.5$.

The perturbed observation (#54) was detected as influential for all the schemes and levels of censoring considered. However, it is worth noting that the response perturbation scheme seems to be more affected by the dependence structure of the data, also detecting the observations #52, #53 and #55 as influential for all levels of censoring.

3.3.2 Second simulation study

The second simulation study is a Monte Carlo experiment intending to evaluate the capability of the methodology to detect atypical points. Here we simulated 100 samples of size $n = 200$ and censored the samples in three different levels (5%, 20% and 40%). For each sample and level of censoring, we generate an atypical point in three different ways: *(a)* substituting $\max(\mathbf{y})$ by $\max(\mathbf{y}) + k * \text{sd}(\mathbf{y})$, $k = 1, 2, 3$ and 5; *(b)* substituting the 80th response observation
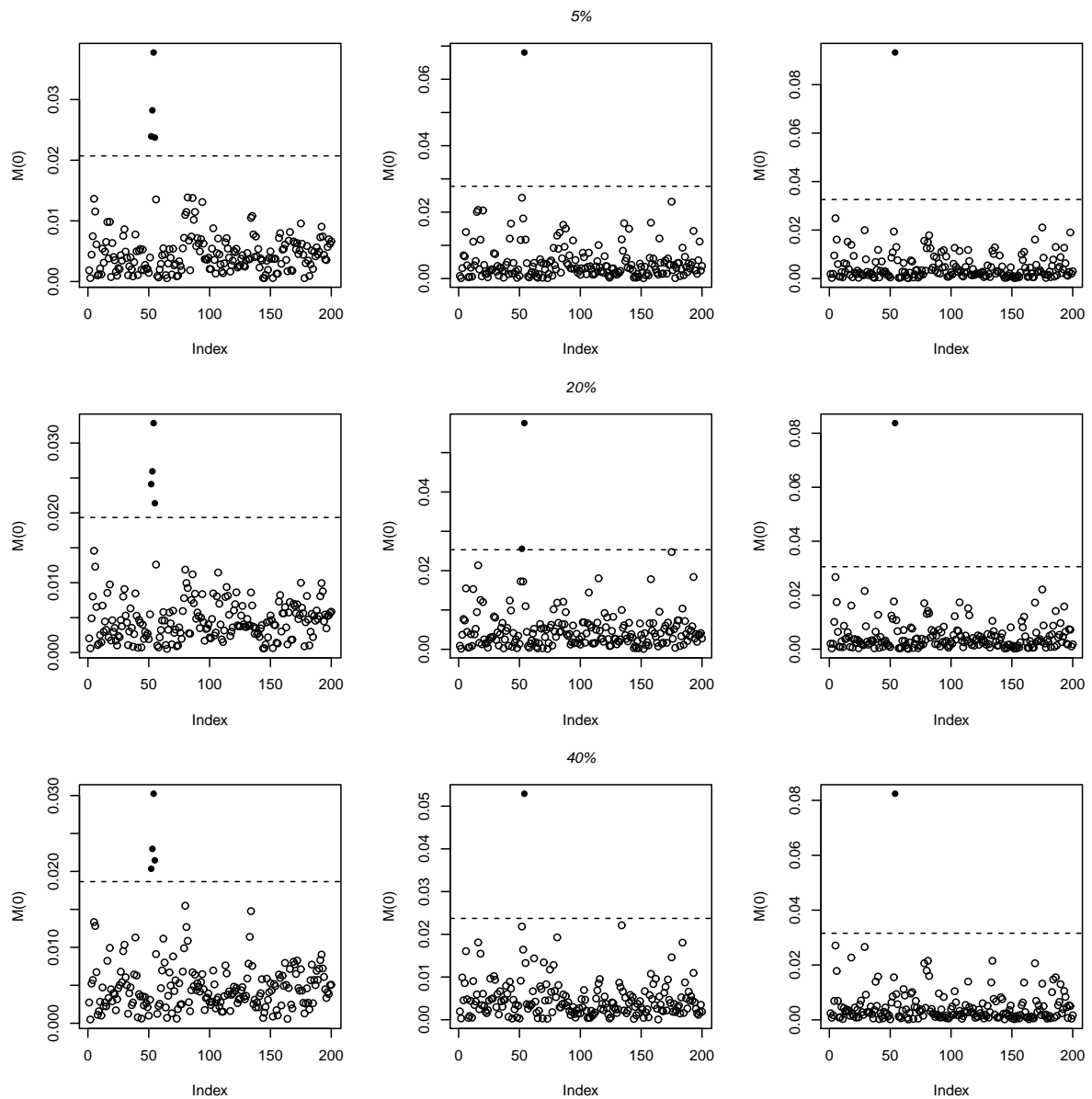
Figure 3.2: Index plot of $M(0)$ for response perturbation (left), scale matrix perturbation (middle) and explanatory variable perturbation (right); with 5% (top), 20% (middle) and 40% (bottom) of censored data.

y_{80} by $y_{80} + k * \text{sd}(\mathbf{y})$, $k = 3, 5$ and 7; and *(c)* substituting the 80th covariate observation x_{80} by $x_{80} + k * \text{sd}(\mathbf{x})$, $k = 5, 7$ and 10.

Following the approach described in Section 3.1 and considering an observation as influential if the respective $M(0)$ is greater than the Lee and Xu (2004) benchmark, given in (3.1.4), computed for $c^* = 3.5$, we counted the number of times that the perturbed observation was identified as influential and the mean number of influential observations identified, for all samples described above.

The obtained results are presented in Tables 3.1 to 3.3. In general, the capability

of the methodology to detect the influential points seems to be reasonable, especially when the percentage of censoring is low, although the capability under the scale matrix perturbation does not seem to decrease as the censoring level increases.

It is worth mentioning that, for the response perturbation scheme, the mean number of observations classified as influential generally increases as the magnitude of the atypical point increases, while it decreases for the other schemes considered.

Table 3.1: Percentage of times that the observation $\max(\mathbf{y}) = \max(\mathbf{y}) + k * \mathrm{sd}(\mathbf{y})$ was identified as an influential observation and mean number of influential observations identified under response perturbation ($\mathbf{y}$), scale matrix perturbation (Σ) and explanatory variable perturbation ($\mathbf{x}$).

		5%			20%			40%	
k	$\mathbf{y}$	Σ	$\mathbf{x}$	$\mathbf{y}$	Σ	$\mathbf{x}$	$\mathbf{y}$	Σ	$\mathbf{x}$
				% of correct identification					
1	61	65	66	57	76	64	54	83	62
2	99	98	99	98	98	99	94	98	96
3	100	100	100	100	100	100	100	100	100
5	100	100	100	100	100	100	100	100	100
			mean number of observations classified as influential						
1	2.53	3.12	2.69	2.46	2.77	2.7	2.37	2.64	2.87
2	2.64	2.76	2.16	2.73	2.49	2.35	2.71	2.37	2.55
3	2.95	1.97	1.66	2.79	1.91	1.8	2.76	2.03	1.96
5	3.27	1.16	1.39	3.2	1.32	1.47	2.99	1.48	1.53

(% of censoring spans the three groups 5%, 20%, 40%)

Table 3.2: Percentage of times that the 80th observation was identified as an influential observation when $y_{80} = y_{80} + k * \mathrm{sd}(\mathbf{y})$ and mean number of influential observations identified under response perturbation ($\mathbf{y}$), scale matrix perturbation (Σ) and explanatory variable perturbation ($\mathbf{x}$).

		5%			20%			40%	
k	$\mathbf{y}$	Σ	$\mathbf{x}$	$\mathbf{y}$	Σ	$\mathbf{x}$	$\mathbf{y}$	Σ	$\mathbf{x}$
				% of correct identification					
3	88	59	92	64	51	68	39	34	42
5	92	92	92	68	68	68	42	42	42
7	92	92	92	68	68	68	42	42	42
			mean number of observations classified as influential						
3	2.85	3.17	2.28	2.55	2.85	2.37	2.43	2.71	2.68
5	3.21	2.85	1.81	2.84	2.75	2.01	2.51	2.53	2.48
7	3.41	2.02	1.44	3.01	2.33	1.83	2.75	2.46	2.35

Table 3.3: Percentage of times that the 80th observation was identified as an influential observation when $x_{80} = x_{80} + k * \text{sd}(\mathbf{x})$ and the average number of influential observations identified under response perturbation ($\mathbf{y}$), scale matrix perturbation ($\mathbf{\Sigma}$) and explanatory variable perturbation ($\mathbf{x}$).

	% of censoring								
	5%			20%			40%		
k	$\mathbf{y}$	$\mathbf{\Sigma}$	$\mathbf{x}$	$\mathbf{y}$	$\mathbf{\Sigma}$	$\mathbf{x}$	$\mathbf{y}$	$\mathbf{\Sigma}$	$\mathbf{x}$
	% of correct identification								
5	37	59	59	38	67	58	28	60	50
7	87	90	91	92	93	94	95	95	95
10	100	100	100	100	100	100	100	100	100
	mean number of observations classified as influential								
5	2.28	2.53	2.72	2.37	2.49	2.8	2.28	2.47	2.87
7	2.55	1.87	2.39	2.44	1.72	2.38	2.54	1.83	2.55
10	1.94	1.23	1.76	1.95	1.11	1.72	1.75	1.1	1.83

3.4 Application on a real dataset: Total phosphorus concentration dataset

Here we perform diagnostic analysis based on the data presented in Subsection 1.1.2 and on the model discussed in Subsection 2.6.2. Following the approach described in Section 3.1, Figure 3.3 shows the index plots of $M(0)$ for the response perturbation, scale matrix perturbation and explanatory variable perturbation (for $\log(Q)$), respectively, along with the Lee and Xu (2004) benchmark, given in (3.1.4), computed for $c^* = 3.5$.

For the considered benchmark, none of the observations was identified as influential under the response perturbation, observations #5, #62 and #66 were identified as influential under the scale matrix perturbation and observations #5, #19 and #138 were identified as influential under the explanatory variable perturbation.

Figure 3.4 presents the log-transformed phosphorus concentration data with the imputed values, estimated using $\widehat{\log(\mathbf{P})} = E\{\log(\mathbf{P})|\mathbf{V}, \mathbf{C}, \widehat{\boldsymbol{\theta}}\}$ obtained from the SAEM algorithm, with the detected influential points emphasized. In general, the detected observations seem to be isolated peaks (see observation #5, for example).

In order to reveal the impact of the detected influential observations (#5, #19, #62, #66 and #138) on the parameter estimates, we refitted the model individually considering each of these five cases as missing observation. In Table 3.4 we show the relative changes (in percentage) of each parameter estimate, defined by

$$RC_\gamma = \left| \frac{\widehat{\gamma} - \widehat{\gamma}_{[t]}}{\widehat{\gamma}} \right|,$$

Figure 3.3: Index plot of $M(0)$ for response perturbation *(a)*, scale matrix perturbation *(b)* and explanatory variable perturbation *(c)* for the total phosphorus concentration dataset.
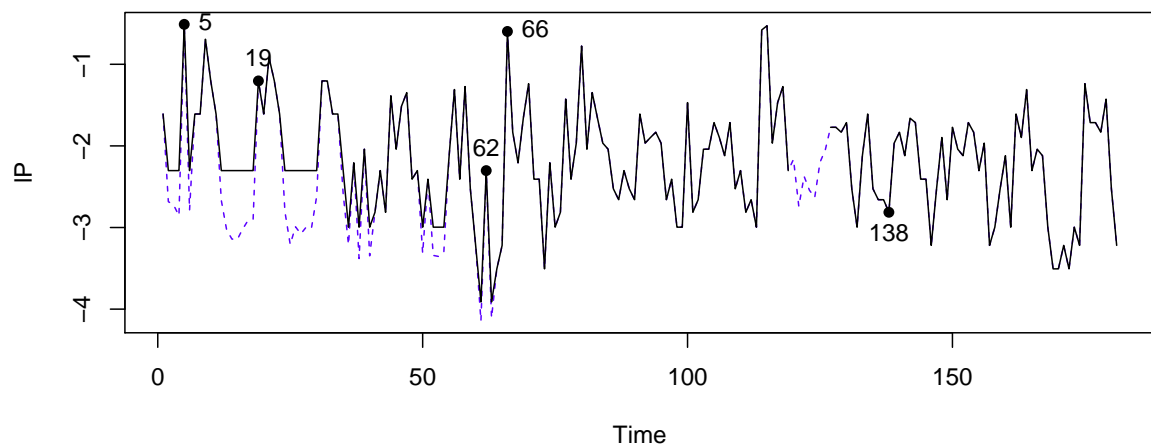
Figure 3.4: Log-transformed phosphorus concentration (lP) with the augmented series based on the fitted AR(1)-CR model (dashed blue line) and detected influential points.

where $\gamma = \beta_{k,l}(k = 0, 1; l = 1, 2, 3, 4), \sigma^2$ or ϕ_1, and $\widehat{\gamma}_{[t]}$ denotes the ML estimate of γ after the tth observation of $\log(P)$ is considered a missing value. It can be noted that observation #19 has a considerable impact on the estimation of the location parameters, while observation #62 has a considerable impact on the estimation of ϕ_1.

Table 3.4: Relative changes (RC in %) for the total phosphorus concentration dataset.

	Considered as missing				
	[#5]	[#19]	[#62]	[#66]	[#138]
$RC_{\beta_{1,1}}$	3.07	0.02	0.24	3.34	7.98
$RC_{\beta_{1,2}}$	1.84	16.39	2.70	2.13	1.70
$RC_{\beta_{1,3}}$	0.15	0.07	0.34	0.02	0.26
$RC_{\beta_{1,4}}$	0.07	0.03	2.47	0.03	0.30
$RC_{\beta_{2,1}}$	9.07	0.03	0.64	9.44	19.77
$RC_{\beta_{2,2}}$	4.47	38.01	6.37	5.01	4.00
$RC_{\beta_{2,3}}$	0.35	0.11	0.66	0.01	0.46
$RC_{\beta_{2,4}}$	0.46	0.02	4.86	0.28	0.62
RC_{σ^2}	8.20	2.09	1.03	6.64	4.19
RC_{ϕ_1}	26.52	7.62	37.06	8.80	18.04

In the interest of identifying influential observations on the regression parameters (β) and on the parameters from the covariance matrix (σ^2 and ϕ_1) separately, Figure 3.5 shows the plot of $M_\beta(0)$ versus $M_{\sigma^2,\phi_1}(0)$ for the three perturbation schemes. In this approach, observations #5 and #114 are detected as influential for all perturbation schemes considered, and observation #5 is the only one that seems to perturb the estimation of both subsets of parameters simultaneously. Besides, observations #19 and #138 seem to impact only the estimation of β, while observations #62 and #66 seem to impact the estimation of σ^2 and ϕ_1.
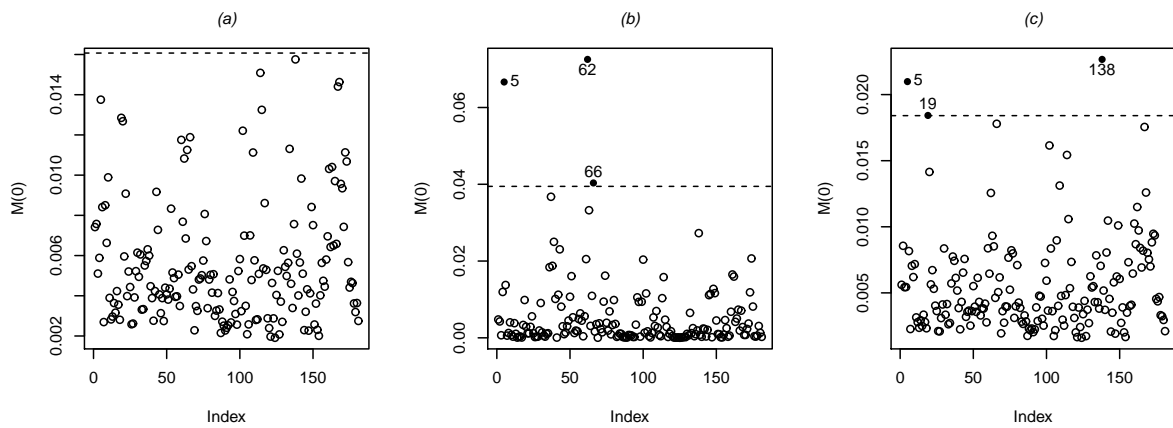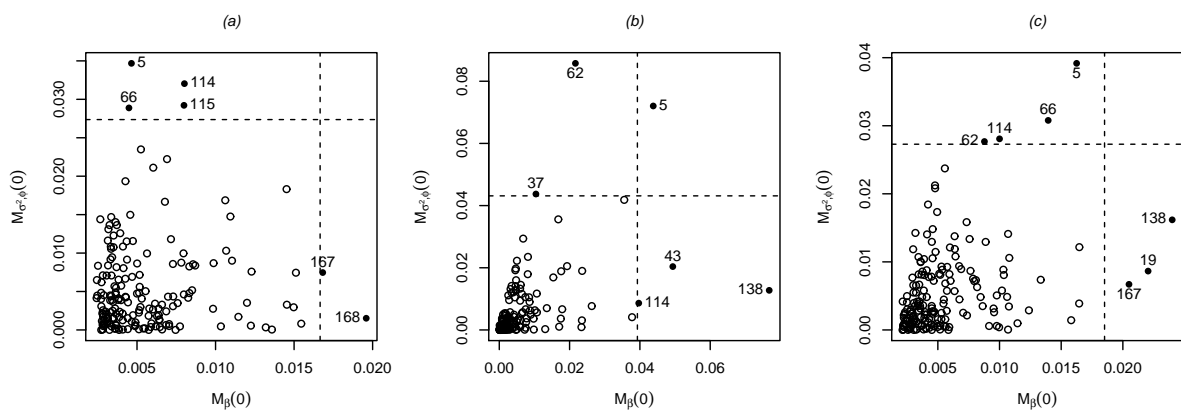

Figure 3.5: Plot of $M_\beta(0)$ versus $M_{\sigma^2,\phi_1}(0)$ for response perturbation *(a)*, scale matrix perturbation *(b)* and explanatory variable perturbation *(c)* for the total phosphorus concentration dataset.

3.5 Conclusions

In this chapter we proposed influence diagnostic tools for detecting influential observations on the context of censored linear models with autoregressive errors. The diagnostic analysis was based on local influence techniques suggested by Zhu and Lee (2001). Using the discussed method, we analyzed a real dataset and carried out extensive simulation studies. We observed that the capability of the methodology in detecting influential points seems to be reasonable. Besides, among the three perturbation schemes considered, the response perturbation seemed to be more affected by the dependence structure of the data.

Chapter 4

Concluding remarks

In this work, we developed a full likelihood approach for censored linear regression models with autoregressive erros, denoted by AR(p)-CR models. We presented the implementation of the SAEM algorithm for maximum likelihood estimation, where the likelihood function, predictions of unobservable response values and the asymptotic standard errors are obtained as byproducts. Next, we developed the diagnostic measures for assessing local influence of these models.

Even though some solutions have been proposed in the literature to deal with the problem of censored responses in AR models, to the best of our knowledge, we consider that this work is the first attempt at exact ML estimation in the context of censored AR models. In order to examine the performance of our proposed methods, we presented various simulation studies and we illustrated the methods through the analysis of two real datasets. The methods developed are implemented in the R package *ARCensReg*, providing practitioners with a convenient tool for further application in their domain.

In addition, following Shi and Huang (2011), we implemented a stepwise local influence method, which intends to deal with masking effects by choosing a perturbation scheme that perturbs only a subset of points (excluding some highly influential observations detected in a previous step from the perturbation scheme), and allowing points that were masked to be uncovered in the new perturbation scheme. However, for appropriate benchmark values, the obtained results were very similar to those obtained by the method described in Section 3.1, and thereby this comparison was not presented in this dissertation.

4.1 Future research

There is a large number of possible extensions of the results obtained in this dissertation, such as:

- The use of scale mixtures of normal distributions to accommodate heavy-tailed features (Lachos et al., 2011).

- To extend the proposed methods to accommodate missing values in addition to censoring using hybrid Bayesian sampling procedures (Wang and Fan, 2012).

- To extend the proposed methods to accommodate multivariate outcomes (Wang et al., 2015).

- To consider other correlation structures, such as ARMA models and space–time correlation structures (Cesare et al., 2001).

4.2 Technical production

In this section, we describe the technical production as result of this dissertation.

4.2.1 Papers

The results from Chapter 2 originated a paper entitled "*Censored regression models with autoregressive errors: A likelihood-based perspective*" that was submitted to the journal Econometrics and Statistics.

The results presented in Chapter 3 are been organized into a paper entitled "*Influence diagnostics for censored regression models with autoregressive errors*" that will be submitted to a statistical journal.

4.2.2 R package

ARCensReg: *Fitting Univariate Censored Linear Regression Model with Autoregressive Errors*

It fits an univariate left or right censored linear regression model with autoregressive errors under the normal distribution. It provides estimates and standard errors of the parameters, prediction of future observations and it supports missing values on the dependent variable. It also performs influence diagnostic through local influence for three possible perturbation schemes. It is available to download for free in the website `https://cran.r-project.org/web/packages/ARCensReg/` and its two main functions are described below.

• *ARCensReg:* Censored Linear Regression Model with Autoregressive Errors

Description

It fits an univariate left or right censored linear regression model with autoregressive errors under the normal distribution using the SAEM algorithm. It provides estimates and standard errors of the parameters, prediction of future observations and it supports missing values on the dependent variable. It also provides convergence plots when exists at least one censored observation.

Usage

R code

```
ARCensReg(cc,y,x,p=1,cens='left',x_pred=NULL,miss=NULL,
  tol=0.0001,show.convergence=TRUE,M=10,perc=0.25,MaxIter=400,
  pc=0.18,show_se=TRUE)
```

Arguments

cc	Vector of censoring indicators of length n, where n is the total of observations. For each observation: 0 if non-censored, 1 if censored.
x	Matrix of covariates of dimension n x l, where l is the number of fixed effects including the intercept, if considered (in models which include an intercept x should contain a column of ones).
y	Vector of responses of length n.
p	Order of the autoregressive process. Must be a positive integer value. For p equal to 0 we suggest to use the function CensReg.SMN from SMNCensReg package.
cens	"left" for left censoring, "right" for right censoring.
x_pred	Matrix of covariates for responses to be predicted. If x_pred = NULL no responses are predicted.
miss	Vector containing the index of missing observations. miss = NULL indicates that no observations are missing.
tol	The convergence maximum error permitted.

```
show.convergence
```
TRUE or FALSE. Indicates if convergence graphs should be built for the parameters estimates (for the case with at least one censored observation). The dashed line indicates the iteration of the SAEM algorithm that simulations start being smoothed. Default=TRUE.

`M`   Size of the Monte Carlo sample generated in each step of the SAEM algorithm. Default=10.

`perc`   Percentage of burn-in on the Monte Carlo sample. Default=0.25.

`MaxIter`   The maximum number of iterations of the SAEM algorithm. Default=400.

`pc`   Percentage of initial iterations of the SAEM algorithm. It is recommended that 50<MaxIter*pc<100. Default=0.18.

`show_se`   TRUE or FALSE. Indicates if the standard errors should be estimated. Default=TRUE.

Details

The initial values are obtained by ignoring censoring and applying maximum likelihood estimation with the censored data simply replaced by their censoring limits. If you want to fit a regression model with autoregressive errors for non-censored data, just set "cc" as a vector of zeros and "cens" as either "right" or "left".

Value

`beta`   Estimate of the regression parameters.

`sigma2`   Estimated variance of the white noise process.

`phi`   Estimate of the autoregressive parameters.

`pi1`   Estimate of the first p partial autocorrelations.

`theta`   Vector of parameters estimate (beta, sigma2, phi).

`SE`   Vector of the standard errors of (beta, sigma2, phi).

`loglik`   Log-likelihood value.

`AIC`   Akaike information criterion.

`BIC`   Bayesian information criterion.

`AICcorr`   Corrected Akaike information criterion.

`time`   Processing time.

`pred`   Predicted values (if x_pred is not `NULL`).

`criteria`   Attained criteria value.

yest	Augmented response variable based on the fitted model.
yyest	Final estimative of $E(Y\%*\%t(Y))$.
iter	Number of iterations until convergence.

Examples

<div align="center">R code</div>

```
##simple example (p = l = 1)
#generating a sample
set.seed(23451)
n=50
x=rep(1,n)
dat = rARCens(n=n,beta=2,pit=.5,sig2=.3,x=x,
                cens='left',pcens=.1)


#fitting the model (quick convergence)
fit0 = ARCensReg(dat$data$cc,dat$data$y,x,tol=0.001,
                  pc=.12,M=5,show_se=FALSE)


##another example (p = l = 2)
#generating a sample
n=100
x=cbind(1,runif(n))
dat = rARCens(n=n,beta=c(2,1),pit=c(.4,-.2),sig2=.5,
                  x=x,cens='left',pcens=.05)
#fitting the model
fit1 = ARCensReg(dat$data$cc,dat$data$y,x,p=2,
                    cens="left",tol=0.0001)
#plotting the augmented variable
par(mfrow=c(1,1))
plot.ts(fit1$yest,lty='dashed',col=4)
lines(dat$data$y)
#simulating missing values
miss = sample(1:n,3)
yMISS = dat$data$y
yMISS[miss] = NA
fit2 = ARCensReg(dat$data$cc,yMISS,x,p=2,miss=miss,
                cens="left",tol=0.0001)
```

- *InfDiag:* Influence Diagnostic in Censored Linear Regression Model with Autoregressive Errors

Description

It performs influence diagnostic by a local influence approach (Cook, 1986) with three possible perturbations schemes: response perturbation (y), scale matrix perturbation (Sigma) or explanatory variable perturbation (x). A benchmark value is calculated that depends on k.

Usage

<div align="center">R code</div>

```
InfDiag(theta,yest,yyest,x,k=3,plots=T,indpar=rep(1,length(theta)),
                perturbation ='y',indcolx = rep(1,ncol(x)))
```

Arguments

theta	Vector of estimated parameters.
yest	Vector of responses of length n with agmented data. Should be the value yest of the ARCensReg function in the case that at least one observation is censored.
yyest	Should be the value yyest of the ARCensReg function in the case that at least one observation is censored. Otherwise, must be y%*%t(y).
x	Matrix of covariates of dimension n x l, where l is the number of fixed effects including the intercept, if considered (in models which include an intercept x should contain a column of ones).
k	Constant to be used in the benchmark calculation: M0+k*sd(M0).
plots	TRUE or FALSE. Indicates if a graph should be plotted.
indpar	Vector of length equal to the number of parameters, with each element 0 or 1 indicating if the respective parameter should be taking into account in the influence calculation.
perturbation	Perturbation scheme. Possible values: "y" for response perturbation, "Sigma" for scale matrix perturbation or "x" for explanatory variable perturbation.

indcolx   If perturbation="x", indcolx must be a vector of length equal
to the number of columns of x, with each element 0 or 1 indicating
if the respective column of x should be perturbed. All columns are
perturbed by default.

Details

The function returns a vector of length n with the aggregated contribution (M0) of all
eigenvectors of the matrix associated with the normal curvature.

Value

M0

Examples

R code

```
#generating the data
set.seed(12341)
x = cbind(1,runif(100))
dat = rARCens(n=100,beta = c(1,-1),pit = c(.4,-.2),sig2=.5,
            x=x,cens='left',pcens=.05)
#creating an outlier
dat$data$y[40] = 5
plot.ts(dat$data$y)

#fitting the model
fit = ARCensReg(cc=dat$data$cc,y=dat$data$y,x,p=2,cens='left',
      tol=0.001,show_se=F)

#influence diagnostic
M0y = InfDiag(theta=fit$res$theta, yest=fit$yest, yyest=fit$yyest,
        x=x, k = 3.5, perturbation = "y")
M0Sigma = InfDiag(theta=fit$res$theta, yest=fit$yest, yyest=fit$yyest,
        x=x, k = 3.5, perturbation = "Sigma")
M0x = InfDiag(theta=fit$res$theta, yest=fit$yest, yyest=fit$yyest,
        x=x, k = 3.5, perturbation = "x",indcolx =c(0,1))

#perturbation on a subset of parameters
M0y1 = InfDiag(theta=fit$res$theta, yest=fit$yest, yyest=fit$yyest,
```

```
          x=x, k = 3.5, perturbation = "y",indpar=c(1,1,0,0,0))
M0y2 = InfDiag(theta=fit$res$theta, yest=fit$yest, yyest=fit$yyest,
          x=x, k = 3.5, perturbation = "y",indpar=c(0,0,1,1,1))
plot(M0y1,M0y2)
abline(v = mean(M0y1)+3.5*sd(M0y1),h = mean(M0y2)+3.5*sd(M0y2),lty=2)
```

Bibliography

Arismendi, J. C. (2013, May). Multivariate truncated moments. *J. Multivar. Anal. 117*, 41–75.

Barndorff-Nielsen, O. and G. Schou (1973). On the parametrization of autoregressive models by partial autocorrelations. *Journal of Multivariate Analysis 3*(4), 408 – 419.

Box, G., G. Jenkins, and G. Reinsel (1994). *Time Series Analysis: Forecasting and Control.* Wiley Series in Probability and Statistics. Wiley.

Cesare, L. D., D. Myers, and D. Posa (2001). Estimating and modeling space-time correlation structures. *Statistics & Probability Letters 51*(1), 9 – 14.

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society, Series B 48*, 133–169.

Couvreur, C. (1996). The EM algorithm: a guided tour. In *Proceedings of the 2d IIEE European Workshop on Computionaly Intensive Methos in Control and Signal Processing*.

De Bastiani, F., A. H. Mariz de Aquino Cysneiros, M. A. Uribe-Opazo, and M. Galea (2014). Influence diagnostics in elliptical spatial linear models. *TEST 24*(2), 322–340.

Delyon, B., M. Lavielle, and E. Moulines (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics 27*(1), 94–128.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B 39*, 1–38.

Galarza, C. E., D. Bandyopadhyay, and V. H. Lachos (2015). Quantile regression in linear mixed models: a stochastic approximation EM approach. *Statistics and its Interface*, In press.

Garay, A. M., V. H. Lachos, H. Bolfarine, and C. R. B. Cabral (2015). Linear censored regression models with scale mixtures of normal distributions. *Statistical Papers*, 1–32.

Genz, A., F. Bretz, T. Hothorn, T. Miwa, X. Mi, F. Leisch, and F. Scheipl (2008). mvtnorm: Multivariate Normal and t Distribution. *R package version 0.9-2, URL http://CRAN. R-project. org/package= mvtnorm*.

Graham, A. (1981). *Kronecker Products and Matrix Calculus: With Applications*. Mathematics and Its Applications. Ellis Horwood Ltd , Publisher.

Hopke, P. K., C. Liu, and D. B. Rubin (2001). Multiple imputation for multivariate data with missing and below-threshold measurements: Time-series concentrations of pollutants in the Arctic. *Biometrics 57*(1), 22–33.

Kim, S.-K. and R. Huggins (1998). Diagnostics for autocorrelated regression models. *Australian & New Zealand Journal of Statistics 40*(1), 65–71.

Kuhn, E. and M. Lavielle (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: Probability and Statistics 8*, 115–131.

Kuhn, E. and M. Lavielle (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis 49*(4), 1020–1038.

Lachos, V., D. Bandyopadhyay, and D. Dey (2011). Linear and nonlinear mixed-effects models for censored HIV viral loads using normal/independent distributions. *Biometrics 67*(4), 1594–1604.

Lee, S. Y. and L. Xu (2004). Influence analysis of nonlinear mixed-effects models. *Computational Statistics & Data Analysis 45*, 321–341.

Matos, L. A., V. H. Lachos, N. Balakrishnan, and F. V. Labra (2013). Influence diagnostics in linear and nonlinear mixed-effects models with censored data. *Computational Statistics & Data Analysis 57*, 450–464.

McLeod, A. I. and Y. Zhang (2006). Partial autocorrelation parameterization for subset autoregression. *Journal of Time Series Analysis 27*(4), 599–612.

Meza, C., F. Osorio, and R. De la Cruz (2012). Estimation in nonlinear mixed-effects models using heavy-tailed distributions. *Statistics and Computing 22*, 121–139.

Ortega, E. M. M., H. Bolfarine, and G. A. Paula (2003, February). Influence diagnostics in generalized log-gamma regression models. *Computational Statistics & Data Analysis 42*(1-2), 165–186.

Osorio, F., G. A. Paula, and M. Galea (2007). Assessment of local influence in elliptical linear models with longitudinal structure. *Computational Statistics & Data Analysis 51*, 4354–4368.

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

Poon, W. Y. and Y. S. Poon (1999). Conformal normal curvature and assessment of local influence. *Journal of the Royal Statistical Society, Series B 61*, 51–61.

R Development Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rao, C. R. et al. (1987). Prediction of future observations in growth curve models. *Statistical Science 2*(4), 434–447.

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*, 167–182.

Schall, R. and T. T. Dunne (1991). *Diagnostics for Regression-Arma Time Series*, pp. 205–221. New York, NY: Springer New York.

Schilling, K., K. Chan, H. Liu, and Y. Zhang (2010, 6). Quantifying the effect of land use land cover change on increasing discharge in the upper mississippi river. *Journal of Hydrology 387*(3-4), 343–345.

Schumacher, F., V. Lachos, and C. Galarza (2016). *ARCensReg: Fitting Univariate Censored Linear Regression Model with Autoregressive Errors*.

Shi, L. and M. Huang (2011, February). Stepwise local influence analysis. *Computational Statistics & Data Analysis 55*(2), 973–982.

Vaida, F. (2005). Parameter convergence for EM and MM algorithms. *Statistica Sinica 15*(3), 831–840.

Vaida, F. and L. Liu (2009). Fast implementation for normal mixed effects models with censored response. *Journal of Computational and Graphical Statistics 18*(4), 797–817.

Wang, C. and K. Chan (2016a). Quasi-likelihood estimation of a censored autoregressive model with exogenous variables.

Wang, C. and K.-S. Chan (2016b). *carx: Censored Autoregressive Model with Exogenous Covariates*.

Wang, W. (2013). Multivariate t linear mixed models for irregularly observed multiple repeated measures with missing outcomes. *Biometrical Journal 55*(4), 554–571.

Wang, W.-L. and T.-H. Fan (2012). Bayesian analysis of multivariate t linear mixed models using a combination of IBF and Gibbs samplers. *Journal of Multivariate Analysis 105*(1), 300–310.

Wang, W.-L., T.-I. Lin, and V. H. Lachos (2015). Extending multivariate-t linear mixed models for multiple longitudinal data with censored responses and heavy tails. *Statistical Methods in Medical Research DOI:0962280215620229*, 1–20.

Wei, G. C. and M. A. Tanner (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association 85*(441), 699–704.

Wilhelm, S. and M. B. G. (2015). tmvtnorm: Truncated Multivariate Normal and Student t Distribution. *R package version 1.4-10, URL https://cran.r-project.org/web/packages/tmvtnorm/*.

Wu, C. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics 11*(1), 95–103.

Zeger, S. L. and R. Brookmeyer (1986). Regression analysis with censored autocorrelated data. *Journal of the American Statistical Association 81*(395), 722–729.

Zhu, H., J. G. Ibrahim, S. Lee, and H. Zhang (2007). Perturbation selection and influence measures in local influence analysis. *The Annals of Statistics 35*(6), 2565–2588.

Zhu, H. and S. Lee (2001). Local influence for incomplete-data models. *Journal of the Royal Statistical Society, Series B 63*, 111–126.

Appendix A

Additional results of Chapter 2 and Chapter 3

A.1 Complementary results of the simulation study from Subsection 2.5.1

Table A.1: Mean square errors of the parameter estimates from the Simulation presented in Subsection 2.5.1 under 5%, 20% and 40% censoring proportions (CP) and different samples sizes (n).

CP	n	β_0	β_1	σ^2	ϕ_1	ϕ_2
	50	0.182	0.4313	0.1436	0.0225	0.0222
	100	0.0735	0.1986	0.0974	0.0085	0.0095
5%	200	0.0436	0.0889	0.0458	0.0048	0.0059
	300	0.035	0.0767	0.0228	0.0029	0.0023
	500	0.0168	0.0377	0.0233	0.0017	0.0016
	50	0.173	0.4378	0.2452	0.0266	0.0227
	100	0.0878	0.2172	0.1039	0.0157	0.0127
20%	200	0.0447	0.0967	0.0483	0.0047	0.0053
	300	0.0284	0.0623	0.0336	0.0035	0.004
	500	0.0177	0.0403	0.0217	0.0024	0.0022
	50	0.2577	0.4916	0.3784	0.0343	0.0307
	100	0.1115	0.2408	0.149	0.0118	0.012
40%	200	0.0692	0.1265	0.0735	0.0066	0.0066
	300	0.0308	0.0797	0.0632	0.0047	0.0039
	500	0.0196	0.0357	0.0357	0.0022	0.0022

Table A.2: Mean absolute errors of the parameter estimates from the Simulation presented in Subsection 2.5.1 under 5%, 20% and 40% censoring proportions (CP) and different samples sizes (n).

CP	n	β_0	β_1	σ^2	ϕ_1	ϕ_2
	50	0.3443	0.5194	0.317	0.1197	0.1246
	100	0.2158	0.3647	0.2535	0.074	0.0772
5%	200	0.1651	0.2256	0.1752	0.0557	0.0604
	300	0.1534	0.2234	0.1207	0.043	0.0396
	500	0.1015	0.1551	0.1254	0.0343	0.0335
	50	0.333	0.5245	0.4117	0.1324	0.1195
	100	0.2343	0.3755	0.2664	0.1001	0.0915
20%	200	0.1673	0.2489	0.1793	0.0554	0.0579
	300	0.1357	0.2024	0.1489	0.0456	0.0512
	500	0.1095	0.1556	0.118	0.0398	0.0364
	50	0.3997	0.5745	0.5273	0.1499	0.1447
	100	0.2749	0.4153	0.3112	0.0883	0.0873
40%	200	0.2145	0.2884	0.2288	0.0619	0.0647
	300	0.1402	0.2283	0.2089	0.0561	0.0508
	500	0.1077	0.1543	0.1533	0.0388	0.0381

A.2 Matrix algebra results

We present some useful matrix algebra results, which may be helpful in understanding some of the results and derivations presented in this work.

Let $\mathbf{X}$ be a random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ of dimension $n \times 1$ and let $\mathbf{A}$ be a non-stochastic $n \times n$ matrix, then

$$
\begin{aligned}
E\left[\mathbf{X}^\top \mathbf{A} \mathbf{X}\right] &= \operatorname{tr}\left(E\left[\mathbf{X}^\top \mathbf{A} \mathbf{X}\right]\right) = E\left[\operatorname{tr}\left(\mathbf{X}^\top \mathbf{A} \mathbf{X}\right)\right] \\
&= E\left[\operatorname{tr}\left(\mathbf{A} \mathbf{X} \mathbf{X}^\top\right)\right] = \operatorname{tr}\left(\mathbf{A} E\left[\mathbf{X} \mathbf{X}^\top\right]\right) \\
&= \operatorname{tr}\left(\mathbf{A}\left(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top\right)\right) = \operatorname{tr}(\mathbf{A}\,\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{\mu}.
\end{aligned} \tag{A.2.1}
$$

The proofs of the next results can be found in Graham (1981). Let $\mathbf{x}$ be a vector of

dimension $n \times 1$ and let $\mathbf{A}$ be a constant matrix of dimension $n \times n$, then

$$\frac{\partial \mathbf{x}^\top \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{A}^\top \mathbf{x}, \tag{A.2.2}$$

$$\frac{\partial \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}^\top, \tag{A.2.3}$$

$$\frac{\partial \mathbf{x}^\top \mathbf{A}}{\partial \mathbf{x}} = \mathbf{A}. \tag{A.2.4}$$

Now, let $\mathbf{A}$ be a symmetric positive definite matrix of dimension $n \times n$ and let x be a scalar, then

$$\frac{\partial \mathbf{A}^{-1}}{\partial x} = -\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial x} \mathbf{A}^{-1}, \tag{A.2.5}$$

$$\frac{\partial \mathrm{tr}(\mathbf{A})}{\partial x} = \mathrm{tr}\left(\frac{\partial \mathbf{A}}{\partial x}\right), \tag{A.2.6}$$

$$\frac{\partial |\mathbf{A}|}{\partial x} = |\mathbf{A}| \mathrm{tr}\left(\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial x}\right), \tag{A.2.7}$$

$$\frac{\partial \log |\mathbf{A}|}{\partial x} = \mathrm{tr}\left(\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial x}\right). \tag{A.2.8}$$

Finally, let $\mathbf{A}$ and $\mathbf{B}$ be matrices of dimension $n \times n$ and let x be a scalar, then

$$\frac{\partial (\mathbf{A}\mathbf{B})}{\partial x} = \left(\frac{\partial \mathbf{A}}{\partial x}\right) \mathbf{B} + \mathbf{A} \left(\frac{\partial \mathbf{B}}{\partial x}\right). \tag{A.2.9}$$