# Censored regression models with autoregressive errors: A likelihood-based perspective

Fernanda L. Schumacher[a], Victor H. Lachos[a] * and Dipak K. Dey[b]

[a] *Departamento de Estatística, Universidade Estadual de Campinas, Brazil*

[b] *Department of Statistics, University of Connecticut, USA*

**Abstract**

In many studies that involve time series variables, limited or censored data are naturally collected. This occurs, in several practical situations, for reasons such as limitations of measuring equipment or from experimental design. Hence, the exact true value is recorded only if it falls within an interval range, so the responses can be either left, interval or right censored. 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 paper, we propose an analytically tractable and efficient stochastic approximation of the EM (SAEM) algorithm to obtain the maximum likelihood estimates of the parameter of censored regression models with autoregressive errors of order $p$. This approach permits easy and fast estimation of the parameters of autoregressive models when censoring is present and as a byproduct, enables predictions of unobservable values of the response variable. The observed information matrix is derived analytically to account for standard errors. We use simulations to investigate the asymptotic properties of the SAEM estimates and prediction accuracy. In this simulation study comparisons are also made between inferences based on the censored data and those based on complete data obtained by crude/ad-hoc imputation methods. Finally, the method is illustrated using a meteorological time series dataset on cloud ceiling height, where the measurements are subject to the detection limit of the recording device. The proposed algorithm and methods are implemented in the new R package *ARCensReg*.

**Keywords** Autoregressive AR(p) models; Censored data; Limit of detection; SAEM algorithm.

## 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

*Corresponding author. Address for correspondence: Departamento de Estatística, Rua Sérgio Buarque de Holanda, 651, Cidade Universitária Zeferino Vaz, Campinas, São Paulo, Brazil. CEP 13083-859. e-mail: hlachos@ime.unicamp.br

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 limit. 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 & 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. 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. Note that 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 paper, our aim is 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. Our empirical results show that the ML estimates based on our proposed SAEM algorithm provide good asymptotic properties. Moreover, the proposed algorithm and methods are implemented in the R package *ARCensReg* (Schumacher *et al.*, 2016; R Development Core Team, 2015).

The rest of the paper is organized as follows. Section 2 gives a brief description of the AR(p) regression (AR(p)-LR) model, including the description of a procedure for ML estimation. Section 3 proposes the AR(p)-CR model and shows how to get the ML estimates through the SAEM algorithm. Section 4 gives a brief sketch of prediction in AR(p)-CR models. Furthermore, the observed

information matrix is derived analytically to account for standard errors. The advantage of the proposed method is illustrated through empirical studies in Section 5. The method is illustrated in Section 6 with the analysis of a meteorological time series dataset on cloud ceiling height. Section 7 concludes with a short discussion of issues raised by our study and some possible directions for a future research.

# 2 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

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

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 variables values, $\varepsilon_t$ is the autoregressive error with Gaussian disturbance $\eta_t$ and $\boldsymbol{\phi} = (\phi_1,\ldots,\phi_p)^\top$ is the vector of autoregressive coefficients. The model defined in (1)-(2), will be denoted the AR(p)-LR model.

To ensure stationarity of the AR(p) model given in Equation (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 $\boldsymbol{\phi}$ such where the process is stationary will be called the admissible region.

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

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

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

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

where $\gamma_0,\ldots,\gamma_{n-1}$ are the theoretical autocovariances of the process and $\rho_k = \gamma_k/\gamma_0$, for $k = 1,2,\ldots$ are the theoretical autocorrelations of the process. Moreover, the coefficients $\phi_1,\ldots,\phi_p$ and $\rho_1,\ldots,\rho_n$ satisfy the Yule-Walker equations (Barndorff-Nielsen & Schou, 1973). To ensure the admissibility of $\boldsymbol{\phi}$ and stabilize the estimating procedure, we reparameterize $\boldsymbol{\phi}$ following Barndorff-Nielsen & Schou (1973), given by:

$$\begin{aligned} \phi_p^{(p)} &= \pi_p, \\ \phi_v^{(p)} &= \phi_v^{(p-1)} - \pi_p \phi_{p-v}^{(p-1)}, & (3) \end{aligned}$$

where $\phi_v^{(p)}$ is the $v$th AR parameter under the AR(p) model (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 $\mathscr{B} : (\pi_1, \ldots, \pi_p) \to (\phi_1, \ldots, \phi_p)$ that is one-to-one, continuous and differentiable inside the admissible region. This parameterization has the advantage that in the $\boldsymbol{\pi}$-space the admissible region is simply the $p$-dimensional cube with boundary surfaces corresponding to $\pm 1$, while in the $\boldsymbol{\phi}$-space it is very complicated (see for instance McLeod & 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), the exact log-likelihood function is given by:

$$\ell(\boldsymbol{\theta} \mid \mathbf{y}) = -\frac{1}{2}\left[n \log \sigma^2 + \frac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{M}_n^{-1}(\boldsymbol{\phi})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \log|\mathbf{M}_n(\boldsymbol{\phi})|\right] + C, \tag{4}$$

with $C$ being a constant independent of the parameter vector $\boldsymbol{\theta}$. Considering the reparameterization given in (3) and dropping constant terms (McLeod & Zhang, 2006), the log-likelihood function can be written as:

$$\ell(\boldsymbol{\theta}|\mathbf{y}) = -\frac{1}{2}\left[n \log \sigma^2 + \frac{1}{\sigma^2} S(\boldsymbol{\pi}, \boldsymbol{\beta}) + \log g_p\right], \tag{5}$$

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

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

where $D(\mathbf{y}, \boldsymbol{\beta})$ is 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:

$$D_{i,j} = D_{j,i} = (Y_i - \mathbf{x}_i^\top \boldsymbol{\beta})(Y_j - \mathbf{x}_j^\top \boldsymbol{\beta}) + \ldots + (Y_{n+1-j} - \mathbf{x}_{n+1-j}^\top \boldsymbol{\beta})(Y_{n+1-i} - \mathbf{x}_{n+1-i}^\top \boldsymbol{\beta}) \tag{6}$$

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

The unknown model parameters can be estimated by maximizing the corresponding log-likelihood function. Since $\frac{\partial \ell(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\pi}} = 0$ 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 (5) is maximized at:

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\pi}) = \left(\mathbf{X}^\top \mathbf{M}_n^{-1}(\mathscr{B}(\boldsymbol{\pi}))\mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{M}_n^{-1}(\mathscr{B}(\boldsymbol{\pi}))\,\mathbf{y}, \tag{7}$$

$$\widehat{\sigma^2}(\boldsymbol{\pi}) = \frac{1}{n} S\left(\boldsymbol{\pi}, \widehat{\boldsymbol{\beta}}\right). \tag{8}$$

Thus, substituting expressions (7)–(8) into the log-likelihood and dropping constant terms, we obtain a concentrated log-likelihood function:

$$\ell_c(\boldsymbol{\pi}|\mathbf{y}) = -\frac{n}{2} \log S\left(\boldsymbol{\pi}, \widehat{\boldsymbol{\beta}}\right) - \frac{1}{2} \log g_p, \tag{9}$$

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

These estimates can be obtained through the R function *arima()* (R Development Core Team, 2015), using the argument *xreg* to declare the regressor variables.

# 3 The censored autoregressive regression model of order $p$

The autoregressive model with exogenous variables is defined by:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{10}$$

where $\mathbf{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ are as defined in Subsection 2.1. Moreover, following Vaida & 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 $t$th time be $(V_t, C_t)$, where $V_t$ represents the uncensored reading or censoring level and $C_t$ is the censoring indicators such that:

$$Y_t \leq V_t \text{ if } C_t = 1 \text{ and } Y_t = V_t \text{ if } C_t = 0, \tag{11}$$

so that equation (11) along with model (10) define the censored autoregressive regression model of order $p$ (hereafter, AR(p)-CR model). 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 $-V_t$.

## 3.1 The log-likelihood function

Following Vaida & Liu (2009), classical inference on the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2, \boldsymbol{\phi}^\top)^\top$ is based on the marginal distribution of $\mathbf{y}$. For complete data, we have marginally that $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$. For responses with censoring pattern as in (11), we have $\mathbf{y} \sim TN_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}; \mathbb{A})$, 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 (10)-(11), 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 $\mathbf{y}^o \sim N_{n^o}(\mathbf{X}^o\boldsymbol{\beta}, \boldsymbol{\Sigma}^{oo})$ and $\mathbf{y}^c|\mathbf{y}^o \sim N_{n^c}(\boldsymbol{\mu}, \mathbf{S})$, where $\boldsymbol{\mu} = \mathbf{X}^c\boldsymbol{\beta} + \boldsymbol{\Sigma}^{co}(\boldsymbol{\Sigma}^{oo})^{-1}(\mathbf{y}^o - \mathbf{X}^o\boldsymbol{\beta})$ and $\mathbf{S} = \boldsymbol{\Sigma}^{cc} - \boldsymbol{\Sigma}^{co}(\boldsymbol{\Sigma}^{oo})^{-1}\boldsymbol{\Sigma}^{oc}$. 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 & Liu (2009) and Matos *et al.* (2013), the likelihood function for the observed data is thus given by (using conditional probability arguments):

$$\begin{aligned} L(\boldsymbol{\theta}) = f(\mathbf{y}|\boldsymbol{\theta}) &= P(\mathbf{V}|\mathbf{C}, \boldsymbol{\theta}) = P(\mathbf{y}^c \leq \mathbf{V}^c|\mathbf{y}^o = \mathbf{V}^o, \boldsymbol{\theta})P(\mathbf{y}^o = \mathbf{V}^o|\boldsymbol{\theta}) \\ &= P(\mathbf{y}^c \leq \mathbf{V}^c|\mathbf{y}^o, \boldsymbol{\theta})f(\mathbf{y}^o|\boldsymbol{\theta}) \\ &= \phi_{n^o}(\mathbf{y}^o; \mathbf{X}^o\boldsymbol{\beta}, \boldsymbol{\Sigma}^{oo})\Phi_{n^c}(\mathbf{V}^c; \boldsymbol{\mu}, \mathbf{S}) = L, \end{aligned} \tag{12}$$

which can be evaluated without much computational burden through the routine *mvtnorm()* available in R; see (Genz *et al.*, 2008; R Development Core Team, 2015). The log-likelihood function $\ell(\boldsymbol{\theta}|\mathbf{y})$ is used to compute different model selection criteria, such as:

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

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

## 3.2 The EM and SAEM algorithm for ML estimation

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 is given by:

$$\ell(\boldsymbol{\theta}|\mathbf{y}_c) \propto -\frac{1}{2}\left[n\log\sigma^2 + \frac{1}{\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top \mathbf{M}_n^{-1}(\boldsymbol{\phi})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \log\left|\mathbf{M}_p(\boldsymbol{\phi})\right|\right]. \tag{13}$$

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:

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

where

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

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

$$\widehat{\mathbf{y}^{(k)}} = E\{\mathbf{y}|\mathbf{V},\mathbf{C},\widehat{\boldsymbol{\theta}}^{(k)}\} \;\text{ and }\; \widehat{\mathbf{y}^{2(k)}} = E\{\mathbf{y}\mathbf{y}^\top|\mathbf{V},\mathbf{C},\widehat{\boldsymbol{\theta}}^{(k)}\}, \tag{14}$$

that is, the first and second moments of a truncated multinormal distribution. Although these can be determined in closed form as a function of multinormal probabilities (for more details on the computation of these moments, one may refer to Vaida & Liu (2009)), 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.

As an alternative to the theoretical calculation, Wei & 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 simulations. Another option 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 MCEM algorithm because the number of required simulations is considerably smaller.

The SAEM algorithm consists, at each iteration, of successively simulating the latent variable 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-step*: sample $\mathbf{y}_{k,l}^c$ $(l = 1, \ldots, m)$ from the truncated normal distribution $TN_{n^c}(\boldsymbol{\mu}, \mathbf{S}; \mathbb{A})$, where $\boldsymbol{\mu} = \mathbf{X}^c\boldsymbol{\beta} + \boldsymbol{\Sigma}^{co}(\boldsymbol{\Sigma}^{oo})^{-1}(\mathbf{y}^o - \mathbf{X}^o\boldsymbol{\beta})$, $\mathbf{S} = \boldsymbol{\Sigma}^{cc} - \boldsymbol{\Sigma}^{co}(\boldsymbol{\Sigma}^{oo})^{-1}\boldsymbol{\Sigma}^{oc}$ and $\mathbb{A} = \{\mathbf{y}^c = (y_1^c, \ldots, y_{n^c}^c)^\top \mid y_1^c \le V_1, \ldots, y_{n^c}^c \le V_{n^c}\}$;

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

$$\hat{Q}_k(\boldsymbol{\theta}) = \hat{Q}_{k-1}(\boldsymbol{\theta}) + \delta_k \left( \frac{1}{m} \sum_{l=1}^{m} \ell(\boldsymbol{\theta} \mid \mathbf{y}_{k,l}) - \hat{Q}_{k-1}(\boldsymbol{\theta}) \right), \tag{15}$$

where $\mathbf{y}_{k,l} = vec(\mathbf{y}^o, \mathbf{y}^c_{k,l})$, $l = 1, \ldots, m$ and $\delta_k$ 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

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

For model (10)–(11), the conditional maximization step becomes:

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \left( \mathbf{X}^\top \mathbf{M}_n^{-1}(\widehat{\boldsymbol{\phi}}^{(k)}) \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{M}_n^{-1}(\widehat{\boldsymbol{\phi}}^{(k)}) \widehat{\mathbf{y}^{(k)}}, \tag{16}$$

$$\begin{aligned} \widehat{\sigma^2}^{(k+1)} &= \frac{1}{n} \left( \operatorname{tr}\left( \widehat{\mathbf{y}^{2(k)}} \mathbf{M}_n^{-1}(\widehat{\boldsymbol{\phi}}^{(k)}) \right) - 2\widehat{\boldsymbol{\beta}}^{(k)\top} \mathbf{X}^\top \mathbf{M}_n^{-1}(\widehat{\boldsymbol{\phi}}^{(k)}) \widehat{\mathbf{y}^{(k)}} \right. \\ &+ \left. \widehat{\boldsymbol{\beta}}^{(k)\top} \mathbf{X}^\top \mathbf{M}_n^{-1}(\widehat{\boldsymbol{\phi}}^{(k)}) \mathbf{X} \widehat{\boldsymbol{\beta}}^{(k)} \right), \end{aligned} \tag{17}$$

$$\begin{aligned} \widehat{\boldsymbol{\pi}}^{(k+1)} &= \underset{\boldsymbol{\pi} \in (-1,1)^p}{\operatorname{argmax}} \left\{ -\frac{n}{2} \log \left[ \left( -1, \mathscr{B}(\boldsymbol{\pi})^\top \right) D \left( \widehat{\mathbf{y}^{(k)}}, \widehat{\boldsymbol{\beta}}^{(k)} \right) \begin{pmatrix} -1 \\ \mathscr{B}(\boldsymbol{\pi}) \end{pmatrix} \right] \right. \\ &- \left. \frac{1}{2} \log \left[ \prod_{j=1}^{p} \left( 1 - \pi_j^2 \right)^{-j} \right] \right\}, \end{aligned} \tag{18}$$

$$\widehat{\boldsymbol{\phi}}^{(k+1)} = \mathscr{B}(\widehat{\boldsymbol{\pi}}^{(k+1)}). \tag{19}$$

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. The initial values were calculated as follows: $\widehat{\boldsymbol{\beta}}^{(0)} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$, $\widehat{\sigma}^{2(0)} = \frac{1}{n-1} \sum_{t=1}^{n} (y_t - \mathbf{x}_t^\top \widehat{\boldsymbol{\beta}}^{(0)})$, $\widehat{\boldsymbol{\pi}}^{(0)} = pacf_p(\mathbf{y} - \mathbf{X}\widehat{\beta}^{(0)})$ and $\widehat{\boldsymbol{\phi}}^{(0)} = \mathscr{B}(\widehat{\boldsymbol{\pi}}^{(0)})$, where $pacf_p(.)$ is the sample partial autocorrelation function for lags $1, \ldots, p$, calculated with the routine *pacf()* available in the R software (R Development Core Team, 2015).

If the smoothing parameter $\delta_k$ 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:

$$\delta_k = \begin{cases} 1, & \text{if } 1 \leq k \leq cW; \\ \frac{1}{k - cW}, & \text{if } cW + 1 \leq k \leq W, \end{cases} \tag{20}$$

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 & 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 $m = 1$.

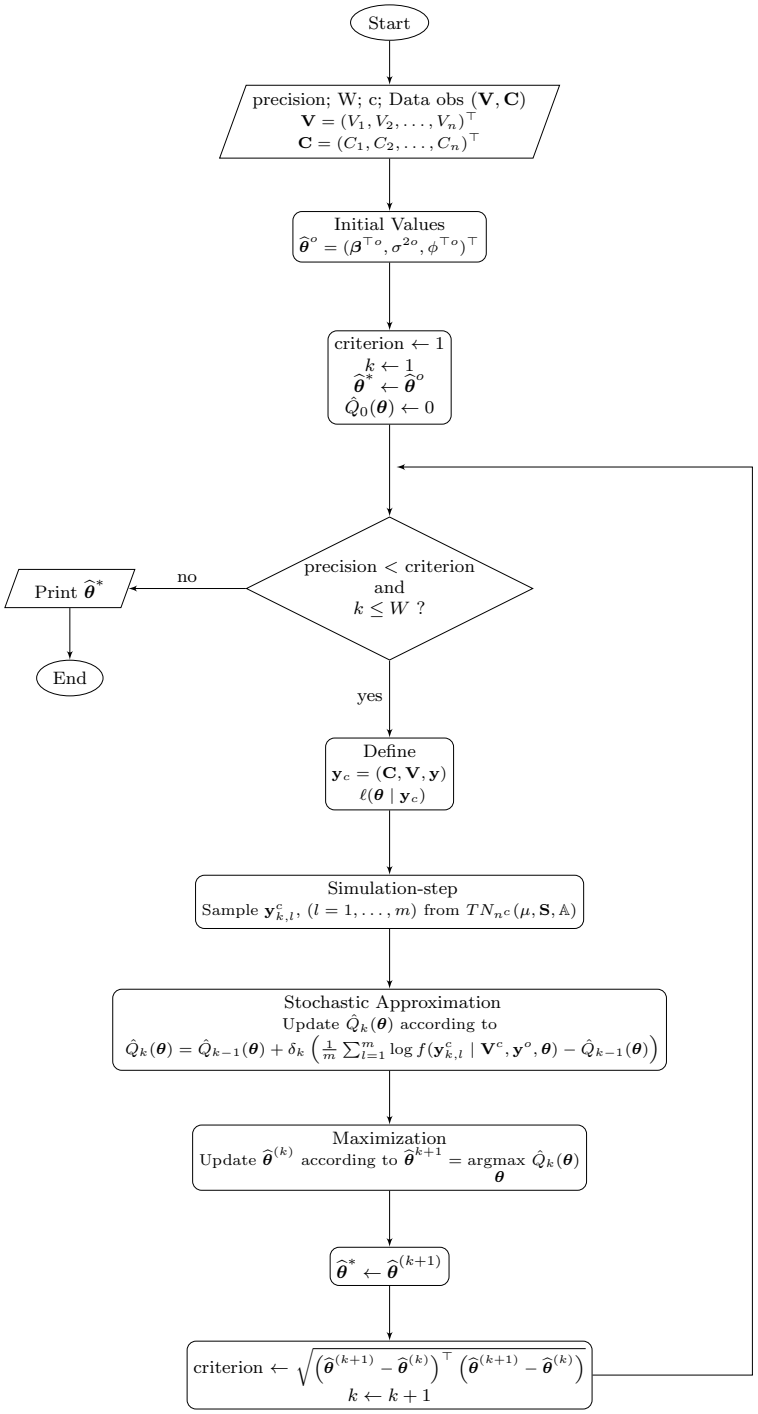


Figure 1: Flow diagram of the SAEM algorithm for the AR(p)-CR models.

### 3.3 Curved exponential family

Following Kuhn & 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

$$\ell(\boldsymbol{\theta}|\mathbf{y}_c) = \exp\left\{-\Psi(\boldsymbol{\theta}) + \left\langle \tilde{S}(\mathbf{y}), \Phi(\boldsymbol{\theta})\right\rangle\right\}, \tag{21}$$

where $\Psi$ and $\Phi$ 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 can be written in terms of the minimal sufficient statistics. Thus, since for model (4) the complete likelihood function belongs to the exponential family, the E-step reduces to the approximation of (14), and the *Stochastic Approximation* step, defined in (15), reduces to:

$$\widehat{\mathbf{y}^{2(k)}} = \widehat{\mathbf{y}^{2(k-1)}} + \delta_k\left(\frac{1}{m}\sum_{l=1}^{m}\mathbf{y}^c_{(k,l)}\mathbf{y}^{c\top}_{(k,l)} - \widehat{\mathbf{y}^{2(k-1)}}\right), \tag{22}$$

$$\widehat{\mathbf{y}^{(k)}} = \widehat{\mathbf{y}^{(k-1)}} + \delta_k\left(\frac{1}{m}\sum_{l=1}^{m}\mathbf{y}^c_{(k,l)} - \widehat{\mathbf{y}^{(k-1)}}\right),\ k = 1,2,3,\ldots. \tag{23}$$

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 1.

## 4 Standard error and prediction

### 4.1 The observed Fisher information matrix

Let $\mathbf{y}_{k,l} = vec(\mathbf{y}^o, \mathbf{y}^c_{k,l})$, $(l = 1,\ldots,m)$, $\partial_\theta\ell(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\partial\ell(\boldsymbol{\theta}|\mathbf{y})}{\partial\boldsymbol{\theta}^\top}$ and $\partial^2_\theta\ell(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\partial^2\ell(\boldsymbol{\theta}|\mathbf{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}$. 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, $\partial^2_\theta\ell(\boldsymbol{\theta} \mid \mathbf{y})$ can be approximated following the stochastic approximation procedure:

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

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 $\mathbf{I}_o(\boldsymbol{\theta}^*) = -\partial^2_\theta \log L(\boldsymbol{\theta}^*)$, where $L(\boldsymbol{\theta})$ is as defined in (12). When $\log L(\boldsymbol{\theta})$ 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).

Hence, let $\ell(\boldsymbol{\theta}|\mathbf{y}_c)$ be as defined in (4) or, equivalently, as defined in (5). Let $D = D(\mathbf{y}_c, \boldsymbol{\beta})$ be partitioned as:

$$D = \begin{bmatrix} D_{11} & D_{\phi 1}^{\top} \\ D_{\phi 1} & D_{\phi\phi} \end{bmatrix}, \tag{24}$$

such that $D_{11}$ is $1 \times 1$, $D_{\phi 1}$ is $p \times 1$ and $D_{\phi\phi}$ is $p \times p$. Then, the sum of squares can the 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, the elements of the gradient vector are:

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\theta}|\mathbf{y}_c)}{\partial \boldsymbol{\beta}} &= \frac{1}{\sigma^2}\left(\mathbf{X}^{\top}\mathbf{M}_n^{-1}\mathbf{y} - \mathbf{X}^{\top}\mathbf{M}_n^{-1}\mathbf{X}\boldsymbol{\beta}\right), \\
\frac{\partial \ell(\boldsymbol{\theta}|\mathbf{y}_c)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\top}\mathbf{M}_n^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \\
\frac{\partial \ell(\boldsymbol{\theta}|\mathbf{y}_c)}{\partial \boldsymbol{\phi}} &= -\frac{1}{2}\mathrm{tr}\left\{\mathbf{M}_p^{-1}\frac{\partial \mathbf{M}_p}{\partial \boldsymbol{\phi}}\right\} - \frac{1}{\sigma^2}\left(-D_{\phi 1} + D_{\phi\phi}\boldsymbol{\phi}\right),
\end{aligned}$$

and the elements of the Hessian matrix are:

$$\begin{aligned}
\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{y}_c)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\top}} &= -\frac{1}{\sigma^2}\mathbf{X}^{\top}\mathbf{M}_n^{-1}\mathbf{X}, \\
\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{y}_c)}{\partial \boldsymbol{\beta}\partial \sigma^2} &= -\frac{1}{\sigma^4}\left(\mathbf{X}^{\top}\mathbf{M}_n^{-1}\mathbf{y} - \mathbf{X}^{\top}\mathbf{M}_n^{-1}\mathbf{X}\boldsymbol{\beta}\right), \\
\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{y}_c)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\phi}^{\top}} &= -\frac{1}{\sigma^2}\frac{\partial(D_{\phi\phi}\boldsymbol{\phi} - D_{\phi 1})}{\partial \boldsymbol{\beta}}, \\
\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{y}_c)}{\partial (\sigma^2)^2} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\top}\mathbf{M}_n^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \\
\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{y}_c)}{\partial \boldsymbol{\phi}\partial \sigma^2} &= \frac{1}{\sigma^4}\left(D_{\phi\phi}\boldsymbol{\phi} - D_{\phi 1}\right), \\
\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{y}_c)}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}^{\top}} &= -\frac{1}{\sigma^2}D_{\phi\phi} - \frac{1}{2}\mathrm{tr}\left\{\frac{\partial}{\partial \boldsymbol{\phi}}\left(\mathbf{M}_p^{-1}\frac{\partial \mathbf{M}_p}{\partial \boldsymbol{\phi}^{\top}}\right)\right\}.
\end{aligned}$$

## 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 $\widetilde{\mathbf{y}} = \left(\mathbf{y}_{obs}^{\top}, \mathbf{y}_{pred}^{\top}\right)^{\top}$.

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:

$$\widetilde{\mathbf{y}}^* = \left(\mathbf{y}^{*\top}_{obs}, \mathbf{y}^{\top}_{pred}\right)^{\top} \sim N_{n_{obs}+n_{pred}}\left(\widetilde{\mathbf{X}}\beta, \boldsymbol{\Sigma}\right), \tag{25}$$

where $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}^{obs^*,obs^*} & \boldsymbol{\Sigma}^{obs^*,pred} \\ \boldsymbol{\Sigma}^{pred,obs^*} & \boldsymbol{\Sigma}^{pred,pred} \end{pmatrix}$. 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:

$$\widehat{\mathbf{y}}_{pred}(\boldsymbol{\theta}) = \mathbf{X}_{pred}\beta + \boldsymbol{\Sigma}^{pred,obs^*}\left(\boldsymbol{\Sigma}^{obs^*,obs^*}\right)^{-1}(\mathbf{y}_{obs} - \mathbf{X}_{obs}\beta). \tag{26}$$

Therefore, the predictor of $\mathbf{y}_{pred}$ can be calculated by substituting $\widehat{\boldsymbol{\theta}}$ in (26), obtaining

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

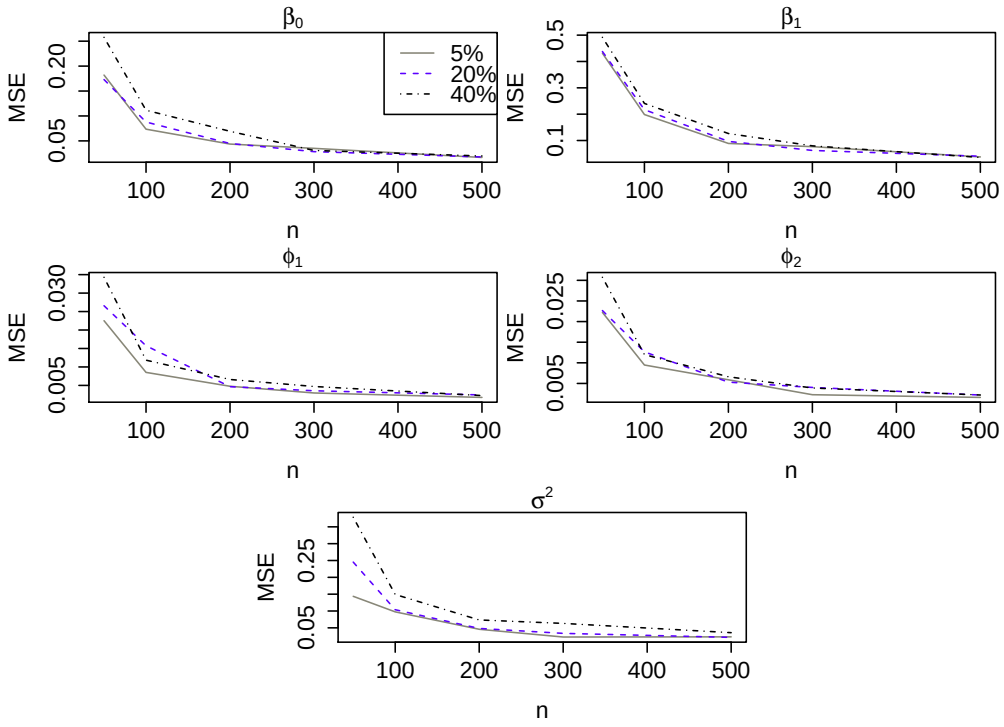


Figure 2: Mean square error of the parameter estimates under 5%, 20% and 40% censoring levels and different samples sizes. The solid line (gray) represents 5% censoring, the dashed line (blue) represents 20% and the dot-dashed line (black) represents 40%.

# 5 Simulation studies

Two simulation studies were conducted to examine the performance of the proposed model, by analyzing the asymptotic properties of the SAEM estimates and prediction accuracy based in the

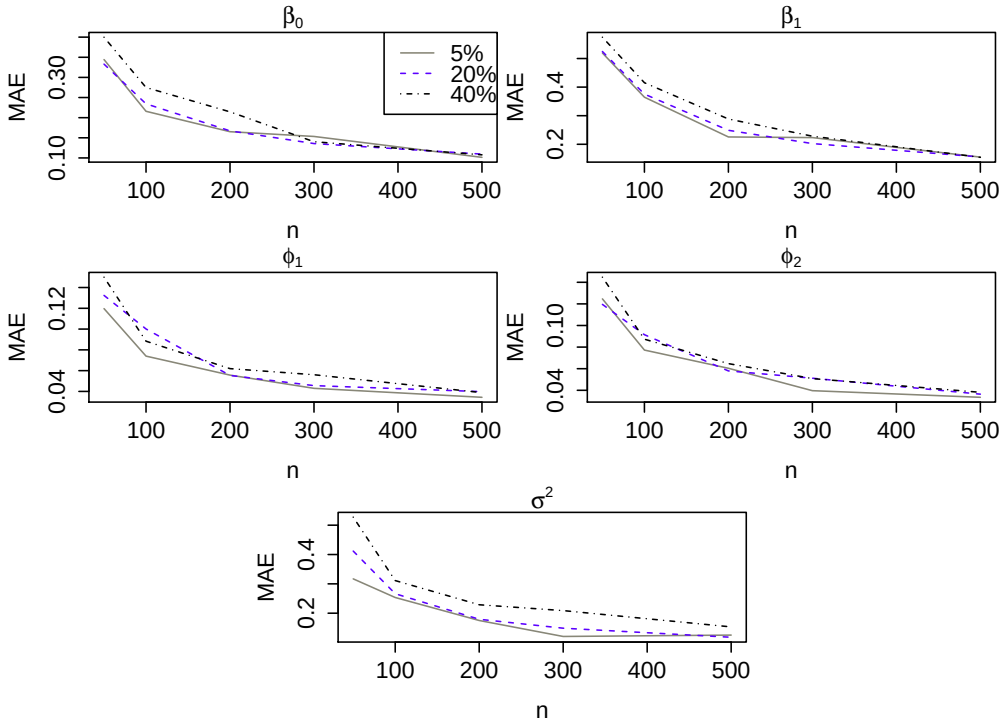


Figure 3: Mean absolute error of the parameter estimates under different censoring levels and different samples sizes. The solid line (gray) represents 5% censoring, the dashed line (blue) represents 20% and the dot-dashed line (black) represents 40%.

plug-in method proposed in Section 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 $\varepsilon_t$ from the model defined in (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 $Y_t = V\, \mathbb{I}_{(Z_t < V)} + Z_t\, \mathbb{I}_{(Z_t \geq V)}$.

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 a cutoff point at $c = 0.12$.

## 5.1 First study

The main goal of this simulation study is to analyze the behavior of the SAEM estimates under different proportions of censoring and sample sizes. The simulated data follow an AR(2)-CR model, as defined in (10)–(11), 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 $\theta_i$ are defined as:

$$\text{MSE}_i = \frac{1}{100}\sum_{j=1}^{100}\left(\widehat{\theta}_i^{(j)} - \theta_i\right)^2 \text{ and } \text{MAE}_i = \frac{1}{100}\sum_{j=1}^{100}\left|\widehat{\theta}_i^{(j)} - \theta_i\right|, \tag{27}$$

where $\widehat{\theta}_i^{(j)}$ is the ML estimate of the parameter $\theta_i$ for the $j$th sample, $j = 1,\ldots,100$. The main goal of this simulation is to provide empirical evidence of the consistency of the ML estimates, for different censoring proportions. Figures 2 and 3show that the MSE and the MAE tend to zero as the sample size increases. Thus, as a general rule the results indicate that the SAEM estimates of the proposed model do provide good asymptotic properties. Note also that simulations for higher censoring rates $(p = 40\%)$, the convergence still behaved well. Table 1 presents the summary statistics for parameter estimation. Here, we also examine the consistency of the approximation method, suggested in Subsection 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

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

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

- The average values of the approximate standard errors of the SAEM estimates obtained through the method described in Subsection 4.1 using the empirical information matrix, denoted by IBM-SE.

Table 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.

## 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 (V) 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, 2015).

The simulated data follow an AR(2)-CR model, as defined in (10)–(11), 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 for prediction comparison purposes 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:

$$\text{MSPE} = \frac{1}{300}\sum_{l=1}^{100}\sum_{j=498}^{500}\left(y_j^{(l)} - \widehat{y}_j^{(l)}\right)^2 \text{ and } \text{MAPE} = \frac{1}{300}\sum_{l=1}^{100}\sum_{j=498}^{500}\left|y_j^{(l)} - \widehat{y}_j^{(l)}\right|, \tag{28}$$

Table 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 Section 4.1.

| n | CP | | $\beta_0$ | $\beta_1$ | $\sigma^2$ | $\phi_1$ | $\phi_2$ |
|---|---|---|---|---|---|---|---|
| | | 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 |

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

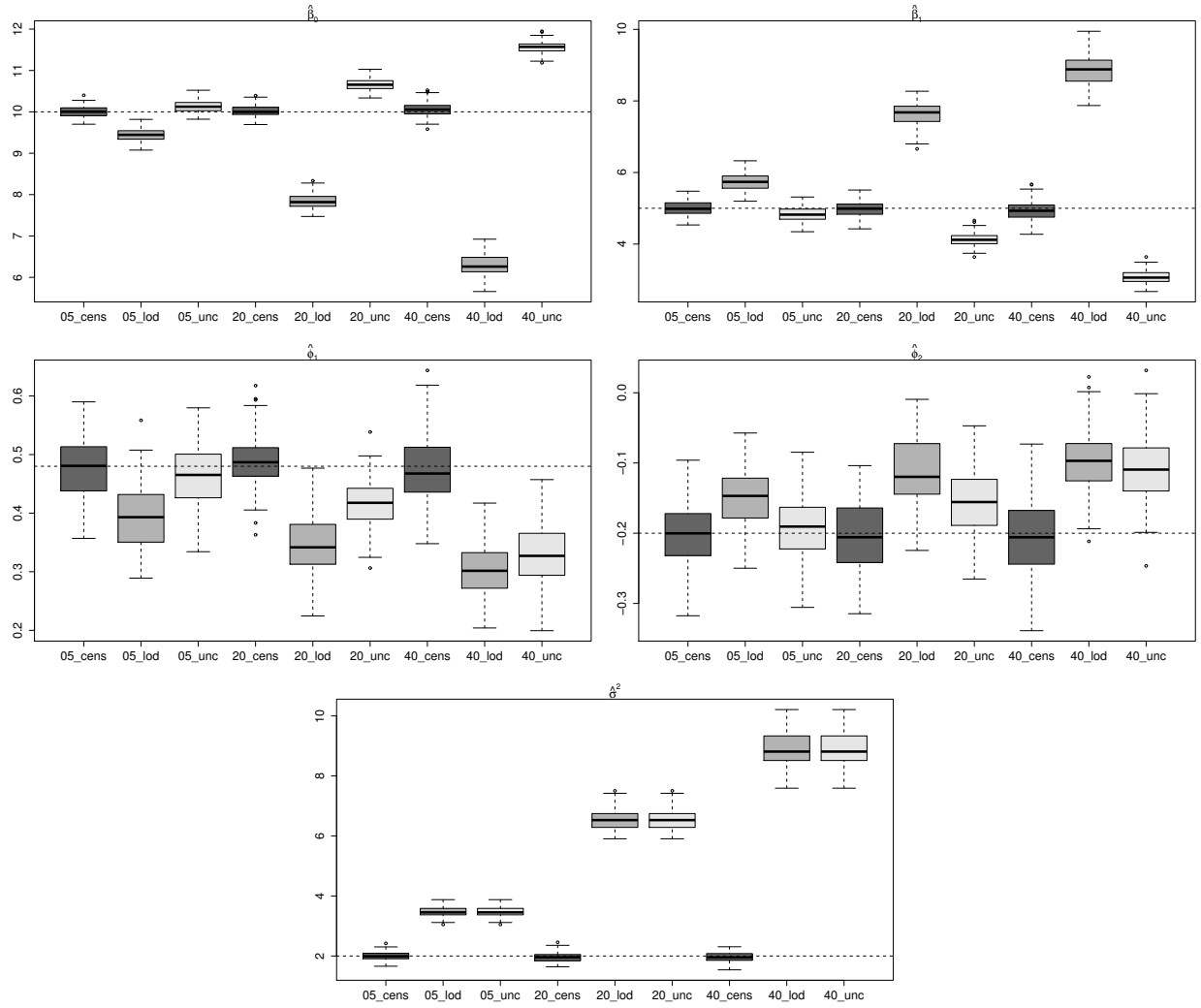


Figure 4: Boxplot of the parameter estimates for $n = 497$. The dotted line indicates the true value of the parameter. *05_cens* indicates the result for the censored case with censoring level of 5%, *05_lod* indicates the result for the LOD procedure with censoring level of 5% and so on.

censoring rates considered and this difference became greater as the censoring rate increased.

Table 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 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.

# 6 Application to real data

In this section we consider the analysis of 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 series.

The data were originally collected by the National Center for Atmospheric Research (NCAR) based of hourly observations in San Francisco, recorded during the month of March 1989, consisting of 716 observations. The censoring rate is 40.5%. Figure 5 shows the log-transformed data whose original scale is hundred feet. There were three missing observations and we considered the censoring interval $A_j = (-\infty, \infty)$ for those observations.

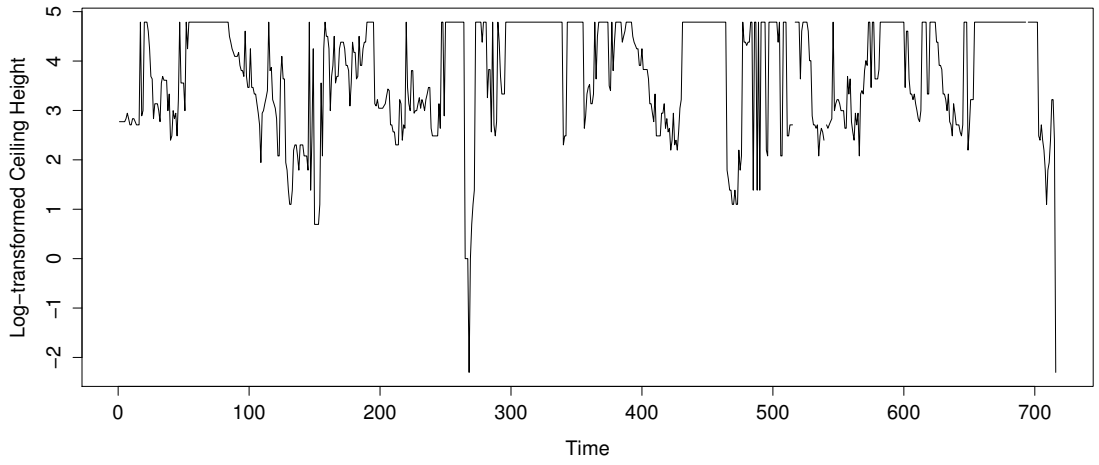


Figure 5: Censored time series of log-transformed hourly cloud ceiling height in San Francisco during March 1989.

For model selection we considered $p = 1, 2$ and 3. The AR(2)-CR model presented the smallest AIC value and therefore was selected. Figure 6 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.

For the selected model (AR(2)-CR) , Figure 7 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 the convergence is attained quickly.

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

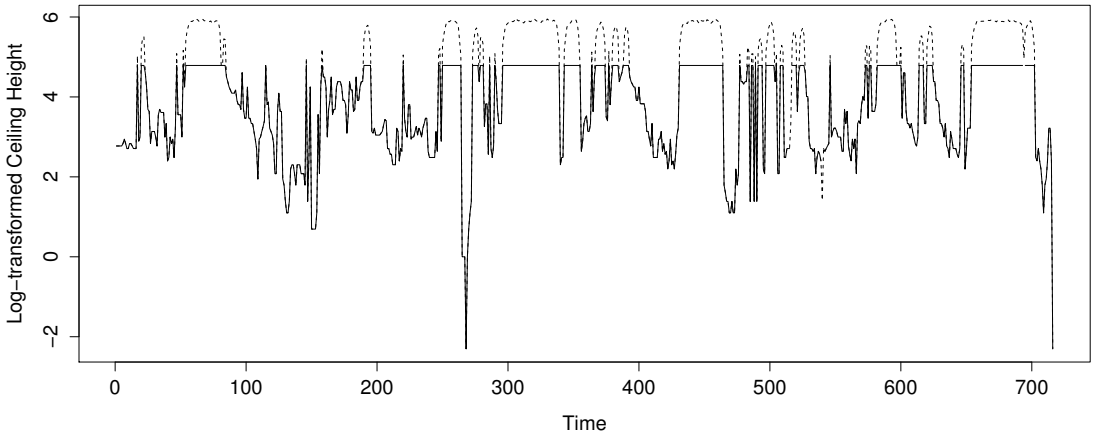


Figure 6: Censored time series of log-transformed hourly cloud ceiling height in San Francisco during March 1989. The dashed line represents the augmented serie based on the fitted AR(2)-CR model.

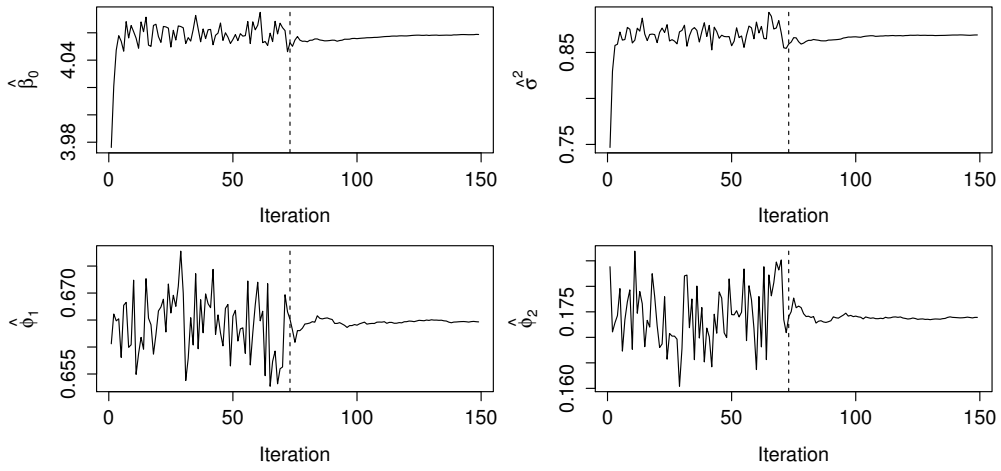


Figure 7: Convergence of the SAEM parameters estimates for the AR(2)-CR model.

# 7 Conclusions

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

tical demonstration, the method is applied to a dataset on cloud ceiling height, measured subject to the detection limit of the recording device. We also use simulation to investigate the properties of predictions and parameter estimates and the robustness of the SAEM algorithm. In this simulation study comparisons are made between inferences based on the censored data and inferences based on complete data obtained by a crude/ad hoc imputation method. We show 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. The proposed methods are implemented in the R package *ARCensReg* (Schumacher *et al.*, 2016; R Development Core Team, 2015), providing practitioners with a convenient tool for further applications in their domain.

Future extensions of the work include the use of scale mixtures of normal distributions to accommodate heavy-tailed features (Lachos *et al.*, 2011). The proposed methods can also be easily applied to other substantive areas where the data being analyzed have censored observations. In line with this, future work includes extending the proposed methods to accommodate missing values in addition to censoring using hybrid Bayesian sampling procedures (Wang & Fan, 2012) and multivariate outcomes (Wang *et al.*, 2015).

# Acknowledgements

Victor H. Lachos acknowledges support from CNPq-Brazil (Grant 305054/2011-2) and FAPESP-Brazil (Grant 2014/02938-9). Fernanda L. Schumacher was supported by CAPES-Brazil. We also thank Dra. Francyelle de Lima Medina from Federal University of Pernanbuco for his help on the earlier version of the paper.

# References

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

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

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

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

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

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

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

Hopke, P. K., Liu, C. & Rubin, D. B. (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.

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

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

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

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

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

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

R Development Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. 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*, pages 167–182.

Schumacher, F., Lachos, V. & Galarza, C. (2016). ARCensReg: Fitting Univariate Censored Linear Regression Model with Autoregressive Errors. *R package version 1.0-0, URL http://CRAN.R-project.org/package=ARCensReg*.

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

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. & Fan, T.-H. (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., Lin, T.-I. & Lachos, V. H. (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. & Tanner, M. A. (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.

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