Censored Multivariate Linear Regression Model

Rodney Sousa (0000-0001-9205-5487)[1], Isabel Pereira (0000-0002-5152-546X)[1], and Maria Eduarda Silva (0000-0003-2972-2050)[2]

[1] University of Aveiro & Center for Research and Development in Mathematics and Applications, Portugal,
rodney@ua.pt

[2] University of Porto & Center for Research and Development in Mathematics and Applications, Portugal

Abstract. Often, real life problems require modelling several response variables together. This work analyses multivariate linear regression model when the data are censored. Censoring distorts the correlation structure of the underlying variables and increases the bias of the usual estimators. Thus, we propose three methods to deal with multivariate data under left censoring, namely, Expectation Maximization (EM), Data Augmentation (DA) and Gibbs Sampler with Data Augmentation (GDA). Results from a simulation study show that both, DA and GDA estimates are consistent for low and moderate correlation. Under high correlation scenarios EM estimates present lower bias.

Keywords: Censored Data, Multivariate Linear Regression

1 Introduction

Linear regression (LR) is one of the most widely used models in Econometrics to analyse the relationship between two sets of variables. Often, real life problems can be best described by considering several (say $m \geq 2$) correlated response variables, that is, experiments are performed to analyse the variation of m characteristics of the same phenomenon. In these cases, we should consider multivariate LR model, which is a natural extension of the univariate regression model. An essential aspect of multivariate analysis is the dependence between the different variables, which may involve the covariance between them [2].

Additionally, some or all of the response variables can be censored, meaning that they are only accessible in a restricted interval. Censored data can arise for a variety of reasons, such as limitations of the measuring device or of the experimental design [12]. Examples occur in environmental studies where mineral concentration in air/water may be subjected to lower detection limits [9], in Medicine, where [3] studied the relationship between two cytokines (pro-inflammatory and anti-inflammatory) when both variables are censored or in Economics where hours worked is usually treated as censored variable [1]. We might note that in the literature, the terminology censored data is also used in the survival data analysis, in which the variable of interest is the time to an event. In these cases, unexpected interruptions of scheduled experiments create fully missing values or censored survival (or failure time) data. The structure of such data and the censored data described above are quite different and require different statistical techniques for their analysis ([13],[5]). Our discussion will focus on the first type of censored data in which the outcome or variable of interest is below (or above) a limit of detection (LOD).

Censoring makes the observed data set incomplete and therefore direct analysis using standard complete data methods inadequate, resulting in inconsistent estimates. To overcome these issues, a variety of methods have been proposed to handle censored univariate data (see [17], [6], [20]). Filling in censored data in order to apply standard complete data methods has a strong intuitive appeal, because this strategy greatly reduces the burden of developing specialized methods and computer code for analysing incomplete data [9].

Methods for creating complete data via filling in censored data can be single imputation (one value for each observation) or multiple imputation. In single imputation, it is common to fill in the censored observation by its expected value, predicted mean or the center of the detection interval. More statistically sound approaches are based on the EM and DA algorithms ([8], [17]). However, extension of methods to handle censored data in multivariate setting confronts a significant practical barrier. Indeed, there are very few works is this subject ([14], [5], [9]). In particular, to the best of our knowledge, there is no specific work in literature about censored multivariate linear regression model (CMLR).

Muthén [15] pointed out that, in addition to inconsistent estimates, censoring also distorts the correlation structure of the response variables. Aiming to develop more suitable methods to handle this problem, in this work we propose three methods to estimate CMLR, mainly Expectation Maximization (EM), Data Augmentation (DA) and Gibbs sampler with Data Augmentation (GDA). All of these methods are based on filling in censored data in order to create a complete data set, which is the most widely used strategy when the data are missing or censored, both in Classical and Bayesian approaches.

The paper is organized as follows: Section 2 presents the CMLR model, Section 3 analyses three methods to estimate CMLR model, in Section 4 we present the simulation study, in which we analyse the accuracy of the proposed methods and, finally, we present some final remarks.

2 Censored Multivariate Linear Regression

In this section we define multivariate linear regression model in order to introduce censored multivariate linear regression.

2.1 The Multivariate Linear Regression Model

In matrix form, the Multivariate Linear Regression Model (MLR) can be written as follows:

$$\mathbf{W} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{W} = [\mathbf{W}_{(1)} \;\; \ldots \;\; \mathbf{W}_{(m)}]$ is a $n \times m$ matrix of m response variables, $\mathbf{X}$ is a $n \times (k+1)$ matrix of k predictors, whose rows are $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ik})'$, $\;i = 1, \ldots, n$, $\boldsymbol{\beta}$ is a $(k+1) \times m$ coefficients matrix and $\boldsymbol{\varepsilon} = [\boldsymbol{\varepsilon}_{(1)} \;\; \ldots \;\; \boldsymbol{\varepsilon}_{(m)}]$ is a $n \times m$ matrix of the errors associated with each response variable, where each $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{im})'$, $\;i = 1, \ldots, n$ is assumed to be iid $m-$variate normal variable with mean $\mathbf{0}$ and $m \times m$ covariance matrix $\boldsymbol{\Sigma} = [\sigma_{ij}]$ [11]. Then, the model (1) may be written as

$$\begin{bmatrix} W_{11} & \ldots & W_{1m} \\ \vdots & \ddots & \vdots \\ W_{n1} & \ldots & W_{nm} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{bmatrix} \cdot \begin{bmatrix} \beta_{01} & \ldots & \beta_{0m} \\ \vdots & \ddots & \vdots \\ \beta_{k1} & \ldots & \beta_{km} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} & \ldots & \varepsilon_{1m} \\ \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \ldots & \varepsilon_{nm} \end{bmatrix} \tag{2}$$

where $E[\boldsymbol{\varepsilon}_{(j)}] = 0$ and $Cov(\boldsymbol{\varepsilon}_{(i)}, \boldsymbol{\varepsilon}_{(j)}) = \sigma_{ij}\mathbf{I}_n$, $\sigma_{jj} = \sigma^2$, $\;i, j = 1, \ldots, m$ and $\mathbf{I}_n$ is the $n \times n$ identity matrix. This is the generalization of multiple LR ($m = 1$), where each response variable $\mathbf{W}_{(j)}$, $\;j = 1, \ldots, m$, follows a multiple LR model.

In the MLR model (2), observations from different individuals are uncorrelated, but the errors for different responses of the same individual can be correlated [11]. By using the multivariate model the covariance of the response variables can be modelled, which is not possible in case of separate univariate regression models.

2.2 The Censored Multivariate Linear Regression Model

Let's assume that the latent variable $\mathbf{W}_i = (W_{i1}, \ldots, W_{im})'$ denotes the m multivariate measure on subject $i = 1, \ldots, n$, and that each component vector $W_{(j)}$ of the hypothetical multivariate data $\mathbf{W}$ is subjected to left censoring at fixed limit of detection (LOD), $L_j \in \mathbf{R}$, $\;j = 1, \ldots, m$. Rather than $\mathbf{W}_i$ we actually observe $\mathbf{Y}_i = (y_{i1}, \ldots, y_{im})'$, where $y_{ij} = \max\{w_{ij}, L_j\}$ and corresponds to the $j-$th record on the subject i, for $i = 1, \ldots, n$ ([14], [5]). Here we are assuming that the censoring patterns vary across the component vectors, but is fixed within each $W_{(j)}$, for $j = 1, \ldots, m$.

Now, given a dataset $\mathbf{Y} = (\mathbf{y}_1', \ldots, \mathbf{y}_n')'$, each observation $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})'$ of the CMLR model can be defined as follows:

$$\begin{aligned} \mathbf{Y} =& [y_{ij}] = [\max(w_{ij}, L_j)], \quad i = 1, \ldots, n \text{ and } j = 1, \ldots, m, \\ \mathbf{W} =& \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \end{aligned} \tag{3}$$

For simplicity of notation, the remaining of the text focus in the bivariate case, $m = 2$, defined as follows:

Censored Bivariate Linear Regression. We assume that the errors term $\boldsymbol{\varepsilon}_i$, $\;i = 1, ..., n$ has bivariate normal distribution $N_2(\mathbf{0}, \boldsymbol{\Sigma})$, the probability density function (pdf) of the latent variable $\mathbf{W}_i$ is $N_2(\boldsymbol{\beta}'\mathbf{x}_i, \boldsymbol{\Sigma})$, and has the form

$$\begin{aligned} f(W_{i1}, W_{i2}) =& \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\Big\{-\frac{1}{2(1-\rho^2)}\Big[\Big(\frac{W_{i1}-\mathbf{x}_i'\boldsymbol{\beta}_1}{\sigma_1}\Big)^2 \\ &+\Big(\frac{W_{i2}-\mathbf{x}_i'\boldsymbol{\beta}_2}{\sigma_2}\Big)^2 - 2\rho\frac{(W_{i1}-\mathbf{x}_i'\boldsymbol{\beta}_1)(W_{i2}-\mathbf{x}_i'\boldsymbol{\beta}_2)}{\sigma_1\sigma_2}\Big]\Big\}, \end{aligned} \tag{4}$$

while the observed $\mathbf{Y}_i$ variable has a bivariate truncated normal distribution, with support $[L_1, \infty] \times [L_2, \infty]$ and pdf

$$f(Y_{i1}, Y_{i2}|W_{i1} \geq L_1, W_{i2} \geq L_2) = \frac{f(W_{i1}, W_{i2})}{P(W_{i1} \geq L_1, W_{i2} \geq L_2)} \times I_{(W_{i1} \geq L_1, W_{i2} \geq L_2)}. \tag{5}$$

Although there are several approaches and methods to estimate CLR in the univariate case, extensions to multivariate settings confront a significant practical barrier. Muthén [15] observed that censoring distorts the correlation structure of the underlying variable and presented results on a general formula for truncation in the standard bivariate normal distributions. Cohen [7] found a maximum likelihood solution for the truncated bivariate normal where the truncation is with respect to only one variable, while Tallis [16] gave general formulas for multivariate truncation from below in the multivariate normal distribution using the moment-generating function.

3 Estimation of CMLR model

In this section we propose three methods to estimate the CMLR model, focusing on left-censored bivariate data. All these methods are based on filling in the censored data in order to obtain complete data.

3.1 EM Algorithm for Multivariate Data

The EM (*Expectation Maximization*) algorithm is an iterative method to maximize the expected value of the likelihood function, given the observed data, $\mathbf{Y}$ [8]. In case of censored bivariate data, the algorithm requires the computation of the expected value of the truncated bivariate variable, in order to fill up the data. If the latent variable $\mathbf{W}_i = (W_{i1}, W_{i2})'$ is left–censored, then the values below the LOD have right-truncated distribution, with expected value given by

$$\begin{aligned} &E[(W_{i1}, W_{i2})'|W_{i1} \leq L_1, W_{i2} \leq L_2] = \\ &\qquad (E[W_{i1}|W_{i1} \leq L_1, W_{i2} \leq L_2], E[W_{i2}|W_{i1} \leq L_1, W_{i2} \leq L_2])'. \end{aligned} \tag{6}$$

for $i = 1, \ldots, n$. Using the moment-generating function, Tallis [16] gave general formulas for truncated multivariate multivariate normal distribution. Let $\alpha =$

$P(W_1 \leq L_1, W_2 \leq L_2) = F(L_1, L_2)$ represent the probability that the random variable $\mathbf{W} = (W_1, W_2)'$ takes on a value less than or equal to $\mathbf{L} = (L_1, L_2)'$. Taking $\mu_j = E[W_{(j)}]$, $\eta_j = (W_j - \mu_j)/\sigma_j$ and $\gamma_j = (L_j - \mu_j)/\sigma_j, \quad j = 1, 2$, the probability α can be written as

$$\alpha = P(\eta_1 \leq \gamma_1, \eta_2 \leq \gamma_2), \tag{7}$$

where $\eta_j, \quad j = 1, 2$, are standardized normal variables, truncated at $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$. Thus, we can write

$$\alpha = \Phi(\boldsymbol{\gamma}; \mathbf{R}), \tag{8}$$

where

$$\mathbf{R} = \begin{bmatrix} 1 \ \rho \\ \rho \ 1 \end{bmatrix} \tag{9}$$

is the correlation matrix with $\rho = corr(\eta_1, \eta_2)$ [16] and the expected value of the truncated standardized bivariate variable, $\boldsymbol{\eta} = (\eta_1, \eta_2)'$, is given by

$$E[\eta_i|\boldsymbol{\eta} \leq \boldsymbol{\gamma}] = \frac{1}{\alpha} \times \left\{\rho_{i1}\phi(\gamma_1)\Phi(A_{12}; \mathbf{R}_1) + \rho_{i2}\phi(\gamma_2)\Phi(A_{21}; \mathbf{R}_2)\right\}, \quad i = 1, 2, \tag{10}$$

where $A_{ij} = (\gamma_j - \rho_{ji}\gamma_i)/\sqrt{1 - \rho_{ji}^2}$, for $i, j = 1, 2$ and $i \neq j$ [16].

From (10) results that

$$\begin{aligned} E[\eta_1|\boldsymbol{\eta} \leq \boldsymbol{\gamma}] &= \frac{1}{\alpha}\left\{\rho_{11}\phi(\gamma_1)\Phi(A_{12}) + \rho_{12}\phi(\gamma_2)\Phi(A_{21})\right\} \\ E[\eta_2|\boldsymbol{\eta} \leq \boldsymbol{\gamma}] &= \frac{1}{\alpha}\left\{\rho_{21}\phi(\gamma_1)\Phi(A_{12}) + \rho_{22}\phi(\gamma_2)\Phi(A_{21})\right\}, \end{aligned} \tag{11}$$

where $\phi(.)$ and $\Phi(.)$ are, respectively, the pdf and distribution function of standard normal variable.

Using the result in the equation (11), the expected value of each component W_{ij} of the truncated variable $\mathbf{W}_i = (W_{i1}, W_{i2})$ is given by

$$E[W_{ij}|\mathbf{W} \leq \mathbf{L}] = \mathbf{x}_i'\boldsymbol{\beta}_{(j)} + \sigma_j \times E[\eta_j|\boldsymbol{\eta} \leq \boldsymbol{\gamma}] \tag{12}$$

where $E[W_j|\mathbf{W} \leq \boldsymbol{\gamma}], \quad j = 1, 2$ are the conditional expected value of standardized normal variables.

At iteration t, after filling up the censored observed data set, the complete dataset $\mathbf{Y}^{(t)}$ is then used to compute the expected log-likelihood function, conditional on $\hat{\boldsymbol{\theta}}^{(t-1)}$,

$$Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(t-1)}) = E[logL(\boldsymbol{\theta}|\mathbf{W}, \boldsymbol{\theta}^{(t-1)})] \tag{13}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Sigma})$ and $L(\boldsymbol{\theta}|\mathbf{W})$ denotes the likelihood function given the complete data. The expected MLE estimates satisfy $\hat{\boldsymbol{\theta}}^{(t)} = argmax \quad Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(t-1)})$. The value of $\boldsymbol{\beta}$ which maximizes (13) is

$$\hat{\boldsymbol{\beta}}^{(t)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(t)}. \tag{14}$$

Given an estimate of $\boldsymbol{\beta}$, an unbiased estimate for $\boldsymbol{\Sigma}$ is

$$\hat{\boldsymbol{\Sigma}}^{(t)} = \frac{1}{n-m-1}(\mathbf{W}^{(t)} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(t)})'(\mathbf{W}^{(t)} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(t)})'. \tag{15}$$

3.2 Data Augmentation algorithm

The data augmentation (DA) algorithm as described here is based on successive updating of the censored observations, and the corresponding ordinary least squares (OLS) estimates are computed using the augmented data.

At each iteration of the DA algorithm, censored values of each response variable $\mathbf{W}_{(j)}$ are sampled from their univariate truncated distribution conditional on the values of the remaining response variables, corresponding to the same subject. This procedure results in a sequence of random $m-$variate variables which converge in probability to the joint distribution of the $m-$variate latent variable $\mathbf{W} = (W_1, \ldots, W_m)$ [4].

In multivariate distributions, the acceptance-rejection algorithms are feasible, but the rate of convergence may be too low to be practical. Thus, a more efficient algorithm is the data augmentation, in which incomplete data is reconstructed using a Gibbs sampler type algorithm [10],[4].

3.3 Gibbs Sampler with Data Augmentation algorithm

The Gibbs sampling with data augmentation (GDA) algorithm [17] allows the use of a Bayesian approach to estimate the CMLR model, where inferences about the model parameters are obtained from the posterior distribution, $\pi(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{W})$, defined by

$$\pi(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{W}) \propto L(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{W}) \times \pi(\mathbf{B}, \boldsymbol{\Sigma}), \tag{16}$$

where $L(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{W})$ is the likelihood function of the observed data and $\pi(\mathbf{B}, \boldsymbol{\Sigma})$ represents the joint prior distribution of the parameters.

The Likelihood Function. As in [18], model (2) may be rewritten equivalently as

$$\mathbf{W}^* = \mathbf{X}^*\mathbf{B} + \boldsymbol{\epsilon}, \tag{17}$$

where $\mathbf{W}^* = (\mathbf{W}'_{(1)}, \ldots, \mathbf{W}'_{(m)})'$ is a $mn \times 1$ vector, $\mathbf{X}^* = diag(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(m)})$ is a $mn \times (mk + m)$ block diagonal matrix, where $\mathbf{X}^{(1)} = \ldots = \mathbf{X}^{(m)} = \mathbf{X}$, $\mathbf{B} = (\boldsymbol{\beta}'_{(1)}, \ldots, \boldsymbol{\beta}'_{(m)})'$ is a $(mk + m) \times 1$ vector of the regression coefficients and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}'_{(1)}, \ldots, \boldsymbol{\epsilon}'_{(m)})'$ is a $mn \times 1$ vector of the disturbances, assumed to be normally distributed, with zero mean and covariance matrix $\boldsymbol{\Sigma} \otimes \mathbf{I}_n$. Then, the likelihood function for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ may be rewritten as

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{W}^*) =&(2\pi)^{-nm/2}|\boldsymbol{\Sigma}|^{-n/2}\exp\Big\{-\frac{1}{2}\boldsymbol{\epsilon}'\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_n\boldsymbol{\epsilon}\Big\}. \\ =&(2\pi)^{-nm/2}|\boldsymbol{\Sigma}|^{-n/2}\exp\Big\{-\frac{1}{2}(\mathbf{W}^*-\mathbf{X}^*\mathbf{B})'\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_n(\mathbf{W}^*-\mathbf{X}^*\mathbf{B})\Big\} \end{aligned} \tag{18}$$

where $\otimes$ is the Kronecker product and $\mathbf{I}_n$ is the identity matrix of order n.

Using the properties of the *trace* of a matrix [11] and considering that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}$ and $\mathbf{A} = (\mathbf{W}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{W}-\mathbf{X}\hat{\boldsymbol{\beta}})$ are jointly sufficient for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ [18], the likelihood function (18) can be simplified to

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{W}) =&(2\pi)^{-nm/2}|\boldsymbol{\Sigma}|^{-n/2} \\ &\times\exp\Big\{-\frac{1}{2}tr\boldsymbol{\Sigma}^{-1}\mathbf{A}-\frac{1}{2}(\mathbf{B}-\hat{\mathbf{B}})'\boldsymbol{\Sigma}^{-1}\otimes\mathbf{X}'\mathbf{X}(\mathbf{B}-\hat{\mathbf{B}})\Big\}. \end{aligned} \tag{19}$$

The Prior Distribution. Now, let's assume that $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are independent [18]. Then, a non-informative prior distribution for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ can be written as

$$\pi(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{\Sigma}). \tag{20}$$

Due to the invariance property [18], we have that

$$\begin{aligned} &\pi(\boldsymbol{\beta}) \propto C, \\ &\pi(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{m+1}{2}}, \end{aligned} \tag{21}$$

where C is a constant. Then, $\pi(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto |\Sigma|^{-\frac{m+1}{2}}$.

The Posterior Distribution. Using the prior distribution in (21) in conjunction with the likelihood function (19), the posterior distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ is given by

$$\begin{aligned} \pi(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{W}) \propto& L(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{W}) \times \pi(\mathbf{B}, \boldsymbol{\Sigma}) \\ \propto&|\boldsymbol{\Sigma}|^{-\frac{n+m+1}{2}}\exp\Big\{-\frac{1}{2}tr\boldsymbol{\Sigma}^{-1}\mathbf{A}-\frac{1}{2}(\mathbf{B}-\hat{\mathbf{B}})'\boldsymbol{\Sigma}^{-1}\otimes\mathbf{X}'\mathbf{X}(\mathbf{B}-\hat{\mathbf{B}})\Big\} \\ \propto&\exp\Big\{-\frac{1}{2}(\mathbf{B}-\hat{\mathbf{B}})'\boldsymbol{\Sigma}^{-1}\otimes\mathbf{X}'\mathbf{X}(\mathbf{B}-\hat{\mathbf{B}})\Big\} \\ &\times|\boldsymbol{\Sigma}|^{-\frac{n+m+1}{2}}\exp\Big\{-\frac{1}{2}tr\big(\boldsymbol{\Sigma}^{-1}\mathbf{A}\big)\Big\} \end{aligned} \tag{22}$$

From the equation (22) and taking only the terms involving each model parameters, the conditional posterior distribution of $\mathbf{B}$ and $\boldsymbol{\Sigma}$ can be expressed as

$$\pi(\mathbf{B}, \boldsymbol{\Sigma}|\mathbf{W}) = \pi(\mathbf{B}|\boldsymbol{\Sigma}, \mathbf{W})\pi(\boldsymbol{\Sigma}|\mathbf{W}), \tag{23}$$

with

$$\pi(\mathbf{B}|\boldsymbol{\Sigma},\mathbf{W}) \propto exp\Big\{-\frac{1}{2}(\mathbf{B}-\hat{\mathbf{B}})'\boldsymbol{\Sigma}^{-1}\otimes\mathbf{X}'\mathbf{X}(\mathbf{B}-\hat{\mathbf{B}})\Big\} \tag{24}$$

and

$$\pi(\boldsymbol{\Sigma}|\mathbf{W}) \propto |\boldsymbol{\Sigma}|^{-\frac{n+m+1}{2}}\exp\Big\{-\frac{1}{2}tr\big(\boldsymbol{\Sigma}^{-1}\mathbf{A}\big)\Big\}. \tag{25}$$

The functional form of (24) and (25) show that

$$\begin{aligned}\pi(\mathbf{B}|\boldsymbol{\Sigma},\mathbf{W}) &\propto N_{(mk+m)}\Big(\hat{\mathbf{B}},\boldsymbol{\Sigma}\otimes(\mathbf{X}'\mathbf{X})^{-1}\Big)\\ \pi(\boldsymbol{\Sigma}|\mathbf{W}) &\propto IW(n,\mathbf{A}),\end{aligned} \tag{26}$$

where $IW(.)$ stands for inverted Wishart distribution [19]. Thus, observations from the joint distribution $\pi(\mathbf{B},\boldsymbol{\Sigma}|\mathbf{W})$ can be drawn, iteratively, through the GDA algorithm.

The GDA Algorithm. The GDA algorithm has two main steps: (1) update the parameters' values from the posterior distributions, based on the data from the previous iterations and (2) use data augmentation (DA) algorithm (see sec. 3.2) to update the censored observations, based on the current parameters' values. The successive updating of the model parameters and censored observations will result in a sequence of random $m-$variate variables which converge to the joint posterior distribution of $\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\Sigma})$ [4],[5].

4 Simulation Study

To analyse the performance of the above procedures consider a bivariate censored LR model ($m=2$) with one predictor[3]. The datasets, of size $n=100, 500$ and 1000, are generated using two sets of regression coefficients $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$, each one combined with three different covariance matrices (low $\boldsymbol{\Sigma}^{(1)}$, moderate $\boldsymbol{\Sigma}^{(2)}$ and high correlation $\boldsymbol{\Sigma}^{(3)}$), as follows:

$$\boldsymbol{\beta}^{(1)}=\begin{bmatrix}2 & 1\\ 0.6 & 0.89\end{bmatrix} \text{ and } \boldsymbol{\beta}^{(2)}=\begin{bmatrix}0.2 & 0.3\\ 0.4 & 0.24\end{bmatrix}, \tag{27}$$

$$\boldsymbol{\Sigma}^{(1)}=\begin{bmatrix}2 & 0.1\\ 0.1 & 1.5\end{bmatrix}, \boldsymbol{\Sigma}^{(2)}=\begin{bmatrix}2 & -0.4\\ -0.4 & 1.5\end{bmatrix} \boldsymbol{\Sigma}^{(3)}=\begin{bmatrix}2 & 0.8\\ 0.8 & 1.5\end{bmatrix}. \tag{28}$$

Values of LOD (L_1 and L_2) were set so that the observed response variables $Y_{(1)}$ and $Y_{(2)}$ have five different pairwise levels of censorship: $A=(5\%,5\%), B=(5\%,20\%), C=(5\%,40\%), D=(20\%,20\%)$ and $E=(40\%,40\%)$. We generate 100 realizations of each of these 90 scenarios to assess the finite sample behaviour

[3] The generalisation of this study to more than one independent variable is trivial for DA and GDA. However, the computation of the EM estimates may be hindered by the need to obtain the moments of the truncated multivariate distributions.

of the estimates.

To illustrate the comparison between the methods, boxplots of biases corresponding to the three scenarios of censorship (low, medium and high) are represented in Figures 1 to 4.

The overall results, illustrated in Figures 1 (weak correlation) and 2 (strong correlation) indicate that the proposed methods produce approximately unbiased estimates for the regression parameters, $\boldsymbol{\beta}$, with decreasing variance as the sample size increases. However, as the correlation increases the estimates present slight bias specially for high censoring.

The DA and GDA approaches yield estimates for $\boldsymbol{\Sigma}$, illustrated in Figures 3 and 4, approximately unbiased and with decreasing variance as the sample size increases under weak correlation $\boldsymbol{\Sigma^{(1)}}$. Under high correlation, $\boldsymbol{\Sigma^{(3)}}$, and high censoring rate, the bias increases for all the approaches, with EM showing lower bias. The results indicate that $\boldsymbol{\Sigma}$ is under estimated in all scenarios but this does not affect the estimates of $\boldsymbol{\beta}$. This behaviour is expected since, in theory, the estimator of $\boldsymbol{\beta}$ is independent of the estimator of $\hat{\boldsymbol{\Sigma}}$ [11].

5 Final Remarks

One of the main features of multivariate LR is cross-correlation among the response variables. The censorship may distorts the correlation pattern in multivariate data. Then, in this work we propose three methods based on filling up data: EM, DA and GDA. Results from the simulation study show that both, DA and GDA estimates are consistent for low and moderate correlation.

This study has been conducted for the bivariate case. The main issue when considering $m > 2$ is related to the computation of the conditional expected value of the multivariate censored variable, needed to compute the EM estimates. Since general expressions for this conditional mean are given in [16] it is our aim to implement higher order cases in the future. Furthermore, we aim to develop methods for censored multivariate time series data.

Acknowledgements

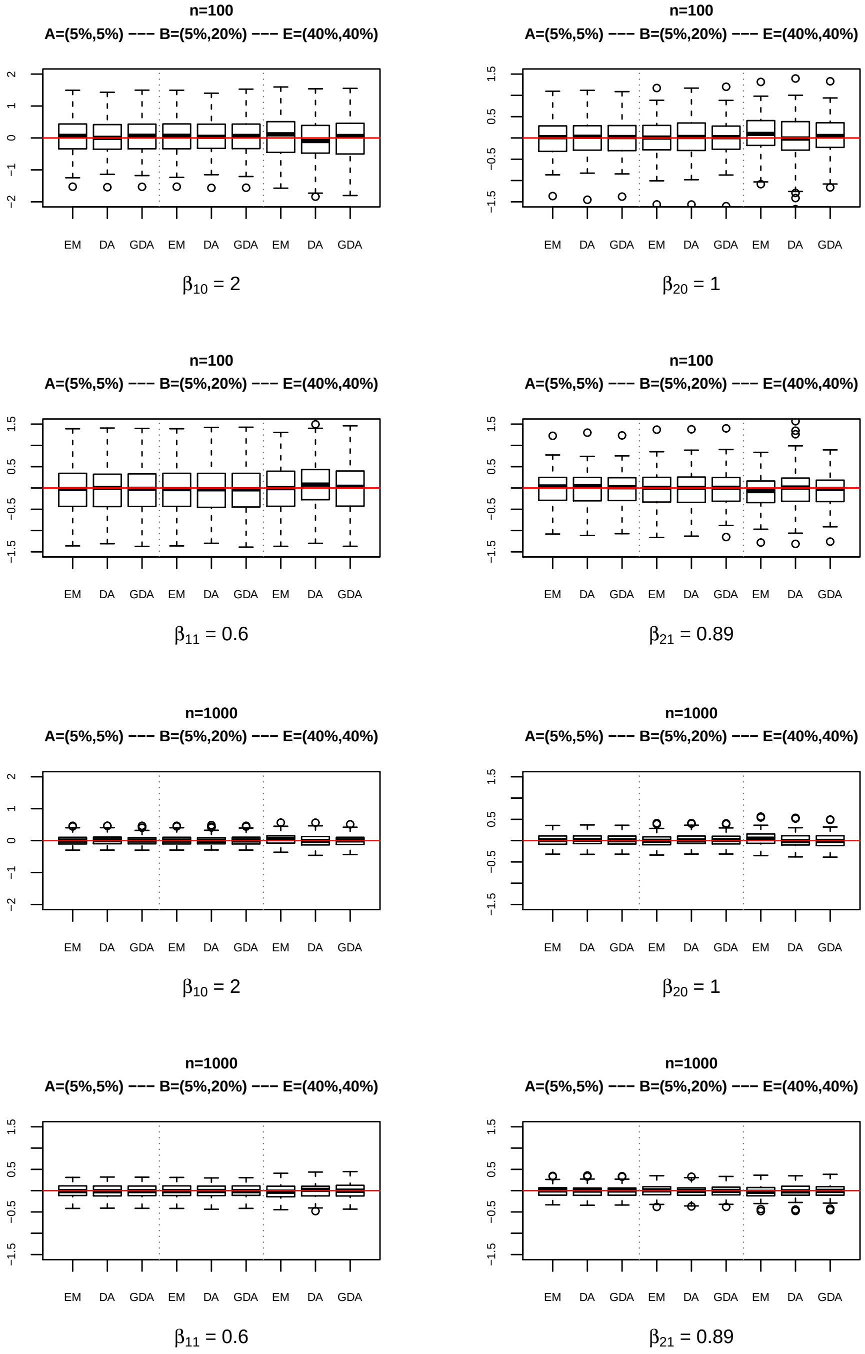
This work is supported by Fundação Calouste Gulbenkian and the Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), reference UIDB/04106/2020.

References

1. Alejo, J. and Montes-Rojas, G. : Quantile regression under limited dependent variable. arXiv:2112.06822v1

2. Anderson, T. W.: An Introduction to Multivariate Statistical Analysis, 3rd Ed. John Wiley & Son, New Jersey (2003).
3. Andersen, A. and Benn, C. and Jorgensen, M. and Ravn, H.: Censored correlated cytokine concentrations: multivariate tobit regression using clustered variance estimation. Statist. Med. 32, 2859–2874 (2013).
4. Breslaw, J. A.: Random sampling from a truncated multivariate normal distribution. Appl. Math. Lett. 7, No. 1, 1-6 (1994).
5. Chen, H. and Quandt, S. and Grzywacz, J. and Arcury, T.: A Bayesian multiple imputation method for handling longitudinal pesticide data with values below the limit of detection. Environmetrics. 24, 132–142 (2013).
6. Chib, S.: Bayes inference in the tobit censored regression model. J. of Econ. 51, 79–99 (1992).
7. Cohen, A.: Restriction and selection in samples from bivariate normal distributions. J. of the Ame. Stat. Ass. 50, No. 271, 884–893 (1955).
8. Dempster, A. and Laird, M. and Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Soc. 39, 1–38 (2013).
9. Hopke, P. and Liu, C. and Rubin, D.: Multiple imputation for multivariate data with missing and below-threshold measurements: Time-series concentrations of pollutants in the Arctic. Biometrics 57, 22–33 (2001).
10. Horrace, W.: Some results on the multivariate truncated normal distribution. J. Multiv. Ana. 94, 209–221 (2005).
11. Johnson, R. and Wichern, D.: Applied Multivariate Statistical Analysis. Prentice Hall, New Jersey (2007).
12. Lee, G. and Scott, C.: EM algorithms for multivariate Gaussian mixture models with truncated and censored data. Comp. Stat. and Data Ana. 56, 2816-2829 (2012).
13. Li, S. and Hu, T. and Tong, T. and Sun, J.: Semiparametric regression analysis of multivariate double censored data. Stat Model. x, 1-25 (2019).
14. Lockwood, J. R. and Schervish, M. J.: MCMC strategies for computing Bayesian predictive densities for censored multivariate data. J. Comp. and Gra. Stat. 14, No. 2, 395–414 (2005).
15. Muthen, B.: Moments of censored and truncated bivariate normal distribution. Bri. J. Math. and Stat. Psy. 43, 131–143 (1965).
16. Tallis, G.: The moment generating function of the truncated multi-normal distribution. J. Royal Stat. Soc. B. 23. No. 1, 223–229 (1961).
17. Tanner, M. and Wong, W.: The calculation of posterior distributions by data augmentation. J. Ame. Stat. Ass. 82, No. 398, 528–540 (1978).
18. Tiao, G. and Zellner, A.: On the Bayesian estimation of multivariate regression. J. Royal Stat. Soc. 26, 277–285 (1964).
19. Wishart, J.: The generalised product moment distribution in samples from a normal multivariate population. Biometrika. 20, 32–52 (1928).
20. Zeger, S. and Brookmeyer, R.: Regression analysis with censored autocorrelated data. J. of Am. Stat. Ass. 81, 722–729 (1986).

Fig. 1. Biases of $\hat{\boldsymbol{\beta}}^{(1)}$ based on data generated from the model with $\boldsymbol{\Sigma}^{(1)}$, for $n = 100$ (*top*) and $n = 1000$ (*down*).



$\beta_{10} = 2$

$\beta_{20} = 1$

$\beta_{11} = 0.6$

$\beta_{21} = 0.89$

$\beta_{10} = 2$

$\beta_{20} = 1$

$\beta_{11} = 0.6$

$\beta_{21} = 0.89$

Fig. 2. Biases of $\hat{\boldsymbol{\beta}}^{(1)}$ based on data generated from the model with $\boldsymbol{\Sigma}^{(3)}$, for $n = 100$ (*top*) and $n = 1000$ (*down*).

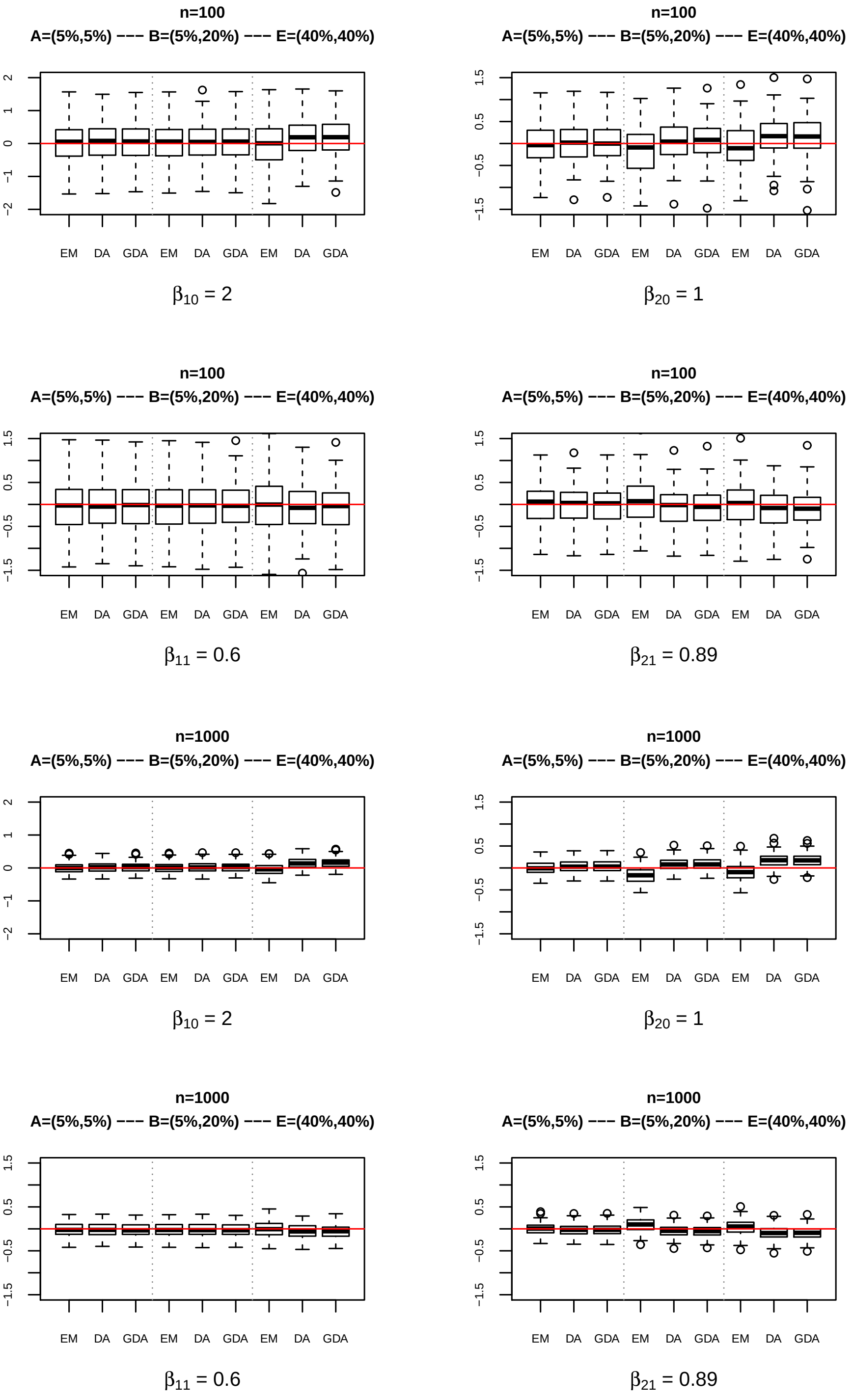


Fig. 3. Biases of $\hat{\boldsymbol{\Sigma}}^{(1)}$ based on data generated from the model with $\boldsymbol{\beta}^{(1)}$, for $n = 100$ (*top*) and $n = 1000$ (*down*).

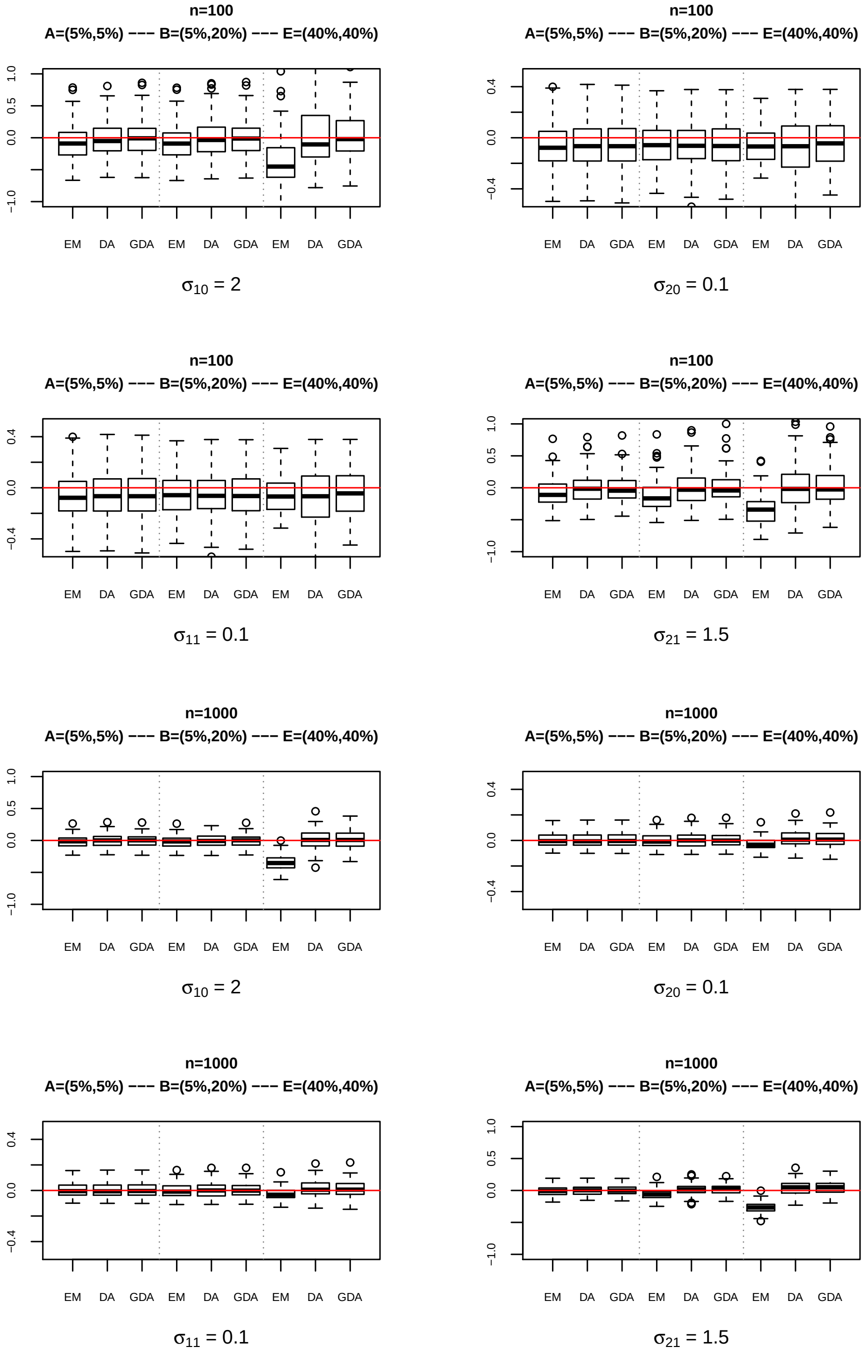
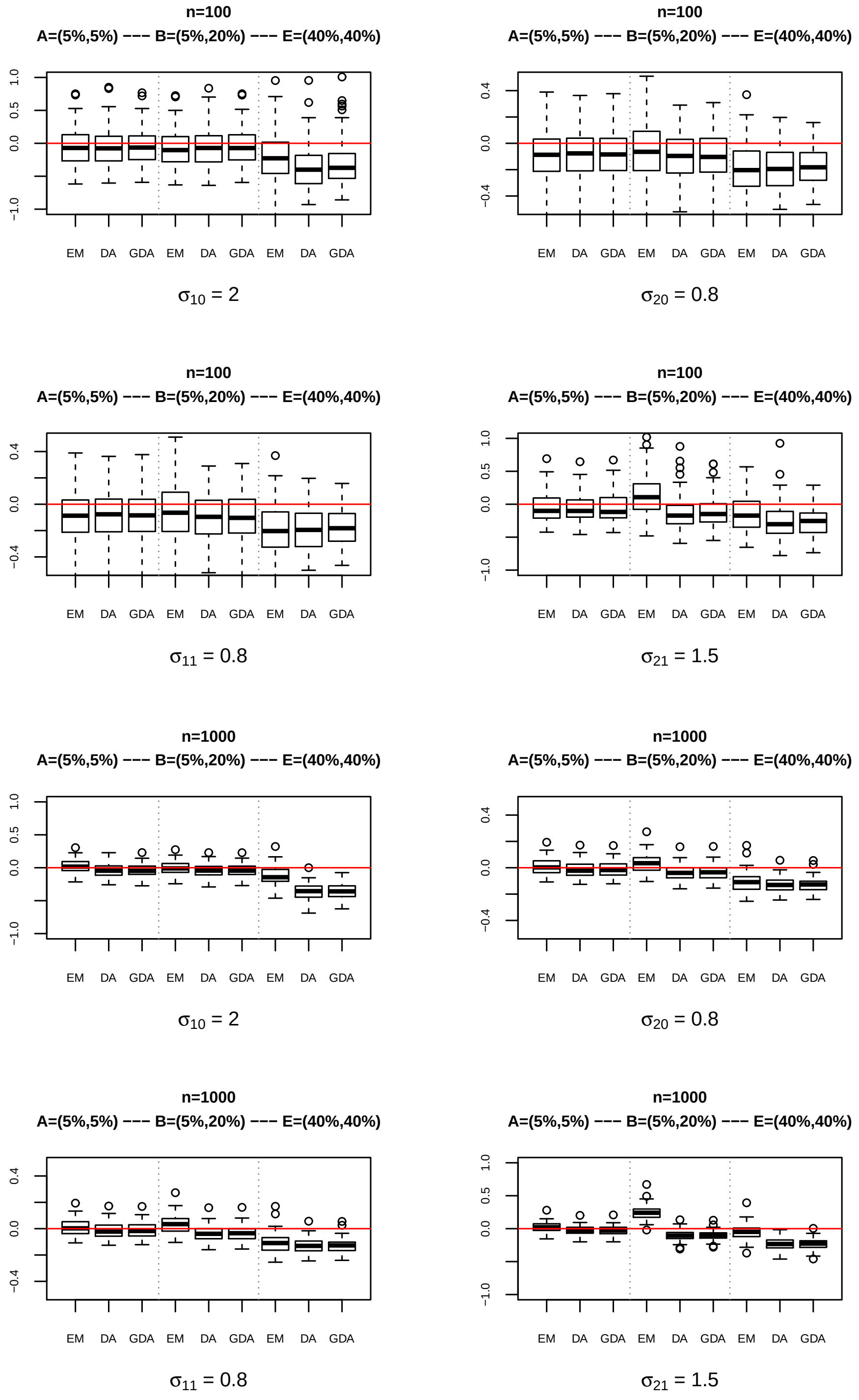


Fig. 4. Biases of $\hat{\boldsymbol{\Sigma}}^{(3)}$ based on data generated from the model with $\boldsymbol{\beta}^{(1)}$, for $n = 100$ (*top*) and $n = 1000$ (*down*).



$\sigma_{10} = 2$

$\sigma_{20} = 0.8$

$\sigma_{11} = 0.8$

$\sigma_{21} = 1.5$

$\sigma_{10} = 2$

$\sigma_{20} = 0.8$

$\sigma_{11} = 0.8$

$\sigma_{21} = 1.5$