

# $\ell_1$ -Penalized censored Gaussian graphical model

LUIGI AUGUGLIARO\*, ANTONINO ABBRUZZO

*Department of Economics, Business and Statistics, University of Palermo, Building 13,  
Viale delle Scienze, 90128 Palermo, Italy  
luigi.augugliaro@unipa.it*

VERONICA VINCIOTTI

*Department of Mathematics, Brunel University London, Kingston Lane, Uxbridge UB8 3PH, UK*

## SUMMARY

Graphical lasso is one of the most used estimators for inferring genetic networks. Despite its diffusion, there are several fields in applied research where the limits of detection of modern measurement technologies make the use of this estimator theoretically unfounded, even when the assumption of a multivariate Gaussian distribution is satisfied. Typical examples are data generated by polymerase chain reactions and flow cytometer. The combination of censoring and high-dimensionality make inference of the underlying genetic networks from these data very challenging. In this article, we propose an  $\ell_1$ -penalized Gaussian graphical model for censored data and derive two EM-like algorithms for inference. We evaluate the computational efficiency of the proposed algorithms by an extensive simulation study and show that, when censored data are available, our proposal is superior to existing competitors both in terms of network recovery and parameter estimation. We apply the proposed method to gene expression data generated by microfluidic Reverse Transcription quantitative Polymerase Chain Reaction technology in order to make inference on the regulatory mechanisms of blood development. A software implementation of our method is available on github (<https://github.com/LuigiAugugliaro/cglasso>).

**Keywords:** Censored data; Expectation-Maximization algorithm; Gaussian graphical model; Graphical Lasso; High-dimensional data.

## 1. INTRODUCTION

An important aim in genomics is to understand interactions among genes, characterized by the regulation and synthesis of proteins under internal and external signals. These relationships can be represented by a genetic network, i.e. a graph where nodes represent genes and edges describe the interactions among them. Genetic networks can be used to gain new insights into the activity of biological pathways and to deduce unknown functions of genes from their dependence on other genes.

Gaussian graphical models (Lauritzen, 1996) have been widely used for reconstructing a genetic network from expression data. The reason of their widespread use relies on the statistical properties of the multivariate Gaussian distribution, which allow the topological structure of a network to be related to the non-zero elements of the concentration matrix, i.e. the inverse of the covariance matrix. Thus, the problem

\*To whom correspondence should be addressed.

of network inference can be recast as the problem of estimating a concentration matrix. The graphical lasso (Yuan and Lin, 2007) is a popular method for estimating a sparse concentration matrix, based on the idea of adding an  $\ell_1$ -penalty to the likelihood function of the multivariate Gaussian distribution. Nowadays, this estimator is widely used in applied research (e.g. Menéndez and others, 2010; Vinciotti and others, 2016) and widely studied in the computational as well as theoretical (e.g. Bickel and Levina, 2008; Friedman and others, 2008; Witten and others, 2011) literature. The interested reader is referred to Augugliaro and others (2016) for an extensive review.

Despite the widespread literature on the graphical lasso estimator, there is a great number of fields in applied research where modern measurement technologies make the use of this graphical model theoretically unfounded, even when the assumption of a multivariate Gaussian distribution is satisfied. A first example of this is Reverse Transcription quantitative Polymerase Chain Reaction (RT-qPCR), a popular technology for gene expression profiling (Derveaux and others, 2010). This technique is used to measure the expression of a set of target genes in a given sample through repeated cycles of sequence-specific DNA amplification followed by expression measurements. The cycle at which the observed expression first exceeds a fixed threshold is commonly called the cycle-threshold (McCall and others, 2014). If a target is not expressed, the threshold is not reached after the maximum number of cycles (limit of detection) and the corresponding cycle-threshold is undetermined. For this reason, the resulting data is naturally right-censored (McCall and others, 2014; Pipelers and others, 2017). Another example is given by the flow cytometer, which is an essential tool in the diagnosis of diseases such as acute leukemias and malignant lymphomas (Brown and Wittwer, 2000). A flow cytometer measures a limited range of signal strength and records each marker value within a fixed range, such as between 0 and 1023. If a measurement falls outside this range, then the value is replaced by the nearest legitimate value; that is, a value smaller than 0 is censored to 0 and a value larger than 1023 is censored to 1023. A direct application of the graphical lasso for network inference from data such as these is theoretically unfounded since it does not consider the effects of the censoring mechanism on the estimator of the concentration matrix.

In this article, we propose an extension of the graphical lasso estimator that takes into account the censoring mechanism of the data explicitly. Our work can be related to Städler and Bühlmann (2012), who propose an  $\ell_1$ -penalized estimator of the inverse covariance matrix of a multivariate Gaussian model based on the assumption that the data are missing at random. As we shall see in the following of this article, failure to take into account the censoring mechanism causes a poor behavior of this approach when compared with our proposal. Our proposal can also be related to the work of Perkins and others (2013), Hoffman and Johnson (2015), and Pesonen and others (2015), who provide a maximum likelihood estimator of the covariance matrix under left-censoring. However, these works do not address the estimation of the precision matrix under a sparsity assumption and for this reason they are applicable only when the sample size is larger than the number of nodes in the network.

The remaining part of this article is structured as follows. In Section 2, we extend the notion of Gaussian graphical model to censored data and in Section 3, we propose the extension of the graphical lasso estimator to a censored graphical lasso estimator and two Expectation-Maximization (EM) algorithms for inference of parameters in the censored Gaussian graphical model. Section 4 is devoted to the evaluation of the behavior of the proposed algorithms and the comparison of the proposed estimator with existing competitors. Finally, in Section 5, we study a real dataset and in Section 6, we draw some conclusions.

## 2. THE CENSORED GAUSSIAN GRAPHICAL MODEL

In order to describe the technical details of the proposed method, let  $\mathbf{X} = (X_1, \dots, X_p)^\top$  be a  $p$ -dimensional random vector. Graphical models allow to represent the set of conditional independencies among these random variables by a graph  $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ , where  $\mathcal{V}$  is the set of nodes associated to  $\mathbf{X}$  and  $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$

is the set of ordered pairs, called edges, representing the conditional dependencies among the  $p$  random variables (Lauritzen, 1996).

The Gaussian graphical model is a member of this class of models based on the assumption that  $\mathbf{X}$  follows a multivariate Gaussian distribution with expected value  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)^\top$  and covariance matrix  $\Sigma = (\sigma_{hk})$ . Denoting with  $\Theta = (\theta_{hk})$  the concentration matrix, i.e. the inverse of the covariance matrix, the density function of  $\mathbf{X}$  can be written as

$$\phi(\mathbf{x}; \boldsymbol{\mu}, \Theta) = (2\pi)^{-p/2} |\Theta|^{1/2} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \Theta (\mathbf{x} - \boldsymbol{\mu}) \right\}. \quad (2.1)$$

As shown in Lauritzen (1996), the pattern of non-zero elements in  $\Theta$  defines the corresponding graph, namely the undirected edge  $(h, k)$  is an element of the edge set  $\mathcal{E}$  of the corresponding conditional independence graph if and only if  $\theta_{hk} \neq 0$ .

Let  $\mathbf{X}$  be a (partially) latent random vector with density function (2.1). In order to include the censoring mechanism inside our framework, let us denote by  $\mathbf{l} = (l_1, \dots, l_p)^\top$  and  $\mathbf{u} = (u_1, \dots, u_p)^\top$ , with  $l_h < u_h$  for  $h = 1, \dots, p$ , the vectors of known left and right censoring values. Thus,  $X_h$  is observed only if it is inside the interval  $[l_h, u_h]$  otherwise it is censored from below if  $X_h < l_h$  or censored from above if  $X_h > u_h$ . Under this setting, a rigorous definition of the joint distribution of the observed data can be obtained using the approach for missing data with ignorable mechanism (Little and Rubin, 2002). This requires the specification of the distribution of a  $p$ -dimensional random vector, denoted by  $R(\mathbf{X}; \mathbf{l}, \mathbf{u})$ , used to encode the censoring patterns. Formally, the  $h$ th element of  $R(\mathbf{X}; \mathbf{l}, \mathbf{u})$  is defined as  $R(X_h; l_h, u_h) = I(X_h > u_h) - I(X_h < l_h)$ , where  $I(\cdot)$  denotes the indicator function. By construction  $R(\mathbf{X}; \mathbf{l}, \mathbf{u})$  is a discrete random vector with support in the set  $\{-1, 0, 1\}^p$  and probability function

$$\text{Prob}\{R(\mathbf{X}; \mathbf{l}, \mathbf{u}) = \mathbf{r}\} = \int_{D_r} \phi(\mathbf{x}; \boldsymbol{\mu}, \Theta) d\mathbf{x},$$

where  $D_r = \{\mathbf{x} \in \mathbb{R}^p : R(\mathbf{x}; \mathbf{l}, \mathbf{u}) = \mathbf{r}\}$ .

Given a censoring pattern, we can simplify our notation by partitioning the set  $\mathcal{I} = \{1, \dots, p\}$  into the sets  $o = \{h \in \mathcal{I} : r_h = 0\}$ ,  $c^- = \{h \in \mathcal{I} : r_h = -1\}$  and  $c^+ = \{h \in \mathcal{I} : r_h = +1\}$  and, in the following of this article, we shall use the convention that a vector indexed by a set of indices denotes the corresponding subvector. For example, the subvector of observed elements in  $\mathbf{x}$  is denoted by  $\mathbf{x}_o = (x_h)_{h \in o}$  and, consequently, the observed data is the vector  $(\mathbf{x}_o^\top, \mathbf{r}^\top)^\top$ . As explained in Little and Rubin (2002), the joint probability distribution of the observed data, denoted by  $\varphi(\mathbf{x}_o, \mathbf{r}; \boldsymbol{\mu}, \Theta)$ , is obtained by integrating  $\mathbf{X}_{c^+}$  and  $\mathbf{X}_{c^-}$  out of the joint distribution of  $\mathbf{X}$  and  $R(\mathbf{X}; \mathbf{l}, \mathbf{u})$ , which can be written as the product of the density function (2.1) and the conditional distribution of  $R(\mathbf{X}; \mathbf{l}, \mathbf{u})$  given  $\mathbf{X} = \mathbf{x}$ . As shown in Schafer (1997), this results in

$$\begin{aligned} \varphi(\mathbf{x}_o, \mathbf{r}; \boldsymbol{\mu}, \Theta) &= \int_{\mathbf{u}_{c^+}}^{+\infty} \int_{-\infty}^{l_{c^-}} \phi(\mathbf{x}_o, \mathbf{x}_{c^-}, \mathbf{x}_{c^+}; \boldsymbol{\mu}, \Theta) d\mathbf{x}_{c^-} d\mathbf{x}_{c^+} I(l_o \leq \mathbf{x}_o \leq \mathbf{u}_o) \\ &= \int_{D_c} \phi(\mathbf{x}_o, \mathbf{x}_c; \boldsymbol{\mu}, \Theta) d\mathbf{x}_c I(l_o \leq \mathbf{x}_o \leq \mathbf{u}_o), \end{aligned} \quad (2.2)$$

where  $c = c^- \cup c^+$  and  $D_c = (-\infty, l_{c^-}) \times (\mathbf{u}_{c^+}, +\infty)$ . The density function (2.2) is used in Lee and Scott (2012) inside the framework of a mixture of multivariate Gaussian distributions with censored data. Using (2.2) the censored Gaussian graphical model can be formally defined.

**DEFINITION 1** Let  $\mathbf{X}$  be a  $p$ -dimensional Gaussian distribution whose density function  $\phi(\mathbf{x}; \boldsymbol{\mu}, \Theta)$  factorizes according to an undirected graph  $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$  and let  $R(\mathbf{X}; \mathbf{l}, \mathbf{u})$  be a  $p$ -dimensional random censoring-data indicator defined by the censoring vectors  $\mathbf{l}$  and  $\mathbf{u}$ . The censored Gaussian graphical model (cGGM) is defined to be the set  $\{\mathbf{X}, R(\mathbf{X}; \mathbf{l}, \mathbf{u}), \varphi(\mathbf{x}_o, \mathbf{r}; \boldsymbol{\mu}, \Theta), \mathcal{G}\}$ .

A closer look at Definition 1 reveals that the proposed notion of censored Gaussian graphical model is characterized by a high degree of generality since it covers also the special cases of the classical Gaussian graphical model ( $l_h = -\infty, u_h = +\infty$  for any  $h$ ), and the cases in which there is only left-censored data ( $u_h = +\infty$  for any  $h$ ) or right-censored data ( $l_h = -\infty$  for any  $h$ ).

### 3. $\ell_1$ -PENALIZED ESTIMATOR FOR CENSORED GAUSSIAN GRAPHICAL MODEL

#### 3.1. The censored graphical lasso estimator

Consider a sample of  $n$  independent observations drawn from the censored Gaussian graphical model  $\{\mathbf{X}, R(\mathbf{X}; \mathbf{l}, \mathbf{u}), \varphi(\mathbf{x}_o, \mathbf{r}; \boldsymbol{\mu}, \Theta), \mathcal{G}\}$ . For ease of exposition, we shall assume that  $\mathbf{l}$  and  $\mathbf{u}$  are fixed across the  $n$  observations, but the extension to the cases where the censoring vectors are specific to each observation is straightforward and does not require a specific treatment. If these values are unknown we suggest to use the smallest and largest observed value for each variable as possible estimates for  $l_h$  and  $u_h$ , respectively.

Let  $\mathbf{r}_i$  be the  $i$ th realization of the random vector  $R(\mathbf{X}_i; \mathbf{l}, \mathbf{u})$ . Then the  $i$ th observed data is the vector  $(\mathbf{x}_{io_i}^\top, \mathbf{r}_i^\top)^\top$ , with  $o_i = \{h \in \mathcal{I} : r_{ih} = 0\}$ . Using the density function (2.2), the observed log-likelihood function can be written as

$$\ell(\boldsymbol{\mu}, \Theta) = \sum_{i=1}^n \log \int_{D_{c_i}} \phi(\mathbf{x}_{io_i}, \mathbf{x}_{ic_i}; \boldsymbol{\mu}, \Theta) d\mathbf{x}_{ic_i} = \sum_{i=1}^n \log \varphi(\mathbf{x}_{io_i}, \mathbf{r}_i; \boldsymbol{\mu}, \Theta), \quad (3.1)$$

where, as before,  $c_i = c_i^- \cup c_i^+$ , with  $c_i^- = \{h \in \mathcal{I} : r_{ih} = -1\}$  and  $c_i^+ = \{h \in \mathcal{I} : r_{ih} = +1\}$ , and  $D_{c_i} = (-\infty, \mathbf{l}_{c_i^-}) \times (\mathbf{u}_{c_i^+}, +\infty)$ . Although inference about the parameters of this model can be carried out via the maximum likelihood method, the application of this inferential procedure to real datasets, such as the gene expression data described in Section 5, is limited for three main reasons. Firstly, the number of measured variables is larger than the sample size, and this implies the non-existence of the maximum likelihood estimator even when the dataset is fully observed. Secondly, even when the sample size is large enough, the maximum likelihood estimator will exhibit a very high variance (Uhler, 2012). Thirdly, empirical evidence suggests that gene networks or more general biochemical networks are not fully connected (Gardner and others, 2003). In terms of Gaussian graphical models this evidence translates in the assumption that  $\Theta$  has a sparse structure, i.e. only few  $\theta_{hk}$  are different from zero, which is not obtained by a direct (or indirect) maximization of the observed log-likelihood function (3.1).

All that considered, in this article, we propose to estimate the parameters of the censored Gaussian graphical model by generalizing the approach proposed in Yuan and Lin (2007), i.e. by maximizing a new objective function defined by adding a lasso-type penalty function to the observed log-likelihood (3.1). The resulting estimator, called censored graphical lasso (cglasso), is formally defined as

$$\{\hat{\boldsymbol{\mu}}^\rho, \hat{\Theta}^\rho\} = \arg \max_{\boldsymbol{\mu}, \Theta \succ 0} \frac{1}{n} \sum_{i=1}^n \log \varphi(\mathbf{x}_{io_i}, \mathbf{r}_i; \boldsymbol{\mu}, \Theta) - \rho \sum_{h \neq k} |\theta_{hk}|. \quad (3.2)$$

Like in the standard graphical lasso estimator, the nonnegative tuning parameter  $\rho$  is used to control the amount of sparsity in the estimated concentration matrix  $\hat{\Theta}^\rho = (\hat{\theta}_{hk}^\rho)$  and, consequently, in the corresponding estimated graph  $\hat{\mathcal{G}}^\rho = \{\mathcal{V}, \hat{\mathcal{E}}^\rho\}$ , where  $\hat{\mathcal{E}}^\rho = \{(h, k) : \hat{\theta}_{hk}^\rho \neq 0\}$ . When  $\rho$  is large enough, some  $\hat{\theta}_{hk}^\rho$  are

shrunk to zero resulting in the removal of the corresponding link in  $\widehat{\mathcal{G}}^\rho$ ; on the other hand, when  $\rho$  is equal to zero and the sample size is large enough the estimator  $\widehat{\Theta}^\rho$  coincides with the maximum likelihood estimator of the concentration matrix, which implies a fully connected estimated graph.

### 3.2. Fitting the censored graphical lasso model

Using known results about the multivariate Gaussian distribution, the conditional distribution of  $\mathbf{X}_{c_i}$  given  $\mathbf{X}_{o_i} = \mathbf{x}_{o_i}$  is also a multivariate Gaussian distribution with concentration matrix  $\Theta_{c_i c_i} = (\theta_{hk})_{h,k \in c_i}$  and conditional expected value equal to  $E_{c_i|o_i}(\mathbf{X}_{c_i}) = \boldsymbol{\mu}_{c_i|o_i} = \boldsymbol{\mu}_{c_i} - \Theta_{c_i c_i}^{-1} \Theta_{c_i o_i}(\mathbf{x}_{o_i} - \boldsymbol{\mu}_{o_i})$ , where  $\Theta_{c_i o_i} = (\theta_{hk})_{h \in c_i, k \in o_i}$ . As we shall show in the next theorem, the conditions characterizing the cglasso estimator are based on the first and second moment of the Gaussian distribution  $\phi(\mathbf{x}_{c_i}; \boldsymbol{\mu}_{c_i|o_i}, \Theta_{c_i c_i})$  truncated over the region  $D_{c_i}$ . To this end, we let

$$x_{i,h}(\boldsymbol{\mu}, \Theta) = \begin{cases} x_{ih} & \text{if } r_{ih} = 0 \\ E_{c_i|o_i}(X_{ih} | \mathbf{X}_{ic_i} \in D_{c_i}) & \text{otherwise,} \end{cases}$$

$$x_{i,hk}(\boldsymbol{\mu}, \Theta) = \begin{cases} x_{ih}x_{ik} & \text{if } r_{ih} = 0 \text{ and } r_{ik} = 0 \\ x_{ih}E_{c_i|o_i}(X_{ik} | \mathbf{X}_{ic_i} \in D_{c_i}) & \text{if } r_{ih} = 0 \text{ and } r_{ik} \neq 0 \\ E_{c_i|o_i}(X_{ih} | \mathbf{X}_{ic_i} \in D_{c_i})x_{ik} & \text{if } r_{ih} \neq 0 \text{ and } r_{ik} = 0 \\ E_{c_i|o_i}(X_{ih}X_{ik} | \mathbf{X}_{ic_i} \in D_{c_i}) & \text{if } r_{ih} \neq 0 \text{ and } r_{ik} \neq 0, \end{cases}$$

where  $E_{c_i|o_i}(\cdot | \mathbf{X}_{ic_i} \in D_{c_i})$  denotes the expected value computed using the conditional distribution of  $\mathbf{X}_{ic_i}$  given  $\mathbf{x}_{io_i}$  truncated over  $D_{c_i}$ . Finally, we let  $\bar{x}_h(\boldsymbol{\mu}, \Theta) = \sum_{i=1}^n x_{i,h}(\boldsymbol{\mu}, \Theta)/n$ ,  $\bar{\mathbf{x}}(\boldsymbol{\mu}, \Theta) = \{\bar{x}_1(\boldsymbol{\mu}, \Theta), \dots, \bar{x}_p(\boldsymbol{\mu}, \Theta)\}^\top$ ,  $s_{hk}(\boldsymbol{\mu}, \Theta) = \sum_{i=1}^n x_{i,hk}(\boldsymbol{\mu}, \Theta)/n - \bar{x}_h(\boldsymbol{\mu}, \Theta)\bar{x}_k(\boldsymbol{\mu}, \Theta)$ , and  $S(\boldsymbol{\mu}, \Theta) = \{s_{hk}(\boldsymbol{\mu}, \Theta)\}$ . Using this notation, Theorem 1 gives the Karush–Kuhn–Tucker conditions for the proposed estimator.

**THEOREM 1** Necessary and sufficient conditions for  $\{\hat{\boldsymbol{\mu}}^\rho, \hat{\Theta}^\rho\}$  to be the solution of the maximization problem

$$\max_{\boldsymbol{\mu}, \Theta \succ 0} \frac{1}{n} \sum_{i=1}^n \log \varphi(\mathbf{x}_{io_i}, \mathbf{r}_i; \boldsymbol{\mu}, \Theta) - \rho \sum_{h \neq k} |\theta_{hk}|,$$

are

$$\left. \begin{aligned} \bar{x}_h(\hat{\boldsymbol{\mu}}^\rho, \hat{\Theta}^\rho) - \hat{\mu}_h^\rho &= 0 \\ \hat{\sigma}_{hk}^\rho(\hat{\boldsymbol{\mu}}^\rho, \hat{\Theta}^\rho) - s_{hk}(\hat{\boldsymbol{\mu}}^\rho, \hat{\Theta}^\rho) - \rho \hat{v}_{hk} &= 0 \end{aligned} \right\} \quad (3.3)$$

where  $\hat{v}_{hk}$  denotes the subgradient of the absolute value function at  $\hat{\theta}_{hk}^\rho$ , i.e.,  $\hat{v}_{hk} = \text{sign}(\hat{\theta}_{hk}^\rho)$  if  $\hat{\theta}_{hk}^\rho \neq 0$  and  $|\hat{v}_{hk}| \leq 1$  if  $\hat{\theta}_{hk}^\rho = 0$ .

A proof of Theorem 1 is reported in the [supplementary material](#) available at *Biostatistics* online. The stationary conditions (3.3) show that, while in the standard graphical lasso the parameter  $\boldsymbol{\mu}$  can be estimated by the empirical average regardless of  $\rho$  and the inference about the concentration matrix can be carried out using the profile log-likelihood function, inside our framework the two inferential problems cannot be separated, since the tuning parameter also affects the estimator of the expected value. Furthermore, the conditions suggest that, for a given value of the tuning parameter, the cglasso estimator can be computed

using the EM algorithm (Dempster and others, 1977). This algorithm is based on the idea of repeating two steps until a convergence criterion is met. The first step, called E-Step, requires the calculation of the conditional expected value of the complete log-likelihood function using the current estimates. The resulting function, called  $Q$ -function, is maximized in the second step, i.e. the M-Step. As explained in McLachlan and Krishnan (2008), the EM algorithm can be significantly simplified when the complete probability density function is a member of the regular exponential family. In this case, the E-Step simply requires the computation of the conditional expected values of the sufficient statistics. In our case, if we denote by  $\{\hat{\mu}_{ini}^\rho, \hat{\Theta}_{ini}^\rho\}$  an initial estimate of the parameters, the E-Step reduces to the computation of  $x_{i,h}(\hat{\mu}_{ini}^\rho, \hat{\Theta}_{ini}^\rho)$  and  $x_{i,hk}(\hat{\mu}_{ini}^\rho, \hat{\Theta}_{ini}^\rho)$ , for  $i = 1, \dots, n$ . From this, conditions (3.3) can be written as

$$\left. \begin{aligned} \bar{x}_h(\hat{\mu}_{ini}^\rho, \hat{\Theta}_{ini}^\rho) - \hat{\mu}_h^\rho &= 0 \\ \hat{\sigma}_{hk}^\rho(\hat{\mu}^\rho, \hat{\Theta}^\rho) - s_{hk}(\hat{\mu}_{ini}^\rho, \hat{\Theta}_{ini}^\rho) - \rho \hat{v}_{hk} &= 0 \end{aligned} \right\} \quad (3.4)$$

which are the stationary conditions of a standard graphical lasso problem (Witten and others, 2011) with  $S(\hat{\mu}_{ini}^\rho, \hat{\Theta}_{ini}^\rho)$  used as the current estimate of the empirical covariance matrix. The conditions (3.4) imply that in the M-Step the parameter  $\mu$  is estimated by  $\bar{x}(\hat{\mu}_{ini}^\rho, \hat{\Theta}_{ini}^\rho)$ , while  $\Theta$  is estimated by solving the following maximization problem:

$$\max_{\Theta > 0} Q(\Theta | \hat{\Theta}_{ini}^\rho) = \max_{\Theta > 0} \log \det \Theta - \text{tr}\{\Theta S(\hat{\mu}_{ini}^\rho, \hat{\Theta}_{ini}^\rho)\} - \rho \sum_{h \neq k} |\theta_{hk}|, \quad (3.5)$$

which is a standard graphical lasso problem. The following steps summarize the proposed EM algorithm for the derivation of the cglasso estimator:

1. Let  $\{\hat{\mu}_{ini}^\rho, \hat{\Theta}_{ini}^\rho\}$  be initial estimates;
2. E-step: Compute  $\bar{x}(\hat{\mu}_{ini}^\rho, \hat{\Theta}_{ini}^\rho)$  and  $S(\hat{\mu}_{ini}^\rho, \hat{\Theta}_{ini}^\rho)$ ;
3. M-step: Let  $\hat{\mu}^\rho = \bar{x}(\hat{\mu}_{ini}^\rho, \hat{\Theta}_{ini}^\rho)$  and solve the problem (3.5) using  $S(\hat{\mu}_{ini}^\rho, \hat{\Theta}_{ini}^\rho)$ ;
4. If a convergence criterion is met then return  $\{\hat{\mu}^\rho, \hat{\Theta}^\rho\}$  else let  $\hat{\mu}_{ini}^\rho = \hat{\mu}^\rho$  and  $\hat{\Theta}_{ini}^\rho = \hat{\Theta}^\rho$ ;
5. Repeat steps 2–4.

Although the maximization problem (3.5) can be efficiently solved using, for example, the algorithm proposed by Friedman and others (2008), Rothman and others (2008), or Witten and others (2011), the previous steps reveal that the main computational cost of the proposed EM algorithm comes from the evaluation of the moments of the multivariate truncated Gaussian distribution, which are needed to compute  $\bar{x}(\hat{\mu}_{ini}^\rho, \hat{\Theta}_{ini}^\rho)$  and  $S(\hat{\mu}_{ini}^\rho, \hat{\Theta}_{ini}^\rho)$ . These can be computed using the methods proposed by Lee (1983), Leppard and Tallis (1989), or Arismendi (2013), but these methods require complex numerical algorithms for the calculation of the integral of the multivariate normal density function (see for a review, Genz and Bretz, 2002), and they are computationally infeasible also for moderate size problems. A possible solution is to approximate the moments using Monte Carlo methods. Our preliminary study, however, shows that the Monte Carlo error causes an increment in the number of EM steps required for convergence of the proposed algorithm, thus removing the gain from using the faster approach for the calculation of the moments. Taking all of this into consideration, we propose a different approximate EM algorithm. In particular, following the idea proposed by Guo and others (2015), we approximate the quantities  $E_{c_i|o_i}(X_{ih}X_{ik} | X_{ic_i} \in D_{c_i})$ , for any  $h \neq k$ , by:

$$E_{c_i|o_i}(X_{ih}X_{ik} | X_{ic_i} \in D_{c_i}) \approx E_{c_i|o_i}(X_{ih} | X_{ic_i} \in D_{c_i})E_{c_i|o_i}(X_{ik} | X_{ic_i} \in D_{c_i}). \quad (3.6)$$



As shown by [Guo and others \(2015\)](#), the approach works well when the real concentration matrix is sparse or the tuning parameter is sufficiently large. The main advantage coming from the approximation (3.6) is that now, in order to evaluate the quantities  $\bar{\mathbf{x}}(\hat{\boldsymbol{\mu}}_{ini}^\rho, \hat{\Theta}_{ini}^\rho)$  and  $S(\hat{\boldsymbol{\mu}}_{ini}^\rho, \hat{\Theta}_{ini}^\rho)$ , we only need to compute  $E_{c_i|o_i}(X_{ih} | \mathbf{X}_{ic_i} \in D_{c_i})$  and  $E_{c_i|o_i}(X_{ih}^2 | \mathbf{X}_{ic_i} \in D_{c_i})$ , for  $h = 1, \dots, p$ . These quantities can be computed by using exact formulas ([Johnson and others, 1994](#)) and require only the evaluation of the cumulative distribution function of the univariate Gaussian distribution. In the following of this article, we denote by  $\bar{S}(\hat{\boldsymbol{\mu}}_{ini}^\rho, \hat{\Theta}_{ini}^\rho)$  the approximate estimate of the empirical covariance matrix resulting from this approach. By an extensive simulation study, in Section 4.1, we shall study the behavior of the proposed algorithm based on the usage of the matrix  $\bar{S}(\hat{\boldsymbol{\mu}}_{ini}^\rho, \hat{\Theta}_{ini}^\rho)$  in steps 2 and 3.

As with graphical lasso, our proposed method requires a sequence of  $\rho$ -values, which should be suitably defined so as to reduce the computational cost needed to compute the entire path of the estimated parameters. Theorem 2 gives the exact formula for the derivation of the largest  $\rho$ -value, denoted by  $\rho_{\max}$ , and the corresponding cglasso estimator.

**THEOREM 2** For any index  $h$  define the sets  $o_h = \{i : r_{ih} = 0\}$ ,  $c_h^- = \{i : r_{ih} = -1\}$  and  $c_h^+ = \{i : r_{ih} = +1\}$  and compute the marginal maximum likelihood estimates  $\{\hat{\mu}_h, \hat{\sigma}_h^2\}$  maximizing:

$$\ell(\mu_h, \sigma_h^2) = \sum_{i \in o_h} \log \phi(x_{ih}; \mu_h, \sigma_h^2) + |c_h^-| \log \int_{-\infty}^{l_h} \phi(x; \mu_h, \sigma_h^2) dx + |c_h^+| \log \int_{u_h}^{+\infty} \phi(x; \mu_h, \sigma_h^2) dx.$$

Then  $\rho_{\max} = \max_{h \neq k} |s_{hk}(\hat{\boldsymbol{\mu}}, \hat{\Theta})|$ , where  $s_{hk}(\hat{\boldsymbol{\mu}}, \hat{\Theta})$  are the elements of the matrix  $S(\hat{\boldsymbol{\mu}}, \hat{\Theta})$  computed using  $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \dots, \hat{\mu}_p)^\top$  and  $\hat{\Theta} = \text{diag}(\hat{\sigma}_1^{-2}, \dots, \hat{\sigma}_p^{-2})$ . Furthermore,  $\hat{\boldsymbol{\mu}}$  and  $\hat{\Theta}$  are the corresponding cglasso estimators.

A proof of Theorem 2 is reported in the [supplementary material](#) available at *Biostatistics* online. Theorem 2 shows how the maximum value of  $\rho$  can be efficiently computed using  $p$  optimization problems which do not involve the tuning parameter, and then calculating  $S(\hat{\boldsymbol{\mu}}, \hat{\Theta})$ . Once  $\rho_{\max}$  is calculated, the entire path of the cglasso estimates is calculated as follow:

1. Compute  $\rho_{\max}$  as specified in Theorem 2 and let  $\rho_{\min}$  be the smallest  $\rho$ -value;
2. Compute a decreasing sequence  $\{\rho_{(k)}\}_{k=1}^K$  of distinct  $\rho$ -values starting from  $\rho_{\max}$  to  $\rho_{\min}$ ;
3. for  $k = 2$  to  $K$  do
4.     Let  $\hat{\boldsymbol{\mu}}_{ini}^{\rho_{(k)}} = \hat{\boldsymbol{\mu}}^{\rho_{(k-1)}}$  and  $\hat{\Theta}_{ini}^{\rho_{(k)}} = \hat{\Theta}^{\rho_{(k-1)}}$
5.     Use the EM algorithm to compute  $\{\hat{\boldsymbol{\mu}}^{\rho_{(k)}}, \hat{\Theta}^{\rho_{(k)}}\}$  with  $\{\hat{\boldsymbol{\mu}}_{ini}^{\rho_{(k)}}, \hat{\Theta}_{ini}^{\rho_{(k)}}\}$  as starting values;
6. end for

The previous description shows that the entire path can be computed using the estimates obtained for a given  $\rho$ -value as warm starts for fitting the next cglasso model. This strategy is commonly used also in other efficient lasso algorithms and R packages ([Friedman and others, 2010](#), [Friedman and others, 2018](#)). In our model, this strategy turns out to be remarkably efficient since using a sufficiently fine sequence of  $\rho$ -values, the starting values defined in Step 4 will be sufficiently close to the estimates computed in Step 5, thus increasing the speed of convergence of the resulting algorithm.

### 3.3. Tuning parameter selection

The tuning parameter plays a central role in the proposed cglasso estimator, since it is designed to control the complexity of the topological structure of the estimated graph. In this article, we propose to select the

optimal  $\rho$ -value of the cglasso estimator by using the extended Bayesian Information Criterion (Foygel and Drton 2010). For our proposed estimator, this is given by:

$$\text{BIC}_\gamma(\hat{\mathcal{E}}^\rho) = -2 \sum_{i=1}^n \log \varphi(\mathbf{x}_{i0_i}, \mathbf{r}_i; \hat{\boldsymbol{\mu}}, \hat{\Theta}(\hat{\mathcal{E}}^\rho)) + a(\rho)(\log n + 4\gamma \log p),$$

where  $\hat{\boldsymbol{\mu}}$  and  $\hat{\Theta}(\hat{\mathcal{E}}^\rho)$  are the maximum likelihood estimates of the Gaussian graphical model specified by  $\hat{\mathcal{E}}^\rho = \{(h, k) : \hat{\theta}_{hk}^\rho \neq 0\}$ , and  $a(\rho)$  denotes the number of nonzero off-diagonal estimates of  $\hat{\Theta}^\rho$ . The measure is indexed by the parameter  $\gamma \in [0, 1]$ , with  $\gamma = 0$  corresponding to the classical BIC measure. Although a grid search can be performed to select the  $\rho$ -value that minimizes  $\text{BIC}_\gamma(\hat{\mathcal{E}}^\rho)$ , the computational burden related to the evaluation of the log-likelihood function can make this strategy infeasible also for moderate size problems. For this reason, following Ibrahim and others (2008), we propose to select the  $\rho$ -value by minimizing the following approximate measure:

$$\overline{\text{BIC}}_\gamma(\hat{\mathcal{E}}^\rho) = -n[\log \det \hat{\Theta}^\rho - \text{tr}\{\Theta S(\hat{\boldsymbol{\mu}}, \hat{\Theta}(\hat{\mathcal{E}}^\rho))\}] + a(\rho)(\log n + 4\gamma \log p),$$

which is defined by substituting the exact log-likelihood function with the  $Q$ -function used in the M-Step of the proposed algorithm, which is easily obtained as a byproduct of the EM algorithm. In the [supplementary material](#) available at *Biostatistics* online, a simulation study is reported where we compare the behavior of the  $\text{BIC}_\gamma(\hat{\mathcal{E}}^\rho)$  and  $\overline{\text{BIC}}_\gamma(\hat{\mathcal{E}}^\rho)$  measures. The results show that the two criteria are equivalent in terms of false discovery rate and true positive rate. Furthermore, as suggested in Foygel and Drton (2010), the optimal results are obtained when  $\gamma = 0.5$ .

#### 4. SIMULATION STUDIES

##### 4.1. Evaluating the behavior of the approximate EM algorithm

In a first simulation, we evaluate the effects of the approximation (3.6) on the accuracy of the estimators obtained with the EM algorithm. As in the real dataset studied in Section 5, we consider right-censored data and we set the right-censoring value  $u_h$  equal to 40, for any  $h = 1, \dots, p$ . For this simulation, where we plan to use the full EM algorithm, we set the number of variables  $p$  to 10 and the sample size  $n$  to 100. To simulate a sample from a sparse right-censored model, we use the following procedure. First, using the method implemented in the R package *huge* (Zhao and others, 2015), we simulate a sparse concentration matrix with a random structure, where we set the probability that a  $\theta_{hk}$  is different from zero to 0.1. Then, the elements of the parameter  $\boldsymbol{\mu}$  are selected to obtain a fixed probability of right censoring for any given random variable. More specifically, we randomly draw a subset  $\mathcal{D}$  from  $\mathcal{I} = \{1, \dots, p\}$  and for each  $h \in \mathcal{D}$  the value of the parameter  $\mu_h$  is such that  $\text{Prob}\{R(X_h; -\infty, 40) = +1\} = 0.25$ , while for each  $h \notin \mathcal{D}$  the parameter  $\mu_h$  is such that the probability of right censoring is approximately equal to  $10^{-11}$ . The cardinality of the set  $\mathcal{D}$ , denoted by  $|\mathcal{D}|$ , is used in the study as a tool to analyze the effects of the number of censored variables on the behavior of the proposed EM algorithm and its approximated version. Finally, we draw a sample from a multivariate Gaussian distribution with parameters given in the previous steps and we treat each value greater than 40 as a missing (censored) value. We simulated 100 datasets from this model and for each simulation we compute a path of cglasso estimates using the proposed EM algorithm and a path using the approximated EM algorithm. In the following of this section, the estimates belonging to the two paths are denoted by  $\{\hat{\boldsymbol{\mu}}_e^\rho; \hat{\Theta}_e^\rho\}$  and  $\{\hat{\boldsymbol{\mu}}_a^\rho; \hat{\Theta}_a^\rho\}$ , respectively. For each path, the largest value of the tuning parameter was computed using the results given in Theorem 2 while the smallest value was set equal to  $1 \times 10^{-3}$ .

The first part of the Table 1 reports the average CPU times for computing the path. As expected, the computational time needed to compute a path is always an increasing function of the number of the



Table 1. Results of the simulation study on evaluating the effect of the approximation (3.6): first part reports the average CPU time for computing a path, whereas second part reports the average of  $\max_{\rho} \|\Delta \hat{\mu}^{\rho}\|^2$  and  $\max_{\rho} \|\Delta \hat{\Theta}^{\rho}\|_F^2$

	Number of censored variables						
	2	3	4	5	6	7	8
Average CPU time (s)							
Exact EM	9.60 (2.79)	25.87 (7.39)	68.36 (18.87)	104.77 (28.53)	140.01 (34.39)	249.40 (57.13)	374.34 (96.26)
Approx. EM	5.08 (1.02)	7.89 (1.44)	13.81 (2.23)	15.31 (2.20)	16.26 (2.33)	21.28 (2.64)	24.90 (3.10)
Difference between the two estimates							
$\max_{\rho} \ \Delta \hat{\mu}^{\rho}\ ^2$	$2.9 \times 10^{-6}$ ( $8.6 \times 10^{-6}$ )	$8.2 \times 10^{-6}$ ( $2.1 \times 10^{-5}$ )	$1.7 \times 10^{-5}$ ( $3.7 \times 10^{-5}$ )	$2.2 \times 10^{-5}$ ( $3.1 \times 10^{-5}$ )	$7.8 \times 10^{-5}$ ( $1.0 \times 10^{-4}$ )	$1.1 \times 10^{-4}$ ( $1.2 \times 10^{-4}$ )	$2.1 \times 10^{-4}$ ( $2.3 \times 10^{-4}$ )
$\max_{\rho} \ \Delta \hat{\Theta}^{\rho}\ _F^2$	$3.0 \times 10^{-5}$ ( $8.5 \times 10^{-5}$ )	$8.3 \times 10^{-5}$ ( $2.0 \times 10^{-4}$ )	$2.0 \times 10^{-4}$ ( $3.7 \times 10^{-4}$ )	$2.6 \times 10^{-4}$ ( $3.8 \times 10^{-4}$ )	$2.3 \times 10^{-3}$ ( $3.1 \times 10^{-4}$ )	$2.6 \times 10^{-3}$ ( $2.4 \times 10^{-3}$ )	$6.6 \times 10^{-3}$ ( $6.0 \times 10^{-3}$ )

Standard deviations are shown in brackets.

censored variables but, when we use the matrix  $S(\hat{\mu}_{ini}^{\rho}, \hat{\Theta}_{ini}^{\rho})$  in the E-step, the table reveals that the form of the relationship between the CPU time and  $|\mathcal{D}|$  is almost exponential implying that the computation of the cglasso estimator is infeasible also for relatively small datasets. In contrast to this, when we use the matrix  $\bar{S}(\hat{\mu}_{ini}^{\rho}, \hat{\Theta}_{ini}^{\rho})$  to approximate the current estimate of the empirical covariance matrix, the table shows an almost linear dependence of the CPU time, which implies a significant reduction of the computational complexity compared with the full EM algorithm. Although the results showed in Table 1 strongly suggest the use of the approximated EM algorithm to compute the cglasso estimator, they do not provide information about the difference between the two estimates. For this reason, we also computed the largest Euclidean distance between  $\hat{\mu}_e^{\rho}$  and  $\hat{\mu}_a^{\rho}$ , denoted as  $\max_{\rho} \|\Delta \hat{\mu}^{\rho}\|^2$ , and the largest Frobenius distance between  $\hat{\Theta}_e^{\rho}$  and  $\hat{\Theta}_a^{\rho}$ , denoted by  $\max_{\rho} \|\Delta \hat{\Theta}^{\rho}\|_F^2$ . The average and standard deviations of these values are reported in the second part of the Table 1. All the results clearly show that the two estimators are sufficiently close to each other, and point to the use of the approximated EM algorithm for the derivation of the cglasso estimator.

#### 4.2. Comparison of methods on data simulated from a censored Gaussian graphical model

In a second simulation study, we compare our proposed estimator with MissGlasso (Städler and Bühlmann, 2012), which performs  $\ell_1$ -penalized estimation under the assumption that the censored data are missing at random, and with the glasso estimator (Friedman and others, 2008), where the empirical covariance matrix is calculated by imputing the missing values with the limit of detection. These estimators are evaluated in terms of both recovering the structure of the true graph and the mean squared error. We use a similar approach to the previous simulation for generating right censored data. In particular, we set the right censoring value to 40 for any variable and the sample size  $n$  to 100. We generate a sparse concentration matrix with random structure and set the probability of observing a link between two nodes to  $k/p$ , where  $p$  is the number of variables and  $k$  is used to control the amount of sparsity in  $\Theta$ . Finally, we set the mean  $\mu$  in such a way that  $\mu_h = 40$  for the  $H$  censored variables, i.e.  $\text{Prob}\{R(X_h; -\infty, 40) = +1\} = 0.50$ , while for the remaining variables  $\mu_h$  is sampled from a uniform distribution on the interval  $[10; 35]$ . At this point, we simulate a sample from the latent  $p$ -variate Gaussian distribution and treat all values greater

than 40 as missing. The quantities  $k$ ,  $p$ , and  $H$  are used to specify the different models used to analyze the behavior of the considered estimators. In particular, we consider the following cases:

- **Model 1:**  $k = 3$ ,  $p = 50$  and  $H \in \{25, 35\}$ . This setting is used to evaluate the effects of the number of censored variables on the behavior of the proposed estimators when  $n > p$ .
- **Model 2:**  $k \in \{1, 5\}$ ,  $p = 50$  and  $H = 30$ . This setting is used to evaluate the effects of the sparsity of the matrix  $\Theta$  on the considered estimators when  $n > p$ .
- **Model 3:**  $k = 3$ ,  $p = 200$  and  $H = 100$ . This setting is used to evaluate the impact of the high dimensionality on the estimators ( $p \gg n$ ).

For each model, we simulate 100 samples from the right-censored model and in each simulation we compute the coefficients path using cglasso, MissGlasso, and glasso. Each path is computed using an equally spaced sequence of 30  $\rho$ -values. Figure 1a shows the precision-recall curves for Model 1 with  $H = 25$ . The curves report the relationship between precision and recall for any  $\rho$ -value, which are defined by:

$$\text{Precision}(\rho) = \frac{\text{number of } \hat{\theta}_{hk}^{\rho} \neq 0 \text{ and } \theta_{hk} \neq 0}{\text{number of } \hat{\theta}_{hk}^{\rho} \neq 0}, \quad \text{Recall}(\rho) = \frac{\text{number of } \hat{\theta}_{hk}^{\rho} \neq 0 \text{ and } \theta_{hk} \neq 0}{\text{number of } \theta_{hk} \neq 0}.$$

The curves show how cglasso gives a better estimate of the concentration matrix both in terms of precision and recall, for any given value of the tuning parameter. Figure 1b shows the distributions of the quantity  $\min_{\rho} \text{MSE}(\hat{\Theta}^{\rho})$ , which gives the minimum value of the mean squared error attained along the path of solutions. These box-plots emphasize that, not only the cglasso has a mean squared error much smaller than glasso and MissGlasso, but also that it is much more stable than its competitors. The same behavior was also observed in the other models used in our numerical study, and can be found in the [Supplementary Material](#) available at *Biostatistics* online. Table 2 reports the summary statistics from all the simulations. In addition to the quantity described above (second meta-column), the first meta-column shows the mean squared error in the estimation of the mean (denoted by  $\min_{\rho} \text{MSE}(\hat{\mu}^{\rho})$ ) and the third meta-column reports the Area Under the precision-recall Curve (AUC). Note that the considered measures allow us to study the behavior of the estimators along the entire path. The distribution of the minimum value of the mean squared errors shows that, not only our estimator is able to recover the structure of the graph but also outperforms the competitors in terms of both estimation of  $\mu$  and  $\Theta$ . We did not report  $\min_{\rho} \text{MSE}(\hat{\mu}^{\rho})$  for glasso since this method does not allow to estimate the parameter  $\mu$ . The results on the AUC suggest that cglasso can be used as an efficient tool for recovering the structure of the true concentration matrix of a Gaussian graphical model from censored data.

#### 4.3. Testing the robustness of the method on more realistic biological data

In a third simulation study, we test the robustness of the method on more realistic biological data. In particular, we consider expression data from [Wille and others \(2004\)](#) on the extensively studied *Arabidopsis thaliana* biological system. The study reports data from  $n = 118$  experiments on  $p = 39$  genes, whose regulatory network is of interest. Although the data are fully observed, we test our method on a dataset where observations are made artificially right censored, similarly to [Städler and Bühlmann \(2012\)](#). In particular, we produce three datasets at different levels of right censoring, by recording as missing the 10%, 20%, and 30% of the highest values, respectively. Then, we compare our method, cglasso, with MissGlasso ([Städler and Bühlmann, 2012](#)) and with three additional methods, which impute missing

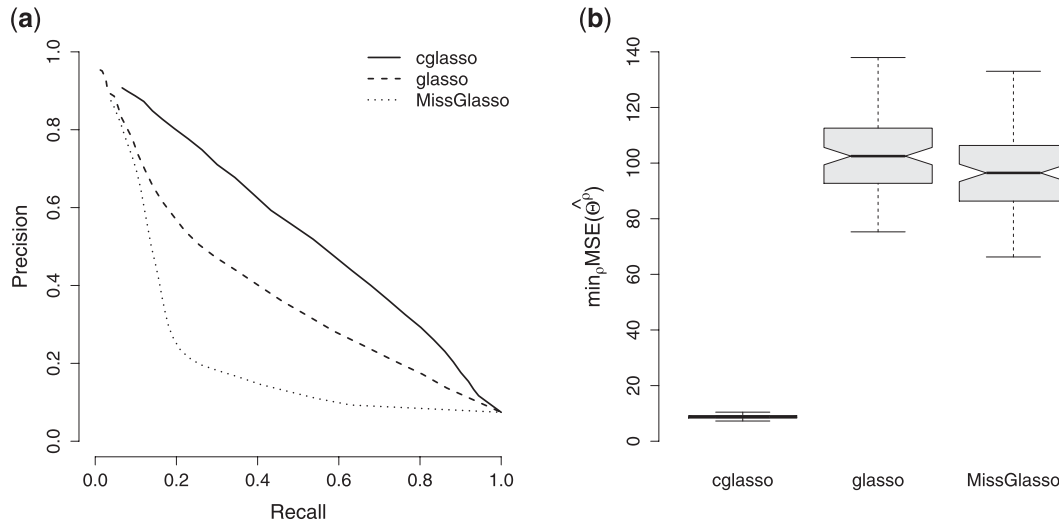


Fig. 1. Results of the comparative simulation study based on Model 1 with  $H = 25$ . (a) The average precision-recall curves are shown; (b) The box-plots of  $\min_{\rho} \text{MSE}(\hat{\Theta}^{\rho})$  for the considered estimators is shown.

Table 2. Comparison between cglasso and two existing methods on simulated data: for each measure the table reports the average value and standard deviation between parentheses

Model			$\min_{\rho} \text{MSE}(\hat{\mu}^{\rho})$		$\min_{\rho} \text{MSE}(\hat{\Theta}^{\rho})$			AUC		
$p$	$H/p$	$k/p$	cglasso	MissGlasso	cglasso	glasso	MissGlasso	cglasso	glasso	MissGlasso
50	0.5	0.06	0.47	14.50	8.76	103.35	96.75	0.48	0.37	0.19
			(0.11)	(0.69)	(0.64)	(14.43)	(16.01)	(0.04)	(0.03)	(0.02)
50	0.7	0.06	0.48	21.00	10.11	139.76	131.99	0.47	0.33	0.15
			(0.10)	(0.76)	(0.84)	(15.94)	(18.81)	(0.05)	(0.03)	(0.02)
50	0.6	0.02	0.47	18.31	6.92	128.60	119.65	0.61	0.42	0.16
			(0.10)	(0.81)	(0.82)	(17.75)	(17.38)	(0.08)	(0.05)	(0.03)
50	0.6	0.10	0.46	17.39	12.02	113.84	105.70	0.43	0.34	0.20
			(0.10)	(0.92)	(0.85)	(14.79)	(15.71)	(0.04)	(0.03)	(0.02)
200	0.5	0.015	1.92	63.34	41.57	398.02	373.86	0.32	0.21	0.13
			(0.19)	(1.54)	(1.52)	(29.78)	(32.46)	(0.02)	(0.02)	(0.01)

values first and then infer the network using graphical lasso on the imputed values. As in Stadler and Buhlmann (2012), we consider k-nearest neighbor imputation, which we denote with missknn, and imputation using random forests, which we denote with MissForest. These methods are based on a missing at random assumption. We further consider a method which relies on multivariate Gaussian data under a censoring mechanism which we denote with imputeLCMD (Lazar, 2015).

Table 3 shows the Euclidean norm between the true observed data and the imputed one for the five methods considered and across the three different levels of censoring. The results show how the performance of all methods decreases the higher the level of censoring and how cglasso is overall superior to the other methods across all comparisons, followed closely in some cases by imputeLCMD. The areas under the precision-recall curves of the networks inferred by the five methods compared with the graphical lasso

Table 3. Comparison of methods on the *Arabidopsis thaliana* expression data: Euclidian distance (ED) between the observed and imputed data and area under the precision-recall curve (AUC) across the five methods and the three levels of censoring

	Percentage censored	cglasso	imputeLMCD	MissGlasso	missknn	missForest
ED	10%	9.95	13.49	27.89	31.37	33.31
	20%	15.22	18.60	37.73	40.16	46.21
	30%	19.76	24.38	46.65	50.08	57.54
AUC	10%	0.93	0.88	0.78	0.79	0.77
	20%	0.90	0.86	0.70	0.72	0.68
	30%	0.87	0.85	0.63	0.70	0.61

network selected by  $\overline{\text{BIC}}_{0.5}(\hat{\mathcal{E}}^\rho)$  from the fully observed data show how the proposed cglasso method leads to a significant gain in network recovery over MissGlasso, missknn, and MissForest even with 30% of censored data.

## 5. APPLICATION TO SINGLE CELL-DATA: MEGAKARYOCYTE-ERYTHROID PROGENITORS

Recent advances in single-cell techniques have provided the opportunity to finely dissect cellular heterogeneity within known populations and to uncover rare cell types. In a study about the formation of blood cells, Psaila and others (2016) have recently identified three distinct subpopulations of cells, which are all derived from hematopoietic stem cells through cell differentiation. One of these sub-populations, denoted by MK-MEP, is a previously unknown, rare population of cells that are bipotent but primarily generate megakaryocytic progeny. In this section, we look closely at this sub-population and investigate the molecular mechanisms of regulation within these cells. To this end, we used data available from Psaila and others (2016) on 87 genes and 48 single human MK-MEP cells profiled by multiplex RT-qPCR.

As discussed in the Introduction, RT-qPCR data are typically right-censored. In this particular study, the limit of detection is fixed by the manufacturer to 40. Raw data have been mean normalized using the method proposed in Pipelers and others (2017) and using the two housekeeping genes provided. Figure 2a shows the relationship between the proportion of right-censored data and the mean of the normalized cycle-threshold. Two main conclusions can be drawn from this figure. Firstly, the proportion of censored data is an increasing function of the mean of the normalized cycle-threshold: this is expected and it means that the assumption of missing-at-random is not justified on this dataset. Secondly, there are some genes with a very high proportion of censoring (see the points above the dashed line): these may be genes whose transcription failed to amplify, in which case they should be treated as missing rather than censored. Following this explorative analysis and considering also the possible computational problems caused by the inclusion of these genes, we filtered out the genes with a proportion of censoring above 85% for subsequent analysis. The resulting data set contains the expression of 63 genes measured on 48 cells.

The normalized cyclic thresholds of the remaining genes are used to fit a right-censored GGM by using the proposed cglasso estimator. Figure 2b shows the path of the  $\overline{\text{BIC}}_{0.5}(\hat{\mathcal{E}}^\rho)$  measure and the vertical dashed line is traced in correspondence of the optimal  $\rho$ -value. The resulting estimated graph (see Figure 2c) contains about 0.6% of all possible edges and about 19% of the considered genes. Consistent with the existing knowledge about this subpopulation, the estimated graph shows the central role of two genes, CD42 (glycoprotein 1b) and MYB (Psaila and others, 2016). The first one, CD42, is a megakaryocyte gene whose expression was found to be significantly high in the MK-MEP subpopulation. In particular, it

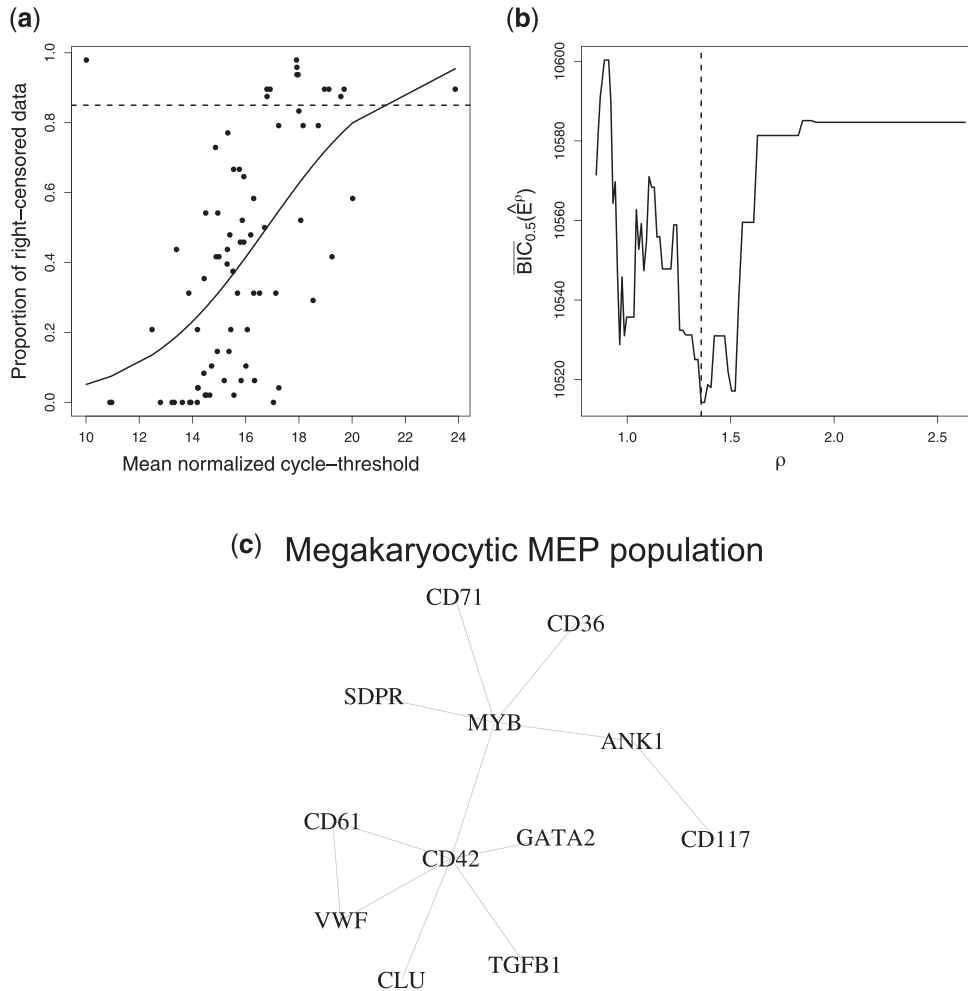


Fig. 2. Real data analysis. (a) The proportion of right-censored data versus the mean of the normalized cycle-threshold; the black line is obtained by fitting a logistic regression model, while the dashed line identifies the threshold used to filter out the genes from the study. (b) Path of the  $\overline{\text{BIC}}_{0.5}(\hat{\mathcal{E}}^\rho)$  measure; the vertical dashed line identifies the optimal value of the tuning parameter. (c) The estimated graph.

is expressed later during megakaryocyte differentiation and has been associated with unipotent megakaryopoietic activity in mouse models. The link between CD42 and VWF, another highly expressed genes in this subpopulation (Psaila and others, 2016), was discovered by Chan and others (2017) as well as other connections, such as the link between the genes CD42 and CD61 and that between CD42 and TGFB1. The second central gene, MYB, is a transcription factor that is known to enhance erythroid differentiation at the expense of megakaryopoiesis.

## 6. CONCLUSION

In this article, we have proposed a computational approach to fit Gaussian Graphical models in the presence of censored data. The approach includes both the cases of right- and left-censored data. Since

classical Gaussian graphical models cannot be used in high-dimensional settings, we also introduced  $\ell_1$ -penalization to produce sparsity (model selection) and parameter estimation simultaneously. The resulting estimator is called censored graphical lasso (cglasso). The computational problem of estimating the mean and conditional independence graph of a Gaussian graphical model from censored data is solved via an EM-algorithm. An extensive simulation study showed that the proposed estimator overcomes the existing estimators both in terms of parameter estimation and of network recovery. The analysis of a real RT-qPCR dataset showed how the method is able to infer the regulatory network underlying blood development under high levels of censoring.

In addition, we have proposed an approximated EM-algorithm, which is computationally more efficient and makes the method feasible for high-dimension settings. This approach relies on an efficient approximation of the mixed moments of the latent variables conditional on the observed. The approximation is exact under conditional independence and has thus been found to work well under sparse settings, as shown by [Guo and others \(2015\)](#) and our own simulations. Future work will investigate the theoretical justifications for this as well as refining the scenarios under which a good performance is to be expected.

#### SUPPLEMENTARY MATERIAL

Supplementary material is available at <http://biostatistics.oxfordjournals.org>.

#### ACKNOWLEDGMENTS

*Conflict of Interest:* None declared.

#### FUNDING

The project was partially supported by the European Cooperation in Science and Technology (COST) [COST Action CA15109 European Cooperation for Statistics of Network Data Science (COSTNET)].

#### REFERENCES

- ARISMENDI, J. C. (2013). Multivariate truncated moments. *Journal of Multivariate Analysis* **117**, 41–75.
- AUGUGLIARO, L., MINEO, A. M. AND WIT, E. C. (2016).  $\ell_1$ -Penalized methods in high-dimensional Gaussian Markov random fields. In: Dehmer, M., Shi, Y. and Emmert-Streib, F. (editors), *Computational Network Analysis with R: Applications in Biology, Medicine, and Chemistry*, Chapter 8. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, pp. 201–267.
- BICKEL, P. J. AND LEVINA, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics* **36**, 199–227.
- BROWN, M. AND WITTWER, C. (2000). Flow cytometry: principles and clinical applications in hematology. *Clinical Chemistry* **46**, 1221–1229.
- CHAN, T. E., STUMPF, M. P. H. AND BARTIE, A. C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Systems* **5**, 251–267.
- DEMPSTER, A. P., LAIRD, N. M. AND RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B* **39**, 1–38.
- DERVEAUX, S., VANDESOMPELE, J. AND HELLEMANS, J. (2010). How to do successful gene expression analysis using real-time PCR. *Methods* **50**, 227–230.
- FOYGEL, R. AND DRTON, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. In: Lafferty, J., Williams, C., Shawe-taylor, J., Zemel, R.s. and Culott, A. (editors), *Advances in Neural Information Processing Systems*, Vancouver, British Columbia, Canada: Curran Associates, Inc., pp. 604–612.



- FRIEDMAN, J. H., HASTIE, T. AND TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441.
- FRIEDMAN, J. H., HASTIE, T. AND TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.
- FRIEDMAN, J. H., HASTIE, T. AND TIBSHIRANI, R. (2018). *glasso: Graphical Lasso-Estimation of Gaussian Graphical Models*. R package version 1.10
- GARDNER, T. S., DI BERNARDO, D., LORENZ, D. AND COLLINS, J. J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301**, 102–105.
- GENZ, A. AND BRETZ, F. (2002). Comparison of methods for the computation of multivariate  $t$  probabilities. *Journal of Computational and Graphical Statistics* **11**, 950–971.
- GUO, J., LEVINA, E., MICHAELIDIS, G. AND ZHU, J. (2015). Graphical models for ordinal data. *Journal of Computational and Graphical Statistics* **24**, 183–204.
- HOFFMAN, H. J. AND JOHNSON, R. E. (2015). Pseudo-likelihood estimation of multivariate normal parameters in the presence of left-censored data. *Journal of Agricultural, Biological, and Environmental Statistics* **20**, 156–171.
- IBRAHIM, J. G., ZHU, H. AND TANG, N. (2008). Model selection criteria for missing-data problems using the EM algorithm. *Journal of the American Statistical Association* **103**, 1648–1658.
- JOHNSON, N. L., KOTZ, S. AND BALAKRISHNAN, N. (1994). *Continuous Univariate Distributions*, 2nd edition, Volume 1. New York: John Wiley & Sons, Inc.
- LAURITZEN, S. L. (1996). *Graphical Models*. Oxford: Oxford University Press.
- LAZAR, C. (2015). *imputeLCMD: A Collection of Methods for Left-Censored Missing Data Imputation*. R package version 2.0.
- LEE, G. AND SCOTT, C. (2012). EM algorithms for multivariate Gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis* **56**, 2816–2829.
- LEE, L.-F. (1983). The determination of moments of the doubly truncated multivariate normal tobit model. *Economics Letters* **11**, 245–250.
- LEPPARD, P. AND TALLIS, G. M. (1989). Algorithm AS 249: Evaluation of the mean and covariance of the truncated multinormal distribution. *Journal of the Royal Statistical Society. Series C* **38**, 543–553.
- LITTLE, R. J. A. AND RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd edition. Hoboken, NJ, USA: John Wiley & Sons, Inc.
- MCCALL, M. N., McMURRAY, H. R., LAND, H. AND ALMUDEVAR, A. (2014). On non-detects in qPCR data. *Bioinformatics* **30**, 2310–2316.
- MCLACHLAN, G. AND KRISHNAN, T. (2008). *The EM Algorithm and Extensions*, 2nd edition. Hoboken, NJ, USA: John Wiley & Sons, Inc.
- MENÉNDEZ, P., KOURMPETIS, Y. A. I., TER BRAAK, C. J. F. AND VAN EEUWIJK, F. A. (2010). Gene regulatory networks from multifactorial perturbations using graphical lasso: application to DREAM4 challenge. *PLoS One* **5**, e14147.
- PERKINS, N. J., SCHISTERMAN, E. F. AND VEXIER, A. (2013). Multivariate normally distributed biomarkers subject to limits of detection and receiver operating characteristic curve inference. *Academic Radiology* **20**, 838–846.
- PESONEN, M., PESONEN, H. AND NEVALAINEN, J. (2015). Covariance matrix estimation for left-censored data. *Computational Statistics & Data Analysis* **92**, 13–25.
- PIPELERS, P., CLEMENT, L., VYNCK, M., HELLEMANS, J., VANDESOMPELE, J. AND THAS, O. (2017). A unified censored normal regression model for qPCR differential gene expression analysis. *PLoS One* **12**, e0182832.

- PSAILA, B., BARKAS, N., ISKANDER, D., ROY, A., ANDERSON, S., ASHLEY, N., CAPUTO, V. S., LICHTENBERG, J., LOAIZA, S., BODINE, D. M. *and others.* (2016). Single-cell profiling of human megakaryocyte-erythroid progenitors identifies distinct megakaryocyte and erythroid differentiation pathways. *Genome Biology* **17**, 83–102.
- ROTHMAN, A., BICKEL, P. J., LEVINA, E. AND ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2**, 494–515.
- SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*, Monographs on Statistics and Applied Probability. Boca Raton, Florida: Chapman and Hall/CRC.
- STÄDLER, N. AND BÜHLMANN, P. (2012). Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing* **22**, 219–235.
- UHLER, C. (2012). Geometry of maximum likelihood estimation in Gaussian graphical models. *The Annals of Statistics* **40**, 238–261.
- VINCIOTTI, V., AUGUGLIARO, L., ABBRUZZO, A. AND WIT, E. C. (2016). Model selection for factorial Gaussian graphical models with an application to dynamic regulatory networks. *Statistical Applications in Genetics and Molecular Biology* **15**, 193–212.
- WILLE, A., ZIMMERMANN, P., VRANOVÁ, E., FÜRHOZ, A., LAULE, O., BLEULER, S., HENNIG, L., PRELIĆ, A., VON ROHR, P., THIELE, L., ZITZLER, E., GRUISSEM, W. *and others.* (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biology* **5**, R92.
- WITTEN, D. M., FRIEDMAN, J. H. AND SIMON, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics* **20**, 892–900.
- YUAN, M. AND LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**, 19–35.
- ZHAO, T., LI, X., LIU, H., ROEDER, K., LAFFERTY, J. AND WASSERMAN, L. (2015). *huge: High-Dimensional Undirected Graph Estimation*. R package version 1.2.7.

[Received January 9, 2018; revised July 2, 2018; accepted for publication July 15, 2018]