Biostatistics (2018), **0**, 0, *pp.* 1–28
doi:10.1093/biostatistics/cglasso˙01˙07˙2018

ℓ_1-Penalized Censored Gaussian Graphical Model

Luigi Augugliaro*

Department of Economics, Business and Statistics, University of Palermo, Building 13, Viale delle Scienze, 90128 Palermo, Italy

luigi.augugliaro@unipa.it

Antonino Abbruzzo

Department of Economics, Business and Statistics, University of Palermo, Building 13, Viale delle Scienze, 90128 Palermo, Italy

Veronica Vinciotti

Department of Mathematics, Brunel University London, Uxbridge UB8 3PH, United Kingdom

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 paper we propose an ℓ_1-penalized Gaussian graphical model for censored data and derive two EM-like algorithms for inference. We evaluate the computational efficiency

*To whom correspondence should be addressed.

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 RT-qPCR 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`).

Key words: 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 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 ℓ_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. Friedman *and others*, 2008; Witten *and others*, 2011; Bickel and Levina, 2008) 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 (Pipelers *and others*, 2017; McCall *and others*, 2014). 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 unfunded since it does not consider the effects of the censoring mechanism on the estimator of the concentration matrix.

In this paper 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 ℓ_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 paper, failure to take into account the censoring mechanism causes a poor behaviour of this approach when compared to 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 paper 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 behaviour 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 $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ 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 $\boldsymbol{X}$ 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 $\boldsymbol{X}$ follows a multivariate Gaussian distribution with expected value $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top$ 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 $\boldsymbol{X}$ can be written as

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

As shown in Lauritzen (1996), the pattern of non-zero elements in Θ 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 $\boldsymbol{X}$ 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 $\boldsymbol{l} = (l_1, \ldots, l_p)^\top$ and $\boldsymbol{u} = (u_1, \ldots, u_p)^\top$, with $l_h < u_h$ for $h = 1, \ldots, p$, 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 $R(\boldsymbol{X}; \boldsymbol{l}, \boldsymbol{u})$, used to encode the censoring patterns. Formally, the hth element of $R(\boldsymbol{X}; \boldsymbol{l}, \boldsymbol{u})$ 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 $R(\boldsymbol{X}; \boldsymbol{l}, \boldsymbol{u})$ is a discrete random vector with support in the set $\{-1, 0, 1\}^p$ and probability function

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

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

Given a censoring pattern, we can simplify our notation by partitioning the set $\mathcal{I} = \{1, \ldots, p\}$ 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 paper, 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 $\boldsymbol{x}$ is denoted by $\boldsymbol{x}_o = (x_h)_{h \in o}$ and, consequently, the observed data is the vector $(\boldsymbol{x}_o^\top, \boldsymbol{r}^\top)^\top$. As explained in Little and Rubin (2002), the joint probability distribution of the observed data, denoted by $\varphi(\boldsymbol{x}_o, \boldsymbol{r}; \boldsymbol{\mu}, \Theta)$, is obtained by integrating $\boldsymbol{X}_{c^+}$ and $\boldsymbol{X}_{c^-}$ out of the joint distribution of $\boldsymbol{X}$ and $R(\boldsymbol{X}; \boldsymbol{l}, \boldsymbol{u})$, which can be written as the product of the density function (2.1) and the conditional distribution of $R(\boldsymbol{X}; \boldsymbol{l}, \boldsymbol{u})$ given $\boldsymbol{X} = \boldsymbol{x}$. As shown in Schafer (2000), this results in

$$\begin{aligned} \varphi(\boldsymbol{x}_o, \boldsymbol{r}; \boldsymbol{\mu}, \Theta) &= \int_{\boldsymbol{u}_{c^+}}^{+\infty} \int_{-\infty}^{\boldsymbol{l}_{c^-}} \phi(\boldsymbol{x}_o, \boldsymbol{x}_{c^-}, x_{c^+}; \boldsymbol{\mu}, \Theta) \mathrm{d}\boldsymbol{x}_{c^-} \mathrm{d}\boldsymbol{x}_{c^+} I(\boldsymbol{l}_o \leqslant \boldsymbol{x}_o \leqslant \boldsymbol{u}_o) \\ &= \int_{D_c} \phi(\boldsymbol{x}_o, \boldsymbol{x}_c; \boldsymbol{\mu}, \Theta) \mathrm{d}\boldsymbol{x}_c I(\boldsymbol{l}_o \leqslant \boldsymbol{x}_o \leqslant \boldsymbol{u}_o), \end{aligned} \tag{2.2}$$

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

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. ℓ_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 $\{\boldsymbol{X}, R(\boldsymbol{X}; \boldsymbol{l}, \boldsymbol{u}), \varphi(\boldsymbol{x}_o, \boldsymbol{r}; \boldsymbol{\mu}, \Theta), \mathcal{G}\}$. For ease of exposition, we shall assume that $\boldsymbol{l}$ and $\boldsymbol{u}$ 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 $\boldsymbol{r}_i$ be the ith realization of the random vector $R(\boldsymbol{X}_i; \boldsymbol{l}, \boldsymbol{u})$. Then the ith observed data is the vector $(\boldsymbol{x}_{io_i}^\top, \boldsymbol{r}_i^\top)^\top$, 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

$$\ell(\boldsymbol{\mu}, \Theta) = \sum_{i=1}^{n} \log \int_{D_{c_i}} \phi(\boldsymbol{x}_{io_i}, \boldsymbol{x}_{ic_i}; \boldsymbol{\mu}, \Theta) \mathrm{d}\boldsymbol{x}_{ic_i} = \sum_{i=1}^{n} \log \varphi(\boldsymbol{x}_{io_i}, \boldsymbol{r}_i; \boldsymbol{\mu}, \Theta), \tag{3.3}$$

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 $D_{c_i} = (-\infty, \boldsymbol{l}_{c_i^-}) \times (\boldsymbol{u}_{c_i^+}, +\infty)$. 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 Θ has a sparse structure, i.e., only few θ_{hk} are different from zero, which is not obtained by a direct (or indirect) maximization of the observed log-likelihood function (3.3).

All that considered, in this paper 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 (3.3). The resulting estimator, called censored graphical lasso (cglasso),

is formally defined as

$$\{\hat{\boldsymbol{\mu}}^\rho, \widehat{\Theta}^\rho\} = \arg\max_{\boldsymbol{\mu}, \Theta \succ 0} \frac{1}{n} \sum_{i=1}^{n} \log \varphi(\boldsymbol{x}_{io_i}, \boldsymbol{r}_i; \boldsymbol{\mu}, \Theta) - \rho \sum_{h \neq k} |\theta_{hk}|. \tag{3.4}$$

Like in the standard graphical lasso estimator, the non-negative tuning parameter ρ is used to control the amount of sparsity in the estimated concentration matrix $\widehat{\Theta}^\rho = (\hat{\theta}^\rho_{hk})$ and, consequently, in the corresponding estimated graph $\widehat{\mathcal{G}}^\rho = \{\mathcal{V}, \widehat{\mathcal{E}}^\rho\}$, where $\widehat{\mathcal{E}}^\rho = \{(h,k) : \hat{\theta}^\rho_{hk} \neq 0\}$. When ρ is large enough, some $\hat{\theta}^\rho_{hk}$ are shrunken to zero resulting in the removal of the corresponding link in $\widehat{\mathcal{G}}^\rho$; on the other hand, when ρ 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 $\boldsymbol{X}_{c_i}$ given $\boldsymbol{X}_{o_i} = \boldsymbol{x}_{o_i}$ 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 $E_{c_i|o_i}(\boldsymbol{X}_{c_i}) = \boldsymbol{\mu}_{c_i|o_i} = \boldsymbol{\mu}_{c_i} - \Theta^{-1}_{c_i c_i} \Theta_{c_i o_i}(\boldsymbol{x}_{io_i} - \boldsymbol{\mu}_{o_i})$, 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 the first and second moment of the Gaussian distribution $\phi(\boldsymbol{x}_{c_i}; \boldsymbol{\mu}_{c_i|o_i}, \Theta_{c_i c_i})$ truncated over the region D_{c_i}. To this end we let

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

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

where $E_{c_i|o_i}(\cdot \mid \boldsymbol{X}_{ic_i} \in D_{c_i})$ denotes the expected value computed using the conditional distribution of $\boldsymbol{X}_{ic_i}$ given $\boldsymbol{x}_{io_i}$ truncated over D_{c_i}. Finally we let $\bar{x}_h(\boldsymbol{\mu}, \Theta) = \sum_{i=1}^n x_{i,h}(\boldsymbol{\mu}, \Theta)/n$, $\bar{\boldsymbol{x}}(\boldsymbol{\mu}, \Theta) = \{\bar{x}_1(\boldsymbol{\mu}, \Theta), \ldots, \bar{x}_p(\boldsymbol{\mu}, \Theta)\}^\top$, $s_{hk}(\boldsymbol{\mu}, \Theta) = \sum_{i=1}^n x_{i,hk}(\boldsymbol{\mu}, \Theta)/n - \bar{x}_h(\boldsymbol{\mu}, \Theta)\bar{x}_k(\boldsymbol{\mu}, \Theta)$ and

$S(\boldsymbol{\mu}, \Theta) = \{s_{hk}(\boldsymbol{\mu}, \Theta)\}$. Using this notation, the next theorem gives the Karush-Kuhn-Tucker conditions for the proposed estimator.

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

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

are

$$\left.\begin{aligned} \bar{x}_h(\hat{\boldsymbol{\mu}}^\rho, \widehat{\Theta}^\rho) - \hat{\mu}_h^\rho = 0 \\ \hat{\sigma}_{hk}^\rho(\hat{\boldsymbol{\mu}}^\rho, \widehat{\Theta}^\rho) - s_{hk}(\hat{\boldsymbol{\mu}}^\rho, \widehat{\Theta}^\rho) - \rho \hat{v}_{hk} = 0 \end{aligned}\right\} \tag{3.5}$$

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 $|\hat{v}_{hk}| \leqslant 1$ if $\hat{\theta}_{hk}^\rho = 0$.

A proof of this theorem is reported in the supplementary material available at *Biostatistics* online. The stationary conditions (3.5) show that, while in the standard graphical lasso the parameter $\boldsymbol{\mu}$ can be estimated by the empirical average regardless of ρ 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 $\{\hat{\boldsymbol{\mu}}_{ini}^\rho, \widehat{\Theta}_{ini}^\rho\}$ an initial estimate of the parameters,

the E-Step reduces to the computation of $x_{i,h}(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini})$ and $x_{i,hk}(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini})$, for $i = 1, \ldots, n$. From this, conditions (3.5) can be written as

$$\left.\begin{aligned} \bar{x}_h(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini}) - \hat{\mu}^{\rho}_h = 0 \\ \hat{\sigma}^{\rho}_{hk}(\hat{\boldsymbol{\mu}}, \widehat{\Theta}) - s_{hk}(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini}) - \rho \hat{v}_{hk} = 0 \end{aligned}\right\} \tag{3.6}$$

which are the stationary conditions of a standard graphical lasso problem (Witten *and others*, 2011) with $S(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini})$ used as the current estimate of the empirical covariance matrix. The conditions (3.6) imply that in the M-Step the parameter $\boldsymbol{\mu}$ is estimated by $\bar{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini})$ while Θ is estimated by solving the following maximization problem:

$$\max_{\Theta \succ 0} Q(\Theta \mid \widehat{\Theta}^{\rho}_{ini}) = \max_{\Theta \succ 0} \log\det\Theta - \mathrm{tr}\{\Theta S(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini})\} - \rho \sum_{h,k} |\theta_{hk}|, \tag{3.7}$$

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

1. Let $\{\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini}\}$ be initial estimates;
2. E-step: Compute $\bar{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini})$ and $S(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini})$;
3. M-step: Let $\hat{\boldsymbol{\mu}}^{\rho} = \bar{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini})$ and solve the problem (3.7) using $S(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini})$;
4. If a convergence criterion is met then return $\{\hat{\boldsymbol{\mu}}^{\rho}, \widehat{\Theta}^{\rho}\}$ else let $\hat{\boldsymbol{\mu}}^{\rho}_{ini} = \hat{\boldsymbol{\mu}}^{\rho}$ and $\widehat{\Theta}^{\rho}_{ini} = \widehat{\Theta}^{\rho}$;
5. Repeat steps 2–4.

Although the maximization problem (3.7) can be efficiently solved using, for example, the algorithm proposed by Friedman *and others* (2008), by Rothman *and others* (2008) or by 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 $\bar{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini})$ and $S(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini})$. These can be computed using the methods proposed by Lee (1983), Leppard and Tallis (1989) or by Arismendi (2013), but these methods require complex numerical algorithms for the calculation of the integral of

the multivariate normal density function (see Genz and Bretz (2002) for a review), 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 $E_{c_i|o_i}(X_{ih}X_{ik} \mid \boldsymbol{X}_{ic_i} \in D_{c_i})$, for any $h \neq k$, by:

$$E_{c_i|o_i}(X_{ih}X_{ik} \mid \boldsymbol{X}_{ic_i} \in D_{c_i}) \approx E_{c_i|o_i}(X_{ih} \mid \boldsymbol{X}_{ic_i} \in D_{c_i})E_{c_i|o_i}(X_{ik} \mid \boldsymbol{X}_{ic_i} \in D_{c_i}). \tag{3.8}$$

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 (3.8) is that now, in order to evaluate the quantities $\bar{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini})$ and $S(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini})$, we only need to compute $E_{c_i|o_i}(X_{ih} \mid \boldsymbol{X}_{ic_i} \in D_{c_i})$ and $E_{c_i|o_i}(X^2_{ih} \mid \boldsymbol{X}_{ic_i} \in D_{c_i})$, for $h = 1, \ldots, p$. 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 paper, we denote by $\bar{S}(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini})$ 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 behaviour of the proposed algorithm based on the usage of the matrix $\bar{S}(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini})$ in step 2 and 3.

As with graphical lasso, our proposed method requires a sequence of ρ-values, which should be suitably defined so as to reduce the computational cost needed to compute the entire path of the estimated parameters. The next theorem gives the exact formula for the derivation of the largest ρ-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:

$$\{\hat{\mu}_h, \hat{\sigma}_h^2\} = \arg\max_{\mu_h, \sigma_h^2} \sum_{i \in o_h} \log \phi(x_{ih}; \mu_h, \sigma_h^2) + |c_h^-| \int_{-\infty}^{l_h} \phi(x; \mu_h, \sigma_h^2) \mathrm{d}x + |c_h^+| \int_{u_h}^{+\infty} \phi(x; \mu_h, \sigma_h^2) \mathrm{d}x.$$

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

A proof of this theorem is reported in the supplementary material available at *Biostatistics* online. Theorem 2 shows how the maximum value of ρ can be efficiently computed using p optimization problems which do not involve the tuning parameter, and then calculating $S(\hat{\boldsymbol{\mu}}, \widehat{\Theta})$. 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 ρ-value;
2. Compute a decreasing sequence $\{\rho_{(k)}\}_{k=1}^K$ of distinct ρ-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 $\widehat{\Theta}_{ini}^{\rho_{(k)}} = \widehat{\Theta}^{\rho_{(k-1)}}$
5. Use the EM algorithm to compute $\{\hat{\boldsymbol{\mu}}^{\rho_{(k)}}, \widehat{\Theta}^{\rho_{(k)}}\}$ with $\{\hat{\boldsymbol{\mu}}_{ini}^{\rho_{(k)}}, \widehat{\Theta}_{ini}^{\rho_{(k)}}\}$ as starting values;
6. end for

The previous description shows that the entire path can be computed using the estimates obtained for a given ρ-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*, 2014). In our model, this strategy turns out to be remarkably efficient since using a sufficiently fine sequence of ρ-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 paper we propose to select the optimal ρ-value of the cglasso estimator by using the extended Bayesian Information Criterion (Foygel and Drton 2010). For our proposed estimator, this is given by:

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

where $\hat{\boldsymbol{\mu}}$ and $\widehat{\Theta}(\widehat{\mathcal{E}}^\rho)$ are the maximum likelihood estimators of the Gaussian graphical model specified by $\widehat{\mathcal{E}}^\rho = \{(h,k) : \hat{\theta}^\rho_{hk} \neq 0\}$, and $a(\rho)$ denotes the number of nonzero off-diagonal estimates of $\widehat{\Theta}^\rho$. 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 ρ-value that minimizes $\mathrm{BIC}_\gamma(\widehat{\mathcal{E}}^\rho)$, 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 ρ-value by minimizing the following approximate measure:

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

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 behaviour of the $\mathrm{BIC}_\gamma(\widehat{\mathcal{E}}^\rho)$ and $\overline{\mathrm{BIC}}_\gamma(\widehat{\mathcal{E}}^\rho)$ 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 behaviour of the approximate EM algorithm*

In a first simulation, we evaluate the effects of the approximation (3.8) 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 $h = 1, \ldots, p$. 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 θ_{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 $\mathcal{I} = \{1, \ldots, p\}$ and for each $h \in \mathcal{D}$ the value of the parameter μ_h is such that $\text{Prob}\{R(X_h; -\infty, 40) = +1\} = 0.25$ while for each $h \notin \mathcal{D}$ the parameter μ_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 behaviour 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 1×10^{-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 censored variables but, when we use the matrix $S(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini})$ in the E-step, the figure 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 $\bar{S}(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini})$ to approximate the current estimate of the empirical covariance matrix, the figure shows an almost linear dependence of the CPU time, which implies a significant reduction of the computational complexity compared to 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 $\hat{\boldsymbol{\mu}}^{\rho}_{e}$ and $\hat{\boldsymbol{\mu}}^{\rho}_{a}$, denoted as $\max_{\rho} \|\Delta\hat{\boldsymbol{\mu}}^{\rho}\|^2$, and the largest Frobenius distance between $\hat{\Theta}^{\rho}_{e}$ and $\hat{\Theta}^{\rho}_{a}$, denoted by $\max_{\rho} \|\Delta\hat{\Theta}^{\rho}\|^2_F$. 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 ℓ_1-penalised 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 Θ. Finally, we set the mean $\boldsymbol{\mu}$ 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 μ_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 behaviour 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 behaviour 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 Θ 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 ρ-values. Panel (a) in Figure 1 shows the precision-recall curves for Model 1 with $H = 25$. The curves report the relationship between precision and recall for any ρ-value, which are defined by:

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

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. Panel (b) shows the distributions of the quantity $\min_\rho \text{MSE}(\widehat{\Theta}^{\rho})$, 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 behaviour 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 $\min_\rho \text{MSE}(\hat{\boldsymbol{\mu}}^\rho)$) and the third meta-column reports the Area Under the precision-recall Curve (AUC). Note that the considered measures allow us to study the behaviour 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 $\boldsymbol{\mu}$ and Θ. We did not report $\min_\rho \text{MSE}(\hat{\boldsymbol{\mu}}^\rho)$ for glasso since this method does not allow to estimate the parameter $\boldsymbol{\mu}$. 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 values first and then infer the network

using graphical lasso on the imputed values. As in Städler and Bühlmann (2012), we consider k-nearest neighbour 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 network selected by $\overline{\mathrm{BIC}}_{0.5}(\widehat{\mathcal{E}}^{\rho})$ 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 2 housekeeping genes provided. Panel (a) in Figure 2 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. Panel (b) in Figure 2 shows the path of the $\overline{\text{BIC}}_{0.5}(\widehat{\mathcal{E}}^{\rho})$ measure and the vertical dashed line is traced in correspondence of the optimal ρ-value. The resulting estimated graph (see Panel (c) in Figure 2) 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 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 CD64

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 paper, 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 ℓ_1-penalisation 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 Materials

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

Funding

The project was partially supported by the *"European Cooperation in Science & Technology"* (COST) funding: action number CA15109.

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). ℓ_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, Peter J. and Levina, Elizaveta. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics* **36**(1), 199–227.

Brown, M. and Wittwer, C. (2000). Flow cytometry: principles and clinical applications in hematology. *Clinical Chemistry* **46**(8), 1221—1229.

Chan, T. E., Stumpf, M. P. H. and Babtie, A. C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. *Cell systems* **5**(3), 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), 1–38.

Derveaux, S., Vandesompele, J. and Hellemans, J. (2010). How to do successful gene expression analysis using real-time PCR. *Methods* **50**(4), 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 23*. pp. 604–612.

Friedman, J. H., Hastie, T. and Tibshirani, R. (2008). Sparce inverse covariance estimation with the graphical lasso. *Biostatistics* **9**(3), 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), 1–22.

Friedman, J. H., Hastie, T. and Tibshirani, R. (2014). *glasso: Graphical lasso- estimation of Gaussian graphical models*.

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**(5629), 102–105.

Genz, A. and Bretz, F. (2002). Comparison of methods for the computation of the multivariate t probabilities. *Journal of Computational and Graphical Statistics* **11**(4), 950–971.

Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2015). Graphical models for ordinal data. *Journal of Computational and Graphical Statistics* **24**(1), 183–204.

Hoffman, Jeather J. and Johnson, Robert E. (2015). Pseudo-likelihood estimation of multivariate normal parameters in the presence of left-censored data. *Journal of Agricultural, Biological, and Environmental Statistics* **20**(1), 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**(484), 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, Cosmin. (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**(9), 2816–2829.

Lee, L.-F. (1983). The determination of moments of the doubly truncated multivariate normal tobit model. *Economics Letters* **11**(3), 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**(3), 543–553.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, second edition 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**(16), 2310–2316.

McLachlan, G. and Krishnan, T. (2008). *The EM Algorithm and Exetnsions*, second edition 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 challange. *PLoS One* **5**(2), e14147.

Perkins, Neil J., Schisterman, Enrique F. and Vexier, Albert. (2013). Multivariate normally distributed biomarkers subject to limits of detection and receiver operating characteristic curve inference. *Academic Radiology* **20**(7), 838–846.

Pesonen, Maiju, Pesonen, Henri and Nevalainen, Jaakko. (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**(8), e0182832.

Psaila, B., Barkas, N., Iskander, D., Roy, A., Anderson, S., Ashley, N., Caputo, V. S., Lichtenberg, J., Loaiza, S., Bodine, D. M., Karadimitris, A., Mead, A. J. *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. (2000). *Analysis of Incomplete Multivariate Data*, Monographs on statistics and applied probability. 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**(1), 219–235.

Uhler, C. (2012). Geometry of maximum likelihood estimation in Gaussian graphical models. *The Annals of Statistics* **40**(12), 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**(3), 193–212.

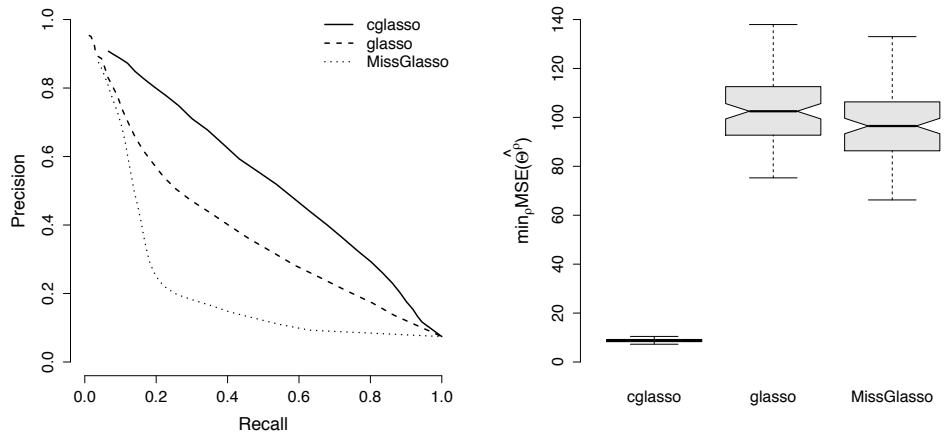
Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelić, A., von Rohr, P., Thiele, L., Zitzlerand, E., Gruissem, W. *and others*. (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biology* **5**(11), R92.

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**(4), 892–900.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**(1), 19–35.

Zhao, T., Li, Xi., Liu, H., Roeder, K., Lafferty, J. and Wasserman, L. (2015). *huge: High-Dimensional Undirected Graph Estimation*. R package version 1.2.7.

[*Received August 1, 2010; revised October 1, 2010; accepted for publication November 1, 2010*]

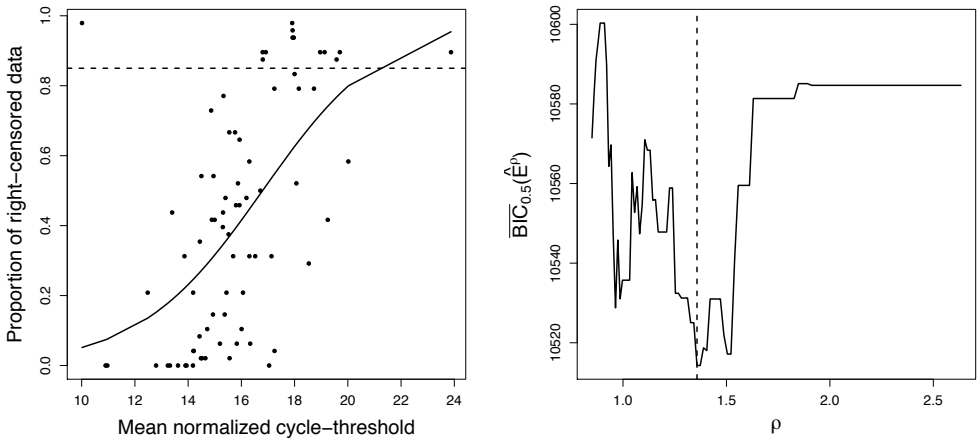


(a) (b)

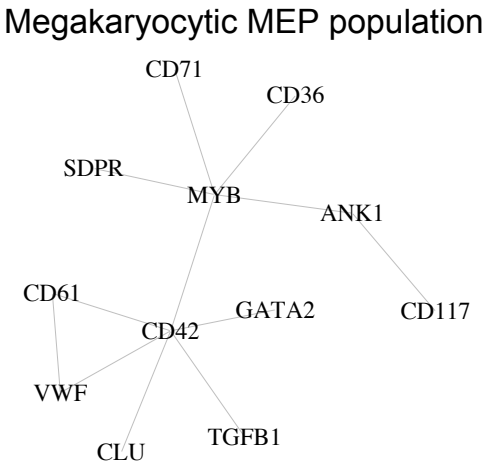
Fig. 1. Results of the comparative simulation study based on Model 1 with $H = 25$. Panel (a) shows the average precision-recall curves; Panel (b) shows the box-plots of $\min_\rho \mathrm{MSE}(\widehat{\Theta}^\rho)$ for the considered estimators.

Table 1. *Results of the simulation study on evaluating the effect of the approximation (3.8): first part reports the average CPU time for computing a path whereas second part reports the average of* $\max_\rho \|\Delta\hat{\boldsymbol{\mu}}^\rho\|^2$ *and* $\max_\rho \|\Delta\hat{\Theta}^\rho\|_F^2$ *. Standard deviations are shown in brackets.*

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



(a) (b)



(c)

Fig. 2. Real data analysis: Panel (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. Panel (b): path of the $\overline{\mathrm{BIC}}_{0.5}(\widehat{\mathcal{E}}^{\rho})$ measure; vertical dashed line identifies the optimal value of the tuning parameter. Panel (c): the estimated graph.

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			$\min_\rho \text{MSE}(\hat{\boldsymbol{\mu}}^\rho)$		$\min_\rho \text{MSE}(\widehat{\Theta}^\rho)$			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)

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

	% 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

Biostatistics (2018), **0**, 0, *pp.* 1–11
doi:10.1093/biostatistics/SupplementaryMaterial˙14˙05˙2018

ℓ_1-Penalized Censored Gaussian Graphical Model Supplementary Materials

Luigi Augugliaro*

Department of Economics, Business and Statistics, University of Palermo, Building 13, Viale delle Scienze, 90128 Palermo, Italy

luigi.augugliaro@unipa.it

Antonino Abbruzzo

Department of Economics, Business and Statistics, University of Palermo, Building 13, Viale delle Scienze, 90128 Palermo, Italy

Veronica Vinciotti

Department of Mathematics, Brunel University London, Uxbridge UB8 3PH, United Kingdom

1. Evaluating the behaviour of the extended Bayesian Information Criteria

In this section we evaluate, by a simulation study, the behaviour of the two measures of goodness-of-fit proposed in Section 3.3 to select the optimal value of the tuning parameter of the cglasso estimator: the extended Bayesian information criterion based on the log-likelihood function and that based on the Q-function, respectively. These are defined by:

$$\begin{aligned}\mathrm{BIC}_\gamma(\widehat{\mathcal{E}}^\rho) &= -2\sum_{i=1}^{n}\log\varphi(\boldsymbol{x}_{io_i},\boldsymbol{r}_i;\hat{\boldsymbol{\mu}},\widehat{\Theta}(\widehat{\mathcal{E}}^\rho)) + a(\rho)(\log n + 4\gamma\log p),\\ \overline{\mathrm{BIC}}_\gamma(\widehat{\mathcal{E}}^\rho) &= -n[\log\det\widehat{\Theta}^\rho - \mathrm{tr}\{\Theta S(\hat{\boldsymbol{\mu}},\widehat{\Theta}(\widehat{\mathcal{E}}^\rho))\}] + a(\rho)(\log n + 4\gamma\log p).\end{aligned}$$

*To whom correspondence should be addressed.

Our simulation study is based on a right-censored Gaussian graphical model with censoring value fixed to 40. Similarly to the simulation study used in Foygel and Drton (2010) in the case of a standard Gaussian graphical model, we vary the sample size n in the set $\{100, 200, 300, 400\}$ and, in each case, we fix the number of variables by $p = n^k$, with $k = 0.5$. As done in Section 4.1, the elements of the parameter $\boldsymbol{\mu}$ are used to fix the probability of right censoring. More specifically, we randomly draw a subset $\mathcal{D}$ from $\mathcal{I} = \{1, \ldots, p\}$ containing the 5% of the p indices then μ_h is set in such a way that $\text{Prob}\{R(X_h; -\infty, 40) = +1\} \in \{0.25, 0.50\}$, for each $h \in \mathcal{D}$. The remaining elements of the parameter $\boldsymbol{\mu}$ are such that the probability of right censoring is approximately equal to 10^{-11}. In order to evaluate the effects of the probability of censoring on the behaviour of the two considered criteria of goodness-of-fit, we model Θ as done in Section 3.1 of Foygel and Drton (2010), i.e., we use the following two models:

- **Model 1**: Θ is a symmetric tridiagonal matrix with $\theta_{h(h+1)} = 0.3$; the remaining off-diagonal elements are equal to zero.

- **Model 2**: in this model we set $\theta_{h(h+1)} = 0.2$, $\theta_{h(h+2)} = 0.1$ and the remaining off-diagonal elements are equal to zero.

The diagonal elements of the two concentration matrices are equal to 1. For each model, we simulate 100 samples from the censored Gaussian graphical model and in each simulation the cglasso model is fitted using a sequence of 100 evenly spaced ρ-values, with values between $\rho_{\max}/100$ and $\rho_{\max}$. Then, we use $\text{BIC}_\gamma(\widehat{\mathcal{E}}^\rho)$ to select the best model in terms of goodness-of-fit, i.e., the estimated model with the lowest $\text{BIC}_\gamma(\widehat{\mathcal{E}}^\rho)$ measure along the path; a second model is selected as previously done, but by using $\overline{\text{BIC}}_\gamma(\widehat{\mathcal{E}}^\rho)$. Finally, the behaviour of these criteria is evaluated computing the true positive rate (TPR)

$$\text{TPR} = \frac{\text{number of } \hat{\theta}^\rho_{hk} \neq 0 \text{ and } \theta_{hk} \neq 0}{\text{number of } \theta_{hk} \neq 0},$$

and the false discovery rate (FDR)

$$\text{FDR} = \frac{\text{number of } \hat{\theta}^{\rho}_{hk} \neq 0 \text{ and } \theta_{hk} = 0}{\text{number of } \hat{\theta}^{\rho}_{hk} \neq 0}.$$

Table 1 and 2 report the results of the simulation study for Model 1 and 2, respectively. A first result arising from the analysis of these tables is that, for any combination of the parameters, the two criteria have the same behaviour both in terms of TPR and FPR. In particular the behaviour seems to be independent on the level of censoring probability. Focusing on the results of the first model, we observe that, for large sample sizes, the recovery of non-zero coefficients is perfect or near perfect for any value of the parameter γ. On the other hand, we have a significant improvement in terms of FDR when we use a γ-value greater than zero. In particular, when the sample size is large, the extended Bayesian information criterion clearly outperforms the classical BIC measure. This behaviour is also shown in Panel(a) of Figure (1). When we look at the results from Model 2, as observed also in Foygel and Drton (2010), we have a noticeable reduction in TPR as γ increases, even if Panel (b) in Figure (1) shows that it is an increasing function of the sample size. Finally, also in this case the classical BIC measure is characterized by an extremely high FDR for any value of the sample size.

In conclusion, for any combination of the parameters, $\text{BIC}_\gamma(\widehat{\mathcal{E}}^{\rho})$ and $\overline{\text{BIC}}_\gamma(\widehat{\mathcal{E}}^{\rho})$ tend to select the same estimated model for any γ-value. This result suggests that in practice the $\overline{\text{BIC}}_\gamma(\widehat{\mathcal{E}}^{\rho})$ measure is preferred to $\text{BIC}_\gamma(\widehat{\mathcal{E}}^{\rho})$ since its computation requires only quantities that are a byproduct of the EM algorithm. Furthermore, the simulation study seems to confirm what was originally obtained in Foygel and Drton (2010) in the case of a standard Gaussian graphical model, suggesting that the optimal asymptotic properties of the extended Bayesian criteria are mantained also for the proposed cglasso estimator.

2. PROOFS

2.1 *Proof of Theorem 1*

We start by deriving the first equation in the system (3.6). Consider the derivative of the Lagrangian function

$$\mathcal{L}_\rho(\boldsymbol{\mu}, \Theta) = \frac{1}{n}\sum_{i=1}^{n} \log \varphi(\boldsymbol{x}_{io_i}, \boldsymbol{r}_i; \boldsymbol{\mu}, \Theta) - \rho \sum_{h\neq k} |\theta_{hk}|, \tag{2.1}$$

with respect to μ_h. Formally

$$\begin{aligned}\frac{\partial \mathcal{L}_\rho(\boldsymbol{\mu}, \Theta)}{\partial \mu_h} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \mu_h} \log \int_{D_{c_i}} \phi(\boldsymbol{x}_{io_i}, \boldsymbol{x}_{ic_i}; \boldsymbol{\mu}, \Theta) \mathrm{d}\boldsymbol{x}_{ic_i} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{\int_{D_{c_i}} \{\partial \phi(\boldsymbol{x}_{io_i}, \boldsymbol{x}_{ic_i}; \boldsymbol{\mu}, \Theta)/\partial \mu_h\} \mathrm{d}\boldsymbol{x}_{ic_i}}{\int_{D_{c_i}} \phi(\boldsymbol{x}_{io_i}, \boldsymbol{x}_{ic_i}; \boldsymbol{\mu}, \Theta) \mathrm{d}\boldsymbol{x}_{ic_i}}. \end{aligned} \tag{2.2}$$

Writing $\phi(\boldsymbol{x}_{io_i}, \boldsymbol{x}_{ic_i}; \boldsymbol{\mu}, \Theta)$ as the exponential of the log-likelihood of the multivariate Gaussian distribution, we have

$$\frac{\partial \phi(\boldsymbol{x}_{io_i}, \boldsymbol{x}_{ic_i}; \boldsymbol{\mu}, \Theta)}{\partial \mu_h} = \phi(\boldsymbol{x}_{io_i}, \boldsymbol{x}_{ic_i}; \boldsymbol{\mu}, \Theta) \sum_{k=1}^{p} \theta_{hk}(x_{ik} - \mu_k),$$

from which expression (2.2) can be written as

$$\begin{aligned}\frac{\partial \mathcal{L}_\rho(\boldsymbol{\mu}, \Theta)}{\partial \mu_h} &= \sum_{k=1}^{p} \theta_{hk} \frac{1}{n}\sum_{i=1}^{n} \int_{D_{c_i}} (x_{ik} - \mu_k) \frac{\phi(\boldsymbol{x}_{io_i}, \boldsymbol{x}_{ic_i}; \boldsymbol{\mu}, \Theta)}{\int_{D_{c_i}} \phi(\boldsymbol{x}_{io_i}, \boldsymbol{x}_{ic_i}; \boldsymbol{\mu}, \Theta) \mathrm{d}\boldsymbol{x}_{ic_i}} \mathrm{d}\boldsymbol{x}_{ic_i} \\ &= \sum_{k=1}^{p} \theta_{hk} \frac{1}{n}\sum_{i=1}^{n} \int_{D_{c_i}} (x_{ik} - \mu_k) \frac{\phi(\boldsymbol{x}_{ic_i}; \boldsymbol{\mu}_{c_i|o_i}, \Theta_{c_ic_i})}{\int_{D_{c_i}} \phi(\boldsymbol{x}_{ic_i}; \boldsymbol{\mu}_{c_i|o_i}, \Theta_{c_ic_i}) \mathrm{d}\boldsymbol{x}_{ic_i}} \mathrm{d}\boldsymbol{x}_{ic_i} \\ &= \sum_{k=1}^{p} \theta_{hk} \left\{ \frac{1}{n}\sum_{i=1}^{n} \int_{D_{c_i}} x_{ik} \frac{\phi(\boldsymbol{x}_{ic_i}; \boldsymbol{\mu}_{c_i|o_i}, \Theta_{c_ic_i})}{\int_{D_{c_i}} \phi(\boldsymbol{x}_{ic_i}; \boldsymbol{\mu}_{c_i|o_i}, \Theta_{c_ic_i}) \mathrm{d}\boldsymbol{x}_{ic_i}} \mathrm{d}\boldsymbol{x}_{ic_i} - \mu_k \right\}, \end{aligned} \tag{2.3}$$

where (2.3) is obtained by writing $\phi(\boldsymbol{x}_{io_i}, \boldsymbol{x}_{ic_i}; \boldsymbol{\mu}, \Theta)$ as the product of the marginal distribution of $\boldsymbol{X}_{io_i}$ and the conditional distribution of $\boldsymbol{X}_{ic_i}$ given $\boldsymbol{X}_{io_i} = \boldsymbol{x}_{io_i}$. Finally, by observing that $\phi(\boldsymbol{x}_{ic_i}; \boldsymbol{\mu}_{c_i|o_i}, \Theta_{c_ic_i}) / \int_{D_{c_i}} \phi(\boldsymbol{x}_{ic_i}; \boldsymbol{\mu}_{c_i|o_i}, \Theta_{c_ic_i}) \mathrm{d}\boldsymbol{x}_{ic_i}$ is the density function of the truncated normal distribution, resulting from the conditional distribution of $\boldsymbol{X}_{ic_i}$ given $\boldsymbol{X}_{io_i} = \boldsymbol{x}_{io_i}$ over the region D_{c_i}, and using the notation introduced in the first part of the Section 3.2, expression (2.3) can

be rewritten as

$$\frac{\partial \mathcal{L}_\rho(\boldsymbol{\mu}, \Theta)}{\partial \mu_h} = \sum_{k=1}^{p} \theta_{hk} \left\{ \frac{1}{n} \sum_{i=1}^{n} x_{i,k}(\boldsymbol{\mu}, \Theta) - \mu_k \right\}.$$

The first equation in the system (3.6) follows by simplifying the system with hth equation

$$\sum_{k=1}^{p} \hat{\theta}^{\rho}_{hk} \left\{ \frac{1}{n} \sum_{i=1}^{n} x_{i,k}(\hat{\boldsymbol{\mu}}^{\rho}, \widehat{\Theta}^{\rho}) - \hat{\mu}^{\rho}_{k} \right\} = 0,$$

for $h = 1, \ldots, p$. The second equation in system (3.6) is derived using the same strategy previously adopted to find expression (2.3) and by observing that the derivative of the Lagrangian function (2.1) with respect to θ_{hk} is equal to

$$\begin{aligned}
\frac{\partial \mathcal{L}_\rho(\boldsymbol{\mu}, \Theta)}{\partial \theta_{hk}} &= \sigma_{hk} - \left\{ \frac{1}{n} \sum_{i=1}^{n} \int_{D_{c_i}} x_{ih} x_{ik} \frac{\phi(\boldsymbol{x}_{ic_i}; \boldsymbol{\mu}_{c_i|o_i}, \Theta_{c_i c_i})}{\int_{D_{c_i}} \phi(\boldsymbol{x}_{ic_i}; \boldsymbol{\mu}_{c_i|o_i}, \Theta_{c_i c_i}) \mathrm{d}\boldsymbol{x}_{ic_i}} \mathrm{d}\boldsymbol{x}_{ic_i} - \mu_h \mu_k \right\} - \rho v_{hk} \\
&= \sigma_{hk} - \left\{ \frac{1}{n} \sum_{i=1}^{n} x_{i,hk}(\boldsymbol{\mu}, \Theta) - \mu_h \mu_k \right\} - \rho v_{hk},
\end{aligned}$$

where v_{hk} denotes the subgradient of the absolute value function at θ_{hk}, i.e., $v_{hk} = \mathrm{sign}(\theta_{hk})$ if $\theta_{hk} \neq 0$ otherwise it is a value inside the closed interval $[-1, 1]$. Finally, the second equation follows using the identity $\hat{\mu}^{\rho}_{h} = \bar{x}_h(\hat{\boldsymbol{\mu}}^{\rho}, \widehat{\Theta})$ and the definition of $s_{hk}(\hat{\boldsymbol{\mu}}^{\rho}, \widehat{\Theta}^{\rho})$ in equation

$$\hat{\sigma}^{\rho}_{hk} - \left\{ \frac{1}{n} \sum_{i=1}^{n} x_{i,hk}(\hat{\boldsymbol{\mu}}^{\rho}, \widehat{\Theta}^{\rho}) - \hat{\mu}^{\rho}_{h} \hat{\mu}^{\rho}_{k} \right\} - \rho \hat{v}_{hk} = 0.$$

2.2 *Proof of Theorem 2*

Let us suppose that we know the maximum value of the tuning parameter, which we denote as $\rho_{\max}$. By definition, at $\rho_{\max}$ the cglasso estimator of the concentration matrix is a diagonal matrix, implying that $\hat{\mu}^{\rho_{\max}}_{h}$ and $\hat{\theta}^{\rho_{\max}}_{hh}$ can be estimated through the maximum likelihood estimators of the parameters of the marginal distributions, i.e.

$$\{\hat{\mu}_h, \hat{\sigma}^2_h\} = \arg\max_{\mu_h, \sigma^2_h} \sum_{i \in o_h} \log \phi(x_{ih}; \mu_h, \sigma^2_h) + |c^-_h| \int_{-\infty}^{l_h} \phi(x; \mu_h, \sigma^2_h) \mathrm{d}x + |c^+_h| \int_{u_h}^{+\infty} \phi(x; \mu_h, \sigma^2_h) \mathrm{d}x,$$

where $o_h = \{i : r_{ih} = 0\}$, $c^-_h = \{i : r_{ih} = -1\}$ and $c^+_h = \{i : r_{ih} = +1\}$. Then $\hat{\mu}^{\rho_{\max}}_{h}$ is equal to $\hat{\mu}_h$ and $\hat{\theta}^{\rho_{\max}}_{hh}$ is equal to the inverse of $\hat{\sigma}^2_h$. Finally, $\rho_{\max}$ comes from the analysis of the

second stationary condition reported in Theorem 1. Since $\widehat{\Theta}^{\rho_{\max}}$ is a diagonal matrix we have, for each $h \neq k$, that $\hat{\sigma}_{hk}^{\rho_{\max}}(\hat{\boldsymbol{\mu}}^{\rho_{\max}}, \widehat{\Theta}^{\rho_{\max}}) = 0$. Then the second equation of the system reported in Theorem 1 can be written as follows:

$$s_{hk}(\hat{\boldsymbol{\mu}}^{\rho_{\max}}, \widehat{\Theta}^{\rho_{\max}}) = -\rho_{\max}\hat{v}_{hk}.$$

Considering the absolute value of both sides of this equation and remembering that $|\hat{v}_{hk}| \leqslant 1$, since $\hat{\theta}_{hk}^{\rho_{\max}} = 0$ for each $h \neq k$, we have that

$$|s_{hk}(\hat{\boldsymbol{\mu}}^{\rho_{\max}}, \widehat{\Theta}^{\rho_{\max}})| = \rho_{\max}|\hat{v}_{hk}| \leqslant \rho_{\max}.$$

The previous inequality is satisfied when we let $\rho_{\max} = \max_{h \neq k} |s_{hk}(\hat{\boldsymbol{\mu}}^{\rho_{\max}}, \widehat{\Theta}^{\rho_{\max}})|$.

3. Additional figures to Section 4.2

In this section, we report additional results to those presented in Section 4.2 of the paper, with Models 1, 2 and 3 defined as in that section.

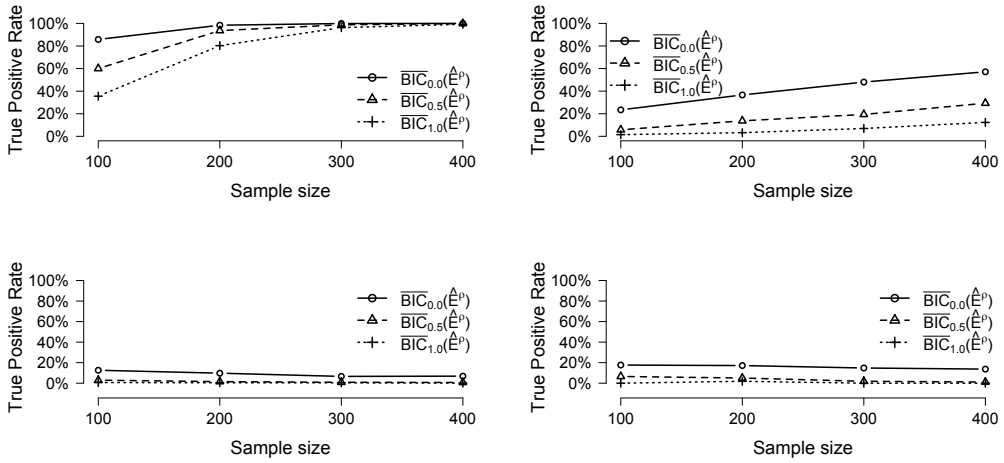
References

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 23*. pp. 604–612.

[*Received August 1, 2010; revised October 1, 2010; accepted for publication November 1, 2010*]

Table 1. *Results of the simulation study based on Model 1: for each measure used to evaluate the behaviour of the extended Bayesian information criteria we report average values across the 100 iterations and the standard deviation between parentheses*

		$\mathrm{BIC}_\gamma(\widehat{\mathcal{E}}^\rho)$						$\overline{\mathrm{BIC}}_\gamma(\widehat{\mathcal{E}}^\rho)$					
		TPR			FDR			TPR			FDR		
	γ	0.0	0.5	1.0	0.0	0.5	1.0	0.0	0.5	1.0	0.0	0.5	1.0
		$\mathrm{Prob}\{R(X_h;-\infty,40)=+1\}=0.25$											
n	100	0.86	0.59	0.35	0.12	0.03	0.01	0.86	0.60	0.36	0.13	0.03	0.01
		(0.11)	(0.20)	(0.20)	(0.10)	(0.06)	(0.04)	(0.11)	(0.21)	(0.20)	(0.10)	(0.07)	(0.04)
	200	0.98	0.93	0.80	0.09	0.01	0.00	0.98	0.94	0.80	0.10	0.02	0.00
		(0.04)	(0.07)	(0.12)	(0.08)	(0.03)	(0.02)	(0.04)	(0.07)	(0.11)	(0.09)	(0.04)	(0.02)
	300	1.00	0.99	0.96	0.06	0.01	0.00	1.00	0.99	0.96	0.07	0.01	0.00
		(0.01)	(0.03)	(0.05)	(0.07)	(0.02)	(0.01)	(0.01)	(0.03)	(0.05)	(0.07)	(0.02)	(0.01)
	400	1.00	1.00	0.99	0.07	0.01	0.00	1.00	1.00	0.99	0.07	0.01	0.00
		(0.00)	(0.01)	(0.02)	(0.07)	(0.02)	(0.01)	(0.00)	(0.01)	(0.02)	(0.07)	(0.02)	(0.01)
		$\mathrm{Prob}\{R(X_h;-\infty,40)=+1\}=0.50$											
n	100	0.84	0.57	0.33	0.13	0.03	0.01	0.85	0.59	0.36	0.14	0.03	0.01
		(0.13)	(0.20)	(0.20)	(0.11)	(0.07)	(0.04)	(0.12)	(0.21)	(0.21)	(0.11)	(0.07)	(0.05)
	200	0.98	0.93	0.78	0.09	0.02	0.01	0.98	0.92	0.80	0.10	0.02	0.01
		(0.04)	(0.07)	(0.12)	(0.09)	(0.04)	(0.02)	(0.04)	(0.08)	(0.12)	(0.09)	(0.04)	(0.02)
	300	1.00	0.98	0.96	0.07	0.01	0.01	1.00	0.99	0.96	0.08	0.02	0.01
		(0.01)	(0.04)	(0.05)	(0.07)	(0.03)	(0.02)	(0.01)	(0.03)	(0.05)	(0.07)	(0.03)	(0.02)
	400	1.00	1.00	0.99	0.07	0.01	0.00	1.00	1.00	0.99	0.07	0.01	0.01
		(0.00)	(0.02)	(0.03)	(0.07)	(0.02)	(0.01)	(0.00)	(0.02)	(0.02)	(0.08)	(0.02)	(0.01)

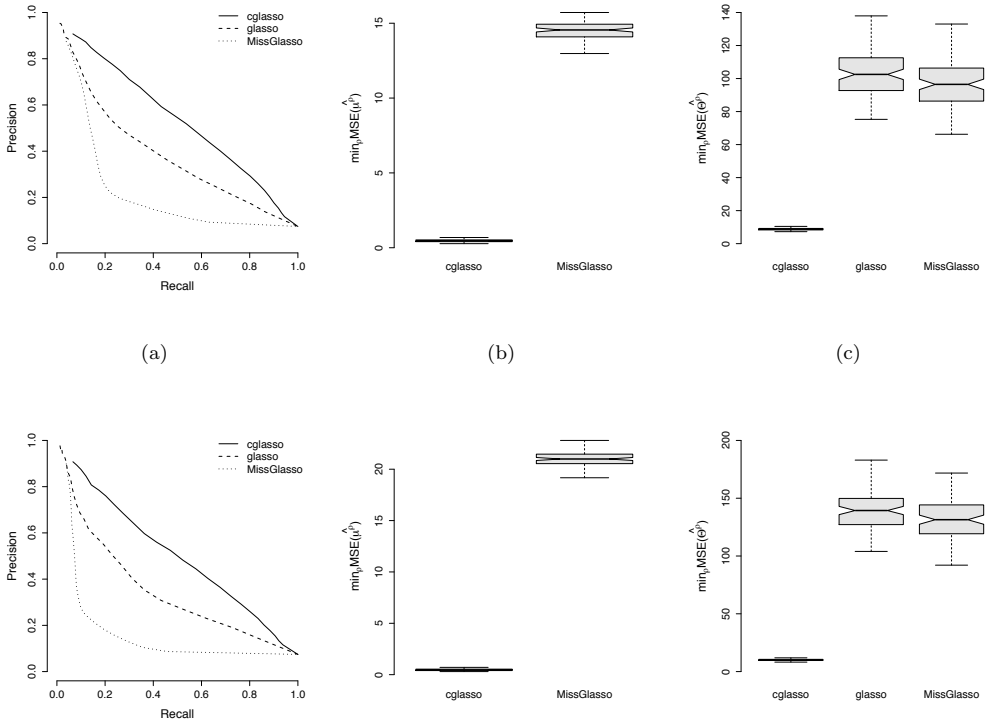


(a) (b)

Fig. 1. True positive rate and false discovery rate as a function of the sample size. Panel (a) refers to Model 1 with censoring probability equal to 0.25. Panel (b) refers to Model 2 with censoring probability equal to 0.25.

Table 2. *Results of the simulation study based on Model 2: for each measure used to evaluate the behaviour of the extended Bayesian information criteria we report average values across 100 iterations and the standard deviation between parentheses.*

		$\mathrm{BIC}_\gamma(\widehat{\mathcal{E}}^\rho)$						$\overline{\mathrm{BIC}}_\gamma(\widehat{\mathcal{E}}^\rho)$					
		TPR			FDR			TPR			FDR		
	γ	0.0	0.5	1.0	0.0	0.5	1.0	0.0	0.5	1.0	0.0	0.5	1.0
		$\mathrm{Prob}\{R(X_h;-\infty,40)=+1\}=0.25$											
n	100	0.23	0.06	0.01	0.18	0.06	0.00	0.23	0.06	0.01	0.18	0.07	0.00
		(0.10)	(0.05)	(0.03)	(0.18)	(0.19)	(0.00)	(0.10)	(0.05)	(0.03)	(0.17)	(0.20)	(0.00)
	200	0.37	0.13	0.03	0.16	0.05	0.02	0.37	0.14	0.03	0.17	0.05	0.02
		(0.10)	(0.07)	(0.04)	(0.11)	(0.15)	(0.14)	(0.10)	(0.07)	(0.04)	(0.10)	(0.15)	(0.14)
	300	0.47	0.19	0.07	0.14	0.02	0.00	0.48	0.19	0.07	0.15	0.02	0.00
		(0.09)	(0.07)	(0.04)	(0.08)	(0.06)	(0.00)	(0.09)	(0.07)	(0.04)	(0.08)	(0.06)	(0.00)
	400	0.57	0.29	0.12	0.13	0.01	0.00	0.57	0.29	0.12	0.14	0.01	0.00
		(0.08)	(0.08)	(0.05)	(0.06)	(0.03)	(0.00)	(0.08)	(0.08)	(0.05)	(0.06)	(0.03)	(0.00)
		$\mathrm{Prob}\{R(X_h;-\infty,40)=+1\}=0.50$											
n	100	0.23	0.06	0.01	0.18	0.07	0.00	0.24	0.06	0.02	0.18	0.07	0.00
		(0.11)	(0.05)	(0.03)	(0.18)	(0.20)	(0.00)	(0.10)	(0.06)	(0.03)	(0.18)	(0.20)	(0.00)
	200	0.35	0.13	0.03	0.17	0.05	0.02	0.36	0.13	0.03	0.18	0.05	0.02
		(0.10)	(0.07)	(0.03)	(0.10)	(0.15)	(0.14)	(0.10)	(0.08)	(0.04)	(0.10)	(0.14)	(0.13)
	300	0.48	0.19	0.07	0.15	0.02	0.00	0.48	0.19	0.07	0.15	0.02	0.00
		(0.09)	(0.07)	(0.04)	(0.08)	(0.06)	(0.00)	(0.09)	(0.08)	(0.04)	(0.08)	(0.06)	(0.00)
	400	0.56	0.28	0.12	0.13	0.01	0.00	0.57	0.29	0.12	0.14	0.01	0.00
		(0.08)	(0.08)	(0.05)	(0.06)	(0.03)	(0.00)	(0.08)	(0.08)	(0.06)	(0.06)	(0.03)	(0.00)



(a) (b) (c)

(d) (e) (f)

Fig. 2. Results of the simulation study based on Model 1: Panel (a), (b) and (c) refer to the setting with $H = 25$ while Panel (c), (d) and (f) refer to the setting with $H = 35$.

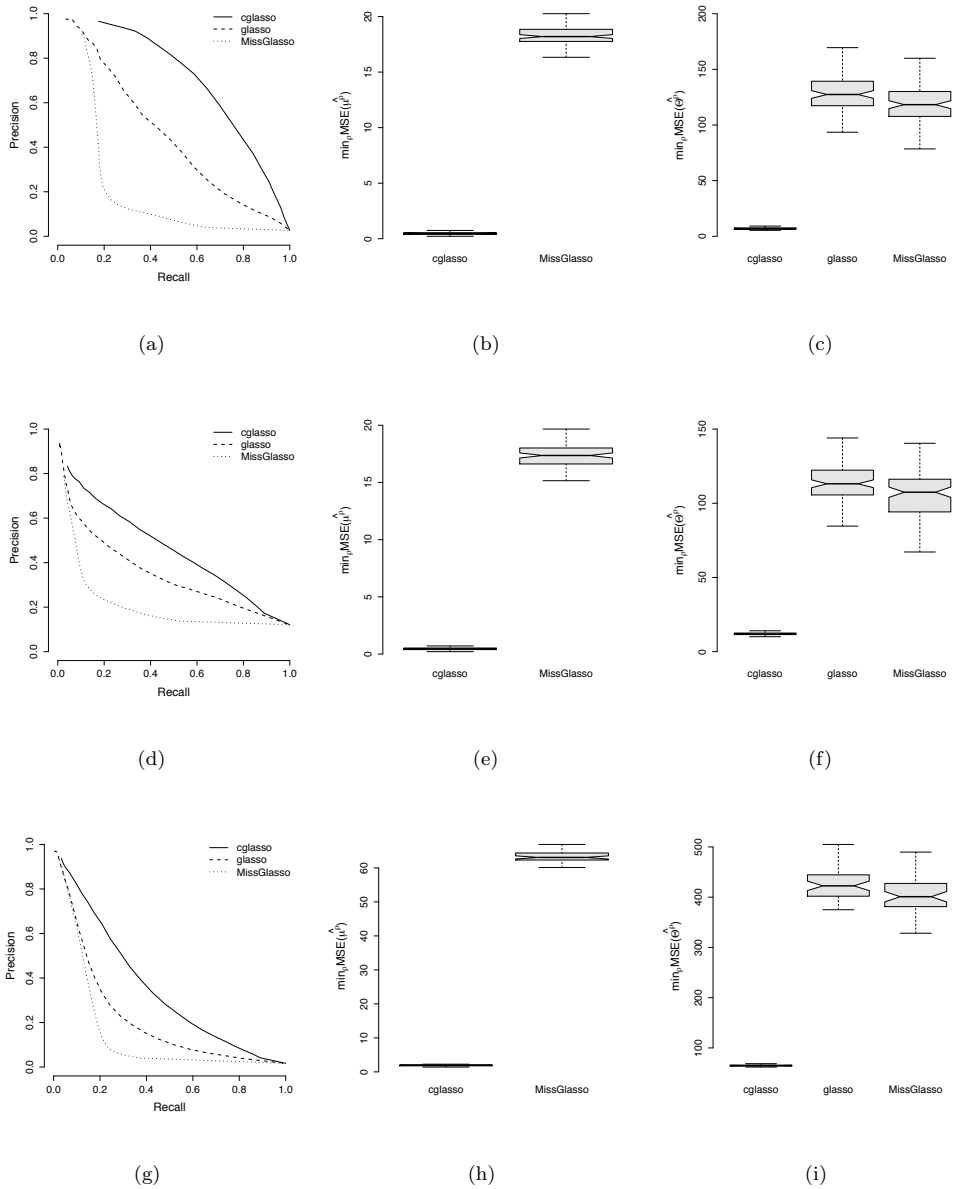


Fig. 3. Results of the simulation study based on Model 2 and 3: Panel (a), (b) and (c) refer to Model 2 with setting $k = 1$; Panel (d), (e) and (f) refer to Model 2 with setting $k = 2$; Panel (g), (h) and (k) refer to Model 3.