CLADAG 2021

BOOK OF ABSTRACTS AND SHORT PAPERS

13th Scientific Meeting of the Classification and Data Analysis Group
Firenze, September 9-11, 2021

edited by
Giovanni C. Porzio
Carla Rampichini
Chiara Bocci

FIRENZE
UNIVERSITY
PRESS

Sparse Inference in Covariate Adjusted Censored Gaussian Graphical Models

Luigi Augugliaro[1], Gianluca Sottile[1] and Angelo M. Mineo[1]

[1] Dep. of Economics, Business and Statistics, University of Palermo, Italy,
(e-mail: luigi.augugliaro@unipa.it, angelo.mineo@unipa.it,
gianluca.sottile@unipa.it)

Abstract: The covariate adjusted glasso 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. In this paper we propose an extension to censored data.

Keywords: censored data, censored glasso estimator, Gaussian graphical model, glasso estimator.

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. Gaussian graphical models (GGM, Lauritzen (1996)) have been widely used for reconstructing a genetic network from expression data. The reason of such diffusion relies on the statistical properties of the multivariate Gaussian distribution which allow the topological structure of a network to be related with 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 covariate adjusted glasso estimator (Yin & Li, 2011) is a popular method for estimating a sparse concentration matrix, based on the idea of adding an ℓ_1-penalty function to the likelihood function of the multivariate Gaussian distribution. Despite the widespread literature on the covariate adjusted glasso estimator, there is a great number of fields in applied research where the use of the graphical model is theoretically unfounded. For example in some cases data are left- or right-censored. In this paper we propose an extension of the covariate adjusted glasso estimator that takes into account the censoring mechanism of the data explicitly.

2 The covariate adjusted censored Gaussian graphical model

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^\top$ 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 $\boldsymbol{Y}$ 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 covariate adjusted Gaussian graphical model (CGGM) is an extension of the classical GGM based on the assumption that the conditional distribution of $\boldsymbol{Y}$ given a q-dimensional vector of predictors, say $\boldsymbol{X} = (X_1, \ldots, X_q)^\top$, follows a multivariate Gaussian distribution with expected value: $\boldsymbol{\mu}(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{x}$, where $\boldsymbol{\beta} = (\beta_{hk})$ is a matrix $q \times p$ coefficient matrix, and covariance matrix denoted by $\Sigma = (\sigma_{hk})$. Denoting with $\Theta = (\theta_{hk})$ the concentration matrix, i.e., the inverse of the covariance matrix, the conditional density function of $\boldsymbol{Y}$ can be written as follows:

$$\phi(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\beta}, \Theta) = (2\pi)^{-p/2} |\Theta|^{1/2} \exp[-1/2\{\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})\}^\top \Theta \{\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})\}]. \tag{1}$$

As shown in Lauritzen (1996), the off-diagonal elements of the concentration matrix are the parametric tools relating the pairwise Markov property to the factorization of the density (1), i.e., two random variables, say Y_h and Y_k, are conditionally independent given all the remaining variables if and only if θ_{hk} is equal to zero.

As done in Augugliaro *et al.* (2020), we assume that $\boldsymbol{Y}$ is a (partially) latent random vector with density function (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, Y_h is observed only if it is inside the interval $[l_h, u_h]$ otherwise it is censored from below if $Y_h < l_h$ or censored from above if $Y_h > u_h$. Using the approach for missing data with nonignorable mechanism (Little & Rubin (2002)) we denote the quantity $R(\boldsymbol{Y}; \boldsymbol{l}, \boldsymbol{u})$, to encode the censoring patterns, such that the hth element of $R(\boldsymbol{Y}; \boldsymbol{l}, \boldsymbol{u})$ is defined as $R(Y_h; l_h, u_h) = I(Y_h > u_h) - I(Y_h < l_h)$, where $I(\cdot)$ denotes the indicator function. By construction $R(\boldsymbol{Y}; \boldsymbol{l}, \boldsymbol{u})$ is a discrete random vector with support the set $\{-1, 0, 1\}^p$ and probability function $\Pr\{R(\boldsymbol{Y}; \boldsymbol{l}, \boldsymbol{u}) = \boldsymbol{r}\} = \int_{D_{\boldsymbol{r}}} \phi(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\beta}, \Theta) d\boldsymbol{y}$, where $D_{\boldsymbol{r}} = \{\boldsymbol{y} \in \mathbb{R}^p : R(\boldsymbol{y}; \boldsymbol{l}, \boldsymbol{u}) = \boldsymbol{r}\}$. Given a censoring pattern, we can simplify our notation by partitioning the set $I = \{1, \ldots, p\}$ into $o = \{h \in I : r_h = 0\}, c^- = \{h \in I : r_h = -1\}$ and $c^+ = \{h \in I : r_h = +1\}$ 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. As done in Augugliaro *et al.* (2020), the probability distribution of the observed data, denoted by $\varphi(\{\boldsymbol{y}_o, \boldsymbol{r}\} \mid \boldsymbol{x}; \boldsymbol{\beta}, \Theta)$, can be defined as follows:

$$\varphi(\{\boldsymbol{y}_o, \boldsymbol{r}\} | \boldsymbol{x}; \boldsymbol{\beta}, \Theta) = \int \phi(\{\boldsymbol{y}_o, \boldsymbol{y}_c\} | \boldsymbol{x}; \boldsymbol{\beta}, \Theta) \Pr\{R(\boldsymbol{Y}; \boldsymbol{l}, \boldsymbol{u}) = \boldsymbol{r} | \boldsymbol{Y} = \boldsymbol{y}\} d\boldsymbol{y}_c, \tag{2}$$

where $c = c^- \cup c^+$. Density (2) can be simplified by observing that $\Pr\{R(\boldsymbol{Y}; \boldsymbol{l}, \boldsymbol{u}) = \boldsymbol{r} \mid \boldsymbol{Y} = \boldsymbol{y}\}$ is equal to one if the censoring pattern encoded in $\boldsymbol{r}$ is equal to the pattern observed in $\boldsymbol{y}$, otherwise it is equal to zero, hence $\varphi(\{\boldsymbol{y}_o, \boldsymbol{r}\} \mid \boldsymbol{x}; \boldsymbol{\beta}, \Theta)$ can be rewritten as

$$\varphi(\{\boldsymbol{y}_o, \boldsymbol{r}\} | \boldsymbol{x}; \boldsymbol{\beta}, \Theta) = \int_{D_c} \phi(\{\boldsymbol{y}_o, \boldsymbol{y}_c\} | \boldsymbol{x}; \boldsymbol{\beta}, \Theta) d\boldsymbol{y}_c I(\boldsymbol{l}_o \leq \boldsymbol{y}_o \leq \boldsymbol{u}_o), \tag{3}$$

where $D_c = (-\infty, \boldsymbol{l}_{c^-}) \times (\boldsymbol{u}_{c^+}, +\infty)$. Using density (3), the covariate adjusted censored Gaussian graphical model (CCGGM) is defined as the set $\{\boldsymbol{Y}, R(\boldsymbol{Y}; \boldsymbol{l}, \boldsymbol{u}), \varphi(\{\boldsymbol{y}_o, \boldsymbol{r}\} \mid \boldsymbol{x}; \boldsymbol{\beta}, \Theta), \mathcal{G}\}$, where $\varphi(\{\boldsymbol{y}_o, \boldsymbol{r}\} | \boldsymbol{x}; \boldsymbol{\beta}, \Theta)$ factorizes according to the undirected graph $\mathcal{G}$.

3 The covariate adjusted censored glasso estimator

Suppose we have a sample of size n independent observations drawn from a CCGGM. For ease of exposition, we shall assume that $\boldsymbol{l}$ and $\boldsymbol{u}$ are fixed across the n observations. To simplify our notation the set of indices of the variables observed in the ith observation is denoted by $o_i = \{h \in I : r_{ih} = 0\}$, while $c_i^- = \{h \in I : r_{ih} = -1\}$ and $c_i^+ = \{h \in I : r_{ih} = +1\}$ denote the sets of indices associated to the left and right-censored data, respectively. Denoting by $\boldsymbol{r}_i$ the realization of the random vector $R(\boldsymbol{Y}_i; \boldsymbol{l}, \boldsymbol{u})$, the ith observed data is the vector $(\boldsymbol{y}_{io_i}^\top, \boldsymbol{x}_i^\top, \boldsymbol{r}_i^\top)^\top$. Using the density function (3), the observed log-likelihood function can be written as

$$\ell(\boldsymbol{\beta}, \Theta) = \sum_{i=1}^{n} \log \int_{D_{c_i}} \phi(\{\boldsymbol{y}_{io_i}, \boldsymbol{y}_{ic_i}\} | \boldsymbol{x}_i; \boldsymbol{\beta}, \Theta) d\boldsymbol{y}_{ic_i} = \sum_{i=1}^{n} \log \varphi(\{\boldsymbol{y}_{io_i}, \boldsymbol{r}_i\} | \boldsymbol{x}_i; \boldsymbol{\beta}, \Theta), \quad (4)$$

where $D_{c_i} = (-\infty, \boldsymbol{l}_{c_i^-}) \times (\boldsymbol{u}_{c_i^+}, +\infty)$ and $c_i = c_i^- \cup c_i^+$. 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 is limited.

We propose to estimate the parameters of the CCGGM by generalizing the approach proposed in Yin & Li (2011), i.e., by maximizing a new objective function defined by adding two lasso-type penalty functions to the observed log-likelihood (4). The resulting estimator, called covariate adjusted censored glasso estimator, is formally defined as

$$\{\hat{\boldsymbol{\beta}}^{\lambda}, \widehat{\Theta}^{\rho}\} = \arg \max_{\boldsymbol{\beta}, \Theta \succ 0} \frac{1}{n} \sum_{i=1}^{n} \log \varphi(\{\boldsymbol{y}_{io_i}, \boldsymbol{r}_i\} | \boldsymbol{x}_i; \boldsymbol{\beta}, \Theta) - \lambda \sum_{h,k} |\beta_{hk}| - \rho \sum_{h \neq k} |\theta_{hk}|, \quad (5)$$

where λ and ρ are two non-negative tuning parameters. The lasso penalty on $\boldsymbol{\beta}$ introduces sparsity in $\hat{\boldsymbol{\beta}}^{\lambda}$, while the tuning parameter ρ controls the amount of sparsity in the estimated concentration matrix $\widehat{\Theta}^{\rho} = (\hat{\theta}_{hk}^{\rho})$.

4 Simulation study

In this section, we compare our proposed estimator with MissGlasso (Städler & Bühlmann, 2012), which performs ℓ_1-penalized estimation under the assumption that the censored data are missing at random, and with the covariate adjusted glasso estimator (Yin & Li, 2011), where the empirical covariance matrix is calculated by imputing the missing values with the censoring values. These estimators are evaluated in terms of both recovering the structure of the true graph. We use the method implemented in the R package `huge` (Zhao *et al.*, 2020), to simulate a sparse concentration matrix with a random structure for $\boldsymbol{Y}$. We set the probability of observing a link between two nodes to k/p, where p is the number of responses and k is used to control the amount of sparsity in $\boldsymbol{\Theta}$. Moreover, we set the right censoring value to 40 for any variable and the sample size n to 100. The predictors matrix $\boldsymbol{X}$ is sampled from a multivariate gaussian distribution with zero expected value and sparse covariance matrix simulated as done for $\boldsymbol{Y}$. Each column of the true matrix of predictors $\boldsymbol{\beta}$ contains only two non-zero regression coefficients sampled from a uniform distribution on the interval $[0.3, 0.7]$. The values of the intercepts are chosen in such a way that H response variables are right censored with probability equal to 0.40. The quantities k, p, q and H are chosen according to the following cases:

- **Scenario 1**: $k = 3$, $p = 50$, $q = 10$ and $H = 25$. This setting is used to evaluate the effects of the number of censored variables on the behavior of the proposed estimators when $n > p$.
- **Scenario 2**: $k = 3$, $p = 150$, $q = 10$ and $H = 75$. This setting is used to evaluate the impact of the high dimensionality on the estimators ($p > n$).

For each scenario, we simulate 50 samples and in each simulation, we compute the coefficients path using cglasso, MissGlasso, and glasso. Each path is computed using an equally spaced sequence of ρ and λ-values. Moreover, the precision-recall curves and the area under the curves (AUCs) are computed for each Scenarios. Table 1 shows how cglasso gives a better estimate of the concentration and coefficient matrices in terms of AUCs, for any given value of the tuning parameters. We report only five evenly spaced values of λ and ρ.

Table 1. *Mean area under the curves across the sequence of* ρ *and* λ*-values under the specification of the two Scenarios (see row blocks). The first column block refers to the concentration matrix (*$\boldsymbol{\Theta}$*) when* λ *is fixed and the second refers to the coefficient matrix (*$\boldsymbol{\beta}$*) when* ρ *is fixed. In the first column (1), (2) and (3) refer to cglasso, MissGlasso and glasso algorithms, respectively.*

	$\lambda/\lambda_{\max}$					$\rho/\rho\text{max}$				
	0.00	0.25	0.50	0.75	1.00	0.00	0.25	0.50	0.75	1.00
(1)	0.546	0.429	0.139	0.103	0.101	0.844	0.877	0.883	0.882	0.885
(2)	0.239	0.199	0.086	0.073	0.073	0.745	0.764	0.766	0.767	0.768
(3)	0.414	0.218	0.097	0.092	0.091	0.813	0.847	0.864	0.866	0.866
(1)	0.418	0.094	0.037	0.035	0.035	0.794	0.930	0.931	0.929	0.933
(2)	0.329	0.098	0.033	0.031	0.030	0.753	0.830	0.831	0.830	0.831
(3)	0.321	0.040	0.033	0.032	0.031	0.751	0.902	0.906	0.907	0.907

References

AUGUGLIARO, L., ABBRUZZO, A., & V., VINCIOTTI. 2020. ℓ_1-Penalized censored Gaussian graphical model. *Biostatitistics.*, **21**(2), e1–e16.

LAURITZEN, S.L. 1996. *Graphical Models.* Oxford University Press, Oxford.

LITTLE, R.J.A., & RUBIN, D.B. 2002. *Statistical Analysis with Missing Data.* John Wiley & Sons, Inc., Hoboken.

STÄDLER, N., & BÜHLMANN, P. 2012. Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing.*, **22**(1), 219–235.

YIN, J., & LI, H. 2011. A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Annals of Applied Statistics.*, **5**(4), 2630–2650.

ZHAO, T., LI, X., LIU, H., ROEDER, K., LAFFERTY, J., & WASSERMAN, L. 2020. *huge: High-Dimensional Undirected Graph Estimation.* R package version 1.3.4.1.