SIS | 2022

51st Scientific Meeting
of the Italian Statistical Society

Caserta, 22-24 June

Università degli Studi della Campania *Luigi Vanvitelli*

Società Italiana di Statistica

www.unicampania.it

Book of the Short Papers

Editors: Antonio Balzanella, Matilde Bini, Carlo Cavicchia, Rosanna Verde

Città di Caserta

1222-2022 800 ANNI

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE STATISTICHE

STATA

sas

UNIVERSITÀ DEGLI STUDI DEL SANNIO Benevento

Pearson

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891932310

The Joint Censored Gaussian Graphical Lasso Model

Inferenza penalizzata del modello grafico Gaussiano congiunto

Gianluca Sottile, Luigi Augugliaro and Veronica Vinciotti

Abstract The Gaussian graphical model is one of the most used tools for inferring genetic networks. Nowadays, the data are often collected from different sources or under different biological conditions, resulting in heterogeneous datasets that exhibit a dependency structure that varies across groups. The complex structure of these data is typically recovered using regularized inferential procedures that use two penalties, one that encourages sparsity within each graph and the other that encourages common structures among the different groups. To this date, these approaches have not been developed for handling the case of censored data. However, these data are often generated by gene expression technologies such as RT-qPCR experiments. In this paper, we fill this gap and propose an extension of joint Gaussian graphical modelling to account for censored, or more generally missing, data.

Abstract *Il modello grafico Gaussiano è uno degli stimatori più utilizzati per fare inferenza sulle reti genetiche. Al giorno d'oggi, i dati raccolti sono spesso generati da diverse fonti o da diverse condizioni biologiche, risultando in dataset eterogenei, la cui struttura complessa viene analizzata utilizzando stimatori con due penalizzazioni per incoraggiare, da un lato, la sparsità all'interno di ciascun grafo e, dall'altro, le strutture comuni tra i grafi. Tuttavia, in diversi campi applicativi i limiti degli strumenti di rilevazione ne rendono teoricamente ingiustificato l'utilizzo, anche quando l'assunzione relativa alla distribuzione normale multivariata è soddisfatta. In questo articolo proponiamo un'estensione ai dati censurati.*

Key words: Gaussian Graphical Models, High-Dimensional Incomplete Data, Graphical Lasso, Heterogeneous Data

---

Gianluca Sottile
Department of Economics, Business and Statistics, University of Palermo, Italy, e-mail: gianluca.sottile@unipa.it

Luigi Augugliaro
Department of Economics, Business and Statistics, University of Palermo, Italy, e-mail: luigi.augugliaro@unipa.it

Veronica Vinciotti
Department of Mathematics, University of Trento, Italy, e-mail: veronica.vinciotti@unitn.it

1 Introduction

Recently, sparse inference of Gaussian Graphical Models (GGMs) defined in a high-dimensional setting has received intense development. Given the assumption of independent and identically distributed data, a penalized estimator of a GGM is the maximizer of a specific objective function where the log-likelihood function is compensated by a sparsity inducing penalty function that controls the amount of shrinkage on the resulting estimators. The estimator proposed in Yuan and Lin (2007), called graphical lasso (glasso), uses a lasso-type penalty function and has been extensively used in applied research and well studied in the computational as well as theoretical literature (e.g. Friedman et al., 2008). The interested reader is referred to Augugliaro et al. (2016) for an extensive review.

Despite a widespread literature on the glasso estimator, the assumption of independent and identically distributed data has heavily reduced its application to modern datasets, which are often generated by more complex sampling schemes. A typical example of complex data structures is the case of data collected from different sources, such as gene expression data measured on multiple tissues. In this setting, sparse inference is typically carried out by using two specific penalty functions. These are chosen to encourage, on the one hand, sparsity within each graph and, on the other hand, the common structures across the graphs.

In this paper, we extend the estimator proposed in Danaher et al. (2014), called joint glasso (jglasso), to the setting studied in Augugliaro et al. (2020). That is, we consider the case where a part of the data is unobserved due to a known censoring mechanism. Gene expression measured by transcription quantitative polymerase chain reaction (RT-qPCR) is a well-known example of highly dimensional right-censored data.

The remaining part of this paper is structured as follows. Section 2 briefly reviews the inference of GGMs under censoring, while Section 3 proposes an extension of the jglasso estimator to the setting of multiple conditions. Computational aspects are addressed in Section 4 whereas in Section 5 we evaluate the performance of the proposed estimator by a simulation study. Finally, in Section 6 we draw some conclusions.

2 Background on censored Gaussian Graphical Models

Suppose that a p-dimensional random vector, say $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^\top$, is distributed according to a multivariate Gaussian density function denoted by:

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

where $E(\boldsymbol{Z}) = \boldsymbol{\mu}$ and $\Theta = (\theta_{ij})$ is the precision matrix, that is, the inverse of the covariance matrix. Now, let $\boldsymbol{z} = (z_1, \ldots, z_p)^\top$ be a realization of $\boldsymbol{Z}$ and assume that a part of this vector is censored. Thus, $\boldsymbol{z}$ can be split into $\boldsymbol{z}^o$ and $\boldsymbol{z}^c$, i.e., the observed

and unobserved part of $\boldsymbol{z}$, respectively. Denoting by l_j and u_j the lower and upper censoring values of Z_j, respectively, it is easy to show that the probability density function can be written as follows (see Augugliaro et al. (2020) for more details):

$$\widetilde{\phi}(\boldsymbol{z}^o;\boldsymbol{\mu},\Theta) = \int_D^c \phi(\boldsymbol{z}^o,\boldsymbol{z}^c;\boldsymbol{\mu},\Theta)\mathrm{d}\boldsymbol{z}^c, \tag{1}$$

where the integral refers only to the variables whose index belongs to the set $c = \{j : z_j \text{ is censored}\}$, whereas the region of integration D is the Cartesian product of $|c|$ intervals, denoted by D_j, whose definition depends on whether z_j is left or right censored. Formally, $D_j = (-\infty, l_j)$ if $z_j \leq l_j$ otherwise we let $D_j = (u_j, +\infty)$.

Using (1), a censored GGM is the set $\{\widetilde{\phi}(\boldsymbol{z}^o;\boldsymbol{\mu},\Theta),\mathscr{G}\}$, where $\mathscr{G} = \{\mathscr{V},\mathscr{E}\}$ is the undirected graph encoding the conditional independences among the p random variables. More specifically, $\mathscr{V}$ is the set of nodes associated with the random variables, and $\mathscr{E}$ is the subset of the Cartesian product $\mathscr{V} \times \mathscr{V}$, such that $(i,j) \notin \mathscr{E}$ iff Z_i and Z_j are stochastically independent given all the remaining variables. As shown in Lauritzen (1996), the topological structure of the graph $\mathscr{G}$ is related to the entries of the precision matrix Θ, i.e., $(i,j) \in \mathscr{E}$ iff $\theta_{ij} \neq 0$, thus the problem of estimating $\mathscr{E}$ is equivalent to the problem of fitting a sparse precision matrix.

3 Sparse estimation of multiple GGMs with censored data

Suppose that n observations are collected from $K \geq 2$ different but related subpopulations, under a censoring mechanism and assuming a GGM for each subpopulation. To simplify our notation, all quantities related to the kth population are indexed by k. For instance, the kth censored GGM is denoted by $\{\widetilde{\phi}(\boldsymbol{z}_k^o;\boldsymbol{\mu}_k,\Theta_k),\mathscr{G}_k\}$. Under the assumption of independent sampling, the average observed log-likelihood function is

$$\bar{\ell}(\{\boldsymbol{\mu}\},\{\Theta\}) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\log\int_{D_{i,k}}^{c_{i,k}} \phi(\boldsymbol{z}_{i,k}^o,\boldsymbol{z}_{i,k}^c;\boldsymbol{\mu}_k,\Theta_k)\mathrm{d}\boldsymbol{z}_{i,k}^c, \tag{2}$$

where $\{\boldsymbol{\mu}\} = \{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K\}$ and $\{\Theta\} = \{\Theta_1,\ldots,\Theta_K\}$. Maximum likelihood estimators are found by maximizing equation (2). However, these estimators have a very high variance when the sample sizes n_k are larger but close to p. Thus, to overcome the inferential problems related with the high-dimensional setting, we propose to extend the jglasso estimator of Danaher et al. (2014), by replacing the log-likelihood function with the function (2). The resulting estimator, that we call censored jglasso, is formally defined as follows:

$$(\{\widehat{\boldsymbol{\mu}}\},\{\widehat{\Theta}\}) = \arg\max \bar{\ell}(\{\boldsymbol{\mu}\},\{\Theta\}) - \rho\left(\alpha\sum_{k=1}^{K}\|\Theta_k\| + (1-\alpha)P(\{\Theta\})\right), \tag{3}$$

where $\alpha \in [0,1]$ is a tuning parameter that controls the trade-off between two convex penalty functions, where the first one is a lasso-type penalty function that encourages sparsity within each estimated precision matrix, while the second one $P(\{\Theta\})$ is chosen to encourage some form of similarity across the K graphs. Following Danaher et al. (2014), two possible choices are:

$$P_F(\{\Theta\}) = \sum_{k<k'} |\theta_{ij,k} - \theta_{ij,k'}| \quad \text{or} \quad P_G(\{\Theta\}) = \sum_{i \neq j} \left(\sum_{k=1}^{K} \theta_{ij,k} \right)^{1/2}.$$

The censored fused glasso (cfglasso) estimator is defined by solving problem (3) with $P_F(\{\Theta\})$ as the second penalty function whereas, using $P_G(\{\Theta\})$, we get the censored group glasso (cgglasso) estimator. Like the standard glasso, both estimators result in sparse precision matrices. However while cfglasso encourages a stronger form of similarity between the precision matrices, enforcing some entries of $\widehat{\Theta}_1, \ldots, \widehat{\Theta}_K$ to be identical, $P_G(\{\Theta\})$ encourages only a shared pattern of sparsity. Finally, the positive tuning parameter ρ controls the overall amount of shrinkage on the resulting estimators.

4 Computational Aspects

The estimator (3) can be efficiently solved by combining the penalized EM algorithm proposed in Augugliaro et al. (2020) with the alternating directions method of multipliers (ADMM) algorithm developed in Danaher et al. (2014).

The EM algorithm is based on the idea of repeating two steps until a convergence criterion is met. Since the multivariate Gaussian distribution belongs to the exponential family, the first step, called E-Step, simply requires the computation of the conditional expected values of the sufficient statistics. In our model, the E-Step requires the computation of the K imputed vectors of empirical means, denoted by $\bar{\boldsymbol{x}}_k$, along with the empirical covariance matrices $\boldsymbol{S}_k$ (see Augugliaro et al. (2020) for more details). In the second step of the EM algorithm, the M-Step, we first update the current estimates of the K expected values using $\bar{\boldsymbol{x}}_k$, and then the precision matrices are updated by solving the following maximization problem:

$$\max_{\{\Theta\}} \sum_{k=1}^{K} \frac{f_k}{2} \{\log|\Theta_k| - \operatorname{tr}(S_k\Theta_k)\} - \rho \left(\alpha \sum_{k=1}^{K} \|\Theta_k\| + (1-\alpha) P(\{\Theta\}) \right), \quad (4)$$

where $f_k = n_k/n$. The maximization problem (4) is equivalent to the problem studied in Danaher et al. (2014) and can be easily solved using the ADMM algorithm.

5 A Simulation Study

In this section, we compare our proposed estimator with jglasso (Danaher et al., 2014). For the latter, we use the R package `JGL` after imputing the censored data

with their limit of detection. The estimators are evaluated both in terms of recovery of the true graph (area under the precision-recall curve) and estimation of the true precision matrix (mean squared error). We set the right-censoring value to 40, the number of groups K to 3, and the sample size n_k to 100. Each right-censored response vector $\boldsymbol{z}_k$ is simulated according to sparse concentration matrices as reported in Fig. 1. The diagonal entries are fixed to 1 while the non-zero partial correlation coefficients $\theta_{h(h+j),k}$, with $h = 1, 6, 11, \ldots, p-4$ and $j = 1, \ldots, 4$, are sampled from a uniform distribution on the interval $[0.30, 0.50]$. In terms of graph theory what we have is a set of stars, i.e., complete bipartite graphs with only one internal vertex and four leaves. The values of the expected means are chosen in such a way that M

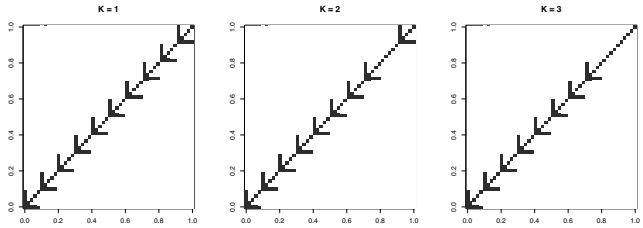

Fig. 1 True graph structure for the $K = 3$ concentration matrices

response variables are right-censored with probability equal to 0.40. The quantities p and M are used to specify the different scenarios used to analyze the behavior of the proposed estimators. In particular, we consider the following cases:

- **Scenario 1:** $p = 50$ and $M = 20$. This setting is used to evaluate the effects of the number of censored variables on the behavior of the estimators when $n > p$.
- **Scenario 2:** $p = 200$ and $M = 80$. 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 coefficient path of cglasso and jglasso, respectively, using the group lasso penalty in both cases. The path is computed using an equally spaced sequence of ρ-values keeping the α parameter equal to 0.50. For each Scenario and along the path of ρ values, the precision-recall curves and the area under these curves (AUC) are computed and then averaged between the K groups. The curves report the relationship between precision and recall for any ρ-value, which are defined as:

$$\text{Precision} = \frac{1}{K}\sum_k \left\{ \frac{\text{TP}}{\text{TP} + \text{FP}} \right\}_k, \quad \text{Recall} = \frac{1}{K}\sum_k \left\{ \frac{\text{TP}}{\text{TP} + \text{FN}} \right\}_k,$$

where TP, FP, and FN are given by the number of correctly selected edges, the number of wrongly selected edges and the number of wrongly selected missing edges, respectively. Table 1 shows how cgglasso gives a better estimate of the concentra-

tion matrices in terms of AUC and a lower mean squared error, for any given value of ρ. We report only five evenly spaced values of ρ.

Table 1 The first 5 columns refer to the average mean squared error of the concentration matrices $\boldsymbol{\Theta}$ for five evenly spaced values of ρ under the specification of the two Scenarios. The last column refers to the mean area under the curves across the sequence of ρ-values. Standard errors are reported in brackets.

	$\rho/\rho_{\max}$					AUC
	0.10	0.25	0.50	0.75	1.00	
Scenario 1						
cgglasso	4.65 (0.54)	8.06 (1.37)	12.10 (3.18)	12.92 (3.82)	12.81 (3.83)	0.97 (0.01)
jglasso	10.19 (1.46)	13.67 (0.94)	16.56 (0.93)	16.84 (1.51)	16.76 (1.57)	0.83 (0.02)
Scenario 2						
cgglasso	17.65 (2.37)	34.85 (4.71)	52.84 (10.72)	55.18 (12.38)	54.65 (12.26)	0.96 (0.01)
jglasso	45.32 (10.09)	62.55 (9.16)	75.53 (4.94)	75.50 (2.65)	75.32 (2.56)	0.81 (0.01)

6 Conclusion

In this paper, we have proposed an extension of the joint glasso estimator to multivariate censored data generated under multiple conditions. A simulation study showed that the proposed estimator performs better than the existing estimators both in terms of parameter estimation and of network recovery.

References

L. Augugliaro, A. M. Mineo, and E. C. Wit. ℓ_1-penalized methods in high-dimensional Gaussian Markov random fields. In M. Dehmer, Y. Shi, and F. Emmert-Streib, editors, *Computational Network Analysis with R: Applications in Biology, Medicine, and Chemistry*, chapter 8, pages 201–267. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2016.

L. Augugliaro, A. Abbruzzo, and V. Vinciotti. ℓ_1-penalized censored gaussian graphical model. *Biostatistics*, 21(2):e1–e16, 2020. doi: 10.1093/biostatistics/kxy043.

P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B*, 76(2):373–397, 2014.

J. H. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

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

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