An extension of the glasso estimator to multivariate censored data

Un'estensione dello stimatore glasso per dati censurati multivariati

Antonino Abbruzzo and Luigi Augugliaro and Angelo M. Mineo

Sommario 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.
Sommario *Lo stimatore glasso è uno degli stimatori più diffusi per fare inferenza sulle reti generiche. Nonostante la sua diffusione, vi sono molti campi della ricerca applicata dove i limiti di misurazione dei moderni strumenti di misurazione rendono teoricamente infondato l'utilizzo di questo stimatore, anche quando l'assunzione sulla distribuzione gaussiana multivariata è soddisfatta. In questo lavoro, proponiamo un'estensione dello stimatore glasso ai dati censurati.*

Key words: Censored data, 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

Antonino Abbruzzo
Department of Economics, Business and Statistics, University of Palermo, Italy, e-mail: antonino.abbruzzo@unipa.it

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

Angelo M. Mineo
Department of Economics, Business and Statistics, University of Palermo, Italy, e-mail: angelo.mineo@unipa.it

nodes represent genes and edges describe the interactions among them. Gaussian graphical models [3] 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 glasso estimator [8] 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 glasso 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. This technique relies on fluorescence-based detection of amplicon DNA and allows the kinetics of PCR amplification to be monitored in real time, making it possible to quantify nucleic acids with extraordinary ease and precision. The analysis of the raw RT-qPCR profiles is based on the cycle-threshold, defined as the fractional cycle number in the log-linear region of PCR amplification in which the reaction reaches fixed amounts of amplicon DNA. If a target is not expressed or the amplification step fails, 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 data. In this paper we propose an extension of the glasso estimator that takes into account the censoring mechanism of the data explicitly.

2 The censored glasso estimator

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 $\mathscr{G} = \{\mathscr{V}, \mathscr{E}\}$, where $\mathscr{V}$ is the set of nodes associated to $\boldsymbol{X}$ and $\mathscr{E} \subseteq \mathscr{V} \times \mathscr{V}$ is the set of ordered pairs, called edges, representing the conditional dependencies among the p random variables [3]. 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 follows

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

As shown in [3], the off-diagonal elements of the concentration matrix are the parametric tools relating the pairwise Markov property to the factorization of the density

(1). Formally, two random variables, say X_h and X_k, are conditionally independent given all the remaining variables if and only if θ_{hk} is equal to zero. This result provides a simple way to relate the topological structure of the graph $\mathscr{G}$ to the pairwise Markov property, i.e., the undirected edge (h,k) is an element of the edge set $\mathscr{E}$ if and only if $\theta_{hk} \neq 0$, consequently the graph specifying the factorization of the density (1) is also called concentration graph.

Let $\boldsymbol{X}$ be 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, 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 nonignorable mechanism [4]. 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 the set $\{-1,0,1\}^p$ and probability function $\Pr\{R(\boldsymbol{X};\boldsymbol{l},\boldsymbol{u}) = \boldsymbol{r}\} = \int_{D_{\boldsymbol{r}}} \phi(\boldsymbol{x};\boldsymbol{\mu},\Theta)d\boldsymbol{x}$, 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 $\mathscr{I} = \{1,\ldots,p\}$ into the sets $o = \{h \in \mathscr{I} : r_h = 0\}, c^- = \{h \in \mathscr{I} : r_h = -1\}$ and $c^+ = \{h \in \mathscr{I} : r_h = +1\}$ 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 $(\boldsymbol{x}_o^\top,\boldsymbol{r}^\top)^\top$. 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 (1) and the conditional distribution of $R(\boldsymbol{X};\boldsymbol{l},\boldsymbol{u})$ given $\boldsymbol{X} = \boldsymbol{x}$. Formally

$$\varphi(\boldsymbol{x}_o,\boldsymbol{r};\boldsymbol{\mu},\Theta) = \int \phi(\boldsymbol{x}_o,\boldsymbol{x}_{c^-},\boldsymbol{x}_{c^+};\boldsymbol{\mu},\Theta)\Pr\{R(\boldsymbol{X};\boldsymbol{l},\boldsymbol{u}) = \boldsymbol{r} \mid \boldsymbol{X} = \boldsymbol{x}\}d\boldsymbol{x}_{c^-}d\boldsymbol{x}_{c^+}. \quad (2)$$

Density (2) can be simplified by observing that $\Pr\{R(\boldsymbol{X};\boldsymbol{l},\boldsymbol{u}) = \boldsymbol{r} \mid \boldsymbol{X} = \boldsymbol{x}\}$ is equal to one if the censoring pattern encoded in $\boldsymbol{r}$ is equal to the pattern observed in $\boldsymbol{x}$, otherwise it is equal to zero, i.e.,

$$\Pr\{R(\boldsymbol{X};\boldsymbol{l},\boldsymbol{u}) = \boldsymbol{r} \mid \boldsymbol{X} = \boldsymbol{x}\} = I(\boldsymbol{x}_{c^-} < \boldsymbol{l}_{c^-})I(\boldsymbol{l}_o \leq \boldsymbol{x}_o \leq \boldsymbol{u}_o)I(\boldsymbol{u}_{c^+} < \boldsymbol{x}_{c^+}),$$

where the inequalities in the previous expressions are intended elementwise. From this, $\varphi(\boldsymbol{x}_o,\boldsymbol{r};\boldsymbol{\mu},\Theta)$ can be rewritten as

$$\varphi(\boldsymbol{x}_o,\boldsymbol{r};\boldsymbol{\mu},\Theta) = \int_{\boldsymbol{u}_{c^+}}^{+\infty}\int_{-\infty}^{\boldsymbol{l}_{c^-}} \phi(\boldsymbol{x}_o,\boldsymbol{x}_{c^-},x_{c^+};\boldsymbol{\mu},\Theta)d\boldsymbol{x}_{c^-}d\boldsymbol{x}_{c^+}I(\boldsymbol{l}_o \leq \boldsymbol{x}_o \leq \boldsymbol{u}_o)$$

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

where $D_c = (-\infty, \boldsymbol{l}_{c^-}) \times (\boldsymbol{u}_{c^+}, +\infty)$ and $c = c^- \cup c^+$. Suppose we have a sample of size n; in order to simplify our notation the set of indices of the variables observed in the ith observation is denoted by $o_i = \{h \in \mathscr{I} : r_{ih} = 0\}$, while $c_i^- = \{h \in \mathscr{I} : r_{ih} = -1\}$ and $c_i^+ = \{h \in \mathscr{I} : r_{ih} = +1\}$ 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 $R(\boldsymbol{X}_i; \boldsymbol{l}, \boldsymbol{u})$, the ith observed data is the vector $(\boldsymbol{x}_{io_i}^\top, \boldsymbol{r}_i^\top)^\top$. Using the density function (3), the observed log-likelihood function can be written as

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

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 for three main reasons. Firstly, the number of measured variables is often 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 [5, 7]. Thirdly, empirical evidence suggests that gene networks or more general biochemical networks are not fully connected [2]. 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 (4).

All that considered, we propose to estimate the parameters of the Gaussian graphical model by generalizing the approach proposed in [8], i.e., by maximizing a new objective function defined by adding a lasso-type penalty function to the observed log-likelihood (4). The resulting estimator, called censored glasso (cglasso), is formally defined as

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

Like in the standard glasso estimator, the non-negative tuning parameter ρ is used to control the amount of sparsity in the estimated concentration matrix $\widehat{\Theta}^\rho = (\hat{\theta}_{hk}^\rho)$ and, consequently, in the corresponding estimated concentration graph $\widehat{\mathscr{G}}^\rho = \{\mathscr{V}, \widehat{\mathscr{E}}^\rho\}$, where $\widehat{\mathscr{E}}^\rho = \{(h,k) : \hat{\theta}_{hk}^\rho \neq 0\}$. When ρ is large enough, some $\hat{\theta}_{hk}^\rho$ are shrunken to zero resulting in the removal of the corresponding link in $\widehat{\mathscr{G}}^\rho$; 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 concentration graph.

Tabella 1 Results of the simulation study: for each measure used to evaluate the behaviour of the considered methods we report average values and standard deviation between parentheses

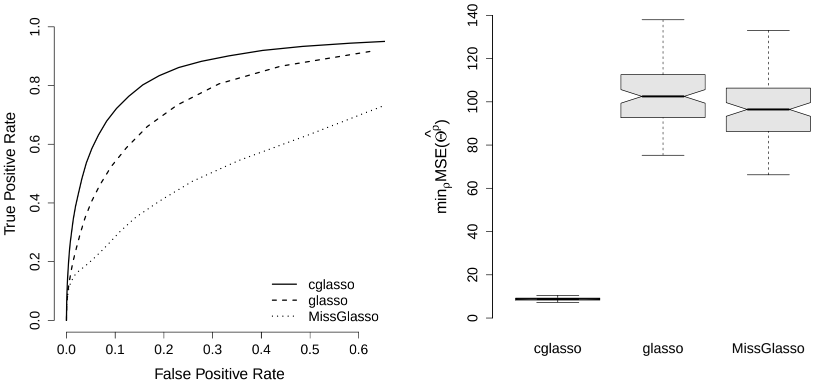
Model	$\min_\rho \text{MSE}(\hat{\boldsymbol{\mu}}^\rho)$		$\min_\rho \text{MSE}(\widehat{\Theta}^\rho)$			AUC		
H/p	cglasso	MissGlasso	cglasso	glasso	MissGlasso	cglasso	glasso	MissGlasso
0.5	0.47	14.50	8.76	103.35	96.75	0.60	0.46	0.37
	(0.11)	(0.69)	(0.64)	(14.43)	(16.01)	(0.02)	(0.02)	(0.02)
0.7	0.48	21.00	10.11	139.76	131.99	0.58	0.39	0.25
	(0.10)	(0.76)	(0.84)	(15.94)	(18.81)	(0.02)	(0.02)	(0.02)

3 Simulation study

By a simulation study, we compare our proposed estimator with MissGlasso [6], which performs ℓ_1-penalised estimation under the assumption that the censored data are missing at random, and with the glasso estimator [1], 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 concentration graph and the mean squared error.

Our study is based on a multivariate Gaussian distribution with $p = 50$ and sparse concentration matrix simulated by a random structure, i.e., the probability of observing a link between two nodes is 0.05. To simulate a censored sample we use the following procedure: we set the mean $\boldsymbol{\mu}$ in such a way that $\mu_h = 40$ for the H censored variables, i.e. $\Pr\{R(X_h; -\infty, 40) = +1\} = 0.50$, while for the remaining variables μ_h is sampled from a uniform distribution on the interval $[10; 35]$. The quantity H is used to evaluate the effects of the number of censored variables on the behaviour of the considered estimators. In particular, we consider $H \in \{25, 35\}$. At this point, we simulate a sample from the latent p-variate Gaussian distribution and treat all values greater than 40 as censored. We use the previous procedure to simulate 100 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 30 ρ-values. Table 1 reports the summary statistics $\min_\rho \text{MSE}(\hat{\boldsymbol{\mu}}^\rho)$, $\min_\rho \text{MSE}(\widehat{\Theta}^\rho)$ and the Area Under the Curve (AUC) for network discovery.

The results on the AUC suggest that cglasso can be used as an efficient tool for recovering the structure of the true concentration matrix. 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 $\boldsymbol{\mu}$. Figure 1 shows a graphical representation of the results obtained with $H = 25$.



(a) (b)

Figura 1 Results of the simulation study with $H = 25$. Panel (a) shows the ROC curves; Panel (b) shows the box-plots of the behaviour of quantity $\min_\rho \mathrm{MSE}(\widehat{\Theta}^\rho)$ for the considered estimators.

4 Conclusions

In this paper, we have proposed an extension of the glasso estimator to multivariate censored data. An extensive simulation study showed that the proposed estimator overcomes the existing estimators both in terms of parameter estimation and of network recovery.

Riferimenti bibliografici

1. Friedman, J. H., Hastie T., Tibshiran, T.: Sparce inverse covariance estimation with the graphical lasso. **9**(3), 432–441 (2008)
2. Gardner T. S., di Bernardo D., Lorenz D., Collins J. J.: Inferring genetic networks and identifying compound mode of action via expression profiling. Science. **301**, 102–105 (2003)
3. Lauritzen, S. L.: Graphical Models. Oxford University Press, Oxford (1996)
4. Little, R. J. A., Rubin, D. B.: Statistical Analysis with Missing Data. John Wiley & Sons, Inc., Hoboken (2002)
5. Schäfer, J., Strimmer, K.: A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology. **4**(1). (2005)
6. Städler, N., Bühlmann, P.: Missing values: sparse inverse covariance estimation and an extension to sparse regression. Stat. Comput. **22**(1), 219–235 (2012)
7. Uhler C.: Geometry of maximum likelihood estimation in Gaussian graphical models. Ann. Statist. **40**(12), 238–261 (2012)
8. Yuan M., Lin, Y.: Model selection and estimation in the Gaussian graphical model. Biometrika. **94**(1), 19–35 (2007)