Learning Bayesian Networks for Nonparanormal Data

Apprendimento di reti bayesiane per dati non parametrici

Abstract In the literature, structural learning procedures for selecting the directed acyclic graph of a Bayesian network are increasingly explored and specified according to the analyzed data typology. With respect to data drawn from a Gaussian Copula model, the Rank PC algorithm, based on Spearman rank correlation, has been introduced. Moreover, we recently proposed a modified version of the well known Grow-Shrink algorithm, the Copula Grow-Shrink one, based on the Spearman rank correlation and the Copula assumption. Here, we show a simulation study to verify the robustness of our Copula Grow-Shrink algorithm and we discuss the performance results in comparison with the baseline and the Rank PC algorithm.
Abstract *In letteratura, le procedure di apprendimento strutturale per la stima di un grafico diretto aciclico di una rete bayesiana sono sempre più esplorate e dettagliate in base alla tipologia di dati analizzati. Per quanto riguarda i dati derivanti da un modello di copula gaussiana l'algoritmo Rank PC, basato sulla correlazione di Spearman, è stato proposto in letteratura. Inoltre, abbiamo recentemente proposto una versione modificata del noto algoritmo Grow-Shrink, l'algoritmo Copula Grow-Shrink, basato sulla correlazione tra ranghi di Spearman e sull'assunzione di Copula. Qui, mostriamo uno studio di simulazione per verificare la solidità del nostro algoritmo Copula Grow-Shrink e discutiamo i risultati delle prestazioni rispetto agli algoritmi Grow-Shrink e Rank PC.*

Key words: joint normal copula, Copula Grow-Shrink algorithm, simulation study, diagnostic measures

1 Introduction

Statistical multivariate data modeling is increasingly carryed out through Bayesian networks, (BN, [2]) that depict the multivariate probability distribution of a set of

variables by a graphical representation of independencies encoded in a directed acyclic graph (DAG). A DAG is a finite set of nodes, standing for random variables, and directed edges, arranged never producing cycles, that point out direct relevance of one variable to another. In a DAG, a parent node has an outcoming arrow pointing to another node namely child; every node is associated with a conditional distribution given its parents and the joint distribution can be factorized according to the DAG.

In this context, a common issue concerns the DAG structural elicitation. When the dependencies are unknown or partially known, DAG structure has to be estimated directly from data. Most often, researchers wish to maximize the learning power respecting the typology of managed data. For nonparanormal data some structural learning algorithms have been discussed in the literature; we recently proposed the Copula Grow-Shrink [1], a modified version of the Grow-Shrink algorithm, based on the recovery of the Markov blanket of the nodes and on the Spearman correlation. The paper, aiming at evaluating the robustness of the our proposal, is organized as follows: nonparanormal graphical models are briefly recalled in Section 2; the Grow-Shrink and the Copula Grow-Shrink algorithms are discussed in Section 3; the simulation study and preliminary results are addressed in Section 4.

2 Nonparanormal Graphical models and their estimations

Nonparanormal data modeling by graphical models has been studied in the literature. Generally speaking, a nonparanormal graphical model is a semiparametric extentions of a Gaussian graphical model useful when the analysed continuous variables follow a Gaussian graphical model only if trasformed by unknown smooth monotone functions preserving the dependencies structure of the underlying multivariate normal distribution. According to [7]:

> Definition 1. Let $f = (f_v)_{v \in V}$ a collection of strictly increasing functions $f_v : R \to R$ and $\Sigma \in R^{V \times V}$ be a positive definite correlation matrix. The nonparanormal distribution NPN(f, Σ) is the distribution of the random vector $(f_v(Z_v))_{v \in V}$ for $(Z_v)_{v \in V} \sim N(0, \Sigma)$.

> Definition 2. The nonparanormal graphical model NPN(G) associated with a DAG G is the set of all distributions NPN(f, Σ) that are Markov with respect to G.

The function f_v realizes a deterministic transformation on Z_v preserving the same dependence structure of the underlying latent multivariate normal distribution also in the nonparanormal model.
If $X \sim NPN(f, \Sigma)$ and $Z \sim N(0, \Sigma)$, for any triple of pairwise disjoint set $A, B, S \subset V$, then $X_A \perp\!\!\!\perp X_B | X_S \Leftrightarrow Z_A \perp\!\!\!\perp Z_B | Z_S$.
For two nodes (u, v) and a separating set S we have $X_u \perp\!\!\!\perp X_v | X_S \Leftrightarrow \rho_{uv|S} = 0$.

A trigonometric transformation on Spearman rank correlation (r) produces latent Normal correlation coefficients accurate estimators. Reference [5] show that if (X,Y) are bivariate normal with $Corr(X,Y) = \rho$, it yelds:

$$P(|2sin(\frac{\pi}{6}\widehat{r}) - \rho|) > \varepsilon) \leq 2exp(-\frac{2}{9\pi^2}n\varepsilon^2) \tag{1}$$

Since $\widehat{r}$ depends on the observations *via* their ranks that are preserved under strictly increasing functions, (1) still holds for nonparanormal graphical models with Pearson correlation $\rho = \Sigma_{xy}$ in the underlying latent bivariate normal distribution. On the basis of the previous result ρ is estimated as:

$$\hat{\rho} = 2\sin(\frac{\pi}{6} \cdot \hat{r}) \tag{2}$$

The same transformation still holds for the partial correlation coefficients.

3 Bayesian Netoworks Structural Learning

Bayesian netoworks structural learning methods are mainly *scoring and searching* techniques or *constraint-based* algorithms; they estimate and depict the unknown independencies relations among variables by a DAG. The most spread algorithm is the PC algorithm [9] that proceeds along three steps: (i) the skeleton identification by testing marginal and conditional independencies by Pearson correlation test for Gaussian data, (ii) the v-structures identification standing for conditional dependence between two nodes given a third and (iii) the orientation of the remaining links without producing additional v-structures and/or directed cycles. If variables are not Gaussian, a PC algorithm rank version named Rank PC (RPC) algorithm is available [7]. RPC algorithm tests conditional independence between two variables given a separating set by computing the rank-based partial correlation estimates (Eq. 2). The RPC algorithm consistency is proved by [7], under some non-strict assumptions. It is shown that RPC works at the same strength of PC algorithm for normal data but considerably better for non-normal data under the *strong* assumption of joint distribution following a normal copula model. The RPC algorithm could be implemented using the `pcalg R` package [4].

A competitive algorithm to these common choices is the Grow-Shrink algorithm (GS, [6]) based on the intuitive concept of the Markov blanket (MB) of a variable, *i.e.* the set of all parents, children and parents of children of the variable of interest, say X. Moreover, the MB(X) d-separates variable X from any other variable outside its Markow blanket. The GS algorithm focuses on the recovery of the MB(X) based on pairwise independence tests by two phases: the growing phase, where, from MB(X) empty set denoted by S, the procedure adds variables to S as long as they are associated with X given the current contents of S; the shrinking phase identifies and removes variables not really belonging to MB(X) eventually added to S. Our Copula GS algorithm (CGS) [1] has the same logical structure as GS but the

marginal and partial correlations coefficients used in the statistical test for independence are computed trough (2). The GS is implemented in `bnlearn R` package [8] so that our proposal is developed in `R` as well.

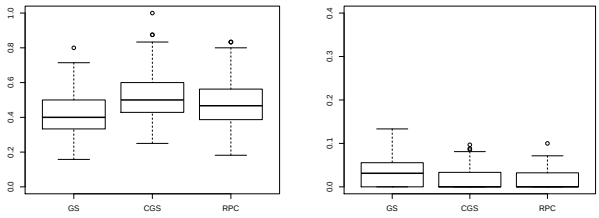
4 Main Results and Conclusions

With the aim to explore the robustenss, a simulation study has been carried out. According to the procedure in [3] and to the simulation plan in [7], we simulated 200 random DAGs with sparsity parameter $s = 0.3$ and we sampled from a Gaussian Copula distribution faithful to them. Considered sample sizes are $n = 50$ and $n = 1000$. On every training set, we performed the structural learning GS, CGS and RPC algorithms with a significance level of 0.05. Algorithm performances have been compared in terms of sensitivity, specificity and precision.
In details, the diagnostic measure *true positive rate* (TPR) is the proportion of edges correctly estimated on the true edges; the closer the value is to 1, the better is the sensitivity. The *false positive rate* (FPR) is the proportion of edges incorrectly found over the number of true gaps; the closer the value is to 0, the better is the specificity. The *true discovery rate* (TDR) is the proportion of edges correctly found on the total number of estimated edges; the closer the value is to 1, the better is the precision. The performance measure distributions from simulations are displayed in the following boxplot (see Figures 1 e 2).

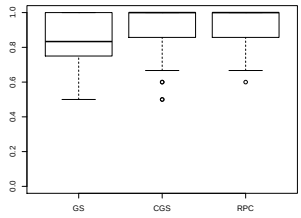
For small sample size $n = 50$ (see Figure 1) the sensitivity of CGS algorithm outperforms the GS and the RPC ones denoting a better capacity to catch the real structure. The specificity of CGS is still better with respect to GS and only slightly more variable in comparison to RPC. Also in terms of TDR, the CGS outperforms the GS and works the same as the RPC. For large sample size $n = 1000$ (see Figure 2) the CGS outperforms the GS and works slightly better than the RPC in terms of specificity and precision but gains a stronger sensitivity.
According to these simulation results the algorithm we propose represents a better choice to estimate a DAG in case of nonparanormal data. Since $1 - TPR$ is equal to the False Negative Rate (FNR), it means that the CGS algorithm prevents from bias in the model due to the absence of a "true" arc. We argue that, as the RPC one, also the CGS algorithm reduces the risk of an overparametrization of model since the FPR is smaller than that of GS for both sample sizes.

References

1. author: (year)
2. Cowell, R.G., Dawid, P., Lauritzen, S.L., Spiegelhalter, D.J.: Probabilistic Networks and Expert Systems. Springer, New York (1999)
3. Kalisch, M., Bühlmann, P.: Estimating high-dimensional directed acyclic graphs with the pc-algorithm. J. Mach. Learn. Res. **8**, 613–636. (2007)

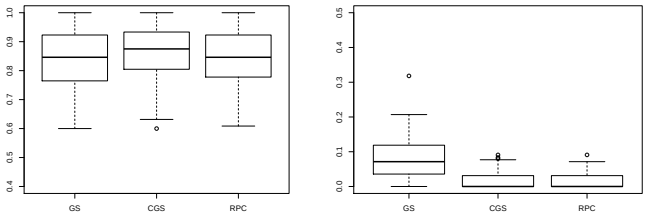


(a) TPR (b) FPR

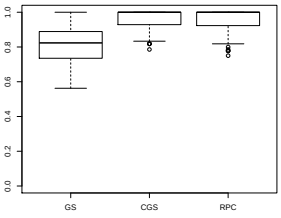


(c) TDR

Fig. 1 Boxplot of diagnostic measures for n=50



(a) TPR (b) FPR



(c) TDR

Fig. 2 Boxplot of diagnostic measures for n=1000

4. Kalisch, M., Mächler, M., Colombo, D., Maathuis, M.H., Bühlmann, P.: Causal inference using graphical models with the R package pcalg. J. Stat. Softw. **47**(11), 1–26 (2012). URL http://www.jstatsoft.org/v47/i11/.
5. Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L.: High-dimensional semiparametric gaussian copula graphical models. The Annals of Statistics **40**(4), 2293–2326 (2012). DOI 10.1214/12-AOS1037. URL http://dx.doi.org/10.1214/12-AOS1037
6. Margaritis, D.: Learning bayesian network model structure from data. Ph.D. thesis, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA (2003). Technical Report CMU-CS-03-153
7. Naftali, H., Drton, M.: Pc algorithm for nonparanormal graphical models. J. Mach. Learn. Res. **14**, 3365–3383. (2013)
8. Scutari, M.: Learning bayesian networks with the bnlearn r package. J. Stat. Softw. **35**(3), 1–22 (2010). URL http://www.jstatsoft.org/v35/i03/
9. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search, 2nd edn. MIT press, Cambridge, Massachusetts (2000)