Sensitivity of graphical modeling against contamination

Sonja Kuhnt and Claudia Becker

Department of Statistics, University of Dortmund

44221 Dortmund, Germany

Abstract

Graphical modeling as a form of multivariate analysis has turned out to be a capable tool for the detection and modeling of complex dependency structures. Statistical models are related to graphs, in which variables are represented by points and associations between each two of them as lines. The usefulness of graphical modeling depends of course on finding a graphical model, which fits the data appropriately. We will investigate how existing model building strategies and estimation methods can be affected by model disturbances or outlying observations. The focus of our sensitivity analysis lies on mixed graphical models, where both discrete and continuous variables are considered.

1 Introduction

Graphical models turned out to be a helpful tool for detecting and modeling dependency structures (e.g. Cox and Wermuth, 1996, Edwards, 2000, Lauritzen, 1996, Whittaker, 1990). Up to now not much work has been done with respect to considering model disturbances or the effect of outlying observations on the estimation in such models. Since usually estimation is performed by the maximum likelihood method, it can be expected that – similar to other model situations – also in graphical models outliers or contaminated data will disturb the estimation. Hence, with the growing acceptance of using graphical models for analyzing dependency structures there will also be a growing need for sensitivity analyses of the existing estimation methods and for the construction of robust estimates for these models.

We consider here first approaches to sensitivity analyses and robustness in graphical models, where we focus on the case of mixed random vectors containing both, discrete and continuous elements. There exists already work on robustness and the effect of outliers for either purely discrete or purely continuous cases – more with respect to the continuous case, less with respect to the discrete case (see e.g. Barnett and Lewis, 1994, for an overview). We try to put together ideas from both branches for the usage in mixed graphical models, concentrating on graphical interaction models, where an undirected graph shows the association structure between the variables.

The paper is organized as follows. In Section 2 we briefly recollect the main ideas of graphical modeling and introduce in more detail the distributional assumptions. Section 3 provides an example data set illustrating the effect of a certain disturbance in the data on the model building process. To expand the findings of the example, we show the results of a simulation study in Section 4, where the effect of contamination on the model building process is investigated in more detail. We conclude with some remarks on the definition of outliers and possible robustifications of the modeling in graphical mixed models.

2 Graphical Models: Distributional Assumptions

The notion of graphical independence models visualizes conditional independences inherent in a statistical model by a graph. A graph in the mathematical sense is a pair $G = (V, E)$, where $V = \{1, ..., n\}$ is a finite set of vertices and the set of edges E is a subset of the set $V \times V$ of ordered tuples of distinct vertices. In an undirected graph it follows from $(a, b) \in E$ that also $(b, a) \in E$.

Given a random vector $X = (X_1, ..., X_n)'$ and an undirected graph $G = (V, E)$ with $V = \{1, \ldots, n\}$ the notion of a graphical independence model is defined by the class of all distributions of X, for which $X_a \perp X_b \mid \mathbf{X}_{V \setminus \{a,b\}}$ holds iff (a, b) is not element of the set E. Here, $X_a \perp X_b \mid \mathbf{X}_{V \setminus \{a,b\}}$ denotes the conditional independence of X_a and X_b given all other variables.

We consider a set $X = (X_1, ..., X_n)'$ of random variables, where the first p variables are discrete and the following q continuous, $n = p + q$. Denote the vector of discrete variables by X_Δ and the vector of of continuous variables by X_Γ. If a purely discrete vector X_Δ is considered, graphical models provide a new way to demonstrate well-established log linear models (Edwards and Kreiner, 1983, Whittaker, 1990). In the purely continuous case graphical models based on the assumption of a normal distribution are characterized by restrictions on the covariance matrix (Edwards, 2000, Chap.3). An extension of the distributional assumptions of the pure cases to the mixed case has been provided by Lauritzen and Wermuth (1989) with the notion of a conditional gaussian (CG-) distribution, where the continuous variables given the discrete variables are normally distributed. Let a typical observation of $X = (X'_\Delta, X'_\Gamma)'$ be written as $(i', y')'$, where i is a p-tuple containing the values of the discrete variables and y is a real-valued vector of length q. Let further $\mathcal{I}$ denote the set of all possible outcomes of X_Δ. The density of a CG-distribution is then defined by $f_{X_\Delta,X_\Gamma} = f_{X_\Delta}(i)\, f_{X_\Gamma|X_\Delta}(i|y)$, yielding

$$f_{X_\Delta,X_\Gamma}(i,y) = p_i\ (2\pi)^{-\frac{q}{2}}\ \det(\Sigma_i)^{-\frac{1}{2}}\ \exp\{-\frac{1}{2}(y-\mu_i)'\Sigma_i^{-1}(y-\mu_i)\}, \qquad (1)$$

where p_i denotes the probability of the occurrence of i, and μ_i, Σ_i are the conditional mean and covariance of Y given i. The distribution is called homogeneous, if for some Σ it holds that $\Sigma_i = \Sigma \quad \forall\, i \in \mathcal{I}$. The set $\{p_i, \mu_i, \Sigma_i\}_{i\in\mathcal{I}}$ is called the set of moment parameters, their structure determines the independence properties between the random variables. Applications of graphical independence models usually aim at finding a simple model in the sense of a sparse dependency structure, which is still consistent with the data. Various strategies have been proposed for the selection of an appropriate model. These are of course to be seen with the appropriate caution and should always be accompanied by background knowledge. Reviews can be found in Edwards (2000, Chap. 6) and Blauth (2002, Chap. 2). They encompass backward and forward selection strategies as well as alternative search algorithms. As a typical example we look at a backward-selection procedure. The procedure starts with a saturated model, where no conditional independence holds, hence every pair of vertices is joined by an edge in the corresponding independence graph. Then, step by step, individual edges are removed. At each step a criterion is calculated for every model resulting from the removal of a further edge from the present graph. Such a criterion can e.g. be Akaike's Information Criterion (AIC), $-2\ln\widehat{\ell}_M + 2r$, where $\widehat{\ell}_M$ is the maximum likelihood unter the model M and r is the dimension (number of free parameters) of the model. Also the χ^2 test based on the deviance difference

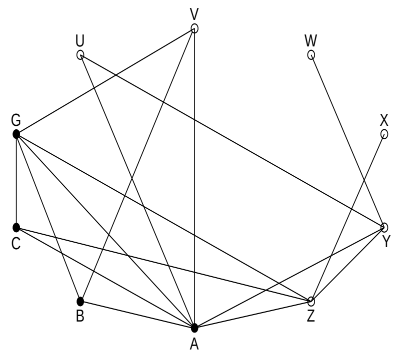

Figure 1: Independence graph resulting from stepwise model selection

$-2(\ln \widehat{\ell}_{M_0} - \ln \widehat{\ell}_{M_1})$ between two models M_0 and M_1, $M_0 \subseteq M_1$, is frequently used. The edge corresponding to the largest value of the AIC-criterion or the largest p-value is deleted and the procedure continues until there is no further improvement in the AIC-criterion or the p-value stays below a given level α. Often, the model search is restricted to models with a decomposable graph, such that maximum likelihood estimates can be explicitly calculated. The aim of this paper is to explore the sensitivity of such a procedure to contaminated data.

3 Data Example: Breast Cancer

As an example we consider a data set dealing with ablative surgery for breast cancer. The main variable (G) classifies each of 186 patients by the treatment success as either successful / intermediate (G=1) or failure (G=2). The data set further contains six continuous variables (U-Z) and three binary variables (A-C) describing various characteristics of the patients. This data set has originally been described by Krzanowski (1975) in the context of discriminant analysis. The data set provides an illustrative example of the mixed case and has already been analyzed using the graphical model approach (see e.g. Edwards, 2000, p. 119 ff.) Starting with the homogeneous saturated model and applying the backwards model selection strategy described in Section 2 based on the χ^2 test, results in the independence model described by the graph in Figure 1. Note, that the principle of coherence has been followed, meaning that an edge with a p-value below the chosen level $\alpha = 0.05$ at any step will not be removed in a later step. Also, only models with decomposable graphs have been considered. The model search has been conducted using the computer program MIM (Edwards, 2000).

We repeat the same procedure after changing the value of variable U for the first observation in the data set. The values of variable U vary between 23 and 69 with a median of 47.14 and a variance of 78.5. The value 35 of the first observation is replaced by the maximum value 69 for U in the data set, hence by an observation not obviously presenting a contaminated value. Still, it suffices to change the result of the model search selection procedure, compare Figure 2. Concentrating on the main variable (G) we see in Figure 1, that the corresponding vertex is connected with the vertices for the variables A,B,C and Z. In Figure 2, however, variable V is also connected with G. Changing a single value in only one observation hence already changed the identified model, yielding an additional edge in the resulting graph.

4 Simulation study for a mixed homogeneous model

The example of the foregoing section gives rise to the question whether this is a singular effect or may happen rather often. In general, we have to in-

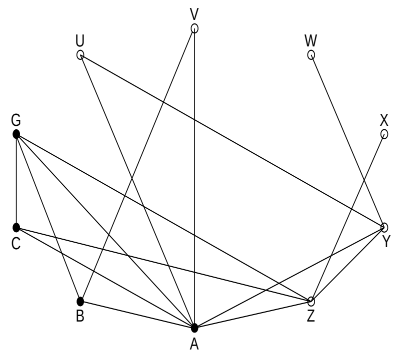

Figure 2: Independence graph resulting from stepwise model selection on the data with a single changed observation

vestigate what happens to the model building process if some observations do not fit into the pattern built by the majority of the data. To expand the findings of the example, we perform a simulation study. We generate 100 data sets with 1000 observations each according to a graphical mixed model with dependency structure according to Figure 3. The four-dimensional random vector X with observation $x = (i', y')' = (i_1, i_2, y_1, y_2)'$ follows a CG-distribution with density according to (1). The moment parameters for the simulation are given in Table 1. The data generated in this way will also be called the "true" data. Next, we disturb the true data in several ways, creating five data situations:

(A) true data

(B) y_1 replaced by -30 in 10 randomly chosen observations

(C) y_1 replaced by 1000 in 10 randomly chosen observations

(D) y_1 replaced by -30 if $i = (1, 1)$

p_{i_1,i_2}	$i_2 = 1$	2	3
$i_1 = 1$	0.3712	0.0087	0.0219
2	0.3799	0.1101	0.0032
3	0.0912	0.0007	0.0131

μ_{i_1,i_2}	$i_2 = 1$	2	3
$i_1 = 1$	$\binom{-15.38}{27.77}$	$\binom{11.41}{19.48}$	$\binom{-18.45}{34.19}$
2	$\binom{-11.07}{18.76}$	$\binom{-7.11}{10.47}$	$\binom{-14.14}{25.18}$
3	$\binom{-17.35}{31.91}$	$\binom{-13.39}{23.62}$	$\binom{-20.42}{38.33}$

$$\Sigma = \begin{pmatrix} 10.78 & -14.22 \\ -14.22 & 29.74 \end{pmatrix}$$

Table 1: Moment parameters of the simulated model

(E) i replaced by $(3, 2)$ in 10 randomly chosen observations.

To each of the simulated data sets we apply the same backwards selection strategy as for the breast cancer data, but without the restriction to decomposable models.

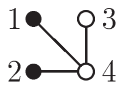

Figure 3: Independence graph of the simulated model

The results of the model selection procedure are reported in Table 2. Several conclusions can be drawn from these results. First, we see that certainly contamination of the data has an influence on the results of the model selection. The observed effects of the data example in Section 3 seem to be no singular event. Second, obviously the degree of contamination is important. Comparing the results for situations (B) and (C), we find that the moderate contamination scheme (B), where we change some observed values

	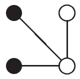	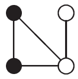	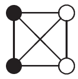	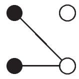	Others
(A)	95	1	0	0	4
(B)	96	1	1	0	2
(C)	6	0	1	82	1
(D)	0	0	100	0	0
(E)	0	98	1	0	1

Table 2: Simulation results: number of samples out of 100 resulting in respective graphs in situations (A) to (E)

to a value which lies at the border of what we may expect under the model, does not change the outcome of the model selection. On the other hand, the very extreme contamination of situation (C) drastically changes the results. We expect that there will be some degree of extremeness in the contamination where the behaviour of the model selection procedure changes. It will be one aspect of further research to determine this “change point” more exactly. Third, the type of contamination matters. Different disturbances have different effects in the sense that different graphs are found to represent the dependency structure. There is no general direction of change, we find graphs with additional edges as well as graphs with less edges than for our true model. Again, further research is needed to investigate these effects in more detail.

5 Further Directions: Outliers and Robust Model Selection

The results of the simulation study show how severely "outliers" in the data can affect the model selection process in graphical modeling. These findings yield two interesting directions of further research: first, the concept of outlyingness has to be investigated conceptionally in the context of graphical modeling, especially with respect to mixed models. Second, a robust alternative for model selection should be found which is less sensitive to contamination in the data.

Concerning the first point, we give a few comments. It is not immediately obvious how to characterize an observation as outlying with respect to a mixed graphical model. If we restrict to the marginal (resp. conditional) distributions, it is of course possible to define outliers with respect to the normal distribution (continuous variables) as well as with respect to the contingency table model (discrete variables), cf. Barnett and Lewis (1994), Becker and Gather (1997, 1999), Gather, Kuhnt and Pawlitschko (2002), Kuhnt (2002). A special aspect of the combination of both is that it is possible to have a single outlier in the continuous variables, whereas in the contingency table we find a whole cell and hence a set of observations to be outlying. From this follows a certain asymmetry. When looking at both types of variables together, it is necessary to define an outlier with respect to the CG-distribution. Several questions arise. Could it happen that we have observations contributing with their discrete part to the same cell of the contingency table, where some of these observations are outliers with respect to the distribution of the continuous variables, but some are not? On the other hand, does it make sense to declare all observations contributing with their discrete part to an outlying cell as outliers, if their continuous parts fit in the structure of the continuous variables quite well? Although we have some sort of intuitive knowledge of what is an outlying observation, a formalization for CG-distributions is a challenge to be followed.

Concerning "robust" ways of model selection, first approaches for general parametric models are given by Ronchetti (1997), especially on robust versions of the Akaike Information Criterion. When fitting graphical models as described above, maximum likelihood estimation is used. It is well known that maximum likelihood estimators in general are sensitive against contamination in the data. Hence, a natural way to robustify the model selection process would be to replace the maximum likelihood estimators by more robust alternatives. This approach may also be applicable in graphical modeling. Look at the case of investigating continuous variables only. The analysis of the dependency structure is based mainly on the concentration matrix, i.e. the inverse of the covariance matrix. This has to be estimated appropriately. In the saturated model, this does not pose any problem, but when deleting edges, the concentration matrix estimation has to be performed under restrictions (certain entries in the matrix have to be equal to zero). When calculating maximum likelihood estimators, the so-called modified iterative proportional scaling is used (Frydenberg and Edwards, 1989), where, starting from an initial estimate, the concentration matrix is adjusted iteratively to reflect the dependency structure of the data on the one hand and to fulfill the restrictions on the other hand. To our knowledge there does not exist a robustified version of the iterative proportional scaling algorithm up to now. Since in this algorithm empirical covariance matrices for certain choices of subsets of the variables are calculated, a robustification seems possible by using robust covariance estimators like e.g. the MCD covariance estimator (Rousseeuw, 1985). Of course, this would be a solution for the continuous variables only. There exists a respective version of iterative proportional scaling for mixed models which has to be modified also with respect to estimating the discrete probabilities and the dependency between discrete and continuous variables robustly. Here, the proposal of a modification is less obvious and still under investigation.

Acknowledgements

The financial support of the Deutsche Forschungsgemeinschaft (SFB 475) is gratefully acknowledged.

References

BARNETT, V. and LEWIS, T. (1994): *Outliers in Statistical Data.* 3rd ed. Wiley, Chichester.

BECKER, C. and GATHER, U. (1997): Outlier Identification and Robust Methods. In: G.S. Maddala and C.R. Rao (Eds.): *Handbook of Statistics 15: Robust Inference.* Elsevier, Amsterdam, 123–143.

BECKER, C. and GATHER, U. (1999): The Masking Breakdown Point of Multivariate Outlier Identification Rules. *Journal of the American Statistical Association, 94, 947–955.*

BLAUTH, A. (2002): *Model Selection in Graphical Models With Special Focus on Genetic Algorithms.* Logos Verlag, Berlin.

COX, D.R. and WERMUTH, N. (1996): *Multivariate Dependencies.* Chapman & Hall, London.

EDWARDS, D. (2000): *Introduction to Graphical Modelling.* 2nd ed. Springer, New York.

EDWARDS, D. and KREINER, S. (1983): The Analysis of Contingency Tables by Graphical Models. *Biometrika, 70, 553-565.*

FRYDENBERG, M. and EDWARDS, D. (1989): A Modified Iterative Proportional Scaling Algorithm for Estimation in Regular Exponential Families. *Comp. Statist. Data Analysis, 8, 143-153.*

GATHER, U., KUHNT, S. and PAWLITSCHKO. J. (2002): Outlier Regions for Various Data Structures. To appear as invited Chapter to the Volume *"Emerging Areas in Probability, Statistics and Operations Research", Mathematical Sciences Series.*

KRZANOWSKI, W.J. (1975): Discrimination and Classification Using Both Binary and Continuous Variables. *Journal of the American Statistical Association, 70(352), 782–790.*

KUHNT, S. (2002): Outlier Identification Procedures for Contingency Tables using Maximum Likelihood and L_1 Estimates. Submitted.

LAURITZEN, S.L. (1996): *Graphical Models.* Clarendon Press, Oxford.

LAURITZEN, S.L. and WERMUTH, N. (1989): Graphical Models for Associations Between Variables, Some of Which are Qualitative and Some Quantitative. *Annals of Statistics, 17, 31–54.*

RONCHETTI, E.M. (1997): Robustness Aspects of Model Choice. *Statistica Sinica 1, 327-338.*

ROUSSEEUW, P.J. (1985): Multivariate Estimation With High Breakdown Point. In: W. Grossmann, G. Pflug, I. Vincze and W. Wertz (Eds.): *Mathematical Statistics and Applications*. Reidel, Dordrecht, 283–297.

WHITTAKER, J. (1990): *Graphical Models in Applied Mathematical Multivariate Statistics*. Wiley, Chichester.