Dipartimento di Statistica
"Giuseppe Parenti"

WORKING PAPER 2006/06

On the impact of contaminations in graphical Gaussian models

Anna Gottard, Simona Pacillo

Università degli Studi
di Firenze

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

On the impact of contaminations in graphical Gaussian models

Anna Gottard and Simona Pacillo
Department of Statistics “G. Parenti”
University of Florence
gottard@ds.unifi.it, pacillo@ds.unifi.it

Abstract: This work analyzes the impact of some kind of contaminants and wrong model assumptions on concentration graph models. The impact is measured in terms of model selection as the correct identification of the conditional independence structure of a vector of gaussian variables. Four different kinds of source of contamination were investigated, in order to consider both the case of occurrence of gross errors and model deviation. It is of interest to assess against which kind of contaminants graphical models have a more robust behavior. The analysis is based on simulated data. The simulation study shows that relatively few contaminated observations in even just one of the variables can have a significant impact on correct model selection, especially when the contaminated variable is a node in a separating set of the graph.

Keywords: Concentration graph models, Contaminants, Graphical models selection, Model deviation, Multivariate Normal distribution, Robustness.

1 Introduction

Graphical models are a key technique for the analysis of the conditional independence structure of a multivariate distribution (see, for example, Cox and Wermuth, 1996). The literature on graphical models has grown considerably, but not much work has been done to check the consequences of outliers, in the sense of occurrence of gross-errors, contaminants and model deviations. Ideally, inferential analysis should be based on correct empirical data and optimal inference procedures. In reality, empirical data are subject to outliers due to gross errors, and model assumptions can be partially or completely wrong. The presence of outliers is a cause of uncertainty, being a signal of a certain weakness in the data quality and/or in the model specification. Different inferential procedures are characterized by different robustness levels against wrong data or model deviations. This paper intends to analyze the consequences of contaminants for the model selection of Gaussian graphical models by means of simulated data.

On this topic, Kuhnt and Becker (2003) conducted a sensitivity analysis against contamination of undirected graphs in the case of conditionally Gaussian distributed variables, limiting to presence of gross errors in the data. In this work, we extend the analysis to different types of contaminants, limited to the multivariate normal case. Moreover, we consider both decomposable and non-decomposable models, applying the SINFUL model selection procedure (Drton and Perlman, 2004), which controls for multiple tests' overall significance level.

In this article, Section 2 describes the class of graphical models analyzed here and the adopted model selection procedure. Section 3 lists the different sources of distortion we considered, due to contaminants or model deviations. Such sources of distortion are evaluated in Section 4 with a simulation study. Section 5 contains concluding remarks.

2 Concentration graph models and model selection

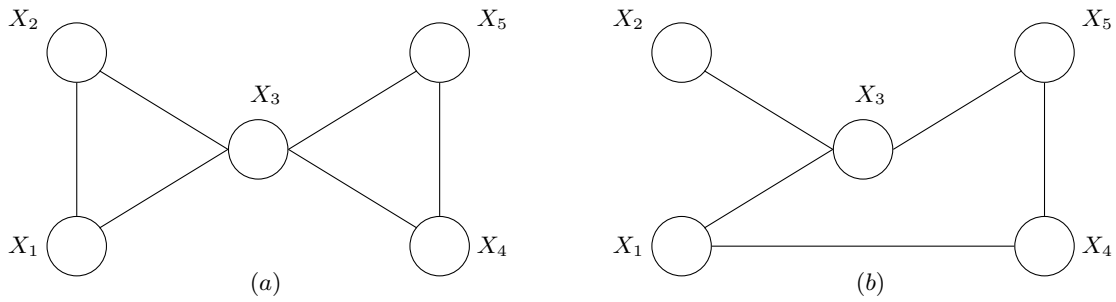
Graphical models are a family of probability distributions whose conditional independence structure is encoded by a graph. A graph $\mathcal{G}$ consists of a pair of sets (V, E), where V is the set of nodes and $E \subset V \times V$ is a set of edges. Each node is represented by a circle or a dot; two nodes v_1 and v_2 are connected by an undirected edge in the graph whenever both the pairs (v_1, v_2) and $(v_2, v_1) \in E$. We will denote an undirected edge as an unordered pair $\{v_1, v_2\} \in E$. If only one of the two pairs are in E, then the two nodes are connected by an arrow; if neither of them are in E, then they are not connected. An undirected graph is a graph allowing only for undirected edges.

In graphical models, each random variable is associated with a node. The independence relation $\perp\!\!\!\perp$ (Dawid, 1979) among subsets of random variables is encoded in the graph as the absence of edges, so that a missing edge between two vertices v_1 and v_2 indicates that the corresponding variables X_{v_1} and X_{v_2} are (conditionally) independent. A set of properties, called Markov properties, establishes the conditional set of variables of each independence constraint.

A *concentration graph model* (Cox and Wermuth, 1993), as considered here, has a vector of random variables $\mathcal{X}_V = (X_1, X_2, \ldots, X_p)$ having a multivariate Gaussian joint distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, whose conditional independence structure can be represented by an undirected graph. A concentration graph model has a missing edge between two nodes, say v_i and v_j, whenever the corresponding element in the concentration matrix $\boldsymbol{\Sigma}^{-1}$ is zero, $\sigma^{ij} = 0$. For this reason, these models are also called *covariance selection models* (Dempster, 1972).

Figures $1(a)$ and $1(b)$ show two graphs used later in the paper to generate data. To illustrate undirected graphs, here, we discuss only Figure 1(a), called a *butterfly graph*. Here, we consider

Figure 1 *Undirected graphs considered as data generating processes*



5 nodes, representing 5 random variables, $\mathcal{X}_V = \{X_1, X_2, \ldots, X_5\}$. The set of edges E consists of $\Big(\{1,2\},\{1,3\},\{2,3\},\ \{3,4\},\{3,5\},\{4,5\}\Big)$. Consequently, the concentration matrix $\mathbf{\Sigma}^{-1}$ has $\sigma^{14} = \sigma^{15} = \sigma^{24} = \sigma^{25} = 0$. Looking at the conditional independence statements encoded by such butterfly graph, one can see that, for instance, because the edge $\{2,5\}$ is missing, we have $X_2 \perp\!\!\!\perp X_5 \mid (X_1, X_3, X_4)$. Such an independence statement derives from the so-called *pairwise Markov property*. Furthermore, if we partition the set V into three subsets, $A = \{X_1, X_2\}$, $B = \{X_4, X_5\}$ and $S = \{X_3\}$, the concentration matrix can be conformably partitioned as

$$\mathbf{\Sigma}^{-1} = \begin{pmatrix} \mathbf{\Sigma}^{AA} & \cdot & \cdot \\ \mathbf{\Sigma}^{SA} & \mathbf{\Sigma}^{SS} & \cdot \\ \mathbf{\Sigma}^{BA} & \mathbf{\Sigma}^{BS} & \mathbf{\Sigma}^{BB} \end{pmatrix} = \begin{pmatrix} \mathbf{\Sigma}^{AA} & \cdot & \cdot \\ \mathbf{\Sigma}^{SA} & \mathbf{\Sigma}^{SS} & \cdot \\ \mathbf{0} & \mathbf{\Sigma}^{BS} & \mathbf{\Sigma}^{BB} \end{pmatrix}$$

that is the sub-matrix $\mathbf{\Sigma}^{BA}$ is a null matrix. This implies, according to the *global Markov property*, that $(X_1, X_2) \perp\!\!\!\perp (X_4, X_5) \mid X_3$. The set S is said to be a *separating set* between A and B, because every path from a node in A to a node in B necessarily passes by the node in S. Similar statements can be specified for the remaining missing edges. The reader is referred to Whittaker (1990), Lauritzen (1996) and Edwards (2000) for a detailed presentation of concentration graph models.

A model selection procedure aims to identify the statistical model giving the best explanation of the data. In graphical models, this corresponds to a procedure for searching among all the possible graphs and detecting which edges are missing. In concentration graph models, this consists of identifying which one of the $p(p-1)/2$ elements σ^{ij} in the concentration matrix is not significantly different from zero.

Most model selection procedures for concentration graphs are based on a stepwise analysis about which edge has to be included or eliminated from the selected graph (Whittaker, 1990; Edwards, 1995; Roverato and Whittaker, 1996) utilizing a series of likelihood-ratio tests or Wald tests. A new model

selection procedure has been recently proposed by Drton and Perlman (2004). This procedure, called SINFUL, takes into account simultaneous confidence intervals for the partial correlations, in order to control the overall error rate for incorrect edge inclusion. The SINFUL procedure utilizes Šidák's inequality (Šidák, 1967) to improve the Bonferroni's adjustment, leading to narrower confidence intervals. If $T_1, \ldots, T_k$ random variables have multivariate normal joint distribution $N_k(\mathbf{0}, \mathbf{\Sigma})$, with $\mathbf{\Sigma}$ positive definite and $\xi_1, \ldots, \xi_k$ arbitrary positive constants, then, according to Šidák's inequality,

$$P(|T_1| \leq \xi_1, \ldots, |T_k| \leq \xi_k) \geq P(|T_1| \leq \xi_1) \cdot P(|T_2| \leq \xi_2) \cdot \ldots \cdot P(|T_k| \leq \xi_k)$$

This inequality is utilized to construct a set of confidence intervals for the partial correlation coefficients ensuring an overall confidence level at least $1 - \alpha$. A test statistic for a partial correlation coefficient is obtained applying the Fisher's z transformation of the partial correlation (Anderson, 1984):

$$T_{ij} = (n - p - 1) \cdot \frac{1}{2} \left(\ln \frac{1 + \hat{\rho}^{ij}}{1 - \hat{\rho}^{ij}} - \ln \frac{1 + \rho^{ij}}{1 - \rho^{ij}} \right) \sim N(0, 1)$$

where ρ^{ij} is the unknown true partial correlation coefficient and $\hat{\rho}^{ij}$ its sample estimate. Given that the set of the $p(1-p)/2$ statistics T_{ij} has an approximate multivariate normal joint distribution, then Šidák's inequality can be applied them for testing the null hypothesis $H_0 : \rho^{ij} = 0$, corresponding to $\ln \frac{1+\rho^{ij}}{1-\rho^{ij}} = 0$. Whenever the data are compatible with the null hypothesis on ρ^{ij} at a prefixed overall confidence level, then the edge between X_i and X_j is removed.

3 Simulation studies to check robustness

As pointed out by Grunert da Fonseca and Fieller (2006), it is important to distinguish between data contamination caused by outliers and by model deviations, because of different implications on the performance of statistical inference procedures. In order to evaluate the impact of contaminations in a concentration graph model, we consider four different situations. The aim is to measure the effect of ignoring such contaminants, assuming a multivariate Normal joint distribution.

In the first case considered, only one univariate variable is measured with error. This case corresponds to the classical presence of outliers due to gross errors or measurement mistakes.

In the second case, we include in the sample a set of multivariate observations from a population having the same conditional independence structure but a different multivariate normal distribution having different variances. This can be viewed both as a model deviation and as presence of outliers on all the variables.

The third case has a model deviation due to contaminants having a different multivariate normal distribution with a different conditional independence structure. This case could correspond to a population being a mixture of two distinct sub-populations, with the same distribution form, but deeply different in the independence structure. The contaminant distribution has variances, covariances and partial correlations that are different from the regular distribution. This can also be viewed as a model deviation.

The fourth case has data contamination by a population having a non-normal distribution. Here we limit to the case of observations coming from an extended skew–normal distribution. This choice derives from the realization that an outlier for a Normal distribution need not be an outlier for a skewed normal distribution.

The Skew-Normal distribution (Azzalini and Della Valle, 1996; Azzalini and Capitanio, 1999) is an extension of the Gaussian distribution having an additional parameter regulating the skewness. This class of distributions fulfills some properties similar to the Normal, such as closure under marginalization and an approximate closure under conditioning. The *extended* Skew-Normal distribution (Azzalini *et al.*, 1996) has the additional property of exact closure under conditioning, useful for graphical models. The conditional independence structure of an extended Skew Normal distribution connects to some restriction on the parameters. For example, denote $\mathcal{X}_V \sim SN_p(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, \tau)$, where $\boldsymbol{\xi}$ is a p-vector of location parameters, $\boldsymbol{\Omega}$ is a full rank covariance matrix, $\boldsymbol{\alpha}$ is a vector of parameters regulating the skewness and τ is an additional real parameter specific for the extended Skew Normal distribution. A zero element in $\boldsymbol{\alpha}$ indicates that the correspondent univariate component has a Gaussian distribution. For the partition of $\mathcal{X}_V$ as $\{X_A, X_B, X_S\}$, $X_A \perp\!\!\!\perp X_B \mid X_S$ if and only if the sub-matrix $\boldsymbol{\Omega}^{AB}$ of $\boldsymbol{\Omega}^{-1}$ is the null matrix and, simultaneously, at least one of α_A and α_B are the null vector. This implies that, whenever $X_A \perp\!\!\!\perp X_B \mid X_S$ it follows that at least one of X_A and X_B has to be Gaussian.

4 Summary of results

We report here the main results of the simulation study to investigate the impact of outliers and contaminants on the SINFUL model selection procedure for concentration graph models. For each one of the four different types of contaminants described in Section 3, we have considered different sample sizes and different proportions λ of contaminants in the sample. For each case, 10000 random samples have been generated.

The main measure with which we compared the impact of contaminants on the inferential procedure is the proportion of correct presence/absence edges identification $\mathcal{I}_E$. Let $k = p(p-1)/2$ be the total number of possible edges in a given graph $\mathcal{G} = (V, E)$ and g the actual cardinality of E. Denote by $\bar{E}$ the set of edges missing in $\mathcal{G}$ and $\bar{g}$ its cardinality. Let $\hat{\bar{g}}$ be the number of edges in $\bar{E}$ assigned as missing by the model selection procedure and let $\hat{g}$ be the number of edges in E assigned as non missing. Then the proportion of correct edges identification is

$$\mathcal{I}_E = \frac{\hat{g} + \hat{\bar{g}}}{k}$$

Then, $\mathcal{I}_E = 1$ if and only if the selected graph coincides with the actual one $\mathcal{G}$. This kind of measure takes into account the overall errors of both types – including edges that should not be, and not including edges that should be included.

As the "true model", we used a graphical model having the butterfly graph shown in Fig.1(a). Here V consists of $p = 5$ nodes, while the set of undirected edges is the one presented in Section 2. The set of missing edges is $\bar{E} = \Big(\{1,4\},\{1,5\},\{2,4\},\{2,5\}\Big)$. Here we assume the joint distribution is $N_5(\mathbf{0}, \mathbf{\Sigma})$, with covariance matrix and partial correlation matrix

$$\mathbf{\Sigma} = \begin{pmatrix} 279.97 & 103.43 & 80.44 & 73.52 & 82.03 \\ 103.43 & 148.68 & 63.03 & 57.61 & 64.27 \\ 80.44 & 63.03 & 95.99 & 87.73 & 97.88 \\ 73.52 & 57.61 & 87.73 & 186.73 & 122.24 \\ 82.03 & 64.27 & 97.88 & 122.24 & 263.74 \end{pmatrix} \qquad \mathbf{R} = \begin{pmatrix} 1.00 & 0.34 & 0.23 & 0.00 & 0.00 \\ 0.34 & 1.00 & 0.28 & 0.00 & 0.00 \\ 0.23 & 0.28 & 1.00 & 0.43 & 0.36 \\ 0.00 & 0.00 & 0.43 & 1.00 & 0.25 \\ 0.00 & 0.00 & 0.36 & 0.25 & 1.00 \end{pmatrix}$$

The assumed $\mathbf{\Sigma}$ matrix is a version of the sample covariance matrix in the math marks data presented by Whittaker (1990), originally analyzed by Mardia *et al.* (1979), modified so to have exactly zero in the proper σ^{ij} elements in $\mathbf{\Sigma}^{-1}$ and the corresponding elements on the partial correlation matrix R.

First of all, to describe the SINFUL model selection procedure and to check its validity, we generated $10,000$ samples from the butterfly graph model. The results are reported in Table 1.

According to the simulation study, we note that the procedure works poorly for relatively small sample size. When $n = 100$ only in the 3.63% of the samples the procedure selects the correct graph. Such percentage rises to 75.28% for $n = 250$, achieving the appropriate limiting level of 95% when n is about 500. As a consequence, we will report here for the contaminated simulated data only the cases with $n \geq 500$.

Table 1 Simulations results on non-contaminated data, showing the distribution of the proportion of correct edges and non-edges as a function of the sample size n.

$\mathcal{I}_E$	$n = 100$	$n = 250$	$n = 500$	$n = 750$	$n = 1000$
0.0-0.3	0.00	0.00	0.00	0.00	0.00
0.4	0.01	0.00	0.00	0.00	0.00
0.5	0.32	0.00	0.00	0.00	0.00
0.6	4.67	0.00	0.00	0.01	0.01
0.7	26.06	0.12	0.03	0.04	0.03
0.8	43.59	2.88	0.48	0.35	0.25
0.9	21.72	21.72	5.21	4.63	3.99
1.0	3.63	75.28	94.28	94.97	95.72

In a first set of simulations with contaminated data we considered the case of presence of univariate gross-errors in the data. In this case, some univariate observations are measured with error occurring on one random variable at a time. The contaminants come from univariate Normal distribution with a doubled standard deviation. The uncontaminate observations come form the 5-variate Normal distribution following the butterfly graph. Proportions $\lambda = 0.05$ and 0.10 of the data were randomly substituted with contaminated observations.

Table 2 Correct edge identification index in the case of data with 1% of univariate contaminants occurring on one of the five random variables in $\mathcal{X}_V$

	$N = 500$					$N = 1000$				
$\mathcal{I}_E$	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
0.0-0.2	0.00	0.00	0.41	0.00	0.00	0.00	0.00	8.72	0.00	0.00
0.3	0.00	0.00	8.38	0.00	0.00	0.00	0.00	29.66	0.00	0.00
0.4	0.00	0.00	31.41	0.00	0.00	0.00	0.00	34.69	0.00	0.00
0.5	0.00	0.01	34.88	0.00	0.00	0.00	0.00	19.40	0.00	0.00
0.6	0.11	0.07	17.13	0.15	0.15	0.06	0.08	6.43	0.04	0.11
0.7	3.02	2.64	5.57	2.42	2.93	2.03	1.72	1.05	1.74	2.28
0.8	85.34	74.26	1.65	74.99	84.40	80.27	60.31	0.05	61.50	78.17
0.9	11.17	20.72	0.48	21.34	12.14	17.35	33.67	0.00	35.07	19.15
1.0	0.36	2.30	0.09	1.10	0.38	0.29	4.22	0.00	1.65	0.29

Table 3 Correct edge identification index in the case of data with 5% of univariate contaminants occurring on one of the five random variables in $\mathcal{X}_V$

	$N = 500$					$N = 1000$				
$\mathcal{I}_E$	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
0.0-0.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.2	0.00	0.00	1.06	0.00	0.00	0.00	0.00	32.46	0.00	0.00
0.3	0.00	0.00	15.50	0.00	0.00	0.00	0.00	49.39	0.00	0.00
0.4	0.00	0.00	46.44	0.00	0.00	0.00	0.00	16.34	0.00	0.00
0.5	0.01	0.00	32.03	0.01	0.00	0.00	0.00	1.73	0.00	0.00
0.6	0.15	0.06	4.71	0.12	0.10	0.05	0.05	0.08	0.09	0.12
0.7	3.06	3.43	0.24	3.08	3.30	2.52	2.61	0.00	2.49	2.85
0.8	94.55	93.41	0.02	93.57	94.10	94.37	91.65	0.00	91.72	93.79
0.9	2.19	3.08	0.00	3.13	2.34	3.04	5.66	0.00	5.66	3.13
1.0	0.04	0.02	0.00	0.09	0.16	0.02	0.03	0.00	0.04	0.11

As illustrated in Table 2, the model selection procedure broke down immediately also with only the 1% of outliers, showing that the estimates and the hypothesis tests are in a sense sensitive to the presence of gross errors. It is worth to note the greatest impact on the correct edge identification due to contaminants regarding the variable X_3, namely the node which is a separating set considering the whole graph. Moreover, comparing Table 2 and 3, one can note that, as expected, the model selection procedure gives worse results when the proportion of contaminants increases, but is not better when the sample size raises for a fixed proportion λ.

For the second type of contaminants, we considered the case of 5-variate observations coming from a different joint Normal distribution with the same conditional independence structure of the butterfly graph. In particular, we took into account the case of doubled standard deviations with respect to the uncontaminated data.

In this case (Table 4), even if the conditional independence structure on contaminants is the same as for the target population, the model selection procedure is slightly affected also by the mild change in the variances, more heavily when λ increase. Note that when the sample size increases, the SINFUL model selection procedure seems more sensitive to model deviances. This behavior was confirmed also in other cases.

The third kind of contaminated data are observations from a Gaussian distribution with a different conditional independence structure. In particular, we considered the case of contaminants having similar variances in terms of order of magnitude to the true population, to isolate the effect of the change

Table 4 Correct edge identification index in case of multivariate contaminants with the same independence structure

	$\lambda = 1\%$			$\lambda = 5\%$		
$\mathcal{I}_E$	$n = 100$	$n = 500$	$n = 1000$	$n = 100$	$n = 500$	$n = 1000$
0.0-0.3	0.00	0.00	0.00	0.00	0.00	0.00
0.4	0.03	0.00	0.00	0.22	0.00	0.00
0.5	0.67	0.00	0.00	2.50	0.00	0.00
0.6	6.04	0.02	0.03	12.13	0.20	0.32
0.7	28.11	0.10	0.12	34.66	0.56	1.36
0.8	43.74	1.07	1.10	38.40	5.89	10.30
0.9	18.46	8.58	9.63	10.80	31.84	41.93
1.0	2.95	90.23	89.12	1.29	61.51	46.09

in the conditional independence structure. The procedure seems to be more robust against this kind of contaminants than to an increment in the magnitude of the variance, unless the proportion of contaminants becomes sizable. In Table 5 we report two different independence structure: the first, depicted in figure $1(b)$, while the second is the complementary graph $\mathcal{G} = (V, \bar{E})$. Contaminants from population with a conditional independence structure more similar to the one of the assumed true population, the butterfly graph, do not affect considerably the procedure, unless λ is near 0.5.

Table 5 Correct edge identification index in case of multivariate contaminants with different independence structure (n=1000)

	Graph in Figure $1(b)$			*Complementary graph*		
$\mathcal{I}_E$	$\lambda = 0.05$	$\lambda = 0.10$	$\lambda = 0.30$	$\lambda = 0.05$	$\lambda = 0.10$	$\lambda = 0.30$
0.0-0.4	0.00	0.00	0.00	0.00	0.00	0.11
0.5	0.00	0.00	0.00	0.00	0.00	11.20
0.6	0.03	0.07	0.23	0.88	9.90	85.02
0.7	0.26	0.89	2.92	18.25	72.53	3.67
0.8	4.40	17.33	63.41	37.74	16.14	0.00
0.9	31.73	71.64	33.44	32.34	1.43	0.00
1.0	63.58	10.07	0.00	10.79	0.00	0.00

The last case considered involves contaminants having a Skew-Normal distribution. Here we took into account the case of a Skew-Normal multivariate distribution having the same butterfly graph as

the target population. In this way, we can observe in Table 6 the effect of asymmetry apart from the effect considered in the previous case (Table 5). The SINFUL model selection procedure appears quite robust unless the amount of contaminate observations become large. Nevertheless, the particular structure of the butterfly graph is such that all the nodes but X_3 are involved in at least one conditional independence constraint. Consequently, as explained in the last paragraph of Section 3, in the assumed contaminant population, only X_1, X_2, X_3 are actually skewed. The alternative choice, assuming X_3, X_4, X_5 as actually skewed, does not reveal tangible differences.

Table 6 Correct edge identification index in case of multivariate contaminants with Skew-Normal distribution

	$\lambda = 1\%$		$\lambda = 5\%$		$\lambda = 100\%$	
$\mathcal{I}_E$	$n = 500$	$n = 1000$	$n = 500$	$n = 1000$	$n = 500$	$n = 1000$
0.0-0.4	0.00	0.00	0.00	0.00	0.00	0.00
0.5	0.00	0.00	0.00	0.00	0.01	0.00
0.6	0.00	0.01	0.00	0.02	0.29	0.02
0.7	0.02	0.03	0.03	0.02	9.86	0.83
0.8	0.41	0.26	0.41	0.30	52.75	21.50
0.9	5.11	4.02	5.16	4.28	33.98	60.41
1.0	94.46	95.68	94.40	95.38	3.11	17.24

5 Final remarks

Concluding, the results of this study suggest a considerable potential effect in the model selection procedure of even light contamination. The simulation study carried out suggests that four kinds of contaminants have different impact on the model selection.

The greatest impact on the model selection procedure observed in this paper is due to gross errors or observations having a larger variance than the target population. Also in concentration graphs, therefore, influential observations are those far away from the majority of the data. In particular, a sizable impact seems to be due to univariate contaminants regarding variables whose nodes are in a separating set of the graph. This great impact can be motivated by the fact that nodes in separating sets have an important role in the factorization of the joint distribution according to the concentration graph. The model selection procedure showed a more robust behavior against small amount of contaminants from a different population having the same order of magnitude of the variances, but a different conditional independence structure or asymmetric joint distribution.

On the basis of the simulation study, we suggest that in practice, an accurate contaminants detection be attempted before conducting the model selection procedure, particularly for those variables that could have a separating role in the concentration graph. However, the simulation study here was limited to a few kinds of graphical models for given correlation matrices, and other analysis are required to generalize the results. Moreover, it is important to remark that the obtained results are strictly connected both to the chosen inferential procedure to estimate the variance covariance $\mathbf{\Sigma}$ and to the chosen model selection procedure. Different choices could bring to different results. In future research, robustness of undirected graphs for log-linear models for categorical responses could be of investigated with respect to different kind of contaminants.

The statistician George Box is often quoted as saying that ”All models are wrong, but some are useful.” In practice, *any* model that we specify does not truly hold in practice. In this sense, when combined with the Gaussian model assumed, the contaminant structures used in our paper could be regarded as different ways also that reality may differ from simple models idealized in practice. With this view, our focus could change from ”Did we choose the correct model?” to ”Are our substantive conclusions about reality correct?” A possible generalization of the study in this paper could attempt to address the effects of different types of contaminants on such substantive conclusions, for example regarding parameters that describe effect sizes. It’s not obvious how to operationalize such an investigation, but this is an important problem for further research.

Acknowledgements The authors would like to thank Matilde Bini and Giovanni Marchetti for having encouraged their interest on this topic. Moreover, we wish to thank Alan Agresti for the valuable comments and suggestions.

References

Anderson T.W. (1984) *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons.

Azzalini A. and Capitanio A. (1999) Statistical applications of the multivariate skew-normal distribution, *Journal of the Royal Statistical Society: Series B*, 61, 3, 579–602.

Azzalini A., Capitanio A. and Stanghellini E. (1996) Graphical models for skew normal variates, *Scandinavian Journal of Statistics*, 30, 129–144.

Azzalini A. and Della Valle A. (1996) The multivariate skew-normal distribution, *Biometrika*, 83, 715–726.

Cox D.R. and Wermuth N. (1993) Linear dependencies represented by chain graphs (with discussion), *Statistical Science*, 8, 204–218.

Cox D.R. and Wermuth N. (1996) Multivariate Dependencies. Models, Analysis and Interpretation, Chapman and Hall, London.

Dawid A.P. (1979) Conditional independence in statistical theory (with discussion), Journal of the Royal Statistical Society, Series B, 41, 1–31.

Dempster A.M. (1972) Covariance selection, *Biometrics*, 28, 157–175.

Drton M. and Perlman M. (2004) A sinful approach to gaussian graphical model selection, avaible in http://www.stat.washington.edu/drton/Papers/2005statsci.pdf.

Edwards D. (1995) *Introduction to Graphical Modelling*, Springer-Verlag Inc.

Edwards D. (2000) *Introduction to graphical modelling. Second edition*, New York: Springer–Verlag.

Grunert da Fonseca V. and Fieller N.R.J. (2006) Distorion in statistical inference: the distinction between data contamination and model deviation, *Metrika*, 63, 169–190.

Kuhnt S. and Becker C. (2003) Sensitivity of graphical modeling against contamination, in: *Between Data Science and Applied Data Analysis*, Springer, ed., Schader, M.; Gaul, W.; Vichi, M. (eds.), Berlin Heidelberg New York, 279–287.

Lauritzen S. (1996) *Graphical Models*, Oxford: Oxford Science Publications.

Mardia K., Kent J. and Bibby J. (1979) *Multivariate Analysis*, Academic Press, London.

Roverato A. and Whittaker J. (1996) Standard errors for the parameters of graphical Gaussian models, *Statistics and Computing*, 6, 297–302.

Šidák Z. (1967) Rectangular confidence regions for the means of multivariate normal distributions, *Journal of the American Statistical Association*, 62, 626–633.

Whittaker J. (1990) *Graphical Models in Applied Multivariate Statistics*, Wiley: New York.

Copyright © 2006
Anna Gottard, Simona Pacillo