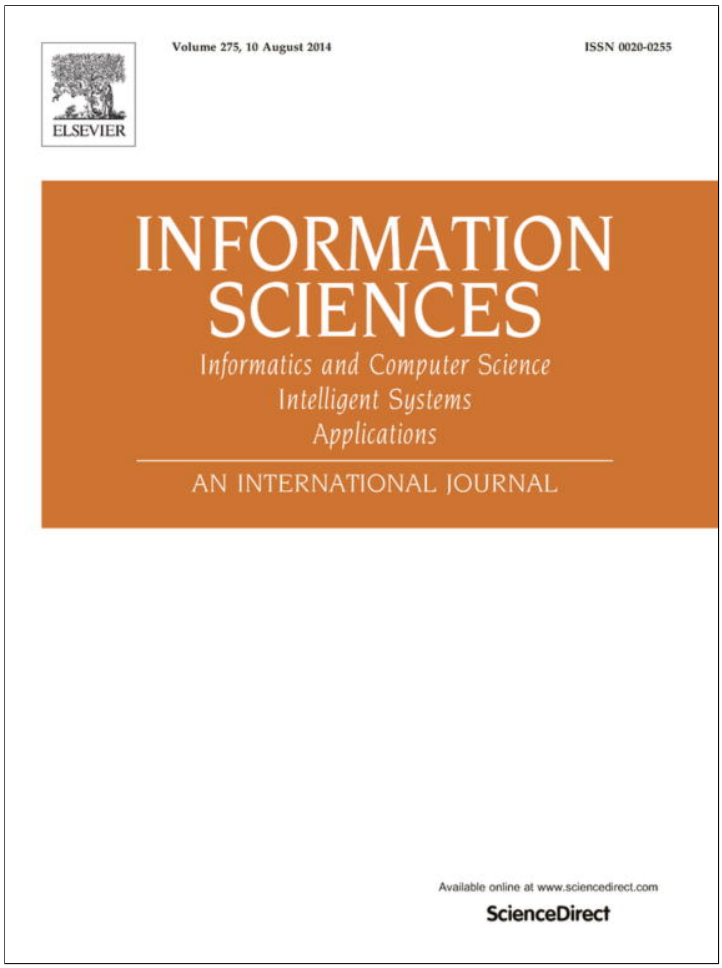

Information Sciences 275 (2014) 115–126

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Sensitivity to evidence in Gaussian Bayesian networks using mutual information

Miguel Angel Gómez-Villegas [a,*], Paloma Main [a], Paola Viviani [b]

[a] Department of Statistics and O.R., Complutense University of Madrid, Spain
[b] Department of Public Health, Faculty of Medicine, Pontificia Universidad Católica de Chile, Chile

ARTICLE INFO

Article history:
Received 22 September 2011
Received in revised form 26 October 2013
Accepted 9 February 2014
Available online 19 February 2014

Keywords:
Entropy
Evidence propagation
Gaussian Bayesian network
Mutual information
Sensitivity analysis

ABSTRACT

We introduce a methodology for sensitivity analysis of evidence variables in Gaussian Bayesian networks. Knowledge of the posterior probability distribution of the target variable in a Bayesian network, given a set of evidence, is desirable. However, this evidence is not always determined; in fact, additional information might be requested to improve the solution in terms of reducing uncertainty. In this study we develop a procedure, based on Shannon entropy and information theory measures, that allows us to prioritize information according to its utility in yielding a better result. Some examples illustrate the concepts and methods introduced.

1. Introduction

A Bayesian network is a probabilistic graphical model that represents the conditional dependencies among a set of random variables through a directed acyclic graph (DAG). Bayesian networks have become an increasing popular representation for reasoning under uncertainty and are widely applied to diverse fields, such as medical diagnosis, image recognition, and decision-making systems, among many others.

Formally, a Bayesian network consists of qualitative and quantitative parts. The quantitative part is given by a DAG, whose nodes represent random variables that may be observable, latent, or a target variable of interest. The qualitative part, specifies the conditional probability distribution for each node given its parents; this allows us to compute the joint probability distribution of the model.

The aim of Bayesian network analysis is usually to obtain the conditional probability distribution of a target variable when a set of observable variables (evidence values) is available. Sometimes the variables defined as evidence are fixed in advance but other times they vary from model to model.

In this context, sensitivity analysis is a method for investigating the relationship between network inputs and the conditional distribution of the target variable, for which inputs can be the parameters considered in the conditional probability distribution or actual values taken by the observed variables. There is a large body of literature dealing with sensitivity analysis techniques for Bayesian networks. Most studies have addressed discrete Bayesian networks. For example, Malhas and Al Aghbari [16] introduced a score based on mutual information increases to discover new interesting patterns. Chan and Darwiche [4] presented a distance measure between the original distribution and a new one in which the parameters have

* Corresponding author. Tel.: +34 91 3944428; fax: +34 91 3944606.
E-mail addresses: ma_gv@mat.ucm.es (M.A. Gómez-Villegas), pmain@mat.ucm.es (P. Main), paolav@mat.puc.cl (P. Viviani).

http://dx.doi.org/10.1016/j.ins.2014.02.025
0020-0255/

been changed. Laskey [14] measured sensitivity by computing the partial derivatives of output probabilities with respect to given parameters. Kostal et al. [13] proposed measures of statistical dispersion based on Shannon and Fisher information. Castillo and Kjaerulff [3] developed a sensitivity analysis for Gaussian Bayesian networks (GBNs) using partial derivatives and symbolic propagation. Gómez-Villegas et al. [7–10] used Kullback Leibler divergence as a measure of sensitivity in GBNs.

Here we focus on a different aspect of sensitivity analysis. As mentioned previously, the set of evidence variables is not specified in advance in many real-life problems. In fact, it is usual practice to try to collect as much information as possible. However, this information always has an associated cost, so it may be desirable to evaluate which of all the available variables are most informative and useful for obtaining the best results. A very important assumption made in this paper is that a *better result* is achieved if the conditional probability distribution of the target variable has the lowest uncertainty, that is, the lowest entropy. Thus, we use information theory to provide tools to prioritize the available information to reduce the uncertainty of the target variable as far as possible.

The remainder of the article is structured as follows. In Section 2 we briefly review GBNs and show how propagation of observable values can be performed in this case. We also introduce our working example. Section 3 presents some general concepts of entropy, mutual information, and normalized measures. In Section 4, we first propose a procedure to study the sensitivity to evidence in GBNs and then perform a sensitivity analysis on our working example. The second contribution of the paper is presented in Section 5, which is an extension of the sensitivity analysis proposed above but incorporating normalized measures. Results are presented for the working example and a supplementary example. Finally, in Section 6 we draw conclusions.

2. Gaussian Bayesian networks

In this section we first recall the definition of a general Bayesian network and then the special case of a GBN. We also present the methodology for evidence propagation in GBNs.

2.1. Definition: Bayesian network

A *Bayesian network* is a pair $(\mathcal{G},\mathcal{P})$, where $\mathcal{G}$ is a directed acyclic graph (DAG) with one node for each random variable of $\mathbf{X}=\{X_1,\ldots,X_n\}$ and edges that represent probabilistic dependencies between them.

$\mathcal{P}=\{p(x_1|pa(x_1)),\ldots,p(x_n|pa(x_n))\}$ is a set of conditional probability distributions and $pa(x_i)$ is the set of parents of node X_i in $\mathcal{G}$. From $\mathcal{P}$, the associated joint probability distribution for $\mathbf{X}$ is defined as

$$P(\mathbf{X})=od_{i=1}^{n}P(X_i|pa(X_i)). \tag{1}$$

The type of random variables, X_i, considered in the problem defines whether we are dealing with discrete, Gaussian, or mixed Bayesian networks. In this paper we develop results for GBNs that are based on continuous variables; these have been studied by Castillo et al. [2], Cowell et al. [6] and Gómez-Villegas et al. [7], among others.

2.2. Definition: GBN

GBNs are a subclass of Bayesian networks in which the joint probability density of $\mathbf{X}$ is a multivariate normal distribution $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$, that is,

$$f(\mathbf{x})=(2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\},$$

where $\boldsymbol{\mu}$ is the n-dimensional mean vector, $\boldsymbol{\Sigma}$ is the positive definite $n\times n$ covariance matrix, $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$, and $\boldsymbol{\mu}^T$ the transpose of $\boldsymbol{\mu}$.

According to the normal distribution properties and the factorization presented in (1), in a GBN the joint probability density can be specified also as a product of conditional probability densities, each of which corresponds to a univariate normal distribution.

2.3. Evidence propagation in a GBN

In real-life problems, information about the state of one or more variables of a Bayesian network, known as evidence variables, may be available. If so, probability distributions for the rest of the variables in the network can be updated given the observed values. This process is called *evidence propagation*.

Different algorithms have been proposed for evidence propagation in GBNs. Here, we consider an incremental method developed by Castillo et al. [2]. This consists of computing the conditional probability density of a normal distribution after introducing one evidential variable at a time. We consider the set of non-evidential variables $\mathbf{Y}$ and the evidential variables $\mathbf{E}$. Then $\mathbf{X}$ can be written as the partition $\mathbf{X}=(\mathbf{Y},\mathbf{E})$, and the conditional distribution of $\mathbf{Y}$ given $\mathbf{E}=\mathbf{e}$ is a multivariate normal distribution with parameters

$$\boldsymbol{\mu}^{\mathbf{Y}|\mathbf{E}=\mathbf{e}} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\Sigma}_{\mathbf{YE}}\boldsymbol{\Sigma}_{\mathbf{EE}}^{-1}(\mathbf{e} - \boldsymbol{\mu}_{\mathbf{E}}) \tag{2}$$

and

$$\boldsymbol{\Sigma}^{\mathbf{Y}|\mathbf{E}=\mathbf{e}} = \boldsymbol{\Sigma}_{\mathbf{YY}} - \boldsymbol{\Sigma}_{\mathbf{YE}}\boldsymbol{\Sigma}_{\mathbf{EE}}^{-1}\boldsymbol{\Sigma}_{\mathbf{EY}}. \tag{3}$$

If we are interested in the posterior marginal density of just one variable considered as a target, $X_i \in \mathbf{Y}$, and one evidence variable, E, after evidence propagation we have,

$$X_i|E = e \sim N\left(\mu_i^{Y|E=e}, \sigma_{ii}^{Y|E=e}\right) = N\left(\mu_i + \frac{\sigma_{ie}}{\sigma_{ee}}(e - \mu_e), \sigma_{ii} - \frac{\sigma_{ie}^2}{\sigma_{ee}}\right),$$

where μ_i and μ_e are the mean of X_i and E, σ_{ii} and σ_{ee} are the variance of X_i and E, respectively and σ_{ie} is the covariance between X_i and E before evidence propagation.

To review the concepts presented in this section we now present an example introduced by Gómez-Villegas et al. [9].

2.4. Example

This problem is about the duration for which a machine is working. The machine is made up of seven elements, connected according to the DAG in Fig. 1.

The target is the variable X_7 and the joint probability distribution of $\mathbf{X}$ is a multivariate normal distribution, with the following parameters given by experts:

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 3 \\ 2 \\ 1 \\ 4 \\ 5 \\ 8 \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \mathbf{D} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix},$$

where $\boldsymbol{\mu}$ is the n-dimensional mean vector, $\mathbf{D}$ is a diagonal matrix with conditional variances v_i, $\mathbf{B}$ is a strictly upper triangular matrix with regression coefficients β_{ji}, and X_j is a parent of X_i, with $j < i$.

It is well known that the covariance matrix $\boldsymbol{\Sigma}$ can be computed as

$$\boldsymbol{\Sigma} = \left[(\mathbf{I} - \mathbf{B})^{-1}\right]^T \mathbf{D}[(\mathbf{I} - \mathbf{B})^{-1}].$$

Thus, for this example we obtain that $\mathbf{X}$ has a multivariate normal distribution with the following parameters,

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 3 \\ 2 \\ 1 \\ 4 \\ 5 \\ 8 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 2 & 2 \\ 0 & 1 & 0 & 2 & 2 & 8 & 8 \\ 0 & 0 & 2 & 0 & 2 & 4 & 4 \\ 1 & 2 & 0 & 6 & 4 & 20 & 20 \\ 0 & 2 & 2 & 4 & 10 & 28 & 28 \\ 2 & 8 & 4 & 20 & 28 & 97 & 97 \\ 2 & 8 & 4 & 20 & 28 & 97 & 99 \end{pmatrix}.$$

Assume that we know the values of X_1, X_2 and X_3, say, $\mathbf{E} = \{X_1 = 2, X_2 = 2, X_3 = 1\}$. Then we can obtain the posterior probability distribution of the non-evidential variables after performing evidence propagation. Thus, we obtain $\mathbf{Y}|\mathbf{E} \sim N(\mu^{\mathbf{Y}|\mathbf{E}=e}, \Sigma^{\mathbf{Y}|\mathbf{E}})$ with

$$\mu^{\mathbf{Y}|\mathbf{E}=e} = \begin{pmatrix} 0 \\ 1 \\ -3 \\ 0 \end{pmatrix} \quad \Sigma^{\mathbf{Y}|\mathbf{E}} = \begin{pmatrix} 1 & 0 & 2 & 2 \\ 0 & 4 & 8 & 8 \\ 2 & 8 & 21 & 21 \\ 2 & 8 & 21 & 23 \end{pmatrix}.$$

Note that the marginal distribution of the target variable X_7 in the original network was $X_7 \sim N(8, 99)$, which implies considerable uncertainty owing to its high variability. However, after evidence propagation, the updated marginal distribution for this variable is $X_7 \sim N(0, 23)$. Thus, the uncertainty has decreased significantly due to actual observations. The question still arises as to whether the variables $\mathbf{E} = \{X_1 = 2, X_2 = 2, X_3 = 1\}$ are the best choice for reducing the variance of X_7. In this paper, we provide the necessary tools to answer this question.

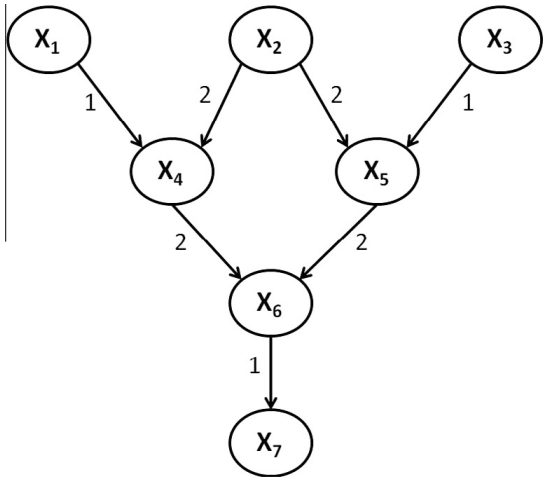

Fig. 1. DAG for Example 2.4.

3. Entropy and mutual information

In this section, we present a brief review of information theory and Shannon entropy [18]. Because we are dealing with normally distributed variables, the definitions are given for the case of continuous variables [5].

3.1. Definition: differential entropy

Differential entropy refers to the entropy of a continuous variable X with probability density function $f(x)$ and is given by

$$h(X) = -\int_S f(x)\ ln\ f(x)\ dx,$$

where S is the support set (the set for which $f(x) > 0$) of the random variable.

For the discrete case, entropy is a measure of uncertainty or randomness of a random variable. However, intuitively speaking, uncertainty or randomness of a continuous variable is infinite [11]. Thus, for discrete variables, entropy measures uncertainty in an *absolute* way, but for continuous variables the measurement is *relative*, and differences in entropy may be compared between two or more variables or between values of the same variable under different models.

The *joint differential entropy* of a set of random variables $X_1,\ldots,X_n$ distributed according to the joint probability density function $f(x_1,\ldots,x_n)$ is defined as

$$h(X_1,\ldots,X_n) = -\int f(x_1,\ldots,x_n)\ ln\ f(x_1,\ldots,x_n)\ dx_1\cdots dx_n.$$

As we are working with GBNs, we need the differential entropy for univariate and multivariate normal distributions [5].

Entropy for a normal distribution

Let $X \sim N(\mu,\sigma^2)$. Then,

$$h(X) = \frac{1}{2}\ ln\ (2\pi e\sigma^2). \tag{4}$$

Entropy for a multivariate normal distribution

Let $X_1, X_2, \ldots, X_n$ have a multivariate normal distribution $N(\mu,\Sigma)$. Then,

$$h(X_1, X_2, \ldots, X_n) = \frac{1}{2}\ ln\ [\ (2\pi e)^n|\Sigma|\ ], \tag{5}$$

The *conditional differential entropy* for two random variables X and Y with joint density function $f(x,y)$ can be computed as

$$h(X|Y) = h(X,Y) - h(Y). \tag{6}$$

The joint entropy measures how much uncertainty there is in a set of random variables $X_1,\ldots,X_n$ taken together, and the conditional entropy is a measure of how much uncertainty remains about the random variable X when Y is known.

Note that if we are working with a normal distribution, then the entropy, the joint entropy and therefore the conditional entropy, all depend only on the covariance matrix of the random variables, which does not consider the evidence value (3). This means that if the conditional entropy is computed, only the conditioning variable matters, and not the value taken by the variable, that is, $h(X|Y = y) = h(X|Y)$.

3.2. Definition: mutual information

The *mutual information* between two continuous random variables X, Y with joint density $f(x,y)$ is defined as

$$I(X;Y) = \int f(x,y)\ ln\ \frac{f(x,y)}{f(x)f(y)}\ dxdy.$$

From this definition, we have

$$I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) = h(X) + h(Y) - h(X,Y). \tag{7}$$

Mutual information measures the information shared by two random variables; this quantifies how much the uncertainty of one variable is reduced provided that the other is known. It could be considered as a measure of dependence, because $I(X;Y) = 0$ only if X and Y are independent. This measure will always be non-negative for both discrete and continuous variables.

For a normal distribution, the mutual information can be specified in terms of the correlation between X and Y. Consider two variables, X and Y, that follow a joint normal distribution with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and correlation coefficient ρ. Then,

$$I(X;Y) = -\frac{1}{2}\ ln(1-\rho^2).$$

The *conditional mutual information* measures the dependence between two random variables X and Y given a third variable Z, and is defined by

$$I(X;Y|Z) = h(X,Z) + h(Y,Z) - h(X,Y,Z) - h(Z). \tag{8}$$

In general terms, if we obtain a small value of $I(X,Y|Z)$ in observing Z, this indicates that this variable decreases the dependence between X and Y. In the limit, if $I(X,Y|Z) = 0$, then X and Y are independent given Z.

3.3. Definition: normalized mutual information

Although all the results referred to in this section were given for discrete variables, we introduce them using the corresponding notation for continuous variables.

The *normalized mutual information* between two variables X and Y was suggested by Strehl and Ghosh [19] as follows:

$$NI(X;Y) = \frac{I(X;Y)}{\sqrt{h(X)h(Y)}}, \tag{9}$$

where $h(X)$ and $h(Y)$ are the differential entropies for these variables.

Then NI is a measure that ranges from 0, for which X is independent of Y, to 1 in the limit of $X = Y$ (in this case, Cover and Thomas [5] show that $I(X,X) = h(X)$). Note that if both differential entropies of X and Y are positive or negative, then the denominator will be a real number, but if only one of them is negative, then the result will be a complex number. In particular, in our specific case in which the random variables have normal distributions, the entropy defined by (4) and (5) will be negative if

$$\sigma^2 < \frac{1}{2\pi e}(\sim 0.0585) \text{ for a univariate normal distribution}$$

and

$$|K| < \frac{1}{(2\pi e)^n} \text{ for a multivariate normal distribution.}$$

This means that if we are comparing two variables X and Y and only one of them has a small variance, (0.0585 for univariate normal distributions or 0.0585^2 for the bivariate case, and so on), then a complex number will be obtained for the denominator, and it is better to use the procedure proposed in Section 4.1. However, in most cases this does not occur because the problem usually starts with high variances rather than great differences between the variables, and the aim is to reduce this uncertainty.

For three random variables, Richiardi [17] proposed the *normalized conditional mutual information* as

$$NI(X;Y|Z) = \frac{I(X;Y|Z)}{\sqrt{h(X|Z)h(Y|Z)}}, \tag{10}$$

where $NI(X;Y|Z)$ is between 0 (X independent of Y given Z) and 1 (X = Y given Z). Similarly, for $NI(X;Y)$, care is necessary to satisfy properly the conditions required to obtain a real number.

Finally, we present the *normalized difference* proposed by Besson et al. [1], which aids interpretation of the results:

$$\Delta I_{XYZ} = \frac{[NI(X;Y) - NI(X;Y|Z)]}{NI(X;Y)}. \tag{11}$$

4. Sensitivity to evidence based on mutual information

For Bayesian networks it is desirable to know the posterior probability distribution of the target variable given a set of evidence. However, this evidence is not always determined; in fact, additional information might be requested to improve the solution in terms of reducing uncertainty. Because any collection of information has an associated cost, it is important to prioritize which information will be of greatest utility.

In this section, we introduce a methodology based on information theory that allows us to measure the potential usefulness of incorporating additional information before the information source is consulted. To this end, Kjaerulff and Madsen [12] developed a procedure for discrete Bayesian networks they called *value of information analysis*. We first present an extension of their results to GBNs and then incorporate normalized measures into the analysis.

4.1. Procedure for prioritizing evidence

The principal aim of this procedure is to identify variables with the highest mutual information with the target variable, since these potential evidence variables will further reduce its uncertainty.

Consider the set of non-evidential variables $\mathbf{Y}$ (initially we can assume that $\mathbf{Y} = \mathbf{X}$, because $E = \phi$) and the target variable $X_i \in \mathbf{Y}$; we refer to $\mathbf{Y}_{-i}$ as the set of non-evidential variables without considering the target variable X_i, that is, $\mathbf{Y}_{-i} = \mathbf{Y} \setminus X_i$. Then the procedure involves the following steps:

1. Calculate the entropy for the target variable $h(X_i)$.
2. Compute the mutual information between the target variable X_i and each non-evidential variable $\mathbf{Y}$ different from X_i, that is, $I(X_i; \mathbf{Y}_{-\mathbf{i}})$.
3. Choose as the evidential variable the X_k from $\mathbf{Y}_{-i}$ that has the highest non-zero mutual information with X_i.
4. Look at the decrease in uncertainty as $h(X_i|\mathbf{E}) - I(X_i; X_k|\mathbf{E}) = h(X_i|\mathbf{E}, X_k)$.
5. Consider that $X_k \in E$.
6. Compute the conditional mutual information between the target variable and the new set of non-evidential variables $\mathbf{Y}_{-\mathbf{i}}$, that is, $I(X_i; \mathbf{Y}_{-\mathbf{i}}|\mathbf{E})$.
7. Go back to 3.

Notes:

- Stop when the uncertainty of X_i is sufficiently small or when there are no more non-evidential variables available with significant non-zero mutual information.
- If the variable with the highest mutual information score is not available, then proceed to observe the variable with the second-highest score.
- Variables with mutual information of zero (or close to zero) should not be considered as evidence, because they will not add any information to the analysis.

We use Example 2.4 to illustrate the procedure proposed in this section.

4.2. Example

Consider the GBN given in Example 2.4 and suppose that no evidential variables have been determined yet. We have defined X_7 as the target variable, so applying the specifications to the normal distributions given in (4) and (5) and the definition for mutual information (7), steps 1 and 2 are computed.

We obtained differential entropy for X_7 of $h(X_7) = 3.7165$ and determined the mutual information scores for the rest of the variables with the target. The results are shown in Table 1.

It is evident from Table 1 that an order of priority for reducing the uncertainty of the target is obtained. The most informative variable for X_7 is X_6, followed by X_5.

To show how the decision on which variable to incorporate as evidence in the network affects the uncertainty of the target variable X_7, Table 2 presents the posterior variance and conditional differential entropy of X_7, computed using (3) and (6), respectively, in the simulated case in which each of the non-evidential variables was observed. We have seen that these measures do not depend on the evidence value, so we consider only the set of evidence variables.

As we saw before, the initial distribution of X_7 was $N(8, 99)$ with differential entropy of $h(X_7) = 3.7165$, so now it is evident that X_6 is the most important variable in decreasing the uncertainty of X_7. In fact, if we could observe only this variable, the problem would be solved with high accuracy. However, it is of worth noting that two variables that were considered as

Table 1
Example of step 2.

Y_{-i}	$h(Y_{-i})$	$h(X_7, Y_{-i})$	$I(X_7; Y_{-i})$
X_1	1.4189	5.1148	0.0206
X_2	1.4989	4.6155	0.5199
X_3	1.7655	5.4399	0.0421
X_4	2.3148	5.4718	0.5595
X_5	2.5702	5.5018	0.7849
X_6	3.7063	5.4718	1.9509

Table 2
Example of the sensitivity of X_7 to evidence.

Y_{-i}	$Var(X_7\|Y_{-i})$	$h(X_7\|Y_{-i})$
X_1	95	3.6959
X_2	35	3.1966
X_3	91	3.6743
X_4	32.3	3.1564
X_5	20.6	2.9316
X_6	2	1.7655

evidence when this example was introduced in Section 2.4, X_1 and X_3, had no significant individual contribution to the reduction in variance of X_7.

To continue with the proposed procedure, imagine that X_6 is not available, but that X_5 (the second most informative variable) is observable. If we introduce this variable as evidence into the network and apply step 3, we obtain the result $Var(X_7|X_5) = 20.6$. Then, proceeding to step 4, the new differential entropy $h(X_7|X_5)$ is calculated, giving a value of 2.9316, which implies a significant reduction in uncertainty of X_7.

According to steps 5 and 6, respectively, $X_5 \in \mathbf{E}$ and all the conditional mutual information scores for the rest of the variables Y_{-i} with X_7 are calculated using definitions (6) and (8). Equivalently, if we calculate the conditional multivariate normal distribution of $(X_1, X_2, X_3, X_4, X_6, X_7|X_5)$, then, the corresponding measures can be obtained as in Table 1. The results for this step are shown in Table 3.

From Table 3 it is evident that X_6 is still the most informative variable for X_7, but as we supposed that this variable was not available, we observe the variable with the second highest mutual information, X_4. Proceeding to step 3, after propagating X_4 as the evidential variable, we get $Var(X_7|X_5, X_4) = 3$, which implies that high accuracy is obtained.

Finally, the conditional entropy for $X_7|X_5, X_4$ is $h(X_7|X_5, X_4) = 1.968$ and Table 4 shows the new measures for entropy and mutual information. Note that the entropy for the target has been reduced significantly, and that no other variable adds much information to the target, not even X_6. In fact, we obtain information values of zero for the first three variables, that is, they are independent of X_7 given X_4 and X_5, as shown in the DAG.

4.3. Example

To show how the proposed method works with larger variable sets, we consider now a GBN introduced by Castillo and Kjaerulff [3] to assess the damage to reinforced concrete structures of buildings. This network consists of 24 nodes, where the node of interest is X_{24}; the qualitative part is the DAG reproduced in Fig. 2. The quantitative part of the network is given by a Multivariate Normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

The covariance matrix $\boldsymbol{\Sigma}$ was calculated from $\mathbf{D}$, the diagonal matrix of conditional variances v_i and $\mathbf{B}$, the upper triangular matrix with regression coefficients β_{ji}. The values for entropy and mutual information are then calculated. For the target variable, X_{24}, we obtained that X_{24} is $N(0, 18.267)$ with an entropy of 2.8715. The results for the rest of the variables are shown in Table 5.

Table 3
Example of step 6.

Y_{-i}	$h(Y_{-i}\|X_5)$	$h(X_7, Y_{-i}\|X_5)$	$I(X_7; Y_{-i}\|X_5)$
X_1	1.4189	4.2426	0.1079
X_2	1.1635	3.7814	0.3137
X_3	1.6539	4.5451	0.0404
X_4	2.1597	4.1279	0.9633
X_6	2.8805	4.6460	1.1661

Table 4
Example of step 6.

Y_{-i}	$h(Y_{-i}\|X_5,X_4)$	$h(X_7,Y_{-i}\|X_5,X_4)$	$I(X_7;Y_{-i}\|X_5,X_4)$
X_1	1.2883	3.2565	0
X_2	0.7643	2.7325	0
X_3	1.6047	3.5729	0
X_6	1.4189	3.1844	0.2027

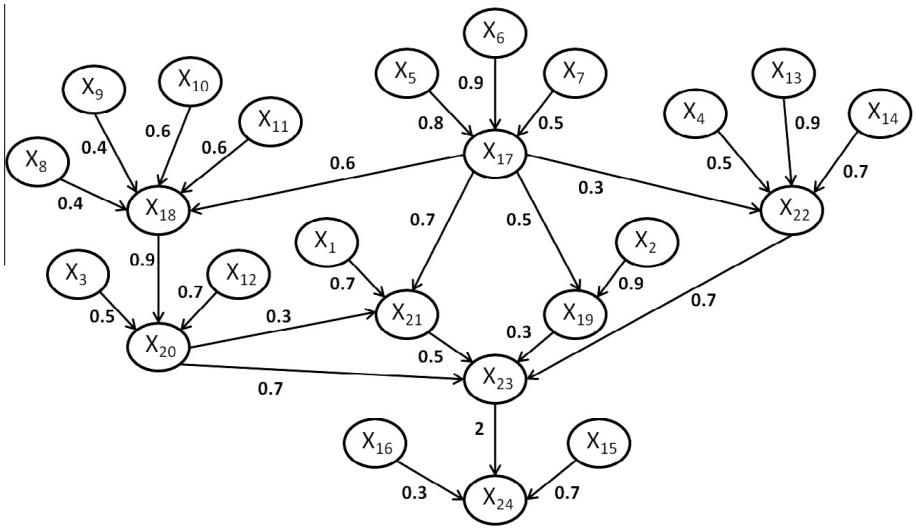

Fig. 2. DAG for Example 4.3.

It can be seen, that the variables that further contribute to entropy reduction of X_{24}, considered as evidence, are first X_{21} and X_{20} and second X_{17} and X_{18}.

In the works of Castillo and Kjaerulff [3] and Main and Navarro [15], the evidence $E = X_1 = 1,\ldots,X_{16} = 1$ is introduced in the network. It is clear from the information given in Table 5, that none of these variables, contribute individually, and significantly to reduce the variance and entropy of X_{24}. However, we have added them to the network to adjust for its effect and further evaluate the importance of other variables.

Table 5
Example of step 2.

Variable Y_{-i}	$h(Y_{-i})$	$h(X_{24},Y_{-i})$	$I(X_{24};Y_{-i})$
X_1	1.4189	4.2768	0.01359
X_2	1.4189	4.2823	0.00804
X_3	1.4189	4.2702	0.02017
X_4	1.4189	4.2768	0.01359
X_5	1.4189	4.1841	0.10629
X_6	1.4189	4.1516	0.13879
X_7	1.4189	4.2515	0.03888
X_8	1.4189	4.2801	0.01035
X_9	1.4189	4.2801	0.01035
X_{10}	1.4189	4.2668	0.02361
X_{11}	1.4189	4.2668	0.02361
X_{12}	1.4189	4.2501	0.04034
X_{13}	1.4189	4.2449	0.04546
X_{14}	1.4189	4.2634	0.02700
X_{15}	1.4189	4.2768	0.01359
X_{16}	1.4189	4.2879	0.00247
X_{17}	1.6843	4.2004	0.35538
X_{18}	1.6699	4.1704	0.37103
X_{19}	1.5245	4.2378	0.15815
X_{20}	1.7847	4.1774	0.47882
X_{21}	1.7387	4.0934	0.51679
X_{22}	1.6851	4.3308	0.22583
X_{23}	2.1622	3.3089	1.72483

Thus, propagating the evidence through the network, a covariance matrix with values close to zero is obtained. In particular, the variable X_{24} is $N(15.42, 0.0019)$. This result has two main consequences. First, the values for entropies corresponding to very small variances are negative, however, the value of the mutual information is positive and makes sense. Second, such precise values for the distribution of X_{24}, remove the need to perform further analysis to reduce uncertainty. However, to complete the case study, the values of entropies and mutual information are shown in Table 6.

5. Extension to sensitivity analysis based on normalized measures

Now we extend our procedure according to the normalized measures presented in Section 3.3. The advantage of the measures used in this new procedure is that values range from 0 to 1 regardless of the unit of measurement of the variables considered, so we can compare the relative importance of evidence variables in a network or even between different networks.

5.1. New procedure for prioritizing evidence

This new procedure is based on the procedure presented in Section 4.1, but we replace the values for mutual information with normalized measures and add a difference-normalized measure. The steps for the analysis are as follows:

1. Calculate the entropy for the target variable $h(X_i)$.
2. Compute the normalized mutual information between the target variable X_i and each non-evidential variable $\mathbf{Y}$ different from X_i, that is, $NI(X_i; \mathbf{Y}_{-\mathbf{i}})$
3. Choose as the evidential variable the X_k from $\mathbf{Y}_{-i}$ that has the highest non-zero normalized mutual information with X_i.
4. Look at the decrease in uncertainty as $h(X_i|\mathbf{E}) - I(X_i; X_k|\mathbf{E}) = h(X_i|\mathbf{E}, X_k)$.
5. Consider that $X_k \in E$.
6. Compute the normalized conditional mutual information between the target variable and the new set of non-evidential variables $\mathbf{Y}_{-\mathbf{i}}$, that is, $NI(X_i; \mathbf{Y}_{-\mathbf{i}}|\mathbf{E})$.
7. Calculate the normalized difference $\Delta I_{X_i Y_{-i} E}$.
8. Go back to 3.

Notes for this procedure are analogous to those presented above.

We do not need new specifications to apply this procedure because the same definitions of entropy for a normal distribution, given in (4) and (5), are adequate for calculations. Next, we show how the procedure works when applied to the same example as before.

5.2. Example

Consider the GBN introduced in Example 2.4, where X_7 is defined as the target variable. First we have to asses the validity of the restrictions required to obtain a real number value $\sqrt{h(X)h(Y)}$. From the covariance matrix Σ, it is evident that all the individual variances are greater than 0.0585. The determinant calculated for Σ is $|\Sigma| = 16 > 1/(2\pi e)^7$, so we can use this new procedure with normalized measures.

We start by computing steps 1 and 2 using (4) and (9), respectively. The initial entropy for X_7 does not change, that is, $h(X_7) = 3.7165$; the normalized mutual information for the non-evidential variables is shown in Table 7.

Note that the results in Table 7 are consistent with those previously obtained: X_6 is the most informative variable for X_7, followed by X_5. For comparative purposes, consider X_5 as an evidential variable, so the results in steps 3 and 4 match the calculations we made in Example 4.2; we have $Var(X_7|X_5) = 20.6$ and differential entropy $h(X_7|X_5) = 2.9316$. Table 8 shows the results for steps 6 and 7 obtained with (10) and (11), respectively.

Table 8 shows that X_6 is still the most informative variable, followed by X_4, as previously. The new result in Table 8 is the calculation of the difference $\Delta I_{X_7 Y_{-i} X_5}$; it can be understood as a measure of the relative loss of influence for each variable on the target after the evidence is incorporated into the network. A negative value indicates that the importance of the variable has increased, and a positive value otherwise. It is interesting to see, for example, that variable X_6 was initially the most

Table 6
Example of step 6.

Variable Y_{-i}	$h(Y_{-i}\|E)$	$h(X_{24}, Y_{-i}\|E)$	$I(X_{24}; Y_{-i}\|E)$
X_{17}	−3.186	−5.069	0.1694
X_{18}	−3.032	−4.948	0.2022
X_{19}	−3.074	−4.859	0.0706
X_{20}	−2.815	−4.847	0.3185
X_{21}	−2.864	−4.856	0.2793
X_{22}	−3.143	−4.976	0.1198
X_{23}	−2.434	−5.620	1.4727

Table 7
Example of step 2.

Y_{-i}	$h(Y_{-i})$	$NI(X_7;Y_{-i})$
X_1	1.4189	0.0089
X_2	1.4989	0.2263
X_3	1.7655	0.0164
X_4	2.3148	0.1907
X_5	2.5702	0.2539
X_6	3.7063	0.5257

Table 8
Example of steps 6 and 7.

Y_{-i}	$h(Y_{-i}\|X_5)$	$NI(X_7;Y_{-i}\|X_5)$	$\Delta I_{X_7Y_{-i}X_5}$
X_1	1.4189	0.0529	−4.9438
X_2	1.1635	0.1698	0.2496
X_3	1.6539	0.0183	−0.1158
X_4	2.1597	0.3828	−1.0073
X_6	2.8805	0.4013	0.2365

important. However, after X_5 is incorporated as evidence, the contribution of X_4 increases and that of X_6 decreases, so after this step we still have a very similar entropy reduction.

Continuing with the procedure, we return to step 3. After again checking the restrictions, we select variable X_4 to be incorporated as evidence. The results, similar to those in Table 4, are shown in Table 9.

Finally, an important advantage of these normalized measures is illustrated by a transformed Bayesian network.

5.3. Example

Consider the same problem introduced in Example 2.4 on the duration for which a machine is working. Now imagine that the machine is made up in part of two new elements that replace two of the former ones. They are connected as shown in the DAG in Fig. 3.

The target is still variable X_7 and the joint probability distribution of $\mathbf{X} = \{X_1, X_2, X_3*, X_4*, X_5, X_6, X_7\}$ (note that variables X_3* and X_4* replace X_3 and X_4) is a multivariate normal distribution with the following parameters given by experts:

$$\boldsymbol{\mu}^* = \begin{pmatrix} 1 \\ 3 \\ 3 \\ 6 \\ 4 \\ 5 \\ 8 \end{pmatrix} \quad \mathbf{B}^* = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \mathbf{D}^* = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

where $\boldsymbol{\mu}^*$ is the n-dimensional mean vector, $\mathbf{D}^*$ is a diagonal matrix with conditional variances v_i^*, $\mathbf{B}^*$ is a strictly upper triangular matrix with regression coefficients β_{ji}^*, and X_j is a parent of X_i.

Similar to the previous examples, we computed $\boldsymbol{\Sigma}^*$ as

$$\boldsymbol{\Sigma}^* = [(\mathbf{I} - \mathbf{B}^*)^{-1}]^T \mathbf{D}^* [(\mathbf{I} - \mathbf{B}^*)^{-1}].$$

Thus, $\mathbf{X}$ has a multivariate normal distribution with the following parameters:

Table 9
Example of step 6.

Y_{-i}	$h(Y_{-i}\|X_5,X_4)$	$NI(X_7;Y_{-i}\|X_5,X_4)$
X_1	1.2882	0
X_2	0.7643	0
X_3	1.6047	0
X_6	1.4189	0.1213

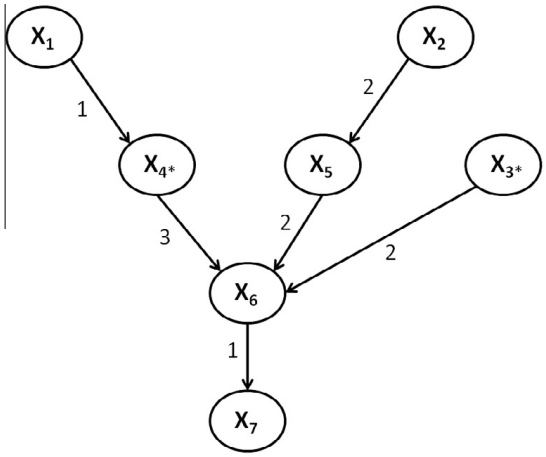

Fig. 3. DAG for Example 5.3.

Table 10
Comparison of the GBN results.

Y_{-i}	$NI_1(X_7; Y_{-i})$	$NI_2(X_7; Y_{-i})$
X_1	0.0089	0.0284
X_2	0.2263	0.0532
X_3	0.0164	
X_{3*}		0.1189
X_4	0.1907	
X_{4*}		0.0224
X_5	0.2539	0.0799
X_6	0.5257	0.6039

$$\boldsymbol{\mu}^* = \begin{pmatrix} 1 \\ 3 \\ 3 \\ 6 \\ 4 \\ 5 \\ 8 \end{pmatrix} \quad \boldsymbol{\Sigma}^* = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 3 & 3 \\ 0 & 1 & 0 & 0 & 2 & 4 & 4 \\ 1 & 0 & 4 & 0 & 0 & 12 & 12 \\ 0 & 0 & 0 & 2 & 0 & 4 & 4 \\ 0 & 2 & 0 & 0 & 7 & 14 & 14 \\ 3 & 4 & 12 & 4 & 14 & 74 & 74 \\ 3 & 4 & 12 & 4 & 14 & 74 & 75 \end{pmatrix}.$$

Let h_j, I_j, and NI_j represent the entropy, mutual information, and normalized mutual information, respectively, of the jth GBN, where $j = 1$ is the problem in Example 2.4 and $j = 2$ is the one introduced here.

Again, we start by checking the restrictions and find that the individual variances are greater than 0.0585 and $|\Sigma^*| = 36 > 1/(2\pi e)^7$.

As we have seen before, the entropy for the target variable of the network presented in Fig. 1 is $h_1(X_7) = 3.7165$, and now we obtain $h_2(X_7) = 3.5777$ for Fig. 3. The results for both GBNs are compared in Table 10.

From Table 10 we can compare the relative influence of the common variables in both networks and draw some conclusions. For example, note that in the presence of X_{3*} and X_{4*}, the importance of X_2 and X_5 decreases significantly. Furthermore, in the second GBN, assuming that X_6 is still not an observable variable, it is better to introduce the variable X_{3*} as evidence rather than X_5.

6. Conclusions

We proposed a new methodology to quantify numerically the contribution of each non-evidential variable to reduction in uncertainty of a target variable. Then, the conditional distribution of the variable of interest is used to select the most informative variables.

The first method proposed is based on mutual information measures and allows us to obtain a prioritization to request additional information. We showed that given a set of evidence variables, if there are some non-evidential variables with mutual information close to zero, they should not be considered as evidence because they do not add any information to the analysis. This is an important contribution, because it reduces the cost of modeling and data collection.

The second method proposed, which is an extension of the first, includes normalized measures of mutual information. Under some restrictions, this procedure, as well as prioritizing unobserved variables, can compare the contribution of the same variable to a different target or to the same target in different Bayesian networks. Therefore, with this method more and better tools are provided for analysis, although the restrictions should be taken into account.

We introduced a Bayesian network as an example to show the results of the first procedure in our sensitivity analysis; we then added a second Bayesian network to make comparisons using the second procedure.

Acknowledgements

This work was supported in part by Grant MTM2008-03282/MTM from Spanish Ministerio de Ciencia e Innovación and Grant 910395 from Universidad Complutense-Banco Santander. We thank the editor and two referees for their constructive comments and suggestions.

References

[1] P. Besson, J. Richiardi, C. Bourdin, L. Bringoux, D.R. Mestre, J.L. Vercher, Bayesian networks and information theory for audio-visual perception modeling, Biolog. Cybernet. 103 (2010) 213–226.
[2] E. Castillo, J.M. Gutiérrez, A.S. Hadi, Expert Systems and Probabilistic Network Models, Springer Verlag, New York, 1997.
[3] E. Castillo, U. Kjaerulff, Sensitivity analysis in Gaussian Bayesian networks using a symbolic-numerical technique, Reliab. Eng. Syst. Safety 79 (2003) 139–148.
[4] H. Chan, A. Darwiche, Sensitivity analysis in Bayesian networks: from single to multiple parameters, in: Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence (UAI), AUAI Press, Arlington, VA, 2004, pp. 67–75.
[5] T. Cover, J. Thomas, Elements of Information Theory, Wiley, Chichester, 1991.
[6] R.G. Cowell, A.P. Dawid, S.L. Lauritzen, D.J. Spiegelhalter, Probabilistic Networks and Expert Systems, Springer, Barcelona, 1999.
[7] M.A. Gómez-Villegas, P. Main, R. Susi, Sensitivity analysis in Gaussian Bayesian networks using a divergence measure, Commun. Statist.-Theory Meth. 36 (2007) 523–539.
[8] M.A. Gómez-Villegas, P. Main, R. Susi, Extreme inaccuracies in Gaussian Bayesian networks, J. Multivar. Anal. (9) (2008) 1929–1940.
[9] M.A. Gómez-Villegas, P. Main, H. Navarro, R. Susi, Evaluating the difference between graph structures in Gaussian Bayesian networks, Exp. Syst. Appl. 38 (2011) 12409–12414.
[10] M.A. Gómez-Villegas, P. Main, R. Susi, The effect of block parameter perturbations in Gaussian Bayesian networks: sensitivity and robustness, Inform. Sci. 222 (2013) 439–458.
[11] S. Ihara, Information Theory for Continuous Systems, World Scientific, Singapore, 1993.
[12] U. Kjaerulff, A. Madsen, Probabilistics Networks for Practitioners, A Guide to Construction and Analysis of Bayesian Networks and Influence Diagrams, Springer, New York, 2007.
[13] L. Kostal, P. Lansky, O. Pokora, Measures of statistical dispersion based on Shannon and Fisher information concepts, Inform. Sci. 235 (2013) 214–223.
[14] K.B. Laskey, Sensitivity analysis for probability assessments in Bayesian networks, IEEE Trans. Syst. Man Cybernet. 25 (1995) 901–909.
[15] P. Main, H. Navarro, Analyzing the effect of introducing a kurtosis parameter in Gaussian Bayesian networks, Reliab. Eng. Syst. Safety 94 (2009) 922–926.
[16] R. Malhas, Z. Al Aghbari, Using sensitivity of a Bayesian network to discover interesting patterns, Int. Conf. Comp. Syst. Appl. 1–3 (2008) 196–205.
[17] J. Richiari, Probabilistics Models for Multi-Classifier Biometric Authentication Using Quality Measures, These no. 3954, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2007.
[18] C.E. Shannon, A mathematical theory of communication, The Bell Syst. Tech. J. 27 (1948). 379–423 and 623–656.
[19] A. Strehl, J. Ghosh, Cluster ensembles – a knowledge reuse framework for combining multiple partitions, J. Mach. Learn. Res. 3 (2002) 583–617.