KNOWLEDGE PRESENTATION AND REASONING WITH LOGLINEAR MODELS

Veska Noncheva, Nuno Marques

Abstract: Our approach for knowledge presentation is based on the idea of expert system shell. At first we will build a graph shell of both possible dependencies and possible actions. Then, reasoning by means of Loglinear models, we will activate some nodes and some directed links. In this way a Bayesian network and networks presenting loglinear models are generated.

Keywords: computer oriented statistics, knowledge discovery, learning Bayesian networks, automatic analysis of multivariate categorical data sets.

Introduction

Our main aim is to link statistical theory to some networks in order to enrich computer's reasoning capability. In this paper we will define a new data structure called LLN and offer an algorithm for learning a Bayesian network by using knowledge obtained from categorical data set. This algorithm finds out the loglinear model, describing data, a presentation of this model by a LLN and a Bayesian network describing the relationship among the variables of interest.

Categorical data is most often modelled using loglinear models. In this paper we provide a principled foundation for reasoning with Loglinear models. We will discuss loglinear models that describe association patterns among two and three categorical variables.

Graphs are natural data structures for digital computers. We will provide a framework presentable by a direct graph for modelling categorical data and interpreting the results. Different directed graphs can represent the same dependence structure for the set of associated variables. Consequently, if the links have no causal interpretations, we will obtain a set of equivalent graphical structures.

Basic concepts and definitions

Suppose we have a set of *n*, *n>1* possibly related categorical variables $V=\{X_1, X_2, \ldots, X_n\}$. This set can be represented pictorially by a set of nodes – one node for each variable of *V*. These nodes can be connected by arcs. The dependency structure could be presented by a Bayesian network. The language of Bayesian networks is described in [Castillo,Gutierrez,Hadi,1997].

Suppose also that we have a set of *k* actions $A=\{A_1, A_2, \ldots, A_k\}$, that could be applied on some of these *n* nodes. The objective of an action is building a loglinear model describing categorical data available. These actions can be applied to some variables from *X*. The application of an action to objects is visualised by directed arcs.

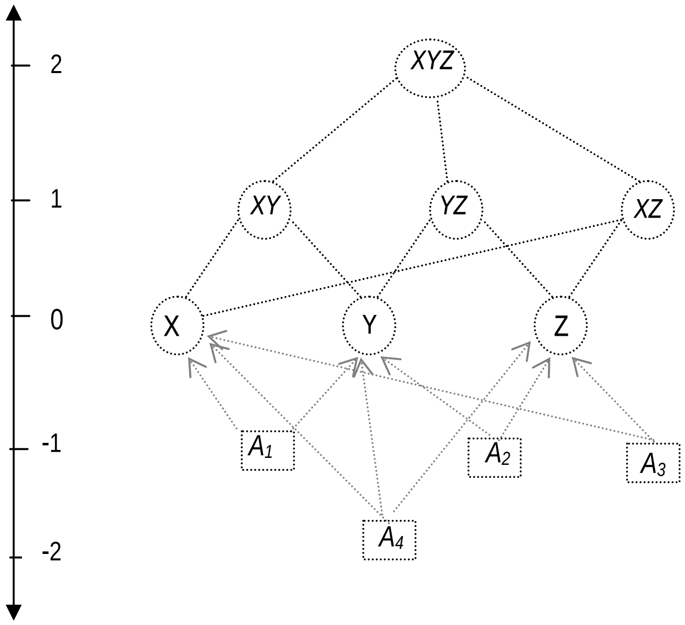


Figure 1. Graph shell.

Let *V*={*X*,*Y*,*Z*}. Actions and nodes are placed on different levels in the graph shell (see *Figure 1*). On level 0 the nodes, presenting the associated categorical variables *X*, *Y*, and *Z* are placed. On level 1 – nodes, presenting new variables *XY*, *YZ*, and *XZ*. The variable *XZ* could be composed of the *IK* combinations of levels of *X* and *Z* or it could present partial association only. On level 2 – nodes, presenting a new variable *XYZ*. The variable *XYZ* could be composed of the *IJK* combinations of levels of *X*, *Y* and *Z* or it could present three-factor interaction only. On level -l – actions applicable to nodes from level 0 and giving results in nodes from level *l*, *l*=1,2. On a sub-network in the level interval [-1, 1] we can reason about the relationship between two variables. On a sub-network in the level interval [-2, 2] we can reason about the relationship between three variables. In the general case on a sub-network in the level interval [-*l*, *l*] we could reason about the relationship between *l+1*, *l=1,2*,… variables.

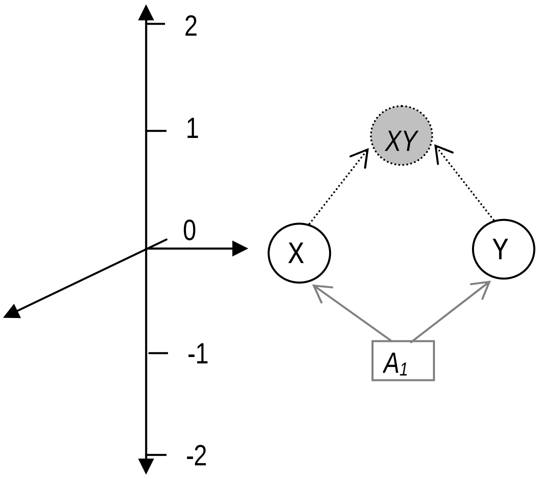
Loglinear models and their interpretations

In two dimensions only two distinct loglinear models can occur, in general. Either the two variables are independent, or they are associated. The loglinear models, presented bellow, present these two cases.

Given a two-dimensional contingency table the models assume a sample of size *n* distributed over *IJ* cells. Under multinomial sampling, the probability that an observation will fall into cell ij is π_{ij} for all *i=1*, … , *I*, *j=1*, …, *J*. The expected value m_{ij} is $n\pi_{ij}$. The expected value for the observed counts in a contingency table could be estimated by independence loglinear model $\log(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y$, where the parameters $\{\lambda_i^X\}$ and $\{\lambda_j^Y\}$ satisfy $\sum\lambda_i^X = \sum\lambda_j^Y = 0$, or by dependence loglinear model $\log(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$, where the parameters $\{\lambda_i^X\}$, $\{\lambda_j^Y\}$ and $\{\lambda_k^{XY}\}$ satisfy $\sum\lambda_i^X = \sum\lambda_j^Y = \sum\lambda_K^{XY} = 0.$

These two mathematical models are graphically presented in *Figure 2* and *Figure 3*. The case of independence is shown in Figure 2. This properly indicates the presence of mutual independence. The case of interaction is depicted in Figure 3.

Given a three-dimensional contingency table the model assumes a sample of size n distributed over *IJK=N* cells. Under multinomial sampling, the probability that an observation will fall into cell *ijk* is π_{ijk} for all *i=1*,… ,*I*, *j=1*,…,*J, k=1*,…,*K*. The expected value m_{ijk} is $n\pi_{ijk}$. The model of mutual independence for a three-dimensional contingency table is $\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$. The parameters $\{\lambda_i^X\}$, $\{\lambda_j^Y\}$ and $\{\lambda_k^Z\}$ satisfy $\sum\lambda_i^X = \sum\lambda_j^Y = \sum\lambda_K^Z = 0.$



$$\log(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y$$

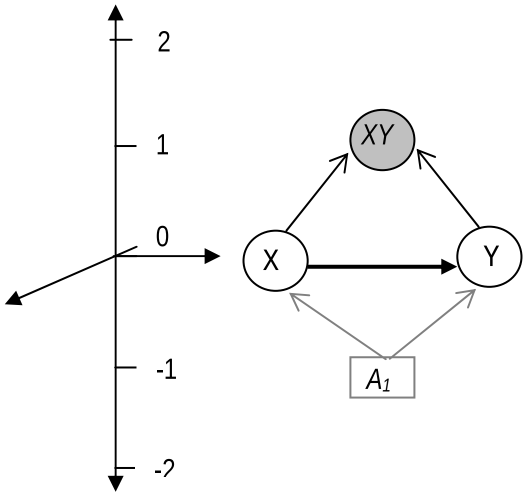
$$p(x, y) = p(x)p(y)$$

Figure 2. Graphical presentation of the independence loglinear model

Interactions between two or all three variables can be modelled by including the additional terms $\{\lambda_{ij}^{XY}\}$, $\{\lambda_{jk}^{YZ}\}$, $\{\lambda_{jk}^{YZ}\}$ and $\{\lambda_{ijk}^{XYZ}\}$ with zero sums over the parameters. The interaction structures are following:

mutual independence, partial independence, conditional independence, no three-way interaction, and three-way interaction.

Mutual independence of the three variables is equivalent to $\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$. The loglinear model of mutual independence is $\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$. All variables are pair-wise (mutually) independent. Thus, only the main effects λ_i^X, λ_j^Y, and λ_k^Z appear in the model. Each pair of variables is also conditionally independent and marginally independent.



$$\log(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

$$p(x, y) = p(x)p(y \mid x)$$

Figure 3. Graphical presentation of dependency loglinear model for two-dimensional contingency table

Partial independence means presence of λ_i^X, λ_j^Y, λ_k^Z, and additional presence of one $\lambda^{AB}, A, B \in \{X, Y, Z\}, A \neq B$. Suppose that the variable X is partially independent of *Y* and *Z*. In accordance with the definition the variable *X* is partially independent of *Y* and *Z*, if $\pi_{ijk} = \pi_{i++}\pi_{+jk}$ for all *i,j,k*. The composite variable *YZ*, which has *JK* different levels combinations of *Y* and *Z*, is mutually independent of X. The corresponding loglinear model is $\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}$. The variable X is jointly independent of Y and *Z*. The variables *X* and *Y* are conditionally independent given *Z*, and *X* and *Z* are conditionally independent given Y. The variables *Y* and *Z* are conditionally dependent. The variable *X* is also independent of *Y* and *Z* in the *X-Y* and *X-Z* marginal tables.There are three models of partial independence.

Conditional independence of *X* and *Y* given *Z* means $\pi_{ij|k} = \pi_{i+|k}\pi_{+j|k}$ for all *i,j,k.* The loglinear model is $\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$. The variable *Z* is conditionally dependent with both *X* and *Y*. The variables *X* and *Y* may be marginally dependent, even though they are conditionally independent. There are three models of conditional independence.

The loglinear model of no three-way interaction is $\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$. All three pairs of variables are conditionally dependent. Although every variable interacts with each other variable, there is no interaction between all three variables. No pair of variables is conditionally independent. When there is an absence of three-factor interaction, the association between two variables is identical at each level of the third variable. The cell probabilities have form $\pi_{ijk} = \psi_{ij}\phi_{jk}\varphi_{ik}$.

All parameters are included in the three-way interactions model. The loglinear model of three-way interaction is $\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$. This is the model where every possible interaction is included. The only interpretation of this model is the fact that apparently all other models failed to represent the data in a suitable way.

The most common problem in loglinear modelling is to find the most suitable model. The better model describing data includes as few interaction terms as possible and declares as much of the deviation from mutual independence as possible. We usually use χ^2 and G^2 statistics to judge the adequacy of a loglinear model. Roy and Mitra gave Person-type statistics for large-sample tests.

We have illustrated ideas using the two and three-variable case. Loglinear models for four-way tables are more complex than for three-way tables, because of the variety of potential partial association, three-factor interaction patterns and four-factor interaction pattern.

We can readily extend the framework to arbitrary multi-way tables.

When the number of dimensions increases, both the number of possible interaction patterns and the number of cells dramatically increase. Frameworks for multy-way tables will be much more complex.

The question whether this interaction is statistically significant or not remains unrevealed, as long as the number of underlying observations is unknown. A way for presenting our belief in a loglinear model is presented in [Noncheva,Marques,2002].

Each of these models are visualised below. The topological shape of each model-type is not invariant against a permutation of the variables.

A visual method based on mosaic plots for interpreting and modelling categorical data is considered in [Theus,Lauer,1999]. Paik suggested circle diagrams for presenting results from three-way tables [Paik,1985].

Types of independence in a three-way cross-classification of X, Y, and Z

A relationship is defined by the join distribution of the associated random variables. The joint distribution determines the marginal and conditional distributions. Simplification occurs in a joint distribution when the component random variables are statistically independent. We will discuss four types of independence for categorical variables.

The three variables are mutually independent when $\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$ for all i, j, and k. On a log scale, mutual independence is the loglinear model $\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$.

Variable Y is jointly independent of X and Z when $\pi_{ijk} = \pi_{i+k}\pi_{+j+}$ for all i, j, and k. This is ordinary two-way independence for X and the new variable YZ composed on the JK combinations of levels of Y and Z. The loglinear model is $\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}$. Mutual independence implies joint independence of any one variable from the others.

If X and Y are independent in the partial table for the kth category of Z, then X and Y are said to be conditionally independent at level k of Z. If $\{\pi_{ij|k} = \pi_{ijk} / \pi_{++k}, i = 1,..., I, j = 1,..., J\}$ denotes the joint distribution of X and Y at level k of Z, then conditional independence at level k of Z is $\pi_{ij|k} = \pi_{i+|k}\pi_{+j|k}$ for all i and j. X and Y are conditionally independent given Z when they are conditionally independent at every level of Z, or equivalently, when $\pi_{ijk} = \pi_{i+k}\pi_{+jk} / \pi_{++k}$ for all i, j, and k. Conditional independence of X and Y is the loglinear model $\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$. If *Y* is jointly independent of *X* and *Z*, then *X* and *Y* are conditional independent.

We say that X and Y exhibit marginal independence if $\pi_{ij+} = \pi_{i++}\pi_{+j+}$. Joint independence of Y from *X* and *Z* (or of *X* from *Y* and *Z*) implies *X* and *Y* are marginally independent.

The relationships among the four types of independence are summarized in Figure 5. The basic question is how these types of independence could be presented graphically within the framework we are building. A solution of this task is given in Figure 6 and Figure 7.

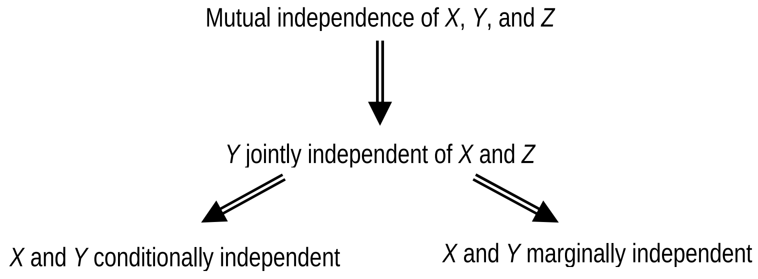

Mutual independence of *X*, *Y*, and *Z*
Y jointly independent of *X* and *Z*
X and *Y* conditionally independent
X and *Y* marginally independent

Figure 5. Relationships among types of X-Y independence.

Decision network for reasoning with Loglinear Models

Decision networks are graphic structures, that represent probability relations and information flows ([Shachter,86], [Shachter,88]). We introduce a kind of decision network for reasoning with loglinear models called loglinear network (LLN). A variable from the loglinear model is presented as a node in the graph of the loglinear network.

Definition: A loglinear network (LLN) comprises of the following set of items: (X,(A,I), (X',P), A*, u, δ). It is within this particular set where they are shown as below:

X is the directed graph of the (in)dependency among all variables X_i, *i=1,…,n*, in the model. It is called LLN graph.

(A,I) is recognized as a directed graph of basic operations, where A={A_i, *i=1,…,k*} is the set of these basic operations, I={(A_i,X'_m), i=1,…,k; m=1,…,*n*} is the set of directed information arcs and X' $\subset$ X is the set of the associate variables that we are interested in. The objective of a basic operation is building a loglinear model describing categorical data available.

(X',P), is a Bayesian network, built for. X' $\subset$ X is the set of the associate variables that we are interested in.

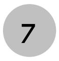
P is the set of conditional probabilities.

A* is the decision set. A* is the set of adequate loglinear models.

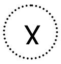
u: X'$\to R$ is the utility function, where R represents the real numbers set. Usually χ^2 and G^2 statistics are used as utility functions.

$\delta:$ A$\to$ A* is the decision rule. Usually the decision rule is based on χ^2 and G^2 statistics.

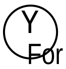
We distinguish between explanatory and independent variables. Independent variables are classified as explanatory if they are involved in the model under study otherwise they remain independent variables that could be involved in a model. Both the explanatory variables and the response variable from the loglinear model are presented as nodes in the LLP graph. The alphabet of the LLP graph language is as follows:



Z is *the response variable*

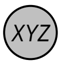


X is an *independent variable*. It is equivalent to setting the model parameter equal to zero in the general model.



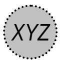
Y is an *explanatory variable*.

For example,



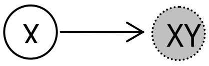
Response variable of the three-way interaction model

$$\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$$

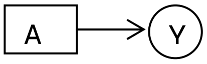


Response variable of the no three-way interaction model

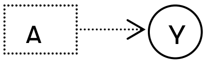
$$\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$



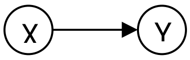
The variable *X* is included in the loglinear model with response *XY* or *X* is an explanatory variable in the loglinear model with response *XY*.



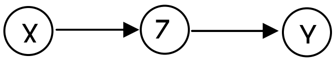
The action *A* is applied to the variable *Y*.



The action *A* could be applied to the variable *Y*.



Variable *Y* depends on variable *X*.



Variables X and *Y* are independent given variable *Z*.

Inference Algorithm

Once the initial knowledge has been presented, one of the most important tasks of a system is to draw conclusions when new information is observed. An algorithm of drawing conclusions about both dependency and probabilistic structure of a Bayessian network is roughed out below.

Inference algorithm

Input: A set of *n* random variables and its graph shell and an *n*-way contingency table (*n*=2,3,…).

Output: A Bayesian network over the set of variables.

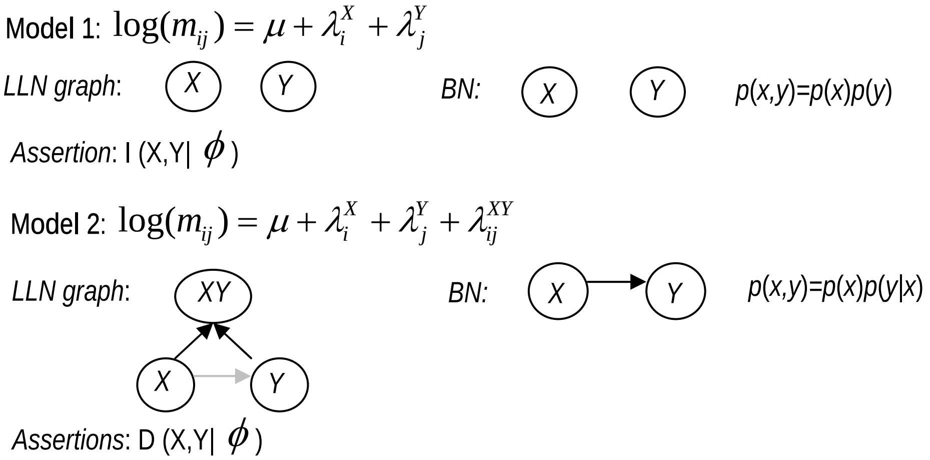
Steps: Generate different tasks for building loglinear models starting with most restricted ones (from mutual independence model to saturated model).

1. Build a restricted loglinear model. Go to step 2.

2. Check for adequacy. If the loglinear model is adequate then activate the appropriate arcs in the graph shell and update the probability distributions of the variables of interests according to the newly available information. Go to step 3. If the loglinear model is not adequate then go to step 1.

3.End.

For example, possible results could be the graphs of Bayesian networks presented in Figure 6 and Figure 7.



Model 1: $\log(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y$

LLN graph: *BN:* *p*(*x,y*)=*p*(*x*)*p*(*y*)

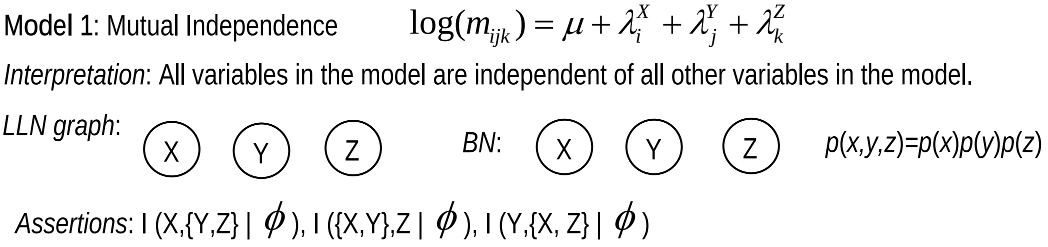
Assertion: I (X,Y| ϕ)

Model 2: $\log(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$

LLN graph: *BN:* *p*(*x,y*)=*p*(*x*)*p*(*y*|*x*)

Assertions: D (X,Y| ϕ)

Figure 6. LLN frameworks for two-way contingency tables.



Model 1: Mutual Independence $\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$

Interpretation: All variables in the model are independent of all other variables in the model.

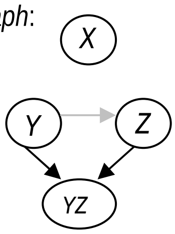
LLN graph: BN: *p*(*x,y,z*)=*p*(*x*)*p*(*y*)*p*(*z*)

Assertions: I (X,{Y,Z} | ϕ), I ({X,Y},Z | ϕ), I (Y,{X, Z} | ϕ)

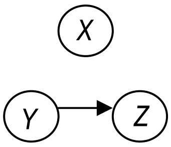
Model 2: Partial Independence $\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}$

Interpretation: One factor is independent of the other factors.

LLN graph:



BN:



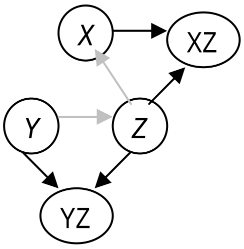
p(*x*,*y*,*z*)=*p*(*x*)*p*(*y*,*z*)=*p*(*x*)*p*(*y*)*p*(*z*|*y*)

Assertions: D (Y,Z| ϕ), I (X,{Y,Z}| ϕ)

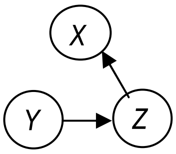
Model 3: Conditional Independence $\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$

Interpretation: We make decision for independence of two factors and there is a relationship between both of those factors and the third factor.

LLN graph:



BN:



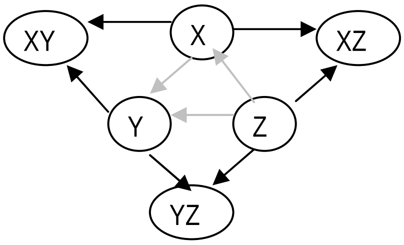
p(*x*,*y*,*z*)=*p*(y)*p*(*z*|y)*p*(*x*|*z*)

Assertions: D (X,Z| ϕ), D (Y,Z| ϕ), I (X, Y | Z)

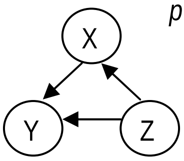
Model 4: No Three-Way Interaction $\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$

Interpretation: There is an association between X and Y that is the same for each level of Z; Y and Z have an association that is the same for each level of X, and X and Z have a relationship that is the same for each level of Y.

LLN graph:



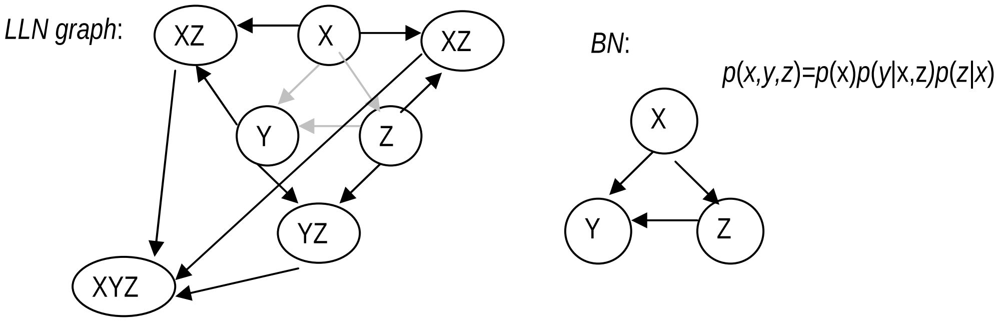
BN:



p(*x*,*y*,*z*)=*p*(z)*p*(*y*|x,z)*p*(*x*|*z*)

Assertions: D (X,Z| Y), D (X,Y| Z), D (Y,Z| X)

Model 5: Three-Way Interaction $\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$



p(*x*,*y*,*z*)=*p*(x)*p*(*y*|x,*z*)*p*(*z*|*x*)

Assertions: D (X,Z| ϕ), D (X,Y| ϕ), D (Y,Z| ϕ), D (X,{Y,Z}| ϕ), D (Y, {X,Z}| ϕ), D ({X,Y},Z | ϕ)

Figure 7. LLN frameworks for three-way contingency tables

Conclusion

We have introduced a new graphical representation of loglinear models. We have presented a framework for reasoning with loglinear models. In our framework both dependences between nodes and actions on nodes enjoy a graphical representation. Loglinear networks are graphs with three types of nodes and two types of arcs, representing dependencies and actions. Using this framework we construct a Bayesian network of associate random variables.

Bibliography

[Castillo,Gutierrez,Hadi,1997] Castillo E., Gutierrez J.M., and A.S. Hadi, Expert Systems and Probabilistic Network models, Springer, 1997.

[Noncheva, Marques,2002] Noncheva V., N. Marques. Agent's Belief: A Stochastic Approach, Proc. of the 14th Belgian-Dutch Conference on Artificial Intelligence BNAIC 2002, pp.227-234.

[Paik,1985] Paik, M. A graphic representation of a three-way contingency table: Simpson's paradox and correlation. Amer. Statist., 1985, 39: 53-54.

[Shachter,86] Shachter R.D. (1986). Evaluating influence diagrams, Operations Research, No.34, pp. 871-882.

[Shachter,88] Shachter R.D. (1988). Probabilistic Inference and Influence Diagrams, Operations Research, 36, 591-604.

[Theus, Lauer,1999] Theus Martin, Stephan R. W. Lauer Visualizing Loglinear Models, Journal of Computational and Graphical Statistics, 1999, Volume 8, Number 3, pp. 396-412.

Author information

Veska Noncheva – Faculty of Mathematics and Informatics, University of Plovdiv, 24 Tzar Assen St. 4000 Plovdiv, Bulgaria; e-mail: wesnon@pu.acad.bg

Nuno Marques Departamento de Informática, FCT, Universidade Nova De Lisboa, Quinta da Torre, 2829 - 516 Caparica, Portugal, e-mail: nmm@di.fct.unl.pt