An approach for knowledge acquisition from a survey data by conducting Bayesian network modeling, adopting the robust coplot method

Derya Ersel and Yasemin Kayhan Atılgan

Department of Statistics, Hacettepe University, Ankara, Turkey

ABSTRACT

This study proposes a methodological approach for extracting useful knowledge from survey data by performing Bayesian network (BN) modeling and adopting the robust coplot analysis results as prior knowledge about association patterns hidden in the data. By addressing the issue of BN construction when the expert knowledge is limited/not available, this proposed approach facilitates the modeling of large data sets describing numerously observed and latent variables. By answering the question of which node(s)/link(s) should be retained or discarded from a BN, we aim to determine a compact model of variables while considering the desired properties of data. The proposed method steps are explained on real data extracted from Turkey Demographic and Health Survey. First, a BN structure is created, which is based solely on the judgment of the analyst. Then the coplot results are employed to update the BN structure and the model parameters are updated using the updated structure and data. Loss scores of the BNs are used to ensure the success of the updated BN that inherits knowledge from coplot.

1. Introduction

In social studies, extracting useful knowledge from a survey is an essential process. Typically, the data questionnaires and interviews include a complex and large number of variables. A researcher is required to build a model which provides a compact and easy to interpret knowledge representation while studying on such datasets. In literature, various statistical methods have been used to extract useful knowledge from survey data. Multivariate analysis techniques such as principal component analysis [25], factor analysis [9,51], regression analysis [20], and some machine learning techniques [18,49] have been used to examine questionnaire data. One of the popular methods among them is logistic regression which is used in case the dependent variable is categorical. The most common problem encountered in this method is that the high number of exploratory variables are involved in the model [52,54]. So, several variable selection approaches based on the forward, stepwise, backward procedures have been recommended for logistic regression.
in the literature. Alternative methods to variable selection apart from these procedures have also been proposed. Pacheco et al. [38] recommended a Tabu search method and showed that it obtained better results than classical variable selection methods. Zhang et al. [54] used genetic algorithms which are heuristic optimization approaches to select variables in logistic regression. However, many studies in the literature show the discrepancy between variable selection methods in logistic regression [52].

Another technique that is widely used in survey studies is structural equation modeling (SEM). This method combines two statistical methods, factor analysis, and multiple regression analysis, and it analyzes the structural relationship between measured variables and latent constructs. When the sample size is not sufficiently large, various variable selection techniques are used in SEM. Zhang et al. [53] proposed an approach for variable selection in SEM in the case that latent variables are linearly regressed on themselves. Fan and Zhong [15] studied on variable selection if endogeneity exists in the model. Jacobucci et al. [26] recommended a regularization method to select variables when observations are limited when the number of parameters is large. The main reason behind the appealing the interest of the researchers in SEM is its ability to easily address large numbers of variables efficiently. However, bibliography or expert knowledge is required to develop the models due to the confirmative nature of the mentioned technique [11,21].

Recently, BN is being used as a powerful tool to acquire accurate and valuable knowledge from questionnaire data. Sebatiani and Ramoni [41] have constructed a BN to analyze a dataset acquired from a British general household survey. Salini and Kenett [40] have studied BNs to analyze customer satisfaction data. Anderson et al. [1] have demonstrated the advantage of BNs by comparing specific statistical methods on the data of transportation service satisfaction. Blodgett and Anderson [6] have presented a BN model to analyze consumer complaints. The use of BNs for survey data analysis offers many advantages for the users. Those are: BNs do not have a restriction on data type as both qualitative and quantitative variables can be analyzed simultaneously; there is no preliminary hypothesis about the nature of the modeled relationships, and they do not have model assumptions. Additionally, BNs can regulate discrepancies in learning data; BNs allow understanding causal relationships between variables, achieving consistent results that are based on probability theory; BNs can make reliable inference in case of any missing data [4,23]. Since BNs provide modeling of expert knowledge as well as information from the data efficiently, in some studies in literature, the success of BN has been verified to be better than other approaches [14,19].

BNs can be constructed directly from the knowledge of domain experts without the time-consuming learning processes. However, it is not always possible to reach the appropriate expert opinion. Thus, it is necessary for constructing BNs to examine the dataset in order to reveal essential associations and hidden structures [23]. This issue is known as the 'learning problem in BNs'. Learning is a very active area of research, and several methods are being developed in this area [4,37]. In literature, the learning problem in BNs is examined under two categories: structural and parameter learning. In structural learning process, the topology of the BN is identified, variables and links in the model are determined. In this study, we address the problem of structural learning of BNs. Creating a model structure from the data is one of the most challenging tasks in building BNs. This task is complicated as it requires searching for the most suitable structure among all
possible structures to be created from the data. To tackle this issue, three classes of methods have been proposed in the literature [5]:

(i) Constraint-based methods identify the best BN structure, which encodes a set of conditional dependence and independence relations between variables. Tests are conducted on conditional independence in the data, and a search is performed that creates a BN consistent with observed dependencies and independencies.

(ii) Score-based methods define a score that measures the fitness of each candidate structure with the data and search for a structure that maximizes this score. In literature, various functions based on entropy, Bayesian scoring, and description length are proposed to create these scores.

(iii) Hybrid methods combine both score-based and constraint-based methods in the construction of BNs [43].

Apart from these methods, other statistical techniques can be used to learn structure of BNs. The obtained results from other statistical methods applied to same dataset can be embedded into the BN model as prior knowledge, specifically in the absence of expert knowledge. For example, Yang and Li [50] benefited from the association analysis results in their work to identify a BN model for evaluating individual bank credits. Sornil and Poonvutthikul [43] also used association analysis, backpropagation neural networks, and traditional data imputation techniques to create BNs. In this study, we present a methodological approach that combines the BN and robust coplot techniques (in the rest of the study, the robust coplot method is briefly named as the coplot method). This study aims to incorporate coplot results as a prior knowledge to determine which variables/links in BN should be retained or discarded when the expert knowledge is limited or non-existent. We present a two-step process to construct a desirable BN representation, which is understandable, easy to interpret, and does not include extra information that has no use. In the first step, the coplot graph of the dataset is examined, important relationship structures, and association patterns hidden in the data are extracted. In the second step, by utilizing the relationship structures extracted from the first step, the critical nodes and links which should be included in the BN structure are decided. In this step, to measure the quality of the BN structure generated by coplot results, loss scores are used.

Graphical display techniques like the coplot method are essential for studying the multidimensional dataset. They provide direct insight into the nature of the data by offering a comprehensive picture of it. When these relations are presented in the form of graphs, they are easier to comprehend and interpret. Though many visualization techniques that are used for multidimensional datasets are defined with a set of new composite variables gathered from the original variables, the variables in coplot are directly taken from the original data. Thanks to this future, the difficulties of interpreting and identifying composite variables are not encountered in the coplot method. The coplot method aims to provide information for simultaneously understanding the similarities between the observations, correlations among the variables, and relations between the observations and variables with a single graph [2].

The procedure of the proposed approach is applied to a dataset extracted from the Turkish Demographic and Health Survey (TDHS), which includes information about the causes and the consequences of domestic violence. We created three latent variables from
the original variables in the survey dataset to define the possible dimension of domestic violence. We examined the relationship between 11 demographic variables, which are selected from the same data and these latent variables by the coplot graph. Based on these relationships, we decided on the BN structure that best represents the dataset.

This paper is structured as follows: Section 2 describes the formulation of latent variables, BNs, and Robust coplot analysis. Section 3 presents a detailed analysis of the methodological process proposed in this paper and an application of this approach on a small example. Section 4 shows how to extract useful knowledge from survey data by performing the recommended approach. The results obtained are explained and discussed in this section. Finally, conclusions are reported in Section 5.

2. Methodology

This section describes all the techniques that have been applied in this study. First, the formulation for latent variables is given then BNs and robust coplot analysis are briefly defined.

2.1. Formulation for latent variables

According to the conducted surveys in social sciences, many characteristics, such as support of a political party, economic progress, racial discrimination, alienation, irregularity, domestic violence, or religious loyalty can not be directly observed. However, we can use some observable characteristics which are indicative of them. For example, even though religious loyalty is not directly observable, the frequency of going to worship and prayer can be measured directly. It is usually thought that the observed variables are affected by latent or unobserved variables. So, the latent variables are defined by taking the relations between observed variables into account. Several statistical methods have been proposed to analyze these relations and define latent variables. One of these methods is the Latent Class Analysis (LCA). In a general sense, LCA is used to identify multi-level discrete latent variables using two or more cross classified observed variables [35].

Many previous studies on latent variables are based on factor analysis. In factor analysis, the continuous latent variable is analyzed by using continuous or binary observed variables. If observed variables are categorical, LCA is adopted instead of factor analysis [24].

LCA was introduced by Lazarsfeld in 1950. Lazarsfeld used this method as a tool for clustering binary observed variables. In 1974, Goodman developed an algorithm to find the Maximum Likelihood (ML) estimates of model parameters. In this study, Lazarsfeld’s approach, which is widely used in the literature, is adopted. Additionally, the method proposed by Goodman is used to find the ML estimates of conditional and latent class probabilities. The Expectation Maximization (EM) algorithm is performed to obtain the ML estimates [22]. poLCA software package, which is implemented in R, is used for the estimation of latent classes [32].

The basic equations of the Latent Class (LC) model for three observed variables \((A, B, C)\) with two or more levels and a latent variable \((X)\) with \(T\) levels can be given as follows.

\[
\pi_{ijk}^{ABC} = \sum_{t=1}^{T} \pi_{ijkt}^{ABCX}
\]  

(1)
The term $\pi_{ijkt}^{ABCX}$ is defined in Equation (2)

$$\pi_{ijkt}^{ABCX} = \pi_t^X \pi_{ijkt}^{ABCX} = \pi_t^X \pi_{ii}^AX \pi_{jj}^BX \pi_{kk}^CX$$

(2)

Here, $\pi_t^X$ is the probability of observation at $t$th level of $X$, which is called latent class probability. $\pi_{ii}^AX$ is the conditional probability for being in the level $i$ of $A$ given the observation is in the level $t$ of latent variable $X$. Similarly, $\pi_{jj}^BX$ is the conditional probability of being in the level $j$ of $B$ given the observation is in the level $t$ of latent variable $X$, $\pi_{kk}^CX$ is the conditional probability of being in the level $k$ of $C$ given the observation is in the level $t$ of latent variable $X$. The $\pi_{ijkt}^{ABCX}$ is the conditional probability for being in level $(i, j, k)$ of joint ABC variable given the observation is in the level $t$ of latent variable $X$ [22].

### 2.2. Bayesian networks

A BN represents a multivariate probability distribution of $p$ random variables in the $X_{n \times p}$ data matrix. The network consists of two components, $G$ and $\Theta$. The first component, $G$, is a graphical structure in which nodes represent $X_1, X_2, \ldots, X_p$ random variables, and the links between nodes which show the direct dependencies between these variables. The second component of BNs, $\Theta$, represents the set of parameters in the network. These parameters are the conditional probability distributions for each random variable in the BN. The conditional probability distribution for a random variable $X_j$ is defined as

$$\theta_{X_j|\pi_j} = P(X_j | \pi_j)$$

where $\pi_j$ is the set of parents of $X_j$ in $G$. If a random variable has no parent in a given BN, then the conditional probability distribution for this random variable corresponds with the marginal probability distribution. In BNs, if a variable is observed, then ordering representation a variable is the evidence node; otherwise, it is hidden or latent node.

BNs satisfy ‘local Markov property’, which indicates that a variable (node) is conditionally independent of its non-descendants given its parents. Using this property, the multivariate probability distribution of $X_1, X_2, \ldots, X_p$ is obtained by ‘chain rule’ given in Equation (3) [7,37].

$$P(X_1, \ldots, X_p) = \prod_{j=1}^{p} P(X_j | \pi_j) = \prod_{j=1}^{p} \theta_{X_j|\pi_j}$$

(3)

Let us give an example for understanding BNs and their important features better [4]. In this example, the incidents that cause back injury are examined. Back injury is indicated with the random variable ‘Back (B)’. Back injury can cause backache and this is represented by ‘Ache (A)’ random variable. A reason for back injury might be doing sports activity wrong. This incident is represented by the ‘Sport (S)’ random variable. Another reason for back injury might be the fact that the person’s chair at the workplace is uncomfortable. The situation of an uncomfortable chair at the workplace is shown as the random variable ‘Chair (C)’. If the cause of the back injury is an uncomfortable chair, then it can also be investigated that whether this person’s colleagues have a similar problem or not. In this case, the corresponding variable can be taken as ‘Worker (W)’. All variables in this problem are binary as ‘True (T)’ and ‘False (F)’. The created BN for this problem [4] is given in Figure 1.
In Figure 1, the Conditional probability tables (CPT)s for each node are given. Multivariate probability distribution of the variables in Figure 1 is as follows,

\[ P(C, S, W, B, A) = P(C) P(S) P(W|S, C) P(B|W, S, C) P(A|B, W, S, C). \]

With the local Markov property, the multivariate probability distribution of variables can be simplified as follows,

\[ P(C, S, W, B, A) = P(C) P(S) P(W|C) P(B|S, C) P(A|B). \]

Thus, the number of parameters in the model reduces from 31 \((2^5 - 1)\) to 10. This reduction in the number of parameters provides excellent convenience for modeling, calculation, and learning. This model, with fewer parameters, is robust against bias and variance effects [4].

Given a BN that helps to factorize the multivariate probability distribution of the variables, inferences about these variables can be carried out by marginalization (summing out over all ‘unnecessary’ variables). There are two types of inference support: predictive support and diagnostic support. Predictive support for a node depends on the evidence nodes connected to it via its parents. In contrast, diagnostic support for a node depends on the evidence nodes to it via its children. Let us consider the example given in Figure 1. When it is observed that a person suffers from backache, diagnostic support regarding the belief that new uncomfortable chairs were bought to that person’s office can be calculated with respect to the equation given below.

\[ P(C = T|A = T) = \frac{P(C = T, A = T)}{P(A = T)} \]
Here the numerator and denominator of the equation are calculated with the help of the following formulas.

\[
P(C = T, A = T) = \sum_{S,W,B \in \{T,F\}} P(C = T) P(S) P(W | C = T) P(B | S, C = T) P(A = T | B)
\]

\[
P(A = T) = \sum_{S,W,B,C \in \{T,F\}} P(C) P(S) P(W | C) P(B | S, C) P(A = T | B)
\]

These calculations become more complicated as the number of nodes in BNs increases. Therefore, marginalization is an NP-hard problem, especially when the number of variables is too high. Some practical exact inference algorithms have been proposed in the literature for performing these calculations, such as message passing, cycle cutset conditioning, and variable elimination. Besides the exact inference algorithms, approximate inference algorithms can also be used. The most preferred approximate inference algorithms are based on Markov Chain Monte Carlo (MCMC) methods [4].

Learning is the most crucial task in BN modeling. This task is divided into two subtasks. The first one is ‘structural learning’ in which the structure of the BN is created, and the second one is ‘parameter learning’ in which the parameters, i.e. conditional probability tables, in the BN are estimated. Both structural and parameter learning can be carried out directly by using expert knowledge. However, it is not always possible to access expert knowledge. Therefore, the structure and the parameters of the BN can be learned from data. In the literature, methods related to learning the structure of BN from data are analyzed under three groups. The first group is constraint-based algorithms which are based on probabilistic semantics of BNs. With these algorithms, links are added or deleted based on the results of statistical tests that define conditional independencies. The second group is score-based algorithms, which are based on scores that measure the candidate network’s quality based on the observed dataset. The third group is hybrid algorithms, which both use constraint-based and score-based algorithms together. In hybrid algorithms, a constraint-based approach is used to reduce the number of candidate BN structures, and then a score-based approach is used to achieve the best structure in this restricted area.

In the process of structure learning by score-based and hybrid approaches, inferred BNs are compared with some scores. In fact, these scores are used as a measurement for how well created BNs fit to the data. These scores evaluate consistency of estimates and diagnostics with real situations through a set of real observations. They can be used to identify weak points on the network, and nodes that make weak correlations with real situations. Thus, the performance of the estimates can be improved by changing the conditional probability tables, increasing the learning data, or updating the network.

Observations in BNs are divided into two categories as ‘observed’ and ‘unobserved’, for the calculation of the scores. The values of the observed nodes are extracted from the data, and they are used to estimate the unobserved node values with the Bayesian belief update. This process is repeated for each case in the data. The predicted values of the unobserved nodes for each case are compared with the actual observed values. Correctly predicted (successful) states and incorrectly predicted (unsuccessful) states are recorded. Successful and unsuccessful states are used to establish criteria for how accurate estimates are made for each unobserved node. Some of these criteria are error matrix, error rate, calibration table, logarithmic loss score, quadratic loss (Brier) score, spherical loss score, confusion index,
test sensitivity. In this study, loss scores are used to measure the quality of the generated BNs. The logarithmic loss score is between 0 and $\infty$, and a value that approaches 0 means high performance. The quadratic loss score takes a value between 0 and 2, and a value that approaches 0 means high performance. The spherical loss score is between 0 and 1, and to take a value that approaches 1 represents high performance [36]. These scores are calculated with Equations (4)–(6), respectively.

\[
\text{Logarithmic Loss} = \text{MOAC} \left( -\ln P_c \right) \tag{4}
\]

\[
\text{Quadratic Loss} = \text{MOAC} \left( 1 - 2P_c + \sum_{i=1}^{n} P_i^2 \right) \tag{5}
\]

\[
\text{Spherical Payoff} = \text{MOAC} \left( \frac{P_c}{\sum_{i=1}^{n} P_i^2} \right) \tag{6}
\]

In Equations (4)–(6), MOAC stands for the mean probability of a given state averaged over all cases; $P_c$ is the probability predicted for the correct state; $P_i$ is the probability predicted for state $i$ where $n$ is the number of states [12,36]. After the structural learning, which is the qualitative part of the BN modeling, is realized, the next step is parameter learning (CPTs), which is the quantitative part of the modeling. In the parameter learning step, it is aimed to get $\hat{\theta}$, the most probable value of $\theta$ that best describes the dataset, where $\theta$ is the set of all CPTs $\theta_{\pi_j, X_j} (j = 1, \ldots, p)$ in BN. Given dataset for $p$ variables as $X = \{X_1, \ldots, X_p\}$ and observations for the variable $X_j (j = 1, \ldots, p)$ as $X_j = \{x_1[j], x_2[j], \ldots, x_n[j]\}$, log-likelihood function denoted as $L(\theta) = \log p(X \mid \theta)$ is used to obtain $\hat{\theta}$ values. If the dataset is complete, this likelihood function can be written as follows using the assumptions of conditional independence and independence of the samples [31].

\[
L(\theta) = \log \left\{ \prod_{j=1}^{p} p \left( x_{1}[j], x_{2}[j], \ldots, x_{n}[j]; \theta \right) \right\}
\]

\[
= \log \left\{ \prod_{i=1}^{n} \prod_{j=1}^{p} p \left( x_{i}[j] \mid \pi_{i}[x_{i}(j)]; \theta \right) \right\} \tag{7}
\]

where $\pi_{i}[x_{i}(j)]$ is the $i$th parent of $x_{i}(j)$. Using Maximum Likelihood Estimation Method, $\hat{\theta}$ is achieved as follows.

\[
\hat{\theta} = \arg \max_{\theta} L(\theta) \tag{8}
\]

However, if the data is incomplete, Equation (7) cannot be performed to obtain $\hat{\theta}$. A standard method that is used to achieve $\hat{\theta}$ is EM algorithm. This algorithm finds a locally optimal maximum-likelihood estimate of $\theta$. MCMC is an alternative approach in the case of incomplete data [4,31].

2.3. Robust coplot method

Robust coplot, a graphical visualization method, is a robust variant of the classical coplot method. Robust coplot works better than classical coplot with datasets containing outliers.
since it is not affected by their presence. Robust coplot, which we briefly called coplot in this study, is a practical tool for visual inspection and rich interpretation of multivariate data. The n-dimensional dataset is embedded in two dimensions, in a way that allows relations between both variables and observations to be analyzed together. It consists of two graphs that are superimposed on each other: the first graph represents the distribution of n observations over two-dimensional space, and the second graph shows the relationships among p variables. The main advantage of this method is enabling a simultaneous investigation of the relations between the observations and between the variables for a dataset. This means that we will be able to see clusters of variables, correlation between the variables, clusters of observations, clusters of observations described by the variable vectors, and a possible outlier(s) in the same graph. This method has been used in different areas with different statistical analyses in literature [2]. In this study, coplot is used as an auxiliary method to produce an accurate BN by gathering useful information about the dataset as the constraints of BN model.

This method consists of three steps [2]. In the first step, the data matrix, \( X_{n \times p} \), is transformed into the standardized matrix, \( Z_{n \times p} \), with Equation (9)

\[
\begin{align*}
    z_{ij} &= \frac{x_{ij} - \text{med} (x_j)}{\text{MAD}(x_j)}
\end{align*}
\]

where, \( z_{ij} \) is the \( i \)th row and \( j \)th column element of the standardized matrix, \( \text{med}(\cdot) \) is the median function, \( x_j \) is the \( j \)th column of the data matrix, and \( \text{MAD}(x_j) = 1.4826 \text{med}(|x_j - \text{med}(x_j)|) \) is the scale estimate based on median absolute deviation.

In the second step, Robust multidimensional scaling (RMDS) [17] is used for finding a proper embedding of the dataset. RMDS tries to minimize the outlier aware cost function as given in Equation (10)

\[
\begin{align*}
    f (O, Y) &= \sum_{i<j} \left[ \delta_{ij} - d_{ij}(Y) - o_{ij} \right]^2 + \lambda \sum_{i<j} |o_{ij}|
\end{align*}
\]

where, \( \delta_{ij} \) corresponds to the city-block metric among each pair rows of \( Z_{n \times p} \), \( d_{ij}(Y) \) shows the Euclidean distance between \( i \)th and \( j \)th row of coordinate matrix \( Y_{n \times 2} \). \( \lambda > 0 \) is the parameter that controls the number of presumed outliers. \( i \)th row and \( j \)th column of the outlier matrix, \( O \), is \( o_{ij} = \text{sgn}(d_{ij} - d_{ij}(Y)) \max(0, |d_{ij} - d_{ij}(Y)| - 1/2) \) which defines the outlier variable.

In the third step, the second graph that consists of \( p \) vectors for \( p \) variables is superimposed on the graph generated by RMDS. Direction and magnitude of a vector are decided by using median absolute deviation correlation coefficient (MADCC) [42] given as;

\[
\begin{align*}
    \rho_{j, \text{MADCC}} &= \frac{\text{MAD}^2(u_j) - \text{MD}^2(k_j)}{\text{MAD}^2(u_j) + \text{MAD}^2(k_j)}
\end{align*}
\]

where, \( u_j \) and \( k_j \) are the robust principal variables which are defined as follows:

\[
\begin{align*}
    u_j &= \frac{z_j - \text{med}(z_j)}{\text{MAD}(z_j)} + \frac{v_j - \text{med}(v_j)}{\text{MAD}(v_j)} \\
    k_j &= \frac{z_j - \text{med}(z_j)}{\text{MAD}(z_j)} - \frac{v_j - \text{med}(v_j)}{\text{MAD}(v_j)}
\end{align*}
\]
where, \( z_j \) is the \( j \)th column of \( Z_{n \times p} \), and \( v_j \) is the projection values of all \( n \) observations in the MDS map on the \( j \)th variable vector for a specific direction.

Any categorical variable can be selected as a color-coded variable to reveal the cluster(s) of observation in the graph better.

In this study robust coplot implementations are performed by the RobCoP package for MATLAB (The MathWorks Inc 2014), a software package that is publicly available [3]. This package offers both classical and robust coplot analysis concerning the selection of Multi-dimensional scaling type, vectors’ correlation coefficient type, and the standardization type of the dataset.

3. Knowledge extraction procedure from a dataset

The methodological process proposed in this paper is based on two steps: (1) Visual investigation of the multidimensional data by coplot (2) Construction of the BN, which takes information about possible retained or discarded variable(s)/link(s) from coplot graph. These steps are described as follows.

STEP 1: The coplot graph is generated from data with all variables and observations. Each variable is represented by a vector emerging from the center of gravity of the observations. Each of the observations is color-coded, according to a selected categorical variable. The quality of the coplot graph is assessed by two types of the goodness of fit measures. The calculated Kruskal stress value (\( \sigma \)) is used as a measure for deciding how good the embeddings of the configuration of \( n \) observations obtained by RMDS is [3,30]. Correlation coefficient values, one for each variable, are used as a measure for the goodness of fit of the \( p \) variables. The direction of each vector is chosen so that the correlation between the actual values of the variable and its projections on the vector is maximized. The length of the vectors is approximately proportional to the correlation values of their associated variables. In the coplot graph, similar observations are located close to each other. Observations with high value in a specific variable should be located in the part of the vector points. With this technique, we will be able to see the clusters of observations, relations between the variables, and the characterization of observations as being above average in specific variables and below in others.

The outcome of the coplot analysis is a simple picture of multidimensional data. The following inferences can be obtained over this picture:

- If two variables are highly positively correlated, the vectors describing them will point in about the same direction and shown as close together. Therefore, if the angle between two vectors is acute, the correlation between the corresponding variables is positive.
- If the correlation between two variables is negatively high, the vectors describing them lay in opposite directions. In other words, if the angle between the two vectors is nearly 180 degrees, the correlation between the corresponding variables is negatively high.
- The vectors associated with uncorrelated variables will tend to be orthogonal.
- The variables that have low correlation values that do not fit into the coplot graph should be considered as the possible redundant variables which are not relate well to the features of the data. If a variable is calculated to have correlation value below 0.50, it may be considered as redundant variable.
Clusters of color-coded observations that are highly characterized by a specific variable will be mapped together and placed in the part of the two-dimensional space the vector points to, while observations with a low value in this variable will be placed at the opposite direction.

STEP 2: After the detailed examination of the coplot graph, the rules for retained or discarded variable(s)/link(s) in BNs can be summarized as follows.

- If two variables are highly positively correlated, a link should be placed between such variables in the BN structure. However, if the correlation between two variables is very high, that is, if the vectors associated with these two variables almost overlap on the coplot graph, one of these variables can be removed from the BN structure. That is because information carried by these two variables is similar, and in the BN to be constructed, the links between each of the other variables, and these two variables will be the same. At this point, the variable with a relatively smaller correlation value can be suggested to be discarded from BN representation. Even though multicollinearity is not one of the significant problems in BNs, having one or more highly correlated variables in the model will unnecessarily increase the number of operations to be performed in the learning of both the BN structure and CPTs.

- If two variables are highly negatively correlated, a link should be placed between such variables in BN representation.

- If two variables are uncorrelated, there should be no link between these two variables in BN representation.

- If the variables have low correlation values, they might be excluded from the BN representation.

- If a cluster of color-coded observations is in the same direction with the vector associated with a variable, a link should be placed between color-coded variable and that variable, in BN representation.

The proposed methodological process is summarized in Algorithm 1.

Before presenting our main application, this section follows the detailed analysis of the methodological process on a simple example. The Fisher’s iris dataset, which is frequently used in BN studies in the literature [27,28,33], from UCI database is used [13]. The dataset gives information on the iris plant type on the basis of four attributes with names sepal-length (slength), sepal-width (swidth), petal-length (plength), and petal-width (pwidth). The dataset has 150 observations of three classes with 50 observations each.

Figure 2 is the coplot graph of the iris data in which the colored observations and the four vectors represent the variables describing them. Each iris plant has been colored according to its type of plant: Setosa (red/o), Versicolour (black/x), and Virginica (blue/+). The Kruskal stress goodness of fit value of the coplot is found as 0.04, which is accepted as a good fit for the representation of multidimensional data in two dimensions [30]. The correlation values of the variables are found to be quite high. Because the length of the vectors is proportional to the correlation value of that variable, from Figure 2 we can observe that all variables contribute significantly to the graph [2]. We can say that all variables should take place in BN representation. Since the angle between the variables pwidth and plengt
Algorithm 1 Construction/Update of BN model using coplot.

**Input:** p-dimensional n points data matrix

**Output:** Optimal BN structure learned from the coplot

**Step 1:** Visualization of the multidimensional data

1. Generate the coplot graph from the data with all variables, one of which is specified as color-coded variable
2. Check the quality of the coplot graph with Kruskal Stress value (if $\sigma \leq 0.10$, coplot graph is appropriate for interpretation)

**Step 2:** Construction of BN

1. Discard a variable from the analysis if:
   (a) a vector is found to have low correlation value (<0.50)
   (b) a vector almost overlaps any other vector and relatively has a smaller correlation
2. Add a link between variables if:
   (a) the angle between two vectors is acute
   (b) the angle between two vectors is nearly 180 degrees
   (c) a cluster of observations of the color-coded variable is in the same direction with a vector
3. Remove/no link between variables if:
   (a) the angle between two vectors is nearly 90 degrees
4. Check the quality of the generated BN with loss scores

**Step 3:** Update the BN structure

1. Apply Step 1 and Step 2 over reduced data set until the BN with the best score value is obtained

is small, there should be a link between these two variables in BN representation. The variables slength and pwidth are in the same direction and the angle between them is an acute angle. Therefore, it is suggested that there should be a link between these two variables in BN representation. The variables swidth is nearly orthogonal to the other variables, there should be no link between swidth with other variables in BN representation. In the coplot graph we also see the variables describing the type of plants. It is evident that clusters of observations are separated from each other with red, black, and blue colors. The setosa (red) type of plants have higher values of the variable swidth. The virginica (blue) type of
plants have higher values of the variables slength, pwidth, and plength. Since all of these variables seem to be important in classifying the types of plants, it is recommended to put a link from class variable to all variables, in BN representation. According to the rules extracted from coplot graph, the constructed BN is given in Figure 3. Before the creation of BN, all the continuous variables we examine on the coplot graph are discretized [16].

---

**Figure 2.** Coplot graph of iris dataset: This graph includes four variables with a different shape notations based on the plant type (where setosa type is (o), versicolour type is (x), and virginica type is (+)).

**Figure 3.** The BN structure for iris data, constructed according to the rules which are obtained by coplot graph (Spherical payoff = 0.9802).
Luksza and Nguyen [33] used this dataset in their proposed algorithm to learn the BN structure. The BN created with respect to their approach is given in Figure A1 in Appendix A. The only difference between the BNs given with Figure 3 and Figure A1 is the link between class and pwidth variables. Luksza and Nguyen did not put a link between these variables. When comparing both BNs according to score values, the spherical payoff value of the BN in Figure 3 is 0.9802, whereas that of the BN in Figure A1 is 0.9278, which means a better BN structure is created with our proposed approach. Similarly, logarithmic and quadratic loss values for BN given in Figure 3 are smaller than those of the BN given in Figure A1. This result indicates that the BN structure learned from the coplot results represents the data better. Using the HUGIN software, several BN structures for the same dataset are created with different structural learning algorithms such as PC [44,45], NPC which is an expansion of PC [46], Greedy Search and Score (GSS) [10], and Rebane-Pearl Polytree [39]. These structures and their score values (spherical payoff) are given in Figure A2 in Appendix A. When all BN structures given in Figure A2 are compared by using these score values, it shows that the best BN structure, Figure 3, is the one obtained from the results of the coplot graph.

To emphasize the success of the proposed method, some changes are made to the dataset, and newly acquired dataset is named as corrupted iris dataset. Firstly, a noise variable with low correlation with the dataset, which is generated from uniform distribution with interval $[-0.3, 0.3]$, is added to data. Then, in order to produce a new variable with high correlation with another variable in the dataset, the noise variable is added to the slength variable. The newly obtained variable is named as slengthnoisy. Thus, the number of variables in the dataset is increased from 4 to 6. The coplot graph, in Figure 5, from the corrupted iris dataset is generated. The Kruskal stress goodness of fit value of the coplot graph is found as 0.10 which is accepted as fair fit for the representation of multidimensional data in two dimensions. From Figure 5 the following rules are obtained.

The correlation value of the noise variable is 0.458, which is compared to the correlation value of other variables. This value remained quite small. Since this value is less than 0.5, it is recommended to discard this variable from BN representation. The angle between the variables slength and slengthnoisy is quite small. This suggests that the information carried by these two variables is quite similar. In order to avoid the possible multicollinearity problem one of the variable can be eliminated. At this point, since the slengthnoisy variable has a relatively smaller correlation value, it can be suggested to be discarded from BN representation. The correlation values of the remaining variables are found to be quite high, meaning that they should be included in the BN representation. The correlation structures of the remaining variables are very similar to the correlation structures obtained from clean data. Therefore, the rules which are given in clean data will be valid for the corrupted data. It is essential to highlight that the coplot technique suggests constructing the same BN which is given in Figure 3 for clean and corrupted datasets. BN structures created by some learning algorithms over corrupted dataset are given in Figure 4. As seen in the figure, noise and slengthnoisy variables are included in BN models created by NPC, PC and Rebane-Pearl Polytree algorithms and links are placed between these variables and slength variable. In the BN structure, which is established with the GSS algorithm, there is no link between the noise variable and other variables, but a link is placed between the slengthnoisy and the slength. As a result, variables and links that are determined unnecessary according to coplot results are involved in BNs created with other algorithms. Thus, it shows that the BN
structure created by using the coplot graph will create simpler BNs since it will not contain unnecessary variables and links. This feature allows us not to have to deal with unnecessary operations, especially when working with a large number of variables.

4. An application of the proposed approach on a survey data

Every five years, a population-based field demographic and health survey, titled Turkish Demographic and Health Survey (TDHS), is conducted by the Institute of Population Studies at Hacettepe University, Turkey. In this study, we used two datasets that are extracted from TDHS 2008 and 2013, respectively.

4.1. A descriptive look at the dataset

TDHS is a nationally representative sample survey that is designed to provide information about the levels and trends on fertility, infant and child mortality, family planning, and
Figure 5. Coplot graph of corrupted iris dataset: This graph includes six (four original, and two artificial) with a different shape notations based on the plant type (where setosa type is (o), versicolour type is (x), and virginica type is (+)).

maternal and child health. Survey results are presented at the national level, including 12 geographical regions (NUTS 1), according to the survey topics. Interviews were carried out with 8003 and 9746 ever-married women between the reproductive ages of 15–49 in TDHS-2008 and TDHS-2013, respectively [47,48].

To get an accurate result from our methodology, the missing values were taking out of the data. After eliminating those cases from the analysis, we got 6008 and 5268 ever-married women between the reproductive ages of 15–49 in 2008 and 2013, respectively.

Since the TDHS covers many important topics about the population, we chose to extract a dataset from the women’s status questionnaire, which provides information on women’s perceptions about domestic violence.

Although the most apparent indicator of domestic violence is beating, it is not the only one. Therefore, in our study, the concept of domestic violence is considered as physical and verbal actions that injure the individuals and make them feel humiliated, frayed, and/or under emotional pressure. In order to investigate domestic violence and to connect it with other demographic variables, it is necessary to define the violence as a measurable variable. In this study, various dimensions of the violence seen in the families are measured according to the following three bases. Each of them corresponds to a latent variable: physical violence against women (S763), the belief that women are more worthless than men; in other words, gender inequality (W129), and the restriction of women’s freedom in some cases (S765). Table 1 summarizes the latent variables and corresponding observed
Table 1. Dimensions of the domestic violence with observed variables.

| Observed Variables | Physical violence (S763) | Gender inequality (W129) | Restriction of women’s freedom (S765) |
|--------------------|--------------------------|--------------------------|-------------------------------------|
| When she burns food | Does not cook food        | The important decisions concerning the family should be made only by men | Preventing the woman from seeing her female friends |
| When she neglects the housework / children | It is better to educate a son than a daughter | A married woman should not be allowed to work outside of the house | Limiting her contact with her family |
| When she wastes money | A woman should not have the permission to state an opinion | Insisting on always knowing where she is | Distrusting her with money |
| When she goes out without telling her husband | Women should be virgins on their wedding night | There should be fewer women politicians | Accusing her of being unfaithful |
| When she refuses to have sex with her husband | | Husbands are not obliged to help with the housework | |
| When she argues with her husband | | | |

Table 2. Demographic variables.

| Demographic variables |
|-----------------------|
| V012: Age             |
| V136: Number of the household members |
| V149: Educational attainment |
| V190: Wealth index    |
| V201: Total number of children ever-born |
| V212: The respondent’s age at 1st birth |
| V218: Number of living children |
| V511: The respondent’s age at first cohabitation |
| V512: Years past since first cohabitation |
| V729: Husband/partner’s educational attainment |
| V730: Husband/partner’s age |

variables. Descriptive information of three latent variables are presented in Tables C1, C2 and C3 in Appendix C.

Specific background characteristics may affect domestic violence, and thus, the relationship between them should be taken into consideration while measuring it. For this reason, 11 demographic variables are selected and added into our analysis. Table 2 describes our demographic variables. Descriptive information gathered from the selected demographic variables are given in Tables D1–D11 in Appendix D.

In this application, for each of the datasets extracted from 2008 and 2013 time periods, BN structures that model each latent variable’s relationship with 11 demographic variables are created by following procedure.

— Firstly, BNs that focus on latent variables for given time periods are built by regarding all variables based on our prior knowledge. Domestic violence is a known topic that we have little knowledge about. Since learning from the data is time-consuming, we prefer to build the first BNs according to our knowledge.
Table 3. Correlations and orientations of 11 demographic variables in 2008.

| Variable | Orientation | Correlation |
|----------|-------------|-------------|
| V012     | 297         | 0.92816     |
| V136     | 51          | 0.63768     |
| V149     | 210         | 0.74413     |
| V190     | 237         | 0.74149     |
| V201     | 1           | 0.86109     |
| V212     | 217         | 0.5859      |
| V218     | 1           | 0.85427     |
| V511     | 216         | 0.59281     |
| V512     | 314         | 0.93096     |
| V729     | 215         | 0.61946     |
| V730     | 300         | 0.90796     |

- Since the expert knowledge is not available about the dataset, a preliminary examination of the relationship structures within it is performed by the coplot method.
- The created BNs are updated by using the coplot results. Updating can be carried out in two ways; Adding new link(s)/node(s) to the first BN and removing the existing link(s)/node(s) from the first BN.
- The success of the updates is measured with the loss scores and then decides whether the resulting new BNs represent the dataset better than the previous one.
- The final BNs are used to predict domestic violence’s value by changes in values in other demographic variables.

As a result, six different coplot graphs and six BN models (based on three latent variables for 2008 and 2013 time periods) are generated for domestic violence. The most obvious result from coplot graphs is that the variable clusters are comprised of the same variables in both 2008 and 2013. Although there are slight differences in the angles between vectors, the interpretations of the correlations between variables are very similar for both years. Since the relation structures are very similar for both years, the BN structures that have been constructed were also the same. For this reason, only the results of 2008 are given in detail here. In addition, since the BN structures obtained for all latent variables are the same, only the physical violence (S763) are presented in detail.

4.2. Coplot findings

Coplot graphs are constructed to visualize our data in two-dimensions. Three latent variables (S763, W129, S765) are considered as color-coded variables, and coplot graphs are drawn separately. In Appendix B, Figures B13 and B14 represent the latent variables’ graphs and 11 demographic variables, according to (a) 2008, and (b) 2013 data. The degree of the variable vectors, which are relative to the x-axis (orientations) and the correlation values of each vector are given in Table 3 for the 2008 time period.

Table 4 presents the angles between the demographic variables for the same dataset. Highly positively and negatively correlated variables and uncorrelated variables can be easily detected from this table. For example, the angle between the educational attainment (V149) and husband’s age (V730) is 90 degrees; thus, these are assumed to be uncorrelated
Table 4. Angles between 11 demographic variables in 2008.

|          | V012 | V136 | V149 | V190 | V201 | V212 | V218 | V511 | V512 | V729 | V730 |
|----------|------|------|------|------|------|------|------|------|------|------|------|
| V012     | 0    | 114  | 87   | 60   | 64   | 80   | 64   | 81   | 17   | 82   | 3    |
| V136     | 0    | 159  | 174  | 50   | 166  | 50   | 165  | 97   | 164  | 111  |      |
| V149     | 0    | 27   | 151  | 7    | 151  | 6    | 104  | 5    | 90   |      |      |
| V190     | 0    | 124  | 20   | 122  | 21   | 77   | 22   | 63   |      |      |      |
| V201     | 0    | 144  | 0    | 145  | 47   | 146  | 61   |      |      |      |      |
| V212     |      |      |      | 97   | 2    | 83   |      |      |      |      |      |
| V218     |      |      |      |      | 145  | 47   | 146  | 61   |      |      |      |
| V511     |      |      |      | 98   | 1    | 84   |      |      |      |      |      |
| V512     |      |      |      | 99   | 14   |      |      |      |      |      |      |
| V729     |      |      |      |      |      | 85   |      |      |      |      |      |
| V730     |      |      |      |      |      |      |      |      |      |      | 0    |

Variables. Also, the angle between wealth status (V190) and the number of household members (V136) is close to 180 degrees, which means that these are highly negatively correlated variables. The angles between respondent’s age at first birth (V212), woman’s educational attainment (V149), the respondent’s age at first cohabitation (V511) and partner’s educational attainment (V729) are less than 10 degrees, so these are highly positively correlated variables.

After examining the relationship between demographic variables, colored observations are examined to reveal the possible relationship between the latent variable (physical violence) and demographic variables. Red circles represent women who normalize the physical violence and say that it is acceptable in some cases. In contrast, black crosses represent women who state that physical violence is unacceptable in any case. Due to the high number of observations, Figure B13 does not have a clear overview of women who agree or disagree with the given opinion. In order to observe this discrimination in more detail, coplot graphs are drawn according to the NUTS I regions (see Figures B1–B12 in Appendix B). The numbers of ever-married women interviewed from the NUTS I regions for 2008 dataset are given in Table 5, and for 2013 dataset is given in Table D12 in Appendix D.
Figure 6. Coplot graph of the Istanbul region for the 2008 time period: This graph includes 11 demographic variables and physical violence latent variable. Each observation is represented based on women’s opinion (those who agree are (o), and those who disagree are (x)).

Coplot graphs, which are drawn according to the NUTS I regions, enable us to assess whether regional differences affect the women’s opinions visually. The change in proportions of the women who agree or disagree with the given opinion (S763) over the regions, can be seen as colored observations. Among NUTS I regions, Istanbul is the largest and the most cosmopolitan region in Turkey. It is the center of commerce, manufacturing, and culture. The coplot graph of Istanbul shows that the relation structures found in Istanbul dataset, are very similar to the relation structures in the Turkey 2008 dataset. Therefore, this region is investigated and presented in detail here, see Figure 6.

An obvious conclusion that can be drawn from Figure 6 is that there are three main variable clusters. In the first cluster, the number of the living children (V218) and the number of the total children (V201) are highly positively correlated variables. The number of household members (V136) has a mild positive correlation with them. In the second cluster, age (V012) and the partner’s age (V730) are highly positively correlated variables, and years passed since first cohabitation (V512) is positively correlated with these two variables. In the last cluster, age at first cohabitation (V511), age at first birth (V212), husband/partner’s educational attainment (V729), and educational attainment (V149) are positively correlated variables, and wealth status (V190) is also positively correlated with these variables. Apart from these conclusions, the number of household members (V136) is highly negatively correlated with wealth index (V190), and also negatively correlated with the respondent’s age at first cohabitation (V511), husband’s educational attainment (V729),
the respondent’s age at first birth (V212), and educational attainment (V149). Highly positively correlated variables may be accepted as duplicated or nearly duplicated to provide information for BN. This knowledge can be used to determine a more parsimonious set of variables.

According to the colored observations in Figure 6, we can roughly say that the number of women who agree with the given statement is lower than women who do not agree. The proportion of women who legitimize violence is higher when there are more members of the household, and the wealth status and the educational level of the spouse are low. This proportion is high among women with more children. It is also high among women who married and got pregnant at a young age. In the same manner, the increase in husband’s educational level, wealth status, and marital age, lead to a decrease in the proportion of women who agree with the physical violence. Similarly, women who are well-educated and married to well-educated men, have a high quality of life, got married at an older age, and had children at an older age, take a stand against physical violence. This knowledge indicates that there should be links between the physical violence and the corresponding demographic variables in the BN representation.

These results are very similar throughout Turkey. For West Marmara, Aegean, East Marmara, East Anatolia, West Anatolia, and East Blacksea, the same interpretations can be given. Unbalance in development levels among the regions affect women’s views on physical violence, and it can be seen in coplot graphs of corresponding regions. In the Mediterranean and Central Anatolia, the proportion of women who agree with the given statements about physical violence is higher than the women who live in Istanbul. There is a notable rise in the proportion of agreement with physical violence among the women who live in Northeast Anatolia (10), Centraleast Anatolia (11), Southeast Anatolia (12), see Appendix B. Coplot graphs unveil interesting results among the NUTS I regions. The proportion of women who agree with the given opinions in both years is higher in rural regions than in urban regions.

4.3. Building a BN

After the examination of the data by coplot analysis, this section explains how to utilize the extracted knowledge to improve the first BN. Since similar correlation structures are found among the variables for each NUTS I region by coplot method, we use all data without dividing them into NUTS I regions to construct BN.

Before the construction of BNs, let us consider the imbalance problem in data. When Table 5 is examined, it is seen that there is a general imbalance between agreeing and disagreeing rates in all regions in the responses given to all physical violence (S763), gender inequality (W129), and restriction of women’s freedom (S765) latent variables. It is expected that women do not express a positive attitude towards the ideas that affirm violence against them, confirm men’s superiority over women, and restrict their freedom. So agree rates for latent variables that reflect these ideas will also be low. This imbalance problem is often encountered in many real datasets. The skewed distribution of data causes problems in learning algorithms, and results are biased towards the group, which has a higher frequency. However, the information carried by the low frequency group is just as important as the high frequency group and gives striking information for end users [8,29,34]. Coplot graphs, which are drawn according to the NUTS
Figure 7. BN structure for the dataset extracted from the 2008 survey, built according to our knowledge: This network includes 11 demographic variables and physical violence latent variable.

Table 6. The scoring rule results for Figure 7.

| Metric                  | Value   |
|-------------------------|---------|
| Logarithmic loss        | 0.2776  |
| Quadratic loss          | 0.1657  |
| Spherical payoff        | 0.9101  |

I regions, enable us to visually assess the relations of the lower frequency group of observations with the demographic variables. Therefore, in this study, the information about lower frequency group of observations are also included in the BN construction process.

For 2008 data, the first constructed BN for the latent variable of physical violence (S763) and 11 demographic variables is given in Figure 7. The scoring rule results for Figure 7 are given in Table 6. We have updated BN, given in Figure 7, according to the following statements:

- **The variable ‘Partner's Age (V730)’ is removed:** Since V730 and Age (V012) are highly positively correlated variables, the lower correlation valued variable, V730, is decided to be removed.
- **The variable ‘Number of the Living Children (V218)’ is removed:** Since the variable ‘Total children ever-born (V201)’ is highly positively correlated with V218, the lower correlation valued variable, V218, is decided to be removed.
- **The variable ‘Respondent's age at first birth (V212)’ is removed:** Since the variables V212, ‘Age at first marriage (V511)’ and ‘Partner’s educational attainment (V729)’ are highly positively correlated to each other, the lower correlation valued variable, V212, is decided to be removed.
- **The link between ‘Educational attainment (V149)’ and ‘Age at first marriage (V511)’ is removed:** V149, V511, and Husband’s educational attainment (V729) are highly positively correlated variables. Since the angle between V729 and V511 is lower than the angle between V149 and V511, we added a link between V729 and V511 instead of between V149 and V511.
Figure 8. The updated BN structure for the dataset extracted from the 2008 survey, built according to coplot analysis results: This network includes eight demographic variables and physical violence latent variable.

Table 7. The scoring rule results for Figure 8.

| Metric                  | Value   |
|-------------------------|---------|
| Logarithmic loss        | 0.2254  |
| Quadratic loss          | 0.1389  |
| Spherical payoff        | 0.9203  |

- The link between ‘Years passed since first marriage (V512)’ and ‘Age at first marriage (V511)’ is removed: Since the angle between these two variables is nearly 90 degrees, V512 and V511 are accepted as uncorrelated variables, and the link between them is removed.
- A link between ‘Educational attainment (V149)’ and ‘Number of the household members (V136)’ is added: Since the angle between these variables is nearly 180 degrees, V149 and V136 are accepted as highly negatively correlated variables, and a link is added.
- A link between ‘Total children ever-born (V201)’ and ‘Physical Violence (S763)’ is added: The variable V201 is in the direction of red circles (the women who accept physical violence). So we added a link between these two variables.
- A link between ‘Partner’s educational attainment (V729)’ and ‘Physical Violence (S763)’ is added: The variable V729 is in the direction of black crosses (the women who are against physical violence). So we added a link between these two variables.

Figure 8 illustrates the updated BN and the scoring rule results of updated BN are given in Table 7.

These two network structures are compared by loss scores. The values of logarithmic and quadratic loss scores of updated BN, given in Figure 8, are lower than that of the first BN, given in Figure 7, and the value of the spherical payoff score of updated BN is higher.
Figure 9. The re-updated version of the BN given in Figure 7, which is improved according to coplot analysis results: This network includes six demographic variables and physical violence latent variable.

Table 8. The scoring rule results for Figure 9.

| Metric            | Value   |
|-------------------|---------|
| Logarithmic loss  | 0.2763  |
| Quadratic loss    | 0.1654  |
| Spherical payoff  | 0.9101  |

This result means that the updated BN represents the data better than the BN, which is constructed based on our knowledge.

After elimination of the three variables, the coplot graph of the remaining dataset is re-evaluated. We have taken the update process one step further with the following statements to understand whether we can achieve more compact and easy-to-interpret BN.

- The variable ‘Age at first marriage (V511)’ is removed: Since the variable ‘Partner’s educational attainment (V729)’ is highly positively correlated with V511, the lower correlation valued variable, V511, is decided to be removed.
- The variable ‘Number of the household members (V136)’ is removed: When compared to the other variables, V136 has a low correlation value, so it is decided to be removed.

The BN, re-updated with statements above, is given in Figure 9, and the scoring rule results for Figure 9 are given in Table 8. Here, we created a BN, given in Figure 9, with the same score value as the BN in Figure 7. These results show that we have obtained a BN, which has fewer nodes/links and it is easily interpretable with the same power as the first BN to represent the data. At this point, the researcher may decide which BN should be used according to the study’s aim.

This application indicates that the presented approach can be used to improve the BN structures when expert knowledge is not available. Once the BN is learned, it can perform many tasks such as prediction, diagnostics, anomaly detection, and much more. In our application, the resulting BN can be used for prediction. In BNs, the prediction is the process of calculating the probability distribution for one or more variables, given the information (evidence) we have about one or more variables. These predictions can be in the form of predicting marginal probability tables, conditional probability tables, or joint probability tables.
The model, Figure 9, not only can be used to see relations among variables which explain reasons of physical violence but also can predict the probability of agree/disagree rates of affirming physical violence in certain conditions that state of specific nodes are determined. Figure A3 in Appendix A is an estimate of the state of affirmation of physical violence on the assumption that a woman and her husband have no education, have 5–10 children, and have the poorest life standard. In Netica, the statuses of ‘no education’ in ‘educational attainment’, ‘no education’ in ‘partner’s educational attainment’, ‘poorest’ in ‘wealth index’, ‘5–10’ in ‘total children ever born’ are set as 100%; that is, the statuses of the evidence variables are determined. Then the probabilities of the whole network are updated and the probability change in the status of ‘physical violence’ is observed. In this case, the probability of ‘agree’ in ‘physical violence’ increases from 10.4% to 35.4%. Therefore, it can be found that the status change of evidence node variables affects the probabilities of ‘physical violence’. The decrease in education level and wealth index, and the increase in the number of children increase the agree rate of affirming physical violence. This result is consistent with the human behavior science.

The existence and frequency of violence against women are some of the most challenging issues to measure. This application demonstrates that the combined use of BN and coplot in a two-step methodology is a powerful tool to grasp and interpret human behaviors on this topic. This study shows that coplot may be accepted as a powerful method for exploring the association between observations and variables in multidimensional data. Coplot results are embedded in the BN as model constraints. The BN is updated with this embedded information. Our analysis of the TDHS data suggests that interpretations of physical violence presented with the updated BN is simpler, more tractable, and easier to understand.

5. Conclusion

BNs can be considered as an advanced and suitable modeling technique describing the complex structure of the survey data, where a high number of observed variables and latent variables frequently appear. Recently, learning BNs has become popular research topic in modeling in this area. In practical application, we often encounter difficulties in learning structure. When the dataset contains too many variables, it is considered time-consuming and challenging to learn the BN structure from data due to the big search space. In the literature, many research studies, to overcome this difficulty, have been proposed to explain how to learn BNs from data, when the expert knowledge does not exist. In this study, we present a method for extracting useful knowledge from survey data by performing a BN, which incorporates coplot results as prior knowledge about association patterns that are hidden in the data. Since the BNs model the relationships among nodes, we use the coplot as an exploratory technique to determine these relations among variables. Besides, coplot graph allows information about the links between the variables where researchers have partial/no knowledge about the research topic being studied. The main contribution of this study is to help the researcher develop a more compact, tractable, and more interpretable model by using the combinatorial coplot method in the BN. The practicality and effectiveness of the proposed method are verified by a dataset extracted from TDHS survey study. This application has demonstrated the usefulness of this methodological approach.
for improving the power of representation data of BNs. Therefore, the limitations associated with expert knowledge or learning structure directly from data are overcome by methodological approach applied in this research.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Derya Ersel http://orcid.org/0000-0003-0393-9198
Yasemin Kayhan Atılgan http://orcid.org/0000-0002-2612-7216

References

[1] R.D. Anderson, R.D. Mackoy, V.B. Thompson, and G. Harrell, A bayesian network estimation of the service-profit chain for transport service satisfaction, Decis. Sci. 35 (2004), pp. 665–689.
[2] Y.K. Atılgan, Robust coplot analysis, Commun. Stat. - Simul. Comput. 45 (2016), pp. 1763–1775.
[3] Y.K. Atılgan and E.L. Atılgan, Robcop: A matlab package for robust coplot analysis, Open J. Stat. 7 (2017), p. 23.
[4] I. Ben-Gal. Bayesian networks, in Encyclopedia of Statistics in Quality and Reliability, F. Ruggeri, F. Faltin and R. Kenett, eds., John Wiley & Sons, 2007.
[5] S. Beretta, M. Castelli, I. Gonçalves, R. Henriques, and D. Ramazzotti, Learning the structure of bayesian networks: A quantitative assessment of the effect of different algorithmic schemes, Complexity 2018 (2018), pp. 1–12.
[6] J.G. Blodgett and R.D. Anderson, A bayesian network model of the consumer complaint process, J. Serv. Res. 2 (2000), pp. 321–338.
[7] S.G. Botcher and D.C. Deal, A package for learning bayesian networks, J. Stat. Softw. 8 (2002), pp. 594–605.
[8] P. Branco, L. Torgo, and R.P. Ribeiro, A survey of predictive modelling under imbalanced distributions, CoRR, Preprint (2015). Available at http://arxiv.org/abs/1505.01658.
[9] T.A. Brown, T.F. Cash, and P.J. Mikulka, Attitudinal body-image assessment: Factor analysis of the body-self relations questionnaire, J. Pers. Assess. 55 (1990), pp. 135–144.
[10] D. Chickering, Optimal structure identification with greedy search, J. Mach. Learn. Res. 3 (2002), pp. 507–554.
[11] F. Díez-Mesa, R. De Oña, and J. De Oña, Bayesian networks and structural equation modelling to develop service quality models: Metro of seville case study, Trans. Res. Part Policy Pract. 118 (2018), pp. 1–13.
[12] W. Dlamini, Application of a Bayesian network for land-cover classification from a landsat 7 etm+ image, Int. J Remote Sens 32 (2011), pp. 1–18.
[13] D. Dua and C. Graff, UCI machine learning repository, Preprint (2017). Available at http://archive.ics.uci.edu/ml.
[14] M. Ducher, E. Kaltbacher, F. Combarnous, J. Vilaine, B. McGregor, D. Fouque, and J. Fauvel, Comparison of a bayesian network with a logistic regression model to forecast iga nephropathy, Biomed. Res. Int. 2013 (2013), p. 686150.
[15] Q. Fan and W. Zhong, Variable selection for structural equation with endogeneity, J. Syst. Sci. Complex 31 (2018), pp. 787–803.
[16] U.M. Fayyad and K.B. Irani, Multi-Interval discretization of continuous-valued attributes for classification learning, in IJCAI, Morgan Kaufmann, San Francisco, CA, 1993, pp. 1022–1027.
[17] P.A. Forero and G.B. Giannakis, Robust multi-dimensional scaling via outlier-sparcity control, in 2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), Pacific Grove, CA, 6–9 November 2011, IEEE, New York, pp. 1183–1187.
[18] Y. Garcia-Chimeno, B. Garcia-Zapirain, M. Gomez-Beldarrain, B. Fernandez-Ruanova, and J. Garcia-Monco, Automatic migraine classification via feature selection committee and machine learning techniques over imaging and questionnaire data, BMC Med. Inform. Decis. Mak. 17 (2017), p. 38.

[19] O. Gevaert, F. De Smet, E. Kirk, B. Van Calster, T. Bourne, S. Huffel, Y. Moreau, D. Timmerman, B. De Moor, and G. Condous, Predicting the outcome of pregnancies of unknown location: Bayesian networks with expert prior information compared to logistic regression, Human Reprod. (Oxford, England) 21 (2006), pp. 1824–1831.

[20] C.C. Goh and G. Hu, Exploring the relationship between metacognitive awareness and listening performance with questionnaire data, Lang. Awareness 23 (2014), pp. 255–274.

[21] T. Golob, Structural equation modeling for travel behavior research, Trans. Res. Part B: Methodol. 37 (2003), pp. 1–25.

[22] J.A. Hagenaars, Loglinear Models with Latent Variables, Vol. 94, Sage, Newbury Park, CA, 1993.

[23] D. Heckerman, Bayesian networks for data mining’data mining and knowledge discovery, Data. Min. Knowl. Discov. 1 (1997), pp. 79–119.

[24] T. Heinen, Latent Class and Discrete Latent Trait Models: Similarities and Differences., Sage Publications Inc, Thousand Oaks, CA, 1996.

[25] M. Husky, Y. Zgueb, U. Ouali, C.I.A. Gonzalez, M. Piras, G. Testa, A. Maleci, A. Mulas, A. Montisci, S. Nujedat, G. Kalcev, I. Teodorov, A. Preti, M. Angermeyer, and M. Carta, Principal component analysis of the well-being at work and respect for human rights questionnaire (wwrrr) in the mediterranean region, Clin. Pract. Epidemiol. Mental Health : CP & EMH 16 (2020), pp. 115–124.

[26] R. Jacobucci, A.M. Brandmaier, and R.A. Kievit, Variable selection in structural equation models with regularized MIMIC models, 2018. Available at https://psyarxiv.com/bxzjf/.

[27] R. Kohavi, Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree hybrid, in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, AAAI Press, Portland, OR, 09, 1996, pp. 202–207.

[28] S.B. Kotsiantis and P.E. Pintelas, Logitboost of simple bayesian classifier, Inform. (Slovenia) 29 (2005), pp. 53–60.

[29] B. Krawczyk, Learning from imbalanced data: Open challenges and future directions, Progress Artif. Intell. 5 (2016), pp. 221–232.

[30] J. Kruskall, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika 29 (1964), pp. 1–27.

[31] W. Liao and Q. Ji, Learning bayesian network parameters under incomplete data with domain knowledge, Pattern Recogn 42 (2009), pp. 3046–3056.

[32] D.A. Linzer and J.B. Lewis, polca: An r package for polytomous variable latent class analysis, J. Stat. Softw. 42 (2011), pp. 1–29.

[33] K. Luksza and H.S. Nguyen, On application of rough sets to Bayesian network construction from data, in Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer–Verlag Berlin Heidelberg, 2006.

[34] M. Maragoudakis, K. Kermainidis, A. Garmpis, and N. Fakotakis, Dealing with imbalanced data using bayesian techniques, Proceeding of 5th International Conference on Language Resources and Evaluation (LREC’06), Genoa, Italia, European Language Resources Association (ELRA), 2006, pp. 1045–1050.

[35] A.L. McCutcheon, Latent Class Analysis, Vol. 64, Sage, Beverly Hills, CA, 1987.

[36] M. Millett Granger and M. Henrion, Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis, New York, Cambridge Univ Press, 1990.

[37] T.D. Nielsen and F.V. Jensen, Bayesian Networks and Decision Graphs, Springer, New York, 2007.

[38] J. Pacheco, S. Casado, and L. Núñez, A variable selection method based on tabu search for logistic regression models, Eur. J. Oper. Res. 199 (2009), pp. 506–511.

[39] G. Rebane and J. Pearl, The recovery of causal poly-trees from statistical data, Preprint (1987). Available at http://arxiv.org/abs/1304.2736.
[40] S. Salini and R.S. Kenett, *Bayesian networks of customer satisfaction survey data*, J. Appl. Stat. 36 (2009), pp. 1177–1189.

[41] P. Sebastiani and M. Ramoni, *On the use of bayesian networks to analyze survey data*, Res. Official Stat. 4 (2001), pp. 53–64.

[42] G. Shevlyakov and P. Smirnov, *Robust estimation of a correlation coefficient: An attempt of survey*, In *Proceedings of the 9th International Conference on Computer Data Analysis and Modeling*, Vol. 1, 2010, pp. 108–115.

[43] O. Sornil and S. Poovutthikul, *Constructing bayesian networks from association analysis*, in *PRICAI 2006: Trends in Artificial Intelligence*, Vol. 4099, Springer, Berlin, Heidelberg, 2006, pp. 231–240.

[44] P. Spirtes and C. Glymour, *An algorithm for fast recovery of sparse causal graphs*, Soc. Sci. Comput. Rev. 9 (1991), pp. 62–72.

[45] P. Spirtes, C. Glymour, and R. Scheines, *Open problems*. in: *Causation, Prediction, and Search*. Lecture Notes in Statistics, vol 81. Springer, New York, 1993. https://doi.org/10.1007/978-1-4612-2748-9_12

[46] H. Steck and V. Tresp, *Bayesian belief networks for data mining*, in *Proceedings of the 2nd Workshop on Data Mining Und Data Warehousing Als Grundlage Moderner Entscheidungsunterstuzende Systeme*, 2000.

[47] TDHS Data, *Turkey Demographic and Health Survey – 2008*, Hacettepe University, Institute of Population Studies, Ankara, Turkey, 2008. Available at https://dhsprogram.com/data/dataset/Turkey_Standard-DHS_2008.cfm?flag=0 (Online; accessed 24 July 2019).

[48] TDHS Data, *Turkey Demographic and Health Survey – 2013*, Hacettepe University, Institute of Population Studies, Ankara, Turkey, 2013. Available at https://dhsprogram.com/data/dataset/Turkey_Standard-DHS_2013.cfm?flag=0 (Online; accessed 24 July 2019).

[49] J. Tomasik, S.Y.S. Han, G. Barton-Owen, D.M. Mirea, N.A. Martin-Key, N. Rustogi, S.G. Lago, T. Olmert, J.D. Cooper, S. Ozcan, P. Eljasz, G. Thomas, R. Tuytten, T. Metcalfe, T.S. Schei, L.P. Farrag, L.V. Friend, E. Bell, D. Cowell, and S. Bahn, *A machine learning algorithm to differentiate bipolar disorder from major depressive disorder using an online mental health questionnaire and blood biomarker data*, Transl. Psychiatry. 11 (2021), p. 41.

[50] D.P. Yang and J.L. Li, *Research on personal credit evaluation model based on bayesian network and association rules*, in *2007 International Conference on Wireless Communications, Networking and Mobile Computing*, IEEE, Honolulu, HI, USA, 2007, pp. 3677–3680.

[51] A. Zaidman-Zait, L. Zwaigenbaum, E. Duku, T. Bennett, P. Szatmari, P. Mirenda, I. Smith, T. Vaillancourt, J. Volden, C. Waddell, C. Kerns, M. Elsabbagh, S. Georgiades, W.J. Ungar, E. Fombonne, and W. Roberts, *Factor analysis of the children's sleep habits questionnaire among preschool children with autism spectrum disorder*, Res. Dev. Disabil. 97 (2020), pp. 103548.

[52] D. Zellner and G. Zellner, *Variable selection in logistic regression models*, Commun. Statist.-Simulat. Comput. 33 (2004), pp. 787–805.

[53] Y.Q. Zhang, G.L. Tian, and N. Tang, *Latent variable selection in structural equation models*, J. Multivar. Anal. 152 (2016), pp. 190–205.

[54] Z. Zhang, V. Trevino, S.S. Hoseini, S. Belciug, A.M. Boopathi, P. Zhang, F. Gorunescu, V. Subha, and S. Dai, *Variable selection in logistic regression model with genetic algorithm*, Ann. Transl. Med. 6 (2018), p. 45. Available at http://atm.amegroups.com/article/view/18292.