Probabilistic SEM: an augmentation to classical Structural Equation Modelling

by

Keunyoung Yoo

Submitted in partial fulfilment of the requirements for the degree
Master of Commerce (Statistics)
in the Faculty of Economics and Management Sciences
University of Pretoria, Pretoria

May 2017

Publication data:

Keunyoung Yoo. Probabilistic SEM: an augmentation to classical Structural Equation Modelling. Master's dissertation, University of Pretoria, Department of Statistics, Pretoria, South Africa, May 2017.

Probabilistic SEM: an augmentation to classical Structural Equation Modelling

by

Keunyoung Yoo
E-mail: kyyou92@gmail.com

Abstract

Structural equation modelling (SEM) is carried out with the aim of testing hypotheses on the model of the researcher in a quantitative way, using the sampled data. Although SEM has developed in many aspects over the past few decades, there are still numerous advances which can make SEM an even more powerful technique. We propose representing the final theoretical SEM by a Bayesian Network (BN), which we would like to call a Probabilistic Structural Equation Model (PSEM). With the PSEM, we can take things a step further and conduct inference by explicitly entering evidence into the network and performing different types of inferences. Because the direction of the inference is not an issue, various scenarios can be simulated using the BN. The augmentation of SEM with BN provides significant contributions to the field. Firstly, structural learning can mine data for additional causal information which is not necessarily clear when hypothesising causality from theory. Secondly, the inference ability of the BN provides not only insight as mentioned before, but acts as an interactive tool as the 'what-if' analysis is dynamic.

Keywords: Structural Equation Modelling, Bayesian Network, Graphical model.

Supervisors : Dr. A. J. de Waal
Department : Department of Statistics
Degree : Master of Commerce

Acknowledgements

I dedicate this research to the following people, without whom this would not have happened:

- My God, for leading my way and using me as a vessel for His glory;
- My father, Seonghyun Yoo, and my mother,Okkyong Hwang, for sacrificing their life in Korea and moving to South Africa so that I could become the man I am today;
- My supervisor Dr de Waal for introducing me to the topic of SEM and BN, being the source of inspiration and supporting me immensely in making this work come to life;
- Prof. Wiese, for gladly providing the Facebook advertisement dataset for application;
- Dr Fletcher, for various guidance and expertise which were crucial with regards to SEM;
- CAIR, for their generous financial grant which made this research possible.

Contents

List of Figures

List of Algorithms

List of Tables

Chapter 1

Introduction

Structural equation modelling (SEM) is carried out with the aim of testing hypotheses on the model of the researcher in a quantitative way, using the sampled data [31]. Although this sampled data consists solely of observed measures, the variables in the model can be unobserved constructs as well [16]. Due to its flexibility, SEM is preferred by researchers in situations where one cannot simply design and conduct experiments, for example, because of ethical concerns, or when data is not observable [1]. Although SEM has developed in many aspects over the past few decades, there are still numerous advances which can make SEM an even more powerful technique.

1.1 Motivation

SEM and Bayesian networks (BN) are graphical models. However, there is limited literature on the relationship between these two techniques. While Xu et al. have used the goodness-of-fit measures from SEM to select the best BN for their analysis [34], we aim to augment SEM using BN. Making the connection between these two ideas in this way can provide SEM researchers with an additional technique for analysis. Additionally, SEM is most often used as a tool for hypothesis testing only and not for inference. By having a BN to work with, we add prediction to the capabilities of SEM.

1.2 Objectives

The objectives of this dissertation are as follows:

- Address the basic aspects of SEM
- Establish a link between SEM and BN from the perspective of graph theory
- Represent classical SEM as Probabilistic SEM (PSEM) which is by definition a BN
- Conduct inferences for SEM research questions using its BN analogue

1.3 Contributions

By clarifying an additional link between SEM and BN, this dissertation provides the SEM practitioners with another avenue with which they can conduct their research. Furthermore, by having a BN representation of the data, practitioners can interact with the data further through the use of what-if analysis. The BN structure for this analysis can be based on the knowledge of the expert or can be mined from the data, using unsupervised learning techniques.

1.4 Dissertation Outline

The outline for the rest of the dissertation is as follows:

- **Chapter 2** sets the scene for SEM and BN, showing developments as well as areas of application for SEM.
- **Chapter 3** covers basic aspects of SEM by discussing its building blocks and how they all come together to form SEM.
- **Chapter 4** deals with graphical structure learning, which looks at how a graphical structure can be learned from data and how this structure is a probabilistic perspective on the data.

- **Chapter 5** applies PSEM and other types of BNs to a dataset on advertisement and social media.
- **Chapter 6** wraps everything up and gives conclusions.

The following items are discussed in the appendices:

- **Appendix A** describes the algorithm for MWST, using a simple step-by-step example.
- **Appendix B** provides the list of questions in the Facebook dataset
- **Appendix C** shows the SEM on the Facebook data which was done in SPSS AMOS

Chapter 2

Literature review

Structural Equation Modelling (SEM) has been used by a growing number of social scientists in the recent years. Since its inception in 1970s, SEM has been put together as a field on its own and has seen many developments [16]. This includes applying Bayesian methods to SEM [19], connecting SEM to causal structure, and integrating generalised linear models and multilevel models into SEM, to name a few [16].

In this chapter we investigate the history SEM, such as how it had developed and what it is used for. Furthermore, we look at the research fields which are applicable to this dissertation and the work that has been done so far. Section 2.1 gives a brief overview of the history of SEM, section 2.2 shows applications for SEM aside from the typical social science field and section 2.3 gives discussion regarding recent extensions in SEM, which extends SEM further into different areas.

2.1 History of SEM

Different aspects of SEM have evolved from different fields of study and their respective challenges faced: path analysis from population genetics, factor analysis from psychology and simultaneous equation models from economics [26].

Matsueda traced the origins of path analysis to genetics and biology [26]. A geneticist by the name of Sewall Wright wanted to see a causal structure in his model of bone sizes in rabbits and thus developed path analysis to achieve this task. This work, published in

1918, was the first application of its kind. Factor analysis, which is a method of finding latent variables that summarises the information of the original variables, was developed in psychology. The aim is to obtain fewer variables which explain the covariance of the original variables as a whole. Spearman is considered to be the developer of this technique [32, 26]. Lastly, simultaneous equations were necessary in economics in order to estimate parameters for demand and supply equation which feed values into each other. This development was a work of many researchers, Haavelmo and the Cowles Foundation (formerly known as the Cowles Commission) [26].

In 1970, the Conference on Structural Equation Models was held, which was attended by economists, statisticians, psychologists, sociologists and political scientists [26]. The SEM which combines path analysis, factor analysis and simultaneous equation models started to come together in this period by academics such as Hauser and Goldberger [15] and Jöreskog [17]. The next progress in SEM was made in the area of discrete outcomes, which allowed for items to be measured in dichotomous or discrete scales.

Recent development in SEM has been in a multitude of areas, such as latent growth and latent class growth models, Bayesian application to SEM, combining generalised linear models and mixed models into SEM, as well as discovering and dealing with causality within the SEM framework.

2.2 Modern applications of SEM

SEM is a very well-known and popular technique of choice in the field of social and business science. However, it has been applied successfully to other fields which shows that SEM can become an important tool for many other researchers.

Although finance, economics and accounting have not seen active use of SEM [3], there are a few researchers who see the value which SEM can add to these fields. Kohn et al. investigated the relationships between various economic variables (CPI, mortgage rate, personal income) and housing prices to study the components which lead to housing bubble in the U.S [21]. Since many of these variables are co-dependent, traditional regression analysis can encounter the problem of multicollinearity, while SEM is able to overcome the issue by capturing this dependency into the structural model [21]. Titman et al. paved the way for using SEM in the field of corporate finance in 1988 to analyse the determinants of capital structure but could not obtain convincing results [33]. Chang et al. continued this work with a different model and more data and found that growth in firms is the strongest determinant in its capital structure [3].

Golob et el. used SEM to model how households with multiple vehicles use them, based on the characteristics of the household and the vehicles as well as the drivers [12]. This was done in order to be able to predict vehicle emissions as well as provide a baseline to forecast demand for alternative fuels. Other applications of SEM in modelling travel behaviour and travel demand can be found in [11]. In the context of education, the effects of using social media for teaching purposes in tertiary education has been studied by Cao et al. using partial least squares (PLS) SEM. Their findings suggest that the use of social media "has a positive effect on student's learning outcomes and their satisfaction." [2]. The continued expansion of use in SEM makes this topic an exciting and a relevant one to study.

SEM is a graphical model consisting of directed arcs between nodes. The broader research field of graphical models provides the opportunity to explore adaptation and augmentation of SEM using the theory of graph theory

2.3 Beyond classical SEM

2.3.1 Learning structure

For this discussion, we depart momentarily from SEM and move into the field of Bayesian Network (BN), which is explained in section 4.1. Hypothesising and finding the correct structure for the measurement model as well as the structural model is what the researcher wants to achieve by using a SEM. This involves extensive knowledge regarding the data on hand as well as the broader field from which the data is acquired. Once a suitable structure is given, the researcher can use softwares to estimate the parameters in the model as well as the model goodness-of-fit to determine whether the hypothesised structure is truly reflected by the data.

With a BN, there is no defined measure of goodness-of-fit, unlike a SEM. However, both can be represented visually using a diagram. In [34], Xu et al. have drawn up simple

BN structures and changed it into SEM by adding disturbance terms and evaluated the goodness-of-fit in a Bayesian approach to select the model with the best fit. This is slightly different from the aim of our research, where we investigate the possibility of finding networks for SEM using BN. The topic of graphical model structure learning is covered in chapter 4. For a more comprehensive discussion of BN and graphical structure learning, see [28].

2.3.2 Probabilistic approach

When parameters are estimated in a SEM, their point estimates are obtained along with the parameter variance, which is same as obtaining estimates in a regression context. However, unlike a regression analysis where the model is often used for prediction, SEM usually stops at the parameter estimation stage, for hypothesis testing.

Conrady et al. makes use of Bayesian Network to construct a graphical structure similar to that of a SEM, where the observed variables are connected to “factors” (or latent variables)and the “factors” are connected to each other [4]. These methods are explained in section 4.4. By changing the network into a BN the network can be used to perform inference, which is discussed in section 4.4.5. In a frequentistic approach, parameters and their variances are estimated for building confidence intervals and significance testing. In a Bayesian approach the parameter can belong to a number of states with certain probabilities, and these probabilities change according to instantiation of different variables in the network. In Bayesian perspective, a credibility interval exists, which is similar to the confidence interval in a frequentist approach [16]. For more details regarding the field of BN, [6], [22] and [24] are a few of many literatures available, which provide a complete tool-kit for BN.

2.3.3 Bayesian SEM

With improving technology and computation power, many statistical techniques have been developed and improved in the last few decades: Bayesian Statistics is no exception. At its simplest, Bayesian estimation comes down to rearranging conditional probability formula and calculating the probability of finding the variable of interest in a certain state (known as the posterior distribution), given the data on hand. An example is given in section 4.4.5.

The estimation of parameters is done by Markov chain Monte Carlo (MCMC) sampling technique. The idea is to "instead of attempting to analytically solve for the moments and quantiles of the posterior distribution, MCMC instead draws specially constructed samples from the posterior distribution of the model parameters" [16]. The sampling process is repeated for many iterations, until the range of values around which the parameters fluctuate stabilises. Once this stabilisation occurs, the necessary statistics, such as the posterior means and standard deviations are calculated, based on the stable set of values [16]. Bayesian SEM comes with its own metrics for goodness-of-fit, such as posterior predictive checks and deviance information criterion [16]. It is also one way of dealing with missing values in SEM [19].

Bayesian SEM, as can be seen, has to do with estimation of SEM parameters by following a Bayesian perspective. This is quite different from what this dissertation aims to do, since the work here is more about graphically connecting SEM to a BN.

2.4 Summary

In this chapter, we had a brief look at some of the current research around SEM and BN and how it differs from what this dissertation is set to investigate. In the next chapter, we go to the basic building blocks of SEM and establish what type of information is gathered by performing a SEM.

Chapter 3

Classical Structural Equation Modelling

SEM is a multidisciplinary field which has been developed by combining elements from Genetics, Psychology and Economics. It is a model which consists of both observed and unobserved variables and depicts the relationship graphically to make the interpretation easier. The graphical depiction of the model which uses arrows to connect variables is known as a path diagram. In this chapter we provide an operational overview of SEM, such as why it was developed and what it is used for. The development of a SEM is a systematic process which consists of several elements. We cover the basic elements of SEM in isolation to see what purpose they serve when used as a part in SEM.

Section 3.1 discusses Factor Analysis (FA), Path Analysis (PA), simultaneous-equation models, Confirmatory Factor Analysis (CFA) and SEM. Section 3.2 looks at how to determine whether the developed model is adequate and in Section 3.4 we wrap up the contents of the chapter.

3.1 Elements of SEM

SEM is often used in the field of social science, business and marketing [28] to model abstract concepts such as intelligence, attitude, and inclination. Typically an observable dataset consists of a test or a survey that measures different aspects of an unobservable concept. Although we cannot directly measure someone's inclination towards becoming an entrepreneur we can be fairly confident that a person is inclined to become an entrepreneur if we know that the respondent strongly agrees to being his/her own boss and taking a lot of risk, when given a survey. The SEM practitioner then performs exploratory factor analysis to determine the appropriate number of factors to include in the model. It takes sufficient knowledge of the data to understand what each factor represents as well as to establish the path between the factors. The model is then fitted and evaluated and the practitioner can interpret the result. A schematic representation of the overall process of conducting a SEM is shown in figure 3.1.

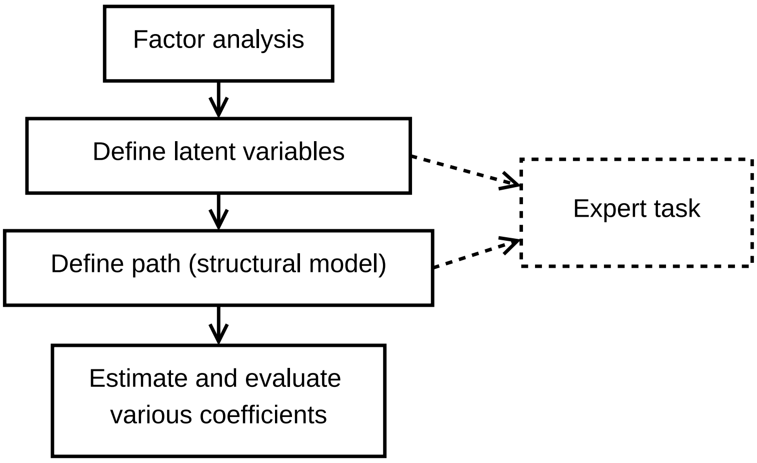

Factor analysis
Define latent variables
Expert task
Define path (structural model)
Estimate and evaluate various coefficients

Figure 3.1: Schematic representation of modelling process.

3.1.1 Factor Analysis

Factor analysis is a statistical technique that allows one to derive latent (unobservable) variables from manifest or indicator (observed) variables. This is done by measuring the correlation or covariance matrix of the manifest variables to see if certain sets of variables tend to move together. Once this is confirmed, one can reason that a variable which is not measured in the data (latent variable) can cause a number of variables to all increase or decrease. These latent variables are also known as factors, hence it is called factor analysis.

3.1.1.1 Latent Variables

Typically in Statistical modelling, one would work with data that was directly observable that could be measured in some units, such as currency, time or any other quantity. With factor analysis, we aim to infer the presence and effect of variables which are unobservable or not directly measurable. These unobservable variables are known as latent variables and they can be used to explain the covariation of manifest variables [30]. Although it is not covered here, it is perhaps worth mentioning that factor analysis can be used as a dimension reduction technique [16]. If there are many observed variables that are highly correlated, they can possibly be summarised into a few factors and make the model parsimonious and handle multicollinearity at the same time. This technique is known as factor analysis regression [23].

3.1.1.2 Exploratory Factor Analysis

There are two kinds of factor analysis - exploratory factor analysis and confirmatory factor analysis [16]. The latter will be elaborated later in the chapter, for now we will focus on the former. As the word exploratory implies, this kind of factor analysis is used to explore the data: one would not know how many factors are appropriate for the data and would thus look for the suitable number of factors to include in the model. This will typically not form a part of a SEM, as one of the theories tested in a SEM is whether the factors are well represented their manifest variables in the confirmatory factor analysis. Hence optimising the model based on the results of EFA and moving on to SEM will artificially improve the results. Nevertheless, the researcher can use EFA to get a better idea of the data on hand.

Typically, one would hypothesize that the factors are the underlying causes of the variations in manifest (indicator) variables. This can be seen in Figure 3.2 as the arrows are drawn from ovals to rectangles. This would be the case when the manifest variables are highly correlated and are called reflective indicators [14]. For example, a person who is generally clean (factor) is expected to have a clean house, dress neatly and be hygienic (indicator variables). Of course, the reverse relationship is also possible where the indicator variables determine the factor. In such cases the indicator variables do not necessarily have to be highly correlated. For example, being physically healthy (factor)

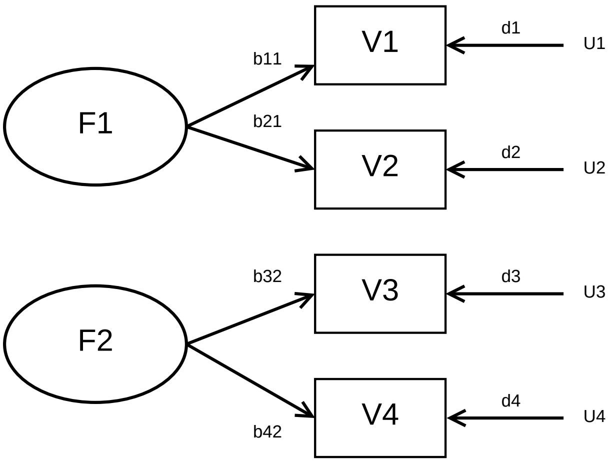


Figure 3.2: 4 variable, 2-factor model.

is a product of a number of conditions, such as exercising often, getting sufficient sleep and limiting consumption of high calorie food products (indicator variables). In this case the manifest variables are called formative indicators [14].

It is convention to represent latent variables in ovals and manifest variables in rectangles. F_1 and F_2 are known as common factors, since they have influence over more than one manifest variable. $V_1 - V_4$ are the manifest variables and $U_1 - U_4$ are called unique factors. These unique factors capture all the influence to a single manifest variable, which is not captured by the common factors and can include disturbance terms.

It can be seen that the arrows for common factors are accompanied by the coefficients such as $b_{11}, b_{21}, \ldots$. These coefficients are known as factor loadings and the two numbers in b_{xy} indicate "the factor loading from factor F_y to variable V_x". For example, b_{32} indicates that this is the value of factor loading from F_2 to V_3 [30]. When one assumes that the factors are orthogonal, or uncorrelated, these coefficients can be seen as standardized regression coefficients (all variables have variance of 1), correlation coefficients (since the factors, which are the predictor variables, are standardized and orthogonal) and path coefficients.

3.1.2 Path Analysis

Path analysis offers the graphical and causal aspect to SEM. One could view path analysis as SEM with only manifest variables. Its first application can be traced to Sewall Wright in 1918 [26]. Path analysis can be thought of as an extension of multiple regression and can be estimated by using ordinary least squares (OLS), maximum likelihood (ML) or Two-Stage Least Squares (2SLS), which is a more developed version of OLS [9].

In figure 3.3, arrows from one variable to another indicate the causal paths. Naturally, the direction of causation is from the variable at the base of the arrow to the tip of it. For example, in figure 3.3 we can say that A causes D while B also causes D via C. Any variable with the tip of the arrow towards it is called an endogenous variable, while any variable with no arrow pointing towards it (with the exception of error terms) is called an exogenous variable. A and B are exogenous, while C and D are endogenous. The double-headed arrow indicates correlation or covariance between exogenous variable and we can see that there A and B covary in figure 3.3.

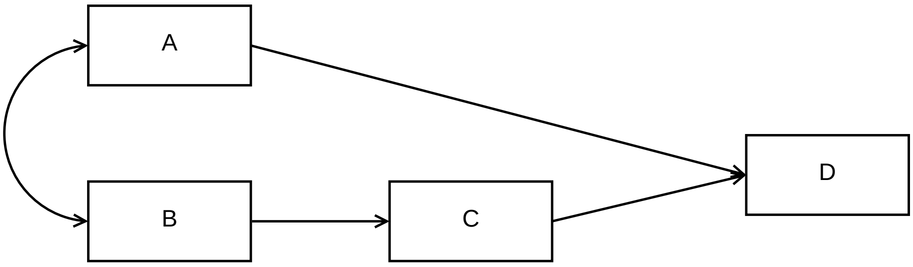


Figure 3.3: Recursive path diagram

The path from A to D is known as a direct path, while the path from B to D is an indirect path, since an intervening endogenous variable is in the path [9]. A path coefficient can be determined for each of these paths, which is a standardized regression coefficient that shows the direct effect of the explanatory variable on the response variable [9]. Note that in figure 3.3 there are no loops created by single-headed arrows. This type of model is referred to as a recursive model [30]. In contrast, there is a cycle in figure 3.4, where C feeds into A, A feeds into D and D feeds into C. These models are referred to as non-recursive models and working with them poses additional challenges.

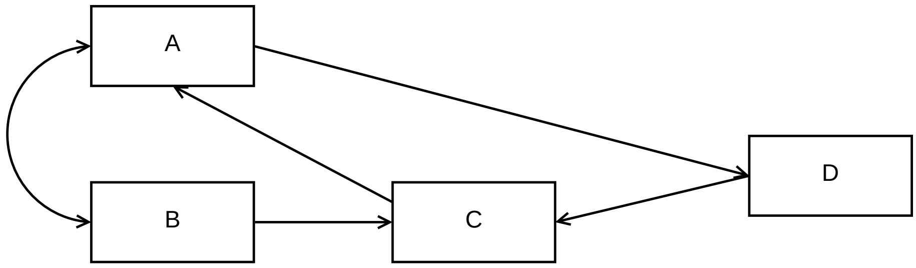


Figure 3.4: Non-recursive path diagram

3.1.3 Simultaneous-Equation Models

While path analysis offers a way to visualise the model, simultaneous-equation model offers the technical capability of parameter estimation. Often in SEM, identification is a crucial topic, which is defined as ”going to the known information to the unknown parameters.” A model with p variables has $p(p+1)/2$ known information. The unknown parameters here include all parameters in the model, such as structural coefficients, variances and covariances [16]. Identification has to do with whether a parameter or a coefficient can be determined. When there are more variables than equations, the system of equations are under-identified and a unique solution cannot be obtained. When there are same number of equations as the variables, the system is just-identified and it will always result in the same solution. When there are more number of equations than variables, the system is over-identified and typically this is the preferred case, since confirmatory tests can be performed the answers for parameters [20].

When working with recursive models, the model will always be identified and specifically over-identified as long as not all variables are linked to every other variables by a single or a double-headed arrow [30]. Figure 3.5 is an altered form of figure 3.3 so that all variables are connected. There are 5 path coefficients and 1 covariance parameter to be calculated. Additionally, the variance of each exogenous variable as well as error variance of each endogenous variable must be obtained, hence there are 10 parameters to be calculated in this diagram. With 4 variables, the number of known information is $4(4+1)/2 = 10$, therefore there are as many unknowns as the knowns and this model is just-identified. Typically we restrict certain paths to have a coefficient of 0, effectively eliminating the path from the diagram and then the model would be over-identified.

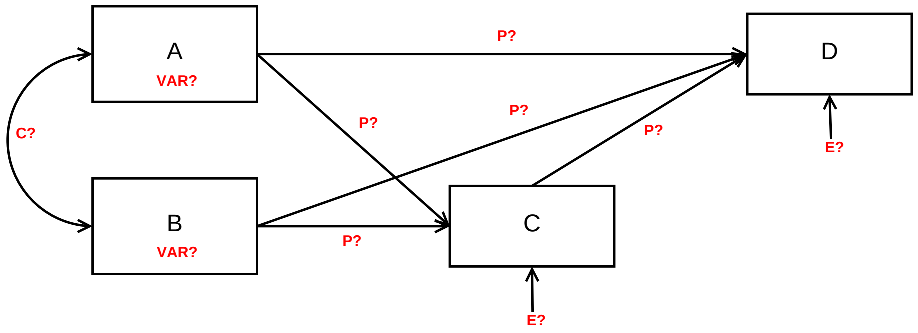


Figure 3.5: Just-identified path diagram

Following the notations of Murphy [28], equation 3.1 represents how a SEM is defined

$$x_i = \mu_i + \sum_{j \neq i} w_{ij} x_j + \epsilon_i \tag{3.1}$$

where x are the variables in the model, μ are mean of each variable, w_{ij} is the coefficient from variable$j$ to i and ϵ are measurement error for each variable, where $\epsilon \sim N(\mathbf{0}, \mathbf{\Psi})$. This model can be rewritten in matrix form as follows

$$\begin{aligned} \mathbf{x} &= \mathbf{Wx} + \boldsymbol{\mu} + \boldsymbol{\epsilon} \\ \mathbf{x} - \mathbf{Wx} &= \boldsymbol{\mu} + \boldsymbol{\epsilon} \\ \mathbf{x}(\mathbf{I} - \mathbf{W}) &= \boldsymbol{\mu} + \boldsymbol{\epsilon} \\ \mathbf{x} &= (\mathbf{I} - \mathbf{W})^{-\mathbf{1}}(\boldsymbol{\mu} + \boldsymbol{\epsilon}) \end{aligned} \tag{3.2}$$

The joint distribution is then given by $p(\mathbf{x}) = N(\boldsymbol{\mu}, \sum)$ where

$$\sum = (\mathbf{I} - \mathbf{W})^{-\mathbf{1}} \mathbf{\Psi} (\mathbf{I} - \mathbf{W})^{-\mathbf{T}} \tag{3.3}$$

where $(\mathbf{I} - \mathbf{W})^{-\mathbf{T}}$ is the transpose of $(\mathbf{I} - \mathbf{W})^{-\mathbf{1}}$

This model implied variance-covariance matrix $\sum$ is then compared with the sample covariance matrix $\mathbf{S}$ and the aim is to reduce the difference between the two matrices as much as possible. A number of techniques, including but not limited to, maximum likelihood, least squares estimation and the Bayesian method, exist to perform this task [16].

3.1.4 Confirmatory Factor Analysis

A SEM consists of two parts: measurement model and structural model. The measurement model looks at the relationship between the latent variables and their manifest variables while the structural model looks at the relationship between the latent variables[16]. In conducting a CFA, we are interested in seeing whether the the observed variables represent the latent variables sufficiently well, before moving on to test the relationships between the latent variables in SEM.

In figure 3.6, which is similar to figure 3.2 the two factors are assumed to covary. When performing a CFA, covariance is estimated for each pair of factors. Furthermore, the variance of each error term E must be calculated and the significance of the factor loadings, variances and covariances are tested at the evaluation step.

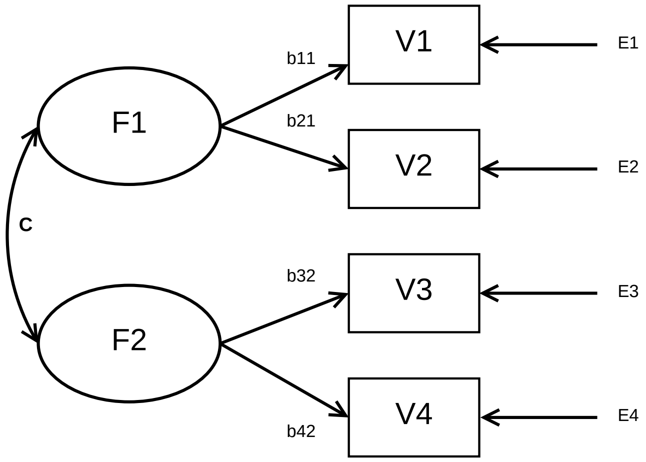


Figure 3.6: Confirmatory factor analysis

3.1.5 Structural Equation modeling

Once a satisfactory measurement model has been obtained, the structural model is estimated and evaluated. While the latent variables were all simply correlated to each other in CFA, they are dependent on each other in SEM, meaning that their relationships are assumed to have a certain direction [30]. These directional assumptions should come from extensive knowledge in the data and the field of study. Ultimately, a researcher performs SEM to fit the measurement and structural model to the data and obtain a good fit so that it can be used for interpretation [30].

It is perhaps worth mentioning that there are two kinds of SEM in use: covariance-based "traditional" SEM and variance-based Partial Least Squares (PLS) SEM. In PLS, the aim is to maximise the variance of dependent variables explained by the independent variables. One instance where it is helpful to use PLS is where latent variables have many indicator variables [14].

Figure 3.7 shows a very simple SEM which is developed from figure 3.6. Typically the latent variables are arranged so that the direction of influence is from left to right. Also, it is advised that each factor gets at least 3 indicator variables. Here, only two manifest variables are used for compact illustration purposes. In the estimation stage of SEM, the measurement models are once again validated, the manifest endogenous variances (denoted by E) as well as the latent endogenous variances (denoted by D) is estimated and evaluated. Additionally, all structural paths (directional paths between latent variables) are estimated and evaluated.

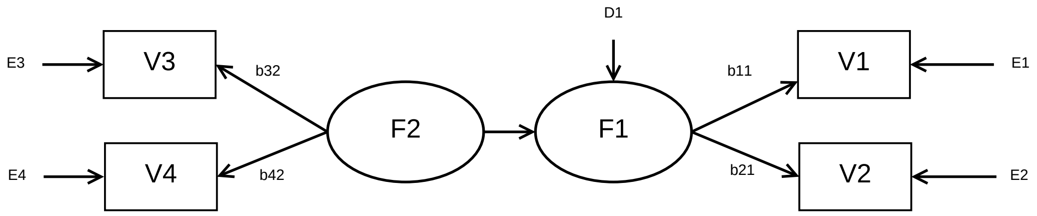


Figure 3.7: Structural equation model

3.2 Evaluation

When working with models with latent variables in social sciences, their reliability is typically confirmed using average variance explained (AVE) and composite reliability (CR). Reliability is a measure which quantifies the percentage of variance in an observed variable explained by the latent variable [30]. If there are little measurement errors, the reliability coefficient will be high. The minimum recommended values are 0.5 and 0.7 for AVE and CR, respectively.

With both CFA and SEM, the significance of factor loading in the measurement model should be tested using the t-statistic. One would hope to see values higher than 1.96 in absolute value for significance at 5% or higher than 2.58 in absolute value for significance at 1% [30].

There are many indices to measure model goodness-of-fit for SEMs. Here we present the most frequently reported few.

3.2.1 Goodness-of-fit indices

Chi-square test tests for significance between actual covariance matrix and estimated covariance matrix. Null hypothesis assumed no significant difference between the two. This is rarely met because of sample size sensitivity (small difference is seen as significant in large samples). It also requires the condition of multivariate normality, hence it is no longer seen as viable option [30].

Three types of indices are reported frequently: absolute fit index, incremental fit index and parsimony-adjusted index. Absolute index takes on values between 0 and 1 and it can be thought of as an R^2. But instead of measuring how much variance is explained by the model, it measures how much the variance-covariance matrices correspond to each other [20]. Naturally, values closer to 1 are preferred. Incremental fit index shows how the model has improved relative to the "baseline model" which essentially assumes the value of 0 for covariances [20]. Parsimony index, as the name suggests, allows us to identify the simpler model among the available models. If all models yield satisfactory results that are of similar level, parsimony index prefers the simplest model[20].

Bentler's CFI is a popular incremental index, where a value of > 0.94 suggests good fit, although values between 0.9 and 0.94 are sufficient. Root mean square error of approximation (RMSEA) is a good choice for parsimony index where values < 0.09 are considered good, and values < 0.055 are actually more ideal. The 90% confidence interval for RMSEA is often accompanied in the report and an interval between 0.09 and 0 is adequate, though an interval between 0.054 and 0 preferred. Standardized root mean square residual (SRMR) is more preferred as absolute index over chi-squared. Although it is an absolute index, the value should be interpreted in reverse, that is, the closer the SRMR is to 0, the better. This should be intuitive since a model which obtains a good fit should have correspondingly low value of residuals. The ideal threshold values are thus same as that of RMSEA [16].

In evaluating the parameter significance, bootstrapping can be used as a non-parametric measure. This is done by sampling random observations with replacement to obtain a new sample or dataset. A model can be fitted and parameter estimation can take place, and all these values can be stored. This can take place many times, say 500 times, each time resulting in different values. Then an empirical distribution can be constructed for each collected parameter and one can verify whether a value of 0 lies within the empirical distribution. If 0 is not in the interval, a conclusion can be made that the parameter is significant [16].

3.3 Definitions

Table 3.1: Definitions for SEM related terminologies

Latent variable	Variable which is not directly observable, also called a factor or construct social network
Manifest variable	Observable variable which can be directly measured, also called indicator variable
Measurement model	The model which depicts the relationships between the factors and their manifest variables
Structural model	The model which depicts the relationships between and among the factors

3.4 Summary

In this chapter we introduced SEM by going through its operational steps. EFA is used to find a suitable number of factors, path analysis and simultaneous-equation models are used to graphically represent and mathematically estimate the parameters in the model, CFA is used to validate the measurement model and SEM is used to validate the structural model. The model is then evaluated using a number of indices and if it survives a disapproval it is seen as a valid model. Appendix C shows an example of a SEM conducted using SPSS AMOS. In the next chapter we take a look at graphical model structure learning, a perspective which is based different ideas but aims to achieve the same goal of finding the right model.

Chapter 4

Graphical Model Structure Learning

Structural equation modelling is often carried out by first building a model based on a set of assumptions or opinions from the researcher or an expert about the structure. The model is then tested against various criteria, such as model goodness-of-fit and parameter significance. If no satisfactory model has been obtained, structural changes will be made with the expectation that the alternative model will perform better. But without any assumptions or knowledge of the field, one may not be able to find a good starting point for creating the model. Additionally, it may be useful to have a tool to validate whether the model which was constructed by the expert could have been obtained by another method. In this chapter, we cover the topic of graphical model structure learning, which provides us with methods to address the two shortcomings that have been mentioned.

A SEM is a graphical model, more specifically, a directed acyclic graph when it has no loops or cycles. It is possible for them to contain a cycle, which can be interpreted as a feedback loop [28]. Therefore, if we want to investigate structural learning for SEM from the data-mining perspective, we can turn to the field of graph theory for valuable insight.

Graphical models offer a condensed visualisation of complicated models by making use of nodes and arcs (which are used to connect nodes) to show how variables interact with each other. An important aspect of using a graphical model is to obtain a meaningful structure which explains the connection as well as the direction of influence between variables. Structure learning involves a data driven process to find the optimal network and does not rely on expert knowledge. This is not an easy task, as there are two major issues with regards to learning structures of a graphical model: dealing with immense number of possible networks and differentiating among equivalence classes of network structures [24].

Section 4.1 covers the basics of Bayesian Networks which is a form of a graphical model. The challenges in learning the structure are covered in section 4.2 and the solutions to those challenges are discussed in 4.3. The application of the ideas are done in section 4.4 using a dataset on survey data for perfumes and the chapter is concluded in section 4.5.

4.1 Bayesian Networks

Before we address challenges with structure learning, we introduce basic graphical model terminology, as well as Bayesian network and related concepts.

4.1.1 Definitions

Table 4.1: Definitions for BN related terminologies

Tree	a graph with no cycles or loops
Node	a visual representation of a single variable
Edge	a line or an arrow connecting two nodes, also referred to as an arc
DAG	directed acyclic graph. A visual representation of a network which uses arrows and does not contain a loop. A recursive model in SEM terminology
Parent node	nodes which have the base of the arrow attached to it. In figure 4.4c, node B is the parent node of A and C
Child node	nodes which have the tip of the arrow attached to it. In figure 4.4b, node B is the child node of A and node C is the child node of B
Root node	a node with no parent nodes
Leaf node	a node with no children nodes
Instantiate	to set evidence onto a node

4.1.2 Introduction

A Bayesian Network (BN) is a graphical model with nodes as variables and arcs (arrows or edges) representing causal connections [22]. Each node has a number of states, which are the possible values the variable can take on. The field of graphical models was developed from graph theory and probability theory [24], and accordingly a BN consists of two parts: graphical part which indicates local dependencies through the use of arcs and probabilistic part which shows the relationship between the two connected nodes [5]. The graphical part, which is the structure of the BN, is also called the directed acyclic graph (DAG)[24].

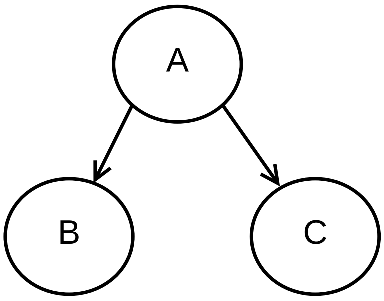


Figure 4.1: A simple DAG

Figure 4.1 is an example of a DAG. Once this is accompanied by a probability table for each node, it becomes a BN. Because node A is a parent node (it does not depend on any other node) the probability table for A will be marginal, meaning that it will only consist of probabilities of observing its states. Nodes B and C are child nodes, since they depend on node A, so the probability of their states will depend on what the state of node A is. Hence their probability tables will be conditional.

A simple illustration of how to use a BN is given in section 4.4.5.

It is possible for independent events to become dependent given new evidence and vice versa. Mathematically, event α is conditionally independent of event β given event γ if and only if $Pr(\alpha|\beta, \gamma) = Pr(\alpha|\gamma)$ [6]. In other words, when event γ is known, event β does not add any more information. This idea of conditional independence is what makes BNs computationally feasible and is further explored in section 4.1.3.

4.1.3 *d*-separation

A chain of nodes can be categorised into 3 groups: chain, fork or v-structure [28].

- chain: The arcs flow in the same direction ($X \rightarrow Y \rightarrow Z$)
- fork: the arcs diverge from a node ($X \leftarrow Y \rightarrow Z$), as shown in figure 4.1
- v-structure: the arcs converge into a node ($X \rightarrow Y \leftarrow Z$)

The nodes X and Z are conditionally independent given that node Y contains evidence (node Y has been set to a specific state), in a chain and a fork. In other words, information cannot flow between X and Z if Y has been set to a particular state. However, X and Z are conditionally independent in a v-structure if neither Y nor any of its descendants contain evidence. This principle is known as d-separation [28]. Of course, multiple nodes can be entered in the place of X, Y and Z in the above example. One must simply find all paths leading from X to Z to confirm whether they are d-separated. When checking for d-separation in a DAG structure, only the directions of arcs for v-structures must be kept- chains and forks can be plain edges without an arrowhead and this can make the graphical verification easier. Building onto this idea, a set of nodes that d-separates a node m from all other nodes in the network is known as m's Markov blanket[28]. Figure 4.2 shows the markov blanket (nodes with red dotted border) for the node $Var3$.

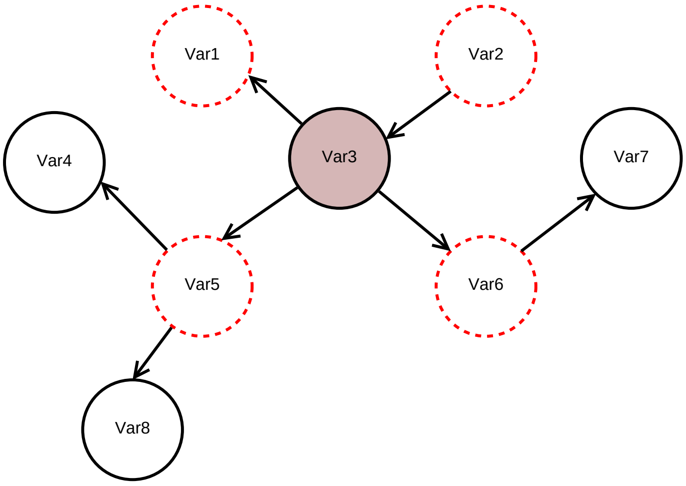


Figure 4.2: Markov blanket of $Var3$

4.2 Challenges with structure learning

4.2.1 Search Space

The first problem encountered in learning the structure of graphical models is that the possible number of graphs increases super exponentially as the number of nodes increases.

The formula is given by a recursive formula [29]:

$$N(d) = \sum_{i=1}^{d} (-1)^{i+1} \binom{d}{i} 2^{i(d-i)} N(d-i) \tag{4.1}$$

where d = number of nodes and $N(d)$= possible number of networks based on d nodes. Table 4.2 shows the values of $N(d)$ for values of $d = 0$ to 7.

Table 4.2: Possible number of networks for different values of d.

d	$N(d)$
0	1
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265

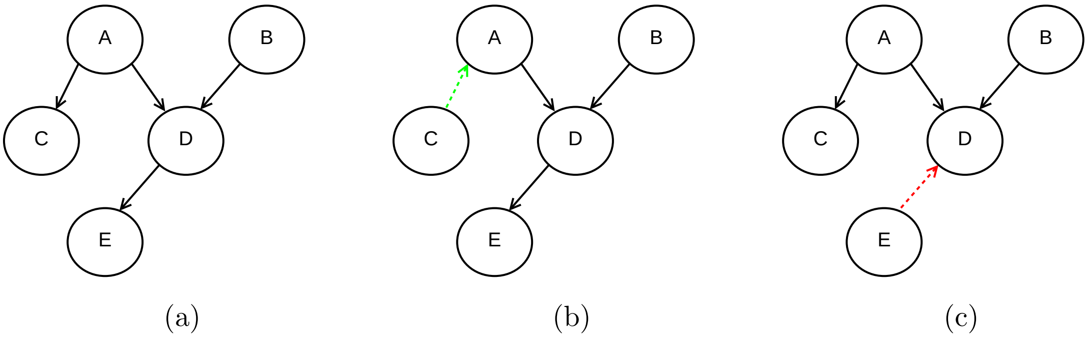
With so many possible networks, it would be impossible to store all the different structures to be analysed, let alone go through all possibilities to obtain the global optimum [28].

4.2.2 Markov equivalence

Two graphs which are Markov equivalent contain the same conditional independence assumptions. This means the graphs which are Markov equivalent belong to the same Markov equivalence class and this happens when the undirected skeleton and the v-structures of the graphs are the same. Put differently, "different graphs that share exactly the same d-separation properties are said to be *Markov equivalent*" [24]. In figure 4.3, we have 3 DAGs which would look identical if their edges were to be undirected. However, the edges going from A to D and B to D should remain as directed, since they represent what is known as a v-structure, where two arrows are going into a single node. In such a case we can say that A and B are independent, but conditionally dependent given their common child node, D.

Figure 4.3a and figure 4.3b belong to the same equivalence class. Even though the direction of the edge from A to C has changed, it did not result in any creation of a new v-structure, hence contains the same conditional independence information as figure 4.3a. However, we cannot say the same for figure 4.3c as this has created new v-structures, $A \rightarrow D \leftarrow E$ and $B \rightarrow D \leftarrow E$.

Therefore, while we will be able to distinguish between figures 4.3a and 4.3b with figure 4.3c in terms of scores, we will not be able to distinguish between figure 4.3a and 4.3b. This is because "Bayesian networks belonging to the same equivalence class have the same scores" [27]. Hence we say that the structure of the DAG can be learned "up to Markov equivalence" [28].



(a) (b) (c)

Figure 4.3: a and b are Markov equivalent DAGs while c is not

In the next section, we introduce algorithms for structure learning, that can be used to overcome these challenges.

4.3 Structure learning solutions

4.3.1 Maximum Weight Spanning Tree (MWST)

In order to overcome the challenge of finding the proverbial needle in a haystack, a clever algorithm can be used to find a tree (a graph without cycles) structure for the network which restricts the number of parents for each node to 1, making the search much simpler.

This algorithm is known as the Chow-Liu algorithm or Kruskal's algorithm [24].

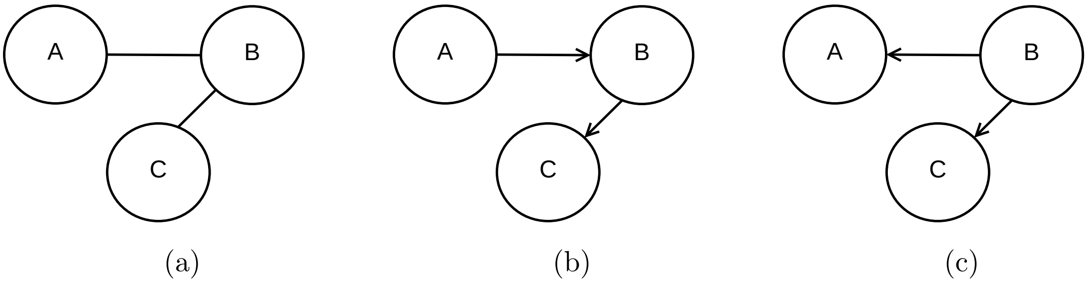


Figure 4.4: Directed and undirected trees.

Note that a tree may be directed or undirected, as shown in figure 4.4. Undirected trees use lines to show connections between different nodes and as such do not assume any directional influence, while directed trees use arrows to show probabilistic relationships in the direction of the arrow. Naturally, structures typically found in measurement models and structural models are directed, since directional relationships are displayed between a factor and its indicator variables and among factors using arrows. However, when SEM as a whole is viewed, exogenous latent variables are allowed to covary, hence it is a partially directed acyclic graph (PDAG). Murphy [28] defines the joint distribution of a directed tree with a single root node r as follows:

$$p(\mathbf{x}|T) = \prod_{t \in V} p(x_t|x_{pa(t)}) \tag{4.2}$$

which simply means that the joint distribution of a directed tree is simply the product of probability of all V nodes x_t, given their respective parent nodes $x_{pa(t)}$. In figure 4.4 b and c, we have

$$\begin{aligned} p(x_a, x_b, x_c|T) &= p(x_a)p(x_b|x_a)p(x_c|x_b) \\ &= p(x_b)p(x_a|x_b)p(x_c|x_b) \end{aligned} \tag{4.3}$$

where the first expansion is for figure 4.4b and the second expansion is for figure 4.4c. The same can be represented for an undirected tree as follows:

$$p(\mathbf{x}|T) = \prod_{t \in V} p(x_t) \prod_{(s,t) \in E} \frac{p(x_s, x_t)}{p(x_s)p(x_t)} \tag{4.4}$$

where (s, t) denotes two nodes which are connected by an edge, and E number of edges are present in the network T. In figure 4.4 a, we have

$$\begin{aligned} p(x_a, x_b, x_c|T) &= p(x_a)p(x_b)p(x_c)\tfrac{p(x_a,x_b)}{p(x_a)p(x_b)}\tfrac{p(x_b,x_c)}{p(x_b)p(x_c)} \\ &= p(x_a, x_b)\tfrac{p(x_b,x_c)}{p(x_b)} \\ &= p(x_a)p(x_b|x_a)p(x_c|x_b) \\ &= p(x_b)p(x_a|x_b)p(x_c|x_b) \end{aligned} \tag{4.5}$$

hence they are all equivalent due to the fundamental rule for probability calculus. Using an undirected tree is preferred for structure learning, because it is symmetric and gives a more general expression. We can turn equation 4.4 into a log-likelihood for a tree as follows:

$$\log p(\mathcal{D}|\theta, T) = \sum_t\sum_k N_{tk}\log p(x_t = k|\theta) + \sum_{s,t}\sum_{j,k} N_{stjk}\log\tfrac{p(x_s=j,x_t=k|\theta)}{p(x_s=j|\theta)p(x_t=k|\theta)} \tag{4.6}$$

where $\mathcal{D}$ is the data matrix, N_{tk} is the number of times node t is in state k and N_{stjk} is the number of times nodes s and t are in state j and k, respectively.

These values can be written in terms of the observed probability, or the empirical probability distribution: $N_{stjk} = N\hat{p}(x_s = j, x_t = k)$ and $N_{tk} = N\hat{p}(x_t = k)$. Setting θ to its maximum likelihood estimates, this becomes

$$\tfrac{\log p(\mathcal{D}|\theta,T)}{N} = \sum_{t\in V}\sum_k \hat{p}(x_t = k)\log\hat{p}(x_t = k) + \sum_{(s,t)\in\varepsilon(T)} MI(x_s, x_t|\hat{\theta}_{st}) \tag{4.7}$$

where $MI(x_s, x_t|\hat{\theta}_{st}) \geq 0$ is the mutual information between x_s and x_t given the empirical distribution:

$$MI(x_s, x_t|\hat{\theta}_{st}) = \sum_j\sum_k \hat{p}(x_s = j, x_t = k)\log\frac{\hat{p}(x_s = j, x_t = k)}{\hat{p}(x_s = j)\hat{p}(x_t = k)} \tag{4.8}$$

Which is none other than the Kullback-Leibler divergence between $p(x_s = j, x_t = k)$ and $p(x_s = j)p(x_t = k)$ that is used to measure mutual information between x_s and x_t [24]. Another way to interpret the Kullback-Leibler divergence is that it measures dissimilarity between two distributions and the smaller this value, the closer the two distributions, with a value of 0 indicating that the two distributions are identical [28]. Using the mutual information between every pair of nodes, the maximum weight spanning tree is created. The algorithm is given in algorithm 4.1

```
for $i = 1$ to $d(d-1)/2$ do ($d$=number of nodes)
        Calculate the Kullback-Leibler divergence for nodes $(i, j)$,
        where $i \neq j$ , $(i, j) \in d(d-1)/2$ (each edge between every pair of nodes)
        Denote these as $b_1, b_2, b_3, ..., b_{d(d-1)/2}$ respectively and store them into a vector $b$.
end for
Sort the vector $b$ largest to smallest
Create a path into the graph structure denoted as the first element in the vector $b$
for $i = 2$ to $d(d-1)/2$ do
        if no path exists between node $i, j$
        Create a path denoted as the $i$th element in the vector $b$
end for
```

Algorithm 4.1: Maximum weight spanning tree algorithm

Once a structure has been obtained we can induce small changes to the network in order to find the local maximum. These small changes are defined as adding, removing or reversing an edge to the network, while making sure that no cycles are created by these changes [6]. It is not guaranteed that a better network will be found after all the local modifications have been performed, in which case the initial network is considered the best network. Alternatively, one can use techniques such as random restarts [6] which starts off the process of local modification with a different initial network (possibly by adding a small noise to the data) to see if many different starting points lead to the same or similar local optimal network.

The main disadvantage of using MWST algorithm is its tree structure. That is, it does not allow networks with more than one parent node per child node. Although this is precisely the reason why MWST algorithm is very fast, it implies that the resulting network cannot arrive at complex scenarios where multiple causes are expected to influenced a single variable.

4.3.2 EQ algorithm

In section 4.2.2, it was noted that networks within the same equivalence class cannot be distinguished, while networks of different equivalence class can be distinguished. We can use this fact to narrow our search space, which entails exploring the space of equivalence classes of BN as opposed to the entire space of BN [27]. Essentially, we are searching for a group of networks that belong to the same family, instead of directly looking for the single best network. By doing so, we can bypass many redundant calculations and only go into detailed modifications once we have obtained the optimum equivalence class. After the optimum equivalence class has been found, we can search for the local optimum by adding or removing directed or undirected edges [27]. The EQ algorithm can be used alongside the MWST algorithm, on a fully unconnected dataset. If both algorithms return the same network, it is quite possible that the network obtained is the optimal network.

4.4 Application

Here we see how graphical model structure learning can be applied to a data for exploration. This is carried out in 4 stages, as proposed by BayesiaLab: Unsupervised Learning (manifest variables), Variable Clustering, Multiple Clustering and Unsupervised Learning (latent variables) [5]. BayesiaLab is a software platform which employs the algorithms mentioned earlier to learn graphical structures from data. The application was thus implemented in BayesiaLab.

The dataset comes from a French market research survey on perfumes and is available from www.bayesia.us/perfume. There are 1320 observations, 40 variables which measure different aspects of the perfume (39 on 1-10 scale and 1 on 1-5 scale) and 1 variable for $Purchase\,Intent$ (measured on 1-6 scale).

Figure 4.5 shows how the data appears in BayesiaLab after it has been imported. All variables which are measured on 1-10 scale have 10 states, which can be condensed for a faster and simpler model. BaysiaLab has the function to discretise continuous variables during the import process and here, equal distance discretisation was used with 5 intervals, since they are all measured on the same scale. Because $Purchase\,Intent$ only

Intensity
Sweet
Elegant
Personality
Corresponds
Romantic
Accessible
Purchase Intent
Easy to wear
Tenacious
Pleasure
Attractive
Sensual
Chic
Fresh
Rich
Evoque quality
It will please others
Is classical
Character
Free
Flowery
Fruity
Evoque joy
Sensuality
Trust
Nice
Passionate
Greedy
Modern
Make attractive
Classical
Bold
Is Modern
In love
Spiced
Original
Universal
High-End
Fulfilled
Is feminine
Radiant
Wooded
Feminine
Trend
Timeless
Active
Optimistic

Figure 4.5: Perfume data: imported into BayesiaLab

has 6 states originally, it is classified as a discrete variable and left at 6 states. Another variable which did not change is the variable *Intensity*, which only has 5 states and is one of the 40 variables that measures the aspect of the perfume.

4.4.1 Initial network

Initially, a network structure should be found which will be used for clustering purpose. Finding the optimal network can be a challenge, as the number of possible networks increases super-exponentially according to the increases in the number of nodes. This relationship is given in equation 4.1. Since heuristic search algorithms can only find local optimum we cannot be certain that we will find the global optimum- hence it is advisable to use different learning algorithms and use cross-validation techniques.

A simple, data driven method of finding a network structure is to use MWST which was discussed in 4.3.1. Of course, using different algorithms for structure learning will most likely lead to slightly different models suggested by each algorithm. The minimum description length (MDL) for each model can be evaluated to select the best model. MDL operates under the logic that regularities within the data can be compressed, meaning certain symbols can be used to describe the data in a more compact way than the actual data. Highly regular data can therefore be highly compressed [13].

4.4.1.1 MDL

MDL is consists of two parts as stated in the following equation:

$$MDL(B, D) = \alpha DL(B) + DL(D|B) \tag{4.9}$$

where B is the model (Bayesian Network) and D is the observed data, hence $DL(B)$ is the complexity (number of bits) of the suggested model and $DL(D|B)$ is the number of bits to describe the log-likelihood of the data given (with the help of) the model, which is none other than the error [13]. Equation 4.9 states that the MDL score is the sum of the complexity of the model and the complexity of the errors. Generally, if the model is highly accurate it will need much description (as it will have many terms) but the resulting error will be small. Conversely, if the model is very simple, its description will be very short but we will need a lot of information to describe its errors. The α is a structural coefficient, or simply a weight, which we can use to reduce the impact of model complexity. Even if the model is highly complex and $DL(B)$ is high, by making α small, model complexity will have reduced impact on the overall MDL [5]. A fully unconnected network will translate to the minimum value for $DL(B)$ and a fully connected network will translate th the minimum value for $DL(D|B)$. Thus obtaining minimum MDL finds the right balance between the two extremes. If we start off with a blank network, an edge will connect two nodes only if the decrease in $DL(D|B)$ is larger than the increase in $DL(B)$ [5].

Cross-validation should be performed to confirm whether the model obtained can be improved. Section 4.3.1 mentioned random restarts where the objective was to see if the final network obtained remains the same when starting from different initial networks. Here, since the initial network is a set of fully unconnected nodes, we add noise to the data and see if the same structure and MDL are obtained. This process of data perturbation serves as a cross-validation step for finding the structure of the network.

As discussed in 4.3, we can employ MWST or EQ algorithm to learn the structure of the data. For now, we do not want purchase intent to be included in the network, since

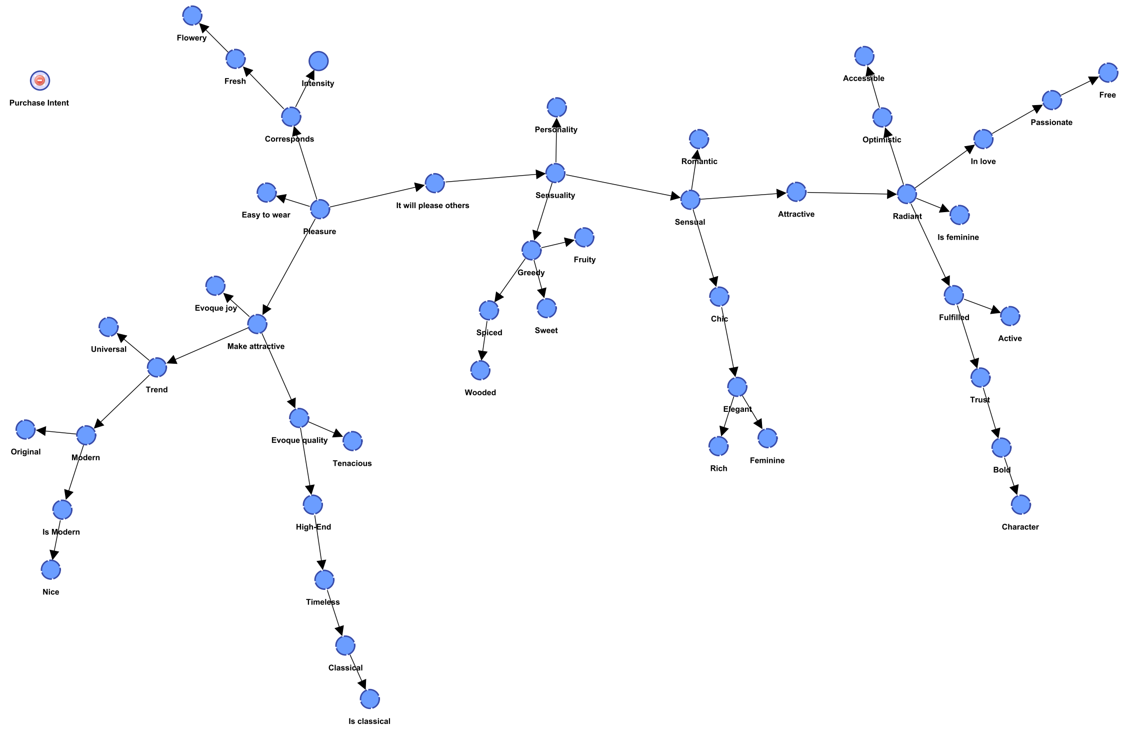


Figure 4.6: Perfume data: Initial network learned using MWST

we are interested in how it interacts with the factors. The resulting learned network is shown in figure 4.6. The MDL score of this network is 98,606.572 and this score is also obtained using the EQ algorithm. Performing data perturbation also returns this score, hence we can proceed to the next step of finding the clusters.

4.4.2 Variable Clustering

Here, the nodes within the network are clustered, based on the Arc Force (Kullback-Leibler Divergence), using hierarchical agglomerate clustering. The Arc Force is calculated for every pair of nodes, which measures how close a node is to every other. At the start of the process, the nodes are all treated as a cluster on its own and two clusters with the smallest “distance” are joined into a single cluster. This process is repeated either until a satisfactory number of clusters have been obtained or until no clusters are deemed “close enough” to be joined into one cluster. This is similar to performing

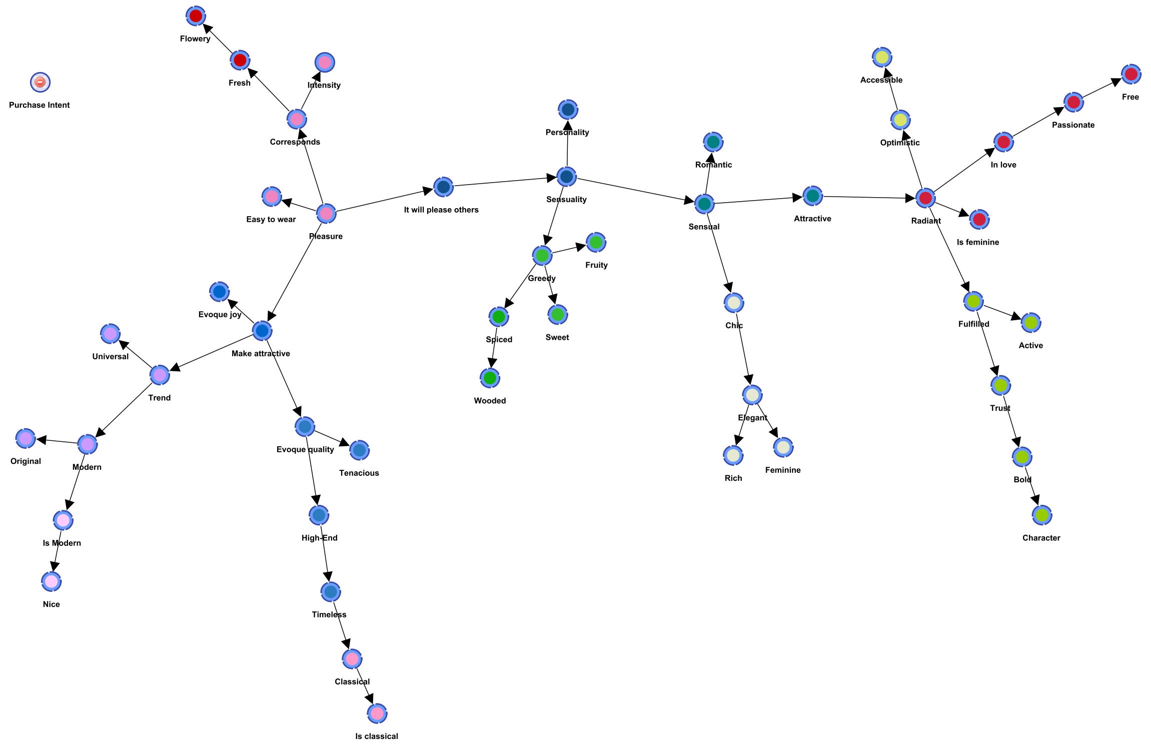


Figure 4.7: Perfume data: Clustering result, max 5 variables per cluster

an exploratory factor analysis where the researcher tries to find the optimal number of factors within the data.

Just as we performed cross-validation on the network, we can perform cross-validation to check whether the clustering groups the same nodes together frequently. To do this we would simply start off by adding small noise to the data, create the network structure, perform clustering on the network and repeat this process many times. If the results show that same nodes are clustered into the same group many times over the iterations, we can safely assume that it is indeed the most likely scenario of clustering outcome.

Pair-wise Kullback-Leibler divergence is calculated for all variables in the network and those values are used to perform hierarchical clustering, which returns the following colour-coded network. By imposing the restriction of maximum 5 variables per cluster, we obtain a 15-cluster solution which is shown in figure 4.7.

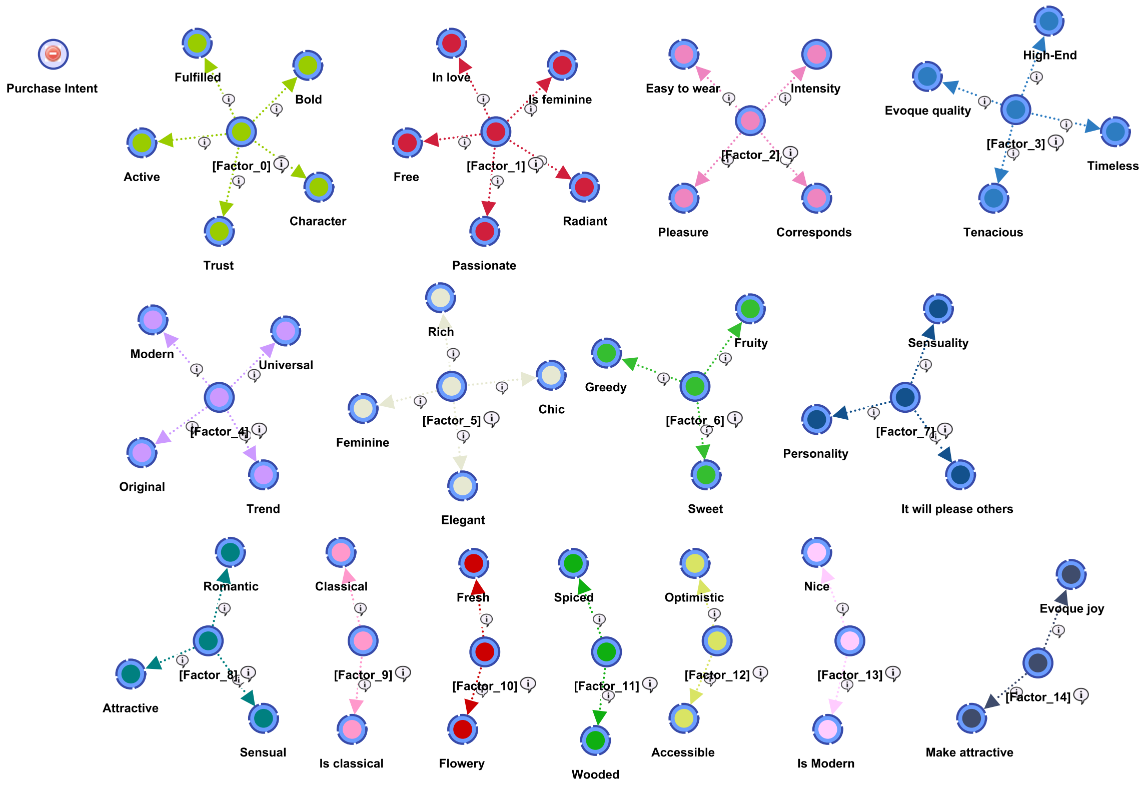


Figure 4.8: Perfume data: Inducing a factor into each cluster

4.4.3 Multiple Clustering

We are now interested in finding factors from the clustering data, with as many factors being introduced into the model as the number of clusters. This process entails inducing factors with discrete states for each cluster, making sure that the factors have high mutual information with their children nodes [28]. The imputing of factor state to each observation can be done using maximum likelihood [18].

This is done by using the Multiple Clustering function of BayesiaLab and it returns each cluster on its own with the factor at the centre. The resulting outcome is shown in figure 4.8. Here, the lines connecting the factors and their manifest variables are dotted to indicate that they are fixed and should not be altered when finding the structural model.

The representative values for each state can be obtained by calculating the weighted average of the manifest variables' means given the specific states, where the weights are relative significance in relation to the factor. The relative significance is defined as follows:

$$RS_i = \frac{I(M_i, F)}{\max\limits_j I(M_j, F)} \tag{4.10}$$

Table 4.3: Example of calculating a weighted average for a state.

Manifest var	Mutual information	Relative significance	Mean Value
V1	0.35	1	2.13
V2	0.33	0.943	2.10
V3	0.30	0.857	2.25
V4	0.29	0.829	1.99
V5	0.28	0.8	2.07

In table 4.3 the highest value for mutual information is 0.35. The relative significance is then each mutual information divided by 0.35. Suppose that the mean values shown here are those of the lowest state for each manifest variable. Then the value for the lowest state of the factor is obtained by $(\sum \text{Relative Sig} \times \text{Mean Val})/\sum \text{Relative Sig} = 9.34/4.43 = 2.11$.

A measure of quality for each factor can be examined using Contingency Table Fit (CTF), which has the following formula:

$$CTF(B) = \frac{\bar{ll}(B_u) - \bar{ll}(B)}{\bar{ll}(B_f) - \bar{ll}(B)} \tag{4.11}$$

where $\bar{ll}(B)$ is the mean log-likelihood of the data given the current network B, $\bar{ll}(B_u)$ is the mean log-likelihood of the data given the fully unconnected network (or the worst case scenario) and $\bar{ll}(B_f)$ is the mean log-likelihood of the data given the fully connected network (or the best case scenario) [5]. Thus, the CTF takes on the value 100 if the current network is able to produce an exact representation of the fully connected joint probability distribution, while it is equal to 0 if the network represents a fully unconnected network, where all variables are marginally independent [5]. The ideal value of CTF depends on the number of variables, the number of states for each variables and the number of states for the factor. For example, a factor with 4 states should be able to fully account for 2 variables with binary states, thus a value of 100 should be expected. However, for 8 variables with 4 states, a factor of 4 states obtaining a CTF of 50% implies that the factor manages to represent JPD of 4^8 cells with a quality of 50% using just 4 states. [5]. On the individual level between each manifest variable and the factor, a test of independence is conducted as to confirm whether it is sensible to create the network with the clustered manifest variable. The G-test is used to obtain the statistic and the p-value, where we conclude that the variable and the factor are not independent if the p-value is low.

So far we have only dealt with how the observed variables interact with factors. This is what is known as the measurement model in the field of SEM [16]. We will now move onto the structural part which deals with how latent variables interact with each other.

4.4.4 Exploratory Bayesian Network (EBN)

Here we must find the relationships between the factors while keeping the relationship between factors and their manifest variables intact. In order to enforce this restriction, we must resort to using the Taboo (also known as Tabu) algorithm, which searches for the local optimum while verifying at every step that it does not belong to a set of restricted networks [10]. Establishing the relationships between the factors returns the structural part of the model and this completes the network structure learning of the data and we can proceed to verify this structure using the classical SEM method and use the what-if analysis for further investigations. Because this network is not based on any input from a researcher, it is only an exploratory result, hence we propose the name of Exploratory Bayesian Network.

The structural model is found by using the Taboo algorithm specifically, since this is the only algorithm that can keep the relationship between the factors and their manifest variables fixed. Here, we bring back the variable *Purchase Intent* because we now want to see the relationship it has with the factors. In figure 4.9 we see that the network has kept the relationships for the measurement model and added the structural model onto it in solid black arrows. We also see that *Purchase Intent* is directly influenced by factor 2. This structure can be redrawn in SPSS AMOS for traditional SEM analysis. Keep in

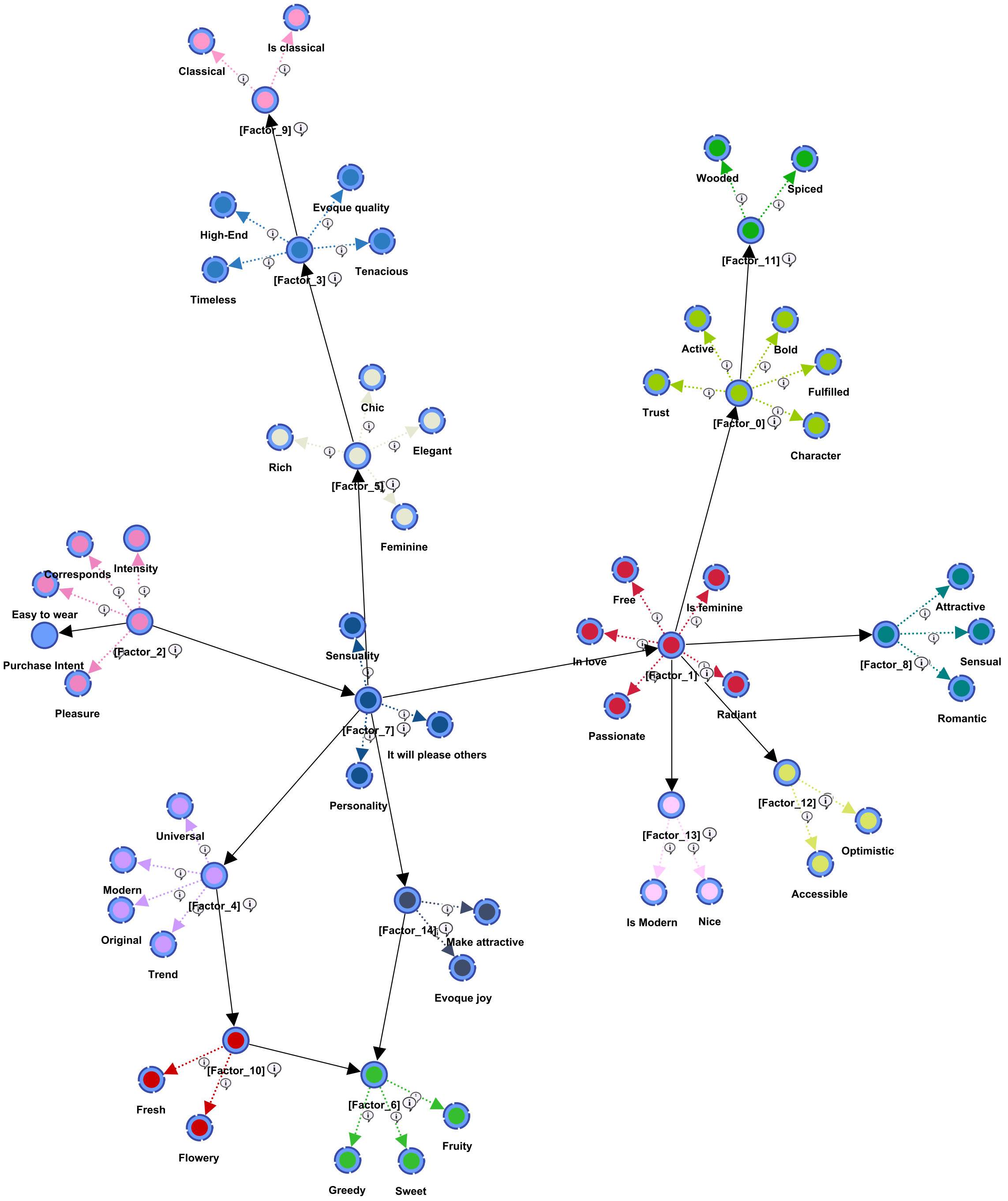


Figure 4.9: Perfume data: Exploratory Bayesian Network

Table 4.4: Conventional SEM model fit indices for Perfume data

CFI	0.904
SRMR	0.0424
RMSEA	0.072
RMSEA upper 90	0.074
RMSEA lower 90	0.071

mind that the 'Perfume' dataset illustrates the structural learning algorithms discussed in this chapter. There is no existing theory to be hypothesised as is the case with SEM. It is, however, interesting to test the reaction of SEM metrics on the learned structure. This by itself also serves as some type of validation for the unsupervised learned structure and parametrisation. The estimation of the model returned coefficients and variances which are all highly significant. Table 4.4 shows some of the model fit indices, which also indicate a satisfactory model fit.

4.4.5 What-if analysis

Typically, researchers using SEM are interested in seeing whether the model they proposed is valid and if so, they want to know what the coefficient values, and interpret them and the analysis comes to an end. By having a Bayesian network representation of SEM, we can take a step further and conduct inference by explicitly entering evidence into the network to see how it changes the value of other variables. The beauty of using Bayesian network is that we can instantiate any variable, regardless of being concerned about the direction of influence. Note that:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{4.12}$$

therefore,

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A) \tag{4.13}$$

This is because the network is probabilistic, as opposed to deterministic, hence it comes down to rearranging the variables for conditional probability calculation. Thus we are not limited to only substituting values for the independent or exogenous variables, we can assign a state to the dependent or endogenous variables and see how that changes behaviour of the remaining variables. When the evidence for a parent node is given, the posterior probability for the child node is simply read off the conditional probability table. We can obtain the opposite, the posterior probability of the parent node given evidence for the child node. Let A denote the parent node and B denote the child node.

$$\begin{aligned} P(A|B)P(B) &= P(A \cap B) = P(B|A)P(A) \\ P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ &= \alpha P(A)\lambda(A) \end{aligned} \tag{4.14}$$

where $\alpha = \frac{1}{P(B)}$ is a normalising constant, P(A) is the prior and $\lambda(A) = P(B|A)$ is the likelihood [22]. This is the famous Bayes Theorem.

Consider a small example from [22], with two variables Flu and $HighTemp$, where the direction of influence is $Flu \rightarrow HighTemp$. The known probability values are prior $P(Flu = T) = 0.05$ and CPT values $P(HighTemp = T|Flu = T) = 0.9$, $P(HighTemp = T|Flu = F) = 0.2$. For a more complete representation, the probability tables are shown in tables 4.5 and 4.6

Table 4.5: Marginal probability of having flu

	True	False
$P(Flu)$	0.05	0.95

Table 4.6: Conditional probability of high temperature given flu

$P(HighTemp\|Flu)$	True	False
$Flu = T$	0.9	0.1
$Flu = F$	0.2	0.8

Let $Bel(A)$ denote the posterior probability, or belief. If the evidence on $HighTemp$ is set to $True$, then the beliefs are calculated as follows:

$$\begin{aligned} Bel(Flu = T) &= \alpha P(Flu = T)\lambda(Flu = T) \\ &= \alpha \times 0.05 \times 0.9 \\ &= 0.045\alpha \\ Bel(Flu = F) &= \alpha P(Flu = T)\lambda(Flu = T) \\ &= \alpha \times 0.95 \times 0.2 \\ &= 0.19\alpha \end{aligned} \tag{4.15}$$

According to the Total probability theorem, $Bel(Flu = T) + Bel(Flu = F) = 1$ [22]. Therefore

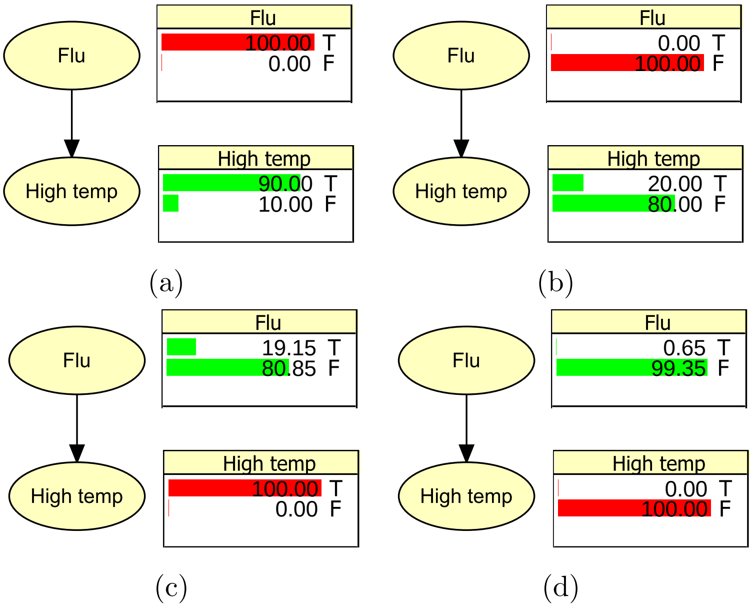
$$\begin{aligned} 1 &= 0.045\alpha + 0.19\alpha \\ 1 &= 0.235\alpha \\ \alpha &= \frac{1}{0.235} \end{aligned} \tag{4.16}$$

Substituting this α back to 4.15,

$$\begin{aligned} Bel(Flu = T) &= \frac{0.045}{0.235} \\ &= 0.1915 \\ Bel(Flu = F) &= \frac{0.19}{0.235} \\ &= 0.8085 \end{aligned} \tag{4.17}$$

Thus, if it is observed that a person has a fever, the probability would be 0.1915 that it is due to flu. Conversely, if it is observed that a person does not have fever, there is a 0.9935 probability that the person does not have flu. There are softwares which can perform such calculations and offer an intuitive graphical interface. The output in figure 4.10 is obtained by using Hugin (www.hugin.com). Although BayesiaLab produces same inference results, Hugin can present the posterior probabilities in a convenient and intuitive way alongside the network.

In figure 4.10, the evidence is entered into the parent node for 4.10a and 4.10b, in other words, the direction of inference is from cause to symptom. This is what is known as a predictive reasoning [22]. Inference in the opposite direction is called a diagnostic reasoning [22] and this helps us to see what causes would have led to a certain outcome. These are shown in 4.10c and 4.10d Another type of inference is called 'prescriptive' and this takes diagnostic reasoning a step further by finding out what the causes should be for the optimal outcome to take place [7].



(a) (b)

(c) (d)

Figure 4.10: Inference by setting evidence on different states

Thereafter we can perform inference on the obtained BN, should we wish to further explore. Figure 4.11 shows the network from figure 4.9, where only the factors and *Purchase Intent* have been carried over to Hugin along with their conditional probabilities. Next to each node there are small windows which show the probability distribution visually as well as numerically: the column of numbers on the right lists the different states of the nodes and the column on the left lists the respective default probability of belonging to the state.

As an example of a predictive inference, an evidence is set to the node *Pleasure* to its highest state in figure 4.12 and we can see how it affects the rest of the network. Compared to figure 4.11, we can see that the distribution has generally shifted towards the higher states. In particular, the probability is 0 that two lowest states in *Purchase Intent* will be observed and the probability of observing the highest state has tripled. Next we can set evidence to *Purchase Intent* for a diagnostic inference. Figure 4.13 shows that in general, the distribution has now shifted towards the lower values of the states. For a prescriptive inference, we change the evidence on *Purchase Intent* to its highest state as shown in figure 4.14. The result, which has a posterior distribution with 86% probability of encountering the highest state in the node *Pleasure*, looks quite

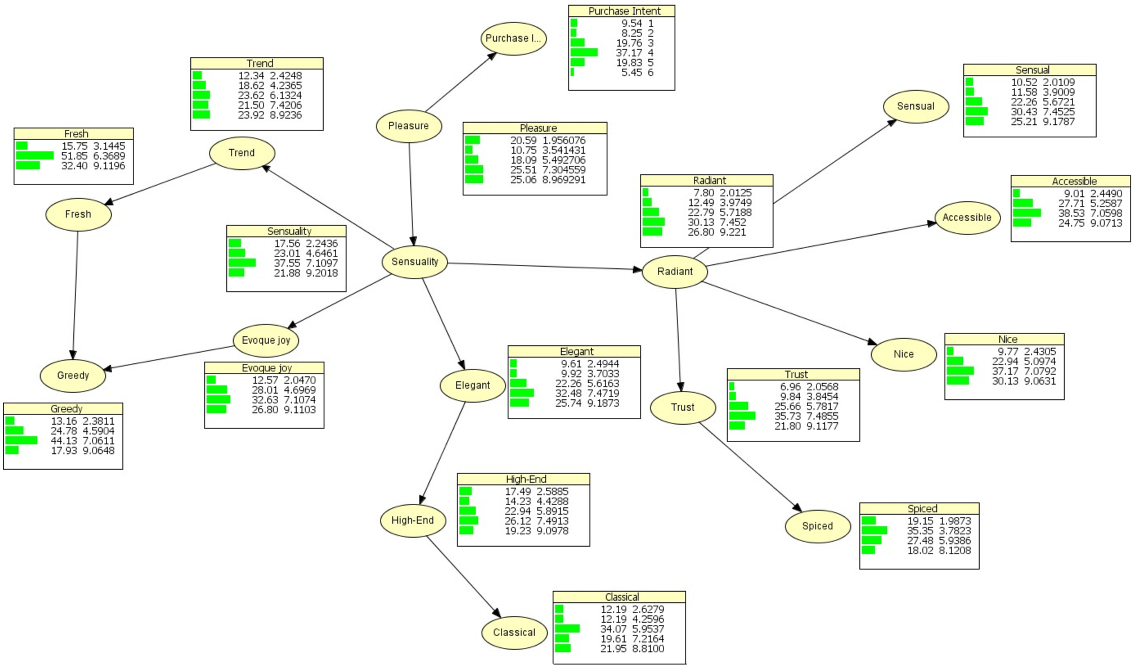


Figure 4.11: Bayesian Network representation of the perfume data

similar to figure 4.12.

4.5 Summary

In this chapter we introduced graphical models, specifically DAGs and BNs. We discussed the workings of a BN briefly by covering the basic concepts with simple examples. The challenges have been identified and solution algorithms have been given with regards to graphical model structure learning from data. The mathematical framework has been applied to a real dataset to show how one can use this as a tool to explore the data. After obtaining the full network with factors, different ways of performing inference was discussed. In the next chapter, we investigate different strategies to arrive at a BN by using the data and the expert knowledge.

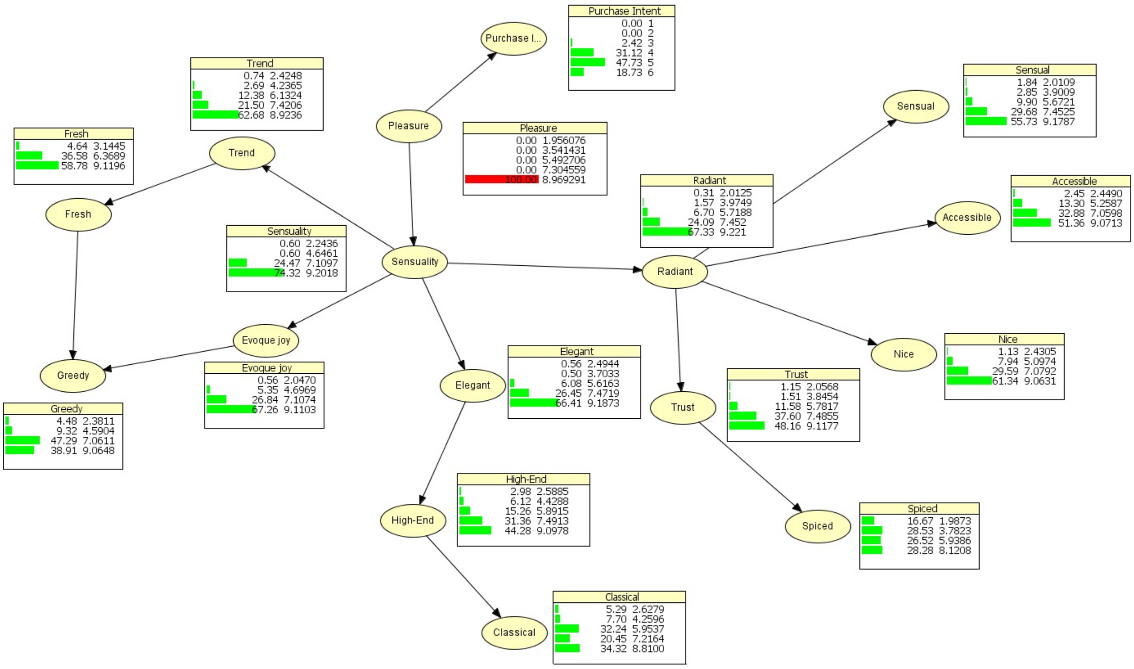


Figure 4.12: What-if analysis: predictive

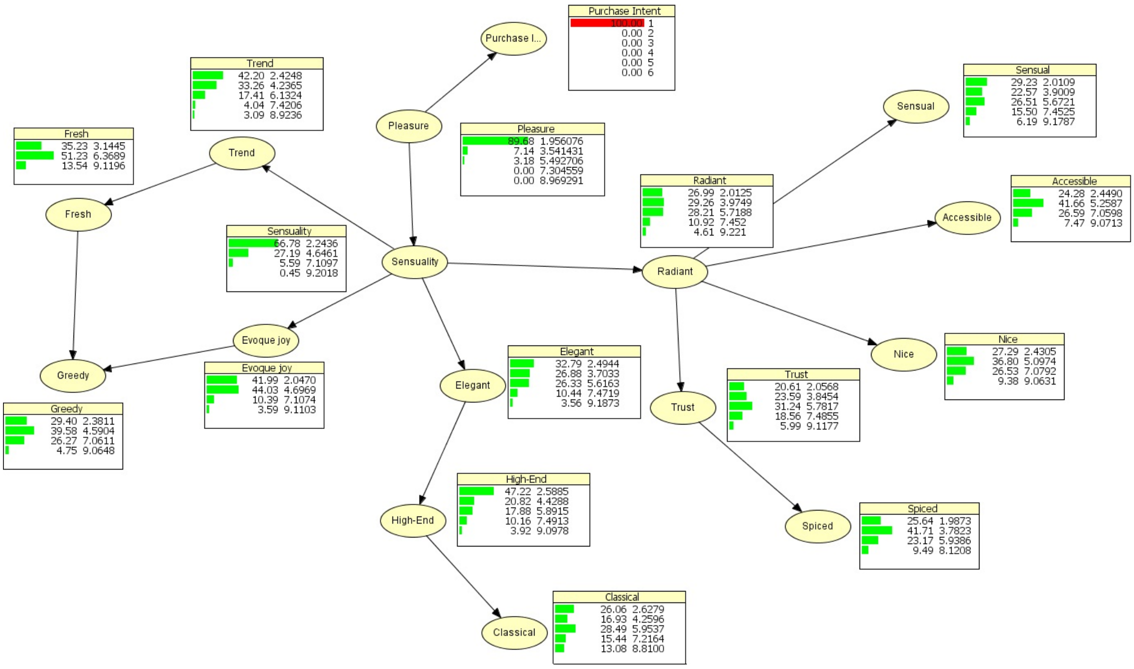


Figure 4.13: What-if analysis: diagnostic

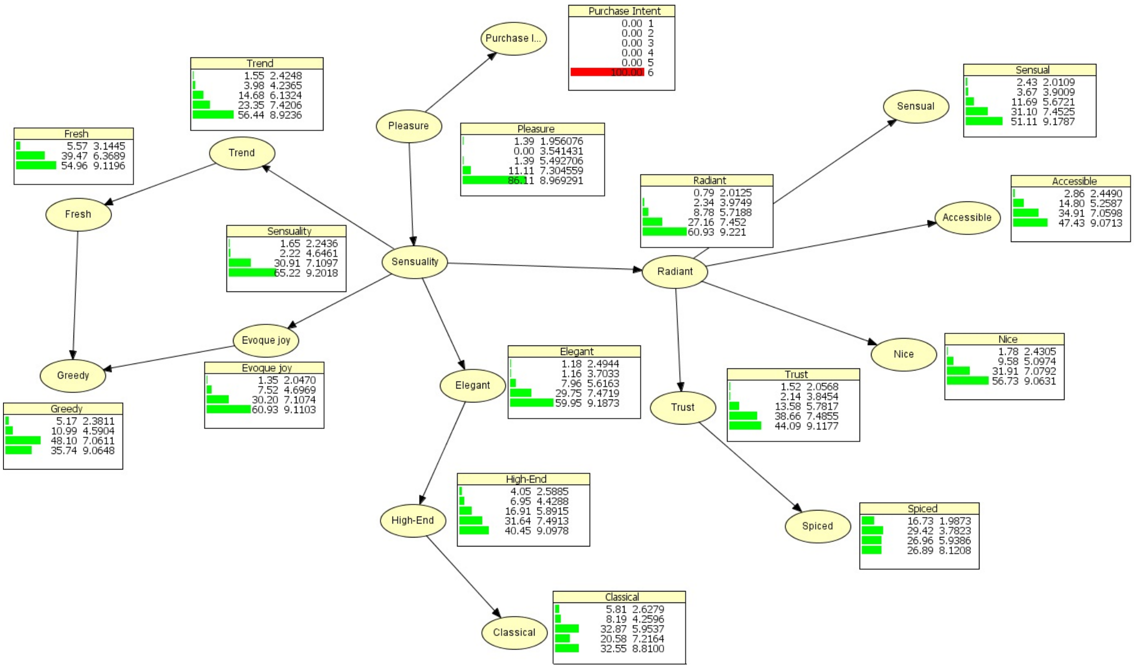


Figure 4.14: What-if analysis: prescriptive

Chapter 5

Probabilistic SEM

A SEM starts out as a theoretical model representing causal paths between variables. The SEM researcher then designs an experiment and collects data to see whether the theoretical model they proposed is valid and if so, they want to know what the coefficient values are and interpret them. In practice, the original theoretical model structure might be adapted slightly based on SEM evaluation results as the process of fitting the data to the models provides new information. This involves checking the model fit as well as parameter estimates and making adjustments (which should be supported by the literature) if the measurement model is unsatisfactory in the CFA. Once the model structure is stabilised and the SEM produces satisfactory results, the calculated coefficients become the discussion points and drives arguments relating back to the theoretical model.

We propose representing the final theoretical SEM by a BN, which we would like to call a Probabilistic Structural Equation Model (PSEM). With the PSEM, we can take things a step further and conduct inference by explicitly entering evidence into the network and performing different types of inferences as listed in 4.4.5. The beauty of using BNs is that we can instantiate any variable, regardless of being concerned about the direction of influence. This was discussed in section 4.4.5.

One can arrive to the BN phase purely from a structure hypothesised from theory, which is then validated and confirmed with data. This is done using classical SEM techniques as discussed in Chapter 3. As we acknowledge the structure as a SEM, the resulting BN is then called a probabilistic SEM (PSEM).

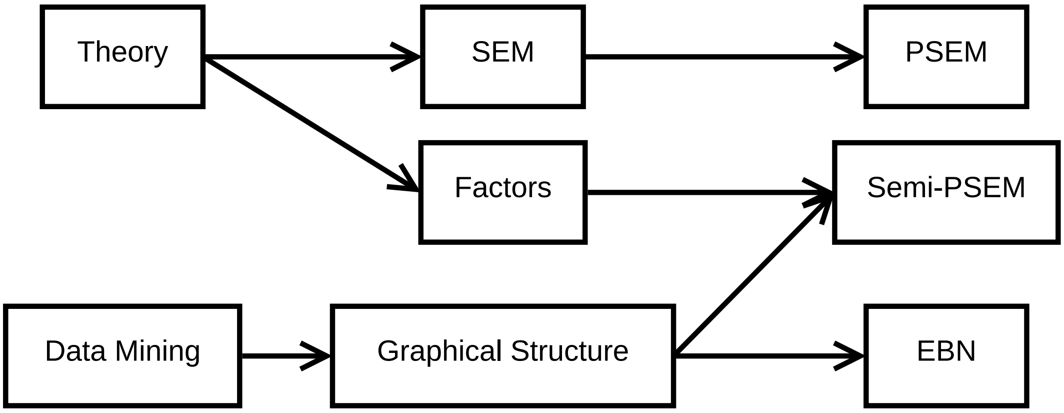


Figure 5.1: Process of obtaining different BNs

The other side of the spectrum is a completely data-driven approach (section 4.4) - assuming no theory driving the structure. This is an explanatory approach and can be useful in scenarios where data exist but no extensive theory is available. The resulting BN is then also called an explanatory BN (EBN), rather than a PSEM as the structure is not based on theory as is the case with SEM.

Finally, we consider the case where the factors are created according to the theory but the structural paths are learned using data. The resulting BN is then called a semi-probabilistic SEM. Figure 5.1 diagrammatically illustrates the process of obtaining different BNs.

5.1 Application

Social Networking Sites (SNS) represent a great opportunity today for companies to advertise their products and services as well as target and personalise their messages based on the data people declared online. However, how people perceive the new advertising methods employed in SNS is still largely unknown, as well as whether they consider such advertisements to be an intrusion on their private, although social, space. In addition, very little empirical research of causal relationships exists, leaving unanswered questions about how SNS users perceive advertisements posted on their own online social profile.

The conceptual framework is based on Theory of Planned Behaviour (TPB). The components of TPB are four general constructs: behavioural intention (BI), attitude (A), subjective norm (SN) and perceived behavioural control (PBC). Following the TPB, we thus predict that behaviour towards SNS advertising will be influenced by the users attitude towards SNS advertising, the subjective norms and perceived behavioural control. The attitude is itself dependent on four main beliefs: trust, attitudes towards advertising in general, advertising value and advertising intrusiveness, which are themselves dependent on other antecedent variables as described in the literature.

The study population comprised of active adult (18 years and older) Facebook users. The survey was developed in English for both (South Africa and Australia) countries and delivered online. Sampling in both countries involved the use of market research firms holding consumer panels where the firms' provided a link to the survey. Participants were incentivised by the market research firms in accordance with their normal practices and a sample of 401were realised in both countries respectively, resulting in a total of 802 respondents.

There are 9 factors in this SEM, namely:

- Privacy Concern
- Trust
- Ad intrusiveness
- Behaviour towards brand
- Behaviour towards ad
- Perceived behaviour control
- Attitude towards ads
- Attitude towards FB ad
- Ad values

The different softwares used are as follows: SPSS AMOS was used for classical SEM, BayesiaLab was used for EBN and Semi-PSEM, and Hugin was used for performing inference with PSEM, EBN and Semi-PSEM. The results of SEM using SPSS AMOS can be found in appendix C.

5.2 Probabilistic SEM

Here, the path between factors and manifest variables are all specified according to the relevant theory and the theoretical or expert-developed SEM is turned into a BN for inference. Figure 5.2 shows the PSEM for the Facebook data. Although the EBN (which is discussed next) has not played a role in developing this network, factors such as $Trust$, $Privacy\,concern$ and $Attitude\,towards\,ads$ are based on the same variables as the EBN in figure 5.10. The corresponding PSEM is shown in figure 5.3. The probabilities, when the network structure is specified, are learned by maximum likelihood, where the occurrence of the variable states are simply counted [5]. Learning probabilities in this manner also makes the network less responsive to changes, as the conditional probabilities will only reflect the changes if there are observations in the data which possess the given state value for the variable.

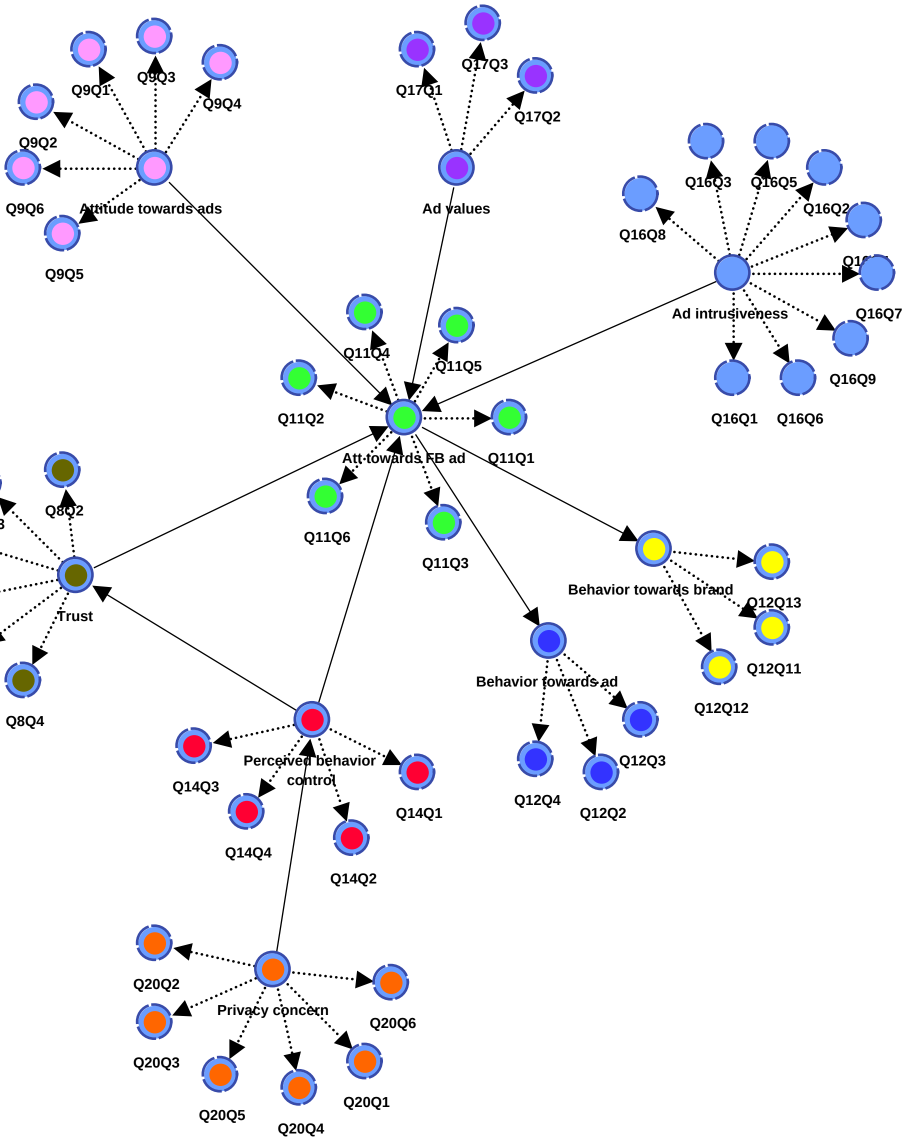


Figure 5.2: Facebook data: PSEM

For the remainder of the chapter, we restrict the focus of our discussion mostly to the nodes that represent *Attitude towards FB ad*, *Behavior towards ad* and *Behavior towards brand*, for compactness. We now proceed to inference to see what can be learned from the PSEM. Figure 5.4 shows changes to the network, given that *Attitude towards FB ad* is at its lowest state. The biggest changes occurred to the nodes *Behavior towards ad* and *Behavior towards brand*, where the probability of finding them in low-valued states have increased drastically. Figure 5.5 illustrates the opposite case, when $Behavior\,towards\,ad$ and $Behavior\,towards\,brand$ have been set to the highest states. This in accordance with the results from SPSS AMOS in appendix C, where the path coefficients from $Attitude\,towards\,FB\,ad$ to $Behavior\,towards\,ad$ and $Behavior\,towards\,brand$ are positive.

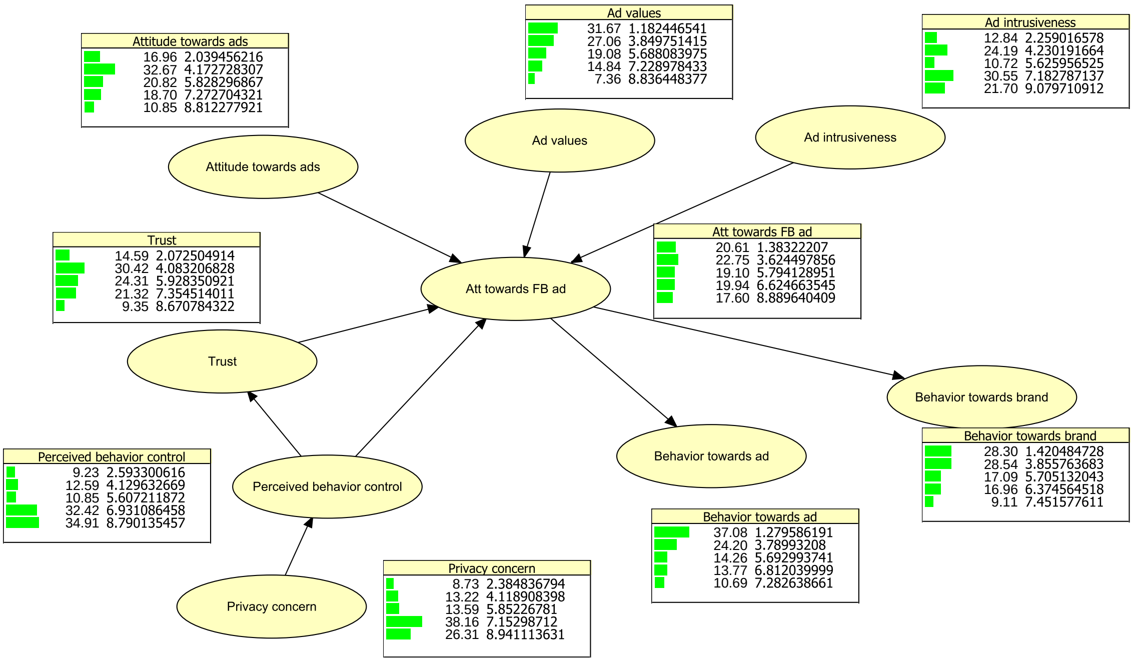


Figure 5.3: Facebook data: PSEM (Factors)

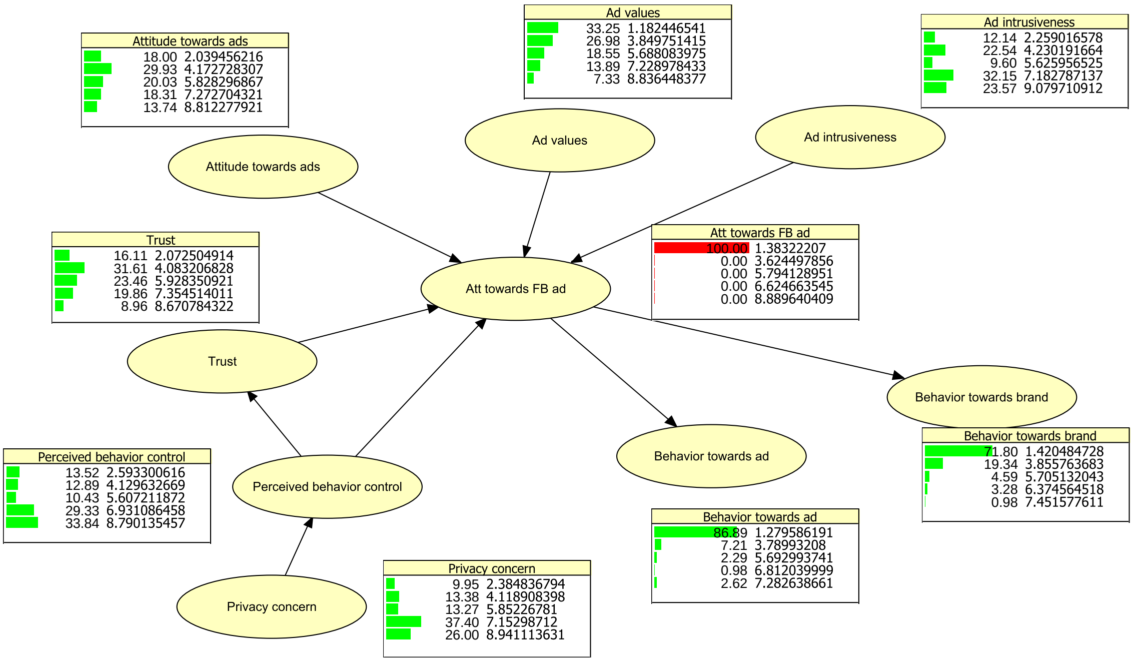


Figure 5.4: Facebook data: PSEM predictive inference

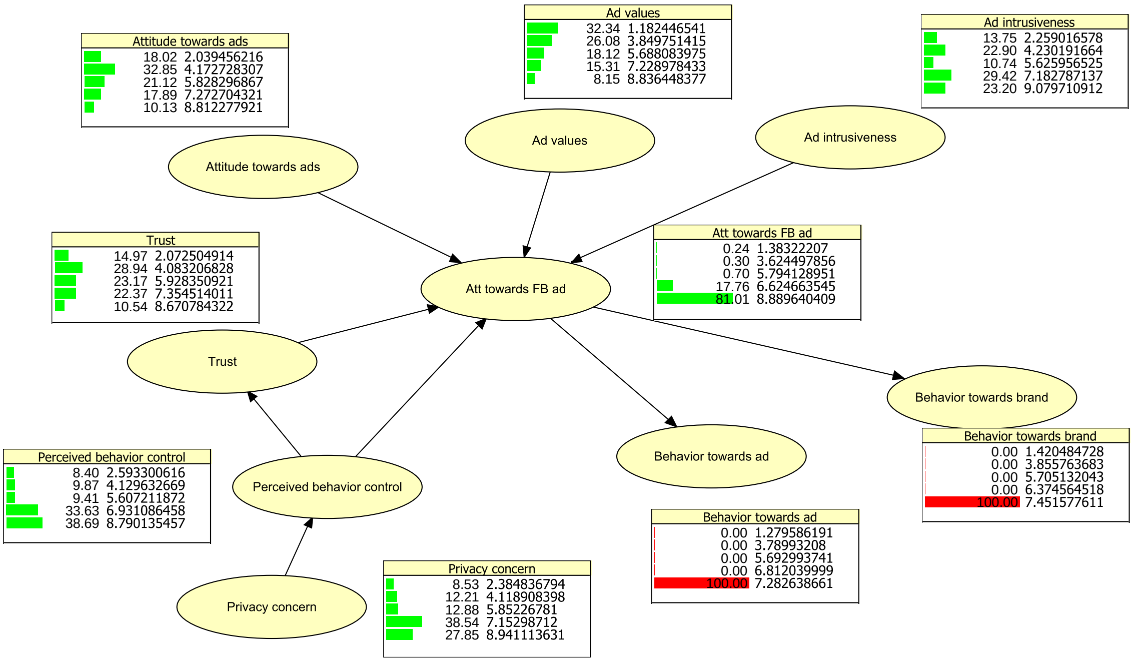


Figure 5.5: Facebook data: PSEM prescriptive inference

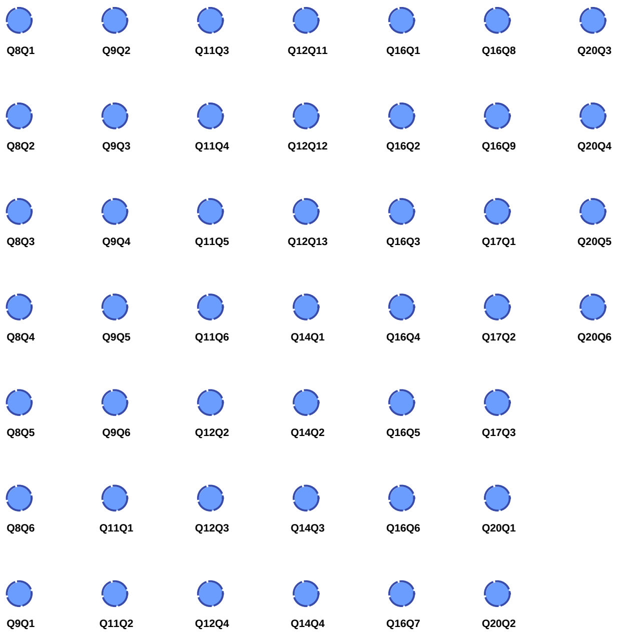


Figure 5.6: Facebook data imported into BayesiaLab

5.3 Exploratory Bayesian Network (EBN)

Figure 5.6 shows how the manifest variables appear upon importing into BayesiaLab. The initial network, shown in figure 5.7 is learned using MWST with MDL score of 66,037.546. This score is obtained when using the EQ algorithm as well and using data perturbation, hence it can be seen as the optimal value.

Next we cluster the nodes based on their Arc Force. The number of clusters can vary. Here, 8 clusters were selected to result in each cluster having 4 to 7 manifest variables. This is shown in figure 5.8. The next step is to induce a factor for each cluster that represents the cluster. After the factor induction we obtain 8 factors as shown in figure 5.9. It is interesting to see that factors 0,1,2,3,5 and 6 are all based on the same group of questions. This can serve as an indication that the question are well structured. Because some variables are clustered differently from the theoretical structure, the factors are not given specific names. The last step in obtaining the EBN is to learn the structure between the factors. Using the Taboo algorithm the network as shown in figure 5.10 is found.

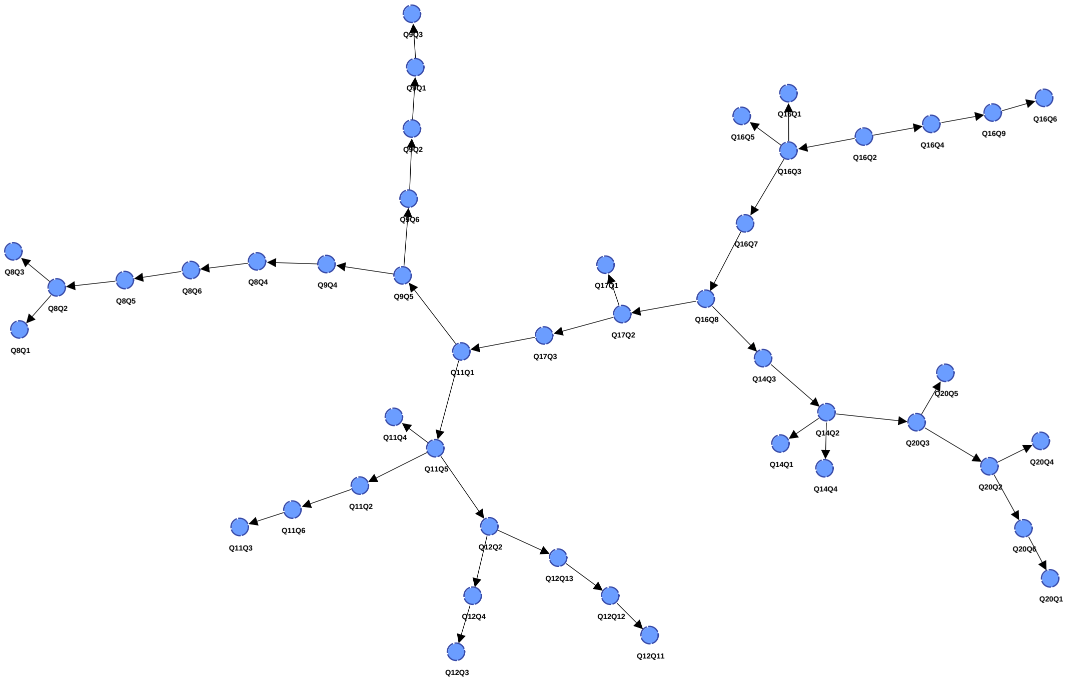


Figure 5.7: Facebook data: Initial network learned using MWST

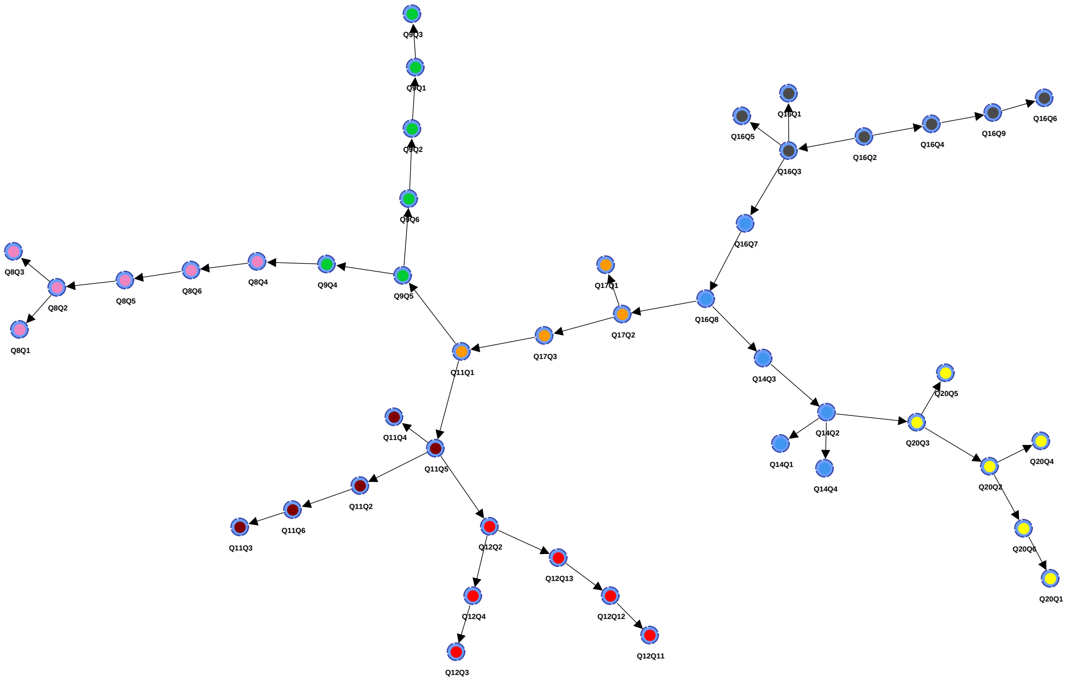


Figure 5.8: Facebook data: Clustering result, max 5 variables per cluster

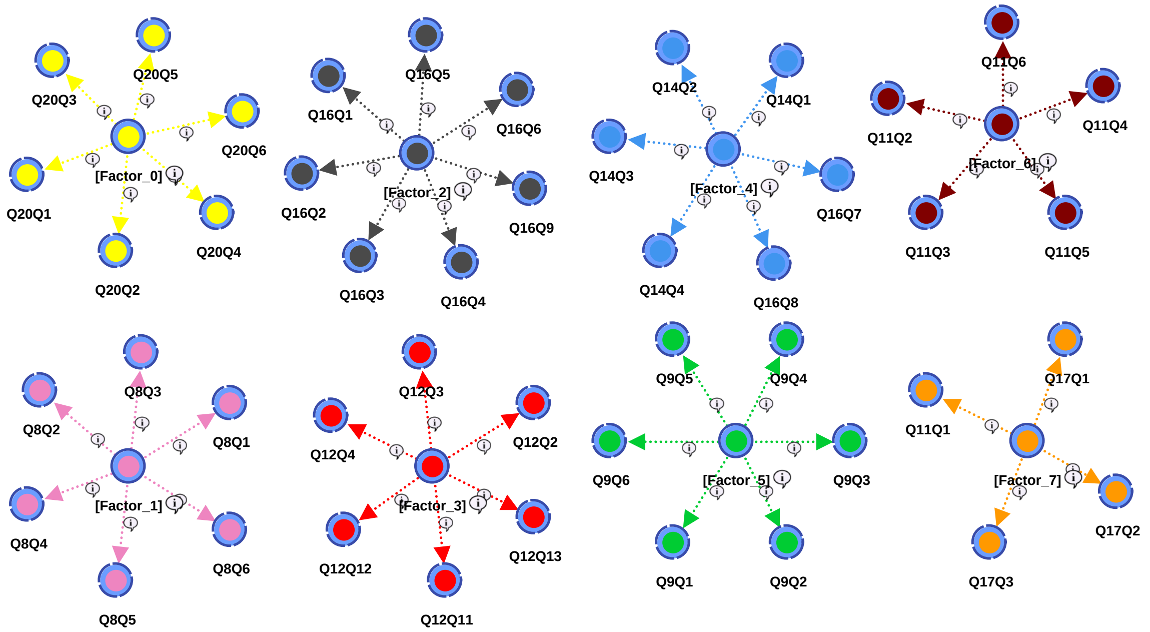


Figure 5.9: Facebook data: Inducing a factor into each cluster

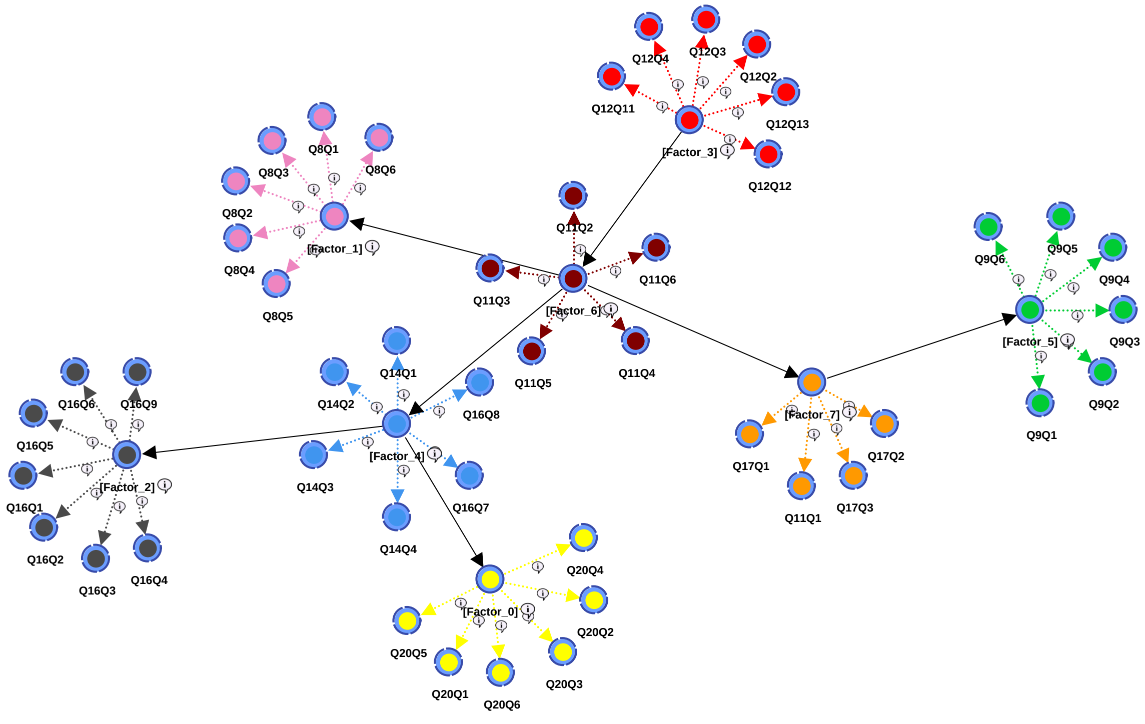


Figure 5.10: Facebook data: Exploratory Bayesian Network

We can take this network and see what happens to the distribution of each factors to better understand the data. Figure 5.11 shows the default probability distributions for each factors. Table 5.1 shows the breakdown of the factors. Notice that the default probabilities are different from that of figure 5.3 because the structure also needed to be learned in EBN. $Factor\,6$ which is comprised of items from question 11, can be seen as the representative factor for $Attitude\,towards\,Facebook\,advertisements$.

In figure 5.12, we set evidence onto $Factor\,6$, and see how it influences other factors in the network. The factors which are influenced the most are $Factor\,1$, $Factor\,3$ and $Factor\,7$. Their distributions have shifted much towards the higher-value states- this is similar to the result we observed for PSEM in figure 5.5. Linking them to the questions, we can conclude that people with a positive attitude towards Facebook advertisements tend to be more trusting, and show favourable behaviours towards the brand and the advertisements while finding the advertisement to be valuable.

It is possible to utilise knowledge from both the data and the theory in order to create a BN; this idea is explored next.

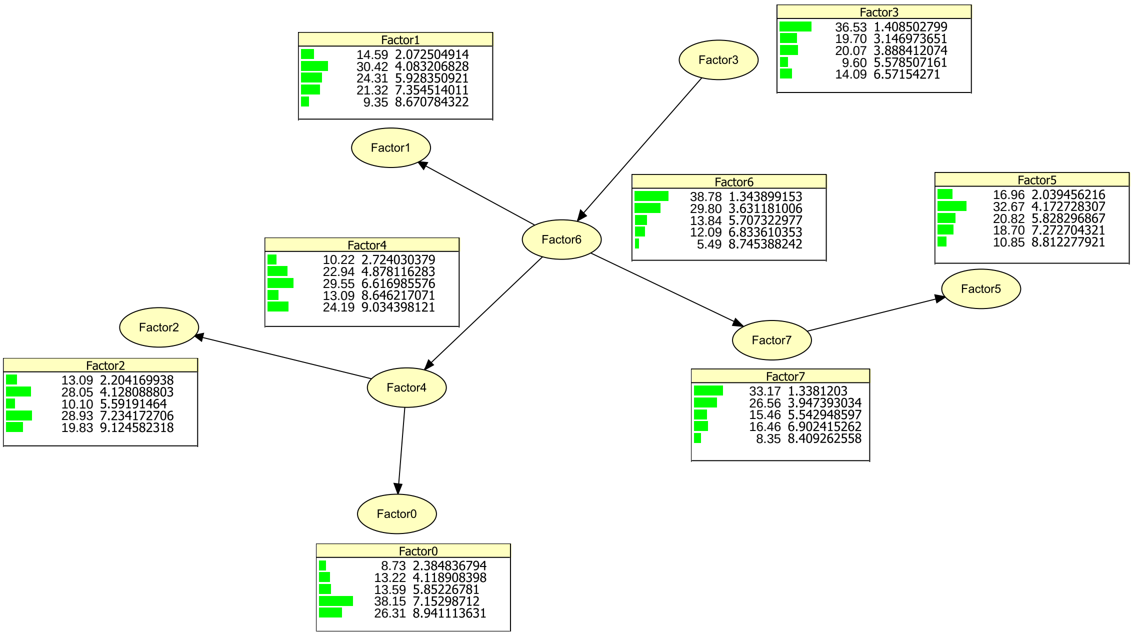


Figure 5.11: Bayesian Network representation of the Facebook data (factors)

Table 5.1: Approximate factors for EBN based on majority manifest variables

Node	Manifest Variables	Closest Theoretical Factor
Factor0	Q20Q1-Q20Q6	Privacy Concern
Factor1	Q8Q1-Q8Q6	Trust
Factor2	Q16Q1-Q16Q6, Q16Q9	Ad intrusiveness
Factor3	Q12Q2-Q12Q4, Q12Q11-Q12Q13	Behaviour towards brand, ad
Factor4	Q14Q1-Q14Q4, Q16Q7-Q16Q8	Perceived behaviour control
Factor5	Q9Q1-Q9Q6	Attitude towards ads
Factor6	Q11Q2-Q11Q6	Attitude towards FB ad
Factor7	Q17Q1-Q17Q3, Q11Q1	Ad values

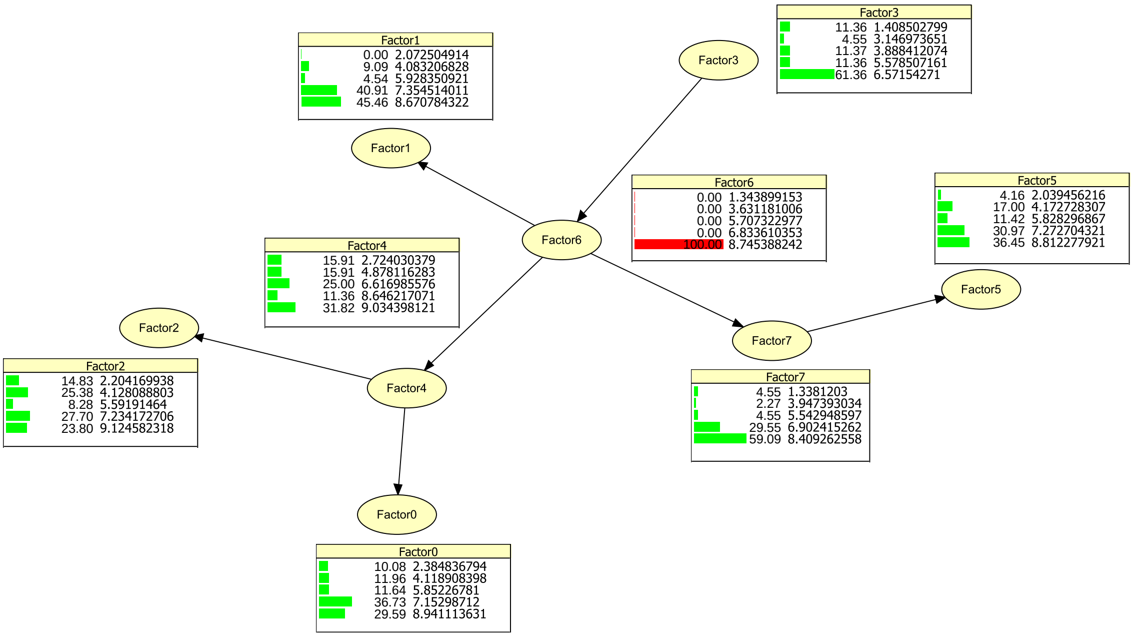


Figure 5.12: Facebook data: Exploratory Bayesian Network inference

5.4 Semi-Probabilistic SEM

In section 5.2 the network was specified completely by the expert, while the network from section 5.3 was learned in a completely data-driven process. It is possible to have a mixture of the two, where the factors are specified according to the literature and theory while the relationship between the factors are learned from the data. This can provide additional perspectives which the researcher may not have thought of. As can be seen in figure 5.13, the factors are based on the manifest variables of figure 5.2, but the network structure between the factors is different.

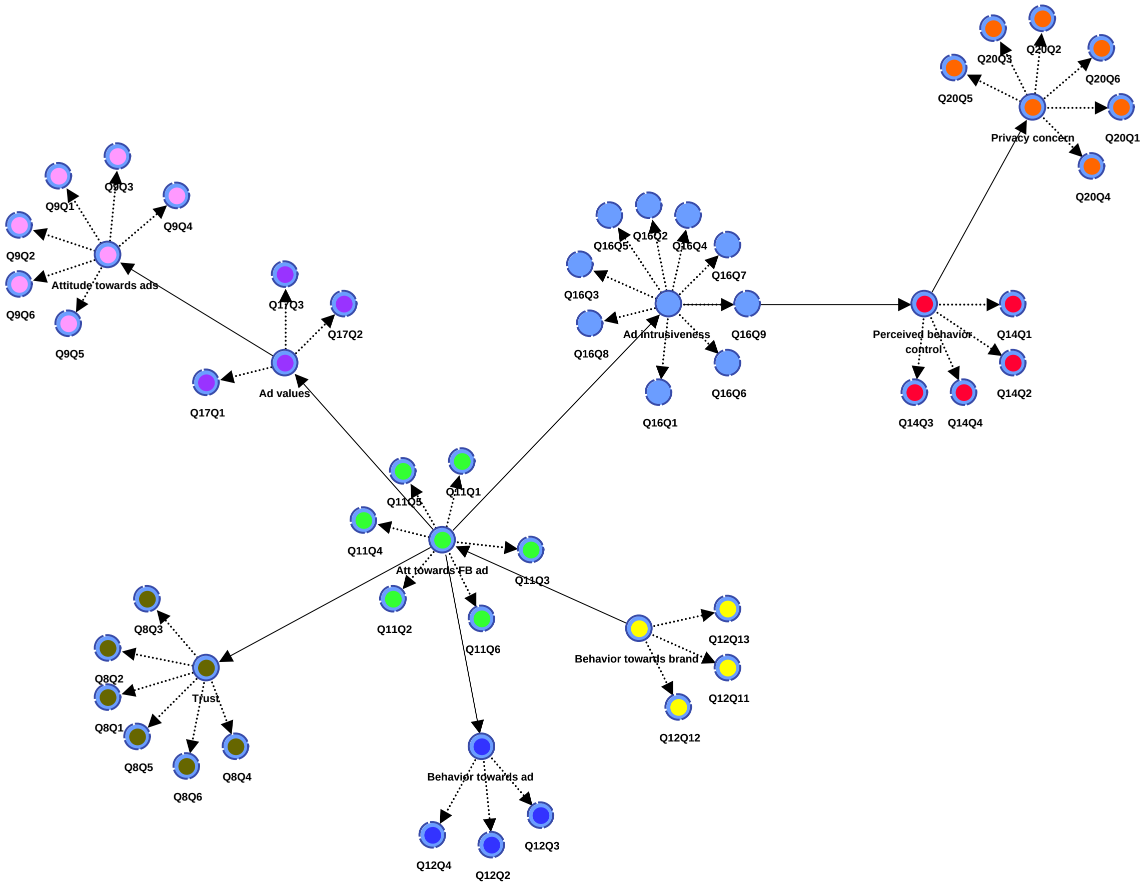


Figure 5.13: Facebook data: Semi-PSEM

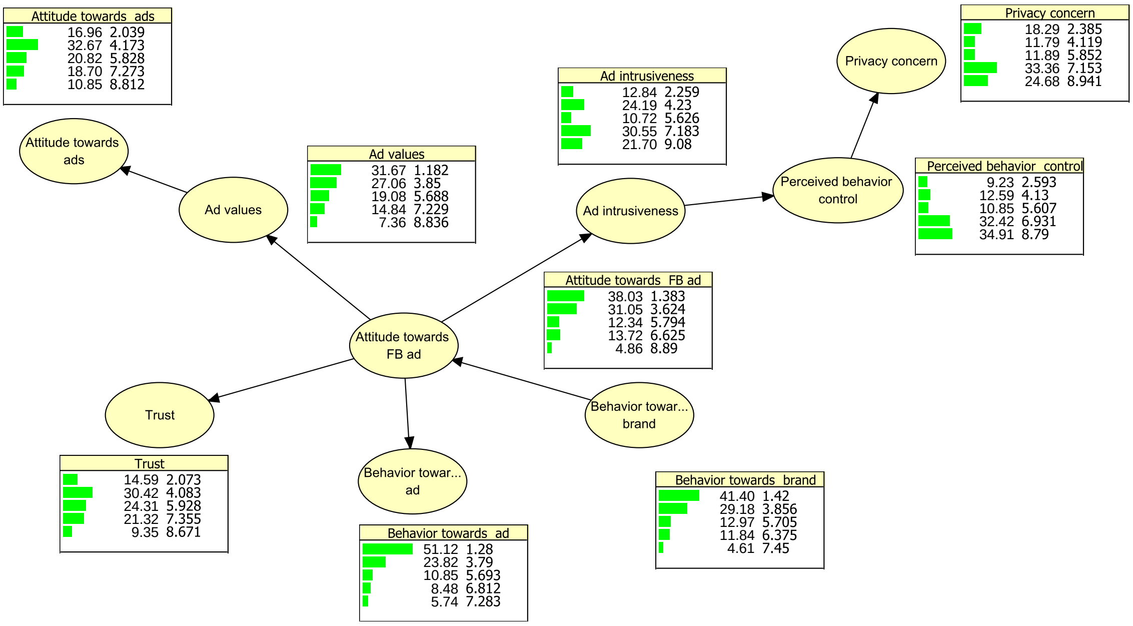


Figure 5.14: Facebook data: Semi-PSEM (factors)

Once again we can perform inference by setting evidence onto the nodes of interest. To make the analysis comparable to EBN, we set evidence for the node $Attitude\,towards\,FB\,ad$. In figure 5.12, the nodes for $Behavior\,towards\,ad$ and $Behavior\,towards\,brand$ are combined into one node, $Factor3$ and its highest-valued state, 6.57, has the highest posterior probability of 0.6136. At first glance, it may seem that this is not reflected in figure 5.15, where $Behavior\,towards\,ad$ and $Behavior\,towards\,brand$ only exhibit probability of around 30% for their highest-valued states. However, it is important to note that the highest state values in figure 5.15 for $Behavior\,towards\,ad$ and $Behavior\,towards\,brand$ are 7.283 and 7.45, respectively. Thus having different bins for the states can severely alter how the results appear to be. The sum of the highest 2 states in both nodes give a probability value higher than 0.5, which is now closer to the value of 0.6136 from figure 5.12.

Figure 5.16 mimics the analysis done in figure 5.5. Even though the probability values are different in absolute terms (the actual probability values are different), both figures have the same general shape and the ordering of the states. That is, both indicate that the state with the highest posterior probability for the node $Attitude\,towards\,FB\,ad$

is 8.89 and thereafter 6.625, where these two states account for the majority of the distribution in both instances.

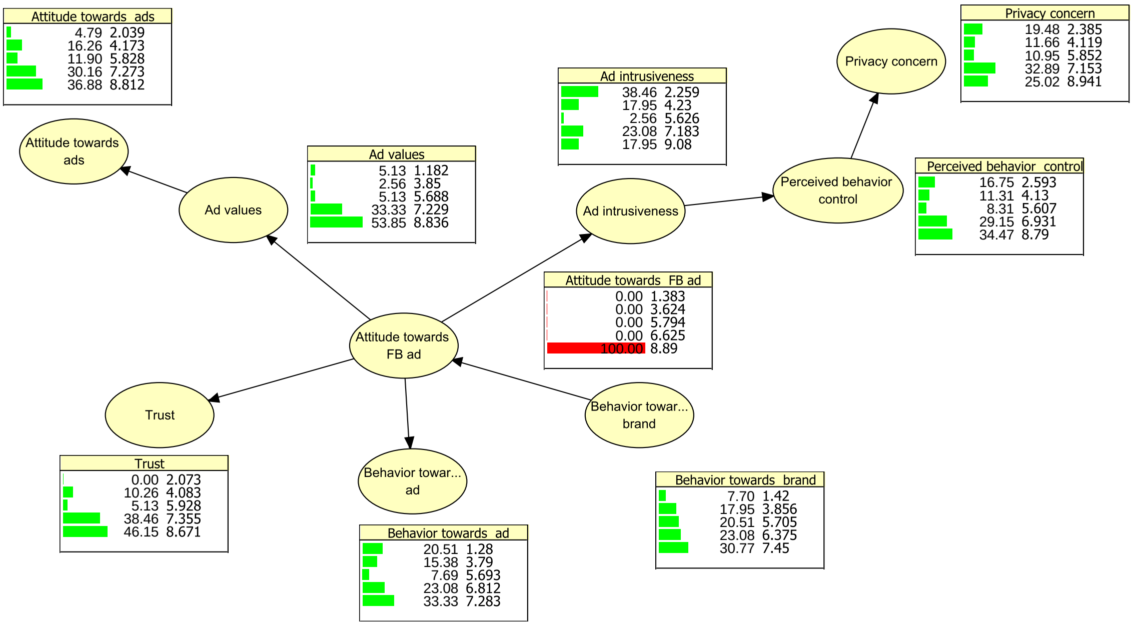


Figure 5.15: Facebook data: Semi-PSEM predictive inference

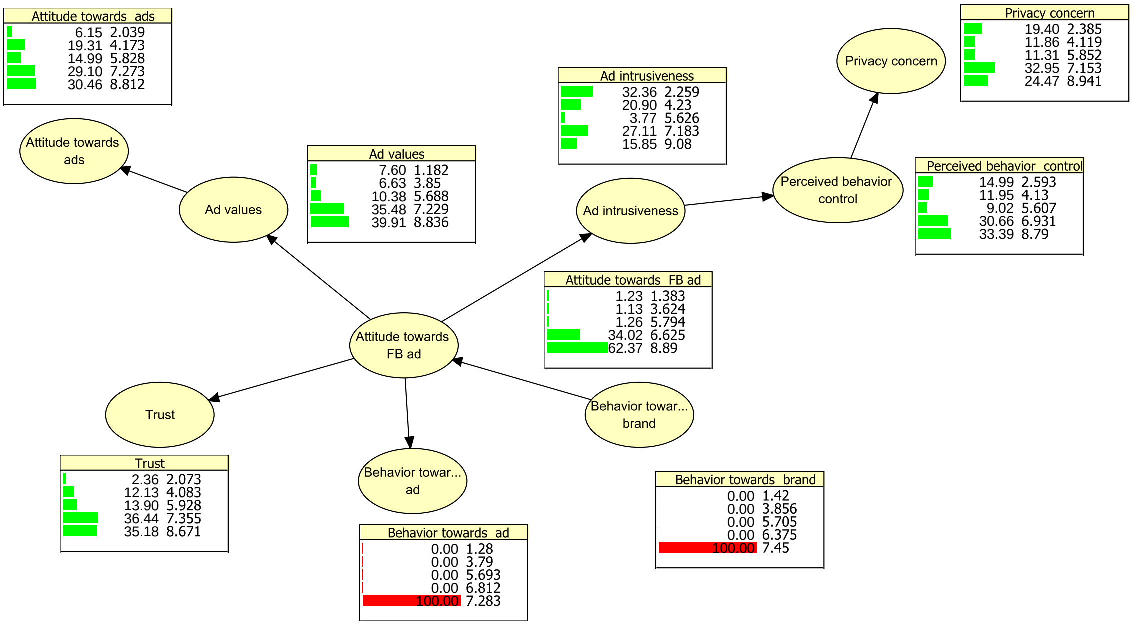


Figure 5.16: Facebook data: Semi-PSEM prescriptive inference

5.5 Summary

In this chapter we discussed three different types of BN which can be found from the data. The PSEM takes the SEM developed from theory by the researcher and converts it into a BN. This allows SEM practitioner to take a step forward by adding the capability to perform what-if analysis onto the network. On the other hand we have the EBN, which can be used without any prior knowledge regarding the data, as the process is purely data-driven. This can assist the practitioner in dynamically exploring the data. We can also generate a network that makes use of both the theory and data, called Semi-PSEM, where factors are defined according to the theory and the structural paths are constructed using a data-driven unsupervised approach. These BNs have been applied to the Facebook advertisement data which are displayed once again in figures 5.17, 5.18, 5.19 for convenience. They have all given results which correspond with a SEM performed with SPSS AMOS.

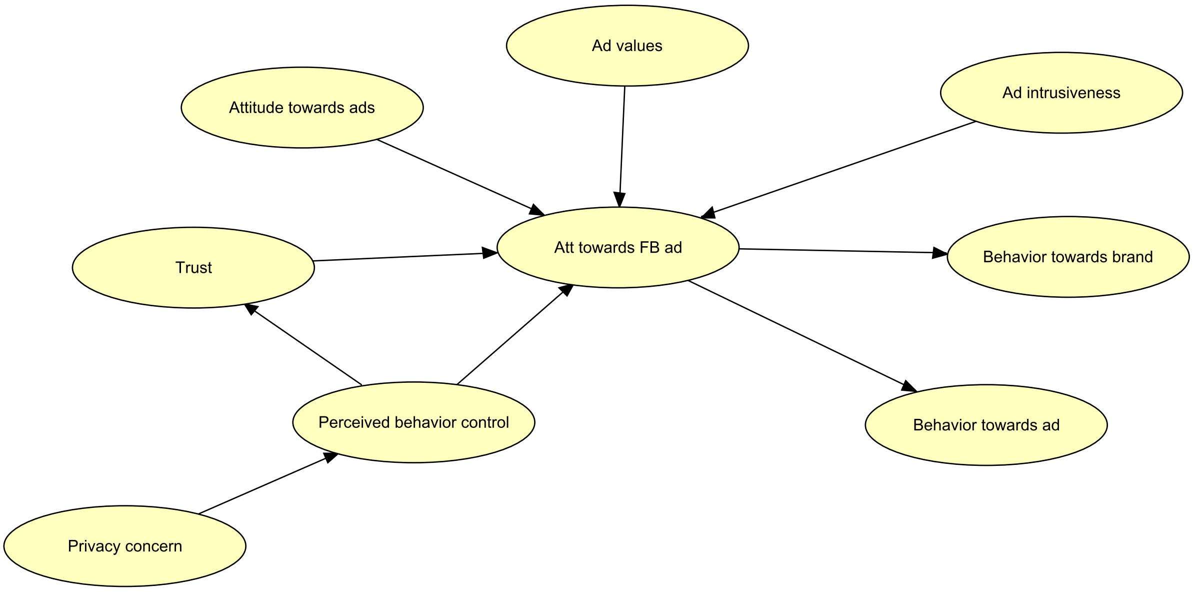


Figure 5.17: PSEM

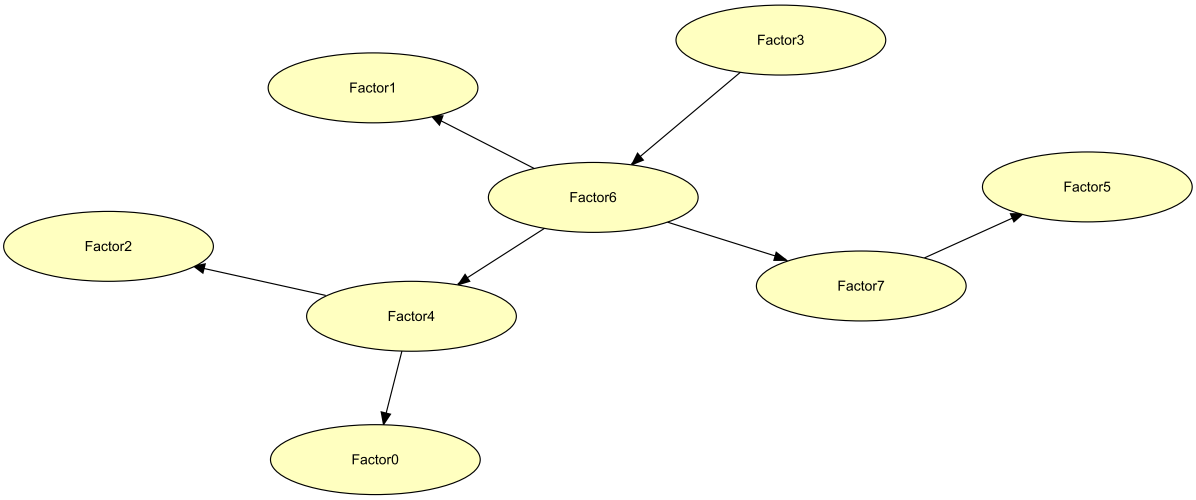


Figure 5.18: EBN

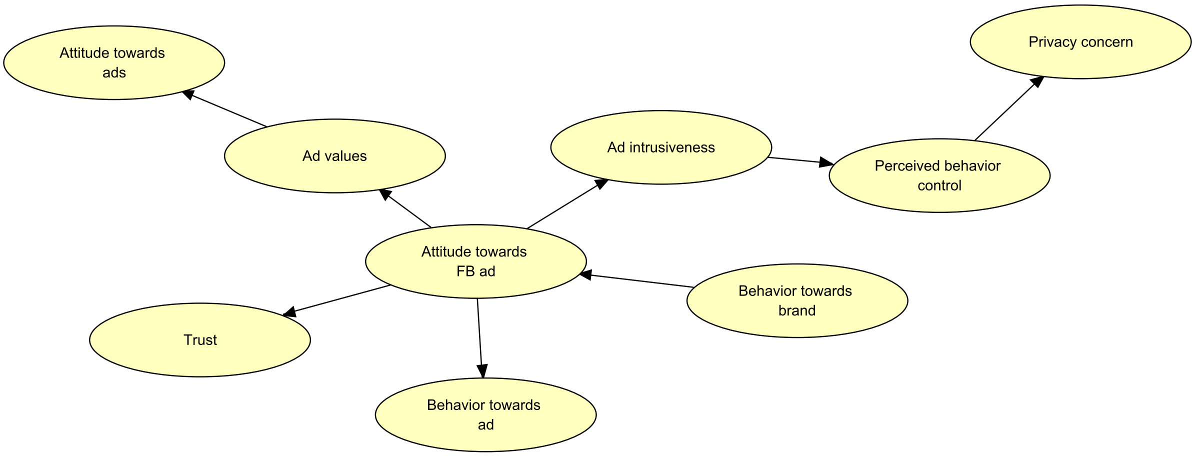


Figure 5.19: Semi-PSEM

Chapter 6

Conclusions

This chapter summarises the main results of the work done in this dissertation. Section 6.1 wraps up the key points of the dissertation and section 6.2 considers future work which can come from this dissertation.

6.1 Summary of Conclusions

Chapter 2 discussed the developments, both historic and current, as well as some interesting applications of SEM in the form of a literature review. The topic of SEM was followed in chapter 3, where the principles and processes behind SEM were covered. We then moved on to the field of graph theory in chapter 4. In chapter 4, we concluded that both SEM and BN are DAGs and it is possible for the structure of a SEM to be learned from data, using algorithms such as maximum weight spanning tree (MWST) and equivalence class (EQ) to find the initial network, cluster the variables according to mutual information, induce a factor for each cluster and learn the network among factors using Taboo algorithm to obtain an exploratory Bayesian Network (EBN).

This idea of EBN was applied along with other types of BNs in chapter 5. EBN is a BN derived from a data-driven perspective, parallel to the researcher's theory-based SEM. The researcher need not necessarily use the information from EBN but instead directly convert the SEM into a PSEM to conduct what-if analysis. It is also possible to specify the factors according to the theory and determine the structural path using the data. This results in a semi PSEM.

These BNs can offer significant insight into the data, as the researcher can then explore the data by instantiating on different nodes of the network (also called ‘what-if’ analysis). Because the direction of the inference is not an issue, various scenarios can be simulated using the BN.

The augmentation of SEM with BN provides significant contributions to the field:

Firstly, structural learning can mine data for additional causal information which is not necessarily clear when hypothesising causality from theory. This is particularly useful when two opposing theories exist (for example, whether the brand or positive media coverage is more effective in improving a company’s image) and the learned structure can confirm one theory above the other.

Secondly, the inference ability of the BN provides not only insight as mentioned before, but acts as an interactive tool as the ‘what-if’ analysis is dynamic. This has been found to be a powerful knowledge transfer platform, specifically in participatory research [8].

6.2 Future Work

Although using a tree structure can quickly find a network for the given variables, it cannot assign more than one parent node for a child node. This implies that data-driven methods such as MWST will not be able to suggest multiple causes for a single variable, unlike theoretical models suggested by the researcher. Therefore a possible research topic can be to find efficient algorithms which can offer network structures which are more complex than tree structures, possibly cyclic [25]

Another way in which this work can be extended is to apply it in other research fields as mentioned in section 2.2, such as finance, investment and economics. The theories which have governed in these fields can, with the help of techniques covered here, be confirmed or be given a new perspective [3].

Bibliography

[1] Barbara M Byrne. *Structural equation modeling with AMOS: Basic concepts, applications, and programming.* Routledge, 2010.

[2] Yingxia Cao, Haya Ajjan, and Paul Hong. Using social media applications for educational outcomes in college teaching: A structural equation analysis. *British Journal of Educational Technology*, 44(4):581–593, 2013.

[3] Chingfu Chang, Alice C Lee, and Cheng F Lee. Determinants of capital structure choice: A structural equation modeling approach. *The quarterly review of economics and finance*, 49(2):197–213, 2009.

[4] Stefan Conrady and Lionel Jouffe. Tutorial on driver analysis and product optimization with bayesialab, 2013.

[5] Stefan Conrady and Lionel Jouffe. *Bayesian Networks and BayesiaLab: A Practical Introduction for Researchers.* Bayesia USA, 2015.

[6] Adnan Darwiche. *Modeling and reasoning with Bayesian networks.* Cambridge University Press, 2009.

[7] A De Waal and T Ritchey. Combining morphological analysis and bayesian networks for strategic decision support. *ORiON*, 23(2):105–121, 2007.

[8] Meike Düspohl, Sina Frank, and Petra Döll. A review of bayesian networks as a participatory modeling approach in support of sustainable environmental management. *Journal of Sustainable Development*, 5(12):1, 2012.

[9] G David Garson. Path analysis. *from Statnotes: Topics in Multivariate Analysis. Retrieved*, 9(05):2009, 2008.

[10] Fred Glover. Tabu search: A tutorial. *Interfaces*, 20(4):74–94, 1990.

[11] Thomas F Golob. Structural equation modeling for travel behavior research. *Transportation Research Part B: Methodological*, 37(1):1–25, 2003.

[12] Thomas F Golob, Seyoung Kim, and Weiping Ren. How households use different types of vehicles: A structural driver allocation and usage model. *Transportation Research Part A: Policy and Practice*, 30(2):103–118, 1996.

[13] Peter Grünwald. Introducing the minimum description length principle. *Advances in minimum description length: Theory and applications*, page 3, 2005.

[14] Michael Haenlein and Andreas M Kaplan. A beginner's guide to partial least squares analysis. *Understanding statistics*, 3(4):283–297, 2004.

[15] Robert M Hauser and Arthur S Goldberger. The treatment of unobservable variables in path analysis. *Sociological methodology*, 3:81–117, 1971.

[16] Rick H Hoyle. *Handbook of structural equation modeling.* Guilford Press, 2012.

[17] Karl G Jöreskog. Analysis of covariance structures. *Multivariate analysis*, 3:263–85, 1973.

[18] Lionel Jouffe and Stefan Conrady. Probabilistic latent factor induction with bayesialab. April 2014.

[19] Sungduk Kim, Sonali Das, Ming-Hui Chen, and Nicholas Warren. Bayesian structural equations modeling for ordinal response data with missing responses and missing covariates. *Communications in StatisticsTheory and Methods*, 38(16-17):2748–2768, 2009.

[20] Rex B Kline. *Principles and practice of structural equation modeling.* Guilford publications, 2015.

[21] Jonathan Kohn and Sarah K Bryant. Factors leading to the us housing bubble: A structural equation modeling approach. *Research in Business and Economics Journal*, 3:D1, 2011.

[22] Kevin B Korb and Ann E Nicholson. *Bayesian artificial intelligence*. CRC press, 2010.

[23] Reinhold Kosfeld and Jørgen Lauridsen. Factor analysis regression. *Statistical Papers*, 49(4):653–667, 2008.

[24] Timo Koski and John Noble. *Bayesian networks: an introduction*, volume 924. John Wiley & Sons, 2011.

[25] Pedro Larrañaga, Mikel Poza, Yosu Yurramendi, Roberto H. Murga, and Cindy M. H. Kuijpers. Structure learning of bayesian networks by genetic algorithms: A performance analysis of control parameters. *IEEE transactions on pattern analysis and machine intelligence*, 18(9):912–926, 1996.

[26] Ross L Matsueda and Guilford Press. Key advances in the history of structural equation modeling. *Handbook of structural equation modeling*, pages 17–42, 2012.

[27] Paul Munteanu and Mohamed Bendou. The eq framework for learning equivalence classes of bayesian networks. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 417–424. IEEE, 2001.

[28] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[29] Richard E Neapolitan et al. *Learning bayesian networks*, volume 38. Pearson Prentice Hall Upper Saddle River, NJ, 2004.

[30] Norm O'Rourke and Larry Hatcher. *A step-by-step approach to using SAS for factor analysis and structural equation modeling*. Sas Institute, 2013.

[31] Randall E Schumacker and Richard G Lomax. *A beginner's guide to structural equation modeling*. Psychology Press, 2004.

[32] Charles Spearman. " general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904.

[33] Sheridan Titman and Roberto Wessels. The determinants of capital structure choice. *The Journal of finance*, 43(1):1–19, 1988.

[34] Xiao-fu Xu, Jian Sun, Hong-tao Nie, De-kui Yuan, and Jian-hua Tao. Linking structural equation modeling with bayesian network and its application to coastal phytoplankton dynamics in the bohai bay. *China Ocean Engineering*, 30(5):733–748, 2016.

Appendix A

Maximum weight spanning tree

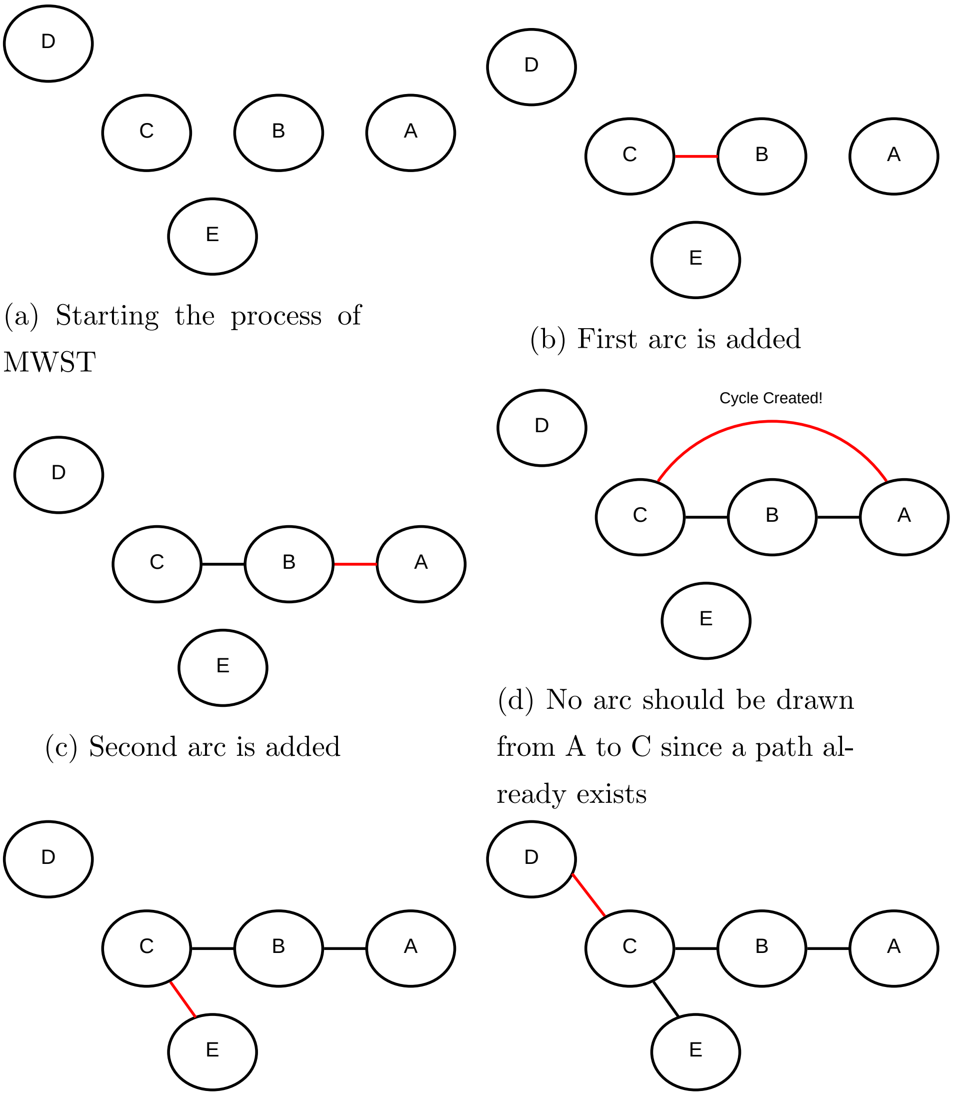
A.1 MWST example

Let us suppose there are 5 variables with which we want to create a MWST. The first step we need to take is to calculate mutual information as given by equation 4.8 for all possible pairs of variables.

Table A.1: Pairwise Mutual information, sorted descending

Var1,Var2	MI
B,C	0.83
A,B	0.71
A,C	0.63
C,E	0.58
B,E	0.22
A,E	0.19
C,D	0.15
A,D	0.13
B,D	0.11
D,E	0.08

Table A.1 shows a fictitious set of values for the mutual information of 5 variables, sorted descending according to their MI.



(a) Starting the process of MWST

(b) First arc is added

(c) Second arc is added

(d) No arc should be drawn from A to C since a path already exists

(e) Third arc is added

(f) Fourth arc is added

Figure A.1: Directed and undirected trees.

Initially we start off with a fully unconnected network as shown in figure A.1a. Next we start connecting the pair of variables as we move down the rows of table A.1. The highest value of MI is between variables B and C and so we draw an arc between those two nodes, illustrated in figure A.1b. The next highest MI is present between A and B so an arc is drawn to connect those two (figure A.1c).

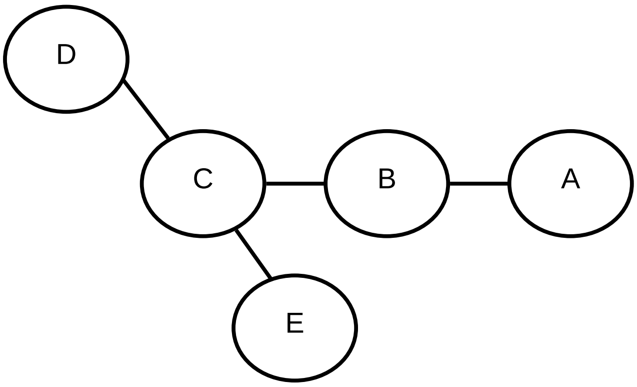


Figure A.2: Final Tree

The next highest MI is between variables A and C but we should not connect those two, as there is already a path between those two variables (through$B$). Put differently, drawing an arc between A and C creates a cycle with variables A, B and C as shown in figure A.1d. Hence we do not change the network and proceed to the next highest MI. The next arc added is between C and E. This means the next two paths (B, E & A, E) will not be drawn (figure A.1e). Finally, an arc is drawn between C, D and this causes all subsequent MIs to be forbidden (figure A.1f). The final tree is given by figure A.2.

Appendix B

Facebook advertisement data questionnaire

Table B.1: Questionnaire for Facebook advertisement

Q8Q1	Facebook is a trustworthy social network
Q8Q2	Facebook can be relied on to keep its promises
Q8Q3	Even if not mentioned, I would trust Facebook to do the job right
Q8Q4	I believe that Facebook would use my data only for purposes that I have approved
Q8Q5	I can count on Facebook to protect my privacy
Q8Q6	I can count on Facebook to protect my personal information from unauthorized use
Q9Q1	I consider advertising a good thing
Q9Q2	In general, I like advertising
Q9Q3	I consider advertising essential
Q9Q4	Having advertisements are important to me
Q9Q5	Advertisements in general are interesting to me
Q9Q6	I would describe my overall attitude towards advertising as favourable
Q11Q1	I consider ads on my Facebook page a good thing
Q11Q2	I like ads on my Facebook page
Q11Q3	I consider ads on my Facebook page essential
Q11Q4	Having ads on my Facebook page are important to me
Q11Q5	Ads on my Facebook page are interesting to me
Q11Q6	I would describe my overall attitude towards ads on my Facebook page as favourable

Table B.2: Questionnaire continued

	When I see an advertisement on my Facebook page, I generally
Q12Q2	click on the ad to find more information
Q12Q3	‘like’ or ‘comment’ on the ad
Q12Q4	‘share’ or ‘repost’ the ad to my friends
Q12Q11	become a fan of the company/brand
Q12Q12	visit the company/brands website
Q12Q13	purchase the advertised product/service
	It is important to me that I can
Q14Q1	only receive ads on my Facebook page if I have previously provided permission
Q14Q2	control the permission to receive ads
Q14Q3	refuse to receive advertising on my Facebook page
Q14Q4	filter advertising on my Facebook page to match my needs
	I find advertisements on my Facebook page
Q16Q1	distracting
Q16Q2	intruding on my privacy
Q16Q3	interfering
Q16Q4	invading my privacy
Q16Q5	deceptive
Q16Q6	confusing
Q16Q7	annoying
Q16Q8	irritating
Q16Q9	compromising my privacy
	Facebook advertising is
Q17Q1	useful
Q17Q2	valuable
Q17Q3	important
Q20Q1	All things considered, the Internet causes serious privacy problems
Q20Q2	Compared to others, I am more sensitive about the way online companies handle my personal information
Q20Q3	To me, it is very important to keep my privacy intact/unharmed from online companies
Q20Q4	I believe other people are not concerned enough with online privacy issues
Q20Q5	Compared to other subjects on my mind, personal privacy is very important
Q20Q6	I am concerned about the threat to my personal privacy today

Appendix C

Facebook data SEM using AMOS

C.1 Summary

Figure C.1 shows how the SEM for the Facebook data was constructed in SPSS AMOS. All path coefficients as well as error variance were significant. Figure C.2 shows the diagram with estimated standardised path coefficients and squared multiple correlation for endogenous variables.

The value of -.13 between $Advertising\,intrusiveness$ and $Attitudes\,towards\,FB\,ad$ indicates that there is an inverse relationship, albeit relatively weak, between the two variables, where an increase of 1 standard deviation in $Advertising\,intrusiveness$ will lead to a decrease of 0.13 standard deviations in $Attitudes\,towards\,FB\,ad$. A strong positive relationship exists between $Attitudes\,towards\,FB\,ad$ and $Behaviour\,towards\,ad$, indicated the coefficient value of .79. The value of .63 for squared multiple correlation of $Behaviour\,towards\,ad$ shows that $Attitudes\,towards\,FB\,ad$ explains 63% of the variance in $Behaviour\,towards\,ad$. Other values in the diagram can be interpreted in the same way.

Furthermore, table C.1 shows values for the model goodness of fit. CFI is larger than 0.9 while $RMSEA$ is less than 0.055, which are both indicative of an adequate overall model fit.

Table C.1: Goodness-of-fit Facebook data SEM

Model fit index	Index value
CFI	0.926
TLI	0.922
RMSEA	0.051
RMSEA upr90	0.054
RMSEA lwr90	0.049

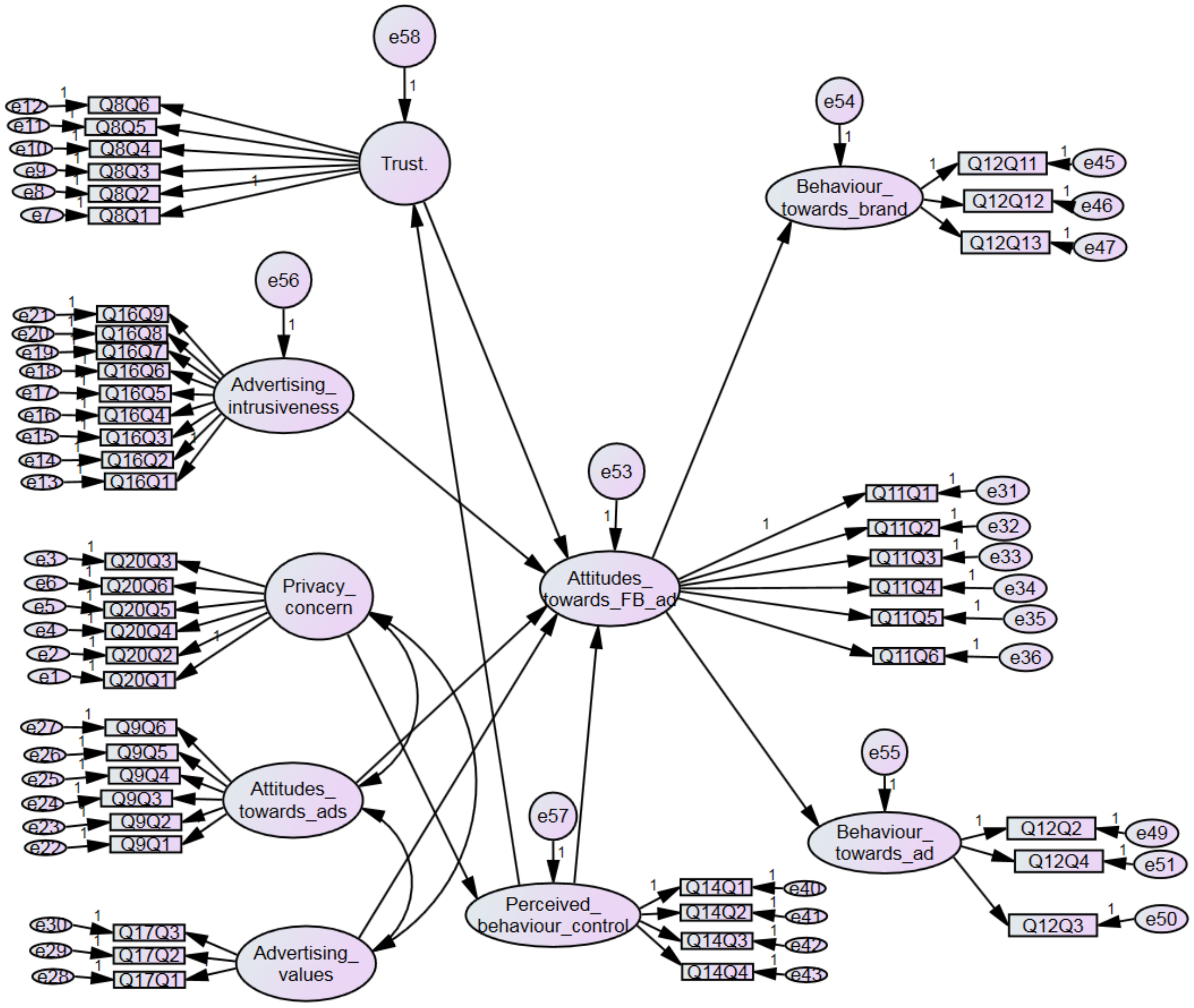


Figure C.1: Theoretical SEM for Facebook data, as drawn in SPSS AMOS

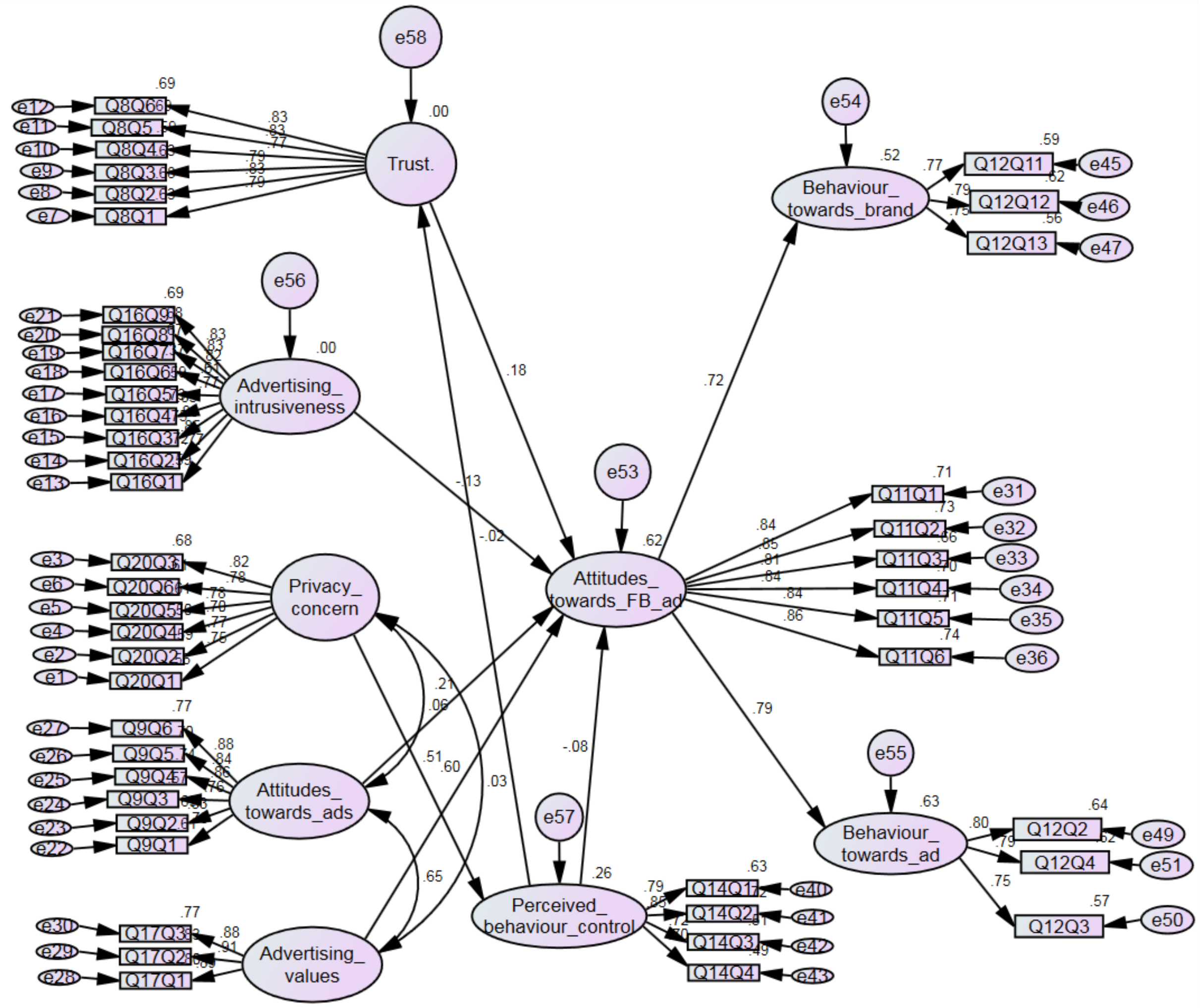


Figure C.2: Estimated coefficients for Facebook data SEM