I²SDS
The Institute for Integrating Statistics in Decision Sciences

Technical Report TR-2008-11
June 30, 2008

Market Basket Analysis Using Bayesian Networks

Xiaojun Li
Capital One, USA

Sanal Mazvancheryl
Kogod School of Business
American University, USA

Srinivas Prasad
Department of Decision Sciences
The George Washington University, USA

Pradeep Rau
Department of Marketing
The George Washington University, USA

Refik Soyer
Department of Decision Sciences
The George Washington University, USA

Market Basket Analysis Using Bayesian Networks

Xiaojun Li
Capital One, Richmond, VA. {xiaojun.li@capitalone.com}
Sanal Mazvancheryl
American University, Washington DC.
Srinivas Y. Prasad, Pradeep Rau, Refik Soyer
School of Business, The George Washington University, Washington DC 20052.
{prasad@gwu.edu, prau@gwu.edu,soyer@gwu.edu}

This paper addresses the question of how promotions work across categories. Promotions in one product category can affect sales of products in another category either directly or indirectly. Given a set of product categories and market basket data, we analyze the presence of cross category impacts using Bayesian Networks. We model the occurrence of a product category, and not the number of units (of a product category) in a basket. The data set we employ is an IRI market basket data set that contains transactions including 22 categories over 2 years for 500 panelists. Bayesian networks are learned from this data and are used to identify the underlying dependencies across product categories. Specifically, we study how the associations across categories vary based on marketing mix activities, and also based on demographics. The results from such an analysis can help in 1) identifying clusters of categories wherein associations exist primarily between categories within a cluster and not across clusters, and 2) in making predictions on basket choices given a set of specific marketing mix activities. The ability of Bayesian networks to learn based on new evidence also makes such an approach possible in an online context when customers' choices can be observed, and marketing activities can be dynamically customized.

Key words: Market Basket Analysis; Bayesian Networks;

1. Introduction

An area of considerable interest among marketing research practitioners and scholars in recent years has been the nature of consumer behavior as reflected in their shopping baskets at the retail level. A number of studies (Russell and Kamakura, 1997; Manchanda, Ansari and Gupta, 1997; Russell and Peterson, 2000) have explored the patterns of occurrence of product categories in individual market baskets and the implications of the category relationships for retailer decisions in the areas of pricing, promotion, store displays etc. These

studies have generally looked at the product categories in pairs based on some predetermined relationship between them ex. complements (ex. detergent and fabric softener) or substitutes (ex. butter and margarine). While these efforts have led to an interesting set of implications for retailer decisions it would appear that a more generalized approach which considers relationships across a wide range of product at the retail level that would uncover pairs of categories that would empirically be of interest and the consequent effects of retailer decisions (pricing, promotion etc.) and consumer demographics on those category pairs would be very useful for retailers in optimizing the effect on sales and profits. In an earlier paper (Li, Mazvancheryl, Prasad and Rau, 2008) we have undertaken an approach to uncovering such product category pairs using a data mining approach where a number of possible category pairs can be arrayed by lift value on a continuum from complements to substitutes with the intervening pairs displaying intermediate levels of relationship based on a large number of shopping baskets. We have argued there that a large number of category pairs may have a moderate level of relationship (based on their lift value) in shopping baskets that could then form the basis for a number of retailer decisions that could optimize the sales/profit performance of those pairs. In a related manner, we have also considered how household (consumer) demographics and shopper decisions impact the likelihood of those product pairs occurring jointly in shopping baskets.

To the best of our knowledge, what has been lacking in most of the existing research described above is a generalized approach to actually picking the product pairs of interest occurring jointly in shopping baskets rather than on a priori criteria such as pairs that are substitutes or complements etc. In this paper, we attempt to develop an algorithm to uncover such pairs of interest using a Bayesian network approach. While such an approach has been employed by at least one recent study (Bezawada, Balachander, Kannan and Shankar, 2009) it was once again designed to uncover the effects of aisle and display placements on pre-selected pairs of categories (ex. cola and chips etc.) whereas our recommended approach involves the market basket data itself suggesting pairs of categories of interest for which the effects of store promotion and consumer demographics could be explored so as to maximize the effectiveness of retailer decisions. As a possible illustration of the usefulness of such an approach, category pairs such as bananas and cereal (which are interestingly being displayed together recently in some supermarkets) could be uncovered and retailer decisions on these pairs could then be made for maximum effectiveness of retailer outcomes.

1.1. Cross-Category Models

Consumers typically purchase multiple products from multiple categories in one shopping trip. Advances in information technology have made the collection and storage of basket level transaction data economically and technically feasible. It is in the retailers' interest to leverage information hidden in these data so as to increase store profit. One useful piece of information is how promotions work, and in particular how they work across categories.

A broader definition of promotion includes not only price discount and couponing, but also display and feature advertising activities, etc. Marketing researchers have found that promotions can either change the magnitude of consumers' purchases and/or enhance store traffic. A thorough review of how promotions work can be found in Blattberg, Briesch, and Edward (1995). One important finding from past research is that promotion in one category affects sales in complementary categories and substitute categories. This would suggest that retail pricing strategy should incorporate demand interdependencies such as complementary and substitute to maximize store profitability (Mulhern and Leone 1991).

Two product categories can be use complements such as cake mix and frosting, substitutes such as butter and margarine, or just independent. Since it is hard to observe use complements by a retailer, we use the idea of purchase complements. A pair of categories are defined to be purchase complements "if marketing actions (price and promotion) in one category influence the purchase decision in the other category"(Manchanda et al. 1999). A more detailed explanation of complements and substitutes is given in (Manchanda et al. 1999, Russell et al. 1999).

According to consumers' purchasing decisions, Seetharaman et al. (2005) classify cross-category models into incidence models, brand choice models, and quantity models. Incidence models can be further classified into "whether to buy" models, "when to buy" models, and "bundle choice" models. This essay extends the "whether to buy" models to explicitly identify category clusters.

Both statistical and data mining models have been proposed to understand promotions among multiple categories. Approaches to measure cross category effects in the marketing literature have been theory driven, and typically based on utility theory models. These utility theory based models usually decompose the utility of a category into its own effects and cross-category effects. Some models go further to isolate cross-category marketing mix effects from cross-category co-incidence effect, examples being multinomial logistic models

(Russell and Peterson, 2000) and multivariate probit models (Manchanda et al. 1999, Chib et al. 2002). These researchers all agree that promotions in one category have impacts on its own sales. However they differ in how to model the cross category promotion effects. There are three paradigms in modeling cross category promotion effects. One is marketing actions effects choices of other categories directly (Manchanda et al. 1999). In this type of model, the consumers' decision process is viewed as a black box. The final choices in a basket are modeled as a function of marketing mix variables. A second paradigm is that promotions in a category only effect its own sales, but presence of this category affects decisions in other categories (Russell et al. 1999, Hruschka et al. 1999). Therefore, marketing mix activities impact indirectly across categories. The last paradigm attempts to model category choices using both promotions and presence of other categories.

The quality of these statistical approaches depends on the practitioner's choice of modeling paradigm. In this essay, we propose a data mining paradigm which learns the cross category promotion effects based on data. We use Bayesian networks to learn the dependencies among variables suggested by the data. Advances in Bayesian network make it possible to learn the multivariate relationship from data. Model uncertainty is also considered using Bayesian networks. Bayesian networks were first introduced as an expert system tool. Because they have both causal and probabilistic semantics, Bayesian networks can represent causal relationships in a problem domain. They can also be used to predict the consequences of intervention. Bayesian networks have several other advantages as a research tool(Heckerman 1996). For example, they can handle missing data readily and avoid overfitting of data.

Bayesian networks have been used to model association among products in past research(Giudici and Passerone 2002, Giudici and Castelo 2003). Note however that our research differs in that we model not only co-occurrence but also how marketing mix works across multiple categories using transaction level market basket data as opposed to aggregated basket data used in earlier research (Giudici and Passerone 2002).

This essay is organized as follows. Section 2 is a brief introduction on Bayesian networks. Section 3 discusses the details of learning a Bayesian network of discrete data. Section 4 applies the Bayesian network to retailing data and summarizes the research.

2. Bayesian networks

Intuitively, a Bayesian network is a graphical representation of conditional independence and dependence among variables regardless numerical or functional details. Also known as Directed Acyclic Graph (DAG) or belief network, a Bayesian network is a type of graphical model that combines the science of statistics and graph theory. As shown in Figure 1, a Bayesian network consists of:

- A directed graph with nodes representing variables and edges representing dependencies.
- A set of probability distributions associated with the edges.

In this figure, A is called the parent of B, and B is parent of C. A and C are said to be conditionally independent given B since $p(C|A \cap B) = p(C|B)$.

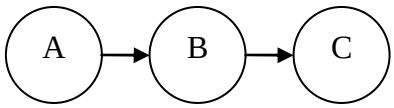

Figure 1: An example of Bayesian network

More rigorously, let $X = x_1, ..., x_n$ represent a set of variables and S represents a network structure. S needs to encode a set of conditional independence statements about X. P represents the set of local probability distribution of each variable in X, $p(x_i|Pa_i, \xi)$. Pa_i denotes the parents of node x_i in structure S and the corresponding variable in X. An important property of Bayesian network is the chain rule, which means the joint probability distribution for X is

$$p(X) = \prod_{i=1}^{n} p(X_i|Pa_i, \xi) \tag{1}$$

As we see, a Bayesian network can be viewed as a collection of local probabilistic/regression models. As we observe the state of some nodes in the network, we can update the probability of other nodes' states.

2.1. Graphical Representation of Cross-Category Effects

To represent cross category marketing effects, we need to define two types of nodes in the network. They are:

1. X_i : Marketing mix variable for product i. It is a binary variable with $X_i = 1$ meaning there is a promotion for the product. We do not need to estimate $p(X_i)$, which is up to the retailer.

2. Y_i : Purchase decision of product i. It is a binary variable with $Y_i = 1$ meaning product purchase.

In this study, we have partial knowledge about the underlying network structure.

- A category node's parents can be either marketing variables or other category nodes.
- A marketing variable does not have any parent node.

Thus, the three cross-category effects modeling paradigms mentioned earlier can be represented as in

1. Figure 2: Direct Paradigm. Sales of category B is directly dependent on category A's promotions.

2. Figure 3: Indirect Paradigm. Sales of category B is dependent on the purchase decision of category A.

3. Figure 4: Mixed Paradigm. Sales of category B is dependent on category A's both promotions and purchase decision.

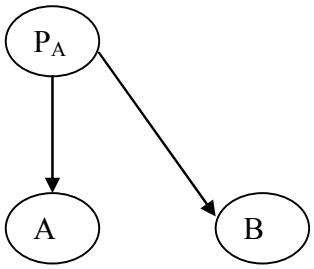

Figure 2: Direct Paradigm

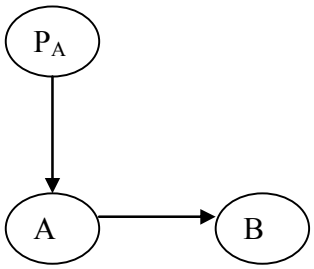

Figure 3: Indirect Paradigm

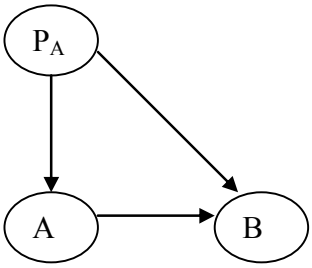

Figure 4: Mixed Paradigm

2.2. Learning Bayesian Networks

A Bayesian network can be constructed based on an expert's knowledge. In sch a context, both statistical parameters and network structure are specified by domain experts. Once actual values of some variables are observed, the Bayesian network updates the posterior probabilities of other variables by an inference process. For example, one can use Bayes' rules to reverse the arcs step by step in the network until the requested probabilities are answered (Shachter 1988). However, either exact inference or approximate inference is NP-hard (Cooper 1990,Dagum and Luby 1993). For Bayesian networks of discrete variables, the most commonly used algorithm is the junction tree algorithm(Lauritzen and Spiegelhalter 1988, Jensen and Lauritzen 1990, Dawid 1992).Probabilistic inference is performed using several mathematical properties of the junction tree.

So far we have assumed the network structure is known. In domains where we have little knowledge, machine learning techniques are enlisted to help learn structures from data. The most straightforward approach is to compare all possible networks based on some measure. The network that optimize this measure will be selected as the best Bayesian network. The challenge in this approach is the number of possible networks given n nodes explodes as n increases.Table 1 gives examples of candidate network counts. There is no closed form formula known for the number of structures. Robinson (1977) gives the following recursive

formula

$$f(n) = \sum_{i=1}^{n} (-1)^i \frac{n!}{(n-i)!i!} 2^{i(n-i)} f(n-i)$$

number of variables	*number of possible DAGs*
2	3
3	25
4	543
5	29,000
10	4.2×10^{18}

Table 1: Number of Network Structures

Given a number of competing networks, we need both a measure and a search strategy to find the most promising network. Measures used include maximum likelihood, predictive assessment, and posterior probabilities. As Chickering (1996) shows, it is NP-hard to learn the structure of a Bayesian network. A variety of heuristic search algorithms has been introduced, such as greedy search, greedy search with restarts, best-first search, and Monte-Carlo methods (Heckerman 1996). The most straightforward search and scoring approach is greedy search. Here is a brief introduction of the greedy search algorithm. Let E represents all eligible changes to a Bayesian network and use log of relative posterior probability as the network score.

$$\log P(D, S^h) = \log P(S^h) + \log P(D|S^h)$$

Let $\delta(e)$ represent the changes of the network score caused by change e.

- Choose an initial network.
- Change one edge in the network at a time and evaluate the change. Pick the one with maximum $\delta(e)$.
- Stop the search when no e can make a positive contribution.

This approach may hit a local maximum. To escape from local maxima, we need to restart the search process randomly with a new initial network. Another way to find a global maxima is to use approximation approach such as Markov Chain Monte Carlo Model Composition, or MC^3 (Madigan and York 1995).

If one has some partial knowledge of causal relations among variables, the search space can be reduced by ruling out unlikely models or by placing a restriction that a node can have at most u parents ($u < n - 1$) (Cooper and Herskovits 1992). It is very likely there is no single dominant model learned. In this case, instead of selecting a single true model, model uncertainty is accounted by averaging all the models.

Let Δ be the quantity of interest. Its posterior distribution conditional on data D is ,

$$pr(\Delta|D) = \sum_{k=1}^{K} pr\left(\Delta|M_k, D\right) pr(M_k|D) \tag{2}$$

To find the M_k and their posterior probability $pr(M_k|D)$, an algorithm called Markov Chain Monte Carlo Model Composition (MC^3) (Madigan and York 1995) can be used. Markov Chain Monte Carlo (MCMC) is a simulation method that generates samples from complex and nonstandard distributions. Developed by Metropolis et al. (1953), and generalized by Hasting, the Metropolis-Hastings algorithm is an implementation of the Markov Chain Monte Carlo method.

A brief introduction of the Metropolis-Hastings algorithm follows - for a more detailed introduction, see Chib and Greenberg (1995). A MCMC algorithm draws samples of a target probability density $\pi(x)$ by constructing a Markov chain, which converges to the target probability distribution. Define the candidate generating density as $q(x, y)$, from which a value y is generated when the process is in state x. When $\pi(x)q(x, y) > \pi(y)q(y, x)$, the process moves from x to y more often than from y to x. Thus we need to specify a probability of move $\alpha(x, y)$ to meet the requirement of reversibility,

$$\pi(x)q(x, y)\alpha(x, y) = \pi(y)q(y, x)\alpha(y, x)$$

Given an initial state $X^{(0)}$,

- Repeat for $j = 1, ..., N$
- Generate y from $q(x^{(j)}, \cdot)$ and u from $U(0, 1)$
- IF $u \leq \alpha(x^{(j)}, y)$, set $x^{(j+1)} = y$
- ELSE set $x^{(j+1)} = x^{(j)}$
- return the values $\left\{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\right\}$

MC^3 is a Metropolis-Hastings based method, which build a Markov chain of graphs that will converge to the distribution of the model that generates the data. MC^3 has been successfully used for linear regression models (Raftery et al. 1997).

3. Bayesian Network of Discrete Data

This section introduces the technical details of learning Bayesian networks of discrete data using the MC^3 algorithm. We assume all variables in the network are discrete and follow Dirichlet distribution (Heckerman 1996). More discussion of the Markov Chain Monte Carlo Model Composition algorithm can be found in Madigan and York (1995), and Giudici and Castelo (2003).

3.1. Network Specifications

Now let g be a Bayesian network of variables $\mathbf{X}$. Each variable $X_i \in \mathbf{X}$ is discrete, $X_i = x_i^1, ...x_i^{r_i}$. Denote X_i's parents as Pa_i

$$p(x_i^k | Pa_i^j, \theta_i, g) = \theta_{ijk} > 0 \tag{3}$$

Assume data completeness and parameter independence, the joint probability of the parameters is

$$p(\theta_g | D, g) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} p(\theta_{ij} | D, g) \tag{4}$$

For discrete variables, we will assume Dirichlet prior distribution:

$$p(\theta_{ij} | g) = Dir(\theta_{ij} | \alpha_{ij1}, ..., \alpha_{ijr_i}) \tag{5}$$

where r_i is the number of realizations under such a configuration. Assume a uninformative assignment with equivalent sample size, $\alpha_{ijk} = 1/(r_i \times q_i)$, q_i being the number of configurations of parents. The Posterior distribution is:

$$p(\theta_{ij} | D, g) = Dir(\theta_{ij} | \alpha_{ij1} + N_{ij1}, ..., \alpha_{ijr_i} + N_{ijr_i}) \tag{6}$$

where N_{ijk} is the number of observations in dataset D with $X_i = x_i^k$ and $Pa_i = Pa_i^j$

Following the four assumptions given in Cooper and Herskovits (1992):

1. All the variables are discrete.

2. Given the Bayesian network model, the observed cases are independent.

3. There is no missing value.

4. Before observing data set D, we are indifferent regarding the numerical probabilities to assign to a given network structure.

Based on the above, Coopers and Herskovits (1992) show that the marginal likelihood of a discrete DAG model is given by

$$L(g) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(N_{ij} + \alpha_{ij})} \prod_{i=1}^{r_i} \frac{\Gamma(N_{ijk} + \alpha_{ijk})}{\Gamma(\alpha_{ijk})} \tag{7}$$

where i is the index of nodes in the network, j is the index of its parents' configurations.. $\alpha_{ij} = \sum \alpha_{ijk}$. and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$.

3.2. Learning network structure using MC^3

Learning the network structure is accomplished by constructing a Markov chain of DAG's. The irreducibility of a DAG is guaranteed in the case of a DAG. However,acyclicity needs to checked for each graph. Define the neighborhood of a given graph g as nbd(g), which is the collection of graphs that can be reached from g in one step by adding or deleting one edge, including itself. There are three types of moves, addition, deletion, and reversal. Reversal of an arc can be viewed as to perform a removal move first followed by an addition move.A move can not generate directed cycles. One way to guarantee no directed cycle is to traverse the whole network.

When the chain moves from g to g', $g' \in nbd(g)$, the acceptance probability of the move is

$$min\{1, \frac{\#(nbd(g))p(g'|D)}{\#(nbd(g'))p(g|D)}\}$$

where $\#(nbd(g)$ represents the cardinality of graph g's neighborhood. If the move is not accepted, the chain stays in state g.

Since g and g' are neighbors, it is reasonable to assume that

$$\frac{\#\left(nbd\left(g\right)\right)}{\#\left(nbd\left(g'\right)\right)} \approx 1$$

Since $p(g|D) \propto p(D|g)p(g)$, the acceptance ratio relates only to the Bayes factor $\frac{p(D|g')}{p(D|g)}$. The Bayes factor in case of addition and deletion is

$$\frac{L(g', i)}{L(g, i)} \tag{8}$$

where i is the variable whose parent set is different in g and g'. Bayes factor in case of reversal is

$$\frac{L(g', i)L(g'', j)}{L(g, i)L(g, j)} \tag{9}$$

Recall that

$$L\left(g, i\right) = \prod_{j=1}^{q_i} \frac{\Gamma\left(\alpha'_{ij}\right)}{\Gamma\left(N_{ij} + \alpha'_{ij}\right)} \prod_{k=1}^{r_i} \frac{\Gamma\left(N_{ijk} + \alpha'_{ijk}\right)}{\Gamma\left(\alpha'_{ijk}\right)} \tag{10}$$

3.3. Average Network

As the number of nodes increase, the result of MC^3 learning are many DAGs with small likelihood. We can apply Bayesian model averaging to present the relation among the nodes in one overall network (as in Giudici and Castelo, 2003). Let e be an edge in a graph. $P(e|D)$ measures the probability of its presence given data D. Only those edges whose $P(e|D) > \mu$ will be drawn in an aggregate graph.

$$P(e|D) = \sum_i I(e|g_i) * P(g_i|D)$$

where

$$I(e|g_i) = \begin{cases} 1 & if\ e \in g \\ 0 & if\ e \notin g \end{cases}$$

These edges can be plotted to create an average network. The average network can be used to identify clusters of variables. Note that the average network constructed in this manner may not be a valid DAG.

4. Application in Marketing

There are two goals in this application. First, we want to identify product clusters using Bayesian network given presence of other products, promotions, or customer demographics. We believe once these clusters are identified, they can provide a retailer with useful information on planning marketing activities or segmenting customers. Second, we specifically want

to evaluate pairwise cross-category relationship given presence of other products, promotions, or customer demographics. This is more in line with traditional marketing research. We build three Bayesian network models. They are:

- Model 1: Using Bayesian network to map the relationship of the products only. This is an equivalent model of (Giudici and Passerone 2002, Giudici and Castelo 2003).
- Model 2: Using Bayesian network to model product relationship with marketing mix variables included.
- Model 3: Using Bayesian network to model product relationship with customer demographics variables included.

The data set include sales data from multiple grocery stores in a Metropolitan area. Also included are marketing data and consumer demographic data. Twelve categories are chosen for the analysis. They are detergent, softener, towel, tissue, yogurt, cereal, soap, cleanser, hotdog, egg, cookie, and cracker. These categories include products which have different functions or closely related.

Section 4.1 introduces implementation and validation the MC^3 algorithm. Section 4.2 and 4.3 discusses details of product clustering and pairwise assessment. Section 4.4 summarizes findings and discusses managerial implications of Bayesian network in marketing research.

4.1. Algorithms Implementation and Validation

The analytic software used is developed on the basis of Bayesian Network Inference with Java Objects or BANJO, which is an open source software originally developed by Duke University researchers. Here is a brief description of the implementation of the MC^3 algorithm.

1. Given an initial network.
2. Randomly propose a move. It can be addition, deletion, or reversal of an edge.
3. Check whether the proposed move leads to cycles. If it does, repeat from step 2.
4. Evaluate the move according to the criteria given in last chapter.
5. Decide if the move is accepted so that Markov chain gets into a new state. Otherwise it stays in the same state. Repeat from step 2 until the chain converges.
6. Maintain a database of Bayesian networks. Two key pieces of information, structure of the network, and its frequency, are kept for calculation of the average network.

As the number of nodes increases, the number of legal networks increases exponentially. The Markov chain will converge very slowly. To speed up, instead of starting the Markov Chain from a random generated network, we start it from a network structure learned using greedy search algorithm. We find that this speeds up the convergence by more than a factor of two.

In this research, we use two random samples to validate the MC^3 algorithm. Two measures are used to validate the algorithm. Let e_{ij} be the edge from node i to node j. Given two datasets D_1 and D_2, E_1 represents all edges learned in D_1 and E_2 represents all edges learned in D_2.

The first measure is on the difference of each edge's posterior probability. Let $P(e_{ij}|D)$ represents the probability of edge e_{ij} being present in data set D. The difference of the two datasets is captured in $\delta_{ij} = P(e_{ij}|D_2) - P(e_{ij}|D_1)$. Using sales data of the twelve categories, there are 41 edges learned from the two networks. Figure 5 is the histogram of the differences, from which we can see there is no major difference between the two edge sets.

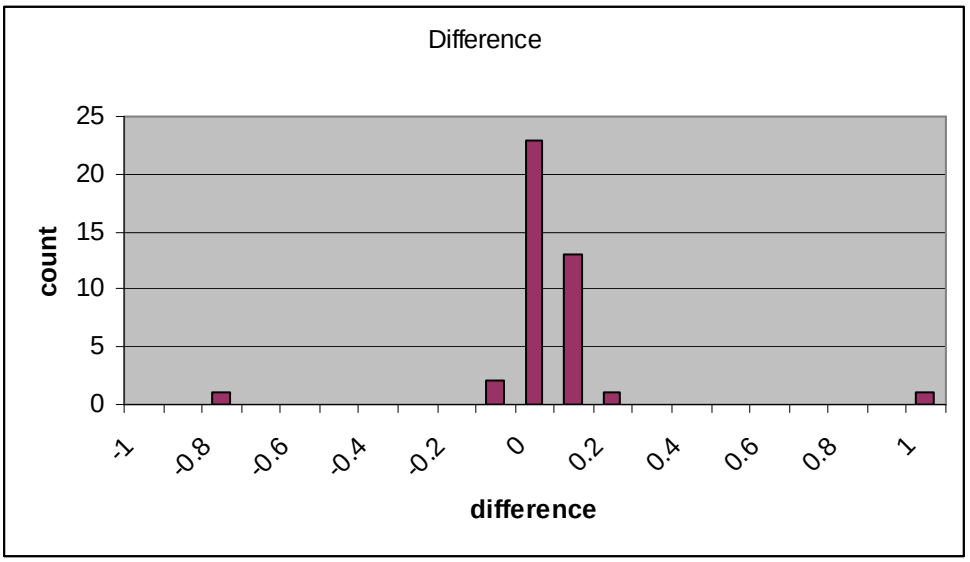

Figure 5: Difference between edge sets

We now present some validation statistics based on the similarity of edges generated in the network structures learned from the two random data sets. At a given cutoff point, suppose the average network from data set D_1 has a set of edges E_1. The size of E_1, which is the number of edges in it, is m_1. Average network from dataset D_2 has a set of edges E_2 with a size m_2. Let m_{12} represent the size of $E_1 \bigcap E_2$. Define validation rate $r = \frac{m_{12}}{m_1}$. r is a

number between 0 and 1. Set cutoff point at $0.9, 0.7$, and 0.5, the validation rate is 0.9, 0.9, and 0.91. Based on the above, we can reasonably conclude that the MC^3 learning algorithm generates valid network structures.

4.2. Product Clusters Identification

4.2.1. Results of Model 1

We start with 12 category sales data only. Based on the average network with a cutoff value 0.9, there are one large network with most categories involved, and several one-category clusters. They are (tissue, towel, cleanser, detergent, softener, cereal, cookie, hotdog, cracker), (egg), (yogurt), and (soap), see Figure 6. This result is different from Giudici and Castelo (2003). In Giudici and Castelo (2003), there are two-category, three-category, and five-category clusters. The difference might be due to two facts. First, their sales data are weekly aggregate data. Second, they use 1 to indicate the weekly sale of some category is greater than median, otherwise 0. The possible explanation for the absence of clusters in our findings can be that most product categories are related to each other at the transaction level.

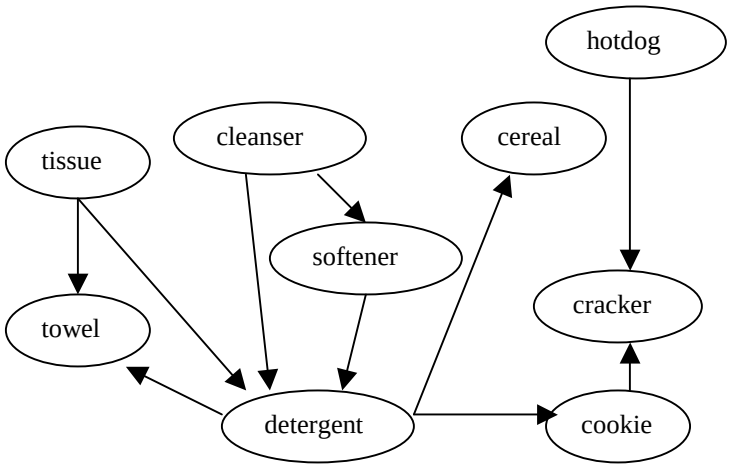

Figure 6: Product cluster

4.2.2. Results of Model 2

We now add promotion information in the learning process. There are three types of marketing activities available in the IRI data. They are price discount, display, and feature. In this application, these three variables are combined into one binary promotion variable to

indicate if there is at least one marketing activity for the category. We set a limit on the MC^3 learning process in this model. That is, there should be no parents for any promotion variable. At cutoff point 0.9, we identify three clusters, which are shown in Figure 7, 8, and 9.

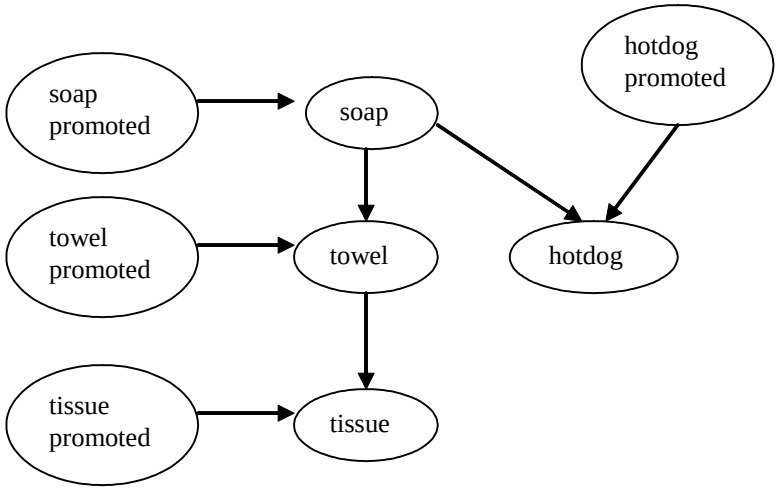

Figure 7: Product cluster one with promotions included, cutoff at 0.9

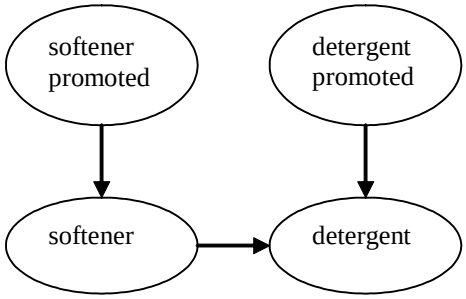

Figure 8: Product cluster two with promotions included, cutoff at 0.9

We find that cross category promotions work through indirect effects primarily. That means, promotions in category A impacts sales of A. In turn the sale or no-sale decision of category A will influence sales of its complement (or substitute) category B. Thus promotions in category A indirectly impact sales of category B. Using a cutoff value 0.9, there are three multi-category clusters. They are (towel, tissue, soap, hotdog), (softener, detergent), and (cereal, cracker, cookie). Using a cutoff value 0.6, the three clusters are merged into two cluster,which are shown in Figure 10, and Figure 11.

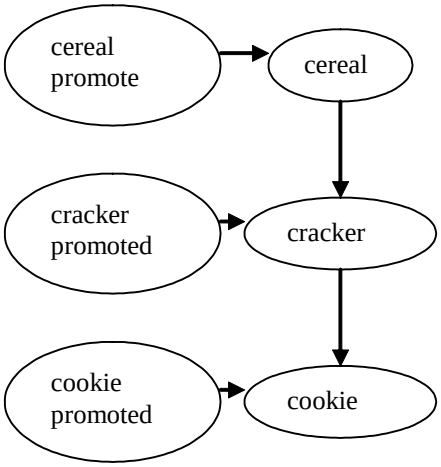

Figure 9: Product cluster three with promotions included, cutoff at 0.9

4.2.3. Results of Model 3

The two types of demographic information used are family size and income. Here is the definition of the two variables. If there are more than two members in a family, the "family size" variables equals to one, otherwise zero. If the family income is over 35 thousand dollars, the "income" variable equals to one, otherwise zero.

Figure 12 illustrates the average network of the model. Family size is related to more categories than income. The possible explanation is that the categories in our data set are most staples. Despite income, most families needs to buy products from these categories. Like model 1, there is one large multi-category cluster. However, the direction of some edges differs, indicating customer segments whose shopping behavior is different from the mass.

These models show Bayesian network learned from data can capture the multi-product relationship. It also shows that with promotion data, Bayesian network can capture a model that is closer to the underlying mechanism.

4.3. Pairwise Assessment

We can see that with more information, Bayesian networks are better able to learn the presence of multi-category relationships. We can identify different typology of complements using Bayesian network combined with localized rule discovery. Table 2 gives examples of category pairs and contrasts each pair's strength of association with the corresponding edge probabilities in the three models.

These results are summarized in Table 3, and show that true use complements always

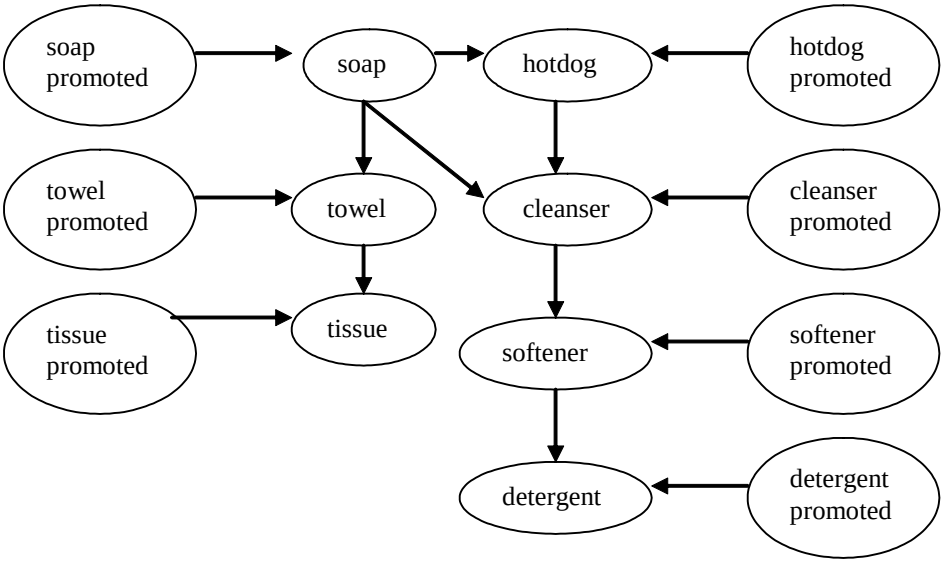

Figure 10: Product cluster one with promotions included, cutoff at 0.6

Pair	Assoc. Rule lift	Prob. in Model 1	Prob. in Model 2	Prob. in Model 3
(Soap, Cleanser)	3.98	0.83	0.63	0.89
(Detergent, Softener)	3.55	0.99	0.99	0.99
(Towel, Tissue)	1.86	0.99	0.99	0.99
(Cracker, Cookie)	1.63	0.99	0.96	0.80
(Cereal, Yogurt)	1.68	0.84	0.62	0.99

Table 2: Comparison across different approaches

have a strong relationship. For example, detergent and softener have both high lift and high presence probability. Spurious complements such as soap and cleanser will disappear when promotion effects are considered. Findings such as towel and tissue, and cracker and cookie can be utilized to cross sell. Another benefit of pairwise assessment is identification of primary and secondary categories in a pair of use complements. For strong complements detergent and softener , three Bayesian networks were learned with their sales data and their promotion data. Figure 13 has a probability of 98% and Figure 14 has a probability of 1%. There is less than 1% probability that the two categories are independent, which is not

Pair	Strong BN Presence	Moderate BN Presence
High lift	(Detergent, Softener)	(Soap, Cleanser)
Moderate lift	(Towel, Tissue)	

Table 3: Comparing Lift to probability of presence in BN

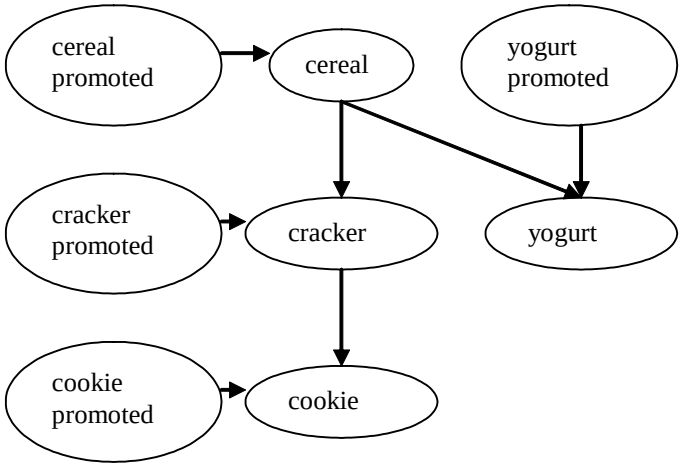

Figure 11: Product cluster two with promotions included, cutoff at 0.6

shown here. We may conclude that softener is the primary item in the Detergent-Softener relationship.

4.4. Summary and Conclusions

There are three interesting findings in this application. First,the promotion variable is found to be more useful in identifying product clusters than just sales data. Second, family size is found to be more important than income in determining product relationships. Finally, complements such as (detergent, softener) and (towel, tissue) can be identified when we compare all the three models. Such findings are useful to retailers in that they enable them to coordinate marketing activities and target specific customers accordingly. We would further extend this research to a multi-period model. In such a model, customers' purchase decisions in current week impact their purchase decisions the following week. In summary, this paper illustrated the use of Bayesian Networks with Market Basket Data. The ability of BNs to identify associations across multiple product categories makes them a powerful tool in being able to better predict buyer behavior in retail contexts. Extensions of the proposed approach would be consider Dynamic BNs that would allow us to model buyer behavior over time. These are being currently investigated.

References

[1] Baesens, B., S. Viaene. 2002. Bayesian neural networks learning for repeat purchase modeling in direct marketing. *European Journal of Operational Research* **138(1)** 191–

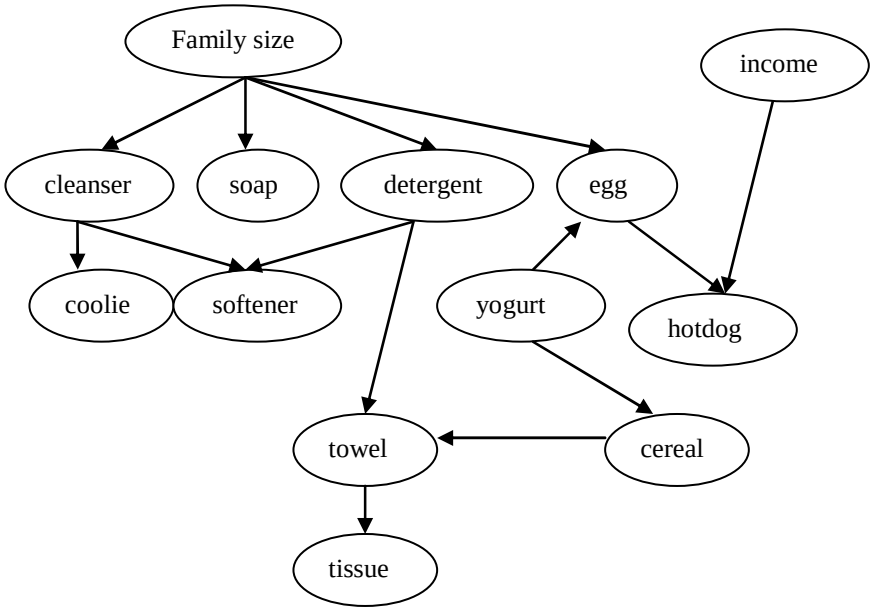

Figure 12: Product cluster with demographics included

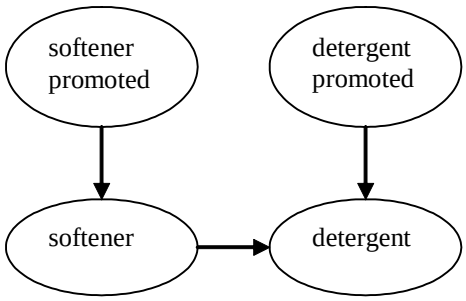

Figure 13: Softener as primary category

211.

[2] Blattberg, R., R. Briesch, F. Edward. 1995. How Promotions Work. *Marketing Science* **14(3)** 122–132.

[3] Chib, S., E. Greenberg. 1995. Understanding the Metropolis-Hastings Algorithm. *The American Statistician* **49(4)** 327–335.

[4] Chib, S., P. Seetharaman, A. Strijnev. 2002. Analysis of Multi-Category Purchase Incidence Decisions Using IRI Market Basket Data. *Econometric Models in Marketing* **16** 65–90.

[5] Chickering, D. M. 1996. Learning Bayesian networks is NP-complete. *LECTURE NOTES IN STATISTICS* **114** 121–130.

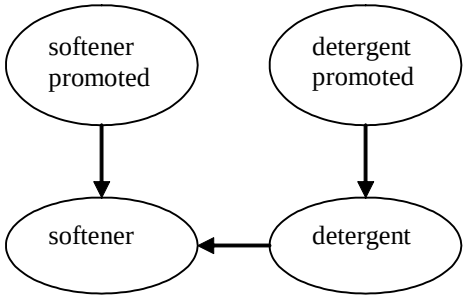

Figure 14: Detergent as primary category

[6] Cooper, G. 1990. Computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence* **42(2/3)** 393–405.

[7] Cooper, G., E. Herskovits. 1992. A Bayesian Method for Induction of Probabilistic Networks from Data. *Machine Learning* **9(4)** 309–347.

[8] Cooper, L. 2000. Strategic marketing Planning for Radically New Products. *Journal of Marketing* **64(1)** 1–16.

[9] Cui, G., M. Wong, 2006. Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming. *Management Science* **52(4)** 597–612.

[10] Dagum, P., M. Luby. 1993. Approximating probabilitic inference in Bayesian belif networks is NP-hard. *Artificial Intelligence* **60** 141–153.

[11] Dawid, P. 1992. Applications of a general propagation algorithm for probabilistic expert systems *Statistics and Computing* **2** 25–36.

[12] Giudici, P., R. Castelo. 2001. Association Models for Web Mining. *Data Mining and Knowledge Discovery* **5** 183–196.

[13] Giudici, P., G. Passerone. 2002. Data Mining of Association Structures to Model Consumer Behavior. *Computational Statistics and Data Analysis* **38** 533–541.

[14] Giudici, P., R. Castelo. 2003. Improving Markov Chain Monte Carlo Model Search for Data Mining. *Machine Learning* **50** 127–158.

[15] Hastings, W.K. 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **57(1)** 97–109.

[16] Heckerman, D. 1996. A tutorial on learning with Bayesian networks. Microsoft Research.

[17] Hruschka, H., M. Lukanowicz, C. Buchata. 1999. Cross category sales promotion effects. *Journal of Retailing and Consumer Services* **6** 99–105.

[18] Jensen, F., S. Lauritzen. 1990. Bayesian updating in recursive graphical models by local computations. *Computational Statisticals Quarterly* **4** 269–282.

[19] Lauritzen, S., D. Spiegelhalter. 1988. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of Royal Statistical Society B*. **50** 157–224.

[20] Madigan, D., J. York. 1995. Bayesian graphical models for discrete data. *International statistical Review* **63(2)** 215–232.

[21] Manchanda, P., A. Asim, G. Sunil. 1999. the "Shopping Basket": a Model of Multicategory Purchase Incidence Decisions . *Marketing Science*. **18(2)** 95–114.

[22] Metropolis, N., A. Rosenblth, M. Rosenbluth, A. Teller, E. Teller. 1953. Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* **21** 1087–1092.

[23] Mulhern, F., R. Leone. 1991. Implicit Price Bundling of Retail Products: a Multiproduct to Maximizing Store Profitability. *Journal of Marketing* **55** 63–76.

[24] Raftery, A., D. Madigan, J. Hoeting. 1997. Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association* **92(437)** 179–191.

[25] Russell, G., S. Ratneshwar, A. Shocker, B. David, A. Bodapati. 1999. Multiple-Category Decision-Making: Review and Synthesis. *Marketing Letters* **10(3)** 319–332.

[26] Russell, G., A. Petersen. 2000. Analysis of Cross Category Dependence in Market Basket Selection. *Journal of Retailing* **76(3)** 367–392.

[27] Robinson, R. W. 1977. Counting unlabelled acyclic digraphs. *Lecture Notes in Mathematics: Combinatorial Mathematics V* Springer-Verlag.

[28] Seetharaman, P.B., S. Chib, A. Ainslie, P. Boatwright, T. Chan, S. Gupta, N. Mehta, V. Rao, A. Strijnev. 2005. Models of Multi-Category Choice Behavior. *Marketing Letters* **16(3/4)** 239–254.

[29] Shachter, R. 1988. Probabilistic inference and influence diagrams. *Operations Research* **36** 589-604.