Structural Plots of Multivariate Binary Data

Ove Frank
Department of Statistics
Stockholm University
Ove.Frank@stat.su.se

ABSTRACT: Data structures comprising many binary variables can be represented graphically in various ways. Depending on the purpose different plots might be useful. Here two ways of showing associations between variables and implications between variables are discussed. The methods are based on conditional independence graphs and lattices of maximal cluster-property pairs. Applications to multivariate samples and network data are briefly discussed.

1. Introduction

A Boolean or binary data structure is typically given as an n by m data matrix (x_{ij}) of entries from $\{0,1\}$. The n rows represent units of some kind, often called individuals, objects, or cases, and the m columns represent binary variables or attributes defined on these units. The presence or absence of the jth attribute at the ith unit is designated by $x_{ij}=1$ and $x_{ij}=0$, respectively. Let $U=\{1,...,n\}$ and $V=\{1,...,m\}$ denote the sets of (integer labels of) units and attributes. The variables and their values are denoted by $x_1,...,x_m$. Thus $x_j=x_{ij}$ is the value taken by x_j for the ith unit. The ordered sequence $x=(x_1,...,x_m)$ of variables represent combinations of attributes which are often called properties. There are 2^m distinct properties. The properties of the n units are given by the rows $a_i=(x_{i1},...,x_{im})$ of the data matrix. The properties of the n units can also be represented as subsets $A_i=\{j\in V : x_{ij}=1\}$ of attributes. Thus the information in the data matrix is given by n binary m-sequences or, equivalently, by n subsets of V. Unless m is very small, it is not straightforward how to gain insight into the structure of properties present in data. When the number of attributes is moderate or large, many properties might have small frequencies, and when n is small compared to 2^m many properties do not occur. The properties that occur might be related in various ways. Useful information about the inherent structure in data might be exhibited in plots of different kinds. Graphical methods as a major tool of efficient data analysis is well known (Tukey (1977),

Chambers et al (1983), Tufte (1983)) and this article focus on plots for describing and testing associations and implications between properties.

The next section introduces further notation and concepts needed in order to explore independence and conditional independence with methods developed for general multivariate loglinear models. The books by Whittaker (1990) and Edwards (1995) give introductions to graphical modeling in statistics. The association graphs introduced in Section 3 are varieties of conditional independence graphs appropriate for random samples of binary data. Section 4 discusses modifications appropriate for non-sample data.

Implications between properties are investigated in Section 5. A single attribute j is present for the units indicated by 1-entries in the jth column $b_j=(x_{1j},...,x_{nj})$ of the data matrix. These units are also given by the subset $B_j=\{i\in U : x_{ij}=1\}$. The subsets $B_1,...,B_m$ are called clusters of units corresponding to the different attributes. The clusters B_j that correspond to the attributes present for the ith unit have one or more units in common. The systematic study of maximal clusters of units having certain attributes in common leads to an interest in the maximal set of attributes in common to the units of a cluster. The maximal cluster and property pairs are partially ordered and can be represented in a lattice diagram (Wille (1982, 1984)). This diagram can be used in order to deduce implications between the attributes. Section 6 embeds the implications in conditional probability statements and shows how bipartite graphs can be used to represent probabilistic implications concerning the presence and absence of different attributes. Section 7 finally gives some brief comments on extensions and applications of the methods.

2. Statistical Theory

Let $U=\{1,...,n\}$ and $V=\{1,...,m\}$ be the sets of units and attributes. The binary variables $x_1,...,x_m$ indicate presence (1) or absence (0) of the attributes, and the sequence $x=(x_1,...,x_m)$ specifies a property. It is natural to talk about properties even if only a subsequence of x is specified. Thus with $m=3$ the property $x_2=0, x_3=1$ is specified by $x=(*,0,1)$ where $*$ is used to indicate an unspecified attribute.

The binary data matrix (x_{ij}) has rows $a_i=(x_{i1},...,x_{im})$ that specify the properties of the units, and columns $b_j=(x_{1j},..,x_{nj})$ that specify the clusters of the attributes. The property as a subset of V is given by A_i for the ith unit, and the cluster as a subset of U is given by B_j for the jth attribute. If we consider the data matrix as an adjacency matrix of a bipartite graph on the set of rows U and columns V, we can also think of A_i as the set of columns after the ith row and B_j as the set of rows before the jth column.

For any binary m-sequence x, the number of units with property x is given by

$$n(x) = \sum_{i=1}^{n} I(a_i = x)$$

where I is an indicator function equal to 1 or 0 according to whether or not the argument of I is true. The relative frequency of property x is denoted by

$$\hat{p}(x) = n(x)/n.$$

The marginal frequencies $\hat{p}_i$, $\hat{p}_{ij}$, ... are obtained by summing $\hat{p}(x)$ over all x with $x_i=1$, with $x_i=x_j=1$, etc. If $\hat{p}(x)$ is considered as an estimate of a probability distribution

$$P(X=x)=p(x)$$

over properties, then the marginal frequencies $\hat{p}_i$, $\hat{p}_{ij}$, ... estimate the expected values $EX_i=p_i$, $EX_iX_j=p_{ij}$,... . More generally we write the marginal probabilities

$$p_i(x_i), p_{ij}(x_i, x_j),...$$

when we need to specify other values than 1 or $x_i,x_j,...$.

Assuming that the properties of the n units are independent observations on the random property X, we get the likelihood function

$$L(p) = \prod_{i=1}^{n} p(a_i) = \prod_{x} p(x)^{n(x)}$$

and it follows that

$$\log L(p) = \sum_{x} n(x) \log p(x) = -n\hat{H} - nD(\hat{p}, p)$$

where

$$\hat{H} = -\sum_x \hat{p}(x) \log \hat{p}(x)$$

is the entropy of the empirical distribution $\hat{p}$ and

$$D(\hat{p}, p) = \sum_x \hat{p}(x) \log[\hat{p}(x) / p(x)]$$

is the information divergence from the empirical distribution $\hat{p}$ to the theoretical probability distribution p. All sums are understood to be over x with p(x)>0. According to information theory (Hamming (1980) is an introductory text and Kullback (1959) a general statistical reference) the divergence $D(\hat{p},p)$ is non-negative, and it is zero if and only if $p=\hat{p}$. If p is restricted according to some model imposing r restrictions, then the maximum likelihood is obtained for minimum information divergence. Let $\tilde{p}$ denote this optimal distribution. It follows that the loglikelihood ratio test statistic is

$$2\log\left[L(\hat{p}) / L(\tilde{p})\right] = 2nD(\hat{p}, \tilde{p}),$$

and it is asymptotically χ^2-distributed with r degrees of freedom when n tends to infinity.

In particular, the model assumption that X_i and X_j are independent imposes one restriction $p_{ij}=p_ip_j$, and it follows that

$$p(x) = p_i(x_i)\, p_j(x_j)\, p_{k|ij}(x_k | x_i, x_j)$$

where k is the subsequence of all attributes except i and j, and x_k is the subsequence of x with x_i and x_j removed. Now $D(\hat{p},p)$ is minimized for $p=\tilde{p}$ given by

$$\tilde{p}(x) = \hat{p}_i(x_i)\, \hat{p}_j(x_j)\, \hat{p}_{k|ij}(x_k | x_i, x_j),$$

and the likelihood ratio can be simplified to

$$\hat{p}(x) / \tilde{p}(x) = \hat{p}_{ij}(x_i, x_j) / \hat{p}_i(x_i)\hat{p}_j(x_j).$$

This leads to

$$D(\hat{p}, \tilde{p}) = \hat{H}_i + \hat{H}_j - \hat{H}_{ij}$$

where $\hat{H}_i$ and $\hat{H}_{ij}$ are entropies of empirical marginal distributions. Thus we have a simple form of the test statistic $2nD(\hat{p}, \tilde{p}) = D_{ij}$. Values of D_{ij} larger than 4 are critical at approximately 5% significance level.

An alternative model specifying that X_i and X_j are conditionally independent when X_k, the rest of X, is fixed implies that

$$p(x) = p_{i|k}(x_i|x_k)\, p_{j|k}(x_j|x_k)\, p_k(x_k)$$

where k is the subsequence of 1,...,m with i and j removed, and x_k is the subsequence of x with x_i and x_j removed. It follows that

$$D(\hat{p}, \tilde{p}) = \hat{H}_{ik} + \hat{H}_{jk} - \hat{H}_{ijk} - \hat{H}_k .$$

There is one restriction for each outcome of x_k, that is $r=2^{m-2}$. Therefore the test statistic

$$D_{ij} = 2nD(\hat{p}, \tilde{p})$$

is asymptotically χ^2-distributed with 2^{m-2} degrees of freedom, and the 5%-level critical value is about $2^{m-2}+2^{(m+1)/2}$.

3. Association Graphs

Guided by the large sample theory in the previous section we first check the variables for pairwise independence by drawing a graph with the attributes as vertices and an edge between the ith and jth vertex if the test statistic

$$D_{ij} = 2n(\hat{H}_i + \hat{H}_j - \hat{H}_{ij})$$

is above some chosen critical value from the χ^2-distribution with 1 degree of freedom. By adjusting the critical value we try to get a graph consisting of complete or almost complete connected components (with no or almost no edge between components). The rationale behind this is that strong pairwise dependence is close to functional dependence, and

pairwise functional dependence between non-constant binary variables is a transitive relation. Vertices in distinct connected components represent independent variables. After having the set of variables split into independent subsets, the analysis can proceed separately with each subset.

The next step is to test a set of dependent variables for pairwise conditional independence. By conditioning successively on an increasing number of variables, we hope to achieve more reliable results than what would be possible by conditioning on all remaining variables when n is relatively small. In order to test conditional independence of X_i and X_j with a conditioning sequence X_k we use the test statistic

$$D_{ijk} = 2n(\hat{H}_{ik} + \hat{H}_{jk} - \hat{H}_{ijk} - \hat{H}_{k})$$

and critical values from the χ^2-distribution with $2^{|k|}$ degrees of freedom, where $|k|$ is the length of k, that is the number of attributes in the conditioning sequence. For each pair (i,j) there are $\binom{m-2}{|k|}$ values D_{ijk} with k of a fixed length $|k|$. A graph can be drawn for each value of $|k|=1,2,...,m-2$ by inserting an edge between i and j if the maximum value D_{ijk} is sufficiently large. Maximum is here taken over $m-2$ values k for the first graph, over $\binom{m-2}{2}$ values k for the second graph, etc. In this way we obtain a sequence of $m-2$ graphs, and only the last one, corresponding to the conditioning on all remaining variables, is a conditional independence graph in the sense understood in loglinear modeling (Whittaker (1990), Edwards (1995)). The choice of critical values for the maximum values of D_{ijk} is not easily based on theoretical grounds, but can perhaps be guided by the requirement that the sequence of graphs should be non-decreasing.

4. Association Under Randomization

If the properties of the n units cannot be considered as a sample of independent observations from a common probability distribution, we might still want to get a description of the multivariate structure of the data. Such a description could be useful as a summary or as a tool for extracting hidden features and patterns.

We consider the empirical distribution as the probability distribution of interest and use the entropies as descriptive measures of spread. There are also natural measures of association based on information theoretic concepts (See for instance Goodman and Kruskal (1979)). We recall the fact that the entropy $H(X)$ of a random m-sequence X satisfies

$$0 \le H(X) \le m \log 2$$

with the extreme values attained for a one-point and a uniform distribution only. Furthermore, for a partition of $X=(X_k,X_l)$ into two parts, the conditional entropy $H(X_k|X_l)$ has an expected value satisfying

$$0 \le EH(X_k \mid X_l) = H(X_k, X_l) - H(X_l) \le H(X_k)$$

with the minimum value attained if and only if X_k is functionally dependent on X_l and the maximum value attained if and only if X_k is stochastically independent of X_l. The difference

$$H(X_k) - EH(X_k \mid X_l) = H(X_k) + H(X_l) - H(X_k, X_l)$$

is symmetrical in the two parts X_k and X_l, and it is a measure of their degree of dependence or association. This measure is equivalent to the information divergence from the simultaneous distribution of X_k and X_l to the product of their marginal distributions. Thus the test statistics D_{ij} used for the association graphs can be interpreted as empirical measures of association. Critical values of such measures can be obtained by introducing some kind of randomization in the data matrix, and using this we can judge whether the actual data matrix is extreme with respect to its value of association. In particular, it is of interest to know what range of association values there is under randomization.

If we randomize by keeping fixed all marginal distributions of single attributes, then $p_1,...,p_m$ and $H_1,...,H_m$ are all fixed, and H_{ij} and $D_{ij}=2n(H_i+H_i-H_{ij})$ vary with p_{ij} only. In order to specify the range of possible values of H_{ij} it is convenient and no restriction to assume that

$$p_1 \le ... \le p_m \le ½ \quad and \quad H_1 \le ... \le H_m.$$

Attributes might have to be replaced by their complements in order to achieve that they have their presence at most as common as their absence. The labeling of attributes according to non-decreasing probabilities implies non-decreasing entropies, since H_i is an increasing function of p_i for $0<p_i \leq ½$. If we define the function

$$\varphi(p) = -p \log p \qquad for \;\; 0 < p \leq 1$$

$$\varphi(0) = 0$$

entropies can be expressed as

$$H(X) = \sum_x \varphi(p(x))$$

$$H_i = \varphi(p_i) + \varphi(1 - p_i)$$

$$H_{ij} = \varphi(p_{ij}) + \varphi(p_i - p_{ij}) + \varphi(p_j - p_{ij}) + \varphi(1 - p_i - p_j + p_{ij}).$$

By differentiating H_{ij} with respect to p_{ij} we find that H_{ij} has a maximum value H_i+H_j at $p_{ij}=p_ip_j$, and H_{ij} is a unimodal function of p_{ij} taking its minimum value at the lower boundary of its domain

$$0 \leq p_{ij} \leq \min(p_i, p_j).$$

The minimum value is

$$min H_{ij} = \varphi(p_i) + \varphi(p_i) + \varphi(1 - p_i - p_j).$$

It follows that D_{ij} is a unimodal function of p_{ij} with minimum value 0 at $p_{ij}=p_ip_j$ and maximum value

$$\max D_{ij} = 2n[\varphi(1 - p_i) + \varphi(1 - p_j) - \varphi(1 - p_i - p_j)]$$

at the lower boundary of the domain $0 \leq p_{ij} \leq \min(p_i,p_j)$. Hence the critical region of p_{ij}-values is one-sided or two-sided depending on whether or not the critical association level is above the value of D_{ij} at the upper boundary, that is

$$2n[\varphi(1 - p_i) + \varphi(p_j) - \varphi(p_j - p_i)] \;\; for \; i<j.$$

The association D_{ij} is generally larger for negative than for positive deviations p_{ij}-p_ip_j of the same absolute value. An alternative association measure given by the Pearson correlation coefficient in this case is

$$\rho_{ij} = (p_{ij} - p_i p_j) / [p_i(1-p_i)p_j(1-p_j)]^{½}$$

and obviously this correlation or its absolute value does not capture the asymmetry of D_{ij}. In fact for $i<j$

$$-[p_i p_j / (1-p_i)(1-p_j)]^{½} \le \rho_{ij} \le [p_i(1-p_j) / (1-p_i)p_j]^{½}$$

and the absolute value of ρ_{ij} at $p_{ij}=0$ is not larger than the value at $p_{ij}=p_i$. As a consequence we need to multiply negative values of ρ_{ij} by a factor larger than 1 if we want to get an association measure that is similar to D_{ij}.

5. Implications

A property specifying the presence or absence of certain attributes is said to imply another property if all units having the first property also have the second property. This first property is called the condition of the implication.

In order to investigate implications systematically we can confine ourselves to implications concerning single attributes added to the condition, since implications concerning several added attributes can be obtained as consequences of those for single attributes.

The specification of attributes that are present or absent can conveniently be referred to as subsets of included and excluded attributes, respectively. Thus a property specifying that all the attributes in a subset $T_1 \subseteq V$ shall be present and all the attributes in a disjoint subset $T_0 \subseteq V$ shall be absent means that we are referring to a cluster $S \subseteq U$ of units such that T_1 shall be included and T_0 excluded from each attribute set A_i for $i \in S$, that is $T_1 \subseteq A_i \subseteq \overline{T}_0$ for $i \in S$.

Implications from one attribute to another are the simplest. For instance, the absence of the ith attribute ($x_i=0$) implies the presence of the jth attribute ($x_j=1$) if the set of units not

having attribute i is contained in the set of units having attribute j. In terms of clusters of units representing the attributes, this implication can be specified as $\overline{B}_i \subseteq B_j$.
Equivalently the implication can be given as $\overline{b}_i \leq b_j$ in terms of the binary column sequences of the data matrix. Here $\overline{b}_i = (1-x_{1i},\ldots,1-x_{ni})$.

Consider a cluster $S \subseteq U$ defined by $T_1 \subseteq A_i \subseteq \overline{T}_0$ for $i \in S$. Another way of expressing this is to say that S is the intersection of the clusters B_j for $j \in T_1$ and $\overline{B}_k$ for $k \in T_0$. If we define the cluster of units sharing the presence of all attributes in T by

$$B(T) = \bigcap_{j \in T} B_j$$

and the cluster of units sharing the absence of all attributes in T by

$$\overline{B}(T) = \bigcap_{j \in T} \overline{B}_j$$

for $T \subseteq V$, we obtain that the cluster of units with property T_1 included and T_0 excluded is given by

$$S = B(T_1) \cap \overline{B}(T_0).$$

It should be noticed that $\overline{B}(T)$ is not the complement of B(T) (which could be denoted $\overline{B(T)}$) but the extension of complements $\overline{B}_j$ to the intersection of such complements.

Analogously we also define the property shared by all units in cluster $S \subseteq U$ as

$$A(S) = \bigcap_{i \in S} A_i .$$

Since $j \in A_i$ and $i \in B_j$ are equivalent, it follows that $S \subseteq B(T)$ and $T \subseteq A(S)$ are equivalent for all $S \subseteq U$ and $T \subseteq V$. Furthermore, $S_1 \subseteq S_2$ implies that $A(S_1) \supseteq A(S_2)$, and $T_1 \subseteq T_2$ implies that $B(T_1) \supseteq B(T_2)$.

If $S \subseteq U$ and $T \subseteq V$ are related according to

$$S = B(T) \text{ and } T = A(S),$$

then the cluster and property pair (S,T) is an important concept that is useful for analyzing the data matrix. The matrix entries x_{ij} are 1 if $i \in S$ and $j \in T$, and neither S

nor T can be extended with further units or attributes without violating this property. We refer to such pairs (S,T) as cliques with clique cluster S and clique property T. If we consider the data matrix as an adjacency matrix of a bipartite graph between U and V, then the cliques are maximal complete bipartite subgraphs. The collection of all cliques is partially ordered by clique cluster inclusion (or, equivalently, clique property inclusion which reverses the order). The partially ordered set of cliques is a lattice, that is there is a smallest clique (S_0,V) and a largest clique (U,T_0). Here S_0 is the set of units with all attributes present, and T_0 is the set of attributes present at all units. These sets may be empty. If (S_1,T_1) and (S_2,T_2) are distinct cliques, we write $(S_1,T_1)<(S_2,T_2)$ and say that (S_1,T_1) is below (S_2,T_2) if $S_1 \subseteq S_2$. For such cliques, $T_1 \supseteq T_2$.

The lattice of cliques is called the concept lattice by Wille (1982, 1984). Algorithms for constructing the lattice from a data matrix are given by Ganter et al. (1986), Luksch et al. (1986) and Duquenne (1987). See also Duquenne (1991). Freeman and White (1993) give a good presentation of the lattice and its usefulness in network data analysis.

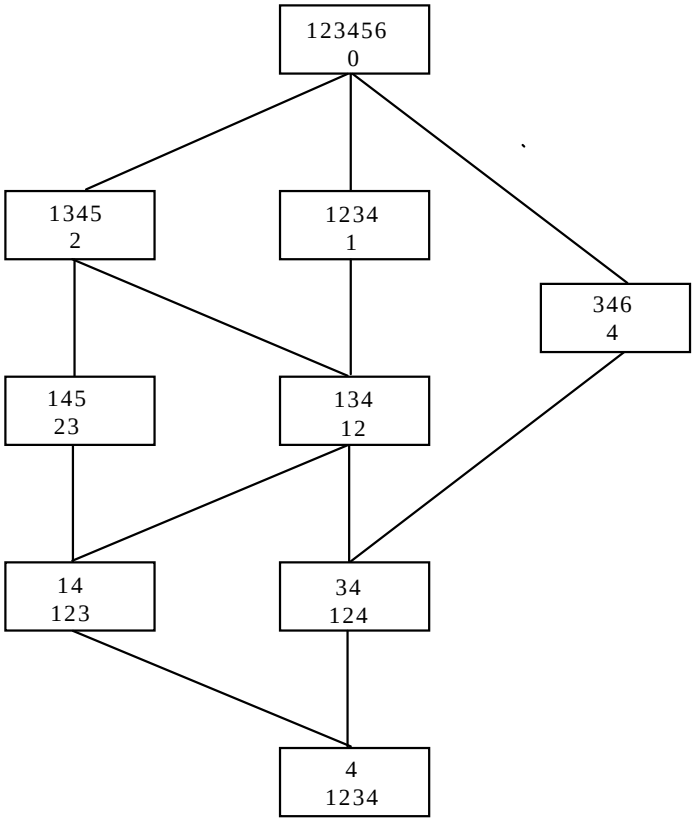
The importance of the lattice of cliques for finding implications should be clear from the following observation. Let (S,T) be a clique and $j \in T$ any attribute of the clique property. If $T-\{j\}$ is not a property of any other clique, then the presence of attribute j is implied by the condition $T-\{j\}$. By checking all cliques for implications of this kind, and keeping only minimal conditions, we can find all implications with conditions referring to the presence of attributes. Implications from conditions referring to both presence and absence of attributes can be found in the same way if the data matrix is enlarged with complementary variables. We illustrate these ideas by a small example.

Consider the data matrix with $m=4$ variables and $n=6$ units given by

1	1	1	0
1	0	0	0
1	1	0	1
1	1	1	1
0	1	1	0
0	0	0	1

There are 9 cliques and their lattice is shown in Figure 1.

Figure 1. Cliques of clusters (above) and properties (below) for a data matrix with 6 units and 4 attributes.

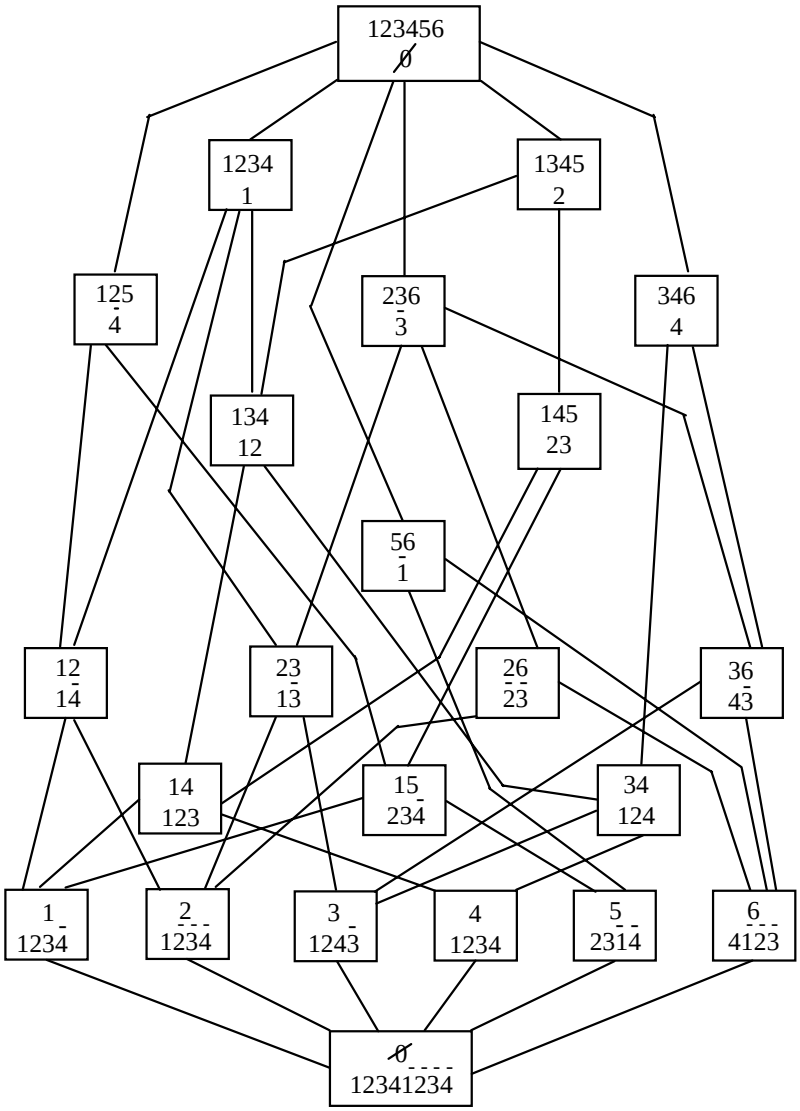


The data matrix extended with 4 complementary variables is given by

1	1	1	0	0	0	0	1
1	0	0	0	0	1	1	1
1	1	0	1	0	0	1	0
1	1	1	1	0	0	0	0
0	1	1	0	1	0	0	1
0	0	0	1	1	1	1	0

Now there are 23 cliques, and their lattice is shown in Figure 2.

Figure 2. Cliques for the data matrix of Figure 1 with the complementary attributes added.



A systematic search for implications can be made like the following. If a subset of a property clique with one attribute removed is not a property clique, then this subset is a condition that implies the presence of the removed attribute. For instance, in Figure 1 the property clique 23 has a subset 3 that is not a property clique, so 3 implies 2. Property clique 123 has a subset 13 that is not a property clique, so 13 implies 2. This implication is redundant since it follows from the previous one. Property clique 124 has subsets 14 and 24 that are not property cliques, so 14 implies 2 and 24 implies 1. Finally, property clique 1234 has subsets 134 and 234 that are not property cliques, so 134 implies 2 and 234

implies 1. Both are redundant. The implications with minimal conditions are given symbolically and with cluster interpretations as follows

$3 \Rightarrow 2$ $\quad B_3 \subseteq B_2$

$14 \Rightarrow 2$ $\quad B_1 \cap B_4 \subseteq B_2$

$24 \Rightarrow 1$ $\quad B_2 \cap B_4 \subseteq B_1$.

From Figure 2 we obtain in a similar manner the following implications: $3 \Rightarrow 2$, $2\bar{4} \Rightarrow 3$, $14 \Rightarrow 2$, $24 \Rightarrow 1$, $1\bar{2}\bar{3} \Rightarrow \bar{4}$, $1\bar{3}\bar{4} \Rightarrow \bar{2}$, $\bar{2}\bar{3}\bar{4} \Rightarrow 1$, $12\bar{3} \Rightarrow 4$, $\bar{1}23 \Rightarrow \bar{4}$, $\bar{1}\bar{3}4 \Rightarrow \bar{2}$, $\bar{2}\bar{3}4 \Rightarrow \bar{1}$, $\bar{1}\bar{2}\bar{3} \Rightarrow 4$ where $\bar{j}$ denotes absence of attribute j, that is presence of the complement of j.

6. Probabilistic Implications

Property $T_1 \subseteq V$ implies property $T_2 \subseteq V$ if $B(T_1) \subseteq B(T_2)$ so that all the units having T_1 also have T_2. By relaxing the requirement that all units in $B(T_1)$ should be included in $B(T_2)$ and focusing on what proportion of units in $B(T_1)$ also are in $B(T_2)$, the implications are generalized to conditional probability statements.

The probability of presence of attribute j conditional on the presence of all attributes in T_1 and the absence of all attributes in T_0 is the conditional expectation of variable X_j given that $X_k=1$ for $k \in T_1$ and $X_k=0$ for $k \in T_0$. Generally there are $m(3^{m-1}-1)$ conditional expectations $E(X_j|X=x)$ where $x=(x_1,...,x_m)$ is an m-sequence of elements from $\{0,1,*\}$ with $x_j=*$. Here, as in Section 2, $*$ indicates an unspecified attribute. A conditional expectation $E(X_j|X=x)$ where x has k specified attributes and m-k unspecified attributes (including attribute j) can also be given as

$$E(X_j|X_i=x_i \ \text{for} \ i=i_1,...,i_k)$$

where $i_1,...,i_k$ are all distinct and distinct from j, and each x_i is chosen from $\{0,1\}$. There are

$$m\binom{m-1}{k}2^k$$

such conditional expectations for $k=1,...,m-1$. Conditional expectations that are close to 0 indicate that X_j is likely to be 0, and conditional expectations that are close to 1 indicate that X_j is likely to be 1.

Such probabilistic implications can be shown in a bipartite graph with edges from condition vertices to attribute vertices if the conditional expectations are sufficiently far from ½. There are 3^m-2^m-1 condition vertices and m attribute vertices. Each condition vertex with k specified attributes has at most $m-k$ incident edges, and each attribute vertex has at most $3^{m-1}-1$ incident edges. Vertices with no incident edges can be omitted. If the graph is still too comprehensive, it can be split into the subgraphs induced by the attribute vertices, but then some condition vertices might be repeated in different subgraphs. Another split of a comprehensive bipartite graph is to separate condition vertices into different subgraphs according to their number of specified attributes. With specification of one, two or three attributes only, the number of condition vertices is substantially reduced for large m.

7. Extensions and Applications

The association graphs discussed in Section 3 need to be tested in practice under varying conditions before recommendations can be made concerning the versions which are not conditional independence graphs with all m variables involved. If we consider whether or not X_i and X_j are conditionally independent given the rest of the sequence X, and this rest is split into two parts X_k and X_l, the question of concern is how misled one can be by considering conditional independence of X_i and X_j given X_k. In order for this to be equivalent to conditional independence given both X_k and X_l, it is required that either X_i or X_j should be conditionally independent of X_l given all the rest, that is given (X_j,X_k) or (X_i,X_k). Obviously we are safe if X_l is independent of the rest, but further research and experience is needed.

All variables treated here are binary, that is, categorical variables with two categories. Categorical variables with more than two categories can be replaced by two or more binary variables by coding the outcomes according to a one-to-one transformation. For instance, a

variable with 4 outcomes can be replaced by two binary variables by coding the outcomes 00,01,10,11. Alternatively, it can be replaced by four binary variables by coding the outcomes 1000, 0100, 0010, 0001. The last coding yields 12 implications between the new variables. There is a vast literature about coding theory, but coding as a tool of data analysis is hardly met in the statistical literature. The French school of data analysis is much concerned about data types as for instance Benzécri (1980) and Jambu and Lebeaux (1983). The psychometric scaling literature pay much attention to data representations; see for instance Kruskal and Wish (1978), Young et al. (1980), and DeLeeuw and Tijssen (1984). The impact of these approaches on descriptive statistics and data analysis ought to be much larger than it is.

The use of entropy and other information theoretic concepts matches both the coding theory and the large sample likelihood theory. Information loss by coding (Frank and Weidenman (1987) and Frank and Öhrvik (1994)) and data transformations to protect individual privacy (Frank (1983, 1988)) are two examples of information theoretic approaches to statistical data dissemination that might be of interest in connection with the fundamental problem of choosing data to be presented.

The data matrix with units and attributes is one of the most common data structures. If units are independent cases, standard statistical multivariate methods are available. Applications are then abundant, and some of the graphical methods presented here might be a natural part of an initial data analysis.

If the units are not independent, then the graphical plots might still be of interest as data displays. The plots can have interpretation of significance in terms of some natural randomization procedure. Units might be individuals, households or countries, for example, and a natural randomization might be a permutation of all the units. If the units are economic transactions between some accounts or social contacts between some actors, then a natural randomization might be a permutation of all the accounts or all the actors. Such a permutation induces a restricted permutation of the units. Data of this type can be obtained from networks with several attributes attached to the edges, that is to the transactions or contacts in the examples mentioned. There are other possibilities to analyze network data which define the data matrix differently. If we take the rows of the data

matrix as the vertices of the network and the columns as the edges, each attribute provides one data matrix, the edge incidence matrix of the network for that attribute. Another possibility is to take the data matrix as the combined matrix of these incidence matrices put side by side with the same rows. Modeling ideas might suggest what is the most appropriate form of the data matrix in any particular case.

References

Benzécri, J.P. (1980). *Pratique de l'analyse des données*. Paris: Dunod.

Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P.A. (1983). *Graphical Methods for Data Analysis*. Belmont, CA: Wadsworth.

De Leeuw, J. and Tijssen, R. (1984). *Multivariate Analysis with Optimal Scaling*. Leiden: Department of Data Theory.

Duquenne, V. (1987). "Contextual implications between attributes and some representation properties for finite lattices." In B. Ganter, R. Wille, and K.E. Wolf (Eds.*): Beitraege zur Begriffsanalyse*. Mannheim: Wissenschaftsverlag. 213-239.

Duquenne, V. (1991). *On the core of finite lattices*. Discrete Mathematics 88, 133-147.
Edwards, D. (1985). *Introduction to Graphical Modeling*. New York, NY: Springer-Verlag.

Frank, O. (1983). "Statistical disclosure control." *Statistical Review* 5, 173-178.

Frank, O. (1988). "Designing classifiers for partial information release." In H.H. Bock (Ed.): *Classification and Related Methods of Data Analysis*. Amsterdam: Elsevier. 685-690.

Frank, O. and Weidenman, P. (1987). "Controlling individual information in statistics by coding." *Journal of Statistical Planning and Inference* 17, 321-336.

Frank, O. and Öhrvik, J. (1994). "Entropy of sums of random digits*." Computational Statistics & Data Analysis* 17, 177-184.

Freeman, L.C. and White, D. R. (1993). "Using Galois lattices to represent network data." In P.V. Marsden (Ed.): *Sociological Methodology 1993*. Washington, DC: American Sociological Association. 127-146.

Ganter, B., Rindfrey, K., and Skorsky, M. (1986). "Software for formal concept analysis." In W. Gaul and M. Schader (Eds.): *Classification as a Tool of Research*. Amsterdam: Elsevier Science. 161-167.

Goodman, L.A. and Kruskal, W.H. (1979). *Measures of Association for Cross-Classifications*. New York, NY: Springer-Verlag.

Hamming, R.W. (1980). *Coding and Information Theory*. Englewood Cliffs, NJ: Prentice-Hall.

Jambu, M. and Lebeaux, M.O. (1983). *Cluster Analysis and Data Analysis*. Amsterdam: North-Holland.

Kruskal, J.B. and Wish, M. (1978). *Multidimensional Scaling*. Beverly Hills, CA: Sage.

Kullback, S. (1959). *Information Theory and Statistics*. New York, NY: Wiley.

Luksch, P., Skorsky, M., and Wille, R. (1986). "Drawing concept lattices with a computer." In W. Gaul and M. Schader (Eds.): *Classification as a Tool of Research*. Amsterdam: Elsevier Science. 269-274.

Tufte, E.R. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.

Wille, R. (1982). "Restructuring lattice theory: an approach based on hierarchies of concepts." In I. Rival (Ed.): *Ordered Sets*. Dordrecht: Reidel. 445-470.

Wille, R. (1984) "Sur la fusion des contexts individuels." *Mathématiques et Sciences Humaines* 85, 57-71.

Young, F.W., De Leeuw, J., and Takane, Y. (1980). "Quantifying qualitative data." In E.D. Lantermann and H. Feger (Eds.): *Similarity and Choice. Papers in Honour of Clyde Coombs*. Berne: Hans Huber.