Data Mining 2020
Counting u-terms in graphical log-linear models

Being able to count u-terms is important for

1. determining the correct degrees of freedom for statistical tests, and
2. computing the correct complexity penalty for a model's AIC or BIC score.

Let $X = (X_1, X_2, \ldots, X_k)$ denote a vector of k discrete random variables. Furthermore, let d_i denote the size of the domain of X_i. For example, if X_i corresponds to the variable "Eye Color", with possible values {blue, brown, green, other}, then $d_i = 4$.

The full log-linear expansion of the probability distribution P_K of $(X_1, X_2, \ldots, X_k)$ is given by

$$\log P_K(x) = \sum_{a \subseteq K} u_a(x_a)$$

where the sum is taken over all possible subsets a of $K = \{1, 2, \ldots, k\}$.
The total number of possible value assignments to $(X_1, X_2, \ldots, X_k)$ (the number of cells in the contingency table) is:

$$\text{ncell} = \prod_{i=1}^{k} d_i$$

The number of u-terms in the full log-linear expansion (the saturated model) is equal to ncell. It can alternatively be computed by the formula

$$\sum_{a \subseteq K} \left(\prod_{i \in a} (d_i - 1) \right),$$

where

$$\prod_{i \in \emptyset} (d_i - 1)$$

is taken to be equal to 1, corresponding to the constant u-term $u_\emptyset$. So apparently

$$\prod_{i=1}^{k} d_i = \sum_{a \subseteq K} \left(\prod_{i \in a} (d_i - 1) \right).$$

In general, a collection of u-terms $u_a(x_a)$ contains

$$\prod_{i \in a}(d_i - 1)$$

u-terms.

Consider as an example the random vector $X = (X_1, X_2, X_3, X_4)$ with $d_1 = 2$, $d_2 = 3$, $d_3 = 5$, and $d_4 = 2$. We compute that

$$\text{ncell} = \prod_{i=1}^{k} d_i = 2 \times 3 \times 5 \times 2 = 60.$$

How many u-terms are there in the graphical model with the cordless 4-cycle $(1{-}2{-}3{-}4{-}1)$ as its independence graph? Recall the following definition:

> "given its independence graph $G = (K, E)$, the log-linear model for the random vector X is a *graphical model* for X if the distribution of X is *arbitrary* apart from constraints of the form that for all pairs of coordinates not in the edge set E, the u-terms containing the selected coordinates are equal to zero".

The following edges are absent in the chordless 4-cycle: $1 - 3$, and $2 - 4$. Because of the absence of $1 - 3$ we have to set the following collections u-terms to zero: u_{13}, u_{123}, u_{134} and u_{1234}. Likewise, because of the absence of $2 - 4$ we have to set the following collections u-terms to zero: u_{24}, u_{124}, u_{234} and u_{1234}. The total number of u-terms that has to be removed is counted in the table below.

u_a	$\prod_{i \in a}(d_i - 1)$
u_{13}	$1 \times 4 = 4$
u_{123}	$1 \times 2 \times 4 = 8$
u_{134}	$1 \times 4 \times 1 = 4$
u_{1234}	$1 \times 2 \times 4 \times 1 = 8$
u_{24}	$2 \times 1 = 2$
u_{124}	$1 \times 2 \times 1 = 2$
u_{234}	$2 \times 4 \times 1 = 8$
Total	36

So to test the chordless 4-cycle against the saturated model, we have to use a chi-square distribtion with 36 degrees of freedom. Also, to compute the AIC or BIC score, we should use $\dim(M) = 60 - 36 = 24$.

Alternatively, we could directly have counted the u-terms present in the model. These are:

u_a	$\prod_{i \in a}(d_i - 1)$
$u_\emptyset$	1
u_1	1
u_2	2
u_3	4
u_4	1
u_{12}	$1 \times 2 = 2$
u_{23}	$2 \times 4 = 8$
u_{34}	$4 \times 1 = 4$
u_{14}	$1 \times 1 = 1$
Total	24