FRIEDRICH-SCHILLER-UNIVERSITY JENA

Faculty of Mathematics and Computer Science

Higher-Order Probabilistic Models: Theory and Algorithms

Masterthesis

A thesis submitted by

Christoph Saffer, born on January 12, 1993 in Bamberg

in fulfillment of the requirements

for the degree Master of Science (M. Sc.)

Supervisor: Prof. Dr. Joachim Giesen

Jena, October 30, 2019

Abstract

Contents

Chapter 1

Introduction

In statistical theory and more applied disciplines as machine learning or artificial intelligence, it is fundamental to reduce the space of parameters. In this work we are going to consider multivariate categorical models, in other words, models that include variables with a discrete set of outcomes. More specifically we are going to look at multivariate binary models or so called Ising Models, whose variables only have two possible outcomes. The parameter space of the full model is growing exponentially with the number of variables. For instance, a binary model (a model with variables that have two possible outcomes) with n variables has $2^n - 1$ parameters, because each combination for all variables is encoded within a parameter. To fit a model to real data, the general principle says that roughly ten times more observations than parameters are necessary in order to get a descriptive, meaningful model, but this requirement can often not be satisfied.

To meet this problem, we either have the option to use regularization methods as sparsity or low rank conditions to reduce the space of parameters while we are fitting the model, for example by setting insignificant parameters to zero. On the other hand we can use a different approach to model the data with fewer parameters from the beginning.

We will look at a specific class of models that we are going to work with, where the focus is set on the interactions between the single variables. These type of modeling data combined with parameter space reduction methods is the objective we are willing to examine.

1.1 Idea of variable interactions

The idea of variable interactions is the based on the assumption that the outcome of one variable depends on the outcomes of the other remaining variables. That is encoded in a statistical model by an additional term which models the interaction. When we consider pairwise interactions, an interaction term between variables x_i and x_j would have the shape

$$q_{ij}\ x_i\ x_j$$

with the interaction parameter q_{ij}. The interaction parameter indicates how much the interaction between x_i and x_j influences the model. An interaction term within a model of interaction order $K = 3$ with variables x_i, x_j, x_k would have the shape

$$q_{ijk}\ x_i\ x_j\ x_k$$

with interaction parameter q_{ijk}. The entire model with the variables x_i, x_j, x_k and interaction order 3 would include the terms

$$q_i x_i + q_j x_j + q_k x_k + q_{ij} x_i x_j + q_{ik} x_i x_k + q_{jk} x_j x_k + q_{ijk} x_i x_j x_k$$

with the parameter space $\Theta = (q_i, q_j, q_k, q_{ij}, q_{ik}, q_{jk}, q_{ijk})$. Here we already get seven parameters for interaction order 3.

Remark. The complexity of a model is equal to the size of the parameter space Θ (number of parameters) of the respective model.

Lemma 1.1. *Let K be the interaction order and n the number of variables, then the complexity of higher-order interaction models can be calculated with*

$$|\Theta| = \sum_{i=1}^{K} \binom{n}{i}, \quad \textit{where } \binom{n}{i} = 0 \textit{ for } i > n$$

Proof. TODO $\square$

Conclusion 1.2. *We consider a binary model and the interaction order K is equal to the number of variables n, then the complexity of the higher-order interaction model is equal to the complexity of the entire categorical model and grows exponentially in n:*

$$|\Theta| = \sum_{i=1}^{n} \binom{n}{i} \ = \ 2^n - 1$$

Var. n \ Order K	2	3	4	5	6	$2^n - 1$
2	3	3	3	3	3	3
3	6	7	7	7	7	7
4	10	14	15	15	15	15
5	15	25	30	31	31	31
6	21	41	56	62	63	63

Table 1.1: Complexity for a interaction model of order K and number of variables n. In comparison on the right, the complexity for a full binary model.

That means, to model higher-order interactions between variables with an interaction order equal to the number of variables, the parameter space grows as fast as the complexity of a full binary model as it is illustrated in table 1.1. Therefore the order of interactions K should be held relatively small in comparison to the number of variables n to get a significant benefit. Therefore, we will mostly consider models with a rather low order of interaction in the following chapters. Later, we are going generalize these approach. Simultaneously, while fitting the model we are going to use methods as regularization as a constraint in the objective function for model selection to reduce the complexity.

1.2 Pairwise interaction binary models

1.2.1 Definitions and conventions

In the following sections, we will define and introduce a multivariate pairwise interaction model on binary variables and try to get a general understanding why it make sense to work with this class of models.

Definition. Let $X = (X_1, ..., X_n)$ be a vector of random variables with the set of outcomes $\Omega = \{0, 1\}^n$. The distribution $p : \Omega \to [0, 1]$ of a *multivariate binary pairwise model* has the form

$$p(x) \propto \exp(\sum_{i=1}^{n} \sum_{j=1}^{n} q_{ij} x_i x_j)$$

with parameters $q_{ij} \in \mathbb{R}$, if the normalization condition

$$\sum_{x \in \Omega} p(x) = 1$$

is satisfied.

Remark. As we mentioned above, the parameter q_{ij} models the interaction between the variables x_i and x_j. Therefore, parameters with permuted index are equal. It is

$$q_{ij} = q_{ji}.$$

That means we presume a symmetric parameter space Θ.

Proposition 1.3. *For the interaction terms $q_i x_i$, $q_{ij} x_{ij}$ applies*

$$q_{ij} x_{ij} = \begin{cases} q_{ij} & , x_i = 1 \wedge x_j = 1 \\ 0 & , x_i = 0 \vee x_j = 0 \end{cases}$$

Before we proceed with model selection, in other words with fitting the parameters to data drawn from the respective model, we will have a closer look at the normalization coefficient and the representation of the parameters.

The model $p(x)$, as defined above, contains a normalization coefficient z to satisfy the normalization condition. The value of the normalization coefficient can be identified by rearranging the equation of the condition:

$$\begin{aligned} &\sum_{x\in\Omega} p(x) = 1 \\ \iff &\sum_{x\in\Omega} \frac{1}{z} \exp(\sum_{i=1}^{n} q_i x_i + \sum_{i=1}^{n}\sum_{j>i}^{n} q_{ij} x_{ij}) = 1 \\ \iff &z = \sum_{x\in\Omega} \exp(\sum_{i=1}^{n} q_i x_i + \sum_{i=1}^{n}\sum_{j>i}^{n} q_{ij} x_{ij}) \end{aligned}$$

Regarding the representation of the distribution of a multivariate binary interaction model, we would like to have a more intuitive, compact description of the parameter space Θ, therefore we write the parameters q_{ij}, in a matrix

$$Q = \{q_{ij}\}_{i,j=1,...,n}$$

and get

$$p(x) = \frac{1}{z} \exp(\sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij} x_i x_j) = \frac{1}{z} \exp(\; x^{\mathrm{T}} \; Q \; x) =: p_Q(x)$$

Remark. We call Θ the parameter space and Q the matrix that contains the parameters. It is $\Theta = \{q_{11}, q_{12}, ..., q_{nn}\} = \{Q\}$. It is p_Q the distribution of the model with matrix Q that contains the parameters of the model.

After we introduced the general shape of a pairwise interaction binary model, we would like to fit the parameters based on data which are distributed by a binary distribution p on Ω.

1.2.2 Model selection

We are given d data points $x^{(1)}, ..., x^{(d)}$ with each $x^{(i)} \in \{0,1\}, i = 1, ..., d$ and $\Omega = \{0,1\}^n$ that are independently drawn from a pairwise interaction binary distribution p on Ω. To find the optimal choice of parameters Q, we are minimizing the negative likelihood function

$$\mathbf{p}_{ML} = \arg\min_{Q} -L(Q) = \arg\min_{Q} - \prod_{i=1}^{m} p_Q(x^{(i)}),$$

respectively minimizing the negative log-likelihood function

$$\mathbf{p}_{ML} = \arg\min_{Q} -\ell(Q) = \arg\min_{Q} - \sum_{i=1}^{m} \log p_Q(x^{(i)}).$$

Unfortunately, as far as we know, the minimization problem does not have an analytical solution, therefore we are going to use optimization algorithms to get a solution for the parameter estimation problem. Because the normalization factor is usually the sophisticated task for implementing a correctly working and high-performance optimization algorithm on categorical models, we are going to use instead the pseudo-likelihood function L_p. The pseudo-likelihood function is an approach to approximate the likelihood function, but with fewer complexity because the normalization constants are easier to track. We basically assume that

$$p_Q(x) = p_Q(x_1, ..., x_n) \approx p_Q(x_1|x_{-1}) \; ... \; p_Q(x_n|x_{-n}),$$

where

$$p_Q(x_r|x_{-r}) = p_Q(x_r|x_1, ..., x_{r-1}, x_{r+1}, ..., x_n) = \frac{\exp(2\sum_{j=1}^{n} q_{rj}x_r x_j - q_{rr}x_r x_r)}{1+\exp(q_{rr}+2\sum_{j=1, j\neq r}^{n} q_{rj}x_j)}.$$

Remark. We call the term $p_Q(x_r|x_{-r})$ the *node conditional* of the $r-th$ variable.

We get the maximum pseudo-likelihood parameter estimation with negative log-pseudo-likelihood function as follows:

$$\begin{aligned}
\mathbf{p}_{ML} &= \arg\min_Q -\ell(Q) \\
&\approx \arg\min_Q -\ell_p(Q) \\
&= \arg\min_Q -\sum_{i=1}^{d}\Big(\sum_{r=1}^{n} \log p_Q(x_r^{(i)}|x_{-r}^{(i)})\Big) \\
&= \arg\min_Q -\sum_{i=1}^{d}\Big(\sum_{r=1}^{n} \log\Big(\frac{\exp(2\sum_{j=1}^{n} q_{rj}x_r^{(i)}x_j^{(i)} - q_{rr}x_r x_r)}{1+\exp(q_{rr}+2\sum_{j=1, j\neq r}^{n} q_{rj}x_j^{(i)})}\Big)\Big) \\
&= \arg\min_Q -\sum_{i=1}^{d}\Big(\sum_{r=1}^{n}\Big(2\sum_{j=1}^{n} q_{rj}x_r^{(i)}x_j^{(i)} - q_{rr}x_r x_r\Big) \\
&\qquad\qquad - \log(1+\exp(q_{rr}+2\sum_{j=1, j\neq r}^{n} q_{rj}x_j^{(i)}))\Big)
\end{aligned}$$

Now that we derived the optimization problem that is going to estimate the parameters, we will look at an example and try to fit a pairwise interaction model to binary data.

1.3 An example of a pairwise interaction binary model

Chapter 2

Interactions of order $K = 3$ for binary models

In the following, we are going to extend the idea of pairwise interaction that we have seen in the previous chapter. From now on, we are going to examine models with interaction order $K = 3$.

2.1 Interaction models of order $K = 3$

Definition. Let $X = (X_1, ..., X_n)$ be a vector of random variables with the set of outcomes $\Omega = \{0, 1\}^n$. The distribution $p : \Omega \to [0, 1]$ of a *multivariate binary model of interaction order* $K = 3$ has the form

$$p(x) \propto \exp(\sum_{i,j,k}^{n} q_{ijk} x_i x_j x_k)$$

with parameters $Q = \{q_{ijk}\}_{i,j,k=1,...,n} \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$, if the normalization condition

$$\sum_{x \in \Omega} p(x) = 1$$

is satisfied.

Remark. It stands $\sum\limits_{i_1,...,i_K}^{n}$ for $\sum\limits_{i_1=1}^{n} ... \sum\limits_{i_K=1}^{n}$.

Remark. We call Q of the previous definition a *tensor* of order $K = 3$ of the multivariate binary model of interaction order $K = 3$. The dimension of the model and of the tensor is n and indicates the number of variables x_i.

Remark. In the following, we presume symmetry of the tensor which means that all parameters with permuted indexes are equal:

$$q_{ijk} = q_{ikj} = q_{jik} = q_{jki} = q_{kij} = q_{kji}$$

2.2 Representation of the node conditional

As we proceeded in the previous chapter, we are also going to look at the pseudo-likelihood function in order to estimate the parameters of the respective model. Therefore, we have to pick up on first the node conditional of the $r-th$ variable in the binary model of interaction order $K=3$. The node conditional $p(x_r|x_{-r})$ describes the slices that we cut out of the tensor Q for the $r-th$ variable. In other words, we are interested in all parameters that include index r. When we regard the permutations (because of symmetry) of the indexes of the single parameters, it turns out that some parameters have more than one occurrence. These occurrences can be counted, categorized into groups and summarized as one expression.

Remark. From now on, we are going to shorten the parameters q_{ijk} in the way that indexes with more than one occurrence are written as one. For example we write:

$$q_{iii} := q_i, \quad q_{iij} := q_{ij}$$

For instance, when we regard a model with interaction order $K=3$, the parameter q_{ij} stands for the parameters q_{iij} and q_{ijj} but we write it as one, because both describe the same interaction between variable x_i and x_j. If we regard higher interaction orders K, a shortened parameter q_i will stand for $q_{\underbrace{i \ldots i}_{K-times}}$.

Example. We consider the tensor Q for $K=3$. Now, we want the node conditional for $r=1$. We list all parameters that appear and categorize them, use their symmetry and write them down with the notation we introduced before:

$$\begin{aligned}
6\ q_{123} &= q_{123} + q_{132} + q_{213} + q_{231} + q_{312} + q_{321} \\
6\ q_{12} &= q_{112} + q_{121} + q_{211} + q_{122} + q_{212} + q_{221} \\
6\ q_{13} &= q_{113} + q_{131} + q_{311} + q_{133} + q_{313} + q_{331} \\
q_1 &= q_{111}
\end{aligned}$$

For the node conditional (without normalization) we get

$$p(x_1|x_{-1}) \propto \exp(\ 6\ q_{123} + 6\ (q_{12} + q_{13}) + q_{111}).$$

Let us generalize the previous approach with the following lemma.

Lemma 2.1. *For the node conditional of the $r-th$ variable in a multivariate binary model of interaction order $K=3$, we get*

$$p_Q(x_r|x_{-r}) \propto \exp(\ 6 \sum_{i<j, i\neq r, j\neq r}^{n} q_{rij} x_r x_i x_j + 6 \sum_{i\neq r}^{n} q_{ri} x_r x_i + q_r x_r)$$

Proof. We are going to count the occurrences for each m partition of $K=3$. To count the combinations for the first term $\sum q_{rij} x_r x_i x_j$, we are going to look at the $m=3$ partition of $K=3$. We only have one partition:

$$3 = 1 + 1 + 1$$

This means all parameters are different. We count the combinations that are represented by q_{rij}:

$$q_{rij} = q_{rji} = q_{irj} = q_{ijr} = q_{jri} = q_{jir}$$

We get $6 = 3!$ different combinations for the $m = 3$ partition of $K = 3$.

The next step is to count combinations for the second term $\sum q_{ri} x_r x_i$ of the $m = 2$ partition of $K = 3$. We get two partitions:

$$3 = 2 + 1 = 1 + 2$$

We count the combinations that are represented by the parameter q_{ri}:

$$q_{rri} = q_{rir} = q_{irr} = q_{rii} = q_{iri} = q_{iir}$$

We get $6 = 2 \cdot \binom{3}{2}$ different combinations.

For the last of the possible partitions, the $m = 1$ partition of $K = 3$, trivially we get only one combination q_{rrr} that is represented by q_r. $\square$

Unfortunately, the formula we just derived is difficult to use in terms of implementation. Because of its restrictions in the sums it would lay kind of skewed in the memory and therefore high-performance techniques like vectorization would be complicated to apply. But, we can also think of the node conditional $p(x_r|x_{-r})$ as slices that we cut out of a K-dimensional tensor at the $r - th$ variable. This considerations leads to a easier formula for the node conditional which we will introduce in the following lemma.

Proposition 2.2. *The sums of the previous lemma can also be expressed as*

$$\sum_{i,j}^{n} q_{rij} x_r x_i x_j = 2 \sum_{i<j, i\neq r, j\neq r}^{n} q_{rij} x_r x_i x_j + 3 \sum_{i\neq r}^{n} q_{ri} x_r x_i + q_r x_r$$

$$\sum_{i}^{n} q_{ri} x_r x_i = \sum_{i\neq r}^{n} q_{ri} x_r x_i + q_r x_r$$

Remark. The findings of proposition 2.2 are simply a result of counting the occurrences of the single terms.

Lemma 2.3. *The node conditional of the $r - th$ variable in a multivariate binary model of interaction order $K = 3$ can also be represented as*

$$p_Q(x_r|x_{-r}) \propto \exp(\ 3 \sum_{i,j}^{n} q_{rij} x_r x_i x_j - 3 \sum_{i}^{n} q_{ri} x_r x_i + q_r x_r).$$

Proof. We will show this by using the terms from proposition 2.2. We rearrange the equations

$$\sum_{i<j, i\neq r, j\neq r}^{n} q_{rij}x_r x_i x_j = \frac{1}{2}\sum_{i,j}^{n} q_{rij}x_r x_i x_j - \frac{3}{2}\sum_{i\neq r}^{n} q_{ri}x_r x_i - \frac{1}{2}q_r x_r$$
$$\sum_{i\neq r}^{n} q_{ri}x_r x_i = \sum_{i}^{n} q_{ri}x_r x_i - q_r x_r$$

and use them for the node conditional:

$$\begin{aligned} p_Q(x_r|x_{-r}) &\propto \exp(\ 6 \sum_{i<j, i\neq r, j\neq r}^{n} q_{rij}x_r x_i x_j + 6\sum_{i\neq r}^{n} q_{ri}x_r x_i + q_r x_r) \\ &= \exp(\ 3\sum_{i,j}^{n} q_{rij}x_r x_i x_j - 3\sum_{i\neq r}^{n} q_{ri}x_r x_i - 2q_r x_r) \\ &= \exp(\ 3\sum_{i,j}^{n} q_{rij}x_r x_i x_j - 3\sum_{i}^{n} q_{ri}x_r x_i + q_r rr) \end{aligned}$$

□

To illustrate the previous proof, we are going to look at a example for $n = 4$.

Example. Let $n = 4$ and $K = 3$. We would like to cut out the node conditional for the $r - th$ variable for $r = 2$. We have the tensor Q and want to extract all parameters $q_{rij}, i, j = 1, ..., 4$, that contain the index $r = 2$. It is illustrated in Figure 2.1 and Figure 2.2.

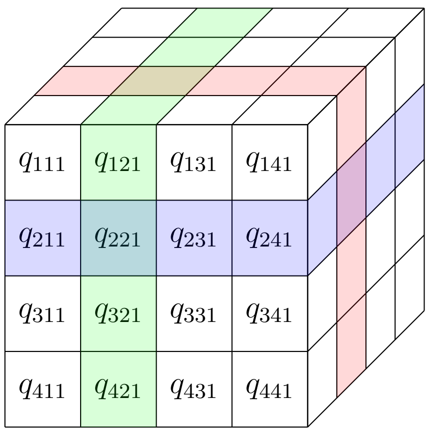

Figure 2.1: A 3-dimensional tensor with marked node conditionals for $r = 2$.

After we cut out the slices that all include index $r = 2$ (first sum in the formula),

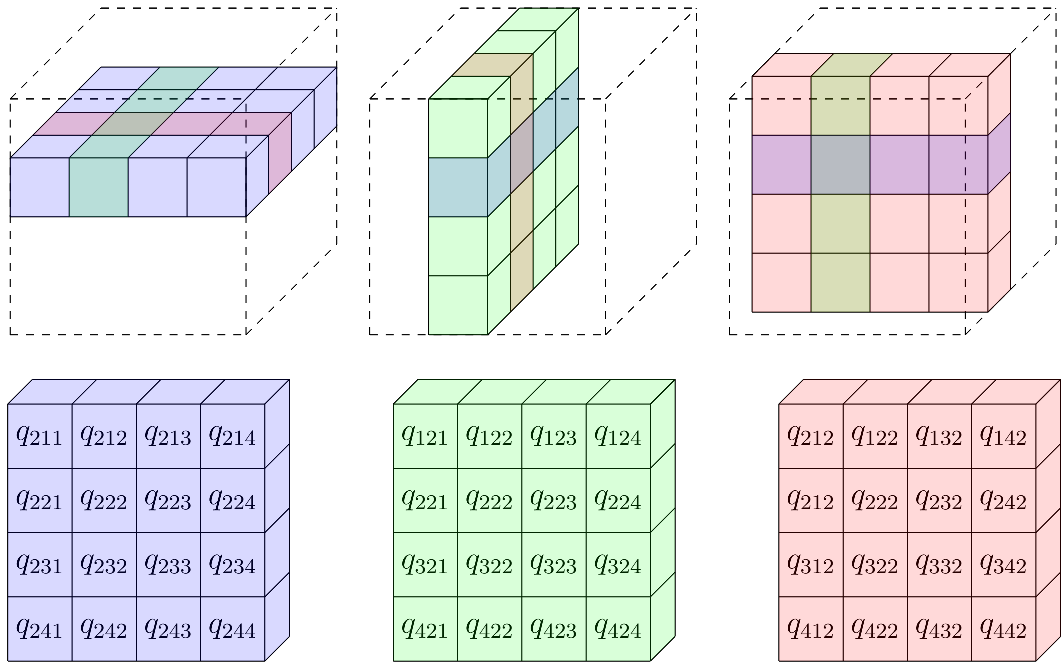

Figure 2.2: Slices that we cut out of the tensor Q that contain index $r = 2$.

we notice that we get too many of some parameters. As we can see in Figure 2.2, the rows $q_{22i} = q_{2i}$, $q_{2i2} = q_{2i}$ and $q_{i22} = q_{2i}, i = 1, 2, 3, 4$ appear twice, that makes

$$2q_{22i} + 2q_{2i2} + 2q_{i22} = 6q_{2i},\ i = 1, 2, 3, 4$$

even though we need each of them only once. Therefore we have to subtract them three times (second summand in the formula). We get

$$q_{22i} + q_{2i2} + q_{i22} = 3q_{2i},\ i = 1, 2, 3, 4$$

which is the correct number of occurrences of q_{2i}. But then, we subtracted the remaining summand q_{222} one time too much, therefore we add it another time (third summand in the formula). That procedure reminds us on the inclusion exclusion principle in combinatorics which generalizes the familiar method of obtaining the number of elements in the union of finite sets. It is clarified by Figure 2.3.

Remark. In the change of representation of the node conditional $p_Q(x_r|x_{-r})$ we went from picking out the single parameters from the tensor Q to using an alternating sum of tensors whose order is getting lower.

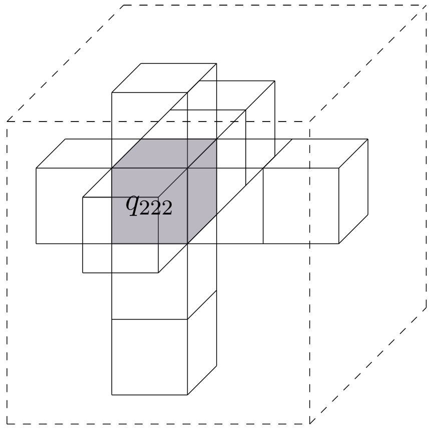

Figure 2.3: Illustration of the parameters that appear more than once when we cut out the slices for the node conditional for $r = 2$.

2.3 Use of einstein sum notation

2.4 Construction of pseudo-likelihood function

Chapter 3

Binary models with arbitrary number of interactions

Now, we move on to generalize the approaches we discussed in the previous chapters. We assume a arbitrary variable interaction order $1 \leq K \leq n$, define the corresponding multivariate binary interaction model and derive the node conditional that depends on K to estimate the parameter in the tensor Q of order K.

3.1 Interaction models of order K

Definition. Let $X = (X_1, ..., X_n)$ be a vector of random variables with the set of outcomes $\Omega = \{0,1\}^n$. The distribution $p : \Omega \to [0,1]$ of a *multivariate binary model of interaction order* K has the form

$$p(x) \propto \exp(\sum_{i_1,...,i_K}^{n} q_{i_1...i_K} \, x_{i_1} ... \, x_{i_K} \,)$$

with parameters $Q = \{q_{i_1...i_K}\}_{i_1,...,i_K=1,...,n} \in \underbrace{\mathbb{R}^n \times ... \times \mathbb{R}^n}_{K-times}$, if the normalization condition

$$\sum_{x \in \Omega} p(x) = 1$$

is satisfied.

Remark. We call Q of the previous definition a *tensor* of order K with dimension n.

Definition. Let Q be a Tensor of order K. Then, Q is *symmetric* when

$$q_{i_1,...,i_K} = q_{\sigma(i_1),...,\sigma(i_K)}$$

for all permutations σ of the set $\{i_1, ..., i_K\}$. That means all entries in Q with permuted indexes are equal.

Remark. In the following, we presume symmetry of the tensor Q.

3.2 Node conditional

Before we get to the general node conditional, we define two different ways to represent them. It follows the findings from the previous chapter where we already defined them for interaction order $K = 3$. No we are going to generalize this approach.

Definition. Let p_Q be the distribution of a multivariate binary model of interaction order K with symmetric tensor Q. We call

$$p_Q(x_r|x_{-r}) \propto \exp(\sum_{m=1}^{K} \tau_{K,m} \sum_{i_1<...<i_m,\ \forall i_l: i_l \neq r}^{n} q_{ri_1...i_m}\ x_r x_{i_1}...\ x_{i_m})$$

the *standard representation* for the $r-th$ node conditional with integer coefficients $\tau_{K,m}, m = 1, ..., K$.

Definition. Let p_Q be the distribution of a multivariate binary model of interaction order K with symmetric tensor Q. We call

$$p_Q(x_r|x_{-r}) \propto \exp(\sum_{m=1}^{K} \pi_{K,m} \sum_{i_1,...,i_m}^{n} q_{ri_1...i_m}\ x_r x_{i_1}...\ x_{i_m})$$

the *slice representation* for the $r-th$ node conditional with integer coefficients $\pi_{K,m}, m = 1, ..., K$.

Remark. The difference between these two ways of representation and their coefficients $\tau_{K,m}$ and $\pi_{K,m}$ lays in the second sum. In contrast to the slice representation we do not have the same index in a single summand twice in the standard representation.

Remark. As we did it in the previous chapter, we are going to shorten the parameters $q_{i_1,...,i_K}$ to distinguish between parameters with a different number of different indexes.

The coefficients $\tau_{K,m}, \pi_{K,m}$ in the previous definitions are the result of counting occurrences of differently categorized parameters, which needs to be done for each K. In the following, we are going to analyze systematically the behavior of these coefficients and try to find a generalized expression to calculate them for each K and m. We will follow a similar strategy as we did it in the chapter before for interaction order $K = 3$.

3.2.1 Standard representation

At first, we are going to analyze the coefficients $\tau_{K,m}$ of the standard representation. When we consider our observations from the previous chapter and take a closer look at the proof of lemma 2.1, we recognize that the coefficients $\tau_{K,m}$ correspond with the m partitions of K.

Definition. Let

$$P_{K,m} := \{(k_1, ..., k_m) \mid K = k_1 + ... + k_m,\ k_1 \geq k_2 \geq ... \geq k_m \geq 1,\ k_i \in \mathbb{N}\}$$

be the set of all m partitions of K.

Definition. Let $(k_1, ..., k_m) \in P_{K,m}$ be a m partition of K. Then we define the *counting vector* for a partition $(k_1, ..., k_m)$ as

$$a_{(k_1,...,k_m)} := (a_1, ..., a_K),\ a_i \text{ is the number of occurrences of } i \in (k_1, ..., k_m),$$

that represents the occurrences of each integer in $(k_1, ..., k_m)$.

Proposition 3.1. *It is*

$$\sum_{i=1}^{K} a_i = m$$

for a counting vector $a_{(k_1,...,k_m)}$ of a m partition of K.

Example. Let us consider the 3 partitions of 6. They are summarized in the set

$$P_{6,3} = \{(4,1,1),\ (3,2,1),\ (2,2,2)\}$$

and its elements have the counting vectors

$$a_{(4,1,1)} = (2,0,0,1,0,0),\ a_{(3,2,1)} = (1,1,1,0,0,0),\ a_{(2,2,2)} = (0,3,0,0,0,0).$$

Before we get to the lemma where we will count the number of permutations of indexes for each m partition of K, we will look at an example to understand how we are going to count and categorize them. The proof afterwards follows the same principle.

Example. Let $K = 5$ and $m = 3$. We count all permutations of indexes for the partitions in $P_{5,3} = \{(3,1,1),\ (2,2,1)\}$. Let us first look at $5 = 3+1+1$. We get 20 elements that are:

$$rrrij,\ rrijr,\ rijrr,\ ijrrr,\ rrirj,\ rirjr,\ irjrr,\ rirrj,\ irrjr,\ irrrj$$

$$rrrji,\ rrjir,\ rjirr,\ jirrr,\ rrjri,\ rjrir,\ jrirr,\ rjrri,\ jrrir,\ jrrri$$

This is the case because we first choose three r's from five, then we choose one i from two and then one j from one:

$$\binom{5}{3} \cdot \binom{2}{1} \cdot \binom{1}{1} = 10 \cdot 2 \cdot 1 = 20$$

This correspondents with the *multinomial coefficient*:

$$\binom{5}{3,1,1} = \frac{5!}{3!1!1!} = \binom{5}{3} \cdot \binom{2}{1} \cdot \binom{1}{1}$$

However, we also need to consider that each index is different. That means we also count the cases where i occurs three times and j three times. That multiplicities can be expressed with the multinomial coefficient of the counting vector

$$a_{(3,1,1)} = (2,0,1,0,0),$$

that can be considered as a 2 partition of 3 with $3 = 2+1$. Its multinomial coefficient would be:

$$\binom{3}{2,1} = \frac{3!}{2!1!} = 3$$

Finally, we get the number of permutations of indexes for the partition $(3, 1, 1)$ as a product of the multinomial coefficient of the partition and the multinomial coefficient of its the counting vector:

$$\binom{5}{3,1,1} \cdot \binom{3}{2,1} = 20 \cdot 3 = 60$$

The same can be done with the other partition $(2, 2, 1) \in P_{5,3}$. Together with its counting vector $a_{2,2,1} = (1, 2, 0, 0, 0)$, we get

$$\binom{5}{2,2,1} \cdot \binom{3}{2,1} = \frac{5!}{2!2!1!} \cdot \frac{3!}{2!1!} = 30 \cdot 3 = 90$$

permutations of the indexes. Overall, for the set $P_{5,3}$, we get

$$\binom{5}{3,1,1} \cdot \binom{3}{2,1} + \binom{5}{2,2,1} \cdot \binom{3}{2,1} = 60 + 90 = 150$$

permutations of indexes.

Lemma 3.2. *Let $K \in \mathbb{N}$ be arbitrary and $m \in \{1, ..., K\}$. Let $P_{K,m}$ be the set of all m partitions of K and $\tau_{K,m}$ the coefficients of the standard representation of the $r-th$ node conditional. Then*

$$\tau_{K,m} = \sum_{(k_1,...,k_m) \in P_{K,m}} \binom{K}{k_1, ..., k_m} \binom{m}{a_1, ..., a_K}$$

where $a_{(k_1,...,k_m)} = (a_1, ..., a_K)$ is the corresponding counting vector of $(k_1, ..., k_m)$.

Proof. Let $(k_1, ..., k_m)$ be a m partition of K and $a_{(k_1,...,k_m)} = (a_1, ..., a_K)$ its corresponding counting vector. To count the permutations of

$$K = k_1 + ... + k_m$$

without regarding the order, we need to calculate:

$$\binom{K}{k_1} \cdot \binom{K-k_1}{k_1} \cdot ... \cdot \binom{K-k_1-...-k_{m-1}}{k_m} = \frac{K!}{k_1! \dots k_m!} = \binom{K}{k_1, ..., k_m}.$$

However we have different indexes, therefore we have to regard the order of the partition as well. For instance for $k_1 \neq k_2$, the permutations of indexes of $(k_1, k_2..., k_m)$ and $(k_2, k_1, ..., k_m)$ must be counted individually. For counting the permutations, we use the counting vector that displays the multiplicities of the single k_i. Because it is $\sum_{i=1}^{K} a_i = m$, the non zero elements of the counting vector can be seen as a partition of m. Therefore we get the multinomial coefficient to count the permutations in the initial permutation $(k_1, ..., k_m)$:

$$\binom{m}{a_1, ..., a_K}$$

Together we get the number of the permutations of the indexes of a m partition of K as a product of

$$\binom{K}{k_1, ..., k_m} \cdot \binom{m}{a_1, ..., a_K}.$$

This needs to be done for all m permutations of K. □

K							
1	1						
2	1	2					
3	1	6	6				
4	1	14	36	24			
5	1	30	150	240	120		
6	1	62	540	1560	1800	720	
7	1	126	1806	8400	16800	15120	5040

Figure 3.1: Coefficients $\tau_{K,m}$ for $K = 1, ..., 7$, in the K-th row and m-th position

The proofs shows us how the single permutation of indexes must be counted and gives us the remaining formula. When we compute the coefficients $\tau_{K,m}$, it generates a triangle that can be seen in figure 3.1.

Proposition 3.3. *For a fixed K ($K - th$ row of the triangle in figure 3.1), the coefficients $\tau_{K,m}$ can also be used to calculate potencies with any basis n as a sum of binomial coefficients:*

$$n^K = \sum_{i=1}^{K} \tau_{K,i} \cdot \binom{n}{i}$$

Proposition 3.4. *The coefficients $\tau_{K,m}$ can also be calculated recursively with*

$$\tau_{K,m} = (\tau_{K-1,m-1} + \tau_{K-1,m}) \cdot m$$

whereby it is $\tau_{K,K} = K!$ and $\tau_{K,1} = 1$.

Remark. The results of propositions 3.3 and 3.4 are solutions of general combinatorial problems that are described in QUELLE (wikipedia: pascalsches dreieck)

Remark. The term in proposition 3.3 also expresses the number of parameters in a n-dimensional tensor of order K and its distribution of the single multiplicities.

Example. Let $K = 4$ and $n = 6$. The parameters of a 6-dimensional tensor of order 4 are distributed as follows:

$$\begin{aligned} 6^4 &= 1 \cdot \binom{6}{1} + 14 \cdot \binom{6}{2} + 36 \cdot \binom{6}{3} + 24 \cdot \binom{6}{4} \\ &= 1 \cdot 6 + 14 \cdot 15 + 36 \cdot 20 + 24 \cdot 15 \\ &= 6 + 210 + 720 + 360 \\ &= 1296 \end{aligned}$$

Thereby we can recognize that there are 6 parameters with one index, 210 parameters with two different indexes, 720 parameters with three different indexes, 360 parameters with four different indexes and all together 1296 indexes.

3.2.2 Slice representation

After we introduced the standard representation and calculated their coefficients $\tau_{K,m}$, we get to the second type of representation and derive the coefficients $\pi_{K,m}$ for the slice representation. As it was pointed out before, the slice representation is mainly used for a highly efficient implementation, because the occurring entire sums can be broken down to simple matrix-vector products.

Definition. Let $K \in \mathbb{N}$ be arbitrary and Q the tensor of order K from the respective model of order K. Then it is $Q_{-m,r}, i \in \{1, ..., K\}, r \in \{1, ..., n\}$ the *m-th subtensor* of Q with order $K - m$ and fixed index r. We also define $Q_{-K,r} = q_r$.

To attain the transition between the two representations, we have to express the entire sums of parameters from the slice representation

$$Q_{-m,r} = \sum_{i_1,...,i_{K-m}} q_{ri_1...i_{K-m}},\ m \in \{1, ..., K\}$$

as a term of sums from the standard representation. As we indicated the terms for $K = 2$ and $K = 3$ in proposition 2.2, we will do it for higher $K's$ as well. Again, it is simply a matter of counting the different types of parameters and assign them to the correct group.

Proposition 3.5. *Let $K \in \mathbb{N}$ be arbitrary and $m \in \{1, ..., K\}$. Then, the sum of the elements of the $m-th$ subtensor of order $K - 1$ can be expressed as*

$$q_r + \sum_{i_1,...,i_{K-m}}^{n} q_{ri_1...i_{K-m}} = \psi_{1,1}\ q_r + \sum_{j=2}^{K-m+1} \psi_{K,j} \sum_{i_1<...<i_{j-1},\ \forall i_l: i_l \neq r}^{n} q_{ri_1...i_{j-1}}$$

with coefficients $\psi_{K,m}$ and $\psi_{1,1} = 1$.

Remark. We suppose a model of interaction order K and therefore a tensor of order K. When we are talking about the $r - th$ node conditional, then we fix one index and the $1st$ subtensor at least. Therefore m can not be lower than 1 in the previous definition.

Example. Let $K = 3$, $n = 3$, $r = 1$ and Q be the tensor of order K of the respective model. Then, the sum of elements of the 1st subtensor can be expressed as

$$\begin{aligned}
\sum_{i,j} q_{1ij} &= q_{111} + q_{112} + q_{113} + q_{121} + q_{122} + q_{123} + q_{131} + q_{132} + q_{133} \\
&= q_1 + q_{12} + q_{13} + q_{12} + q_{12} + q_{123} + q_{13} + q_{132} + q_{13} \\
&= q_1 + 3q_{12} + 3q_{13} + 2q_{123} \\
&= q_1 + 3\sum_{i\neq 1}^{3} q_{ri} + 2\sum_{i<j,\ i,j\neq 1}^{3} q_{rij}
\end{aligned}$$

We get the coefficients $\psi_{3,1} = 1$, $\psi_{3,2} = 3$, $\psi_{3,3} = 2$.

Lemma 3.6. *Let $K \in \mathbb{N}$ be arbitrary and $m \in \{1, ..., K\}$. Let $P_{K,m}$ be the set of all m partitions of K and $\psi_{K,m}$ the coefficients for the representation of the $m-th$ subtensor of Q. Then*

$$\psi_{K,m} = \frac{1}{m} \sum_{(k_1,...,k_m) \in P_{K,m}} \binom{K}{k_1,...,k_m} \binom{m}{a_1,...,a_K} = \frac{\tau_{K,m}}{m}$$

where the $\tau_{K,m}$'s stand for the coefficients of the standard representation of the $r-th$ node conditional and $a_{(k_1,...,k_m)} = (a_1, ..., a_K)$ is the corresponding counting vector of $(k_1, ..., k_m) \in P_{K,m}$.

Proof. TODO □

Proposition 3.7. *The coefficients $\psi_{K,m}$ can also be calculated recursively with*

$$\psi_{K,m} = (\ (m-1)\ \psi_{K-1,m-1} + m\ \psi_{K-1,m}\ )$$

whereby it is $\psi_{K,K} = K!$ and $\psi_{K,1} = 1$.

K							
1	1						
2	1	1					
3	1	3	2				
4	1	7	12	6			
5	1	15	50	60	24		
6	1	32	180	390	360	120	
7	1	63	602	2100	3360	2520	720

Figure 3.2: Coefficients $\psi_{K,m}$ for $K = 1, ..., 7$, in the K-th row and m-th position

Proposition 3.8. *For a fixed K ($K-th$ row of the triangle in figure 3.2), the coefficients $\psi_{K,m}$ can also be used to calculate potencies with any basis n as a sum of binomial coefficients:*

$$n^{K-1} = \sum_{i=1}^{K} \psi_{K,i} \cdot \binom{n}{i-1}$$

Remark. The results from propositions 3.7 and 3.8 follow directly from the corresponding propositions about the coefficients $\tau_{K,m}$ in the previous section.

After calculating the coefficients $\psi_{K,m}$, we are going to examine the coefficients $\pi_{K,m}$. As we did it in section 2.2, we are going from the standard representation of the node conditional to the slice representation. The triangle in figure 3.2 will support us. First, we will look at an example.

Example. Let $K = 4$ and $n > K$ be arbitrary. We get the term of the standard representation for the $r-th$ node conditional as follows:

$$24 \sum_{i<j<k,\, i,j,k\neq r}^{n} q_{rijk} x_r x_i x_j x_k + 36 \sum_{i<j,\, i,j\neq r}^{n} q_{rij} x_r x_i x_j + 14 \sum_{i\neq r}^{n} q_{ri} x_r x_i + q_r x_r$$

Now, we can derive the coefficients $\pi_{K,m}$ from the coefficients $\tau_{K,m}$ with the help of the coefficients $\psi_{K,m}$ in the triangle in figure 3.2:

$$\begin{aligned}
&= 4 \sum_{i,j,k}^{n} q_{rijk} x_r x_i x_j x_k - 12 \sum_{i<j,\, i,j\neq r}^{n} q_{rij} x_r x_i x_j - 14 \sum_{i\neq r}^{n} q_{ri} x_r x_i - 3\; q_r x_r \\
&= 4 \sum_{i,j,k}^{n} q_{rijk} x_r x_i x_j x_k - 6 \sum_{i,j}^{n} q_{rij} x_r x_i x_j + 4 \sum_{i\neq r}^{n} q_{ri} x_r x_i + 3\; q_r x_r \\
&= 4 \sum_{i,j,k}^{n} q_{rijk} x_r x_i x_j x_k - 6 \sum_{i,j}^{n} q_{rij} x_r x_i x_j + 4 \sum_{i}^{n} q_{ri} x_r x_i - q_r x_r
\end{aligned}$$

We get the slice representation of the $r-th$ node conditional for $K = 4$:

$$p_Q(x_r|x_{-r}) \quad \propto \exp(\, 4 \sum_{i,j,k}^{n} q_{rijk} x_r x_i x_j x_k - 6 \sum_{i,j}^{n} q_{rij} x_r x_i x_j + 4 \sum_{i}^{n} q_{ri} x_r x_i - q_r x_r)$$

Lemma 3.9. *Let $K \in \mathbb{N}$ be arbitrary and $m \in \{1, ..., m\}$. Let $\pi_{K,m}$ be the coefficients of the slice representation of the $r-th$ node conditional of a multivariate binary model of interaction order K, then*

$$\pi_{K,m} = (-1)^{K-m} \tbinom{K}{m-1}$$

Proof. TODO □

Remark. The coefficients $\pi_{K,m}$ correspond with the values of the pascal's triangle.

K													
1							1						
2						-1		2					
3					1		-3		3				
4				-1		4		-6		4			
5			1		-5		10		-10		5		
6		-1		6		-15		20		-15		6	
7	1		-7		21		-35		35		-21		7

Figure 3.3: Coefficients $\pi_{K,m}$ for $K = 1, ..., 7$, in the K-th row and m-th position

3.3 Use of einstein sum notation

3.4 Construction of the pseudo-likelihood function

Chapter 4

Interaction order n for categoricals

Chapter 5

Some algorithm part

Chapter 6

Conclusion