High-Dimensional Log-Linear Model Selection with Applications in Molecular Biology

A dissertation submitted to
ETH ZURICH

for the degree of
Doctor of Sciences

presented by
CORINNE DAHINDEN
Dipl. Math. ETH
born May 16, 1978
citizen of Weggis (LU)

accepted on the recommendation of
Prof. Dr. Peter Bühlmann, examiner
Prof. Dr. Leonhard Held, co-examiner
PD Dr. Peter Schraml, co-examiner

2009
Acknowledgements

First, I would like to thank Prof. Peter Bühlmann for supervising this thesis. His office door has always been open to me and I have felt welcome anytime to discuss new ideas or when I was in need for technical or conceptual support. Prof. Bühlmann has created a very motivating working atmosphere, providing excellent advise and ideas but also leaving enough room to pursue own ideas. I have always appreciated his encouraging and reassuring spirit and during these four years, I would probably have been lost many times without him.

I also thank my thesis committee member and co-examiner Prof. Leonhard Held for providing input and having taken time to evaluate my work.

My appreciation further goes to PD Dr. Peter Schraml for the nice collaboration, refereeing my thesis and for acting as co-examiner. I know it must have been hard sometimes to work together with a mathematician.

I also thank the Competence Center for Systems Physiology and Metabolic Diseases for financial support and providing an interdisciplinary environment.

And of course, I want to thank each and every member of the current and former SfS team I had the pleasure to work with. You have all contributed towards providing such a stimulating and fun environment. I always had the impression that whenever and on whatever topic I was stuck, we had the expert to ask and everybody was always very helpful and willing to help. Whether it had been the secretaries concerning organizational matters, the professors concerning statistical matters or the fellow PhD colleagues for social matters, cheering me up when I was stuck and laughing many times about the innumerable jokes we have played on each other. I have always enjoyed their company. A quotation
goes that “a journey is always easier when you travel together”, and we really have travelled together! Not only the journey through our PhD projects but also our joint journeys into the world where we have shared countless great adventures. Had it been deep down in the Atlantic or Indian Ocean or high up on many mountain tops, sledging down the slopes of Rigi, riding roller coasters somewhere in the world, skiing in Savognin or just being lucky to fly Business-Class to someplace. It was a great time and I will miss you!

I am tempted to individually thank all of my friends but by fear of leaving someone out, I will simply say thank you very much to you all. You have provided a good work-life balance while spending many weekends and vacations together. Special thanks goes to my ASVZ friends for having cleared my mind many times after hard thinking and for lifting weights and hopping around in group fitness lessons together with me. It has always been fun.

Last but not least I want to thank my parents. I can hardly express how grateful I am for everything they have done for me, their unconditional support and their confidence throughout the past thirty years.
# Contents

Abstract xi

Zusammenfassung xiii

1 Introduction 1

1.1 Contingency Tables 2

1.2 Log-Linear Model 3

1.2.1 Independence Model 4

1.2.2 Saturated Model 4

1.2.3 Correspondence to Conditional Independence 5

1.2.4 Hierarchical Log-Linear Models 6

1.3 Graphical Models 7

1.4 Model Selection in Log-Linear Models 8

1.4.1 Parameter Estimation with MLE 8

1.4.2 Problem with Zero Cell Counts 9
1.4.3 Testing .............................................. 10
1.5 Thesis Overview ...................................... 11

2 Penalized Likelihood for Sparse Contingency Tables with an Application to Full-Length cDNA Libraries 13

2.1 Introduction ........................................... 14
2.2 Terminology and Notation ............................. 17
  2.2.1 General Introduction to Contingency Tables and Log-Linear Models .......................... 17
  2.2.2 Graphical Models ................................. 20
  2.2.3 Model Selection - Non-Hierarchical Versus Hierarchical Models ............................ 20
2.3 Algorithms .............................................. 21
  2.3.1 \( \ell_1 \)-Regularized Model Selection .............. 21
  2.3.2 Level-\( \ell_1 \)-Regularized Model Selection .......... 24
  2.3.3 Non-hierarchical Bayesian Model Selection ....... 25
  2.3.4 Hierarchical Bayesian Model Selection .......... 28
2.4 Implementation .......................................... 29
  2.4.1 Prior Specification for Bayesian Methods .......... 31
2.5 Simulation Study ........................................ 32
  2.5.1 Data ............................................. 32
  2.5.2 Criteria .......................................... 33
  2.5.3 Results of Simulation Study .................... 33
2.6 Application to Single-Gene Libraries .................................. 36
  2.6.1 Datasets ........................................................................... 36
  2.6.2 Results of Application to Single-Gene Libraries ............... 37

3 Decomposition and Model Selection in Log-Linear Interaction Models for Large Contingency Tables 45

  3.1 Background ................................................................. 46
  3.2 Preliminaries ............................................................... 48
    3.2.1 Log-Linear Interaction Model ................................. 49
    3.2.2 Graphical Models .................................................. 51
    3.2.3 Graphical Models and Hierarchical Log-Linear Models ...... 52
    3.2.4 Collapsibility ....................................................... 54
    3.2.5 Decomposability .................................................. 55
  3.3 Divide-and-Conquer Algorithm ........................................... 57
    3.3.1 Selection of Marginal Models to Collapse on ............... 58
    3.3.2 Combination of results ........................................... 61
  3.4 Graphical Model Selection Procedures ............................... 62
    3.4.1 Model Selection Procedures used in Combination with the Decomposition Approach ............................... 62
    3.4.2 Global Model Selection Procedures ........................... 65
  3.5 Simulation Study .......................................................... 67
    3.5.1 Results .............................................................. 67
3.6 Application to Tissue Microarray Data ................. 77
   3.6.1 Tissue Microarray Technology ...................... 77
   3.6.2 Data ........................................... 77
   3.6.3 Graphical Model ................................ 78
3.7 Discussion .......................................... 80

4 Mining Tissue Microarray Data to Unhide Disregulated Tumor Suppressor Pathways in Clear Cell Renal Cell Cancer 83
   4.1 Introduction ....................................... 84
   4.2 Materials and Methods ............................... 87
      4.2.1 Tissue specimen and TMA construction ............ 87
      4.2.2 Immunohistochemistry ........................... 88
      4.2.3 Data Description and Preprocessing ............... 90
      4.2.4 Recursive Bootstrap Elimination of Variables in Cox Regression ......................... 91
      4.2.5 Log-Linear- and Graphical Model ................. 91
   4.3 Results ............................................ 93
      4.3.1 Protein Expression Patterns, Tumor Phenotype and Survival .......................... 93
      4.3.2 Recursive Bootstrap Elimination .................. 94
      4.3.3 Graphical Log-Linear Model ....................... 96
   4.4 Discussion ......................................... 97
5 Appendix

5.A Technical Details of $\ell_1$-Regularization Algorithm . . . . 107
5.B Details of Model Selection Results Using MCMC . . . . 109
5.C Definition of the Index . . . . . . . . . . . . . . . . . . . . . . 109

Bibliography

6 Curriculum Vitae

Curriculum Vitae
Abstract

The joint analysis of several categorical variables is a common task in many areas of research and is becoming central to systems biology investigations whose goal is to identify potentially complex interactions among variables belonging to a network. Interactions of arbitrary complexity are traditionally modeled in statistics by log-linear models and the structure of a log-linear model can be visually represented by a graphical model. It is challenging to extend these concepts to high dimensional and potentially sparse data arising in computational biology. An important example, which provided the motivation for the first part of this thesis, is the analysis of so-called full-length cDNA libraries of alternatively spliced genes, where relationships among the presence of various exons in transcript species are investigated.

We develop methods to perform model selection and parameter estimation in log-linear models for the analysis of sparse contingency tables for analyzing the interaction of two or more factors. Maximum Likelihood estimation of log-linear model coefficients might not be appropriate because of the presence of zeros in the table’s cells and new methods are required. We propose a computationally efficient $\ell_1$-penalization approach extending the Lasso algorithm to this context and compare it to other procedures in a simulation study. We then illustrate these algorithms on contingency tables arising from full-length cDNA libraries.
In addition to the issue that high-dimensional tables naturally feature a large number of sampling zeros, which often leads to the nonexistence of MLE estimation, we also face the problem of the exponentially growing number of parameters in the saturated model as the number of variables increases. This means a heavy burden as far as computational power is concerned. Even if we restrict ourselves to models of lower order interaction, we still face problems as the number of cells remains unchanged which prevents us from using standard methods or the regularization method described above in high-dimensional log-linear model selection. We therefore present a divide-and-conquer approach where we first divide the problem into several lower-dimensional problems, solve these with the proposed method, which includes the $\ell_1$-penalization approach that avoids the sampling zero problem, and then combine these solutions to form a global solution to the problem. In a simulation study we demonstrate the performance of the proposed method and we apply it to a bio-medical problem in cancer research.

For the bio-medical cancer research problem, we applied a recursive bootstrap elimination scheme in combination with Cox regression to detect patterns of molecular markers which are favourable as far as survival is concerned. This pattern was then analyzed together with biomedical background knowledge and the information from the graphical model fitted with the divide-and-conquer algorithm described above. Some interesting biological insight was gained whose validation will be part of future research.
Zusammenfassung


Wir entwickeln Methoden für die Modellwahl und Parameterschätzungen in Log-Linearen Modellen, speziell für die Analyse von dünnbesetzten Kontingenztafeln mit welchen wir Interaktionen von zwei oder mehreren Faktoren studieren wollen. Die Anwendung von Maximum-
Zusammenfassung


Ferner wurde ein rekursiver Bootstrap-Eliminationsansatz in Kombination mit Cox Regression auf denselben bio-medizinischen Datensatz angewandt, um Muster von molekularen Markern zu erkennen, welche die Überlebenszeit positiv beeinflussen. Diese Muster wurden dann zusammen mit dem bio-medizinischen Hintergrundwissen und dem
Zusammenfassung

Chapter 1

Introduction

Analyzing categorical data is a common task in many areas of biology. A categorical variable is one for which the measurement scale consists of a set of categories. For instance in molecular biology, gene expression may be categorized into “not expressed”, “moderately expressed” and “highly expressed”. However, categorical scales are by no means restricted to biology. They are common in many other areas such as in medicine, where factors such as stage of a disease, severity of an injury or recovery from surgery are measured on a categorical scale or in social sciences where people are cross-classified according to gender, race, marital status or social class.

Analyzing associations between categorical variables is generally done by studying the corresponding contingency table: The cell probabilities are either estimated directly or by expanding the log of the probability by a linear model, a so-called log-linear model. Then the result may be translated into a so-called graphical model where the dependencies between the underlying distribution of the variables can be visually represented and the conditional independence structure can be directly
read off the graph.

In this chapter where the basic concepts are explained, we restrict ourselves to three categorical variables and we study associations among these to keep the notation simple. Once we understand log-linear models for three variables, we can readily extend the concepts to multi-way tables involving any number of variables.

1.1 Contingency Tables

Let $X$, $Y$ and $Z$ be three categorical variables, $X$ having $I$ levels, $Y$ having $J$ levels and $Z$ having $K$ levels. When we cross-classify objects according to these three variables, there are $IJK = N$ possible combinations of classifications. The variables $X,Y,Z$ have a joint distribution. We display this distribution in a $I \times J \times K$ contingency table. The cells of the table represent the $IJK$ possible outcomes and a cell probability is denoted by $p_{ijk}$, where $p_{ijk}$ is the probability that variable $X$ is in category $i$, variable $Y$ in category $j$ and $K$ in $k$. A table containing frequency counts is called a contingency table and $p_{ijk}$ represents the joint distribution of $X,Y$ and $Z$.

The marginal distribution of a variable is the sum over the remaining variables, denoted e.g. by $p_{i++}$ for the marginal distribution of $X$:

$$p_{i++} = \sum_{j} \sum_{k} p_{ijk}.$$ 

Suppose we have observed counts $\{n_{ijk}, i \in \{1, \ldots, I\}, j \in \{1, \ldots, J\}, k \in \{1, \ldots, K\}\}$ in the $IJK = N$ cells of the contingency table. It is generally assumed that the observed data has been generated by a multinomial model with probabilities $p_{ijk}$, meaning that the probability having observed the counts $n_{ijk}$ given that the total number of counts is fixed
1.2 Log-Linear Model

\[ \sum_{i,j,k} n_{ijk} = n \] is the following expression:

\[
\left( \frac{n!}{\prod_{i,j,k} n_{ijk}!} \right) \prod_{i,j,k} p_{ijk}^{n_{ijk}}.
\]

By specifying a distribution for the cell counts, we can estimate cell probabilities by maximizing the log-likelihood function which is up to a constant (only depending on \( \{n_{ijk}\} \)) the function \( \ell \) defined as follows:

\[
\ell = \sum_{i,j,k} n_{ijk} \log (p_{ijk}). \tag{1.1.1}
\]

It can easily be shown that the argument which maximizes (1.1.1) is the vector of sample proportions, meaning that \( p_{ijk} \) is estimated by \( \hat{p}_{ijk} = n_{ijk}/n \).

If we want to model cell probabilities in a contingency table in terms of associations among variables, generally a log-linear approach is used.

### 1.2 Log-Linear Model

Suppose there is a multinomial sample of size \( n \) over the \( IJK = N \) cells of a contingency table. The probabilities \( p_{ijk} \) for that multinomial distribution form the joint distribution. Any relationship between the three variables is fully defined by the joint distribution. For example if the three variables are independent, then it holds that \( p_{ijk} = p_{i++} p_{+j+} p_{++k} \) for all \( i, j \) and \( k \). On a logarithmic scale, independence is translated to:

\[
\log (p_{ijk}) = \log (p_{i++}) + \log (p_{+j+}) + \log (p_{++k}), \tag{1.2.2}
\]

meaning that the logarithm of the probability is an additive function of effects defined by the corresponding categories of the variables \( X, Y \) and \( Z \).
1.2.1 Independence Model

By adopting the notation of Agresti (1990), expression (1.2.2) is equivalent to
\[
\log(p_{ijk}) = \mu + \lambda^X_i + \lambda^Y_j + \lambda^Z_k,
\]
where
\[
\mu = \sum_h \log(p_{h++})/I + \sum_h \log(p_{+h+})/J + \sum_h \log(p_{++h})/K,
\]
\[
\lambda^X_i = \log(p_{i++}) - \left(\sum_h \log(p_{h++})\right)/I,
\]
\[
\lambda^Y_j = \log(p_{+j+}) - \left(\sum_h \log(p_{+h+})\right)/J,
\]
\[
\lambda^Z_k = \log(p_{++k}) - \left(\sum_h \log(p_{++h})\right)/K.
\]

The parameters \(\{\lambda^X_i\}, \{\lambda^Y_j\}\) and \(\{\lambda^Z_k\}\) satisfy
\[
\sum_i \lambda^X_i = \sum_j \lambda^Y_j = \sum_k \lambda^Z_k = 0.
\]

These sum-to-zero constraints are a popular way to make parameters identifiable. But other parameter definitions are possible as well. The parameter \(\mu\) can be viewed as a normalizing constant which ensures that the cell probabilities add up to one.

1.2.2 Saturated Model

Suppose all cell probabilities are positive and \(\eta_{ijk} = \log(p_{ijk})\). Further let
\[
\eta_{ij} = \frac{\sum_k \eta_{ijk}}{K}, \eta_{i.} = \frac{\sum_j \eta_{ijk}}{J}, \eta_{.j} = \frac{\sum_i \eta_{ijk}}{I}.
\]
1.2. Log-Linear Model

\[ \mu = \eta \ldots \]
\[ \lambda_i^X = \eta_{i.} - \eta \ldots, \quad \lambda_i^Y = \eta_{.j} - \eta \ldots, \quad \lambda_i^Z = \eta_{..k} - \eta \ldots \]
\[ \lambda_{ij}^{XY} = \eta_{ij} - \eta_{i.} - \eta_{.j} + \eta \ldots \]
\[ \lambda_{ik}^{XY} = \eta_{i.k} - \eta_{i.} - \eta_{..k} + \eta \ldots \]
\[ \lambda_{jk}^{XZ} = \eta_{.jk} - \eta_{.j} - \eta_{..k} + \eta \ldots \]
\[ \lambda_{ijk}^{XYZ} = \eta_{ijk} - \eta_{i.} - \eta_{i.k} - \eta_{.jk} + \eta_{..k} + \eta_{.j} + \eta_{..k} - \eta \ldots \]

The sum of the parameters for any index equals zero. That is:

\[ \sum_i \lambda_i^X = \sum_j \lambda_j^Y = \ldots = \sum_i \lambda_{ij}^{XY} = \ldots = \sum_k \lambda_{ijk}^{XYZ} = 0. \] (1.2.3)

The saturated log-linear model can then be written:

\[ \log(p_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}. \] (1.2.4)

The parameters with single subscript are called main effects, the parameters with double subscript are first order (or two-factor) interactions and the parameter with a triple subscript is a three-factor interaction.

1.2.3 Correspondence to Conditional Independence

A general log-linear model sets certain parameters equal to zero. For example \( \lambda_{ij}^{YZ} = 0 \) and \( \lambda_{ijk}^{XY} = 0 \) \( \forall i, j, k \). The corresponding log-linear model is

\[ \log(p_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}. \] (1.2.5)
If we fix a value \( k \) for \( Z \), then the joint probability of \( X \) and \( Y \) given \( Z = k \) has the following form:

\[
\log (p_{ij|Z=k}) = \log \left( \frac{p_{ijk}}{p_{++k}} \right) = \log \left( \frac{\exp (\mu + \lambda_i^X + \lambda_j^Y + \lambda_{ik}^XZ + \lambda_{jk}^YZ)}{\sum_{i,j} \exp (\mu + \lambda_i^X + \lambda_j^Y + \lambda_{ik}^XZ + \lambda_{jk}^YZ)} \right) (1.2.6)
\]

We define new interaction coefficients \( \tilde{\mu}, \tilde{\lambda}_i^X, \tilde{\lambda}_j^Y \):

\[
\tilde{\mu} = -\log \sum_{i,j} \exp (\lambda_i^X + \lambda_j^Y + \lambda_{ik}^XZ + \lambda_{jk}^YZ)
\]

\[
\tilde{\lambda}_i^X = \lambda_i^X + \lambda_{ik}^XZ
\]

\[
\tilde{\lambda}_j^Y = \lambda_j^Y + \lambda_{ij}^YZ
\]

The two main effects \( \tilde{\lambda}_i^X \) and \( \tilde{\lambda}_j^Y \) still satisfy the identifiability constraints (1.2.3) and \( \tilde{\mu} \) ensures that the conditional cell probabilities add up to one. Expression (1.2.7) can then be written by

\[
\log (p_{ij|Z=k}) = \tilde{\mu} + \tilde{\lambda}_i^X + \tilde{\lambda}_j^Y,
\]

which is the 2-dimensional analogue to the independence model explained in Section 1.2.1. This means that for a fixed value of \( Z \), the variables \( X \) and \( Y \) are independent. We say that \( X \) is conditionally independent of \( Y \) given \( Z \) and the terminology is \( X \perp \perp Y | Z \).

### 1.2.4 Hierarchical Log-Linear Models

Hierarchical log-linear models are models such that whenever a higher-order interaction is included, lower-order interactions composed of these variables must also be included. For example model (1.2.5) is a hierarchical model. The first order interaction \( \lambda_{ik}^{XZ} \) forces the main effects \( \lambda_i^X \)
and $\lambda^Z_k$ to be contained in the model as well. In hierarchical models, a symbol that lists the highest-order terms for each variable is assigned. These are called generators of the model. For example the symbol for the saturated model (1.2.4) is $(XYZ)$ or for the hierarchical model (1.2.5) it is $(XZ, YZ)$. Nonhierarchical models are sensible in very few applications. Assume for example that the saturated model (1.2.4) has $\lambda^X_i = 0$, whereas all other terms remain. This implies that there are interactions between all variables, but still, the sum of the log probabilities over all cells with a fixed $i$ is the same for all $i$. But this somewhat special constraint depends on the chosen identifiability constraints (1.2.3). With different identifiability constraints, different constraints for the sum of log probabilities are imposed. Therefore zero terms in a non hierarchical model can only be interpreted together with a specific parameterization. Zero terms in a hierarchical model on the other hand do have a direct interpretation in terms of (conditional) independence.

1.3 Graphical Models

A graph can be used to graphically display the conditional independence structure for a joint distributions of random variables. Each node in the
graph represent a random variable and there is an edge between two variables if there is an interaction between these two variables. Consider for example the hierarchical log-linear model (1.2.5). The graph associated with this model builds up by linking all variables appearing together in a common generator (highest-order interactions). The generators for this model are \((XZ, YZ)\) and by linking the corresponding nodes the graph displayed in Figure 1.1 builds up. This graph is called interaction graph of the hierarchical log-linear model. We have shown in Section 1.2.3 that the hierarchical log-linear model (1.2.5) yields the conditional independencies \(X \perp \perp Y | Z\). In the general case, the Hammersley and Clifford theorem (see Lauritzen (1996)) ensures that the interaction graph of a hierarchial model can be used to read off conditional independences implied by the model: Whenever two variables are not linked they are conditionally independent given all other variables. In fact it can be shown that the implication is even stronger, knowing the concept of separation, which will be introduced in Chapter 3.2.2. But for the moment it is sufficient to know that the interaction graph of a log-linear model enables to directly read off the (conditional) independencies implied by the model from the graph.

1.4 Model Selection in Log-Linear Models

1.4.1 Parameter Estimation with MLE

MLE is the maximization of the log-likelihood or which is equivalent the maximization of \(\ell\) (1.1.1) with \(p_{ijk}\) given by the log-linear model under the constraint that the cell probabilities add up to 1. For model (1.2.5) \(\ell\) is:

\[
\ell = n\mu + \sum_{i} n_{i++}\lambda_i^X + \sum_{j} n_{j++}\lambda_j^X + \sum_{k} n_{++k}\lambda_k^Z + \sum_{i,k} n_{i+k}\lambda_{ik}^{XZ} \\
+ \sum_{j,k} n_{j+k}\lambda_{jk}^{YZ} - n \sum_{i,j,k} \exp \left( \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \right)
\]
We see here on the basis of an example that the minimal sufficient statistics for fitting log-linear models with MLE are the marginal cell counts corresponding to the maximal interaction terms. In our example this is \(\{n_{i+k}, i \in \{1, \ldots, I\}, k \in \{1, \ldots, K\}\}\). We obtain the likelihood equations by differentiating with respect to a parameter and setting the results equal to zero:

\[
\frac{\partial \ell}{\partial \mu} = n - n \sum_i \sum_j \sum_k \exp(\mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ})
\]

\[
= n - n \sum_i \sum_j \sum_k \hat{p}_{ijk},
\]

which means that the fitted cell probabilities have to add up to 1. Next,

\[
\frac{\partial \ell}{\partial \lambda_i^X} = n_{i++} - n \sum_j \sum_k \hat{p}_{ijk} = n_{i++} - n\hat{p}_{i++},
\]

and similar for \(\lambda_j^Y\) and \(\lambda_k^Z\). The derivatives for the first order interactions are

\[
\frac{\partial \ell}{\partial \lambda_{ik}^{XZ}} = n_{i+k} - n\hat{p}_{i+k} \quad \text{and} \quad \frac{\partial \ell}{\partial \lambda_{jk}^{YZ}} = n_{+jk} - n\hat{p}_{+jk}.
\]

Since (1.4.9) imply (1.4.7) and (1.4.8), these two equations determine the ML estimates. We see that the likelihood equations match the minimal sufficient statistics to their expected values. Birch (1963) showed that this is the case for a general log-linear model.

### 1.4.2 Problem with Zero Cell Counts

The existence of the MLE is not guaranteed in the case of empty cells. In Habermann (1973) and Habermann (1974) this existence is studied. To illustrate the problem, consider the saturated model for three variables. As proven by Birch (1963), the ML estimate fulfills \(\hat{p}_{ijk} = n_{ijk}/n\) and in case of any empty cell, the log of that estimated probability no longer exists.
Table 1.1: Data for which ML estimates do not exist for the model \((XY,YZ,YZ)\). Cells containing \(*\) can contain any positive number.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>X</td>
<td>1</td>
<td>*</td>
</tr>
</tbody>
</table>

For unsaturated models the situation is a bit more complex. The ML estimate exists when all cell counts are positive and does not exist when any cell count equals zero in the sufficient marginal tables. But even empty cells whose sufficient marginals are positive are problematic as an example taken from Agresti (1990) shows, where the MLE does not exist (see Table 1.1).

### 1.4.3 Testing

We assess goodness of fit by comparing the fitted probabilities to the observed frequencies in terms of the likelihood. Chi-squared statistics are used to test the hypothesis that the observed frequencies satisfy a given model. For our model with 3 variables the likelihood-ratio statistic is

\[
G^2 = 2 \sum_i \sum_j \sum_k n_{ijk} \log \left( \frac{n_{ijk}}{\hat{p}_{ijk}} \right),
\]

which is asymptotically Chi-squared distributed. The Pearson statistic is

\[
X^2 = \sum_i \sum_j \sum_k \frac{(n_{ijk}/n - \hat{p}_{ijk})^2}{\hat{p}_{ijk}},
\]

which also has an asymptotic chi-squared distribution. For both statistics the degrees of freedom (df) equal the difference in dimension be-
tween the full model and the tested model. In general, if two nested models $M_1$ and $M_2$ are compared where $M_2$ is a special case of $M_1$, the likelihood ratio statistic is $G^2(M_2) - G^2(M_1)$, which is chi-squared distributed with degrees of freedom equal to the difference in dimension between $M_1$ and $M_2$.

Fitting all possible models becomes computationally impossible when the dimensionality exceeds a certain number. Stepwise procedures are a sensible choice to handle this problem. At each stage, the term with maximal improvement in fit as far as the $p$-value for the $G^2$ statistic is concerned is selected.

However, the assumption of a chi-squared distribution only holds for $n \to \infty$ and fixed value of $N$. According to Agresti (1990) it was suggested by Cochran (1954) that the chi-squared approximation for $X^2$ is permissible if a minimum expected value for the cell counts of 1 holds as long as no more than about 20% of the cells have expected values below 5. As shown in Koehler and Larntz (1980) the chi-squared approximation for $G^2$ is poor if $n/N < 5$ and that the corresponding test gives highly conservative results, meaning the reported $p$-values are larger than the true ones. For both $X^2$ and $G^2$, Koehler and Larntz (1980) showed that for sparse tables with both small and moderately large expected frequencies, the approximation is poor.

In high-dimensional contingency tables these problems naturally occur because of the large number of cells. Alternative approaches to handle sparse high-dimensional contingency tables are presented in this thesis.

1.5 Thesis Overview

In Chapter 2, a new model selection approach for log-linear model estimation in sparse contingency tables is presented. This is basically the
work which is published in Dahinden et al. (2007).

In Chapter 3, an approach to handle high-dimensional log-linear models is presented. The results of this chapter are currently available as research report (Dahinden and Bühlmann, 2008).

In Chapter 4, an application in molecular biology is studied, partially using techniques described in Chapter 2 and 3. This work has been submitted to Cancer Research.
Chapter 2

Penalized Likelihood for Sparse Contingency Tables with an Application to Full-Length cDNA Libraries

The joint analysis of several categorical variables is a common task in many areas of biology, and is becoming central to systems biology investigations whose goal is to identify potentially complex interaction among variables belonging to a network. Interactions of arbitrary complexity are traditionally modeled in statistics by log-linear models. It is challenging to extend these to the high dimensional and potentially sparse data arising in computational biology. An important example, which provides the motivation for this chapter, is the analysis of so-called
full-length cDNA libraries of alternatively spliced genes, where we investi-
gate relationships among the presence of various exons in transcript spe-
cies. We develop methods to perform model selection and parameter esti-
mation in log-linear models for the analysis of sparse contingency tables, to study the interaction of two or more factors. Maximum Likelihood estimation of log-linear model coefficients might not be appropriate because of the presence of zeros in the table’s cells, and new methods are required. We propose a computationally efficient $\ell_1$- penalization approach extending the Lasso algorithm to this context, and compare it to other procedures in a simulation study. We then illustrate these algorithms on contingency tables arising from full-length cDNA libraries.

We propose regularization methods that can be used successfully to detect complex interaction patterns among categorical variables in a broad range of biological problems involving categorical variables.

### 2.1 Introduction

One of the most striking discoveries of the genomic era is the unexpectedly small number of genes in the human genome. This amount has decreased from more than 100000 (Liang et al., 2000) to an estimated number of roughly between 20000 and 25000, tens of thousands less than initially expected and essentially the same number as found in phenotypically much simpler organisms (International Human Genome Sequencing Consortium, 2004; Southan, 2004). A question of overriding biological significance is, how complex phenotypes of higher organisms arise from limited genomes. Part of the explanation may be that many genes undergo a process called alternative RNA splicing, which can generate many distinct proteins from a single gene.

RNA splicing is a post-transcriptional process that occurs prior to mRNA translation. After the gene has been transcribed into a pre-
messenger RNA (pre-mRNA), it consists of intronic regions destined to be removed during pre-mRNA processing (RNA splicing), as well as exonic sequences that are retained within the mature mRNA. After transcription occurs the actual splicing process, where it is decided which exons are retained in the mature message and which are targets for removal. In general, exons and introns are retained and deleted in different combinations to create a diverse array of mRNAs from a common coding sequence. This process is known as alternative RNA splicing. Depending on the source, the percentage of alternatively spliced genes lies between 35% and 60% (Mironov et al., 1999; Brett et al., 2000; International Human Genome Sequencing Consortium, 2001; Brett et al., 2002; The FANTOM Consortium, 2005; Zavolan et al., 2003; Imanishi et al., 2004). By screening many full-length cDNAs it is possible to record the complete cDNA from a mature RNA for the same gene again and again and a full-length cDNA library, also known as single-gene library (SGL), builds up. The library contains detailed information about how specific exon combinations go together. This information is directly related to the functional regions of the proteins as they are grouped in domains which in many cases correspond to a single exon which encodes these domains. For example a transcription factor consists of a DNA binding domain and a regulatory domain. Thus the alteration of the exon structure corresponds to an alteration in the function of this particular domain. The central premise is that a dependency in the domains points to a functional association. If domains interact functionally then their splicing should be co-regulated. And this co-regulation has direct biological significance because it shows us which variable components also interact in the expressed protein. Because the polypeptide is intricately folded and tightly packed, segments that are separated by dozens of introns in the primary transcript may encode domains that interact functionally within the protein. These domains need not be structural neighbors even in the folded protein, but may interact through electrical or van der Waals forces, effects of global conformational changes, or even associations with other proteins. Because of these intricacies,
there are no inherent distance restrictions, or limits on the number of interacting sites, and separate domains may combine their functional effect in unpredictable ways.

Due to the large number of potential combinations in highly alternatively spliced genes, any library will only comprise a small portion of the total theoretically possible inventory of combinations. Statistically, this leads to sparse contingency tables in which dimensions represent exons and cells represent variants. The investigation of interactions among categorical variables where not all possible combinations are observed, means addressing a model selection problem that is challenging both inferentially and computationally.

As far as alternative splicing is concerned, there is an important reason to determine this interaction structure: searching for intrapeptide interactions in functional assays is a very difficult, open-ended problem, where statistical analysis of the splicing interaction structure in the transcriptome can simplify this task enormously by identifying the sets of interacting domains. And as more investigators become interested in this type of information, and large-scale single-gene libraries become available, there is a strong need for reliable statistical methods for analyzing the resulting datasets.

We develop different statistical methods to analyse sparse contingency tables in order to determine the underlying interaction pattern and we use graphical models to visualize these patterns. The methods are compared in a simulation study and illustrated on full-length cDNA libraries.
2.2 Terminology and Notation

2.2.1 General Introduction to Contingency Tables and Log-Linear Models

In this section we provide general definitions and notations.

Assume we have $q$ categorical random variables or factors, $C = \{C_1, \ldots, C_q\}$, where each $C_j$ can take on a finite number $g_j$ of possible values, called levels. The vector $(c_1, \ldots, c_q)$ represents a particular combination of levels of the joint random variable $C = \{C_1, \ldots, C_q\}$. The total cardinality of $C$ is $m = \prod_{j=1}^{q} g_j$, which corresponds to the $m$ different combinations of levels ($m = 2^q$ when all $C_j$ are dichotomous, as in our splicing example).

We simplify the notation by mapping each configuration of $C$ to a unique natural number $i \in \{1, \ldots, m\}$ with a (bijective) function $f$:

$$f : (c_1, \ldots, c_q) \leftrightarrow i \in \{1, \ldots, m\},$$

so we may write $c_i = (c_1, \ldots, c_q)$. For $n$ observations of $C$, the corresponding $q$-way contingency table has $m$ cells, each listing the frequency of a particular configuration $c_i$:

$$n_{c_1, \ldots, c_q} = n_i, \quad \sum_{i=1}^{m} n_i = n.$$ 

A general introduction to contingency tables can be found in Everitt (1992). If the observations are independent, with $p_i$ the probability of sampling configuration $c_i$, the distribution of the cell counts $(n_1, \ldots, n_q)$ is multinomial with probability $\mathbf{p} = (p_1, \ldots, p_q)$.

In the splicing example, we may consider the $C_j$ as dichotomous random variables representing $q$ sites of alternative splicing, each with two levels, denoted by $c_j \in \{1, -1\}$, corresponding to the presence or absence of exon $j$ in a transcript. The contingency table therefore has
$m = 2^q$ cells, with each cell represented by the $q$-dimensional binary vector $c_i = (c_1, \ldots, c_q)$. A log-linear model for the cell probabilities can be written the following way:

$$\log p_i = \beta_{\emptyset} + \sum_{l \in \{1, \ldots, q\}} \beta_l c_l + \sum_{j, k}^{j < k \in \{1, \ldots, q\}} \beta_{jk} c_j c_k + \ldots + \beta_{12\ldots q} c_1 c_2 \cdots c_q.$$  

(2.2.1)

A general log-linear model represents $p$ as:

$$\log (p) = X\beta,$$  

(2.2.2)

where $\beta$ is a vector of unknown coefficients and $X$ a suitable design matrix as indicated below.

Let us assume that the cell probabilities are expressed in the following way:

$$\log p_{c_1, \ldots, c_q} = \delta_{\emptyset} + \delta_{c_1}^{C_1} + \ldots + \delta_{c_q}^{C_q} + \delta_{c_1, c_2}^{C_1, C_2} + \ldots + \delta_{c_1, \ldots, c_q}^{C_1, \ldots, C_q},$$  

(2.2.3)

where $\delta_{\emptyset}$ is the global mean, $\delta_{c_1}^{C_1}$ is the main effect of the first variable and only depends on the distribution of $C_1$. Similarly $\delta_{c_1, c_2}^{C_1, C_2}$ is the first order interaction between the first two variables and its value only depends on the joint distribution of these two variables.

We now look for a suitable parameterization $\tilde{X}^{C_i}$ of the vector spaces spanned by the main effects $\delta^{C_i}$, a parameterization $\tilde{X}^{C_i, C_j}$ for the vector spaces spanned by the first order interactions $\delta^{C_i, C_j}$ and so on. To ensure identifiability, we impose constraints on these matrices and denote the resulting matrices by $X^{C_i}$, $X^{C_i, C_j}$ and so on. The design matrix $X$ finally consists of these submatrices. The constitution of the design matrix $X$ for factors with two levels can directly be derived from (2.2.1). The derivation of the design matrix $X$ from (2.2.3) in the case of more than two levels per factor is basically an analysis of variance (ANOVA) parameterization with poly-contrasts.
One possible parameterization of the vector space spanned e.g. by \( \delta^{C_1} \), \( \text{span}(\delta^{C_1}) = \tilde{X}^{C_1} \in \mathbb{R}^{m \times |C_1|} \), is given as follows: the \( j \)th column takes on the value 1 if the variable \( c_1 \) of the corresponding cell takes on level \( j \) and 0 otherwise. Similarly for \( \tilde{X}^{C_1,C_2} \in \mathbb{R}^{m \times (|C_1| \cdot |C_2|)} \), there is a 1 in the first column, if the corresponding cell has \( c_1 \) as well as \( c_2 \) taking on its first level. Of course the resulting matrix is highly overparameterized and to ensure identifiability, we have to impose constraints. By definition the vector space spanned by the constant vector \( \delta_\emptyset \) is a subspace of the column space of \( \tilde{X}^{C_1} \), analogously the column space spanned by \( \tilde{X}^{C_1} \) and \( \tilde{X}^{C_2} \) is a subspace of the column space spanned by \( \tilde{X}^{C_1,C_2} \) and the same holds for higher order interactions: The vector space spanned by an interaction of order \( k \) is always a superspace of all vectorspaces spanned by all corresponding interactions of order \( k - 1 \). We then finally define the matrix \( X^{C_{j_1},...,C_{j_k}} \) by the orthogonal complement of the vector space spanned by \( \{ \tilde{X}^{C_{j_2},...,C_{j_k}}, \tilde{X}^{C_{j_1},C_{j_3},...,C_{j_k}}, \tilde{X}^{C_{j_1},...,C_{j_{k-1}}} \} \) in \( \tilde{X}^{C_{j_1},...,C_{j_k}} \). The dimensionality of the resulting submatrix is \( \mathbb{R}^{m \times (|C_{j_1}| - 1) \cdots (|C_{j_k}| - 1)} \). We can easily prove that the resulting design matrix is an orthogonal matrix. We generally normalize the column vectors to length \( \sqrt{m} \), such that the first column is a column of 1’s.

The log-linear interaction model can then be represented in matrix notation (2.2.2).

Sometimes we may assume a smaller model without some of the interaction terms. It is of the form as in (2.2.2) with some columns removed from the design matrix \( X \). We denote matrices of the form \( X^{C_{j_1},...,C_{j_k}} \) by \( X_a \), with \( a = \{ C_{j_1}, \ldots, C_{j_k} \} \subseteq C \). The corresponding subvector of \( \beta \) is denoted by \( \beta_a \).
2.2.2 Graphical Models

A powerful way for visualizing conditional dependencies among variables is given by a graph. A graph \( G = (V, E) \) consists of a finite set \( V \) of vertices and a finite set \( E \) of edges between these vertices. In our context, the vertices correspond to the different discrete random variables. We form the so-called conditional independence graph by connecting all pairs of vertices that appear in the same generator, that is the maximal terms \( a \subseteq C \) which are present in the model. To translate a vector \( \beta \) into a graphical model we look for \( \beta_a \neq 0 \) with \( \beta_b = 0 \) \( \forall a \subset b \) (where \( b \) is a strict super-set of \( a \) and \(|a| > 1\)) and we draw edges between all vertices corresponding to \( a \). From this graph we can directly read off all marginal and conditional independences by the global Markov property for undirected graphs which states: if two sets of variables \( a \) and \( b \) are separated by a third set of variables \( c \) then \( a \) and \( b \) are conditionally independent given \( c \) \( (a \perp \perp b | c) \), where for three subsets \( a, b \) and \( c \) of \( V \), we say \( c \) separates \( a \) and \( b \) if all paths from \( a \) to \( b \) intersect \( c \). For details, see Lauritzen (1996).

2.2.3 Model Selection - Non-Hierarchical Versus Hierarchical Models

Hierarchical models are a subclass of models such that if an interaction term \( \beta_a \) is zero, then all higher order interaction terms \( \beta_b \) for \( b \supseteq a \) are also zero. If we consider the example above with 2 levels, this means for example that if the first order interaction coefficient \( \beta_{ij} = 0 \) then all higher order interaction coefficients including \( i \) and \( j \) are also zero, i.e. \( \beta_{ijk} = 0, \forall k \). While it is possible that the true underlying interaction model may not be hierarchical from a biological standpoint, a difficulty in the use of non-hierarchical models arises from the fact that they are not invariant under reparametrization. We have chosen the design matrix \( X \) with some constraints to ensure identifiability, and we
used a specific, namely an orthonormal basis. In terms of ANOVA, this choice is equivalent to choosing a poly-contrast. We could have imposed different constraints or have chosen a different basis, and this would have resulted in a different design matrix \( X \) or in terms of ANOVA, a different choice of contrast. Suppose we have found an interaction vector \( \beta \) for one parameterization of the log-linear model and that this vector corresponds to a non-hierarchical model, meaning there is at least one lower order interaction term \( \beta_a \) equal to zero, while \( \beta_b \neq 0 \) for at least one \( b \supseteq a \). If we reparametrize the model, using a different design matrix, the coefficient for the model term \( a \) may no longer be zero. On the other hand, by reparametrizing a hierarchical model, all zero terms remain zero after reparametrization. Therefore, hierarchicity is preserved after reparametrization while non-hierarchicity depends on the parameterization. This is a distinct advantage of working within the hierarchical class. In a hierarchical model, all zero coefficients can directly be interpreted in terms of conditional independence, while this is not true for non-hierarchical models.

### 2.3 Algorithms

In this section we introduce different model selection strategies for log-linear models. We first develop an \( \ell_1 \)-regularization model selection approach, which is then expanded to the new so-called level-\( \ell_1 \)-regularization approach. In addition, different Bayesian model selection strategies are introduced.

#### 2.3.1 \( \ell_1 \)-Regularized Model Selection

The Lasso, originally proposed by Tibshirani (1996) for linear regression, performs regularized parameter estimation and variable selection
Chapter 2. Penalized Log-Linear Model Selection

at the same time. The Lasso estimate is defined as follows:

\[ \hat{\beta}^\lambda = \arg \min_\beta \left[ \sum_i (Y - X\beta)_i^2 + \lambda \sum_j |\beta_j| \right], \]

where \( Y = (Y_1, \ldots, Y_n) \) is the response vector. This can also be viewed as a penalized Maximum Likelihood estimator, as \( \sum_i (Y - X\beta)_i^2 \) is proportional to the negative log-likelihood function for Gaussian linear regression. While the MLE for the general regression model is no longer uniquely defined and very poor in the case of more variables than observations, the Lasso estimator is still reasonable as long as \( \lambda > 0 \). For our analysis, we have a similar problem, namely that the MLE does not exist in case of zero counts in the contingency table: a detailed description of the existence of the MLE in general log-linear interaction models is given in Christensen (1991). Inspired by the Lasso, we estimate our parameter vector \( \beta \) by the following expression:

\[ \hat{\beta}^\lambda = \arg \min_\beta \left[ -\frac{1}{n} l(\beta) + \lambda \sum_j |\beta_j| \right], \quad (2.3.4) \]

where \( l(\beta) = \sum_{i=1}^m n_i (X\beta)_i = \log P\beta[n] + c \). The function \( l(\beta) \) is up to an additive constant \( c \), which does not depend on \( \beta \), the log-likelihood function. Therefore in the following we will denote the log-likelihood function by \( l(\beta) \). This minimization has to be calculated under the additional constraint that the cell probabilities add up to 1:

\[ \sum_{i=1}^m \exp \{(X\beta)_i\} = 1. \quad (2.3.5) \]

A problem of the optimization (2.3.4) is that the solution is no longer independent of the choice of the orthogonal subspaces \( X_a \). That is, if any set of orthogonal columns \( X_a \) of \( X \) is reparametrized by a different orthogonal set, we get a different solution. To avoid this undesirable outcome we use a penalty that is intermediate between the \( \ell_1 \)- and the \( \ell_2 \)-penalty. This penalty, called group-\( \ell_1 \)-penalty, has the following
2.3. Algorithms

form:

\[ \sum_{a \subseteq C} \| \beta_a \|_{\ell_2}, \quad \text{where} \quad \| \beta_a \|_{\ell_2}^2 = \sum_j (\beta_{a,j})^2 \]

Originally, this has been proposed by Yuan and Lin (2006) for the linear regression problem with factor variables. The estimator of \( \beta \) then becomes

\[
\hat{\beta}^\lambda = \arg \min_{\beta} \left[ -\frac{1}{n} l(\beta) + \lambda \sum_{a \subseteq C \atop a \neq \emptyset} \| \beta_a \|_{\ell_2} \right],
\]

subject to the constraint in (2.3.5). By imposing a penalty function on the coefficients of the log-linear interaction terms, overfitting as it might occur by using MLE is reduced. Furthermore, the \( \ell_1 \)-penalty encourages sparse solutions for the single components of \( \beta \), the group \( \ell_1 \)-penalty encourages sparsity at the interaction level, meaning that the vector \( \beta_a \), which corresponds to the interaction term \( a \) is either present or absent in the model as a whole. In case of factors with only 2 levels, the group \( \ell_1 \)-penalty and the \( \ell_1 \)-penalty are equivalent.

For both the \( \ell_1 \)-, and the group \( \ell_1 \)-regularization, the parameter \( \lambda \) can be assessed by cross-validation: we divide the individual counts into a number of equal parts and in turn leave out one part for the rest to form a training contingency table with cell counts \( n_{\text{train}} \). The solution for an array of values for \( \lambda \), the so-called solution path, is calculated according to an algorithm described in Section 2.4. The corresponding vectors of cell probabilities are denoted by \( p(\hat{\beta}^\lambda) \). We then use the remainder of the cell counts \( n_{\text{test}} \) to calculate the predictive negative log-likelihood score

\[
-\sum_{i=1}^m n_{\text{test},i} \cdot \log \left( p_i(\hat{\beta}^\lambda) \right) \sum_{i=1}^m n_{\text{test},i},
\]

which is proportional to the out-of-sample negative log-likelihood. This score is on the same scale when varying the number of observations and may therefore be used to compare contingency tables of the same
dimension but with different numbers of cell entries. The parameter \( \lambda \) is chosen as the value which minimizes the cross-validated score in (2.3.7). We use a ten-fold cross-validation in our example.

The resulting model does not necessarily have to be hierarchical and if we consider the hierarchical model induced by this procedure, it might happen that the final model is large for example if a single high order interaction is estimated to be active. To address this, we set up an algorithm described in the next Section.

### 2.3.2 Level-\( \ell_1 \)-Regularized Model Selection

In order to prevent the procedure from choosing single high-order interactions, we alter the \( \ell_1 \)-regularized algorithm described in the previous section: we do not exclusively apply it to the fully saturated model but also to submodels with lower order interactions. Precisely, a model is fitted with main effects only, and the predictive negative log-likelihood score (2.3.7) is calculated for the best main effects model (level 1). The same is done for the model including all main effects and first order interactions (level 2). Proceeding accordingly, we get \(|C|\) log-likelihood scores corresponding to the \(|C|\) levels. The level with minimal score (2.3.7) is then chosen (and within this selected level, we have an \( \ell_1 \)-regularized estimate).

With this procedure the tendency of including a single high-order interaction while most of its lower order interactions are absent is decreased, and the inclusion is only forced if the predictive negative log-likelihood score strongly speaks in favour of the inclusion. Therefore we tend to select sparser models which can be better hierarchized and interpreted in terms of conditional independence, in contrast to the ordinary \( \ell_1 \)-model selection procedure.
2.3. Algorithms

2.3.3 Non-hierarchical Bayesian Model Selection

The Bayesian approach we choose is essentially the same as chosen by Kuo and Mallick (1998) and closely related to what was proposed by Ntzoufras et al. (2000); George and McCulloch (1993); Geweke (1994). We use a Markov chain Monte Carlo algorithm based on Stochastic Search Variable Selection (SSVS): SSVS is a procedure proposed by George and McCulloch (1993) to perform variable selection in the standard linear regression model. We adapt this procedure to log-linear models. But instead of assuming a normal mixture model for the coefficients of interest as in SSVS, we follow an approach proposed by Geweke (1994), and assume the coefficients to be a mixture of a point mass at zero and a normal distribution. Following the notation introduced earlier, the complete model can be described as follows:

\[ \mathbf{n} \sim \text{Multinom}(\mathbf{p}) \text{ with } \log(\mathbf{p}) = \mathbf{X}\mathbf{\beta}, \]
\[ \beta_a | \gamma_a \sim (1 - \gamma_a)\mathbf{I}_0 + \gamma_a\mathcal{N}(0, \sigma_a^2\mathbf{1}_d) \text{ independent for all } a \subseteq C, \]
\[ \gamma_a \sim \text{Ber}(\text{pr}_{\gamma_a}) \text{ independent for all } a \subseteq C, \]
\[ \sigma_a^2 \sim \Gamma^{-1}(l, u) \text{ independent for all } a \subseteq C, \]

(2.3.8)

where \( I_0 \) is a point mass at zero and \( \gamma_a \) is a Bernoulli variable with probability parameter \( \text{pr}_{\gamma_a} \) reflecting prior belief that the corresponding interaction term \( \beta_a \) is not equal to zero. The parameters \( \sigma_a^2 \) follow an inverse gamma distribution with parameters \( l \) and \( u \). In our simulation study, we also considered fixed values for \( \sigma_a^2 \). The choice of the prior parameter \( l, u \) and \( \text{pr}_{\gamma_a} \) is discussed in Section 2.4.1. In the absence of strong prior belief, it is reasonable to assume that all \( \sigma_a^2 \) are identically distributed. By imposing prior distributions on the log-linear parameters \( \beta_a \), it would be possible to incorporate further prior knowledge in the form of existence of correlation or signs of correlation between the different criteria \( C \). One way is to use a prior with expectation different from zero for the corresponding log-linear term (\( \mathbb{E}[\beta_a | \gamma_a = 1] \neq 0 \)). See for example Dellaportas and Forster (1999) for a more detailed discussion on normal priors for the log-linear parameters \( \beta_\alpha \).
We introduce variables $\alpha_a$, where $\alpha_a \sim N(0, \sigma^2_a 1_{d_a})$ and we set $\beta_a = \alpha_a$ if $\gamma_a = 1$ and $\beta_a = 0$ if $\gamma_a = 0$ independent of the value of $\alpha_a$: $\beta_a = \alpha_a \gamma_a$ has then the desired distribution in (2.3.8). This construction is mentioned, but not implemented, in Geweke (1994).

The calculation of the posterior distribution $f(\gamma, \alpha, \sigma^2|n)$ is now required. This cannot be done directly and Monte Carlo approximations are needed, for example from Gibbs sampling. We first calculate the univariate conditional distributions of the parameters $\alpha_a$ or components of $\alpha_a$ if it is a vector:

$$f(\alpha_a|n, \gamma, \alpha_{\setminus a}, \sigma^2) \propto f(n|\gamma, \alpha) f(\alpha_a|\sigma^2_a) \propto \exp\{n \cdot (X_0 \alpha_0 + X_a \alpha_a \gamma_a)\} f(\alpha_a|\sigma^2_a).$$

Although this univariate conditional density is not of any recognized form, we can prove that it is log-concave (see Lemma 1 at the end of this section for details) and therefore sampling from it can be efficiently done using adaptive rejection sampling, as proposed by Gilks and Wild (1992). Sampling $\sigma^2_a$ is straightforward, as

$$f(\sigma^2_a|n, \gamma, \alpha, \sigma^2_{\setminus a}) = f(\sigma^2_a|\alpha_a) \propto f(\alpha_a|\sigma^2_a) f(\sigma^2_a),$$

and we can easily show that $\sigma^2_a|\alpha_a \sim \Gamma^{-1}(\alpha_a^2/2 + l, u + 1/2)$. Therefore we can sample $\sigma^2_a$ from an inverse gamma distribution. In the case where $\sigma^2_a$ is assumed to be fixed, this sampling step can be omitted. To sample from $f(\gamma_a|n, \gamma_{\setminus a}, \alpha, \sigma^2)$, we compute the conditional Bayes factor $BF$ in favour of $\gamma_a = 1$ versus $\gamma_a = 0$. The conditional posterior distribution of $\gamma_a$ is Bernoulli with $p_{\gamma_a} = \frac{BF}{1 + BF}$. Thus we can sample

$$\gamma_a \sim Ber(p_{\gamma_a}).$$

The Bayes factor $BF$ is given by

$$BF = \frac{f(n|\gamma_a = 1, \gamma_{\setminus a}, \alpha)p_{\gamma_a}}{f(n|\gamma_a = 0, \gamma_{\setminus a}, \alpha)(1 - p_{\gamma_a})}.$$
The parameters $\alpha_a$, $\sigma_a^2$ and $\gamma_a$ are updated in turn for all $a \subseteq C$. In this way we are able to efficiently sample from the full posterior $f(\alpha, \gamma, \sigma^2|n)$ and derive from it the posterior of $f(\beta, \gamma, \sigma^2|n)$. From the marginal posterior distribution $f(\gamma|n)$, we can estimate the model probabilities by the sample proportions for $\gamma$, with the most promising models corresponding to the most frequently observed $\gamma$. From $f(\beta|n, \gamma)$ we can derive the distribution for the interaction strength vector $\beta$ conditional on the model $\gamma$.

**Lemma 1.** The function $f(\alpha_a|n, \gamma, \alpha_{\setminus a})$ is log-concave for the prior distributions chosen as described in (2.3.8).

**Proof.** Without loss of generality we assume that $\alpha_a$ is univariate. The proof for the case that $\alpha_a$ is a vector is exactly the same but for a single component of $\alpha_a$. We have to prove that the function $h(\alpha_a)$ is concave for

$$h(\alpha_a) = n\alpha_\emptyset + n^t X_a \alpha_a \gamma_a - \frac{1}{2 \sigma^2} \alpha_a^2,$$

where $\alpha_\emptyset$ is the normalizing constant ensuring that all cell probabilities add up to one. This constant depends on $\alpha_a$. As the last two terms are concave it remains to be shown that $n\alpha_\emptyset(\alpha_a)$ is concave. For $\gamma_a = 0$ this term is constant and $h(\alpha_a)$ is therefore concave. For $\gamma_a = 1$, we set $X' = X \setminus \emptyset$ and $\alpha' = \alpha_{\setminus \emptyset}$, it then holds

$$h(\alpha_a) = n\alpha_\emptyset = -n \log \sum_{i=1}^{m} \exp ((X' \alpha')_i),$$

$$h'(\alpha_a) = -n \frac{X_a^t \exp (X' \alpha')}{\sum_{i=1}^{m} \exp ((X' \alpha')_i)},$$

$$h''(\alpha_a) = -n \frac{(X_a^2)^t \exp (X' \alpha') \sum_{i=1}^{m} \exp ((X' \alpha')_i) - (X_a^t \exp (X' \alpha'))^2}{(\sum_{i=1}^{m} \exp ((X' \alpha')_i))^2},$$

where $\exp (X' \alpha')$ has to be understood as the componentwise application of the exponential function and likewise for $X_a^2$. We now have to
Chapter 2. Penalized Log-Linear Model Selection

show that \( h''(\alpha_a) \) is less than zero. If we denote \( \exp(X'\alpha') \) by \( u \) and \( X_a \) with \( x \), it is sufficient to prove that

\[
\sum_{j=1}^{m} x^2_j u_j \sum_{i=1}^{m} u_i - (x'u)^2 \geq 0.
\]

The above expression is

\[
\sum_{i,j, j<j}(x^2_j u_j u_i + x^2_i u_i u_j) - (2x_i u_i x_j u_j) = \sum_{i,j, j<j}(x^2_j + x^2_i - 2x_i x_j)u_i u_j
\]

\[
= \sum_{i,j, i<j}(x_j - x_i)^2 u_i u_j,
\]

which is greater than zero, as \( u > 0 \). This proves Lemma 1. □

2.3.4 Hierarchical Bayesian Model Selection

We adapt the algorithm described above in a way that allows only moves from one hierarchical model to another, so that we never leave the class of hierarchical models. A hierarchical model is determined by its generators. The only individual model term which may be removed from a hierarchical model so that it remains hierarchical is a generating term. In addition, Edwards and Havranek (1985) defines the dual generators, which are the minimal terms that are not present in the model. The only individual model terms which may be added to the model so that it remains hierarchical are the dual generators.

We consider all hierarchical models to be equally likely and denote the set of generators and dual generators of a hierarchical model corresponding to \( \gamma \) with \( G_\gamma \). We use a Metropolis Hastings algorithm to sample from the full posterior distribution \( f(\gamma, \alpha, \sigma^2|n) \). We propose a move from one model \( \gamma^t \) to the next model \( \gamma^{t+1} \) by choosing an element \( G_{\gamma^t} \). Thus we randomly sample an element \( a \in G_{\gamma^t} \) and the corresponding \( \gamma_a \) is set to one or zero respectively. The resulting \( \gamma \) is
denoted as $\gamma^{t+1}$. The corresponding move is accepted with acceptance probability:

$$\min \left( 1, \frac{f(n|\gamma^{t+1}, \alpha^t)|G_{\gamma^t}|}{f(n|\gamma^t, \alpha^t)|G_{\gamma^{t+1}}|} \right),$$

where $|G|$ refers to the number of models included in each set of generators $G$ and this refers to the probability of proposing each model. The sampling procedure for $\alpha_a$ and $\sigma_a^2$ is performed exactly as in the non-hierarchical case described in the previous section.

### 2.4 Implementation

**Algorithm for $\ell_1$-Regularization for Factors With Two Levels**

For the regularization approaches we calculate $\hat{\beta}^\lambda$ over a large number of values of $\lambda$ in order to do some cross-validation using (2.3.7). For this purpose, an efficient algorithm is required. As one can easily verify by introducing Lagrange multipliers, finding the solution to (2.3.6) under the constraint (2.3.5) is equivalent to minimizing an unconstrained function $g(\beta)$:

$$g(\beta) = -\frac{1}{n} l(\beta) + \sum_{i=1}^m \exp(\mu_i) + \lambda \sum_{a \subseteq C, a \neq \emptyset} \|\beta_a\|_{\ell_2}, \tag{2.4.10}$$

with $\mu = X\beta$ and $\frac{1}{n} l(\beta) = \sum_i \frac{n_i}{n} (X\beta)_i$. Here, $g$ is a convex function. If each factor has two levels only, as in our application with single-gene libraries, we can set up an algorithm, which efficiently yields the estimates for a whole sequence of parameters $\lambda$. Let $A$ denote the set of active interaction terms, which means for $a \in A$ it holds that $\beta_a \neq 0$; $X_A$ is the corresponding sub-matrix of $X$, $\beta_A$ the corresponding sub-vector of $\beta$ and $g_A$ is $g$ restricted to the subspace $\beta_A$. We restrict
ourselves to the currently active set $\mathcal{A}$, where $\nabla g_{\mathcal{A}}$ and $\nabla^2 g_{\mathcal{A}}$ are well-defined:

$$
\nabla g_{\mathcal{A}}(\beta_{\mathcal{A}}, \lambda) = -X^t_{\mathcal{A}}\{\frac{n}{n} - \exp(X_{\mathcal{A}}\beta_{\mathcal{A}})\} + \lambda(0, \text{sign}(\beta_{\mathcal{A}}))^t
$$

$$
\nabla^2 g_{\mathcal{A}}(\beta_{\mathcal{A}}, \lambda) = X^t_{\mathcal{A}}\text{diag}\{\exp(X\beta)\} X_{\mathcal{A}}.
$$

The algorithm, which is an adaption of the path following algorithm proposed by Rosset (2005), is set up as follows:

1. Start with $\hat{\beta} = (-\log(m), 0, \ldots, 0)$
2. Set: $\lambda_0 = 1, \mathcal{A} = \{\emptyset\}$ and $t = 0$.
3. While ($\lambda_t > \lambda_{\text{min}}$)
   3.1 $\lambda_{t+1} = \lambda_t - \epsilon$
   3.2 $\mathcal{A} = \mathcal{A} \cup \{j \notin \mathcal{A} : |(X^t \cdot \frac{n}{n} - \exp(X\beta))_j| > \lambda_{t+1}\}$
   3.3 $\hat{\beta}$ is updated as
      $$
      \hat{\beta}_{t+1} = \hat{\beta}_t - \nabla^2 g_{\mathcal{A}}(\hat{\beta}_t, \lambda_{t+1})^{-1} \cdot \nabla g_{\mathcal{A}}(\hat{\beta}_t, \lambda_{t+1}).
      $$
   3.4 $\mathcal{A} = \mathcal{A} \setminus \{j \in \mathcal{A} : |\hat{\beta}_{t+1,j}| < \delta\}$
   3.5 $t = t + 1$

The pairs $(\hat{\beta}_t, \lambda_t)$, obtained from the algorithm above, represent the estimates from (2.3.6) under the constraint (2.3.5) for a range of penalty parameters $\lambda_t$ e.g. $(t = \epsilon, 2\epsilon, \ldots)$. The choice of the step length $\epsilon$ represents the tradeoff between computational complexity and accuracy. To increase accuracy, one can perform more than one Newton step (3.3) if the gradient starts deviating from zero. The coefficient $\delta$ is also flexible.
Typically it is chosen in the order of $\epsilon$. The lowest $\lambda$ for which one wants the solution to be calculated is denoted by $\lambda_{\text{min}}$.

Technical details concerning the algorithm can be found in the Appendix 5.A.

### 2.4.1 Prior Specification for Bayesian Methods

For the Bayesian estimation of the parameter vector, we must specify the parameters for the prior distribution of $\sigma^2_a$: $\sigma^2_a$ plays a role that is similar to that of the parameter $\lambda$ in the Lasso. The lower $\sigma^2_a$, the smaller the estimated coefficient $\hat{\beta}_a$. An empirical Bayes approach to the implementation could be to specify this parameter by cross-validation. While feasible for the $\ell_1$-regularization approaches, cross-validation becomes prohibitive for the MCMC approaches because of the computational demands. Dellaportas and Forster (1999) proposed a fixed value of 2 for all $a$ in $C$, e.g. $\sigma^2_a = \sigma^2$. Placing a normal prior with mean zero and variance two on each $\alpha_a$ means that with probability 0.95, each of these effects will increase or decrease the ratio of any two cell probabilities by a factor of no more than 10. This is a relatively vague prior, and can be appropriate when no prior information is available. However, our simulation study will illustrate that the final results can be highly sensitive to the choice of this value. To mitigate this sensitivity, we assume $\sigma^2$ to have an inverse gamma distribution with mean and variance equal to one.

In addition, for non-hierarchical model selection, we have to specify the prior distribution for $\gamma_a$. We set $\gamma_a \sim Ber(pr_{\gamma_a})$, where $pr_{\gamma_a}$ reflects prior belief that the corresponding interaction term $U_a$ is present. Without prior knowledge, we assume here that all possible models are a priori equally likely, corresponding to $pr_{\gamma_a} = 1/2$ for all $a \subseteq C$.

This prior is especially attractive when coupled with MAP estimation, as done here, because it effectively cancels out of the MAP cal-
calculation. In other situations, this prior may be less compelling. For example, it may be of interest to report posterior probabilities of properties of sets in the model space, such as marginal posteriors of the inclusion of certain coefficients or marginal posteriors of the presence of high order interactions. Then one has to evaluate carefully the mass that priors give to those sets, and one might have to reconsider the choice of the prior distributions to get reasonable posterior probabilities of these sets. In addition, as \( q \), the number of exons, increases, estimating the MAP in the model space becomes difficult and marginal posteriors of summaries such as the model size or the maximum order of interaction may be all that can be reliably estimated. In those circumstances, we suggest graphing these posteriors along with the corresponding priors probabilities, and/or to report Bayes factors.

2.5 Simulation Study

2.5.1 Data

We choose the true underlying interaction vector \( \beta \) consisting of 5 factors of 2 levels. By enumerating the factors from 1 to 5, the generators of the model are \{345, 235, 234, 135, 123, 14\}, which means that all third and fourth order interactions are absent, only five of ten second order interactions and all first order interactions are present. The corresponding coefficients of \( \beta \) are independently simulated using a normal distribution with mean zero and variance one.

Then, 250 draws from a multinominal distribution with probability vector \( p \) where \( \log (p) = X\beta \), are taken. This corresponds to a reasonable number of cDNAs in a single-gene library. This is then repeated 10 times. With our choice of \( \beta \), the resulting contingency tables are sparse. With the simulated cell counts, \( \hat{\beta} \) is estimated with different methods described in the previous sections and these methods are then
compared as follows:

### 2.5.2 Criteria

As a model selection score (MSS), we consider the fraction of correctly assigned model terms:

\[
MSS = 1 - \frac{1}{m} \sum_{i=1}^{m} |1_{\{\beta_i \neq 0\}} - 1_{\{\hat{\beta}_i \neq 0\}}|.
\]

Moreover, we consider the root mean squared error for the interaction coefficients,

\[
RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (\hat{\beta}_i - \beta_i)^2}.
\]

For assessing how much the estimation of \( \beta \) varies over multiple datasets, we calculate for every coefficient \( \hat{\beta}_i \) the estimated standard deviation \( \hat{\sigma}_i \). The means of these standard deviations are reported as

\[
SPREAD = \frac{1}{m} \sum_{i=1}^{m} \hat{\sigma}_i,
\]

a measure of variability.

To compare the different procedures for estimation of probabilities \( p = \exp(X\beta) \), we calculate the negative log-likelihood score (NLS) similar to the score in (2.3.7):

\[
NLS(\hat{\beta}) = -\sum_{i=1}^{m} p_i \cdot \log(p_i(\hat{\beta})).
\]

### 2.5.3 Results of Simulation Study

The results of the simulation study are summarized in Table 2.1. We notice that the penalty-based regularization approaches proposed in
Table 2.1: Comparison of different methods to estimate the interaction strength vector $\beta$. MSS, NLS, RMSE and SPREAD are described in Section 2.5.2. The additional methods relaxed $\ell_1$-regularization and $\ell_2$-regularization listed in the Table are explained in the Results Section.

<table>
<thead>
<tr>
<th>Method</th>
<th>MSS</th>
<th>NLS</th>
<th>RMSE</th>
<th>SPREAD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Penalty-based regularization methods:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\ell_1$-regularization</td>
<td>69.7%</td>
<td>2.20</td>
<td>0.228</td>
<td>0.144</td>
</tr>
<tr>
<td>Level-$\ell_1$-regularization</td>
<td>89.7%</td>
<td>2.22</td>
<td>0.237</td>
<td>0.179</td>
</tr>
<tr>
<td>Relaxed $\ell_1$-regularization</td>
<td>82.2%</td>
<td>2.22</td>
<td>0.233</td>
<td>0.154</td>
</tr>
<tr>
<td>$\ell_2$-regularization</td>
<td>-</td>
<td>2.20</td>
<td>0.238</td>
<td>0.130</td>
</tr>
<tr>
<td><strong>MCMC without model selection:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2 = 2$</td>
<td>-</td>
<td>2.32</td>
<td>0.747</td>
<td>0.401</td>
</tr>
<tr>
<td>$\sigma^2 = 1$</td>
<td>-</td>
<td>2.27</td>
<td>0.467</td>
<td>0.287</td>
</tr>
<tr>
<td>$\sigma^2 = 1/2$</td>
<td>-</td>
<td>2.24</td>
<td>0.294</td>
<td>0.201</td>
</tr>
<tr>
<td><strong>MCMC with model selection:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2 \sim \Gamma^{-1}(2, 3)$</td>
<td>81.5%</td>
<td>2.23</td>
<td>0.294</td>
<td>0.231</td>
</tr>
<tr>
<td>$\sigma^2 = 2$</td>
<td>76.6%</td>
<td>2.25</td>
<td>0.431</td>
<td>0.342</td>
</tr>
<tr>
<td>$\sigma^2 = 1$</td>
<td>78.4%</td>
<td>2.24</td>
<td>0.331</td>
<td>0.265</td>
</tr>
<tr>
<td>$\sigma^2 = 1/2$</td>
<td>76.6%</td>
<td>2.23</td>
<td>0.281</td>
<td>0.225</td>
</tr>
<tr>
<td><strong>MCMC with hierarchical model selection:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2 \sim \Gamma^{-1}(2, 3)$</td>
<td>84.1%</td>
<td>2.22</td>
<td>0.255</td>
<td>0.180</td>
</tr>
<tr>
<td>$\sigma^2 = 2$</td>
<td>80.6%</td>
<td>2.29</td>
<td>0.415</td>
<td>0.284</td>
</tr>
<tr>
<td>$\sigma^2 = 1$</td>
<td>83.4%</td>
<td>2.26</td>
<td>0.308</td>
<td>0.221</td>
</tr>
<tr>
<td>$\sigma^2 = 1/2$</td>
<td>83.4%</td>
<td>2.24</td>
<td>0.247</td>
<td>0.178</td>
</tr>
<tr>
<td>$\sigma^2 = 1/10$</td>
<td>86.3%</td>
<td>2.20</td>
<td>0.236</td>
<td>0.097</td>
</tr>
<tr>
<td>$\sigma^2 = 1/100$</td>
<td>69.7%</td>
<td>2.28</td>
<td>0.420</td>
<td>0.033</td>
</tr>
</tbody>
</table>

This article leads to comparable or better results than the Bayesian approaches with respect to the NLS-score, RMSE and the variation (SPREAD), though the results of Bayesian approaches vary with the
prior and the set of possible priors has not been extensively explored.

The level- $\ell_1$-regularization and the relaxed $\ell_1$-regularization (see below) are both competitive and can be better than MCMC for model selection.

The results of the MCMC procedures are sensitive to the choice of the prior value or the prior distribution for $\sigma^2$. A flat prior for $\alpha_a$ ($\sigma^2 = 2$) results in worse performance than that of a prior that shrinks the coefficients more towards zero ($\sigma^2 = 1/2$). This suggests that specification of this prior hyperparameter may be difficult in practice, while we can easily optimize $\lambda$ in the regularization approach by cross-validation.

The MCMC approaches without model selection perform poorly, as should be expected from data generated by a sparse model. MCMC methods based on a non-hierarchical model selection are also clearly inferior to the hierarchical counterpart. This is not surprising, as we have simulated data from a hierarchical model.

In Table 2.1 we have also added an additional approach, denoted by $\ell_2$, the equivalent to the $\ell_1$-regularization but using an $\ell_2$-penalty instead of an $\ell_1$-penalty on the coefficients of the log-linear model. This method is equivalent to the MAP estimator with Gaussian priors on $\beta_a$, with the parameter of the distribution optimized by cross-validation. This Ridge-type method does not perform variable selection, but it is competitive for all other criteria that we assessed.

In addition we consider the relaxed $\ell_1$-regularization approach. Rather than using a single penalty parameter $\lambda$, the idea of this method is to control variable selection and parameter estimation by incorporating two penalty parameters. For linear regression it has been proven theoretically as well as empirically by Meinshausen (2007) that under suitable conditions the relaxed $\ell_1$-regularization is better than Lasso.

Overall, the level-$\ell_1$-regularization has good model selection performance (high MSS score) in combination with low negative log-likelihood
score (NLS) and a low mean squared error for the true $\beta$ (RMSE). In addition, it is feasible to optimize the tuning parameter $\lambda$ by cross-validation as the computational cost is very low compared to the MCMC approaches. On the other hand, posterior distributions of estimates from MCMC methods provide additional information about uncertainty in the model space, compared to point estimates from $\ell_1$- or $\ell_2$- regularization.

2.6 Application to Single-Gene Libraries

2.6.1 Datasets

Single-Gene Library of itpr1 Gene

We estimate the splicing interaction pattern for a dataset corresponding to the itpr1 gene, one of three mammalian genes encoding receptors for the second messenger inositol 1,4,5-trisphosphate (InsP$_3$). This gene is subject to alternative RNA splicing, with seven sites of transcript variation, 6 of these within the ORF and among these, $q = 5$ were completely assessed in the single-gene libraries. Five single-gene libraries were built, one for adult rat cerebrum as well as four for different stages of postnatal cerebellar development, namely on days 6, 12, 22 and 90, the latter being considered as adult. Each library consists of between 179 and 277 transcripts which were assessed, i.e. $\sum_{j=1}^{m} n_j \in [179, 277]$, where $n_j$ is the number of counts in cell $j$. This gene is 89% identical at the cDNA level and 95% identical at the amino acid level with the human receptor gene. The complete dataset can be found in Regan et al. (2005).
Single-Gene-Library of CACNA1G Gene

We consider two cDNA libraries from two different developmental stages of human brain, for the gene CACNA1G encoding the low voltage-activated calcium channel gene $\text{Ca}_{v}3.1$. This gene is known to be alternatively spliced at $q = 9$ sites. Detailed information on this dataset can be found in Emerick et al. (2006).

2.6.2 Results of Application to Single-Gene Libraries

Unless stated differently, we report the results using the level $\ell_1$-penalization method.

Results for the itpr1 Gene

We display the interaction vector $\hat{\beta}$ graphically by plotting the components $\hat{\beta}_j$ for the different tissue and development stages in Figure 2.1. Our results suggest that the exons interact mainly in pairs and there is no reliably estimated higher order interaction in the splicing interaction pattern of rat cerebellum. We further notice that the main interaction pattern is very well conserved over different developmental stages. A strong mutual interaction between exons number three, four and five can be observed in all development stages of rat cerebellum as well as in the cerebral tissue. The biggest changes in the interaction pattern during development of rat cerebellum occur from postnatal day six to day 12. This can be seen at position number 10 on the x-axis in Figure 2.1, and it corresponds to the first order interaction between exons two and three, and from day 12 to day 16, the first main effect changes in sign and magnitude. The first main effect decreases progressively from day 6 to adult, reversing in sign between day 12 and 22. Between day 22 and 90, the interaction pattern is strongly conserved. Comparing the splicing interaction patterns between cerebellum and cerebrum in the
adult rat, we see a much more complex pattern in the cerebrum, involving several second order interactions, and therefore a clear distinction from that of the cerebellum.

The conditional independence graphs for the estimated log-linear models are drawn in Figure 2.2, where the thickness of the edges are proportional to the corresponding coefficient of the interaction vector $\hat{\beta}$ (the largest, if there are several giving rise to the same edge) and the radius of the vertices are chosen proportional to the corresponding main effect coefficient. Figure 2.2 graphically exploits the strongly conserved interactions between exons three, four and five. Except for a rather strong interaction between exon two and three on day six, all other interactions appear to be rather small. The graphical representation of the interaction pattern of adult rat cerebrum reveals a more complex interaction pattern with no conditional independences.

The approaches and results presented here can provide valuable insight into the underlying processes in alternative splicing in general, and specifically in the brain development experiments considered here. Most striking is the strong conservation over developmental stages at day 12, 22 and 90 (adult); some differences are showing between postnatal day six and day 12. Also, the conservation between the cerebellum and cerebrum is less pronounced than over developmental stages. Finally, second- or higher-order interaction terms seem to be of minor relevance, suggesting that in this gene/tissue combination, direct interaction mainly happens between pairs of exons, but not combinations of three or more exons.

We have also estimated $\beta$ with the hierarchical Bayesian approach using MCMC. For the choice of $\sigma^2 = 1$ this resulted in very similar interaction patterns as for the level $\ell_1$-penalization method. For $\sigma^2 = 2$ it led to remarkably different results. The according graphical representations of the interaction vector can be found in the Appendix 5.B.
Results for the CACNA1G Gene

We estimate the interaction pattern $\hat{\beta}$ with the step $\ell_1$-regularization method. For the fetal as well as for the adult tissue, a model involving first order interactions only is estimated. By looking at the interaction graphs in Figure 2.3, we clearly see that the patterns exhibit differences. While in the fetal tissue, exon eight interacts with most other exons, this is no more the case for adult tissue where exon number five seems to play a key role. Exon five and eight correspond to exon 30B and exon 35 in Emerick et al. (2006).

We already know that exon 30B plays a key role, because the deletion of segment 30B causes a frameshift, resulting in premature chain termination caused by an early stop codon downstream of this splice site in the new translation reading frame. Transcripts with this condition are often eliminated before they can be translated into proteins through a process called nonsense-mediated decay (NMD). Thus the frequency of these splice variants in the cDNA library may significantly underrepresent their rate of production through transcription and splicing. NMD is an efficient way to use alternative splicing to turn off expression of a single gene in a specific class of cells, without altering gene expression in ways that might effect other genes or even the same gene in neighboring cells of the same tissue. This type of activity shows as a fairly high degree of splicing interactions between this site and other sites, reflecting splicing details in the particular classes of cells where this gene is inactivated by NMD.

Conclusions

We have developed an efficient method for identifying interaction patterns of categorical variables. This can be used to fit a graphical model which is a valuable tool to visualize the conditional dependence struc-
ture among the random variables. In a simulation study, the results of the new level-$\ell_1$-regularization method are superior in comparison to $\ell_1$-regularization and slightly better than the MAP estimator from some of the MCMC methods we considered. With real data, the level $\ell_1$-regularization and hierarchical Bayesian approach led to similar results, subject to a specific choice of priors for the Bayesian method. An important computational advantage of the level-$\ell_1$-method in comparison to MCMC, is that cross-validation becomes feasible which in turn allows for an empirical choice of the tuning parameter.

While the methodology described in this article is motivated by the study of exon splicing interactions in single-gene transcriptomes, it provides a general and flexible toolbox for regularization analysis in relatively high dimensional, sparse contingency tables. Model selection in high dimensional contingency tables has been a traditionally challenging area, and we hope that our generalization of regularization methodologies to this context will prove useful in a variety of areas of computational biology and biostatistics. Several technologies generate categorical data: these include SNP chips that provide genotype and copy number information at the DNA level, sequencing technologies, assays that study binding properties of proteins and binding of RNA to DNA, a variety of disease phenotypes, and more. In most of these contexts the interactions among the variables are critical features in systems biology investigations that aim at studying how the components of complex systems work together in influencing biological outcomes. For example, the log-linear models described here provide a natural approach for fitting very general classes of networks to discrete data. The level-$\ell_1$-regularization is a general tool which can be applied to a wide variety of problems involving sparse contingency tables.

An R package called “logilasso” is available for download on the Comprehensive R Archive Network (CRAN).
2.6. Application to Single-Gene Libraries

Figure 2.1: The upper panel shows the estimated splicing interaction vectors \( \hat{\beta} \) of rat cerebellum tissues at postnatal days 6, 12 and 22. The lower panel shows the splicing interaction vector \( \hat{\beta} \) of rat cerebellum tissues at the age of 90 days, which is considered adult, as well as the splicing interaction vector \( \hat{\beta} \) of rat cerebral tissue at the age of 90 days. Within an interaction degree, the sequence of coefficients is ordered from left to right as follows: e.g. for 2nd order interactions, 123, 124, 125, \ldots, 345, where 1, \ldots, 5 represent exons 12, 23B, 40, 41, and 42 in the rip3r1 gene, as described in Regan et al. (2005).
Figure 2.2: Conditional independence graphs for the estimated log-linear models for the \textit{itpr1} gene. For each graph, the predictive probability score (2.3.7) is reported as a goodness of fit measure. Note the strong mutual interaction between exons three, four and five.
Figure 2.3: Top: Comparison between the interaction pattern $\hat{\beta}$ of fetal and adult tissue. Below: Independence graphs of the estimated log-linear models. On the left are the estimated models, on the right with the strongest interactions only. Within an interaction degree, the sequence of coefficients is ordered from left to right as follows: e.g. for 2nd order interactions, 123, 124, 125, ..., 789, where 1, ..., 9 represent exons 25A, 14, 25B, 26, 30B, 31A, 34, 35, 38B as described in Emerick et al. (2006). We clearly see differences between fetal and adult tissue, especially if we look at the reduced models.
Large contingency tables summarizing categorical variables arise in many areas. For example in biology when a large number of biomarkers are cross-tabulated according to their discrete expression level. Interactions of the variables are generally studied with log-linear models and the structure of a log-linear model can be visually represented by a graph from which the conditional independence structure can then be read off. However, since the number of parameters in a saturated model grows exponentially in the number of variables, this generally comes with a heavy burden as far as computational power is concerned. If we restrict ourselves to models of lower order interactions or other sparse structures
we face similar problems as the number of cells remains unchanged. This is in sharp contrast to high-dimensional regression or classification procedures because, besides a high-dimensional parameter, we have to deal with the analogue of a huge sample size. In addition, high-dimensional tables naturally feature a large number of sampling zeros which often leads to the nonexistence of the MLE estimate. We therefore present a divide-and-conquer approach, where we first divide the problem into several lower-dimensional problems and then combine these to form a global solution. Our methodology is computationally feasible for log-linear interaction modeling with many categorical variables each or some of them having many categories. We demonstrate the proposed method on simulated data and apply it to a bio-medical problem in cancer research.

3.1 Background

We consider the problem of estimation and model selection in log-linear models for large contingency tables involving many categorical variables. This problem encompasses the estimation of the graphical model structure for categorical variables. This structure learning task for discrete graphical models has lately received considerable attention as it plays an important role in a broad range of applications and the resulting models can be used as probabilistic representation of the underlying distribution (Lauritzen, 1996). The conditional independence structure of the distribution can be read off directly from the graph and hence a graphical representation of the distribution is easy to understand. Graphical models for categorical variables correspond to a class of hierarchical log-linear interaction models for contingency tables. Thus, fitting a graph corresponds to model selection in a hierarchical log-linear model.

The difficulties concerning computational aspects can be ranked in the following increasing order. Graphical modeling for discrete categor-
3.1. Background

ical variables is easiest but it doesn’t allow to infer the magnitude of the coefficients $\beta$ in the log-linear model

$$\log(p) = X\beta,$$

see also formula (3.2.3). The next level of difficulty is the estimation of the unknown parameter vector $\beta$ in a log-linear model whose full dimension equals the number of cells in the contingency table. For large tables, the dimension of $\beta$ is huge but under some sparsity assumptions it is possible to accurately estimate such a high-dimensional vector using suitable regularization. The major problem is here that besides the high-dimensionality of $\beta$, the analogue of the sample size (the row-dimension of $X$) is huge, e.g. $3^{40}$ for 40 categorical variables having 3 levels each. Finally, the most difficult problem is the estimation of the probability vector $p$ whose dimension equals again the number of cells in the table. It is rather unrealistic to place some sparsity assumptions on $p$ in the sense that many entries would equal exactly zero which would enable feasible computation. Therefore, it is impossible to ever compute an estimate of the whole probability vector $p$ (e.g. having dimensionality $3^{40}$). Nevertheless, thanks to sparsity of the parameter vector $\beta$ and the junction tree algorithm, it is possible to compute accurate estimates $\{\hat{p}(i); i \in C\}$ for any reasonable-sized collection $C$ of cells in the contingency table. Our presented methodology allows to do all these tasks for categorical variables having possibly different number of levels: inference of a graphical model for discrete variables, of a sparse parameter vector in a log-linear model and of a collection of cell probabilities. There is hardly any other method which can handle all these tasks for contingency tables involving many, say more than 20, variables. In Jackson et al. (2007) some dimensionality reduction is achieved by reducing the number of levels per variable. The reduction is accomplished by collapsing two categories by aggregating their counts if the two categories behave sufficiently similar. If $d$ variables are considered, this method reduces the problem at best to $d$ binary variables. For this special case with binary factors, an approach based on many logistic regressions can be used for fitting log-linear interaction models
whose computational complexity is feasible even if the number of variables is large (Wainwright et al., 2007). Another method to address the log-linear modeling problem for large contingency tables is proposed in Kim (2005) where the variables are grouped such that they are highly connected within groups but less between groups and graphical models are fitted for these subgroups. A graph-theoretical proof is given in Kim (2005) showing that the prime separators of the subgroups appear as prime separators of the full graph. The models are then combined using so-called graphs of prime separators. The implementation of the combination however is not an easy task and no exact algorithm is given on how to combine the models.

Motivated by the approach in Kim (2005), we also propose a so-called divide-and-conquer algorithm, where the dimensionality reduction is achieved by collapsing the table on certain variables and thereby reducing the problem to smaller tables which can be handled more easily. All the fitted lower-dimensional log-linear models are combined appropriately to represent an estimation of the joint distribution of all variables. The procedure enables us to handle very large tables e.g. up to hundreds of categorical variables, where some or all of them can have more than two categories. This multi-category framework cannot be treated by the approach in Wainwright et al. (2007) which can handle large binary tables only.

3.2 Preliminaries

We discuss here log-linear interaction models for contingency tables and establish the connection between the theory of graphical models and log-linear interaction models.
3.2. Preliminaries

3.2.1 Log-Linear Interaction Model

We adopt here the notation of Darroch et al. (1980). Assume we have some factors or categorical variables, indexed by a set $V$. Each factor $v \in V$ has a number of levels $I_v$. The table is the set $I = \prod_{v \in V} I_v$. An individual cell is denoted by $i = (i_v, v \in V)$ and the corresponding cell count by $n_i$. For example, assume we have 2 binary variables, then $V = \{1, 2\}$, $I_1 = I_2 = \{0, 1\}$ and the individual cells are denoted by $(0, 0), (0, 1), (1, 0), (1, 1)$. The total number of cells in a table is $m = |I| = \prod_{v \in V} |I_v|$, which equals 4 in our example. A natural way of representing the distribution of the cell counts is via a vector of probabilities $p = (p(i), i \in I)$. In our example, this would correspond to defining probabilities for all four possible combinations. If a total number of $n$ individuals is observed which are classified independently, then the distribution of the corresponding cell counts $n = (n_1, n_2, \ldots, n_m)^t$ is multinomial with probability $p$. The general log-linear interaction model specifies the unknown distribution $p$ as follows:

$$\log p(i) = \sum_{a \subseteq V} \xi_a(i_a), \quad (3.2.1)$$

where $\xi_a$ are functions of cell $i$ which only depend on the variables in $a$, i.e. on the so-called marginals $i_a$. These functions are called interactions between the variables in $a$. If $|a| = 1$, $\xi_a$ is called main effect, if $|a| = 2$ first order interaction and an interaction of order $k - 1$ if $|a| = k$. For identifiability purposes we impose constraints on the functions, namely that $k^{th}$-order interaction functions are orthogonal to interaction functions of lower order.

A log-linear interaction model usually sets some of these functions equal to zero. Often, it is convenient to work with hierarchical models. These are interaction models where a vanishing interaction forces all interactions of higher order to be zero as well:

$$\xi_a = 0 \implies \xi_b = 0 \quad \text{for all } b \supseteq a$$
Hierarchical models can be specified via the so-called generators or generating class $G$ which is a set of subsets of $V$ consisting of the maximal interactions which are present. Precisely, the generating class $G$ has the following property:

$$\xi_a = 0 \iff \text{there is no } q \in G \text{ with } a \subseteq q. \quad (3.2.2)$$

In our example above with two binary factors, we may consider a model consisting of all main effects, an interaction between 1 and 2 and an interaction between 1 and 3: this corresponds to $G = \{\{1, 2\}, \{1, 3\}\}$.

If we go back to formula (3.2.1) and rewrite it in matrix formulation, we get:

$$\log (p) = X\beta, \quad (3.2.3)$$

where $\beta$ is a vector of unknown regression coefficients and $X \in \mathbb{R}^{m \times m}$ the design matrix. Each row of $X$ corresponds to a certain cell and the columns of $X$ correspond to the functions $\xi_a(i_a)$. The number of columns needed to represent the function $\xi_a$ depends on the number of different states $i_a$ can take on. For example consider a categorical variable $a$ that can take on 3 levels. Then, $\xi_a$ is called a main effect (as $|a| = 1$) and $X_a$ (the columns of $X$ corresponding to $a$) is 2-dimensional. Originally, it would be 3-dimensional but for identifiability purposes, the subspace spanned by $X_a$ is chosen orthogonal to the already existing columns of lower order interaction (here orthogonal to the intercept) and we further choose it orthogonal within the subspace. By choosing the identifiability constraints this way, the parameterization of the matrix used in (3.2.3) is equivalent to choosing a poly contrast in terms of ANOVA. If we go back to our example with two binary factors where
3.2. Preliminaries

$m = 4$, (3.2.3) becomes:

$$
\log p \begin{pmatrix}
(0, 0) \\
(0, 1) \\
(1, 0) \\
(1, 1)
\end{pmatrix} = X\beta,
$$

(3.2.4)

with

$$X = \begin{pmatrix}
1 & 1 & 1 & 1 \\
1 & 1 & -1 & -1 \\
1 & -1 & 1 & 1 \\
1 & -1 & -1 & -1
\end{pmatrix}, \text{ and } \beta = (\beta_{\emptyset}, \beta_1, \beta_2, \beta_{12}),$$

where the first column is the intercept and belongs to $a = \emptyset$, the second column belongs to $a = 1$ and has entry 1 whenever variable $a = 1$ takes on the first level and -1 else and similarly for the third column. The fourth column belongs to the interaction between variable 1 and 2. A description of $X$ for the general case can be found in Dahinden et al. (2007). In the following we will denote the components of $\beta$ belonging to $X_a$ with $\beta_a$.

In the next section, we will make the link between the generating class and graphical models.

### 3.2.2 Graphical Models

We first introduce some terminology. A graph is defined as a pair $G = (V, E)$, where $V$ is the set of vertices or nodes and $E \subseteq V \times V$ is the set of edges linking the vertices. Each node represents a (categorical) random variable. Here we only consider undirected graphs which means that $(u, v) \in E$ is equivalent to $(v, u) \in E$. A path from $u$ to $v$ is a sequence of distinct nodes $v_0 = u, \ldots, v_n = v$ such that $(v_i, v_{i+1}) \in E$ for all $i \in \{0, 1, \ldots, n - 1\}$. Given three sets of variables $A, S, B \subseteq V$, we say that $S$ separates $A$ from $B$ in $V$ if all paths from vertices in $A$ to vertices in $B$ have to pass through $S$. Consider a random vector $Z =$
{Z_v, v \in V} with a certain distribution. We say that the distribution of Z is globally Markov with respect to a graph G if for any 3 disjoint subsets A, S, B \subseteq V the following property holds:

\[ S \text{ separates } A \text{ from } B \implies Z_A \perp Z_B | Z_S, \quad (3.2.5) \]

where the symbol “⊥” denotes (conditional) independence. This states that we can read off conditional independence relations directly from the graph if the distribution is globally Markov with respect to the graph and graphical models therefore provide a way to represent conditional (in)dependence relations between variables in terms of a graph structure. We say that a set of nodes of G forms a complete subgraph of G if every pair in that set is connected by an edge. A maximal complete subgraph is called a clique.

### 3.2.3 Graphical Models and Hierarchical Log-Linear Models

The undirected graphical model represented by a graph G corresponds to a hierarchical log-linear model where the cliques of the graph are the generators of the model. If we go back to our example in the previous section and assume that \( \beta_{12} \neq 0 \) in formula (3.2.4), then the hierarchical log-linear model (3.2.4) can be represented by the graphical model on the left side in Figure 3.1; and if \( \beta_{12} = 0 \), then the corresponding graphical model is the one on the right side in Figure 3.1.

On the other hand, assume that the generators of a log-linear model
are given by a set $\mathcal{G}$. By connecting all the vertices appearing in the same generator, the so-called interaction graph builds up. By the definition of the interaction graph and by looking at formula (3.2.1), it becomes clear that the distribution induced by the log-linear model is Markov with respect to the interaction graph and we can read off conditional independences directly from the graph. It is also clear that $\mathcal{G}$ corresponds to a graphical model via its interaction graph if and only if $\mathcal{G}$ is the set of cliques of this graph. In that case we say that $\mathcal{G}$ is a graphical generating class. If there are cliques in the interaction graph which are not in $\mathcal{G}$, the hierarchical log-linear interaction model is not graphical and its interaction structure cannot be adequately represented by the graph alone. However, the graph may still completely represent all conditional independencies of the underlying distribution. The simplest example of a hierarchical log-linear model which is not graphical is $V = \{1, 2, 3\}$ and $\mathcal{G} = \{\{1, 2\}, \{2, 3\}, \{3, 1\}\}$. Its interaction graph is shown in Figure 3.2 which has as its clique the complete graph $\{1, 2, 3\}$ and obviously, it is not equal to the set of generators $\mathcal{G}$. Any joint probability distribution of discrete random variables can be expanded in terms of a log-linear interaction model. For some distributions it is possible to represent all (conditional) independencies in an undirected graphical model and these distributions are called faithful

**Figure 3.2:** Example of an interaction graph corresponding to a hierarchical log-linear model which is not graphical.

![Interaction Graph Example](image)
to their interaction graph $G$ or we say that the graph is a perfect map of the distribution. In other words, the graph captures all and only the conditional independence relations of the distribution.

### 3.2.4 Collapsibility

Classifying objects into cells according to many categorical variables leaves us with high dimensional contingency tables. Analyzing such large contingency tables directly turns out to be difficult for two reasons. First, some of the cells are likely to be empty because of the exponentially growing amount of cells with growing number of variables, and this may lead to non-existence of the MLE estimator. Secondly, not only the number of cells is growing but also the number of parameters to estimate. Therefore it is often tempting to reduce the dimension of the table by collapsing over some variables if we want to determine the association between the remaining variables. Collapsing over a variable simply means summing over that variable and thereby to reduce (collapse) the table to the remaining dimensions. However in doing so, wrong associations between the variables may be introduced and original associations can vanish. Therefore, one has to be careful when collapsing is permitted and when it is not.

First, we define collapsibility in a precise sense.

**Definition (Collapsibility):** We say that a variable is collapsible with respect to a specific interaction $\xi_a$, when the interaction in the original contingency table is identical to the interaction in the collapsed contingency table.

The general result regarding collapsibility which goes back to a theorem stated in Bishop et al. (1975) can be summarized as follows:

By collapsing a table over a variable which interacts with $s$ other variables, then $s$ and higher order interactions between the remaining variables are not changed in the collapsed table. Conversely, lower-order
interactions between the remaining variables are affected by collapsing.

For example, if we collapse over a variable which only interacts with one other variable no interaction changes by collapsing over that variable, but main effects may be changed. If we collapse over a variable which is independent of all other variables, then neither any main effects nor any interactions change.

### 3.2.5 Decomposability

As mentioned in Section 3.1, we will use a divide-and-conquer algorithm to perform model selection in log-linear interaction models. As we have seen, a graphical log-linear model has a corresponding graph which represents all conditional independencies. The graph may be decomposed into several subgraphs, where each of the subgraphs is analyzed individually and the results can then be combined. The definition of such a decomposition is as follows:

**Definition (Decomposition):** A triple of disjoint subsets \((A,S,B)\) of the vertex set \(V\) forms a decomposition if

1. \(V = A \cup S \cup B\).
2. \(S\) separates \(A\) from \(B\).
3. \(S\) is complete.

Decomposability is defined recursively: A graph is decomposable if it is complete or if there exists a decomposition \((A,S,B)\) where the subgraphs of \(G\) restricted to the vertex sets \(A \cup S\) and \(S \cup B\) are decomposable.
Chapter 3. Log-Lin. Model Selection with Decomposition

**Figure 3.3:** Left: Separating node 2 has corresponding index 2. Right: Separating node 2 has index 1.

Denote by $C$ the set of all cliques of a decomposable graph and by $S$ the set of all separators. A decomposition consists of all cliques and separators $C, S$. As is proved for example in Lauritzen (1996), for a decomposable graph with decomposition into cliques $C$ and separators $S$, the probability of a cell $i$ is given by the following formula:

$$
p(i) = \frac{\prod_{C \in C} p(i_C)}{\prod_{S \in S} p(i_S)^\nu(S)},
$$

(3.2.7)

where $\nu(S)$ is the so-called index of the separator. The formal definition of the index is a bit cumbersome and is given in the Appendix 5.C. However, intuitively it can be thought of as the number of times the set $S$ acts as a separator. For example: on the left side of Figure 3.3, node 2 separates $\{1\}$ and $\{4\}$ (the cliques consisting of single nodes $\{1\}$ and $\{4\}$) and also separates $\{1\}$ and $\{3\}$. Therefore the index of the separating node 2 is 2, as the node 2 acts twice as separator. If we look at at the right side of Figure 3.3, we see that the node 2 only separates $\{1\}$ from the clique $\{3, 4\}$ as the single nodes 3 and 4 are no longer cliques since there is an edge between them and therefore, the node 2 only acts once as separator and thus the corresponding index is 1.
Figure 3.4: Example of a non-decomposable graph.

Remark

It might not be possible to decompose the graph into decomposable (smaller) components. By definition, this is the case for non-decomposable graphs. The simplest example of a non-decomposable graph is given in Figure 3.4. But the analogue in the case of non-decomposable graphs is a decomposition into prime (irreducible) subgraphs. Prime subgraphs are graphs which do not admit a proper decomposition (3.2.6). A maximal prime subgraph is a prime subgraph with vertex set $U$ where the subgraph with vertex set $W$ is decomposable for all $W$ with $U \subset W \subseteq V$. There exists a maximal prime subgraph decomposition (MPD) and it is unique. A proof as well as an algorithm to get the MPD is given in Olesen and Madsen (2002). The main idea behind the algorithm is to add a minimal number of edges to the graph, such that becomes decomposable (this step is called minimal triangulation), perform the decomposition and then aggregate the cliques again whose separators are incomplete. The remaining components are the prime components. An analogue of formula (3.2.7) with an MPD can then be derived.

3.3 Divide-and-Conquer Algorithm

The main problem of fitting a log-linear model, doing model selection in a log-linear model or fitting a graphical model to data is that it quickly
becomes computationally infeasible with growing number of variables and growing number of levels per variable.

If we knew that the underlying graph is sparse, we could use a decomposition and collapse the contingency table on sub-tables given by the cliques $C$ and the separators $S$ given by the decomposition. Then we could perform model selection in the collapsed tables and combine the estimates according to formula (3.2.7). The problem is that we don’t know the graph and therefore we don’t know $C$ and $S$ for the decomposition. In the next section we propose a method how to come up with an initial graph estimate.

### 3.3.1 Selection of Marginal Models to Collapse on

A log-linear models measures the association among the variables. The association between two variables can be measured by doing regression from one variable upon the others. It is thus reasonable to apply a regression method to find groups of variables which are highly associated within a group but only weakly between groups. These groups of variables are the ones which we will collapse on. Inspired by Kim (2005), we use a nonparametric regression approach for finding these groups. But instead of using a single regression tree, we use a Random Forests approach Breiman (2001). The reason to use a regression- or classification-tree or a Random Forests approach consisting of such trees is that trees can naturally incorporate interactions between variables without running severely into the curse of dimensionality. For a detailed description of Random Forests see Breiman (2001).

### Recursive Thinning of the Graph

Random Forests incorporates a way of measuring the importance of individual variables in explaining the response variable in a regression.
3.3. Divide-and-Conquer Algorithm

For example, the importance measure can be the number of times a variable has been chosen as split variable (selection frequency) or the decrease in the so-called Gini index or the permutation accuracy which measures the prediction accuracy before and after permuting a variable. By doing regression from each variable on all others, a so-called importance matrix builds up. Assume we have a discrete random variable $Z$ consisting of $Z_1, \ldots, Z_d$. The algorithm to build up the importance matrix is:

- **Set** $impat = d \times d$ matrix consisting of 0

  **For** $i$ **in** 1:d
  Do regression $Z_i \sim Z \setminus Z_i$
  **Set** $impmat[i,\neg i] \leftarrow$ importance of regression

  **Return** $impmat$

  A high entry in the importance matrix indicates a strong association between the corresponding row and column variable. However, one has to be careful in choosing the importance criteria as far as comparability of the importance between various predictor variables as well as between different regressions is concerned. It has been shown that the popular importance criteria in Random Forests such as the Gini index, selection frequency or the permutation accuracy importance measure are all strongly biased towards variables with more categories (see Strobl et al. (2007)). We therefore propose to use the cforest method proposed by Strobl et al. (2007) which provides a variable importance measure that can be reliably used for variable selection even in situations where the predictor variables vary in their scale of measurement or their number of categories. However, the importance measures are only consistent within rows but not between rows as the variable importance not only depends on the predictor variables in a regression but also on the response variable and therefore, they cannot be directly compared between rows. For that reason we only consider the ranks of the importance matrix entries within rows, meaning that the smallest entry in a
row has rank 1, the next higher has rank 2 and so on. The idea is that if two variables $i$ and $j$ are strongly (conditionally) dependent, then $j$ has high rank in explaining $i$ and the other way round, meaning that the $\text{impmat}_{i,j}$ and $\text{impmat}_{j,i}$ entries are both large.

We then recursively eliminate the edges with lowest corresponding importance matrix entry. Thereby we only eliminate the edge if both entries $(i,j)$ and $(j,i)$ pointing to the variables $i$ and $j$ are below a certain threshold, as the importance matrix is not symmetric. When doing so, we start from the full graph and successively eliminate edges.

**Decomposition**

Whenever an edge is deleted in the process of the recursive thinning procedure from Section 3.3.1, we decompose the graph according to formula (3.2.6). For non-decomposable graphs we decompose the minimal triangulation of the graph, which means we add edges to the graph until it becomes decomposable (minimal triangulation) and don’t perform the aggregation step described in Section 3.2.5. Formula (3.2.7) also holds for such a triangulated graph. The reason for doing so is that at this point we are not interested in an optimal decomposition, but we only want to decompose the graph into smaller subgraphs for which we have the computational capacity to fit a log-linear model. Therefore it is not crucial that we have the sparsest possible subgraphs to collapse on. It is only important that the size of the graph to collapse on is appropriate for our problem and then, any reasonable log-linear model selection procedure can be applied for these subgraphs. If a clique in the (triangulated) thinned graph is small enough, e.g. for binary variables this might be the case for around 10 variables, we split off part of this clique and only proceed with the remaining sub-graph. In detail, if $A \cup S$ corresponds to a clique in the (triangulated) thinned graph, where $S$ separates $A$ from $V \setminus \{A \cup S\}$ and say the number of nodes in $A \cup S$ is less or equal to 10, then we split off the nodes in $A$. We
3.3. Divide-and-Conquer Algorithm

proceed that way (recursively deleting edges, decomposing the thinned graph and trying to split off cliques) until the graph decomposition of the remaining variables consists of cliques which are small enough to be able to perform log-linear model selection.

3.3.2 Combination of results

Assume we have collapsed the table on the cliques \( \mathcal{C} \) and separators \( \mathcal{S} \) of the graph induced by the recursive thinning and decomposition procedure. Furthermore, assume that we have fitted a model for each of these sub-tables (the collapsed tables on \( \mathcal{C} \) and \( \mathcal{S} \)). We then get the log-linear model corresponding to the full graph by using formula (3.2.7):

\[
\log p(i) = \sum_{C \in \mathcal{C}} \log p(i_C) - \sum_{S \in \mathcal{S}} \nu(S) \log p(i_S) \\
= \sum_{C \in \mathcal{C}} X_C \beta_C - \sum_{S \in \mathcal{S}} \nu(S) X_S \beta_S,
\]

(3.3.8)

where \( X_C \) and \( X_S \) are the design matrices resulting from restricting the total design matrix to nodes in \( \mathcal{C} \) and \( \mathcal{S} \) respectively. The same notation applies to \( \beta_C \) and \( \beta_S \). Formula (3.3.8) describes how to aggregate the results of the collapsed tables. In addition, one can derive from (3.3.8) that if we have 3 disjoint subsets \( A, S, B \) where \( S \) separates \( A \) from \( B \), then we can safely collapse over \( B \) without changing an interaction between variables in \( A \) or between variables consisting of a mix of \( A \) and \( S \). The only interactions which might change are the ones between variables which are exclusively in \( S \) (in the following denoted by separator interactions). This is in accordance with the result stated in Section (3.2.4). However, if we do model selection for all sets in \( \mathcal{C} \) and \( \mathcal{S} \), we typically get a model which is too big. The reason for this is that, as mentioned above, if we have a decomposition \( (A, S, B) \) of \( V \), then collapsing on the set \( A \) might introduce interactions between variables which are exclusively in \( S \) (separator interactions might
be changed). But as formula (3.3.8) holds, the introduced interactions have a very small $\beta$ coefficient. We therefore expect that if we threshold the interaction vector, most of the introduced zeros belong to so-called separator interactions $\xi$: \( \exists S \in S \text{ with } \xi \in S \), i.e. interactions exclusively contained in a separator. Consequently, we set the threshold that the introduced zeros belong to equal parts to separator- and non-separator interactions. See Figure 3.5 for a graphical illustration of the procedure. In Section 3.5.1 we will argue empirically that such a thresholding rule works well.

### 3.4 Graphical Model Selection Procedures

We distinguish between divide-and-conquer (decomposition) and global approaches: thereby, divide-and-conquer (decomposition) methods refer to approaches where the graph is decomposed as described in Section 3.3.1 and any graphical model selection procedure described in Section 3.4.1 below is used. In contrast, a global method does not involve such a decomposition and the global approaches used here are described in Section 3.4.2.

#### 3.4.1 Model Selection Procedures used in Combination with the Decomposition Approach

**$\ell_1$-Regularized Model Selection**

Inspired by the Lasso, originally formulated by Tibshirani (1996) for estimation and variable selection in linear regression, a model selection approach for log-linear models has been developed in Dahinden et al. (2007).

The coefficient vector $\beta$ is estimated with the group-$\ell_1$-penalty (Yuan
Figure 3.5: Illustration of how many separator edges to take up into the model. $x$-axis: the fraction of separator interactions $\xi$ with $\hat{\beta}_\xi = 0$ among all separator interactions. $y$-axis: the fraction of non-separator interactions with estimated interaction coefficient equal zero. The points correspond to different levels of thresholding. We see that if we threshold 30% of the non-zero $\hat{\beta}$ coefficients, we have almost exclusively thresholded separator interactions, as we would expect.
and Lin, 2006):

$$\hat{\beta}^\lambda = \arg \min_{\beta} \left[ -\frac{1}{n} l(\beta) + \lambda \sum_{a \subseteq C, a \neq \emptyset} \|\beta_a\|_{\ell^2} \right], \quad (3.4.9)$$

where $l(\beta) = \sum_{i=1}^{m} n_i (X\beta)_i = \log P[\beta | n] + c$. Therefore $l(\beta)$ is up to an additive constant $c$, which does not depend on $\beta$ the log-likelihood function. This minimization has to be calculated under the additional constraint that the cell probabilities add to 1:

$$\sum_{i=1}^{m} \exp \{(X\beta)_i\} = 1.$$

The group-$\ell_1$-penalty

$$\sum_{a \subseteq C, a \neq \emptyset} \|\beta_a\|_{\ell^2}, \text{ where } \|\beta_a\|_{\ell^2}^2 = \sum_{j} (\beta_a)_j^2,$$

has the property that the solution of (3.4.9) is independent of the choice of the orthogonal subspace of $X_a$ and furthermore, the penalty encourages sparsity at the interaction level, meaning that the vector $\hat{\beta}_a$ corresponding to the interaction $\xi_a$ has all components either non-zero or zero. In addition, by using group-$\ell_1$ model selection we avoid the sampling zero problem, which is problematic regarding the existence of the MLE (see e.g. Christensen (1991)).

The tuning parameter $\lambda$ can be assessed by cross-validation: we divide the individual counts into a number of equal parts and in turn leave out one part and use the rest to form a training contingency table with cell counts $n_{train}$.

**Stepwise Forward**

The stepwise forward procedure aims to minimize the AIC-type criterion $sk - 2 \log(l)$, where $l$ is the maximized value of the likelihood function for
the corresponding model with \( k \) degrees of freedom; \( s = 2 \) corresponds to the genuine AIC. Here we also vary the parameter \( s \); a large parameter leads to sparser models and if \( s = 0 \) generally the saturated model is chosen.

### 3.4.2 Global Model Selection Procedures

**Wainwright et al. Method**

In Wainwright et al. (2007) the problem of estimating the graph structure of binary valued Markov networks is considered. They propose to estimate the neighborhood of any given node by performing \( \ell_1 \)-penalized logistic regressions on the remaining variables. Assume we have \( d \) binary random variables and observations thereof \( z = (z_1, \ldots, z_d) \in \{0,1\}^d \).

Furthermore, we assume that the data are generated under the so-called Ising model:

\[
\log p(z) = \sum_{s,t=1}^{d} \theta_{st} z_s z_t + \Psi(\Theta),
\]

(3.4.10)

where \( \Theta \) is a symmetric \( d \times d \) matrix and \( \Psi(\Theta) \) is a normalizing constant which ensures that the probabilities add up to one. This constant is also known as the log-partition function. If we go back to the log-linear interaction model described in Section 3.2.1 with binary variables, i.e. the cell \( i \in \{0,1\}^d \), then by comparing formula (3.2.1) to (3.4.10) we see that the Ising model is a log-linear model whose highest interactions are of order one and the parameterization is, in terms of ANOVA, with Helmert instead of poly contrasts. Therefore, the interaction graph builds up by connecting the nodes \( s \) and \( t \) for which \( \theta_{st} \neq 0 \). If we go back to the example in Section 3.2.1 where the two random variables
are binary and rewrite formula (3.4.10) in matrix formulation we get:

\[
\log p \begin{pmatrix}
(0,0) \\
(0,1) \\
(1,0) \\
(1,1)
\end{pmatrix}
= X \begin{pmatrix}
\Psi(\theta) \\
\theta_{11} \\
\theta_{22} \\
2\theta_{12}
\end{pmatrix},
\quad \text{with } X = \begin{pmatrix}
1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 \\
1 & 1 & 0 & 0 \\
1 & 1 & 1 & 1
\end{pmatrix}.
\]

(3.4.11)

By comparing (3.4.11) to (3.2.4) we see that the normalizing constant \(\Psi(\theta)\) corresponds to the intercept \(\beta_0\), \(\theta_{11}\) to the main effect \(\beta_1\), \(\theta_{22}\) to the main effect \(\beta_2\) and \(2\theta_{12}\) corresponds to the first order interaction \(\beta_{12}\).

It holds that:

\[
\log \left( \frac{p(z_s = 1|z\setminus s)}{p(z_s = 0|z\setminus s)} \right) = \theta_{ss} + \sum_{t=1, t\neq s}^d 2\theta_{st}z_t,
\]

(3.4.12)

and therefore, we can infer the matrix \(\theta\) by doing \(d\) logistic regressions from each variable on the remaining variables: \(\theta_{ss}\) is then the intercept in the logistic regression from \(z_s\) on the remaining variables \(z\setminus s\) and \(2\theta_{st}\) equals the regression coefficient corresponding to \(z_t\). Wainwright et al. (2007) apply \(\ell_1\)-penalized logistic regression to estimate the \(\theta\)-matrix. However, the resulting \(\theta\) is not necessarily symmetric and it has to be symmetrized for example by taking the maximum of \(\theta_{st}\) and \(\theta_{ts}\). Furthermore, Wainwright et al. (2007) prove that under certain sparsity assumptions their method correctly identifies the underlying graph structure. We emphasize that the approach in Wainwright et al. (2007) works for binary variables only, while the decomposition approaches in Section 3.4.1 work for general multi-category variables.

**Random Forests (global)**

Random Forests (Breiman, 2001) is a regression method which internally calculates variable importances and these variable importances can be used to build an importance matrix. The exact procedure how
3.5 Simulation Study

We simulate from a log-linear interaction model corresponding to a graph with 40 nodes and 91 edges. Each node corresponds to a binary variable (and thus, we can compare with the method in Wainwright et al. (2007)). The graph is displayed in Figure 3.6. This is the same simulation setting as was used in Kim (2005). We generate 10 datasets each consisting of 10000 observations according to the graph in Figure 3.6. We restrict the log-linear model selection procedure to first order interaction models even though the true underlying graph includes interactions of higher order. The reason is that projecting a log-linear model involving higher order interactions to the best (w.r.t. the Kullback-Leibler divergence) first-order interaction model usually yields a log-linear model with non-zero first-order interaction coefficients.

3.5.1 Results

In Section 3.5.1 we compare the approaches with respect to performance in estimating the correct model structure and in Section 3.5.1 with respect to accuracy for estimating the parameter vector.

Model Selection Performance for the Graph

First, we assess in Figure 3.7 the optimal decomposition size. We decompose (with Random Forests) the graph into several subgraphs using
different decomposition sizes (maximal size of cliques equal to 3, 5 and 10), then use the $\ell_1$-penalization approach described in Section 3.4.1 to estimate the log-linear model for each subgraph and combine these results as described in Section 3.3.2. An ROC curve is drawn, where the endpoints of the curves correspond to the selected model with $\lambda$ in (3.4.9) chosen by cross-validation. The curves correspond to models which arise by successively eliminating edges with smallest interaction vector coefficient. We see here that larger decomposition sizes lead to slightly more favorable ROC curves. The picture remains qualitatively the same if we use stepwise forward instead of $\ell_1$-penalization.

In Figure 3.8, $\ell_1$-penalization and the stepwise forward method are compared after the decomposition with Random Forests using maximal clique size equal to 10. For the stepwise forward method the line builds
3.5. Simulation Study

Figure 3.7: Comparison of decomposition sizes. Decomposition into cliques of maximal size 3, 5 and 10 with Random Forests and subsequent model selection with $\ell_1$-penalization. The curves corresponds to models which arise by thresholding the final $\hat{\beta}$-coefficient.

up by varying $s$ (compare Section 3.4.1), while the $\ell_1$-penalization approach starts from the CV-solution (endpoint of the curve) and uses hard-thresholding of $\hat{\beta}$ for obtaining the values on the ROC curve. We see that the stepwise forward and the $\ell_1$-penalization approach lead to models which have approximately the same number of false positive and false negative edges, but the $\ell_1$-approach is slightly favorable. However, if we threshold the $\hat{\beta}$ coefficients of the AIC ($s=2$) solution, the line of the stepwise forward approach is very close to the $\ell_1$ solution (not
shown here). Note that in Figure 3.8 the black cross indicates the final selected $\ell_1$-solution if the thresholding procedure described in Section 3.3.2 is applied.

**Figure 3.8:** Comparison of decomposition approaches. The graph is decomposed into cliques of maximal size 10 and then $\ell_1$-penalization and stepwise forward model selection is applied to the subgraphs before they are combined using formula (3.3.8). The dotted line corresponds to the difference in the true positive rate for the two procedures when the two procedures are compared at the false positive rate of the final $\ell_1$-penalization model.

In Figure 3.9, the global approaches (Wainwright et al. (2007) and global Random Forests regression from Section 3.4.2) are compared with
the $\ell_1$-penalization decomposition approach. In order to keep a simple overview, the line for the stepwise procedure is no longer drawn. We see that our decomposition approach slightly outperforms the global approaches (but the global Random Forests does not yield an estimate for the parameter vector and the method from Wainwright et al. (2007) does so only for binary variables). If we assess the optimal $\lambda$ parameter for the method from Wainwright et al. (2007) in each regression with cross-validation, we get a model which is close to the saturated model as we can see by the red cross indicated in Figure 3.9. Even though Figures 3.8 and 3.9 only represent one simulated dataset, the picture doesn’t change for other simulations. The lines for the stepwise and the $\ell_1$ approach are always very close, with the $\ell_1$-penalization method slightly better and both methods clearly superior to the global approaches. The reason why Figures 3.8 and 3.9 only display results from one dataset is that the single final models cannot be averaged over different datasets as they have different positions on the curves for different datasets (different numbers of true and false positives), and if we average over all these values, the result is not very meaningful anymore. However, we can average the differences of true positive rates for e.g. the final $\ell_1$-penalized model and the stepwise model with the same number of false positives as the $\ell_1$-solution (dotted line in Figure 3.8). The results of such comparisons are summarized in Table 3.1. We see that the final $\ell_1$-penalization decomposition solution yields a significantly higher true positive rate than the corresponding stepwise solution. On the other hand, the comparison between the AIC solution and the corresponding (in terms of false positives) $\ell_1$-solution shows that there is no significant difference.

The stepwise decomposition method, the $\ell_1$-decomposition approach as well as the global method in Wainwright et al. (2007) yield estimates of the interaction vector $\beta$ (for the $\ell_1$- and the stepwise forward approach) or $\theta$ (for the Wainwright approach). In the next section we will compare these methods with respect to the performance for estimating the parameter vector $\beta$. 
Figure 3.9: Comparison of global approaches (Wainwright et al. (2007) and global Random Forests from Section 3.4.2) with ℓ₁-penalization decomposition and stepwise forward decomposition approaches. For the ℓ₁-penalization decomposition approach, the line builds up by thresholding the \( \hat{\beta} \)-coefficient where \( \lambda \) is chosen by CV. For Random Forests, the edges with least importance are successively eliminated and for the Wainwright et al. method, the tuning parameter \( \lambda \) is varied.

Comparison of Parameter Estimation

All approaches considered here yield the interaction vector \( \beta \) only up to a constant (except for global Random Forests which does not yield an estimate of \( \beta \)). For the method in Wainwright et al. (2007), the
Table 3.1: Comparisons of true positive rates between the final \( \ell_1 \)-penalization decomposition solution (denoted by “\( \ell_1 \) final”) and the corresponding (in terms of false positives) solutions of other methods (denoted by “Stepwise”, “Wainwright” and “Random Forests”). In addition the AIC solution from the stepwise decomposition approach is compared to the corresponding \( \ell_1 \)-penalization decomposition solution (“Stepwise AIC - \( \ell_1 \)”.

<table>
<thead>
<tr>
<th></th>
<th>Mean difference</th>
<th>p-value (t-test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \ell_1 ) final - Stepwise</td>
<td>0.052</td>
<td>0.036</td>
</tr>
<tr>
<td>( \ell_1 ) final - Wainwright</td>
<td>0.060</td>
<td>0.011</td>
</tr>
<tr>
<td>( \ell_1 ) final - Random Forests</td>
<td>0.013</td>
<td>0.256</td>
</tr>
<tr>
<td>Stepwise AIC - ( \ell_1 )</td>
<td>0.002</td>
<td>0.512</td>
</tr>
</tbody>
</table>

constant \( \Psi(\Theta) \) given in formula (3.4.10) is not computed at all. For the stepwise as well as for the \( \ell_1 \)-penalization method we ensure that the probabilities add up to one in each submodel, but when combining the models according to formula (3.3.8) this property only holds approximately. The reason is that we do not consistently estimate the marginal distributions, i.e. for a decomposition \( V = (A,S,B) \) we require that \( \sum_{x_A} \hat{p}_{A,S}(x_A, x_S) = \sum_{x_B} \hat{p}_{B,S}(x_B, x_S) \), but this is only fulfilled approximately. Thus, we need to normalize. For sparse graphs, we can use the junction tree algorithm to calculate the constant. We know that the probability distribution can be represented by a product of clique marginal distributions divided by the product of separator distributions. The junction tree algorithm constructs a junction tree from the graph, we can then assign potentials for each cluster of the tree and by message propagation on the tree, the potentials are transformed to consistent marginal probabilities such that we can calculate the normalizing constant. For a detailed description see Lauritzen (1996).

We compare the estimated probabilities using an expression which is
Chapter 3. Log-Lin. Model Selection with Decomposition

up to a constant the Kullback-Leibler divergence between the estimated and the true probability (non-normalized Kullback-Leibler divergence):

\[- \log \left( \prod_i \hat{p}_i^{p_i} \right) = - \sum_i p_i \log \hat{p}_i, \quad (3.5.13)\]

where \( \hat{p} \) is the estimated probability vector and \( p \) denotes the true probability vector. As this sum requires the calculation of \( 2^{40} \approx 10^{12} \) components of \( \hat{p} \) and \( p \) and the summation of the two huge vectors, this is computationally not feasible. To avoid this problem, we calculate an empirical version by simulating one million observations from the graph in Figure 3.6 and summing over these values only. The results are summarized in Table 3.2. In this table we have included our three estimators: the two decomposition approaches denoted by “\( \ell_1 \)-penalization” and “stepwise forward” and the global “Wainwright et al.” approach. In addition we have included the “full decomposition model”, where no model selection is performed after decomposition and MLE on the decomposed model is used. Again, we see that the \( \ell_1 \)-penalization approach and the stepwise forward approach perform similarly and the approach in Wainwright et al. (2007) is clearly inferior. The Wainwright solution listed in Table 3.2 is the solution for the minimal \( \lambda \) for which the normalization constant could be computed. This solution is also indicated in Figure 3.8. For the CV-optimal solution, which is indicated in Figure 3.8, the normalizing constant cannot be computed as the junction tree algorithm does hardly provide a simplification due to the fact that the CV-optimal solution almost corresponds to the full model. However, the maximal computable solutions correspond to very large models, which on average involve 22.05% of all possible edges, compared to 17.01% for the \( \ell_1 \)-solution and 11.66% for the true graph.

Table 3.3 provides further insight about significance of the differences in Table 3.2. All methods are compared against each other using a paired t-test for the empirical (non-normalized) Kullback-Leibler divergences. The \( p \)-values are listed in Table 3.3. One can see that there is no significant difference for the decomposition approaches (\( \ell_1 \)-
3.5. Simulation Study

Figure 3.10: Mean empirical (non-normalized) Kullback-Leibler distance between true and estimated probability in dependence of the percentage of thresholded coefficients. The vertical line indicates the average of percentages of thresholded coefficients using the thresholding rule from Section 3.3.2.

penalization, stepwise forward and the full decomposition model using MLE for the decomposed models), whereas they are all superior over the approach in Wainwright et al. (2007). This provides evidence that the decomposition of the model is more crucial than the effective choice of the log-linear model fitting procedure afterwards.

Furthermore, it is worthwhile stating that the thresholding of the coefficients (see Section 3.3.2) does hardly influence the likelihood as we can see in Figure 3.10. On average, 35% of the coefficients are thresholded
as indicated by the dotted line. However, for these threshold-optimal solutions, calculated as described in Section 3.3.2 (see also Figure 3.5), the empirical (non-normalized) Kullback-Leibler divergence is approximately the same as for the non-thresholded model (see Figure 3.10).

Table 3.2: *Mean empirical (non-normalized) Kullback-Leibler divergence between true and estimated probabilities. The p-values for all the pairwise comparisons are given in Table 3.3.*

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \ell_1 ) final</td>
<td>4.0080</td>
<td>0.0204</td>
</tr>
<tr>
<td>Stepwise AIC</td>
<td>4.0242</td>
<td>0.0217</td>
</tr>
<tr>
<td>Full decomp. model</td>
<td>4.0223</td>
<td>0.0099</td>
</tr>
<tr>
<td>Wainwright et al.</td>
<td>4.3360</td>
<td>0.1094</td>
</tr>
</tbody>
</table>

Table 3.3: *All possible pairwise comparisons between models: p-values of a paired t-test for the equality of (non-normalized) Kullback-Leibler divergence.*

<table>
<thead>
<tr>
<th></th>
<th>Stepwise AIC</th>
<th>Full decomp. model</th>
<th>Wainwright</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \ell_1 ) final</td>
<td>0.1030</td>
<td>0.0677</td>
<td>4.7 ( 10^{-6} )</td>
</tr>
<tr>
<td>Stepwise AIC</td>
<td>-</td>
<td>0.8080</td>
<td>6.0 ( 10^{-6} )</td>
</tr>
<tr>
<td>Full decomp. model</td>
<td>-</td>
<td>-</td>
<td>7.5 ( 10^{-6} )</td>
</tr>
</tbody>
</table>
3.6 Application to Tissue Microarray Data

3.6.1 Tissue Microarray Technology

The central motivation that led to this work was to fit a graphical model to discrete expression levels of biomarkers resulting from Tissue Microarray (TMA) experiments. Tissue microarray technology allows rapid visualization of molecular targets in thousands of tissues at a time, either at DNA, RNA or protein level. Tissue microarrays are composed of hundreds of tissue sections from different patients arrayed on a single glass slide. With the use of immunohistochemical staining, they provide a high-throughput method to analyze potential biomarkers on large patient samples. The assessment of the expression level of a biomarker is usually performed by the pathologist on a categorical scale: expressed/not expressed, or the level of expression.

Tissue Microarrays are powerful for validation and extension of findings obtained from genomic surveys such as cDNA microarrays. cDNA microarrays are useful to analyze a huge number of genes, e.g. a couple of thousands in one specimen at a time. In contrast, TMAs are applicable to the analysis of one target at a time, denoted as biomarker, but in up to 1000 tissues on each slide. The analysis of the interaction pattern of these biomarkers and in particular the estimation of the graphical model associated with the underlying discrete random variables are of bio-medical importance. These graph-based patterns can deliver valuable insight into the underlying biology. A detailed description of the Tissue Microarray technology can be found in Kallioniemi et al. (2001).

3.6.2 Data

Our TMA dataset consists of tissue microarray measurements from renal cell carcinoma patients. We have information from 1116 patients,
831 thereof having a clear cell carcinoma tumor, which is the tumor of interest here. We have identified 18 biomarkers from which we have information for the majority of the patients. Among 831 ccRCC (clear cell renal cell carcinoma) observations, 527 observations are complete with all biomarker measurements available. For 87 observations one measurement was missing, 64 and 30 observations had 2 or 3 missing values, respectively. 123 observations contained more than 3 missing values and were ignored in the analysis. For the observations with 1, 2 or 3 missing values, multiple imputation was applied using the R package mice (Van Buuren and Oudshoorn, 2007). From 18 biomarkers, 9 are binary and 9 have 3 levels.

3.6.3 Graphical Model

The graphical model corresponding to the TMA data is displayed in Figure 3.11. The thickness of the line corresponds to the $\ell_2$-norm of the respective interaction coefficients. Two biomarkers connected by a thick line, as is the case for nuclear p27 and cytoplasmic p27, indicates a strong interaction. The kinase inhibitor p27 exhibits its function in the cell nucleus and therefore recent studies have focused on nuclear p27 expression. Our graphical log-linear model however shows a tight association between nuclear and cytoplasmic expression of p27. Therefore it can be speculated that both nuclear and cytoplasmic presence is required to ensure proper function of p27. It has been shown that in renal tumors, the von Hippel-Lindau protein (VHL protein) is upregulating the expression of the tumor suppressor p27 (Osipov et al., 2002). The graphical model here provides supporting evidence that VHL indeed regulates p27, and the corresponding $\beta$ coefficient (not displayed here) implies that it is a positive regulation.

Furthermore it has been shown in vitro by Roe et al. (2006) that VHL increases p53 expression which is a tumor suppressor. In our model it seems as if p53 is conditionally independent of VHL. Indeed, it has
long been known that p53 activates expression of p21 (e.g. Kim (1997)). This dependence is displayed very clearly in the graphical model. We can therefore view the p53-p21 pathway with its strong interaction as one unit and it is therefore very reasonable that nuclear VHL interacts with p53. As nuclear VHL is only expressed in 14% of the tumors, and it further makes sense from a biological point of view that the strong interaction between VHL and the p21-p53 pathway is in fact a causal relation, we can indeed speculate that the loss of VHL deactivates the tumor suppressor p53 which in turn favors tumor development.

CA9, Glut1 and Cyclin D1 are all hypoxia-inducible transcription factor (HIF) target genes (Wenger et al., 2005). HIF has not been measured but we can clearly see that all these HIF targets are connected by a rather thick line implying that they might react to a common gene. In addition, CD10 strongly interacts with Glut1 a known HIF target which suggests that CD10 might also be regulated by HIF. The reduction of E-Cadherin expression has been found to be negatively correlated with HIF expression in Imai et al. (2003). This is supported by a strong negative interaction between E-Cadherin and CA9 which is positively correlated with HIF expression (not measured).

A lot of supporting evidence has been delivered for already existing theories. However, two strong interactions, one between PAX2 and nuclear p21 and the other between PAX2 and Cyclin D1 cannot be immediately explained. PAX2 is absent in normal renal tubular epithelial cells but expressed in many clear cell renal cell carcinoma tumours (see Mazal et al. (2005)). Its frequent expression together with the strong interaction with the p21-p53 pathway, Cyclin D1 and PTEN make PAX2 an interesting and possibly important molecular parameter whose exact function and role still remains to be elucidated.
3.7 Discussion

We have proposed a divide-and-conquer procedure to estimate log-linear models for large contingency tables and for fitting discrete graphical
models. In a simulation study we have compared various algorithms and concluded that the divide-and-conquer procedures are very powerful. It seems that the decomposition of the problem is much more crucial than the choice of the algorithm to handle the smaller decomposed datasets; no matter whether $\ell_{1}$-penalized model selection, stepwise forward model selection or no model selection but only parameter estimation is applied after the decomposition, the resulting models are clearly superior to global (non-decomposition) approaches for model selection as well as for probability or parameter estimation.

Maybe most important is the computational feasibility of our procedure for large contingency tables with factors having more than two levels. The proposed method is scalable to orders of realistic complexity (e.g. dozens up to hundreds of factors) where most or all other existing algorithms become infeasible. In particular, our procedure is not only capable of handling binary data but can easily deal with factors with more levels. Furthermore, with the $\ell_{1}$-penalization procedure one doesn’t risk the nonexistence of the parameter estimator in case of sampling zeroes in the contingency table as this might arise in MLE. In summary, we propose a procedure which is capable of handling a large amount of variables and there is hardly any limitation for the sample size $n$. The procedure not only fits a graphical model but also yields an estimation of the interaction vector $\beta$ in a log-linear model and therefore of the cell probabilities. All this is achieved with good performance in comparison to other methods. As a drawback, if the true underlying graph has a clique which is larger than our decomposition size, then some of the edges in the graph are necessarily lost.

We apply the proposed approach to a problem in molecular biology and we find supporting evidence for dependencies between biomarkers which have already been found to exist in vitro or some even in renal tumors, the domain of our application. However, some strong interactions cannot be explained immediately and therefore, new biological hypotheses arise.
An R package called `decompgraph` will be available for download on the Comprehensive R Archive Network (CRAN).
Chapter 4

Mining Tissue Microarray Data to Unhide Disregulated Tumor Suppressor Pathways in Clear Cell Renal Cell Cancer

Clear cell Renal cell carcinoma (ccRCC) is characterized by its high mortality and increasing incidence. Tumor stage and nuclear differentiation grade are the most important prognostic parameters of ccRCC. The risk of progression of ccRCC remains difficult to predict especially for tumors with organ-confined stage and intermediate differentiation grade.
Chapter 4. Detecting Pathway Disregulation with TMAs

Elucidating molecular pathways deregulated in ccRCC might point to powerful prognostic parameters enabling to better predict the clinical course and to facilitate planning of therapeutic approaches. Here we used tissue microarray to analyze pathways in ccRCC specimens that are normally controlled by the von Hippel-Lindau, PTEN, p27, and p53 tumor suppressors. The expression patterns of 18 different molecular parameters were evaluated in over 800 ccRCC patients from which tumor stage, grade, and survival data were available. Tumor staging and grading were refined by performing variable selection using Cox regression and a recursive bootstrap elimination scheme. Patients with pT2 and pT3 tumors that were p27 (nuclear) and CA9 positive had a better outcome than those with combined loss of these proteins (5-year survival for pT2: 60.9% overall versus 73.3% and 51.4%; pT3: 39% overall versus 45% and 32.4%). Among patients with intermediate grade (Thoenes grade 2) both loss of nuclear p27 and cytoplasmic PTEN expression, as well as the presence of phosphorylated ribosomal protein S6 correlated with shortened survival (5-year survival: 49% versus 71%). By applying graphical log-linear modeling for over 700 ccRCC for which the 18 molecular parameters were available, we deduced only a weak conditional dependence between between p27, PTEN, CA9 and p-S6 expression suggesting that in ccRCC the dysregulation of several pathways are crucial for tumor progression. The use of recursive bootstrap elimination, as well as graphical log-linear modeling for comprehensive TMA data analysis permits to disentangle complex molecular contexts and improves predictive evaluations for cancer patients.

4.1 Introduction

The last decades showed an incidental increase of patients diagnosed with renal cell carcinoma (RCC) with clear cell RCC (ccRCC) as the most frequent and aggressive subtype (Kovacs et al., 1997; Moch et al., 2000). Whereas the 5-year survival rate of metastatic ccRCC remains
poor, patients with local tumors have a significant better outcome as these cancers can be treated with radical or partial nephrectomy. However, recurrence develops in about one third of patients with organ-confined ccRCC (Mickisch, 2002).

The best available predictor of the postoperative clinical course of localized RCCs is tumor stage at presentation (Michael and Pandha, 2003). Because a significant difference in outcome within the same stage exists, additional prognostic parameters are used to better predict the outcome in these patients. After tumor stage, the second most important prognostic parameter is the nuclear differentiation grade (Srigley et al., 1997). For tumor grading, four-tired and three-tired grading systems are commonly applied (Ficarra et al., 2005; Goldstein, 1997). More than 50% of the tumors are classified as moderately differentiated, which is not informative for clinical decision making because of an intermediate risk of recurrence (Lohse et al., 2002). Therefore, the identification of molecular prognostic markers is highly desirable to better predict the clinical outcome of RCC. Hence, many efforts have been made to uncover molecular pathways that are critical for the initiation and progression of ccRCC.

Alteration of von Hippel-Lindau protein (pVHL) regulated pathways due to mutation of its coding gene is the main characteristic feature of most sporadic clear cell renal cell carcinoma (ccRCC). The multi-functional protein pVHL acts as an adaptor for transcriptional regulation, the extracellular matrix and the microtubule cytoskeleton. The best-characterized function of pVHL is as a substrate recognition component of the VBC-Cul2 E3 ubiquitin protein ligase complex that targets the hypoxia inducible factor a (HIF-a) for proteolytic degradation. In ccRCC, loss of pVHL tumor-suppressive function leads to the upregulation of HIF-a-mediated transcriptional programs that favour metastatic processes (reviewed in Frew and Krek (2008); Kaelin Jr. (2004)).

The high mutation frequency in many human tumors makes the p53
tumor suppressor to one of the most important genes in human cancer. Because mutations are relatively rare, p53 is thought to play a rather minor role in ccRCC (Contractor et al., 1997; Uchida et al., 1993). The finding that pVHL increases p53 expression by suppressing proteolytic degradation and interacting with p53-activating proteins ATM and p300 (Roe et al., 2006) suggests that in ccRCC p53 inactivation and accumulation of genetic abnormalities are provoked by loss of pVHL.

Several reports showed that pVHL is also capable of weakening Cyclin D1 expression (Bindra et al., 2002) and upregulating the expression of p27 (KIP1) in renal tumors (Osipov et al., 2002). Interestingly, phosphorylation by AKT impairs the nuclear import of p27 thus relieving CDK2 from p27-induced inhibition (Shin et al., 2002). The association of AKT driven cytosolic retention of p27 with an aggressive phenotype in human breast cancer suggests that in ccRCC activation of AKT might also be crucial for tumor progression.

The phosphatase and tensin homologue (PTEN) negatively regulates numerous growth factor receptor mediated signal transductions by dephosphorylating phosphatidylinositol-3,4,5-triphosphate that is a substrate of activated phosphatidylinositol-3-kinase (PI3K) enzymes (Mae-hama and Dixon, 1999). The loss of PTEN function leads to constitutive activation of PI3K downstream components including AKT and mTOR kinases (Sansal and Sellers, 2004). Although PTEN mutations are rare in ccRCC, allele deletions and loss of PTEN protein expression occur in about 30% of the tumors (Alimov et al., 1999; Kondo et al., 2001) indicating the potential role of PTEN in a considerable subset of tumors.

Clear cell Renal cell carcinoma with combined inactivation of tumor suppressor genes might be characterized with a significantly more aggressive phenotype and, consequently, with worse clinical outcome. To test this hypothesis we analyzed the protein expression patterns of 15 proteins involved in VHL/PTEN/p27/p53 regulated pathways using a tissue microarray with more than 800 ccRCC from which informations
4.2. Materials and Methods

about tumor stage, nuclear grade and survival were available. We used Cox regression with recursive bootstrap elimination to statistically explain survival time with biological and clinical parameters. In addition, graphical log-linear modeling was applied to analyze the dependence structure among biological parameters.

4.2 Materials and Methods

4.2.1 Tissue specimen and TMA construction

We constructed TMAs comprising 831 nephrectomy ccRCC collected from the University Hospital of Zurich (Zurich, Switzerland), the Kantonsspital St. Gallen (St. Gallen, Switzerland) and the University Hospital of Basel (Basel, Switzerland) as previously described (Kononen et al., 1998). All ccRCC samples were fixed in buffered (PBS, pH 7.4) formalin between 16 to 24 hours, histologically reviewed by one pathologist (H.M.) and selected for the study on the basis of hematoxylin and eosin-stained tissue sections. This study was approved by the local commission of ethics (ref. number StV 38-2005). Survival time was obtained from 519 patients. The mean age of patients was 62.7 (15-88) and the mean follow-up of patients was 51.9 months (0.1-229). Tumors were graded according to the Thoenes grading system and histologically classified according to the World Health Organization classification (Eble et al., 2004). There were 164 grade I (20%), 414 grade II (50%), 248 grade III (30%), 161 pT1 (20%), 223 pT2 (27%), 412 pT3 (50.5%), and 20 pT4 (2.5%) tumors. Fifteen tumors had no information about tumor stage and tumor grade was missing for 5 tumors.
4.2.2 Immunohistochemistry

Fifteen antibodies, 14 of which are known to recognize proteins being involved in the VHL/p27/PTEN/p53 pathways (Frew and Krek, 2008; Shin et al., 2002; Frew and Krek, 2007; Esteban et al., 2006; Brugarolas, 2007; Stuart et al., 1995; el Deiry et al., 1993; Yan and Shao, 2006; Semenza, 2003), were selected for this study. As CD10 is expressed in the majority of ccRCC (Chu and Arber, 2000) and might therefore be regulated by the pVHL/HIF-axis it was also included in our TMA analysis. A schematic overview of these pathways is illustrated in Figure 4.1. TMA sections (2.5\(\mu\)m) were transferred to glass slides and treated using Benchmark XT, Bond-maX automat systems or manual protocols. Antibodies, dilutions, and the criteria used for interpreting immunohistochemical stainings are listed in Table 4.1.
### 4.2. Materials and Methods

**Table 4.1:** Overview of antibodies, dilutions, and the criteria used for interpreting immunohistochemical stainings.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Antibody-no.</th>
<th>Species</th>
<th>Supplier</th>
<th>Dilution</th>
<th>Treatment</th>
<th>Compartment</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA9</td>
<td>M75</td>
<td>mouse</td>
<td>J. Zavada</td>
<td>1:200</td>
<td>BondmaX</td>
<td>Membr.</td>
<td>neg., weak, strong</td>
</tr>
<tr>
<td>CCND1</td>
<td>P2D11F11</td>
<td>mouse</td>
<td>Ventana</td>
<td>prediluted</td>
<td>BenchMark XT (Ventana)</td>
<td>Nucl.</td>
<td>neg., &lt;5%; &lt;10%; &gt;10%</td>
</tr>
<tr>
<td>CD10</td>
<td>NCL-CD10-270</td>
<td>mouse</td>
<td>Novocastra</td>
<td>1:30</td>
<td>BenchMark XT</td>
<td>Membr.</td>
<td>neg., weak, strong</td>
</tr>
<tr>
<td>E-CDH</td>
<td>ECH-6</td>
<td>mouse</td>
<td>Cell Marque</td>
<td>1:10</td>
<td>BenchMark XT</td>
<td>Membr.</td>
<td>neg., weak, strong</td>
</tr>
<tr>
<td>EGFR</td>
<td>3C6</td>
<td>mouse</td>
<td>Ventana</td>
<td>prediluted</td>
<td>UView DAB</td>
<td>Membr.</td>
<td>neg., weak, strong</td>
</tr>
<tr>
<td>GLUT-1</td>
<td>AB1341</td>
<td>rabbit</td>
<td>Chemicon</td>
<td>1:1000</td>
<td>BenchMark XT</td>
<td>Membr.</td>
<td>neg., weak, strong</td>
</tr>
<tr>
<td>Ki-67</td>
<td>MIB-1</td>
<td>mouse</td>
<td>DAKO</td>
<td>1:20</td>
<td>BenchMark XT</td>
<td>Nucl.</td>
<td>neg., &lt;5%; &lt;10%; &gt;10%</td>
</tr>
<tr>
<td>p21</td>
<td>sc-397</td>
<td>rabbit</td>
<td>Santa Cruz</td>
<td>1:50</td>
<td>BenchMark XT</td>
<td>Nucl.</td>
<td>neg. &lt;10%; pos. &gt;10%</td>
</tr>
<tr>
<td>p27</td>
<td>sc-528</td>
<td>rabbit</td>
<td>Santa Cruz</td>
<td>1:30</td>
<td>BenchMark XT</td>
<td>Nucl.</td>
<td>neg. &lt;10%; pos. &gt;10% neg., weak, strong</td>
</tr>
<tr>
<td>p53</td>
<td>DO-7</td>
<td>mouse</td>
<td>DAKO</td>
<td>1:150</td>
<td>BenchMark XT</td>
<td>Nucl.</td>
<td>neg., &lt;5%; &lt;10%; &gt;10%</td>
</tr>
<tr>
<td>PAX2</td>
<td>Z-RX2</td>
<td>rabbit</td>
<td>Zymed</td>
<td>1:50</td>
<td>BondmaX</td>
<td>Nucl.</td>
<td>neg. &lt;10%; pos. &gt;10%</td>
</tr>
<tr>
<td>Periostin</td>
<td>RD1840-45100</td>
<td>rabbit</td>
<td>BioVendor</td>
<td>1:500</td>
<td>BondmaX</td>
<td>Tumor Stroma</td>
<td>neg., weak, moderate, strong neg., weak, moderate, strong</td>
</tr>
</tbody>
</table>
4.2.3 Data Description and Preprocessing

Of 831 ccRCC 527 observations (63.4%) had no missing values. For 87 observations one measurement was missing, 64 and 30 observations had 2 or 3 missing values, respectively. 123 observations contained more than 3 missing values and were ignored in the statistical analysis. For observations with 1, 2 or 3 missing values, multiple imputation was applied. Using this Monte Carlo technique, we simulated complete observations for each sample with a missing value. These were then combined with the observations without missing values to form complete datasets. Each of the generated datasets was analyzed as described below. The results were then combined to form a final model by averaging parameter estimates. A schematic overview of the process is given in Figure 4.2. The R package MICE (Multivariate Imputation by Chained Equations [Van Buuren and Oudshoorn, 2007]) was used to impute the data. After binarizing the 18 molecular parameters (-1 encoding low
and 1 encoding high expression, respectively), all missing values were imputed. We have chosen $m = 5$, which implies we have 5 datasets each containing 708 complete observations. The datasets were stored in matrices $X \in \mathbb{R}^{n \times p}$, where $n = 708$ and $p = 18$. Observation $i \in \{1, \ldots, n\}$ was stored in the $i$th row of $X$ and $X_{ij} \in \{-1, 1\}$ reads that molecular parameter $j$ of observation $i$ is not expressed or expressed respectively. For most observations also the survival time $t_i$ is known, for some the survival time is right censored.

### 4.2.4 Recursive Bootstrap Elimination of Variables in Cox Regression

To find molecular parameters and specific combinations which significantly influence survival, a Cox proportional hazard model was used to estimate a regression coefficient $\hat{\beta}$ corresponding to the full model with all molecular parameters and one or two clinical variables (tumor grade and/or stage). We then applied a recursive bootstrap elimination scheme for variable selection where $p$-values for individual components of $\beta$ (Efron and Tibshirani, 1993) were calculated. Components of $\beta$ with highest $p$-values were recursively eliminated, as long as the parameters were above the 5% significance level.

### 4.2.5 Log-Linear- and Graphical Model

We used log-linear- and graphical models to quantify the association between the molecular markers in a graph and to uncover possible dependence structures. We cross-classified the observations in a contingency table and expanded the probability of an observation of falling into a certain cell by a linear function in $\beta$: $\log p_i = X_i \beta$, where $i \in \{1, \ldots, 2^{18}\}$ for the $2^{18}$ different expression patterns (18 binary molecular parameters). Here, the $X_i$ are the rows of a design matrix whose columns consist of so-called main effects and interactions between variables. Here
we only considered first order interactions which is sufficient to derive a conditional independence graph. The construction of the design matrix has recently been described Dahinden et al. (2007). A hierarchical log-linear model directly corresponds to a graphical model, where the nodes correspond to the variables (molecular markers) and two nodes are connected whenever the corresponding first order interaction coefficient (i.e. a component of β) is non-zero in the log-linear model. In a graphical model a missing edge between two nodes implies that the corresponding variables are conditionally independent given all remaining variables (i.e. having accounted for the effect of all remaining variables). A missing edge does not necessarily imply independence between the variables, since they can still be marginally dependent, if there is a connection through other variables. On the other hand, an edge between two variables indicates immediate dependence which cannot be explained by all remaining variables.

We performed variable selection and parameter estimation for by an $\ell_1$-penalization approach (Dahinden et al., 2007), where we estimated β with the following formula:

$$\hat{\beta} = \arg\min_{\beta} \{-\log L(\beta) + \lambda \|\beta\|_1\}, \quad \text{where} \quad \|\beta\|_1 = \sum_{i=1}^{p} |\beta_i|,$$

under the additional constraint that all cell probabilities add up to 1: $\sum_{i} \exp (X\beta)_i = 1$. Here, $L(\beta)$ is the likelihood function assuming a multinomial distribution for the data with probabilities parameterized by the log-linear model. The value for the tuning parameter $\lambda$, which controls the sparsity of the graphical model, was chosen by ten-fold cross-validation of the likelihood. A large value of $\lambda$ implies a sparse solution with many components of $\hat{\beta}(\lambda)$ equal to zero. In other words, by estimating $\beta$ with this formula, parameter estimation and variable selection are done at the same time.

For our setting with 18 parameters, estimating $\beta$ with this formula was already computationally very intensive since there are $2^{18}$ different
states (different molecular marker constellations). We therefore used an adaptation of this formula which decomposes the problem into smaller subproblems where this formula can be used. The solutions of the subproblems were then combined in a suitable way to form the final solution. The exact description of the procedure can be found in Dahinden and Bühlmann (2008).

4.3 Results

4.3.1 Protein Expression Patterns, Tumor Phenotype and Survival

The number of interpretable cases decreased only slightly with the number of the cut TMA sections. Only between 5% (CCND1) and 15% (GLUT-1) of the 15 IHC analyses were non-informative because of the absence of tissue on the TMA, or lack of unequivocal tumor cells in the arrayed sample. Examples of ccRCC with strong nuclear (PAX2), membranous (CA9), and combined nuclear/cytoplasmic (p27) positivity are shown in Figure 4.3. The expression of p27 and PAX2 decreased strikingly with advanced pT category and higher differentiation grade. In contrast, epithelial-specific Periostin and phospho-S6 increased highly significantly with tumor stage and grade. The expression patterns of all 4 proteins were also highly associated with overall survival. Combined nuclear and cytoplasmic VHL as well as p-mTOR expression correlated inversely with tumor stage and grade but not with survival. CA9, CD10, p21, and stromal-specific Periostin positivity was associated with higher differentiation grade. From these only CA9 survival correlated with survival. No significant associations were found for CCND1, E-CDH, EGFR, GLUT1, p53 and PTEN expression neither with grade, stage or survival.
Figure 4.3: Scanned examples of immunohistochemically stained ccRCC using the SpotBrowser software version 2 (Alphelys, France). Strong protein expression of PAX2 (nuclear), CA9 (membranous), and p27 (nuclear and cytoplasmic) (magnification 10x).

4.3.2 Recursive Bootstrap Elimination

As displayed in Figure 4.4(a), cytoplasmic PTEN, nuclear p27, p-S6 expression and tumor grade were found to be significant in the survival analysis using the recursive bootstrap elimination in Cox regression with grade and all molecular markers. The lines correspond to the estimated 95% bootstrap confidence intervals for these coefficients. Based on our finding, we can deduct that for a variable with a positive regression coefficient, such as for p-S6, the risk for patients having expressed such a variable is increased compared to patients having not expressed that variable but otherwise sharing the same expression pattern (whereas for p27 and PTEN the risk is decreased). The same procedure was repeated for the clinical variables stage (Figure 4.4 (b)) and stage and grade (Figure 4.4 (c)). As shown in Figure 4.4 (b), a higher risk is predicted for patients with late tumor stage, absence of nuclear p27, EGFR, and CA9 as well as expression of p-S6. A similar result was obtained with stage
and grade (Figure 4.4 (c)). Using the molecular parameters nuclear p27, cytoplasmic PTEN, and p-S6 for intermediate grade 2 ccRCC, we were able to classify the Kaplan-Meyer survival curve for this group into two sub-populations. Compared to the overall estimated percentage of grade 2 patients (54%), 71% of the patients with the favorable biomarker constellation (PTEN and p27 expressed, P-S6 not expressed) but only 49% of the patients with the complementary molecular parameter combinations survive 5 years (Figure 4.5 (a)). Because the same data was used for estimating and evaluating the model fit, the difference of 22% is not an unbiased estimate. We conducted a cross-validation study to get an unbiased estimate of the difference of five year survival from the favorable and the remaining grade 2 patients. For this purpose the dataset was randomly divided into two parts consisting of two thirds and one third of the observations. Two thirds were used to fit the model and the remaining data was used to evaluate the model. This procedure was repeated ten times and the estimated error was averaged. The cross-validation analysis revealed an unbiased estimated difference in five-year survival of 15%. This is still a very substantial difference which is due to the protein expression patterns only.

The same procedure was repeated for tumor stages pT2, pT3 and all molecular markers (Figure 4.5 (b)). According to Figure 4.4, the relevant markers were nuclear p27, CA9, p-S6 and EGFR. But asking for this specific biomarker combination, the upper curves in Figure 4.5 (b) became insignificant (i.e. not statistically distinguishable from the lower corresponding curve) with e.g. only 13 pT2 tumors having the favorable combination. However, when only nuclear p27 and CA9 were included, a balanced classification into favorable and non-favorable combinations was obtained. Each of the groups of pT2 and pT3 tumors was divided into two curves. The upper curves belong to ccRCC with nuclear p27 and CA9 positive and the lower curves to the complementary constellation. The 5-year survival difference was 73.3% versus 51.4% for pT2 tumors and 51.3% versus 39.0% for pT3 tumors for the favorable and the non-favorable combination respectively. The cross-validated differ-
ence was estimated to be 11% and 3% for pT2 and pT3 tumors, with the restriction that the favorable constellation has to consist at least of 20 observations.

4.3.3 Graphical Log-Linear Model

The graphical model associated with the estimated main effects and interactions vector $\beta$ is displayed in Figure 4.6. The width of the edge between $i$ and $j$ corresponds to the absolute value of the corresponding estimated coefficient $|\beta_{ij}|$.

Figure 4.6 shows that the molecular parameters chosen for modeling survival times (nuclear p27, p-S6, cytoplasmic PTEN, and CA9) are not directly linked to each other. There is no short path with thick edges from one of these parameters to the other implying that (conditionally) independent variables have been selected in the survival analysis. On the other hand, the most striking dependences exist between cytoplasmic and nuclear p27, PAX2 and p21, as well as PAX2 and p-mTOR. Somewhat weaker dependences are seen between cytoplasmic and nuclear PTEN, cytoplasmic PTEN and VHL, PAX2 and CCND1, PAX2 and nuclear PTEN, HIF targets CA9 and GLUT1 but also between GLUT1 and CD10, CA9 and E-Cadherin (inverse). Notably, no obvious link is shown for VHL and HIF targets and only weak dependences are yielded between PTEN, m-TOR and p-S6. Mathematically, we cannot directly conclude a causal relationship between the biomarkers, but we can deduce a dependence which might point to a relationship where one biomarker directly influences the other.
4.4 Discussion

The high frequency of VHL mutations found in sporadic ccRCC (Banks et al., 2006) indicates that loss or altered function of pVHL strongly contributes to tumor initiation and, based on its function as negative HIF regulator, most likely to metastasis. However, results of cytogenetic (Moch et al., 1996), molecular and functional studies (Struckmann et al., 2004; Frew et al., 2008; Thoma et al., 2007) suggest that the dysregulation of additional factors besides pVHL are required to promote ccRCC progression. In previous studies the prognostic value of numerous tumor markers belonging to pVHL dependent and independent pathways was evaluated in ccRCC. In most of these studies the expression of one or only few proteins was analyzed in a rather limited number of cases. As a consequence, in ccRCC the interrelations of different molecular pathways on clinicopathological parameters have hardly been considered so far. In an attempt to address this problem, we analyzed the expression data of 15 functionally well characterized proteins on a large number of arrayed ccRCC specimens using advanced statistical modeling, i.e. Cox regression with recursive Bootstrap elimination and graphical log-linear modeling. The expression frequencies and associations with clinicopathological parameters obtained from our TMA analyses were mostly concordant with data from the literature (Leibovich et al., 2007), CCND1 (Aaltomaa et al., 1999; Lin et al., 1998), p53 (Bot et al., 1994; Zigeuner et al., 2004; Ljungberg et al., 2001), p27 (Hedberg et al., 2003; Migita et al., 2002), VHL (Schraml et al., 2003), GLUT1 (Ozcan et al., 2007), CD10 (Pan et al., 2004), E-Cadherin (Katagiri et al., 1995), p-S6 (Pantuck et al., 2007), PTEN (Kim et al., 2005) and demonstrated that tiny amounts of tissue material are sufficient to reliably evaluate the expression of immunostained proteins in a high number of tumors. Discrepant results were observed for EGFR, p-mTOR, and PAX2 from which contradictory data in the literature also exist (Cohen et al., 2007; Langner et al., 2004a; Campbell et al., 2008; Gnarra and Dressler, 1995). This problem might be explained by the use of different antibodies, pro-
tocols and/or the relative low numbers of tumors analyzed in previous studies. No published data were found for p21 expression in ccRCC. The frequency of p21 positive tumors was low (23%) and only slightly higher than that observed for p53 expressing ccRCC (20%). Compared to early stage tumors, the fraction of p21 positive ccRCC decreased significantly in late stage tumors suggesting that loss of p21 expression is critical for aggressive tumor behaviour. The finding that in ccRCC epithelial rather than stroma-specific expression of Periostin correlated highly significantly with nuclear differentiation grade, tumor stage and survival is novel and indicates the protein’s importance for metastatic processes as it was described in breast cancer (Shao et al., 2004).

As an extension of a Cox proportional-hazard model, which is commonly used to analyze the effect of risk factors on survival, a recursive Bootstrap approach was used for selecting important molecular parameters. The advantage of this method is that the influence of correlated non-informative predictor variables is reduced. Recursive bootstrapping with TMA expression data from 15 proteins enabled us to refine tumor grading and staging. ccRCC of histologic grade 2 (Thoenes) with nuclear p27, cytoplasmic PTEN and inactive, non-phosphorylated S6 had an outcome that was nearly comparable with that of grade 1 tumors (80% versus 71%). In contrast, the 5-year survival rate for the remaining grade 2 tumors was below 50%. The association between favorable outcome and combined expression of p27 and PTEN suggests that the inactivation of both or at least one of these proteins is a critical step towards tumor progression.

A lack of p27 and PTEN was also described for most advanced prostate cancers (Cairns et al., 1997; Macri and Loda, 1998). In mice it was shown that concomitant inactivation of PTEN and p27 are required to accelerate spontaneous neoplastic transformation (Di Cristofano et al., 2001). It was therefore concluded that the combined tumor-suppressive activity of PTEN and p27 is crucial for the control of cell-cycle progression. Our data imply a mechanism for ccRCC progression
that is similar to that described in prostate cancer. The identification of phosphorylated ribosomal protein S6, a downstream target of activated mTOR, as potential risk factor is highly likely a consequence of PTEN and p27 loss and confirms the importance of an activated AKT/mTOR pathway in a considerable subset of ccRCC patients.

In contrast to prostate cancer in which p27 expression appears to be dependent on PTEN activity (Di Cristofano et al., 2001), our estimated graphical model shows only a weak association. Moreover, the significant association between late stage, high grade tumors and absence of nuclear p27 expression suggests that in ccRCC the down regulation of p27 rather than the loss of PTEN probably has more severe consequences for patient’s survival. It also indicates that p27 expression might only be partially dependent on PTEN’s functional integrity. Blocking the NOTCH1 receptor, a key player of the Notch signaling pathway in human ccRCC cell lines, was accompanied by elevated levels of p27 and p21 (Sjolund et al., 2008). Thus, for knocking down p27 the synergized activation of both the NOTCH1 and PI3K/AKT pathways might be necessary to drive ccRCC progression.

Because the Cyclin dependent kinase inhibitor p27 exhibits its function in the cell nucleus all recent immunochemical studies of ccRCC tissues have focused on nuclear p27 expression (Hedberg et al., 2003; Migita et al., 2002; Langner et al., 2004b). It is known that ubiquitinylation and degradation (Carrano et al., 1999) as well as sequestration (Shin et al., 2002) of p27 occurs mainly in the cytoplasm. Our graphical log-linear model shows a tight association between nuclear and cytoplasmic expression of p27. We therefore speculated that both nuclear and cytoplasmic presence is required for nuclear-cytoplasmic trafficking thus ensuring proper function of p27. In fact, the number of tumors that were p27 negative in one or both cellular compartments markedly increased with tumor stage and grade and also had a worse outcome (data not shown). These results imply that alteration of subcellular p27 trafficking is of potential relevance for the biological behavior of
CA9 is upregulated in the vast majority of ccRCC (Leibovich et al., 2007; Bui et al., 2003; Atkins et al., 2005). Notably, the expression of HIF targets CA9 and GLUT1 correlated well with CD10 positivity and suggests that CD10 might also be regulated by HIF. CA9’s role as prognostic and predictive marker for ccRCC is contradictory. Our finding that patients with positive CA9 tumors had a better outcome compared to those which were CA9 negative confirmed the results of a recent study in which 730 ccRCC were analyzed (Leibovich et al., 2007). Although loss of pVHL function considerably supports HIF-mediated transcription of CA9, it is unclear why in metastasizing ccRCC CA9 is increasingly lacking. Carbonic anhydrases, such as CA9, catalyze intracellular carbon dioxide, resulting from buffering the lactazidosis by bicarbonate, into carbonic acid outside the cells which importantly contributes to the acidic micro-environment in tumors (Svastova et al., 2004). Based on a previous study which showed considerable genetic heterogeneity between organ-confined and metastatic RCC (Bissig et al., 1999), it is tempting to speculate that in advanced tumors HIF-independent carbonic anhydrases other than CA9 become activated.

Strong associations were seen between PAX2, activated mTOR, p21, CCND1, and nuclear PTEN. PAX2 belongs to the PAX gene family of developmental transcription factors. Induction of nephrogenesis by the ureter is accompanied by the upregulation of PAX2 followed by cell proliferation of the ureteric bud. Transactivation of the promoter of the tumor suppressor gene WT1 leads to repression of PAX2 and the differentiation of the nephrogenic mesenchyme (Ryan et al., 1995). As PAX2 is absent in normal renal tubular epithelial cells but expressed in many ccRCC (Mazal et al., 2005; Daniel et al., 2001; Memeo et al., 2007), the upregulation of PAX2 might be caused by cancer-linked hypomethylation of its promoter as it was already shown in endometrial cancer (Wu et al., 2005). In humans with kidney malformations PAX2 is highly expressed in cystic and hyperproliferative epithelial cells (Stayner et al.,
4.4. Discussion

2006; Winyard et al., 1996). Cysts also occur frequently in hereditary and sporadic ccRCC which are thought to be a result of uncontrolled cell proliferation and represent tumor precursor lesions (Neumann et al., 1998). The ability to transcriptionally repress the tumor suppressor p53 (Stuart et al., 1995), which may indirectly targets its mediator p21, and its joint action with CCND1 and mTOR makes PAX2 a possibly important promoter of cyst formation in ccRCC. A potential interaction between PAX2 and nuclear PTEN still remains to be elucidated.

Our TMA data consisted of measurements of 18 different molecular parameters (15 proteins including nuclear and cytoplasmic expression for VHL, p27, and PTEN) each of which were coded either with 1 (expressed) or -1 (not expressed) resulting in possible -1/1-combinations. For a saturated model in a regression analysis in which all possible interactions between the markers are considered we would have to fit 262144 parameters. Therefore, complexity was reduced by simplifying the model such as not allowing for interactions, as it was done in our survival analysis. In a regression setting, as a rough rule of thumb, at least 10 observations are needed to reasonably fit one non-zero regression coefficient (depending on the signal to noise ratio and the design of the covariates). As consequence, for an expression study of say 10 or more relevant biomarkers (but we do not know the number of relevant biomarkers beforehand), a minimum of 100 specimens are required. The use of large TMAs containing several hundred biopsies is a prerequisite to reliably measure associations between clinical parameters and expression patterns of many proteins.

In summary, the use of recursive Bootstrap elimination as well as graphical log-linear modeling is ideally suited for comprehensive TMA data analysis. Here we show that ccRCC progression is probably the result of a sequential switch-off of different tumor suppressive programs beginning with pVHL and followed by PTEN and p27. Our data also indicate that in subsets of ccRCC HIF, PI3K-AKT-mTOR, and p27 controlled pathway(s) are activated in a separate or concerted manner.
This might explain why the treatment of advanced ccRCC with kinase inhibitors sunitinib malate or sorafenib (Motzer and Bukowski, 2006), temsirolimus (Hudes, 2007) or CA9 antibodies (Davis et al., 2007) is still rather unsuccessful.
Figure 4.4: Results for recursive bootstrap procedures. On the left: One step before the remaining biomarkers are all considered significant. On the right: Significant biomarkers

(a) Results for grade and all molecular markers. Significant variables are grade 1, grade 3, nuclear p27, cytoplasmic PTEN and p-S6.

(b) Results for stage and all molecular markers. Significant variables are stage 1, stage 3, nuclear p27, p-S6, CA9 and EGFR.

(c) Results for grade&stage and all molecular markers. Significant variables are grade 1, grade 3, stage 1, stage 3, nuclear p27, CA9 and p-S6.
Figure 4.5: Estimated survival time in dependence of pathological and molecular parameters. The dotted vertical lines indicate five year survival.

(a) The upper curve for grade 2 represents patients with PTEN (cytoplasmic), p27 (nuclear) positive and/or p-S6 negative ccRCC. The logrank test comparing the two grade 2 curves yields a $p$-value of 0.0028.

(b) The upper curves for stage 2 and stage 3 represent patients with p27 (nuclear) and CA9 positive ccRCC. The logrank test comparing the two grade 2 curves, yields a $p$-value of 0.089 and for the two grade 3 curves 0.0074.
Figure 4.6: Estimated graphical log-linear model for 15 analyzed proteins including nuclear and cytoplasmic expression of PTEN, p27, and VHL (18 molecular parameters).
Chapter 5

Appendix

5.A Technical Details of \( \ell_1 \)-Regularization Algorithm

We note that if \( \beta \) is a minimum of \( g \), then \( \beta_A \) is a minimum of \( g_A \).

In our application with single-gene libraries, all factors have two levels only, which allows to construct an efficient algorithm. Since the gradient

\[
\nabla \left[ -l(\beta) + \sum_{i=1}^{m} \exp (\mu_i) \right] = -X^t \cdot \left( \frac{n}{n} - \exp (X\beta) \right),
\]

where \( \exp(X\beta) \) is understood as the componentwise exponential function, it follows that for a minimum \( \beta_A \) of \( g_A \), the following equation holds:

\[
\nabla g_A(\beta_A) = -X_A^t \cdot \left( \frac{n}{n} - \exp (X_A\beta) \right) + (0, \text{sign}(\beta_A))^t \cdot \lambda = 0 \quad (5.A.1)
\]
Without loss of generality, we can restrict ourselves to the subspace $\beta \in \mathbb{R}^{-} \times \mathbb{R}^{m-1}$, because the constraint (2.3.5) can only be satisfied for $\beta_0 < 0$ as is proved in the following Lemma 2. Therefore $\beta_0 \in \mathcal{A}$.

**Lemma 2.** $\beta_0 < 0$ for a minimum of $g(\beta)$ for all $\lambda \in \mathbb{R}^+$.\[\]

*Proof.*

$log(p) = X\beta < 0$ which yields $(1, \ldots, 1)X\beta = m\beta_0 < 0$,

which implies that $\beta_0 < 0$. This holds because $(1, \ldots, 1)$ is orthogonal to all columns of $X$ except for the first one. \qed

Additionally for $\beta$ being a minimum, a necessary condition is:

$$|(X^t \cdot (\frac{n}{n} - \exp(X\beta)))_j| < \lambda, \forall j \notin \mathcal{A}. \quad (5.A.2)$$

Conditions (5.A.1) and (5.A.2) are sufficient for $\beta$ being a minimum of (2.4.10). To find the $\beta$'s that solve these equations for an array of values for $\lambda$, we set up a so-called path following algorithm. The idea is to start from an optimal solution $\beta^{\lambda_0}$ for $\lambda_0$, and follow the path for decreasing $\lambda$, using a second-order approximation for $\beta_{\mathcal{A}}$. In the following, we restrict ourselves to the currently active set $\mathcal{A}$, omitting the index $\mathcal{A}$. It then holds:

$$\nabla g(\beta_{t+1}, \lambda_{t+1}) = 0 \approx \nabla g(\beta_t, \lambda_{t+1}) + \nabla^2 g(\beta_t, \lambda_{t+1}) \delta \beta.$$  

$$\downarrow$$

$$\delta \beta = -\nabla^2 g(\beta_t, \lambda_{t+1})^{-1} \nabla g(\beta_t, \lambda_{t+1}). \quad (5.A.3)$$

The algorithm tries to follow the optimal path as close as possible. At each step, it aims to meet the conditions (5.A.1) and (5.A.2). In step (3.2), the active set $\mathcal{A}$ is identified, which forces $\hat{\beta}$ to meet the condition (5.A.2). In step (3.3), a Newton step as described in (5.A.3) is performed. Starting from a solution which meets condition (5.A.1), the new $\hat{\beta}^\lambda$ approximately meets (5.A.1) again.
5.B Details of Model Selection Results Using MCMC

The interaction vectors \( \hat{\beta} \) estimated with MCMC methods, are graphically displayed in Figure 5.1 and Figure 5.2.

5.C Definition of the Index

To define the index of a complete subset of a graph, we first have to define a connectivity component. Two vertices are in the same connectivity component if and only if there is a path between them. If a graph only consists of one connectivity component, i.e. there is a path between any two nodes in the graph, the graph is called connected. Let \( G = (V, E) \) be a connected graph, and \( S \) be a complete subset of \( V \). The pieces of \( G \) relative to \( S \) are defined as follows: Remove \( S \) from \( G \) and form the subgraph with vertices in \( V \setminus S \) and edges which are the same as those in \( E \) restricted to \( V \setminus S \). This subgraph may have one or more connectivity components \( A_t \). Now, rejoin the set \( S \) to the subgraphs \( A_t \) and call the resulting subgraphs \( G_t \): \( G_t \) has vertex set \( A_t \cup S \) and edges those in \( E \) restricted to \( A_t \cup S \). \( G_t \) are now called the pieces of \( G \) relative to \( S \) and we define the index \( \nu(S) \) as follows:

\[
\nu(S) = \text{the } \# \text{ of pieces of } G \text{ relative to } S \text{ in which } S \text{ is not a clique} - 1
\]

Let’s go back to our example in Figure 3.3 with \( S = \{2\} \). If we remove node 2, on the left side all that remains are the three connectivity components consisting of the single nodes \( \{1\}, \{3\} \) and \( \{4\} \). If we rejoin node 2, then then we get the three pieces of \( G \) relative to node \( \{2\} \). And obviously in none of them, the node \( \{2\} \) is a clique and therefore the index is \( \nu(S) = 3 - 1 = 2 \) as we got by the intuitive definition given in Section 3.2.5. If we remove the node 2 on the right side of Figure 3.3, then we get 2 connectivity components and by rejoining the node
Figure 5.1: Interaction vectors $\hat{\beta}$ for the gene itpr1 estimated by the hierarchical MCMC estimator with $\sigma_a^2 = 1$ for all $a$. Note the close similarity between this interaction pattern and the one from the step $\ell_1$-regularization estimator in Figure 2.1.

{2} we get 2 pieces of $G$ relative to $S$. Again, in none of them, {2} is a clique and therefore the index is 1.
Figure 5.2: Interaction vectors $\hat{\beta}$ for the gene itpr1 estimated by the hierarchical MCMC estimator with $\sigma^2_a = 2$ for all $a$. This interaction pattern remarkably differs from the one in Figure 2 in the main text.
Bibliography


Sjolund, J., Johansson, M., Manna, S., Norin, C., Pietras, A., Beckman, S., Nilsson, E., Ljungberg, B. and
Bibliography


Chapter 6

Curriculum Vitae

I was born on May 16, 1978 in Zug. After attending primary school from 1985 to 1991 in Rotkreuz (ZG), I spent seven years at the Kantonsschule Zug and obtained a “Matura, Typus C” in 1998.

I started my studies in Mathematics at ETH Zurich in October 1998 and graduated in March 2004 with a diploma thesis on “Schätzung des Vorhersagefehlers und Anwendungen auf Genexpressionsdaten”.

In November 2004 I enrolled as a PhD student and worked since as a teaching assistant at the Seminar für Statistik (SfS).