Technische Universität München

Department of Mathematics

Diploma thesis
(Corrected version)

Statistical analysis of asymmetric cell fate choice on cellular genealogies

Matthias Sachs

Supervisor: Prof. Dr. Dr. Fabian Theis

Advisor: Michael Strasser M.Sc., Dr. Carsten Marr

Submission Date: September 14th, 2012

I assure the single handed composition of this Diploma thesis only supported by declared resources.

Garching, September 14th, 2012

Abstract

Asymmetric cell division is presumed to be an intrinsic mechanism which keeps the population of self-renewal capable hematopoietic stem cells constant within the blood system. Cellular genealogies originating from time-lapse microscopy experiments potentially encode information about cellular development and division patterns of the observed cell colonies. In this thesis we develop statistical methods which are suitable to find evidence of asymmetric division processes in cellular genealogies. We focus on the modeling of dichotomous fates of cell siblings. We apply our models on genealogies with annotated cell death originating from in vitro experiments with mice hematopoietic stem cells under TGF-β1 positive culture conditions. In our analysis we use cell death as an indicator for the loss of self-renewal capacity. We could not find clear evidence of asymmetric cell division.

Zusammenfassung

Es wird angenommen, dass sich die Population von zu Selbsterneuerung fähigen hämatopoetische Stammzellen im Blutsystem durch asymmetrische Zellteilung konstant hält. Zellgenealogien, die gewöhnlich mithilfe von Videomikroskopieaufnahmen erstellt werden, enthalten Informationen über Zellentwicklungs und Zellteilungsprozesse der beobachteten Zellkolonien. In dieser Arbeit entwickeln wir statistische Verfahren, die auf solche Zellgenealogien angewendet werden können und mit deren Hilfe Evidenz für asymmetrische Zellteilung gezeigt werden kann. Wir beschränken uns auf die Modellierung von dichotomischen *cell fates* von Geschwisterzellen. Die Modelle wenden wir auf Zellgenealogien an, die aus in vitro Experimenten mit hämatopoetischen Mäuse-Stammzellen unter TGF-β1 positiven Kulturbedingungen stammen. Wir verwenden dabei Zelltod als Indikator für den Verlust der Selbsterneuerungsfähigkeit der betreffenden Zellen. Wir konnten keinen eindeutigen Hinweis auf asymmetrische Zellteilung finden.

Acknowledgments

First, I want to thank Michael Strasser for his great support and for his time and effort in supervising my work.

Big thanks also to Carsten Marr for the great supervision and for initially proposing the topic of this thesis.

Thanks to Dirk Loeffler for providing me the experimental data analyzed in this thesis.

I would also like to thank all CMB members for providing a very nice atmosphere and for answering all my queries.

Finally I am very grateful to Fabian Theis for his final comments on my thesis and in particular for giving me the opportunity to write this thesis at his group.

Contents

1 Introduction **11**

1.1 Haematopoesis 11

1.2 Asymmetric cell division 12

1.2.1 Cell fate 12

1.2.2 Formal definition of asymmetric cell division 12

1.2.3 Functional evidence for asymmetric cell division in model organisms . . . 12

1.2.4 Asymmetric cell division in haematopoiesis 13

1.2.5 Statistical evidence for asymmetric cell division 13

1.3 Content of this thesis 14

1.4 Structure of this thesis 14

2 Methods **15**

2.1 Mathematical preliminaries 15

2.1.1 Probability theory 15

2.1.2 Statistics 17

2.1.3 Graph theory 20

2.2 Mathematical modeling of cellular genealogies 20

2.2.1 Graph theoretic description of cellular genealogies 20

2.2.2 Stochastic description of cell siblings 21

2.3 Parameter inference on binary pairings 26

2.3.1 Basic model 27

2.3.2 Treatment-Control model 33

2.3.3 Special mixture model 37

2.4 Mixture effects 40

3 Data **41**

3.1 Experimental setup description 41

3.1.1 Experiments 41

3.1.2 Tracking and data representation 42

4 Results **47**

4.1 Model 1 (application of the basic model) 48

4.2 Model 2 (application of the treatment-control model) 51

4.3 Model 3 (application of the special mixture model) 54

5 Discussion **59**

5.1 Fate determining process 59

5.2 Composite model 62

5.3 Discussion of inferred results 62

5.3.1 Contamination of experimental data 63

5.3.2 Missing vertical dependencies 63

5.3.3 Mixture effects 63

6 Outlook **67**

Chapter 1

Introduction

1.1 Haematopoesis

The blood system in mammalian organisms undergoes a continuous replacement process of cells. Every second millions of blood cells die and are replaced by the right amount of new cells [16]. The underlying process, which describes the regeneration and formation of new blood cells is referred to as haematopoesis. In the classical understanding of haematopoesis, the cells of the blood system are hierarchically ordered according to their differentiation potential. Hematopoietic stem cells (HSCs) have the highest differentiation potential within in this hierarchy. That means, that progenitors of this cells have the capacity to differentiate into any type of mature blood cells. During this differentiation process cells are assumed to undergo subsequent levels of differentiation. Cells of this intermediate levels are vastly classified as multipotent progenitors (MPP), oligopotent progenitors, and unipotent progenitors. Unipotent progenitors merely have the potential to differentiate to one specific mature cell type, whereas oligopotent progenitors can differentiate to a specific group of unipotent cell types. Multi potent progenitors have the same differentiation potential as haematopoietic stem cells. However, they can maintain their differentiation potential only for a limited number of cell divisions, as well as all other cell types with differentiation potential below haematopoietic stem cell level. Hematopoietic stem cells take on a unique role in the haematopoietic system, since they are the only cells which can keep there differentiation potential over an unlimited number of divisions. This property is referred to as self-renewal capacity. In healthy organisms the pool of HSCs stays approximately constant over the whole lifetime [16]. The presence of HSCs in the blood system is crucial for the survival an organism. Since all other cells in the haematopoietic system loose their differentiation potential after a limited number of divisions, these cells need to be replaced by differentiating HSCs. Until now it is poorly understand by what mechanism a balance is kept between HSCs and more mature cell types. One possible mechanism is extrinsic control like cell to cell communication and cell-niche interactions [5]. However, there are also attempts to explain this balancing by

intrinsic differentiation patterns of haematopoietic stem cells. In particular, it is presumed that the balance could be maintained by asymmetric cell division of the haemeatopoietic stem cells.

1.2 Asymmetric cell division

The term of asymmetric cell division is not consistently used in literature. In order to avoid misunderstandings we first formally specify this term as well as the related concepts of asymmetric segregation and asymmetric cell fate before discussing the presumed function of asymmetric cell division in hematopoiesis.

1.2.1 Cell fate

Any feature assigned to a cell either by direct observation of this cell or retrospectively by observation of its progenitors can be referred to as the fate of this cell. For example, we will later refer to haematopoietic stem cells in terms of their survival fate (cell died/died not), or self-renewal fate (cell lost-self renewal capacity/maintained self renewal capacity).

1.2.2 Formal definition of asymmetric cell division

Cell division can produce two daughter cells of identical or distinct fates. Identical fates of the daughter cells are referred to as symmetric fates. Distinct fates of the daughter cells are called asymmetric fates. During cell division it can happen, that certain components (e.g. proteins or even organelles) are distributed asymmetrically on the daughter cells. We refer to this physical process as asymmetric segregation. In case there is a causal relation between asymmetric segregation and asymmetric fate, we refer to the corresponding division as an asymmetric division and the asymmetrically distributed components causing this asymmetry are called fate determinants.

1.2.3 Functional evidence for asymmetric cell division in model organisms

The difficult part in the identification of asymmetric cell division is to show evidence for the causal relation between the asymmetric segregation and asymmetric fates. In well understood model organisms like *Drosophila melanogaster* or *Caenorhabditis elegans* functional evidence for this causality was shown by identification of the underlying biological processes. For example the division of neuroblast progenitor cells during the early development of the central nervous system in *Drosophila melanogaster* has been shown to be asymmetric. The basic process, which causes asymmetric fates in this division has been identified as the asymmetric segregation of Numb proteins along the daughter cells [12]. Numb proteins are synthesized in neuroblast cells, but suppress the Notch signaling pathway of the cell in case of high concentration. The Notch

signaling pathway is essential for the cell to stay in undifferentiated state. Therefore the Numb proteins function as the cell fate determinants.

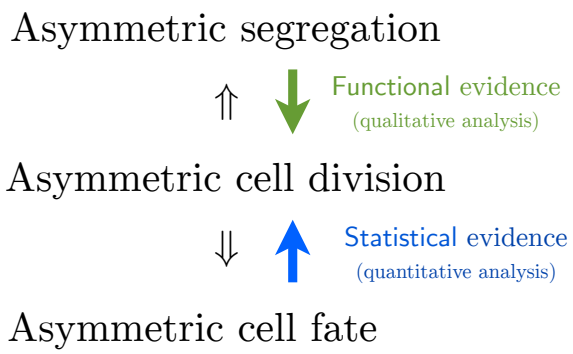


Figure 1.1: Inference methods for asymmetric cell division: By definition asymmetric cell division implies asymmetric fates of the corresponding daughter cells as well as asymmetric segregation. In case of observed asymmetric segregation functional evidence for asymmetric cell division can be shown by identification of the underlying biological processes. In case of observed asymmetric cell fates statistical evidence for asymmetric cell division can be shown by quantitative analysis of cellular genealogy data.

1.2.4 Asymmetric cell division in haematopoiesis

Asymmetric cell division has been proposed as an intrinsic mechanism which could sustain a constant pool of HSCs in the blood system [16]. The idea behind this assumption is that by asymmetric cell division of haematopoietic stem cells one daughter cell always remains it self-renewal capacity, while the other one looses it. Whereby a perfect balance of differentiating cells and HSCs would be assured by this intrinsic mechanism. Various studies have been made in this field [17], which have shown the existence of asymmetric segregation processes of specific components during cell division in the haematopoietic system, but could not proof any functional evidence for this asymmetries, i.e. it is not known, if these asymmetries induced different fates of the daughter cells.

1.2.5 Statistical evidence for asymmetric cell division

Instead of showing functional evidence for asymmetric cell division by analysis of the underlying biological mechanism an alternative approach is to show statistical evidence for asymmetric cell division by a quantitative analysis of differentiation patterns in cellular genealogies. The basic idea of this approach is that asymmetric cell divisions cause statistical dependencies between cell siblings, which potentially are encoded in the structure of the cellular genealogies. Work in

this field has been done in the past by [20, 6]. Glauche et al. proposed to measure the dependency between sibling cells by their mutual information. Global measures for the symmetry or asymmetry of trees have been extensively studied in context of the analysis of phylogenetic trees [11, 1, 10, 18]. However as described in [6] these measures have only weak sensibility for asymmetries in the genealogies caused by only a few asymmetric cell division events and are therefore presumably not usefull in order to find evidence for asymmetric cell divisions.

1.3 Content of this thesis

In this thesis we pursue the approach of a quantitative analysis of cellular genealogy data. We focus on the modeling of dichotomous fates of sibling cells. This approach is conceptual closely related to the analysis of genotype frequencies in populations. In particular the core concept in this field, the so called Hard-Weinberg principle [9], will also take a central role within this thesis. We put particular attention on a well-founded theoretic basis of the methods, which we developed in context of this thesis. We suppose that this aspect distinguish our work from most other attempts in this field. Furthermore it makes generalization of our methods possible. This is, as we will see in our application results, necessary in order to deal with the complexity of the processes encoded in cellular genealogies. The latent variable structures described in our methods are motivated by the experimental data, which was provided by Dirk Loeffler from the Stem Cell Dynamic group at Helmholtz-Zentrum, München. He performed several time-lapse microscopy experiments with hemeatopoetic stem cells, where he added the cytokine TGF-β1 in the culture condition. The assumed effect of TGF-β1 is, that it increases the mortality of committed progenitor cells but does not significantly affect the mortality of HSCs. Therefore it functions as an indicator for the self-renewal fate of cells. The analysis on this data is done in context of the question, whether HSCs divide asymmetrically. So far, our statistical analysis yielded no evidence for this hypothesis.

1.4 Structure of this thesis

This thesis comprises four parts. In chapter 2 we revise some basic stochastic and statistical concepts and describe the main generic models developed in context of this thesis. In chapter 3, we give a general overview on the experimental data used for our analysis. In chapter 4 we show how our models can be applied on the experimental data and infer the parametrization of these models. In chapter 5 we try to interpret the inferred results in terms of the question of possible asymmetric cell division processes encoded in the genealogies. In the chapter 6 we give a brief description of a possible generalization of our methodical approach.

Chapter 2

Methods

2.1 Mathematical preliminaries

In this section, we present some basics from probability theory, statistics and graph theory, which we will use in the later sections.

2.1.1 Probability theory

Let Ω be a non empty set.

Definition 2.1.1 (σ-Algebra)**.** *Let $\mathcal{P}(\Omega)$ be the power set of Ω and let $\mathfrak{A}$ be a subset of $P(\Omega)$. $\mathfrak{A}$ is called a* **σ-Algebra** *if*

(i) $\emptyset \in \mathfrak{A}$, $\Omega \in \mathfrak{A}$

(ii) $A \in \Omega \Rightarrow \Omega \setminus A \in \mathfrak{A}$

(iii) $(A_n)_{n\in\mathbb{N}} \subseteq \mathfrak{A} \Rightarrow \bigcup_{n\in\mathbb{N}} A_n$

Let $\mathcal{E} \subset \mathcal{P}(\Omega)$, then $\sigma(\mathcal{E})$ denotes the smallest σ-algebra on Ω, which contains the family $\mathcal{E}$.

Definition 2.1.2 (Probability measure)**.** *Let $\mathfrak{A}$ be a σ-Algebra. A map $\mu : \mathfrak{A} \to [0, \infty]$ is called a* **measure** *if*

(i) $\mu(\emptyset) = 0$

(ii) For $(A_n)_{n\in\mathbb{N}} \subseteq \mathfrak{A}$ *pairwise disjoint* $\Rightarrow \mu\big(\bigcup_{n\in\mathbb{N}} A_n\big) = \sum_{n\in\mathbb{N}} \mu(A_n)$

If $\mu(\Omega) = 1$, we refer to μ as a **probability measure**.

Let $\mathfrak{A}$ be a σ-algebra on Ω and μ a measure defined on $\mathfrak{A}$. A 2-Tuple of the form $(\Omega, \mathfrak{A})$ is called a **measurable space**. *A 3-tuple $(\Omega, \mathfrak{A}, \mu)$ is referred to as a* **measure space**. *In particular, if P is a probability measure, $(\Omega, \mathfrak{A}, P)$ is called a* **probability space** *and subsets $A \in \mathfrak{A}$ are called* **events**.

Definition 2.1.3 (Random variable)**.** *Let $(\mathcal{E}, \mathcal{F})$ be a measurable space and $(\Omega, \mathfrak{A}, P)$ a probability space. A map $X : \Omega \to \mathcal{E}$ is called a random variable, if it is measurable, i.e.:*

$$\forall A \in \mathcal{F} : X^{-1}(A) \in \mathfrak{A},$$

where

$$X^{-1}(A) := \{\omega \in \Omega : X(w) \in A\}.$$

In the above setting, a set $A \in \mathcal{F}$ is referred to as a **realization** or **observation** of the random variable X. The Random variable X induces a probability measure P^X on $(\mathcal{E}, \mathcal{F})$, where

$$P^X(A) := P(X^{-1}(A)) = P(\{\omega \in \Omega : X(\omega) \in A\}).$$

Let Y be a random variable, which maps onto the same measure space as X. We say that X and Y are identically distributed, if $P^Y = P^X$. It is customary to use the notation

$$P(\{X \in A\}) := P^X(A)$$

and

$$P(X = a) := P^X(\{a\})$$

for sets consisting of only one element.

Definition 2.1.4 (Density function)**.** *Given a measure space $(\Omega, \mathfrak{A}, \mu)$, let $f \to [0, \infty)$ be a non negative $\mathfrak{A}$-measurable map. Then*

$$\mu_f(A) := \int_A f d\mu, A \in \mathfrak{A}$$

defines a measure on $(\Omega, \mathfrak{A})$, which is referred to as the measure with **density** *f with respect to μ. In case μ_f is a probability measure, then f is referred to as a* **probability density**.

In most practical cases, there is a measure μ' defined on $(\mathcal{E}, \mathcal{F})$. In case $\mathcal{E} \subset \mathbb{R}^n, n \in \mathbb{N}$, this is usually the restriction of the Lebesque-Borel measure on $\mathcal{E}$. In case the cardinality of $\mathcal{E}$ is at most infinit countable, this is usually the counting measure. In case that X is a multidimensional real valued random variable and $P^X = \lambda_f$ with respect to the corresponding product measure of the Lebesque-Borel measure λ, then we denote this by $X \sim f$. Analogously, if X is a discrete random variable and f is a density function with respect to the counting measure on $\mathcal{E}$ so that $P^X = \lambda_f$, then we write $X \sim f$.

Note! Not every probability measure can be represented by a probability density. A sufficient condition for the existence of a probability density is given by the Radon-Nykodym theorem. [14]

2.1.2 Statistics

A finite sequence $X = X_1 \dots X_n$ of independent and identical (i.i.d) distributed random variables is called a **random sample**. If $X_i \sim f, i = 1 \dots n$, then we refer to f as the underlying density of X. All random variables mentioned in this section are assumed to be real valued or discrete.

Definition 2.1.5. *Let $X = X_1 \dots X_n$ be a random sample parameterized by $\theta \in \Theta$ i.e. $X_i \sim f_\theta, i = 1 \dots n$. Let $\alpha \in [0, 1]$. A statistic C with values in $\mathcal{S} \subset \mathcal{P}(\Theta)$ is called a confidence region estimator of significance level α, if*

$$P(C(X) \supset \{\theta\}) \geq 1 - \alpha$$

Let X be a random sample of $(\mathcal{E}, \mathcal{F})$-valued random variables and let $(f_\theta)_{\theta \in \Theta}$ be a family of probability density functions defined on $(\mathcal{E}, \mathcal{F})$. A parametric test problem is defined by two mutually exclusive hypotheses on the parametrization θ of the density f_θ. We formally write this as

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_1 : \theta \in \Theta_1$$

where Θ_0 and Θ_1 are disjoint subsets of the parameter space Θ. We refer to H_0 and H_1 as the null hypothesis and the alternative hypothesis, respectively.

Definition 2.1.6 (Hypothesis test according to Neyman-Pearson [13])**.** *Given a parametric test problem*

$$H_0 : \theta \in \Theta_0 \text{ v.s. } H_1 : \theta \in \Theta_1,$$

a statistic

$$\phi : \mathbb{X} \to \{0, 1\},$$

where values $0, 1$ are interpreted as "Null Hypothesis accepted" and "Null Hypothesis rejected", respectively, is referred to as a test statistic of significance level $\alpha \in [0, 1]$ with respect to the corresponding test problem, if

$$P(\phi(X) = 1 | H_0) := \sup_{\theta \in \Theta_0} P_\theta(\phi(X) = 1) \leq \alpha,$$

where P_θ refers to the probability measure induced by the density function f_θ for $\theta \in \Theta$, respectively.

Definition 2.1.7 (P-value)**.** *For a given test problem*

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_1 : \theta \in \Theta_1$$

and a statistic T, which defines the extremeness of an outcome in terms of deviation from the null hypothesis, the p-value statistic $p : \mathbb{X} \to [0,1]$ is defined as

$$p(x) = \sum_{y \in \tilde{\mathbb{X}} : T(y) \geq T(x)} P(X = y | H_0)$$

where $\tilde{\mathbb{X}}$ is a certain subset of $\mathbb{X}$.

The choice of $\tilde{\mathbb{X}}$ depends on the specific test problem. The p-value can be interpreted as the probability of observing a result at least as extreme as the observed one under the assumption that the null-hypothesis is true.

Remark 2.1.8. *Let $T(y) := f(y)$ and $f \sim X|H_0$, then*

$$\phi(x) := \begin{cases} 0 : p(x) \leq \alpha \\ 1 : p(x) > \alpha \end{cases}$$

defines a test statistic of significance level $\alpha \in [0,1]$ for the corresponding test problem.

The relation between the concepts of parametric testing and confidence region estimators is given in the following theorem:

Theorem 2.1.9 (Correspondence theorem)**.** *Given a test problem $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$ and let $C_\alpha : \mathbb{X} \to S \subset \mathcal{P}(\Theta)$ be a confidence region estimator of confidence level $1 - \alpha$, then*

$$\phi(x) = \begin{cases} 1 : C_\alpha(x) \cap \Theta_0 = \emptyset \\ 0 : C_\alpha(x) \cap \Theta_0 \neq \emptyset \end{cases}$$

defines a test statistic of significance level α.

Conversely, let $(\phi_i)_{i \in I}$ be a family of test functions of significance level α for the test problems

$$H_{i0} : \theta \in \Theta_i \text{ vs. } H_{i1} : \theta \in \Theta \setminus \Theta_i,$$

where $\bigcup_{i \in I} \Theta_i = \Theta$ and $\Theta_i \cap \Theta_j = \emptyset$ for $i \neq j$. Then

$$C : \mathbb{X} \to \mathcal{P}(\Theta), C(x) = \bigcup_{\phi_i(x) = 0} \Theta_i$$

defines a confidence region estimator of confidence level $1 - \alpha$.

Likelihood methods

In the following, let X be a random sample with corresponding density function f_{θ_0} parametrized by $\theta_0 \in \Theta$:

$$X \sim f_{\theta_0}$$

.

Definition 2.1.10 (Likelihood function)**.** *Given a realization x of the random variable X, the function*

$$\mathcal{L}(x|\ .\ ) : \Theta \to [0,1],\ \theta \mapsto f_\theta(x)$$

is referred to as the likelihood function with respect to the realization x.

The value $\mathcal{L}(x|\theta)$ is called the **likelihood** of the parameter θ. Global maxima of $\mathcal{L}$ are called **maximum likelihood estimates** of θ_0 and are denoted as $\hat{\theta}$.

Absolute values of the likelihood function $\mathcal{L}$ bear no meaning. Only by comparison of the likelihood values statistical properties of corresponding parametrizations can be derived. The following statistic is central for likelihood-based statistical methods.

Definition 2.1.11 (Likelihood ratio statistic)**.** *Given a parametric test problem $H_0 : \theta \in \Theta_0$ vs. $\theta \in \Theta \setminus \Theta_0$. The statistic*

$$\Lambda(x) := \frac{\sup_{\theta\in\Theta_0} \mathcal{L}(x|\theta)}{\sup_{\theta\in\Theta} \mathcal{L}(x|\theta)}$$

is referred to as the likelihood ratio statistic of H_0.

Usually the exact distribution of Λ is unknown. Under certain regularity conditions, one can show by Taylor expansion of the corresponding log-likelihood function the following asymptotic distribution of $-2\log\Lambda$:

Theorem 2.1.12 (Wilks theorem, [19])**.** *Let $\Theta \subset \mathbb{R}^p$. If the corresponding maximum likelihood estimator $\hat{\theta}$ is unique, consistent and asymptoticly normal distributed and if there exists $\Delta \subset \mathbb{R}^r, r < p$ and a diffeomorphism $\psi \in C_2(\Delta, \Theta_0)$, then the statistic $-2\log\Lambda(X)$ is asymptoticly $\chi^2(p-r)$ distributed under assumption of H_0.*

In many situations, the asymptotic χ^2-distribution of the transformed likelihood ratio statistic makes construction of tests and confidence region estimators straight forward. In the models, presented later in the thesis, we will make extensive use of theorem 2.1.12.

In case θ is multi-dimensional and we are only interested in certain components of θ, the so called **profile likelihood** function is used as a generalization of the normal likelihood function.

Definition 2.1.13 (Profile likelihood)**.** *Let $X \sim f_{\theta,\eta}$, where η is considered as a* **nuisance parameter***. Given the normal likelihood function $\mathcal{L}(x|\theta,\eta)$ with respect to the parameters (θ,η) and a realization x of the random variable X. The function L_p defined as*

$$L_p(x|\theta) := \max_\eta \mathcal{L}(x|\theta,\eta)$$

is called the profile likelihood with respect to θ.

The profile likelihood has similar statistical properties as the likelihood. In particular, under some regularity assumptions for the density f and the parameter spaces corresponding to the test problem, theorem 2.1.12 is also valid for the profile likelihood L_p [8].

2.1.3 Graph theory

We give a short summary of elementary graph theory notations and concepts, which we will need in the next section in order to formally describe cellular genealogies.

An ordered pair $G := (V, E)$ comprising a set V of **vertices** and a set $E \subset V \times V$ of **edges** is referred to as a **directed graph**. Any map defined on the vertex set of a graph is referred to as a **(vertex) labeling**. A finite sequence $(v_i)_{1 \leq i \leq n} \subset V$, where $(v_i, v_{i+1}) \in E, 1 \leq i \leq n-1$, is called a **path** from the vertex v_1 to the vertex v_n of **length** $n-1$. A directed graph, which has exactly one vertex r with the property, that there is exactly one path from r to every other vertex of the graph is referred to as a **rooted tree** and r is called the **root** of the tree. Let $G := (V, E)$ be a rooted tree. A vertex $v_1 \in V$ is called a **child node** of $v_1 \in V$, if $(v_1, v_2) \in E$. A rooted tree, where each vertex has at most two child nodes is referred to as a **binary tree**. An ordered pair $G' = (E', V')$ of subsets $V' \subset V$,$E' \subset E$ is referred as **subtree** of G, if G' is a rooted tree. Let $v \in V$, we denote the biggest subtree of G, which has the vertex v as a root as $G[v]$.

2.2 Mathematical modeling of cellular genealogies

In this section, we will first introduce the graph theoretical description of a cellular genealogy as a binary rooted tree. In order to perform statistical inference on cellular genealogy data on a consistent basis, we need to embedded this graph structure in a measure theoretic framework. The aim of our work is to infer evidence for asymmetric cell division processes from the occurrence of specific cell fates in the genealogy data. We therefore focus on a formulation, which enables us to characterize the dependency between the fates of cell siblings. This issue will be covered by the introduction of unordered random pairings as the central concept of this thesis.

2.2.1 Graph theoretic description of cellular genealogies

A cellular genealogy is a labeled binary rooted tree $G = (C, D)$. Cells are represented by the vertices C of the tree and parent-daughter relations are described by the edges D of the graph. Within this structure, cells are ordered into subsets C^g according to their generation g, which corresponds to the path length from the root cell c_1 to the specific cell in the graph. Cell specific features are given by the labelings of the tree. For a cell $c_i \in C$ of the genealogy, the subtree $C[c_i]$ is referred to as the colony of c_i.

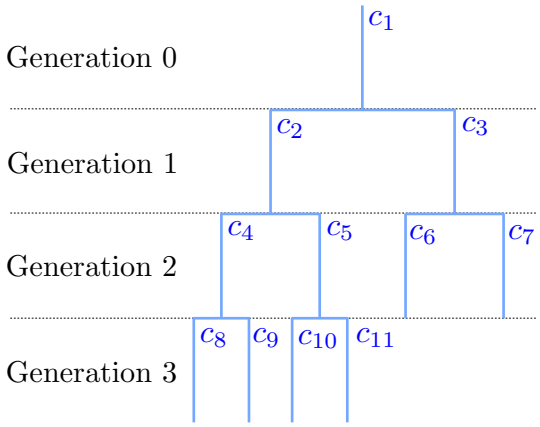


Figure 2.1: Graph representation of a cellular genealogy. Vertical lines correspond to the nodes of the underlying rooted tree, which are associated with the cells $c_i \in C$. Horizontal lines correspond to cell division events and indicate the parent-child relations of the rooted tree. Usually the canonical indexing, where the child nodes of a node $i \in \mathbb{N}$ are indexed by $2i$ and $2i+1$, is used. For time labeled trees the cell cycle time of a cell is indicated by the length of the corresponding line in the graph.

2.2.2 Stochastic description of cell siblings

The graph structure of a cellular genealogy does not define an order on the daughter cells. An adequate stochastic description is given by the random variable structure of an unordered random pairing, which models the commutability of cell sibling pairs.

Unordered pairings

Let X_1 and X_2 be two identically distributed random variabels. We refer to an ordered pair

$$X := (X_1, X_2)$$

as a **random pairing** and we call X_1, X_2 the **pair components** of X.

Let $X_i, i = 1, 2$ be $(\mathcal{E}, \mathcal{F})$-valued. The standard approach would be then to choose X to be $(\mathcal{E} \times \mathcal{E}, \mathcal{F} \otimes \mathcal{F})$-valued. In this case we refer to X as an **ordered random pairing**. If we assume that we cannot distinguish between realizations of the components X_1 and X_2 in our observation, we can model this by assuming X to be $(\mathcal{E} \times \mathcal{E}, \mathfrak{A})$ valued, where the set of observable events $\mathfrak{A}$ is chosen as the biggest σ-algebra contained in $\mathcal{F} \otimes \mathcal{F}$, which has the property

$$\forall E \in \mathfrak{A} : (a, b) \in E \Rightarrow (b, a) \in E \tag{2.1}$$

In the latter case, we call X an **unordered random pairing**. In this thesis, we will only deal with unordered random pairings where the components X_1 and X_2 are either binary or positive real valued, with corresponding measure spaces $(\{0,1\}, \mathcal{P}(\{0,1\}))$ and $([0,\infty), \mathcal{B}^+)$, respectively. $\mathcal{B}^+$ denotes the trace of the Borel σ-Algebra on $[0,\infty)$. In the following example we illustrate condition (2.1) for the binary case.

Example 1. *Let X be an unordered random pairing with $(\{0,1\}, \mathcal{P}(\{0,1\}))$-valued components*

$X_i, i = 1, 2$. The measure space of the corresponding ordered pairing is given by

$$(\{0,1\} \times \{0,1\}, \mathcal{P}(\{0,1\} \times \{0,1\}))$$

According to condition 2.1 the measure space of the corresponding unordered pairing is given by

$$\left(\{0,1\} \times \{0,1\}, \sigma(\left\{\{(0,0)\}, \{(0,1),(1,0)\}, \{(1,1)\}\right\})\right),$$

Note that in case of the unordered binary pairing the elements $(0,1)$ and $(1,0)$ are grouped in one set, within the set of observable events. This feature models the non distinctiveness of this events.

To make notation more compact we will denote the sets $\{(0,0)\}, \{(0,1),(1,0)\}$ and $\{(1,1)\}$ as $00, 01$ and 11 respectively. Moreover we will refer to realizations in 00 or 11 as **symmetric** outcome and to realizations in 01 as **asymmetric** outcomes. For the sake of simplicity we refer to unordered random binary valued pairings just as **binary pairings** .

Parametrizations of binary pairings

Let X be an unordered pairing with binary valued components. Due to the structure of our observation space, we can treat X as multivariate bernoulli distributed with corresponding categories $00, 01, 11$. We denote this in terms of a three categorial multinominal distribution with one draw:

$$X \sim \text{Mult}_3(1, \theta),$$

where the probabilities $\theta \in \Delta^2$ of the categories 11, 01 and 00 is given as an element of the standard 2-simplex

$$\Delta^2 := \{(x_1, x_2, x_2) \in [0,1]^3 : x_1 + x_2 + x_3 = 1\}.$$

In order to emphasize the correspondence between the categories 11, 01, 00 and the components of θ, we use the indexing

$$\theta := (\theta_{11}, \theta_{01}, \theta_{00}),$$

where

$$\theta_A = P(X \in A), A \in \{00, 01, 11\}.$$

Since we assumed in the definition of a binary pairing, that the components are identical distributed, we can define the marginal probability of the components $X_i, i = 1, 2$ as $\pi := P(X_i = 1) = \theta_{11} + \frac{1}{2}\theta_{01},\ i = 1, 2$ in a consistent way. In the following, we will sometimes use the notation $\pi(\theta)$ for the marginal probability, in order to emphasize the dependency to the parametrization θ.

Theorem 2.2.1. *Let X be a binary pairing and let $X \sim Mult_3(1, \theta)$, where $\theta \in \Delta^2 \setminus \{(1,0,0),(0,0,1)\}$, then the correlation $\rho := corr(X_1, X_2)$ is given by*

$$\rho = \frac{\theta_{11} - \pi^2}{\pi(1-\pi)}, \tag{2.2}$$

and the parametrization θ can be represented as

$$\theta = \begin{pmatrix} \pi^2 + \rho\pi(1-\pi) \\ 2\pi(1-\pi) - 2\rho\pi(1-\pi) \\ (1-\pi)^2 + \rho\pi(1-\pi) \end{pmatrix}^\top \tag{2.3}$$

Proof. It is $E[X_i] = \pi, i = 1, 2$ and $\mathrm{Var}(X_i) = \pi(1-\pi), i = 1, 2$. Therefore the correlation of the components X_1 and X_2 is given by

$$\mathrm{corr}(X_1, X_2) = \frac{E[(X_1 - E[X_1])(X_2 - E[X_1])]}{\sqrt{\mathrm{Var}(X_1)\mathrm{Var}(X_2)}} = \frac{E[(X_1 - \pi)(X_2 - \pi)]}{\pi(1-\pi)} = \frac{\theta_{11} - \pi^2}{\pi(1-\pi)},$$

which shows equation (2.2). Equation (2.3) directly follows by substitution of ρ by the use of the just shown equality. □

We refer to the right side of equation (2.3) as the **correlation based representation** of θ. We use the short notation

$$\theta(\rho, \pi) := \begin{pmatrix} \pi^2 + \rho\pi(1-\pi) \\ 2\pi(1-\pi) - 2\rho\pi(1-\pi) \\ (1-\pi)^2 + \rho\pi(1-\pi) \end{pmatrix}^\top$$

for $\rho \in [0,1]$ and $\pi \in [0,1]$. From the above theorem it follows that the correlation based representation is well defined, if

$$\theta \in \Delta^2 \setminus \{(1,0,0),(0,0,1)\}.$$

In particular, we have

$$\{\theta(\rho, \pi) : \rho \in (0,1), \pi \in (0,1)\} = \mathring{\Delta}^2,$$

where $\mathring{\Delta}^2$ denotes the topological interior of Δ^2.

Classification of parametrization of binary pairings

We define the set $\mathcal{K}_0$ as

$$\mathcal{K}_0 := \{\theta(\rho, \pi) : \pi \in [0,1], \rho = 0\}$$

From the foregoing explanations it follows that the components of an unordered binary pairing are independent if and only if the parametrization lies in $\mathcal{K}_0$, which is equivalent to the case that $\rho = 0$ i.e. the components X_1 and X_2 are uncorrelated.

The set $\mathcal{K}_0$ partitions Δ^2 into two subsets

$$\mathcal{K}_+ := \{\theta \in \Delta^2 : \theta_{01} < 2\pi(\theta)(1 - \pi(\theta))\}$$

and

$$\mathcal{K}_- := \{\theta \in \Delta^2 : \theta_{01} > 2\pi(\theta)(1 - \pi(\theta))\},$$

For parametrizations with values in $\mathcal{K}_+$ the probabilities for the symmetric outcomes 00 and 11 are higher than they would be under the assumption of independence and same marginal probability. For parametrizations with values in $\mathcal{K}_+$ the probability for the asymmetric outcome 01 is higher than it would be under the assumption of independence and same marginal probability. We therefore refer to parametrization with values in $\mathcal{K}_+$ as symmetric and to parametrization with values in $\mathcal{K}_-$ as asymmetric parametrizations. The notation is motivated by the fact that parametrizations in $\mathcal{K}_+$ can also be characterized by their property to describe binary pairings with positive correlated components, whereas parameterizations in $\mathcal{K}_-$ describe binary pairings with negative correlated components. This classification is graphically illustrated in the following Figure 2.2. We will use this setup to present parametrizations of binary pairings throughout this thesis.

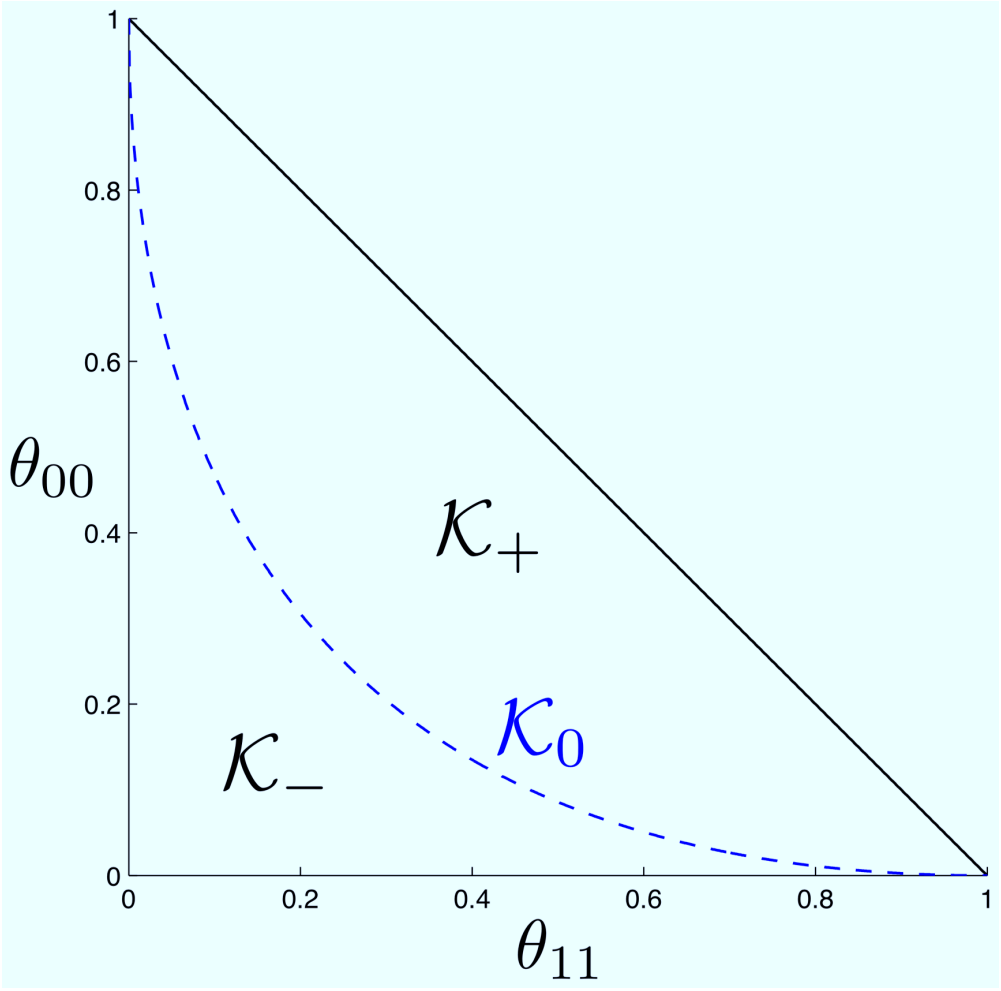


Figure 2.2: Graphical representation of classification of parametrizations of binary pairings: Projection of the standard 2-simplex Δ^2 on the span of θ_{00} and θ_{11}. For parametrizations with values in $\mathcal{K}_+$ the probabilities for the symmetric outcomes 00 and 11 are higher than they would be under the assumption of independence and same marginal probability. For parametrizations with values in $\mathcal{K}_+$ the probability for the asymmetric outcome 01 is higher than it would be under the assumption of independence and same marginal probability. In context of population genetics the set $\mathcal{K}_0$ is referred to as Hardy-Weinberg equilibrium.

2.3 Parameter inference on binary pairings

In the following section we will present some basic methods for parameter inferences for binary pairings in case of observable states as well as in some simple latent variable structures. The focus lies on the construction of confidence regions for the parametrization θ and on the construction of tests for the null hypothesis

$$H_0 : \theta \in \mathcal{K}_0$$

stating independence of the components. One-tailed tests for the null hypothesis

$$H_+ : \theta \in \mathcal{K}_+$$

stating positive correlation and for the null hypothesis

$$H_- : \theta \in \mathcal{K}_-$$

stating negative correlation can be constructed by simple variations. The main methods for each model are summarized in a figure at the end of each model section.

General notations

For the description of the following inference methods, we need to consider random samples of independent and identical distributed unordered random pairings. In order to avoid confusion between the indices of the sample components and the indices of the pair components of the unordered random pairings, we use the following notation.

Definition 2.3.1 (Notation of random samples of unordered random pairings)**.** *Let $n \in \mathbb{N}$ and let X be a random sample of n unordered random pairings. We denote the sample components of X with superscript inidices in round brackets, i.e. $X = X^{(1)} \dots X^{(n)}$. Let $X^{(j)}$ be a sample component of X. We denote the pair components of $X^{(j)}$ with subscript indices, i.e. $X^{(j)} = (X_1^{(j)}, X_2^{(j)})$.*

In case X is a random sample of binary pairings, then we denote the absolute frequencies of the types 00, 01, and 11 in the random sample as X_{00}, X_{01} and X_{11} respectively. Since the components in the random sample are independent and identical distributed the statistic

$$N(X) := (X_{11}, X_{01}, X_{00}) \sim \text{Mult}_3(n, \theta),$$

is multinominal distributed, whereby n denotes the number of sample components and θ refers to the parametrization of this components. Furthermore, let $n_1 := 2x_{11} + x_{01}$ denote the absolute number of pair components with value 1 in the sample.

2.3.1 Basic model

We refer to the case of observable pair components as our basic model. In this case inferences are based on realizations $(x_{11}x_{01}, x_{00})$ of the multinominal distributed statistic $N(X)$. We present two different statistical tests for the null hypotheses H_0, H_+ and H_- and four different types of confidence region estimators. The presented banana confidence region estimators are specific for binary pairings.

Permutation Test for H_0, H_+ and H_-

This test is a variation of the exact fisher test. It is well known as an exact test for Hardy-Weinberg equilibrium [3].

Under assumption of the null hypothesis H_0, the restriction of $N(X)$ on the set

$$\Gamma_{n_0,n} := \{(n_{00}, n_{01}, n_{11}) \in \mathbb{N}^3 : 2n_{11} + n_{01} = n_1, n_{00} + n_{01} + n_{11} = n\}$$

is (Fisher noncentral) hypergeometricly distributed with odds ratio 2 i.e.

$$P\left(N(X) = (n_{00}, n_{01}, n_{11}) | N(X) \in \Gamma_{n_0,n}\right) = \frac{\binom{n}{n_{00},n_{01},n_{11}} 2^{n_{01}}}{\binom{2n}{n_1}} =: T((n_{11}, n_{01}, n_{00}))$$

This makes it possible to directly calculate the corresponding p-value p_0 under the Null-Hypothesis H_0 by

$$p_0(x) := \sum_{T(y) \leq T(x)} T(x) \tag{2.4}$$

According to Remark 2.1.8, p_0 can be used to construct a test statistic for H_0. With some additional arguments the same can be shown for the the p-value p_- corresponding to the Null-hypothesis H_-, if it is defined by

$$p_-(x) := \sum_{T_-(y) \leq T_-(x)} T(x),$$

where

$$T_-(X) := \begin{cases} 1 : N(X) \in S_- \\ T(X) : \text{otherwise} \end{cases} \tag{2.5}$$

and

$$S_- := \{(n_{11}, n_{01}, n_{00}) \in \mathbb{N}^3 : n_{01} \leq 2\sqrt{n_{00}n_{11}}\}.$$

Likelihood ratio test for H_0, H_+ and H_-

For sufficiently large sample size and under the assumption that θ lies in the topological interior of Δ^2, the hypothesis H_0, H_+ and H_- can easily be tested by the corresponding likelihood ratio test: Since $N(X) \sim \text{Mult}_3(n, \theta)$ and by definition 2.1.11 the likelihood ratio statistic Λ_0 corresponding to the test problem H_0 vs. H_1 is given by

$$\Lambda_0(x) := \frac{\sup_{\theta \in \mathcal{K}_0} \mathcal{L}(x|\theta)}{\sup_{\theta \in \Delta^2} \mathcal{L}(x|\theta)} = \frac{\sup_{\theta \in \mathcal{K}_0} \theta_{11}^{x_{11}} \theta_{01}^{x_{01}} \theta_{00}^{x_{00}}}{\sup_{\theta \in \Delta^2} \theta_{11}^{x_{11}} \theta_{01}^{x_{01}} \theta_{00}^{x_{00}}}. \tag{2.6}$$

Using the correlation based representation for θ the statistic Λ can be written as

$$\Lambda_0(x) := \frac{\sup_{\pi \in (0,1), c=0} \theta_{11}^{x_{11}} \theta_{01}^{x_{01}} \theta^{x_{00}}}{\sup_{\pi \in (0,1), c \in (-1,1)} \theta_{11}^{x_{11}} \theta_{01}^{x_{01}} \theta^{x_{00}}}. \tag{2.7}$$

It follows by theorem 2.1.12 that $-2 \log \Lambda_0(N(X))$ is asymptotical $\chi^2(1)$ distributed. For a given significance level $\alpha \in [0, 1]$ this yields the test statistic

$$\phi(X) := \begin{cases} 0 : -2 \log \Lambda_0(N(X)) \leq \chi^2_\alpha(1) \\ 1 : -2 \log \Lambda_0(N(X)) > \chi^2_\alpha(1) \end{cases} \tag{2.8}$$

Test statistics for the one tailed tests corresponding to the null hypothesis H_+ and H_- are given by substitution of $\mathcal{K}_0$ in (2.6) by $\mathcal{K}_+$ and $\mathcal{K}_-$, respectively.

Exact banana confidence region

The permutation test described above can be generalized to certain differentiable submanifolds of Δ^2. Using the correspondence theorem 2.1.9, we can construct exact confidence regions as unions of this submanifolds.

Definition 2.3.2 (γ-curve)**.** *Let $\gamma \in [0, \infty)$, we refer to the set*

$$\Delta^2_\gamma := \{\theta \in \Delta^2 : \theta_{11} + 2\gamma\sqrt{\theta_{11}\theta_{00}} + \theta_{00} = 1\}$$

as a **γ-curve** *.*

Examples of γ-curves are illustrated in Figure 2.3. It can be easily validated that

$$\bigcup_{\gamma \in [0,\infty)} \Delta^2_\gamma = \Delta^2 \setminus \{\theta \in \Delta^2 : \theta_{00} = 0 \vee \theta_{11} = 0\}.$$

In particular, it follows for the topological closure that

$$\overline{\bigcup_{\gamma \in [0,\infty)} \Delta^2_\gamma} = \Delta^2.$$

Let $\gamma \in [0, \infty)$. Under the assumption of the null hypothesis $H_\gamma : \theta \in \Delta_\gamma^2$, the restriction of $N(X)$ on $\Gamma_{n_0,n}$ is (Fisher noncentral) hypergeometricaly distributed with odds ratio 2γ i.e.

$$P\left(N(X) = (n_{00}, n_{01}, n_{11}) | N(X) \in \Gamma_{n_0,n}\right) = \frac{\binom{n}{n_{00},n_{01},n_{11}}(2\gamma)^{n_{01}}}{D_{\gamma,n_0,n}} =: T_\gamma\left((n_{11}, n_{01}, n_{00})\right) \quad (2.9)$$

where

$$D_{\gamma,n_0,n} := \sum_{(m_{11},m_{01},m_{00}) \in \Gamma_{n_0,n}} \binom{n}{m_{00}, m_{01}, m_{11}} (2\gamma)^{m_{01}}.$$

Analogously to (2.4) we can define the corresponding p-value by

$$p_\gamma(x) := \sum_{T_\gamma(y) \leq T_\gamma(x)} T(x).$$

Under the general asumption $\theta \in \bigcup_{\gamma \in [0,\infty)} \Delta_\gamma^2$ it follows for given $\alpha \in [0, 1]$ by theorem 2.1.9 and remark 2.1.8 that the set

$$C_{1-\alpha}(x) := \bigcup_{\gamma : \phi_\gamma(x) = 0} \Delta_\gamma^2, \quad (2.10)$$

where

$$\phi_\gamma(x) := \begin{cases} 0 : p_\gamma(x) \leq \alpha \\ 1 : p_\gamma(x) > \alpha \end{cases} \quad (2.11)$$

defines a confidence region of confidence level $1 - \alpha$. As shown in Figure 2.4b, confidence regions constructed by this method have shapes similar like a banana.

Simultaneous confidence intervals

For large sample-sizes it can be reasonable to construct confidence regions for θ by the use of simultaneous confidence intervals for two of the three components $\theta_{00}, \theta_{01}, \theta_{11}$ due to computational reasons. The corresponding confidence regions have rectangular shapes (Figure 2.4d). Common methods are based on variations of the customary χ^2 statistic

$$\sum_{A \in \{11,00,01\}} \frac{(x_A - n\theta_A)^2}{n\theta_A}.$$

For details we refer to [7].

Likelihood based confidence regions

For sufficiently large sample size we can use likelihood methods to construct confidence regions for θ.

Likelihood confidence region: From theorem 2.1.12 it follows that

$$-2\log\frac{\theta_{00}^{X_{00}}\theta_{01}^{X_{01}}\theta_{00}^{X_{00}}}{\hat{\theta}_{11}^{X_{11}}\hat{\theta}_{01}^{X_{01}}\hat{\theta}_{00}^{X_{00}}}$$

is asymptoticaly $\chi^2(2)$ distributed. By theorem 2.1.9 a confidence region estimator for θ of confidence level $\alpha \in [0,1]$ is therefore given by

$$C_{1-\alpha}(X) := \left\{\theta \in \Delta^2 : -2\log\frac{\theta_{00}^{X_{00}}\theta_{01}^{X_{01}}\theta_{00}^{X_{00}}}{\hat{\theta}_{11}^{X_{11}}\hat{\theta}_{01}^{X_{01}}\hat{\theta}_{00}^{X_{00}}} \leq \chi^2_{1-\alpha}(2)\,\right\} \tag{2.12}$$

As demonstrated in Figure 2.4c the corresponding confidence regions are shaped similar to twisted ellipsoids.

Likelihood based banana confidence region: With the same arguments as for Λ_0 in (2.7) one can show that

$$\Lambda_r(x) := \frac{\sup_{\pi\in(0,1),c=r}\theta_{11}^{x_{11}}\theta_{01}^{x_{01}}\theta_{00}^{x_{00}}}{\sup_{\pi\in(0,1),c\in(-1,1)}\theta_{11}^{x_{11}}\theta_{01}^{x_{01}}\theta_{00}^{x_{00}}} \tag{2.13}$$

is asymptotically $\chi^2(1)$ distributed under the assumption that the correlation between the components of the corresponding binary pairings has the value r. By the use of theorem 2.1.9 a confidence region estimator of confidence level $1-\alpha$ is then given by:

$$C_{1-\alpha}(X) := \left\{\theta(c,\pi) : -2\log\Lambda_c(X) \leq \chi^2_{1-\alpha}(1)\right\}. \tag{2.14}$$

Particularly, for parametrization θ with corresponding correlation close to zero this confidence regions have similar shapes as the above presented exact banana confidence regions (Figure 2.4a). However, for parametrization θ with corresponding correlation close to -1, the shapes of exact banana confidence regions and correlation based confidence regions can strongly differ (Figure2.3).

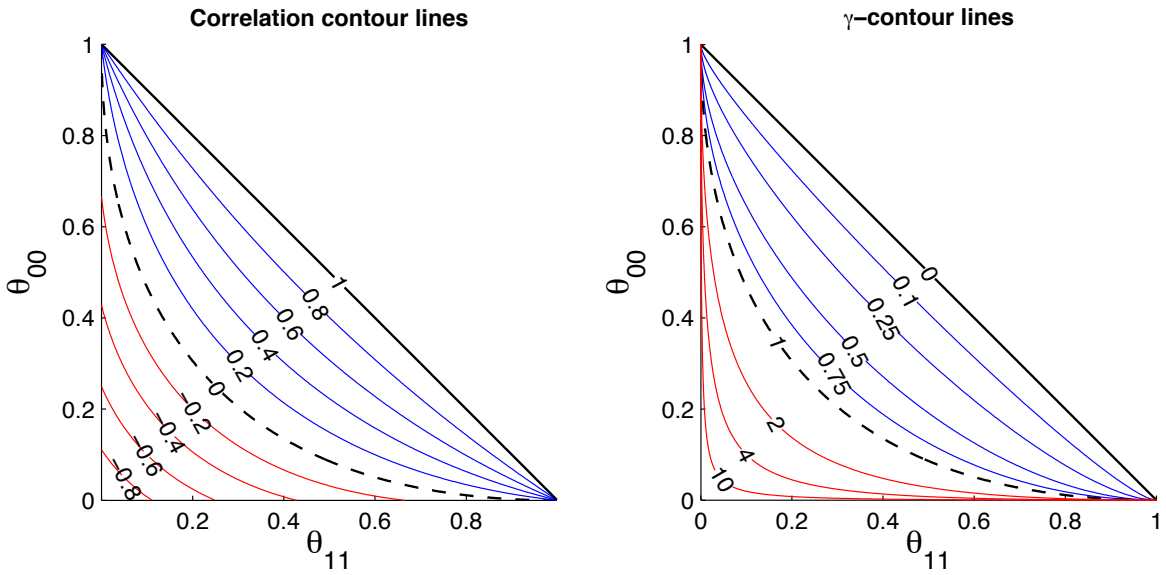


Figure 2.3: Comparison of γ-curves and correlation level curves: Line labels indicate corresponding the correlation level or γ-level, respectively. **a)** Contour-plot of selected correlation levels: $\mathcal{K}_0$ corresponds to the contour line of correlation level 0. Curves corresponding to negative or positive correlation levels lie in $\mathcal{K}_-$ or $\mathcal{K}_+$, respectively. **b)** Contour-plot of selected γ-curves: $\mathcal{K}_0$ corresponds to the contour line of γ-level 1. Curves corresponding to γ-level > 1 or < 1 lie in $\mathcal{K}_-$ or $\mathcal{K}_+$, respectively. Correlation level curves corresponding to negative correlation values only cover a proper subset of $[0, 1]$ of corresponding marginal probabilities, whereas γ-curves cover the whole range $[0, 1]$ for all γ-levels. As a consequence the curvature of γ-curves is unbounded for increasing γ-level.

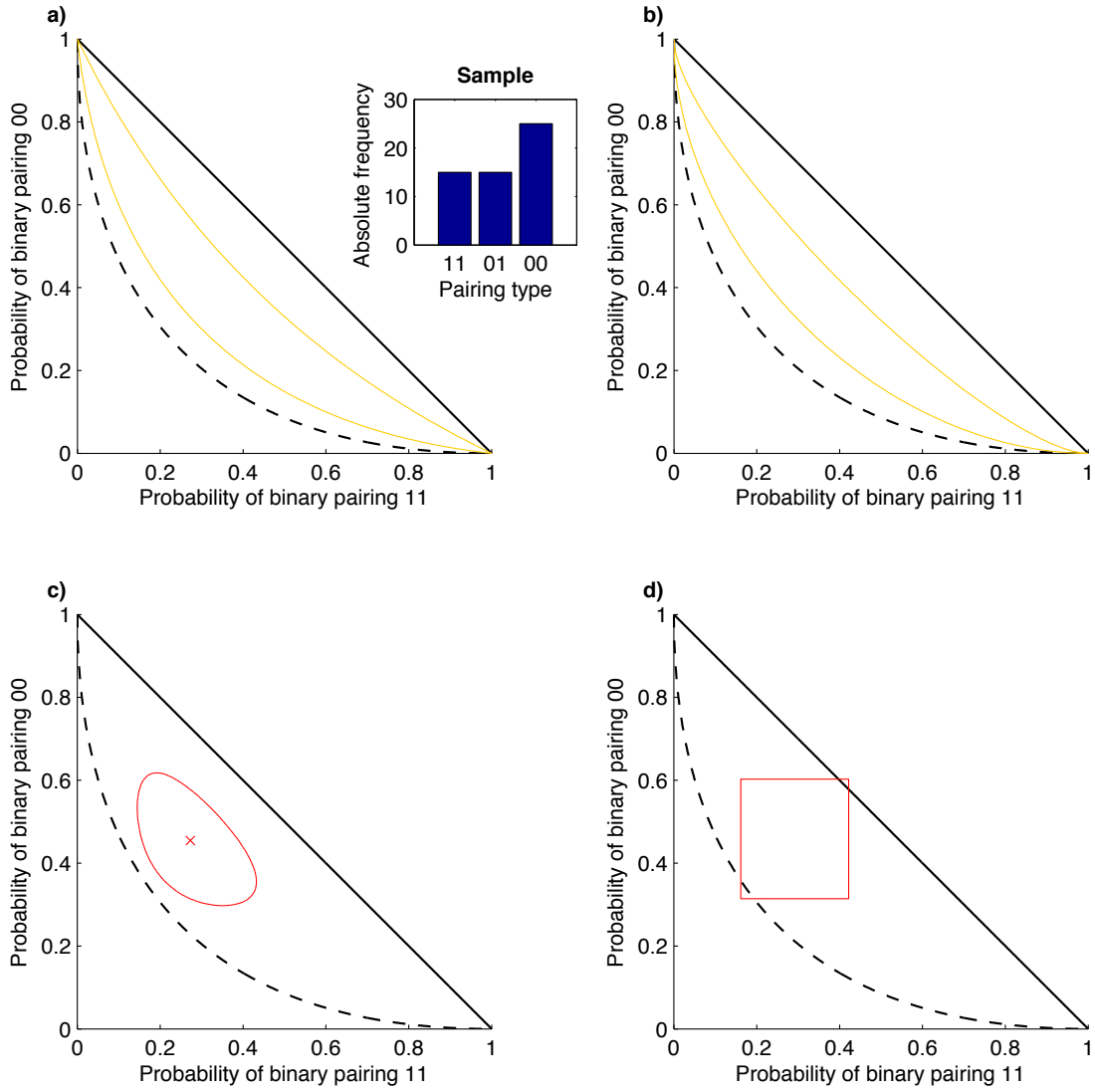


Figure 2.4: Concept figure for basic model: Illustration of different types of 95%-confidence region estimators based on a sample with absolute frequencies 15, 15 and 25 for the binary pairing values 11, 01 and 00 respectively. **a)** Banana confidence region calculated by the use of the correlation based estimator stated in (2.14) **b)** Banana confidence region calculated by the use of the exact banana confidence region estimator stated in (2.10) **c)** Likelihood confidence region based on the estimator stated in (2.12) **d)** Confidence region based on simultaneous confidence intervals for multi nominal proportions proposed by Goodman in [7]

2.3.2 Treatment-Control model

In this model we assume that the values of the binary pairing of interest can not be directly observed. Under certain assumptions, we can instead infer the corresponding parametrization by the comparison of the outcomes of a control and a treatment experiment.

General latent variable structure

Let Z and Y be two equal sized independent random samples, whereas each sample consists of identical distributed and independent binary pairings $Z^{(1)} \dots Z^{(n)}, Z^{(i)} \sim \text{Mult}_3(1, \theta^Z)$ and $Y^{(1)} \dots Y^{(n)}, Y^{(i)} \sim \text{Mult}_3(1, \theta^Y)$ respectively. We can associate Z and Y with the outcomes of an experiment conducted under two different conditions. Z is referred to as the outcome under treatment condition and Y is referred as the outcome under control condition.

For each sample component in Z and each sample component in Y we assume two latent binary valued states and we assume that the underlying distribution for this latent states is identical under treatment and control conditions. Formally we can denote this latent states as random samples of binary pairings $X := (X^{(i,j)})_{1 \le i \le 2, 1 \le j \le n}$ and $E := (E^{(i,j)})_{1 \le i \le 2, 1 \le j \le n}$ where the sample components $(X^{(1,j)})_{1 \le j \le n}$ and $(E^{(1,j)})_{1 \le j \le n}$ correspond to the latent states in the control sample and $(X^{(2,j)})_{1 \le j \le n}$ and $(E^{(2,j)})_{1 \le j \le n}$ correspond to the latent states in the treatment sample. Moreover the sample components $X^{(i,j)} \sim \text{Mult}_3(1, \theta^X)$ and $E^{(i,j)} \sim \text{Mult}_3(1, \theta^E)$ are independent and identical distributed respectively.

We now assume that the sample components in $Y^{(1)} \dots Y^{(n)}$ and $Z^{(1)} \dots Z^{(n)}$ are the result of a condition specific deterministic interaction between the corresponding latent states i.e. there are statistics $S_c, S_t : \{0,1\}^2 \times \{0,1\}^2 \to \{0,1\}^2$ so that $Y^{(i)} := S_c(X^{(i)}, E^{(i)})$ and $Z^{(i)} := S_t(X^{(i)}, E^{(i)})$

Multiple, though finite numbers of interactions modeled by S_c and S_t are possible. Under certain conditions for this interactions, we can infer the parametrization of the latent states X and E by comparison of the treatment and the control sample. In the following we will present parameter inference methods for a special case of interactions S_c and S_t, which we will later use in our application.

Latent variable structure with logical AND interaction

In the following, we set

$$S_c(X^{1,j}, E^{1,j}) := E^{1,j} \tag{2.15}$$

$$S_t(X^{2,j}, E^{2,j}) := (\min\{X_1^{2,j}, Y_1^{2,j}\}, \min\{X_2^{2,j}, Y_2^{2,j}\}) \tag{2.16}$$

for the statistics S_c and S_t. According to (2.15) we have $Y = E$ and E can be treated as an observable variable under treatment conditions. In particular we have $\theta^X = \theta^E$. The interaction

S_t between the components of the binary pairings E^j and X^j can be interpreted as logical AND interaction, if we associate the values $0, 1$ with boolean states *true, false* respectively. This interpretation is in most cases of application more intuitive than the above stated definition with component wise minima.

Parameter inference for logical AND interaction

Since θ^Y and θ^Z can be directly inferred from observations under control conditions and treatment conditions respectively and according to (2.15) we have $\theta^E = \theta^Y$. In the following we show how the parametrization θ^X of the latent states X can be inferred.

The interactions stated in (2.15) and (2.16) induce the following dependencies between the parametrizations θ^X, θ^Y and θ^Z.

$$\begin{aligned}
\theta^Z_{11} &= \theta^X_{11}\theta^Y_{11} \\
\theta^Z_{01} &= \frac{1}{2}\theta^X_{01}\theta^Y_{01} + \theta^X_{11}\theta^Y_{01} + \theta^X_{01}\theta^Y_{11} \\
\theta^Z_{00} &= \theta^X_{00} + \theta^Y_{00} + \frac{1}{2}\theta^X_{01}\theta^Y_{01}
\end{aligned}$$

which in matrix-vector notation can shortly be written as:

$$\theta^Z = R(\theta^X){\theta^Y}^\top \tag{2.17}$$

where

$$R(\theta^X) := \begin{pmatrix} \theta^X_{11} & 0 & 0 \\ \theta^X_{01} & \theta^X_{11} + \frac{1}{2}\theta^X_{01} & 0 \\ \theta^X_{00} & \theta^X_{00} + \frac{1}{2}\theta^X_{01} & 1 \end{pmatrix} \tag{2.18}$$

We refer to the matrix $R(\theta^X)$ as the **effectmatrix** of θ^X.

For given observations $y = y_1 \dots y_n$ and $z = z_1 \dots z_n$ of Y and Z respectively, the corresponding likelihood function is given by

$$\begin{aligned}
&\mathcal{L}(\theta^Y, \theta^Z | N(y), N(z)) = \mathcal{L}(\theta^Y | N(y))\mathcal{L}(\theta^Z | N(z)) = \\
&\binom{n}{z_{11}, z_{01}, z_{00}} {\theta^Z_{11}}^{z_{11}} {\theta^Z_{01}}^{z_{01}} {\theta^Z_{00}}^{z_{00}} \binom{n}{y_{11}, y_{01}, y_{00}} {\theta^Y_{11}}^{y_{11}} {\theta^Y_{01}}^{y_{01}} {\theta^Y_{00}}^{y_{00}},
\end{aligned}$$

since by assumption the samples Y and Z are independent and $N(Y)$ and $N(Z)$ are multi nominal

distributed. By substitution of θ^Z with $R(\theta^X){\theta^Y}^\top$, we can derive the profile likelihood for θ^X

$$L_p(y,z|\theta_X) := \max_{\theta^Y} \mathcal{L}(R(\theta^X)\theta^Y|z)\mathcal{L}(\theta^Y|y) =$$
$$\max_{\theta^Y} \binom{n}{z_{11},z_{01},z_{00}} \left(R(\theta^X){\theta^Y}^\top\right)_{11}^{z_{11}} \left(R(\theta^X){\theta^Y}^\top\right)_{00}^{z_{01}} \left(R(\theta^X){\theta^Y}^\top\right)_{00}^{z_{00}}$$
$$\binom{n}{y_{11},y_{01},y_{00}} {\theta_{11}^Y}^{y_{11}} {\theta_{01}^Y}^{y_{01}} {\theta_{00}^Y}^{y_{00}} = \max_{\theta^Y} \binom{n}{z_{11},z_{01},z_{00}} (\theta_{11}^X\theta_{11}^Y)^{z_{11}} (\frac{1}{2}\theta_{01}^X\theta_{01}^Y + \theta_{11}^X\theta_{01}^Y + \theta_{01}^X\theta_{11}^Y)^{z_{01}}$$
$$(\theta_{00}^X + \theta_{00}^Y + \frac{1}{2}\theta_{01}^X\theta_{01}^Y)^{z_{00}} \binom{n}{y_{11},y_{01},y_{00}} {\theta_{11}^Y}^{y_{11}} {\theta_{01}^Y}^{y_{01}} {\theta_{00}^Y}^{y_{00}}$$

As stated in the statistical preliminary section for sufficient large sample size the profile likelihood can be treated as in the same way as the common likelihood. Likelihood based confidence regions for θ^X and tests for the null hypothesis is H_0, H_+ and H_- stating $\theta^X \in \mathcal{K}_0, \theta^X \in \mathcal{K}_+$ and $\theta^X \in \mathcal{K}_-$ respectively can therefore be constructed in the same way as described in the basic model part.

Separating orbits

In the following we will sketch an alternative approach to construct tests for the null hypothesis $H_0 : \theta^X \in \mathcal{K}_0$, $H_+ : \theta^X \in \mathcal{K}_+$ and $H_- : \theta^X \in \mathcal{K}_-$, which makes use of so called separating orbits. For fixed θ^Y, equation (2.17) can be interpreted as a mapping rule $\theta^X \mapsto R(\theta^X)\theta^Y$, where the corresponding map maps the parametrization θ^X on the parametrization θ^Z of the treatment sample. Under the assumption of the null hypothesis this map can be parametrized with the domain $[0,1]$.

$$s(\ .\ )\theta^Y : [0,1] \to \Delta^2, \pi \mapsto R(\pi)\theta^Y, \tag{2.19}$$

where

$$s(\pi) := R\left((\pi^2, 2\pi(1-\pi), (1-\pi)^2)\right)$$

Sets $s([0,1])\theta^Y, \theta^Y \in \Delta^2$ are uniquely defined by their **start point** θ^Y. We use the short notation $\tau(\theta) := R([0,1])\theta$ and refer to the sets $\tau(\theta), \theta \in \Delta^2$ as **orbits**. This designations is motivated by the fact that sets $\tau(\theta), \theta \in \Delta^2$ are invariant under multiplication with matrices $s(\pi), \pi \in [0,1]$ i.e.

$$\forall \pi \in [0,1] \forall \theta \in \Delta^2 : s(\pi)\tau(\theta) \subset \tau(\theta) \tag{2.20}$$

It follows by the definition of $s(\ .\ )\theta^Y$, that under the null hypothesis $H_0 : \theta^X \in \mathcal{K}_0$, the parametrization θ^Y must be contained in the image of $s(\ .\ )\theta^Y$ i.e. with the above notation we can rewrite the null hypothesis H_0 as

$$H_0 : \theta^Z \in \tau(\theta^Y) \tag{2.21}$$

Theorem 2.3.3. *Let* $\tilde{\Delta}^2 := \Delta^2 \setminus \{\theta \in \Delta^2 : \theta_{00} = 0 \vee \theta_{01} = 0\}$. *The restriction of the family* $(\tau(\theta))_{\theta\in\tilde{\Delta}^2}$ *of* **separating orbits** *on* $\Delta^2 \setminus \{(0,0,1)\}$ *forms a partition of* $\Delta^2 \setminus \{(0,0,1)\}$ *and each orbit* $\tau(\theta)$ *separates* Δ^2 *into two subsets*

$$\mathcal{K}_{\tau(\theta)^+} := conv\big(\gamma(\theta) \cup \{\theta \in \Delta^2 : \theta_{00} = 0\}\big)$$

and

$$\mathcal{K}_{\tau(\theta)^-} := \Delta^2 \setminus \mathcal{K}_{\tau(\theta)^+}$$

whereas $conv(A)$ *denotes the convex hull of the corresponding set* A.

Proof. Can be derived by the use of (2.20), the continuity of the mapping $R(\,.\,)\theta^Y$ and the fact that $(0,0,1) \in \tau(\theta)$ for all $\theta \in \Delta^2$. □

The following theorem states sufficient criteria for the rejection of the null hypothesis H_0 by the use of separating orbits, which have been defined within the above theorem.

Theorem 2.3.4. *Let* $\alpha \in [0,1]$ *and let* C^Y *and* C^Z *be two confidence region estimators of simultaneous confidence level* $1-\alpha$ *for the parametriztions* θ^Y *and* θ^Z *respectively. For two realizations* y *and* z *of* Y *and* Z *respectively* $H_0 : \theta^X \in \mathcal{K}_0$ *can be rejected with significance level* α *if there exists a separating orbit* $\tau(\theta), \theta \in \tilde{\Delta}^2$ *so that either*

$$C^Y(y) \subset \mathcal{K}_{\tau(\theta)^-} \text{ and } C^Z(z) \subset \mathcal{K}_{\tau(\theta)^+} \tag{2.22}$$

or

$$C^Z(z) \subset \mathcal{K}_{\tau(\theta)^-} \text{ and } C^Y(y) \subset \mathcal{K}_{\tau(\theta)^+} \tag{2.23}$$

i.e. there is no separating orbit that intersects the interior of $C^X(x)$ *and* $C^Y(y)$.

Proof. Under the assumption of H_0, the parametrization θ^Z has to be contained in the image of $s(\,.\,)\theta^Y = \gamma(\theta^Y)$ i.e.

$$H_0 : \theta^Z \in \gamma(\theta^Y) \tag{2.24}$$

By definition the simultaneous covering probability of the confidence region estimator C^X and C^Y with respect to the parametrizations θ^X and θ^Y does not fall below $1-\alpha$. In particular it follows by (2.24) that the probability for the existence of an orbit intersecting C^X and C^Y is equal or greater than $1-\alpha$. □

With some additional arguments one can show, that in case of (2.22) we can reject the null hypothesis H_- and in case of (2.23) we can reject the null hypothesis H_+ with correspondinging significance α.

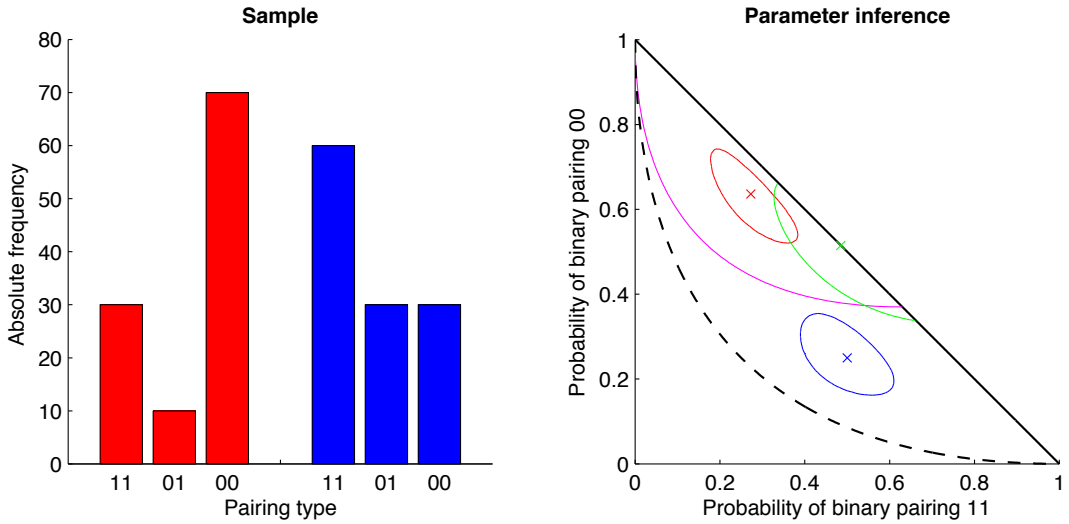


Figure 2.5: Treatment-control model with logical AND interaction: **a)** Barplot of the observed samples under treatment conditions (red) and control conditions (blue). **b)** Parameter inference for sample presented in a): Maximum likelihood estimates and likelihood based 97.5%-confidence regions for the parametrization of the control sample Y and the treatment sample Z are colored in blue and red respectively. The maximum likelihood estimate and the 95%-confidence region for the parametrization of the latent states X is colored in green. A separating orbit for significance level $\alpha = 0.05$ is colored in cyan. The p-value calculated on basis of the corresponding likelihood ratio statistic for the null-hypothesis H_0, which states independent components of X, is $2.3 \cdot 10^{-9}$. For given significance level $\alpha = 0.05$ the hypothesis H_0 can either be rejected by arguments of the corresponding p-value, existence of a separating orbit fulfilling condition (2.22), or no intersection of the 95%-confidence region for the parametrization of the latent effect state with $\mathcal{K}_0$.

2.3.3 Special mixture model

This model describes the case of a mixture distribution with binary pairings as underlying latent variables.

Latent variable structure

Let $X^{(1)} \dots X^{(n)}$ be a random sample of identical and independent distributed binary pairings and let $Y^{(1)} \dots Y^{(n)}$ where $Y^{(j)} \sim f, 1 \leq j \leq n$ be a random sample of independent and identical distributed real valued unordered pairings. We assume that the dependency between the pairing components of $Y^{(j)}$ and $X^{(j)}$ can be described by $Y_i^{(j)} | (X_i^{(j)} = 0) \sim f_0, i = 1, 2$ and $Y_i^{(j)} | (X_i^{(j)} = 1) \sim f_1, i = 1, 2$. Moreover we assume the components $Y_1^{(j)}$ and $Y_2^{(j)}$ of the observable variables $Y^{(j)}$ to be independent, given the corresponding component values $X_1^{(j)}$ and $X_2^{(j)}$ of the latent

states $X^{(j)}$, respectively. That means:

$$Y^{(j)}|(X^{(j)} = A) \sim \theta_A f_A, A \in \{11, 01, 00\},$$

where

$$X \sim \text{Mult}_3(1, \theta)$$

and

$$\begin{aligned} f_{11}(y^{(j)}) &:= f_1(y_1^{(j)}) f_1(y_2^{(j)}) \\ f_{01}(y^{(j)}) &:= \frac{1}{2}\big(f_0(y_1^{(j)}) f_1(y_2^{(j)}) + f_0(y_1^{(j)}) f_1(y_2^{(j)})\big) \\ f_{00}(y^{(j)}) &:= f_0(y_1^{(j)}) f_0(y_2^{(j)}) \end{aligned}$$

for realizations $y^{(j)} = (y_1^{(j)}, y_2^{(j)})$ of $Y^{(j)}$.

Likelihood function

According to the above stated dependencies the likelihood function with respect to the parameters $\theta, \lambda_1, \lambda_2$ for a realization $y := (y^{(j)})_{1 \le j \le n}$ of Y is given by

$$\mathcal{L}(y|\theta, \lambda_1, \lambda_2) = \prod_{j=1}^{n} \sum_{A \in \{00,01,11\}} \theta_A f_A(y^{(j)}|\lambda_0, \lambda_1), \tag{2.25}$$

where

$$\begin{aligned} f_{11}(y^{(j)}|\lambda_0, \lambda_1) &:= f_1(y_1^{(j)}|\lambda_1) f_1(y_2^{(j)}|\lambda_1) \\ f_{01}(y^{(j)}|\lambda_0, \lambda_1) &:= \frac{1}{2}\big(f_0(y_1^{(j)}|\lambda_0) f_1(y_2^{(j)}|\lambda_1) + f_1(y_1^{(j)}|\lambda_1) f_0(y_2^{(j)}|\lambda_0)\big) \\ f_{00}(y^{(j)}|\lambda_0, \lambda_1) &:= f_0(y_1^{(j)}|\lambda_0) f_0(y_2^{(j)}|\lambda_0). \end{aligned}$$

Analogously to the situation in the forgoing treatment-control model, we can infer the parameter θ by using the corresponding profile likelihood

$$L_p(y|\theta) := \max_{\lambda_0, \lambda_1} \prod_{j=1}^{n} \sum_{A \in \{00,01,11\}} \theta_A f_A(y^{(j)}; \lambda_0, \lambda_1). \tag{2.26}$$

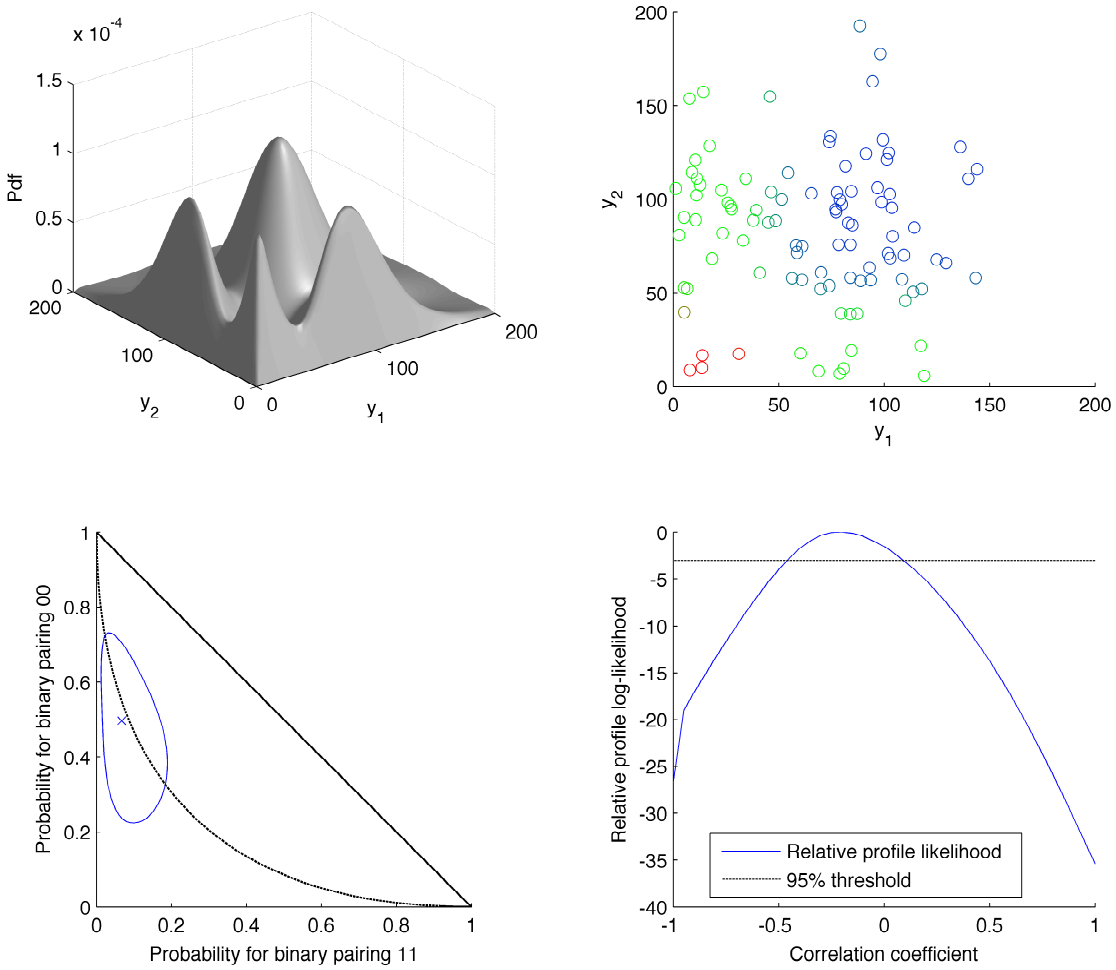


Figure 2.6: Special mixture model: **a)** Probability density plot of special mixture distribution with component densities $f_0 := \log(\mathcal{N})(3,1)$ and $f_1 := \log(\mathcal{N})(4.5, 0.3)$, and mixture proportion $\theta = (0.1, 0.4, 0.5)$ **b)** Scatterplot of raw data generated as toy data on basis of the special mixture distribution graphically illustrated in a). Soft classification based on the maximum likelihood estimatea $\hat{\theta} = (0.0670, 0.4361, 0.4970), \hat{\lambda_0} = (3.0344, 1.0323), \hat{\lambda_1} = (4.4853, 0.3163)$ for the parametrization of the underlying mixture distribution is indicated by the coloring. Blue, green and red color indicates high probability for a latent state of type 11, 01 and 00 respectively. **c)** Maximum likelihood estimate of θ and 95% cofidence region calculated on basis of the profile likelihood (2.26) **d)** Profile likelihood for the correlation of the latent binary pairing under the assumption of log-normal distributed components (blue).

Choice of the density function class

If we statistically infer the parameterization of the binary pairing in this latent variable structure, we have to assume the class of parameterized density functions on which we perform the maximization of the likelihood. The choice of this class has huge impact on the point estimate for θ. Therefore a carefull choice for the class of density functions is crucial. We presume that a combination of this method with a suitable equivalence test, which ensures that the induced error doesn't exceed a certain threshold is a possible approach to deal with this problem.

2.4 Mixture effects

In applications we often face the situation, that we don't know wether the data we observe really originate from (at least approximate) identical distributed binary pairings. In such situations attempts to estimate the correlation of the of the binary components are questionable, since the correlation tends to be overestimated:

Theorem 2.4.1 (Wahlund effect, [2])**.** *Let $Y, X^{(i)}, 1 \leq i \leq n$ binary pairings, where $X^{(i)}, 1 \leq i \leq n$ are independent with corresponding parametrizations $\theta^{(i)}$. Moreover let the parametrization θ^Y of Y be a convex combination of the parametrizations of the $X^{(i)}, 1 \leq i \leq n$ i.e.*

$$\theta^Y := \sum_{i=1}^{n} \alpha_i \theta^{(i)}.$$

Then the inequality

$$\sum_{i=1}^{n} \alpha_i corr(X^{(i)}) \leq corr(Y)$$

holds. In particular, we have equality if and only if the components of all $X^{(i)}$ have the same marginal probability π i.e.

$$\pi(\theta^Y) = \pi(\theta^{(i)}) \text{ for all } i \in \{1, \ldots, n\}.$$

Proof. Follows by the convexity of the correlation level curves (Figure 2.3). □

Chapter 3

Data

3.1 Experimental setup description

3.1.1 Experiments

The genealogical data analyzed in this thesis comprises four datasets, which originate from time-lapse movie experiments with mice haematopoietic cells conducted by Dirk Löffler from the Stem Cell Dynamics (SCD) group of T. Schroeder at the Helmholtz Zentrum München.

Nomenclature

Each dataset corresponds to one experiment. SCD internal designations of the experiments are **111012DL6**, **110603DL5**, **110722DL6**, **111210DL2**. For clarity, we refer to these experiments as **experiment 1**, **experiment 2**, **experiment 3**, **experiment 4**. For experiment 4, two different cell sortings schemes were used, while the time lapse movie was recorded on the same cell culture dish. With respect to this two different cell sortings, we distinguish between experiment 4a and experiment 4b.

Cell sorting

The cells in the experiments 1-3, 4a were sorted by FACS (Fluorescence-activated cell sorting) using the marker combination $CD150^{+}CD48^{-}CD34^{-}Lin^{-}MAC\text{-}1^{-}GR\text{-}1^{-}TER\text{-}119^{+}B220^{-}CD3E^{-}CD19^{-}CD41^{+}$. Transplantation experiments showed that sorting by this markers leads to an enrichment of long-term haematopoietic stem cells (LT-HSC) of approximate 50% [16]. The cells in experiment 4b were sorted using the marker combinations $CD150^{-}CD48^{-}CD34^{+}$, which leads to an enrichment of short-term haematopoietic stem cells (ST-HSC) and multi potent progenitors (MPP)[16].

Culture conditions

The basic cell culture condition in the experiments was SFEM (Serum-Free Expansion Medium), 100ng/ml SCF (stem cell factor), 100ng/ml TPO (thrombopoietin), 1% penicillin. The cell culture dishes used for this experiments were subdivided in physically separated segments. In half of the segments TGF-β1 (Transforming growth factor beta 1) was added to the cell culture. We refer to the culture condition where TGF-β1 was added as the **treatment condition** and to the culture condition, where TGF-β1 was not added as the **control condition**.

Time Lapse microscopy movies

The experiments were performed over a period of about 6-9 days. During this time the cell culture dish was continuously filmed by a camera taking pictures of each position on the cell culture dish around every 90 seconds. Compared to the average cell-cycle time of the observed cell, which is around 12 hours, this time resolution is more than sufficient.

3.1.2 Tracking and data representation

The generation of cellular genealogies i.e. the tracking of the cell lines in the time lapse movies was done manually with Timm's Tracking Tool (TTT)[15], a software for interactive manual tracking of single cells. The documented cellular genealogies have three labelings: The time labelings t_{start} and t_{end} are non negative and real valued. The values $t_{start}(c)$ and $t_{end}(c)$ are the time points at which the observation of a cell $c \in C$ started and terminated, respectively. The third labeling, denoted as s encodes the reason, why the observation of a cell terminated:

$$s(c) := \begin{cases} 0 : \text{cell tracking aborted} \\ 1 : \text{cell divided} \\ 2 : \text{cell died} \end{cases} \tag{3.1}$$

Depending on the experiment, the tracking was performed in different ways: Abortion of the tracking of single cells occurred, if the cell was lost or exceeded the end of the time lapse movie. The duration of the movies are 145, 142, 233 and 161 hours for the experiments 1-4, respectively. In experiment 1 and 3, only genealogies comprising at least two generations have been considered. Experiment 4 was performed to analyze the colony survival time t_{colony} of the root cell. In this experiment the tracking of a tree was not continued after a descendant cell that reached the end of the movie was found. Basic data information are summarized in table 3.1. Plots of typical genealogies with the described labeling are presented in Figure 3.2.

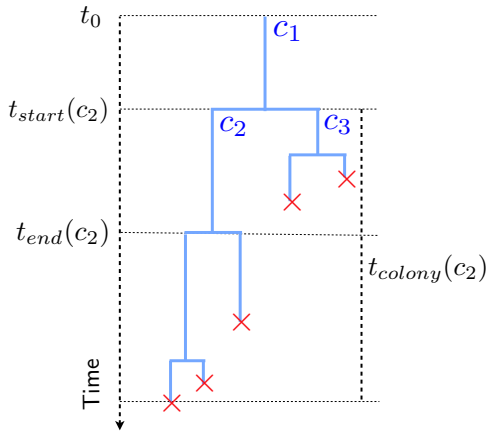


Figure 3.1: Time labeling of genealogies: Red crosses indicate cell death. Let c_i be a cell in the genealogy. Time labelings t_{start} and $_{end}(c_i)$ refer to the time point of begin and end of the observation of cell c_i, respectively. The colony survival time $t_{colony}(c_i) := t_{start}(c_i) - \max_{c \in C[c_i]} t_{end}(c)$ is defined as the time from cell birth of the corresponding cell until the death of its last descendant.

	Experiment 1	Experiment 2	Experiment 3	Experiment 4a	Experiment 4b
Sorted for cell-type	HSC	HSC	HSC	HSC	MPP
# Trees	41|84	110|99	59|61	41|30	37|50
# Complete tracked trees	2|35	85|92	7|58	36|29	37|49
# 1-Generation trees	0	85|76	0	35|21	16|36
Movie duration	145h	142h	233h	161h	161h

Table 3.1: Overview table Data: Left and right values in each column refer the numbers in the control and treatment condition, respectively. Row label **Sorted for cell type** refers to the target cell type in the preprocessing of the cells with FACS. For detailed FACS parameters see section 3.1.1. Labels #**Trees**, #**1-Generation trees** and #**Complete tracked trees** refer to the total number of trees, trees comprising only of one cell and trees, where all leaves comprise dying cells, documented in the experiment, respectively. Row label **Movie duration** refers to the duration of the corresponding time-lapse movie of the experiment.

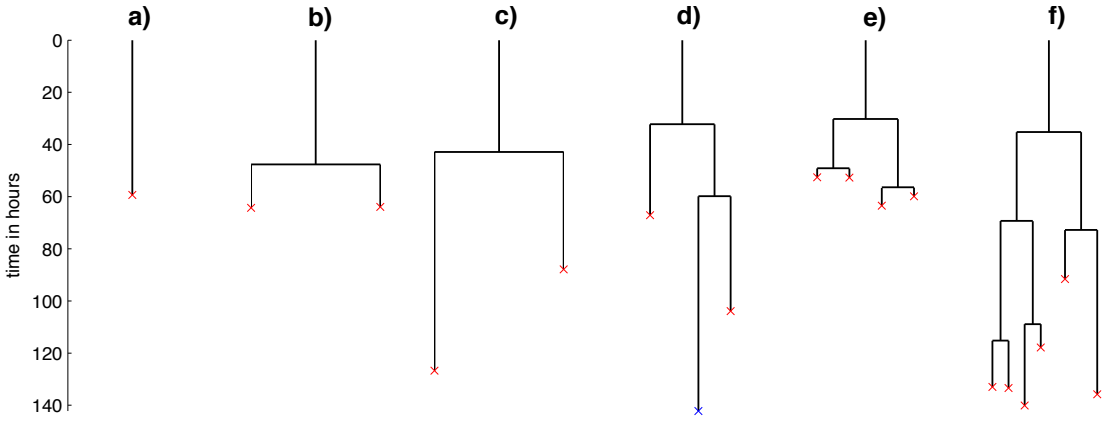


Figure 3.2: Treeplot of selected genealogies in experiment 2: Each vertical bar corresponds to one cell and indicates the life time of the corresponding cell. A red cross indicate that the corresponding cell died at the end of observation. A blue cross indicates that the tracking of the corresponding cell terminated, because the cell was lost or the duration of the movie was exceeded.

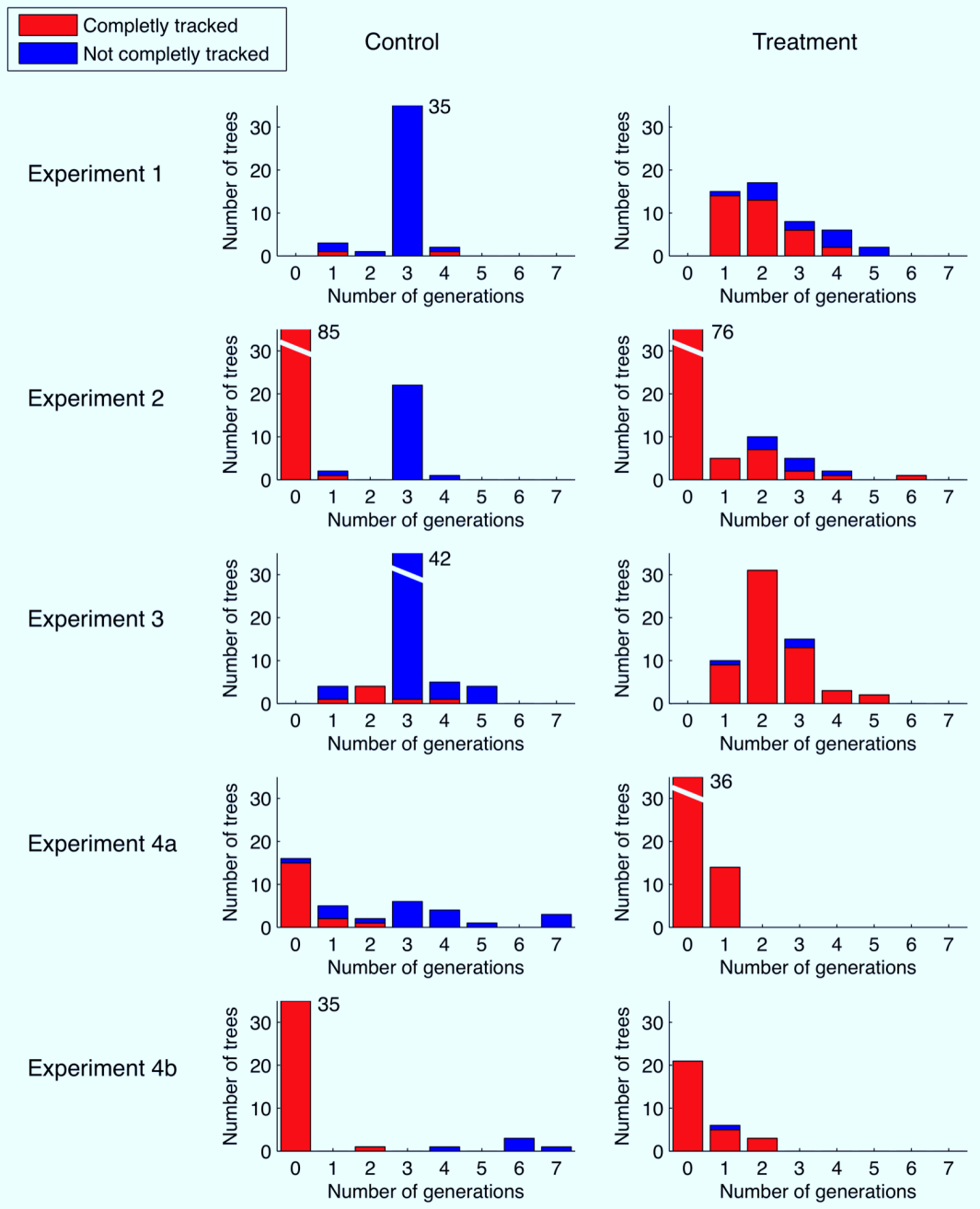


Figure 3.3: generation structure of genealogies: Experiment 1-4: Stacked barplot of the number of trees observed in the experiment comprising of exactly the number of observed generations indicated on the x-axis. Trees comprising only one generation have not been documented in experiments 1, 3. Trees in Experiment 4 comprise at most two generations, which is interesting for modeling since it justifies a generation based model approach for the death kinetic of MPP cells. Experiment 4 features an even higher number of cells dying in the first generation than Experiment 3. This phenomena could be explained by a high amount of differentiated cells in the beginning of the experiment.

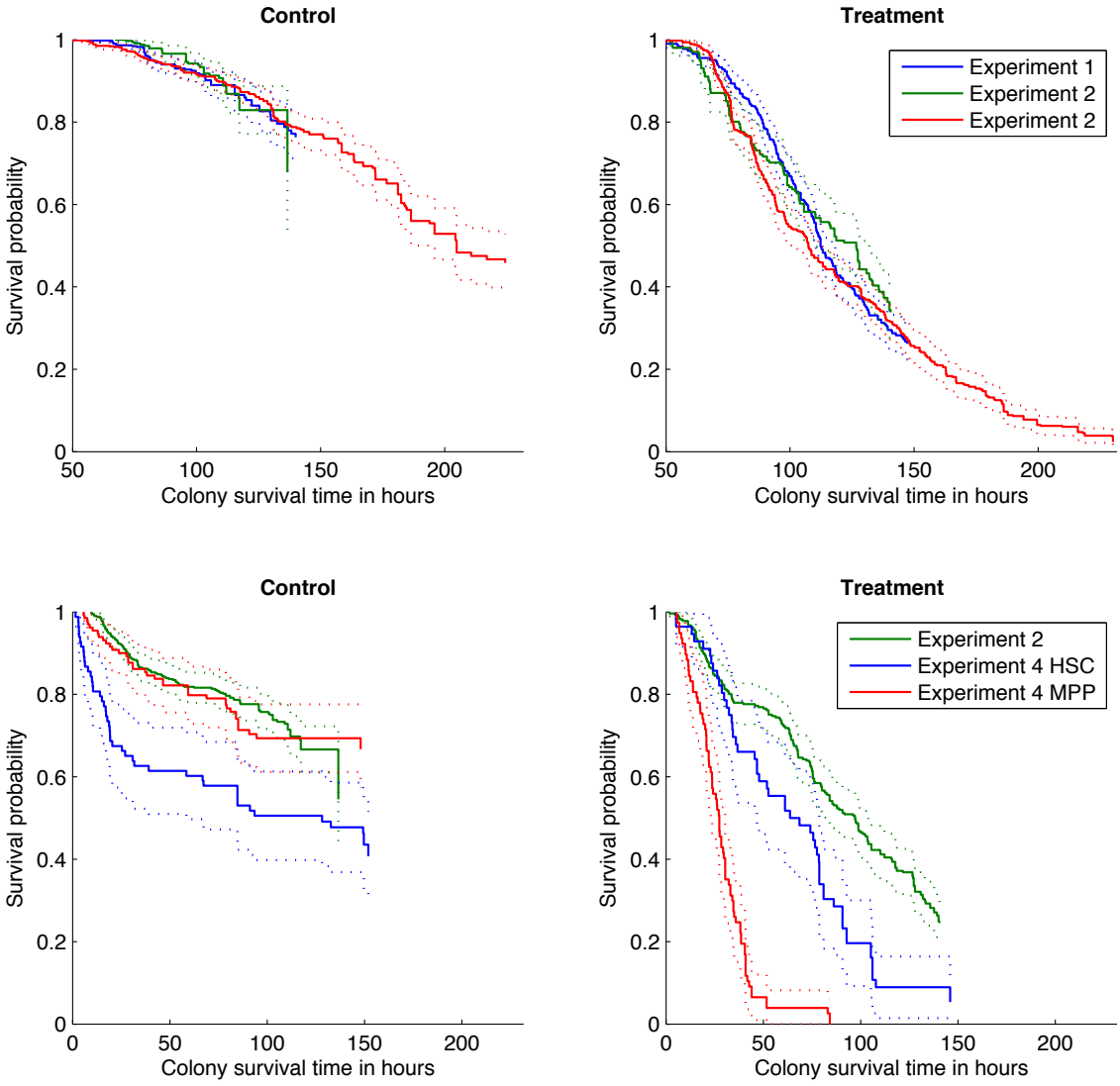


Figure 3.4: Kaplan-Meier estimators (lines) and point-wise 95% confidence intervals (dashed lines) for the colony survival time of cells in Experiment 1, 2, 3 (first row) and Experiment 2, 4a, 4b (second row), respectively. Cells in experiment have approximately similar survival curves. Under treatment condition cells live longer, than under control conditions. This effect is particular strong in experiment 4b, which mainly comprises of MPP cells. This supports the assumption, that MPP cells are more effected by TGF-β1 than HSC.

Chapter 4

Results

In the following we apply the generic models of chapter 1 on the genealogy data we presented in the previous chapter.

The experiments described in the previous chapter 2 have been performed in order to analyze cell fate choice in terms of transitions from HSC state to ST-HSC or MPP state. This three cell types differ in their self-renewal capacity. Unfortunately the self-renewal capacity of a cell can only be determined by elaborate transplantation experiments and can not be observed during experiments. The approach on which the experimental setup of the experiment 1-4 is based, is therefore to deduce the cell fate by death events under treatment (TGF-$\beta 1^{+}$) conditions. The approach is based on the assumption that cells without self-renewal capacity have a lower chance to survive under treatment conditions than cells with self-renewal ability. This assumption is justified by the comparison of the death kinetic of the cells in experiments 1-3 with the death kinetic of the cells in experiment 4 under treatment conditions. (Figure 3.3, Figure 3.4)

The models presented in the following are applications of the generic models described in chapter 2. Each model is based on different assumptions regarding the dependence between cell fates and death events. The basic methodological approach of corresponding generic models is to assume that a cells state can be characterized by one dichotomous feature. In this case of application this is the self-renewal ability. Instead of discriminating between sub categories of cells with different levels of self renewal capacity, we assume the self renewal capacity of a cell to be a dichotomous feature. Due to simplicity reasons we will use in the term HSC and the term MPP synonymously for cells with self-renewal capacity and for cells without self-renewal capacity, respectively. We associate HSC and MPP cells with the binary values 1 and 0, respectively. The goal of our statistical analysis is to infer the parametrization $\theta := (\theta_{11}, \theta_{10}, \theta_{00})$ of the binary pairing X describing the states of sibling cells after a division.

General proceeding

For each presented model, the assumptions and the basic structure are summarized in Figures 4.1, 4.3, 4.5. The graphical representation of the random variables can be interpreted in terms of a random markov field. Random variables in squared box are assumed to be observable, random variables in round ellipses are assumed to be latent. All parameter inferences were done accordingly to the methods of the corresponding generic models described in chapter 2. Parameter inferences were performed on experiments 1-3 separately in order to avoid mixing effects as described in theorem 2.4.1(Wahlund effect), which could lower the power of our statistical analysis. Moreover we fix $\alpha = 0.05$ as the significance level for our test decisions.

4.1 Model 1 (application of the basic model)

Biological assumptions

In this introductory model we assume that cells with self-renewal capacity never die under treatment conditions, whereas cells which loose their self-renewal capacity will immediately die in the same generation i.e. we assume observable states in the sense that under treatment conditions death events correspond to the loss of the self-renewal capacity of the corresponding cell.

Corresponding generic model and parameter inference

The situation of observable states corresponds to the setup of the basic model described in chapter 2. According to the biological assumptions stated above a pair of two dividing cells corresponds to a pair of two HSC. A pair of one dividing cell and one dying cell corresponds to a pair of one HSC and one MPP cell . A pair of two dying cells corresponds to a pair of two MPP cells. In binary notations we associate the latter states with the pairings 11, 01 and 00, respectively. We performed parameter inference on all generations separately. Only division events comprising two completely tracked cells have been considered. For the genealogy trees presented in Figure 3.2, that would mean that for the parameter inference on generation 1 level trees b), c) would be accounted as pairing types 00. Tree

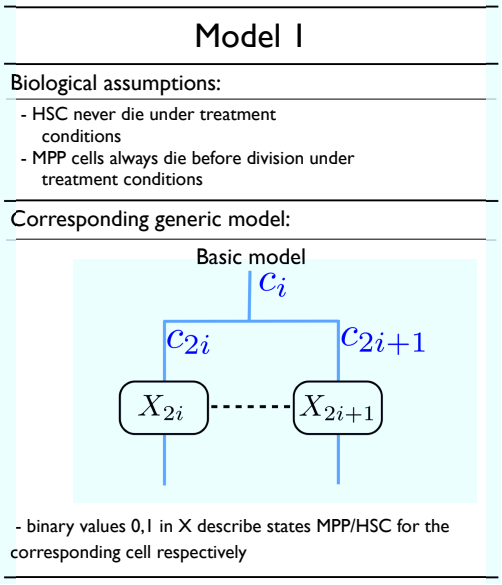


Figure 4.1

d) would be accounted as pairing type 01. The trees e), f) would be accounted as pairing types 11. In generation 2 the subtree $C[c_3]$ of tree d) would be not considered, since it is not completely tracked. Subtrees $C[c_2]$ and $C[c_3]$ of tree e) would both be accounted as binary pairing types 00 and subtrees $C[c_2]$ and $C[c_3]$ of tree f) would be accounted as binary pairings 11 and 00 respectively.

Parameter inference

Likelihood based confidence region for the parametrizations of the corresponding binary pairings are illustrated in Figure 4.1. In all three experiments the maximum likelihood estimates for generation 1-3 lie in the region of symmetric parametrizations $\mathcal{K}_+$. In all cases, this behavior is statistical significant (Table 4.1). Only for parametrizations in the fourth generation the maximum likelihood estimates in experiment 2 and 3 fall in K_-. However, there is no statistical significance given for this behavior, since the corresponding sample size is very small and the corresponding confidence regions cover a big range of $\mathcal{K}_+$ as well as of $\mathcal{K}_-$. In general throughout all experiments one can see an increase of the marginal probability π with higher generations. One interesting feature is, that in experiment 3 the difference between the marginal probability corresponding to the parametrization in generation 1 and the marginal probability for the parametrization in generation 2, respectively is much bigger than in the other experiments.

	Gen.1	**Gen.2**	**Gen.3**	**Gen.4**	**all Gen.**
Experiment 1	5.5696e-07	1.9735e-06	0.033341	0.089899	2.5924e-16
Experiment 2	0.019117	0.00038319	0.005364	1	1.8676e-08
Experiment 3	0.010103	8.8018e-06	0.0051999	1	2.4225e-16
All Data	6.3378e-09	1.3656e-14	4.4155e-06	0.13034	7.6484e-39

Table 4.1: Model 1: p-values under the null hypothesis H_0 stating independent components in generations 1-4, respectively. P-values have been calculated on basis of the permutation test described in **??**. It follows that in all experiments H_0 can be rejected for generations 1-3. Only in generation 4 the corresponding p-values exceed the given significance level $\alpha = 0.05$. However this is assumed to be a consequence of small sample sizes.

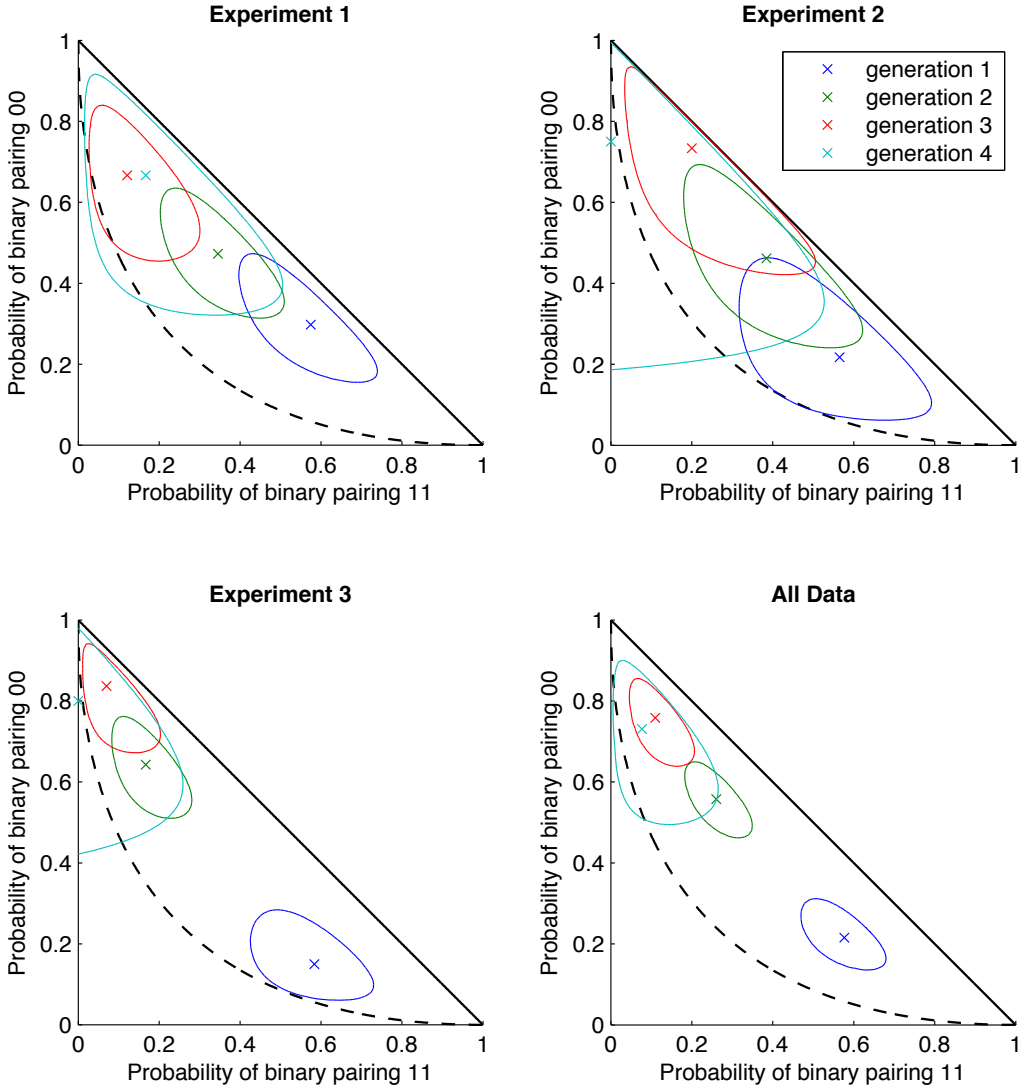


Figure 4.2: Model 1 parameter inference on generations 1-4: Experiment 1-4: 95%-Confidence region and maximum-likelihood estimates for the parametrization of the binary pairing describing the occurrence of death versus cell-division events under treatment conditions in sibling pairs in generations 1-4 and the occurrence averaged over all generations respectively. Under the assumptions of Model 1 this parametrization corresponds to the parametrization of the binary pairing describing the occurrence of differentiation events in sibling pairs in generation 1. P-values for the null hypothesis stating independence of the components of the pairing have been calculated on basis of the permutation test described in 2.3.1 and are listed in Table 4.1

4.2 Model 2 (application of the treatment-control model)

In the foregoing model we assumed that HSCs do not die under treatment conditions at all. This assumption is very questionable and in case it is violated dying HSCs are falsely accounted as differentiated MPP cells, which leads to a systematic error in the parameter inference of θ. In this model we maintain the assumption, that MPP cells always die during cell-cycle but make broader assumptions for the death kinetic of HSCs. We infer the unbiased correlated fate choice probability by the comparison of the correlated death probability under treatment and control conditions with the use of the generic treatment-control model.

Biological assumptions

As already mentioned above, we maintain in this model the assumption, that MPP cells always die during cell cycle. Under this assumption a minimum requirement in order to eliminate the above described bias is, that we can estimate the bias effect of dying HSC in our sample i.e. we need to be able to estimate the correlated death probability of HSC under treatment conditions. We approach this problem by estimating the correlated death probability of HSC on basis of the control dataset and assume the same correlated death probability of HSC under treatment conditions. To our knowledge this assumption does not stands in contradiction to any experimental results so far. As stated before the cells used in the experiments can not be assumed to be all of HSC type. Therefore this approach is only valid under the additional supposition that the correlated death probability of HSC and MPP cells is equal under control conditions. We inferred the corresponding parametrization of the correlated death probability in experiment 1-3 under control conditions and compared them with the corresponding parametrization in experiment 4b. On this basis of data, we could not find any contradicting evidence for the latter assumption.

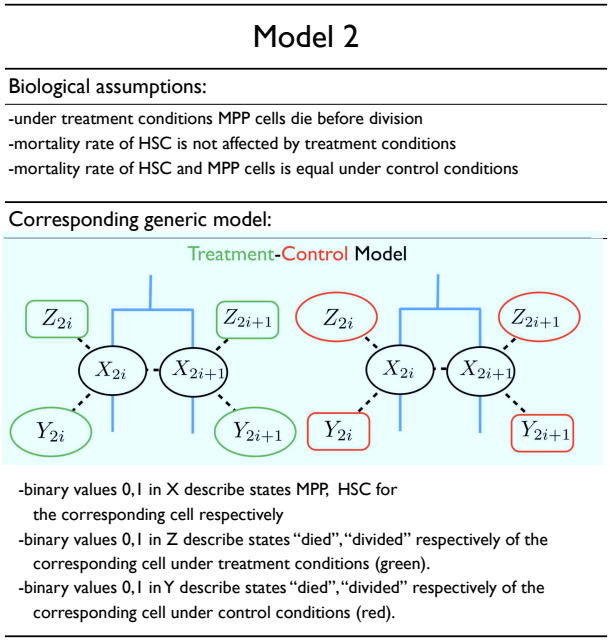


Figure 4.3

Analogy to the treatment control model with logical AND interaction

According to the above stated biological assumptions we can estimate the correlated death probability of HSCs under treatment conditions by inference on the control sample. However our actual aim is to infer the correlated death probability of cells, which died, because they differentiated to MPP cells. In order to infer this probability we need to formulate how the correlated death probability of HSCs influences the overall outcome of paired death events under treatment conditions, which is a combination of dying HSCs and cells, which die, because they differentiated to MPP cells. This can be done within the framework of the treatment control model with logical AND interaction.

With the notation used in section 2.3.2, we describe the control sample with Y and the treatment sample with Z. We associate the binary pairings describing the HSC dying by chance with the hidden binary pairing E and the binary pairing describing the fate of the cells with X. Under control conditions HSC and MPP have the same correlated death probability, therefore $Y = S_c(E, X) := E$. Under treatment condition a cell will only survive, if it is a HSC and does not die by chance i.e. $Z = S_t(E, X) := (\min\{E_1, X_1\}, \min\{E_2, X_2\})$. Apparently, the dependencies formulated by the mapping S_c and S_t exactly match with the dependencies formulated in the case of the treatment-control model with logical AND interaction. The following parameter inferences are based on the methods described in context of this model.

Parameter inference

Parameter inferences for this model with respect to the data of experiment 1-3 and the first generation, respectively are graphically illustrated in Figure 4.4. For all experiments we see, that the inferred confidence regions for the parametrization of the unordered parings describing the transitions from HSC state to MPP state are not remarkably different to the inferred parametrizations in model 1. This can easily be explained by the fact, that the number of cells dying under control conditions is very low. This leads to an inferred marginal survival probability in the treatment-control model which is close to 1 and which nearly matches with the assumption of no dying HSCs in model 1. Similar to the case in the basic model the maximum likelihood estimated for θ lie all in the symmetric part $\mathcal{K}_+$. This behavior is significant for experiment 1 and 3, due to the corresponding p-values for the null-hypothesis H_0 stating $\theta \in \mathcal{K}_0$.

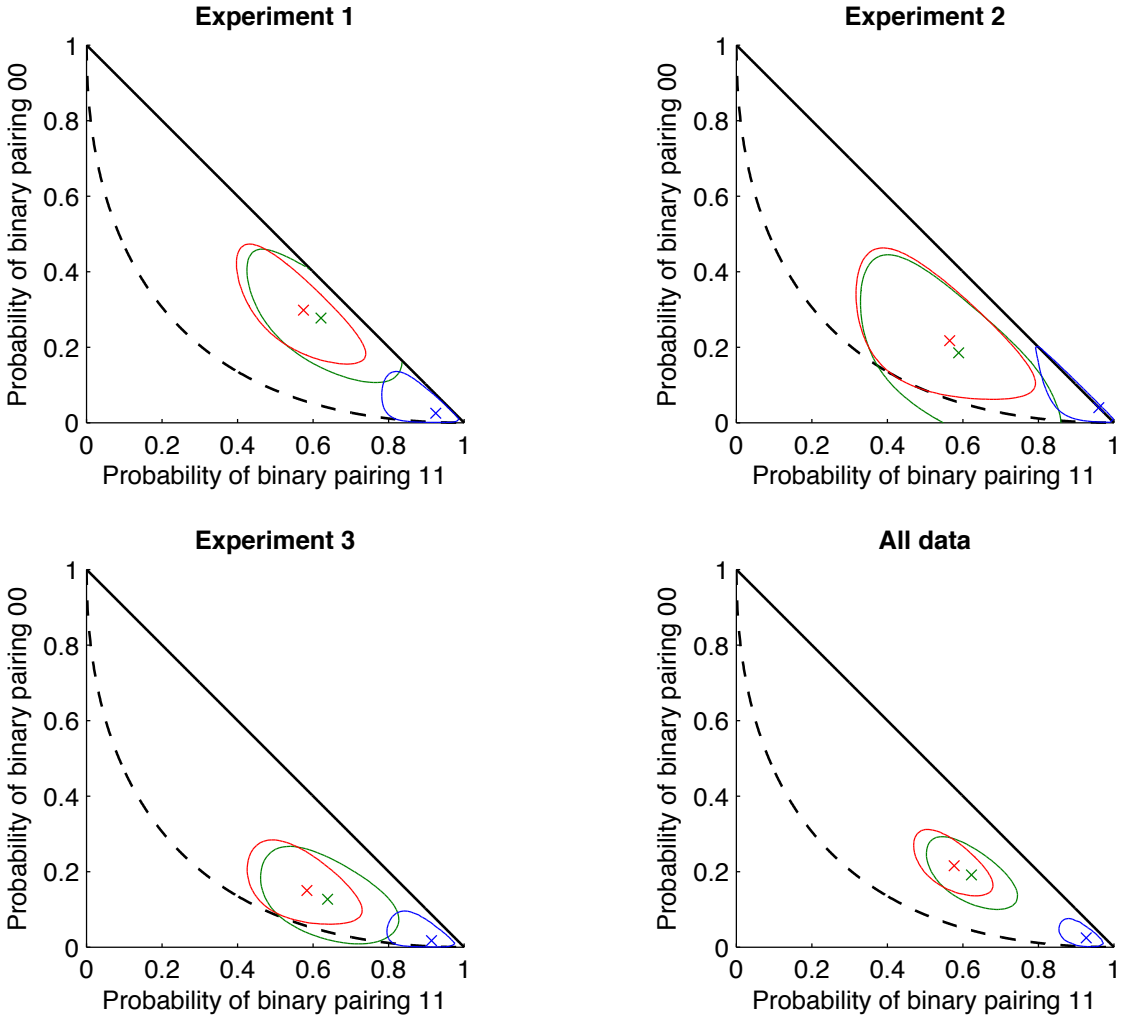


Figure 4.4: Experiment 1-3, All Data: 95%-Confidence region and ML-estimates for the parametrization of the binary parings describing the occurrence of differentiation events (green) under the assumptions of Model 2 and death versus cell-division events under control conditions (blue) and treatment conditions (red) in sibling pairs in generation 1.

4.3 Model 3 (application of the special mixture model)

In the previous two models we assumed, that all information needed to infer the parametrization of the latent states of the daughter cells is given by 2-generation death labeled trees. This has been enforced by assuming that cells, which loose their self-renewal ability will immediately die in the same generation. This assumption stands in contradiction to the results (see Figure 3.3) of experiment 4 in which the survival times of purified MPP cells under treatment conditions have been examined. The relative amount of cells dividing lies around 30 percent in this experiment, which emphasize that our assumption are a too vast simplification, since it is a reasonable argument that this purified MPP cells should be effected even more by TGF-β1 than cells we assume to just have undertaken state transition to MPP cells. However, so far there is no data which could give us direct evidence for the death kinetic of cells which just have undertaken state transition to MPP cells. The motivation of model 3 therefore is to make only very vague assumptions about the exact effect TGF-β1 has on this cells. We do this by application of the special mixture model. As observable unordered random pairing Y we use the colony survival time t_{colony} (Figure 3.1) of the corresponding cell sibling pairs. The critical point of this approach is to make a reasonable assumption for the class of distributions describing the colony survival time of HSC and MPP cells. Since survival processes are commonly described by lognormal or gamma distributions we did two model inferences: One based on gamma distributed components and one based on lognormal distributed components. This parameter inferences are based on the colony survival time of cell siblings of generation 1 and cells of generation 0 are assumed to be HSC.

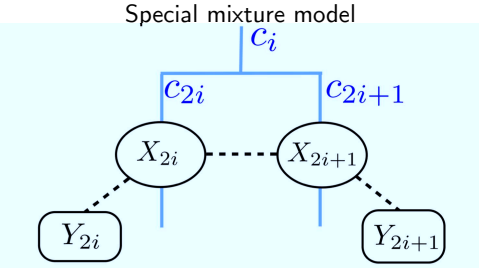
Model 3

Biological assumptions:

-mother cells of considered cell sibling pairs are assumed to be HSC
-colony survival time in treatment conditions of a cell is dependent on the cell's differentiation state
-conditional distributions of the colony survival time, given the differentiation state of the cell are assumed to be log-normal/gamma distributed.

Corresponding generic model:

Special mixture model



-binary values 0,1 in X describe the self-renewal fate (MPP, HSC) for the corresponding cell respectively
-non negative real values in Y describe the colony survival time under treatment conditions of the corresponding cell.

Figure 4.5

A minimal condition for a reasonable parameter inference based on this approach is, that under the assumption of the distribution class of our components, the colony survival time of the considered cells is significantly better described by a bimodal mixture distribution than by a not identifiable unimodal distribution.

At least under the assumption of lognormal distributed components we could show a significant better fit for the case of two components using the approximation of the corresponding likelihood

ratio statistic by the χ^2-distribution with 5 degrees of freedoms proposed in [4]. However, as it can be seen in Figure 4.6 this does not really coincide with the visual impression. In fact, in about 50% of all runs the likelihood optimization terminated in local minima or failed to converge at all. We assume this instability to be the result of an inadequat description of the data by the assumed mixture distribution.

The results of the parameter inference of this approach are graphically illustrated in Figure 4.8. Since the complete parameter inference of the parametrization of the unordered binary pairing describing the fate of our cells is computational not feasible in reasonable time, we only inferred the corresponding correlation level of the parametrization. Apparently the estimated confidence intervals for the correlation lie above zero and the hypothesis of independent components is rejected according to the corresponding p-values. However, we assume the bad convergence properties to originate in a bad model fit. Conclusions on this basis seem to be questionable.

	μ	σ	α	μ_1	σ_1	μ_2	σ_1
111012DL6	4.0031	0.41934	0.53213	3.6166	0.4093	4.4427	0.067585
110603DL5	4.0099	0.40716	0.6166	3.6541	0.32333	4.582	0.011157
110722DL6	4.114	0.41514	0.037824	2.3573	0.36809	4.183	0.29092
All Data	4.0555	0.41822	0.18042	3.3498	0.63796	4.2109	0.23608

Table 4.2: Parametrization of maximum likelihood fits of plots in Figure 4.6 μ, σ refer to the maximum likelihood estimates for the mean and the variance of the corresponding normal distribution in the one component lognormal-fit. α refers to the mixing proportions in two component model. $\mu_i, \sigma_i, i = 1, 2$ refer to the maximum-likelihood estimates for the mean and the variance of the corresponding normal distributions of the lognormal distributed components i=1,2 in the two component model.

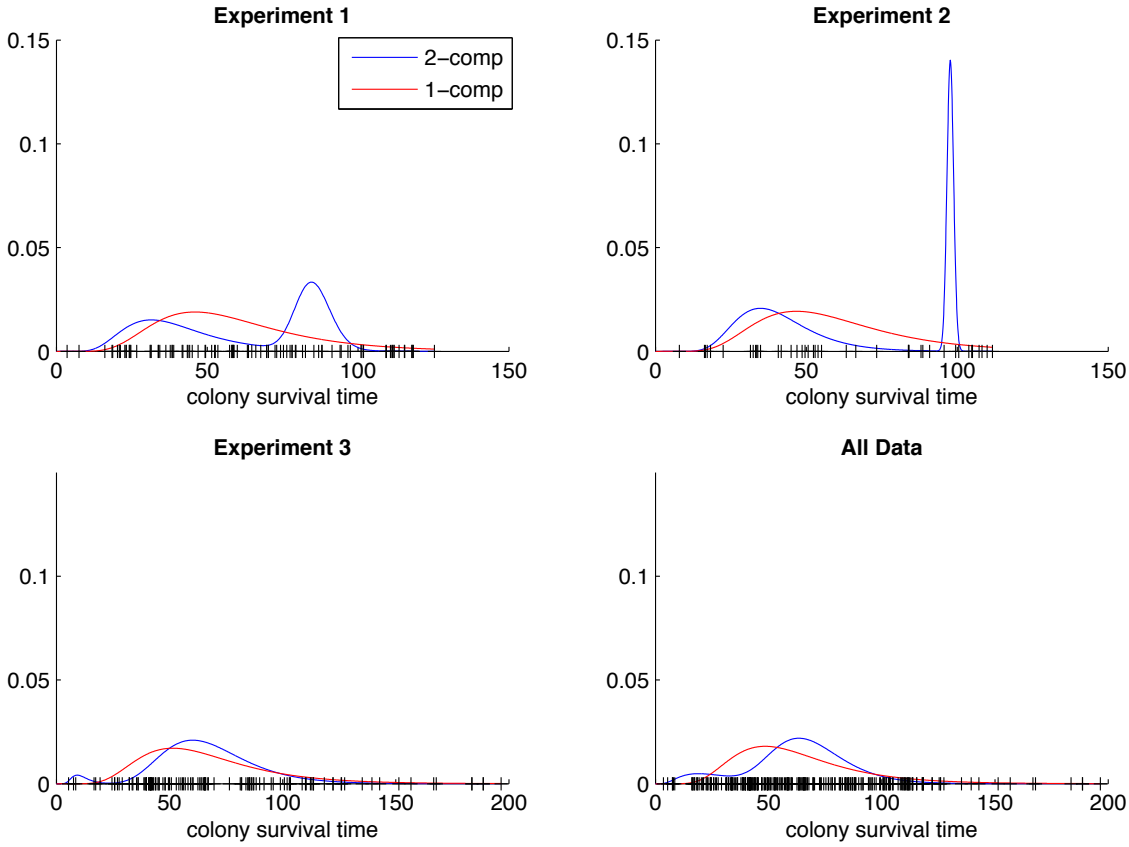


Figure 4.6: Experiment 1-3, All Data: Estimated densities for the colony survival time of cells of the first generation in the corresponding experiment under treatment conditions. Densities have been estimated by using maximum likelihood fits under the assumption of one lognormal distributed component (red) and two lognormal distributed components (blue), respectively. Exact values of parametrization are given in table 4.2. The p-values for the Null hypothesis of one component are $1, 1 \cdot 10^{-3}$, $2.2 \cdot 10^{-4}$, $5.7 \cdot 10^{-3}$ for the experiments 1-3 respectively and $3 \cdot 10^{-5}$ for all data. Black markers indicate observations considered for the inference.

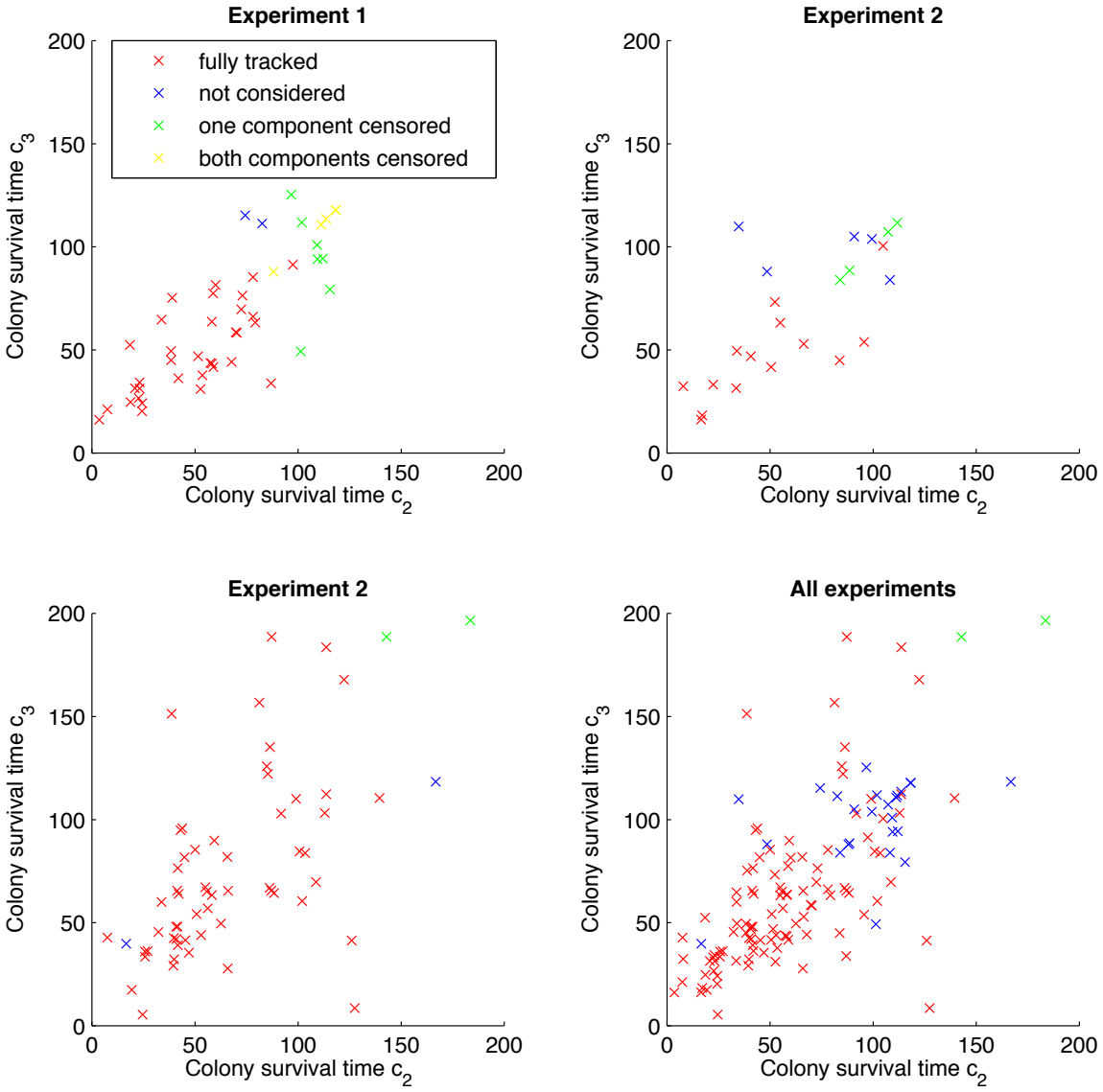


Figure 4.7: Data cleaning for model 3 inference: Experiment 1-3, All Data: Scatter plot describing the colony survival time t_{colony} of cell sibling pairs in generation 1 under treatment conditions. Each scatter point represents the colony survival of one cell sibling pair. Colony survival times of cell sibling pairs which have both been fully tracked are colored in red color. Colony survival times of cell sibling pairs where the tracking of at least one sibling's colony has been aborted before the end of the time lapse movie has been reached are not considered in the model inference and are colored in blue color. Colony survival times where the tracking of the colony survival time of both siblings terminated by the end of the time lapse movie are colored in yellow. Colony Survival times where the observation of at least one sibling has been terminated by the end of the time lapse movie and which fall not in one of the above mentioned categories are considered as censored and are colored in green color.

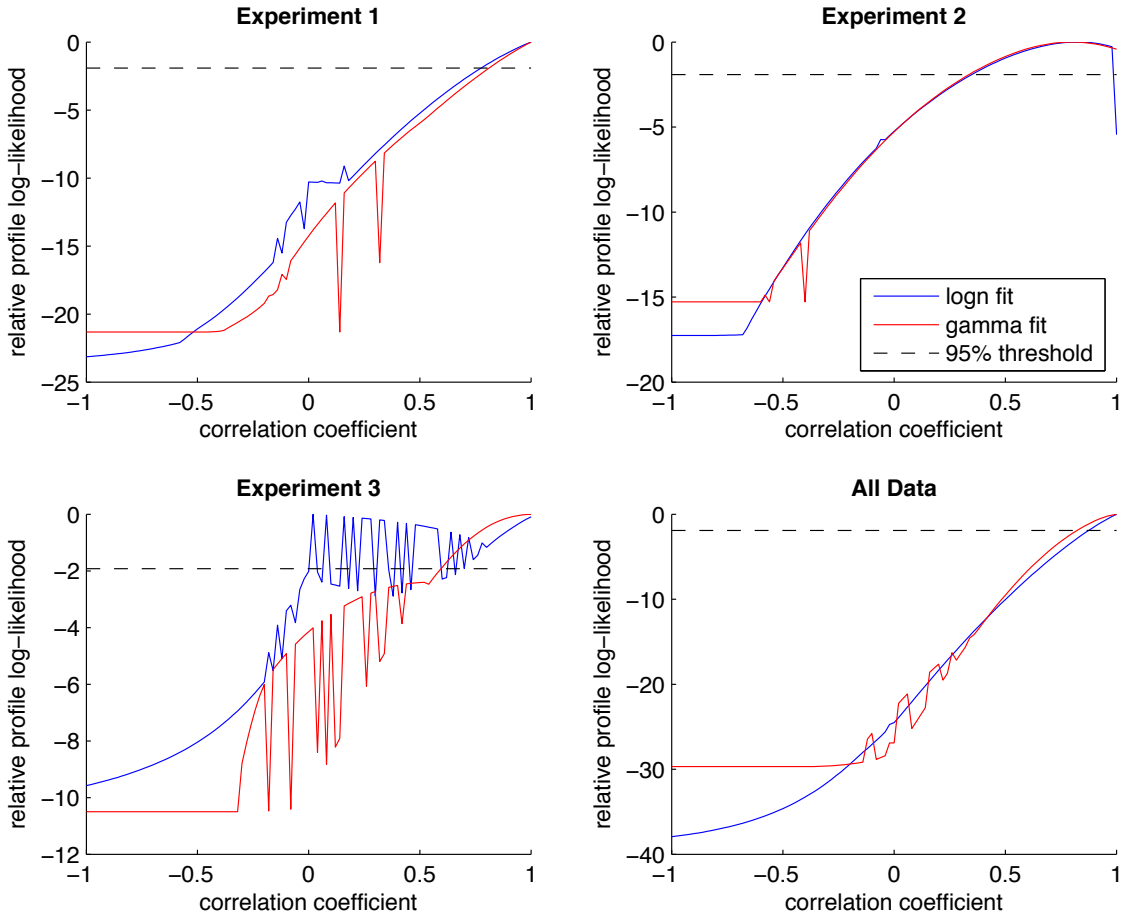


Figure 4.8: Parameter inference for special mixture model: Experiment 1-3, All Data: Relative profile log likelihoods with respect to correlation of the binary pairing describing the occurrence of differentiation events in sibling pairs in generation 1 under the assumptions of gamma distributed (red) and lognormal distributed (blue) components. Confidence intervals for the correlation comprise exclusively positive values in experiment 1,2 and in the combination of the experiments. Optimization of the profile log-likelihood in experiment 3 shows very bad convergency rates. Constant behavior of the profile likelihood for negative correlation values are due to the situation of a non identifiable one component best fit for the corresponding correlation level.

Chapter 5

Discussion

In the previous chapter we inferred the parametrization $\theta = (\theta_{11}, \theta_{01}, \theta_{00})^\top$, which describes the probabilities that both cells, one cell or none of the cells in a sibling pair remain HSC. In this part we will discuss the inferred results in context of our original question, which is whether we can infer statistical evidence for asymmetric cell division in haematopoietic stem cells. To do so, we first need to define how we link values of θ with asymmetric or symmetric division processes. We do this in context of a model describing the assumed underlying biological processes. We refer to this model as a the fate determining process. In the introduction we defined asymmetric cell division as an asymmetric segregation of fate determinants during the cell division process. In the following we formalize these terms as part of the fate determining process, which describes the self renewal fates of the sibling cells and which can be represented as a binary pairing.

5.1 Fate determining process

We assume that a certain protein functions as a fate determinant in the differentiation process. The self-renewal fate of a cell is determined by the amount of this fate determinant in the cell. During the division process this fate determinant can either be segregated symmetrically or asymmetrically along the daughter cells with probabilities β and $1 - \beta$, respectively. If the fate determinant is segregated asymmetrically, one daughter cell inherits the whole amount of this fate determinant and is therefore determined to maintain its self renewal capacity until the next division event. The other daughter cell does not inherit the fate determinant and therefore is determined to loose its self renewal capacity (Figure 5.1a). If the fate determinant is segregated symmetrically, it is distributed in approximately same amounts along the daughter cells. In this case, we assume, that the amount of the fate determinant in each daughter cell is neither enough to make it deterministically maintain its self renewal capacity nor is low enough to make it deterministically loose its self renewal capacity. Instead, the daughter cells are undecided. Each

daughter cell eventually maintains it self renewal capacity with a fixed probability π or looses it with probability $1 - \pi$ until the next division (Figure 5.1b).

Corresponding parametrizations

The above described fate determining process can be identified with a binary pairing with parametrization

$$\theta = \beta \begin{pmatrix} \pi^2 \\ 2\pi(1-\pi) \\ (1-\pi)^2 \end{pmatrix}^\top + (1-\beta) \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}^\top .$$

In case all cells divide symmetrically, i.e. $\beta = 1$ the parametrization θ of the binary pairing lies in $\mathcal{K}_0$ (Figure 5.1c). In case all cells divide asymmetrically i.e. $\beta = 0$ the parametrization of the binary pairing is $(0, 1, 0)^\top$ (Figure 5.1d). If the probability for a cell to divide asymmetrically is not zero i.e. $\beta < 1$, then the parametrization lies in $\mathcal{K}_-$ (Figure 5.1e). Note that values $\theta \in \mathcal{K}_+$ can not be associated with any valid parameter values $\beta \in [0, 1], \pi \in [0, 1]$ of this fate determining process!

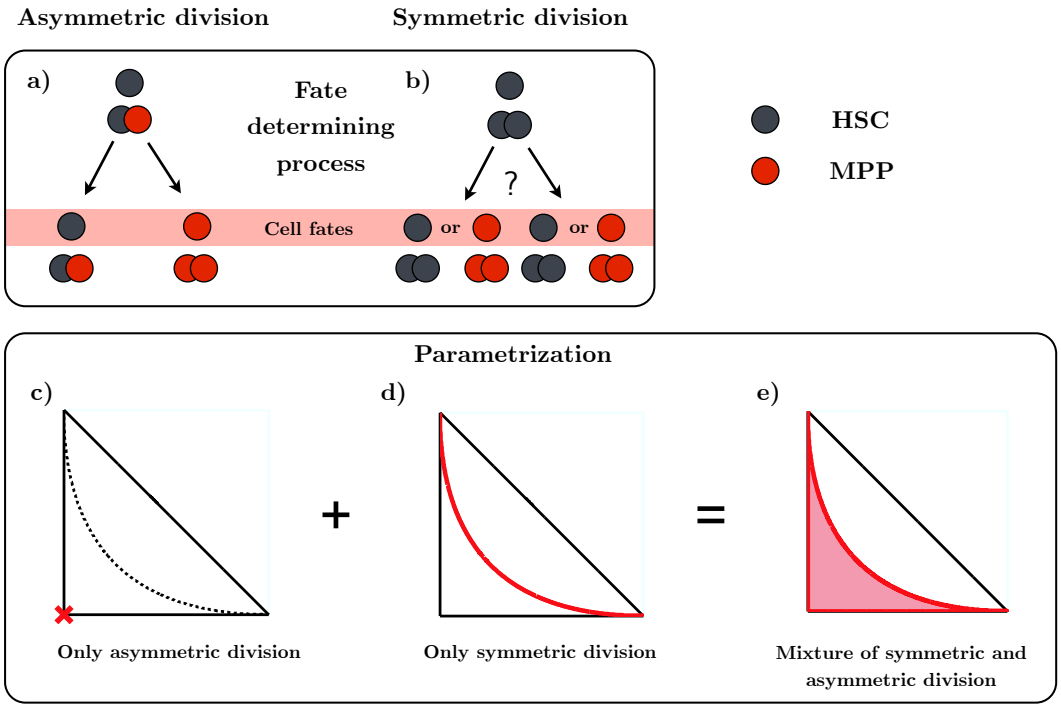


Figure 5.1: **Fate determinig process:** Grey and red colored balls indicate HSCs or MPP cells, respectively **a)** Cells dividing asymmetrically produce two cells of distinct cell fate by asymmetrical segregation of the fate determinant **b)** Cells dividing symmetrically produce two cells with undetermined cell fate. The daughter cells loose their self-renewal capacity independently from each other by chance. **Parametrization: c)** Parametrization of the corresponding binary pairings in case all cells divide asymmetrically, **d)** Parametrization in case all cells divide symmetrically, **e)** Parametrization in case of a mixture of asymmetrically and symmetrically dividing cells.

5.2 Composite model

In our application we performed the parameter inference for θ under assumption of certain dependencies between self-renewal and survival fates. We formally described these dependencies as the models 1-3. By combination of these models with the above described 3-state fate determining process, we get composite models (Figure 5.2). Within this framework parameter inference of θ can be interpreted in the following way. If there is statistical significance for $\theta \in \mathcal{K}_-$ i.e. the null-hypothesis $H_+ : \theta \in \mathcal{K}_+ \cup \mathcal{K}_0$ is rejected, then under assumption of the survival-self-renewal fate dependency on which the inference of θ is based and the fate determining process described above, we can conclude, that some of the observed cells divided asymmetrically i.e. there is evidence for asymmetric cell division. It can also happen that there is statistical significance for $\theta \in \mathcal{K}_+$. However, as described in section 5.1 parameter values $\theta \in \mathcal{K}_+$ can not be associated with any valid parametrizations of the fate determining process i.e. the real biological process determining the fate of the cells can not be described by the composite model. This can either be because

1. the assumptions about the survival fate self-renewal dependency is correct, but the assumption of the fate determining process is wrong.
2. the assumptions of the fate determining process is correct, but the assumption about the survival fate self-renewal dependency is wrong.
3. the assumptions of both model parts are wrong.

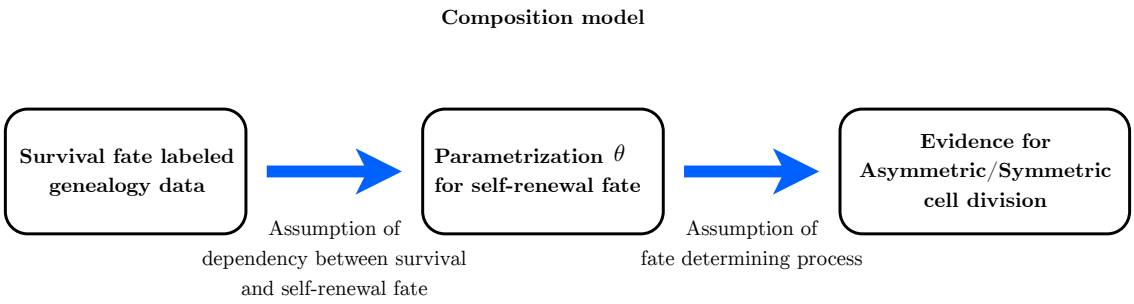


Figure 5.2

5.3 Discussion of inferred results

Regardless which self-renewal fate-survival fate dependency was assumed (model1-3), in almost all cases we found statistical significance for $\theta \in \mathcal{K}_+$. In few cases the null-hypothesis $H_0 : \theta \in \mathcal{K}_0 \cup \mathcal{K}_-$ was not rejected i.e. the fates of sibling cells are positive correlated. There was no case, where we found statistical significance for $\theta \in \mathcal{K}_-$.

That means that in almost all cases of inference the real biological process determining the fate of the cells can not be described by the corresponding composite model. In particular, there is no case, where we can deduce evidence for asymmetric cell division. Under the assumption of our models used for the inference of θ, we can conclude, that we can not describe the self-renewal fates of sibling cells with this fate determining process. Possible explanation for this case are cell to cell communication, extrinsic factors, or any effect, which makes sibling cells choose correlated fates. However it is also very likely that we inferred this results because of to simple assumptions for our parameter inference. Here, we discuss potential shortcomings of the presented methods.

5.3.1 Contamination of experimental data

As mentioned in Chpater 4, the cells in the experiments used in our analysis are assumed to comprise 30%-70% HSCs[16]. The other cells in the experiment are assumed to be MPP cells or cell with even lower differentiation potential. However, in all models we assume that the mother cells of the cell sibling pairs are HSCs. Therefore, we presumably accounted up to 70% percent of the cells falsely as HSCs. This might bias the estimate of the correlation in the inference on the first generation into positive direction: It seems likely that the high amount of trees, where both siblings died in generation 1 are a consequence of this contamination. This is supported by the fact that in experiment 4b which comprises only MPP cells, all cells died at latest in generation 1 (Figure 3.3).

5.3.2 Missing vertical dependencies

Contamination of the experimental data might explain the symmetric results for the estimate of θ in the first generation, but it is unlikely that it has effect on the inference of θ in higher generations. However, HSCs, which differentiated to MPP cells in the first generation can be assumed to have a similar contamination effect on the inference of θ i.e. cells, which differentiated to MPP in generation 1 will be falsely accounted as HSC for the inference of θ in the second generation. As we demonstrated in Figure 5.3, such delay effects can completely distort the results of the inference of θ. The reason why we neglected this kind of effects so far is that modeling of vertical dependencies between cells of different generations requires a more general approach, than binary pairings, which can only be used to model horizontal dependencies between sibling cells. In the outlook chapter of this thesis, we will briefly sketch a suitable approach which can be seen as generalization of binary pairings.

5.3.3 Mixture effects

For a given sample $x^{(1)}, \ldots, x^{(n)}$ inference of the parameter θ only makes sense, if we can assume that $x^{(1)}, \ldots, x^{(n)}$ are realizations of a random sample of binary pairings $X^{(1)}, \ldots, X^{(n)}$, where all sample components are independent and identically distributed. If the underlying sample

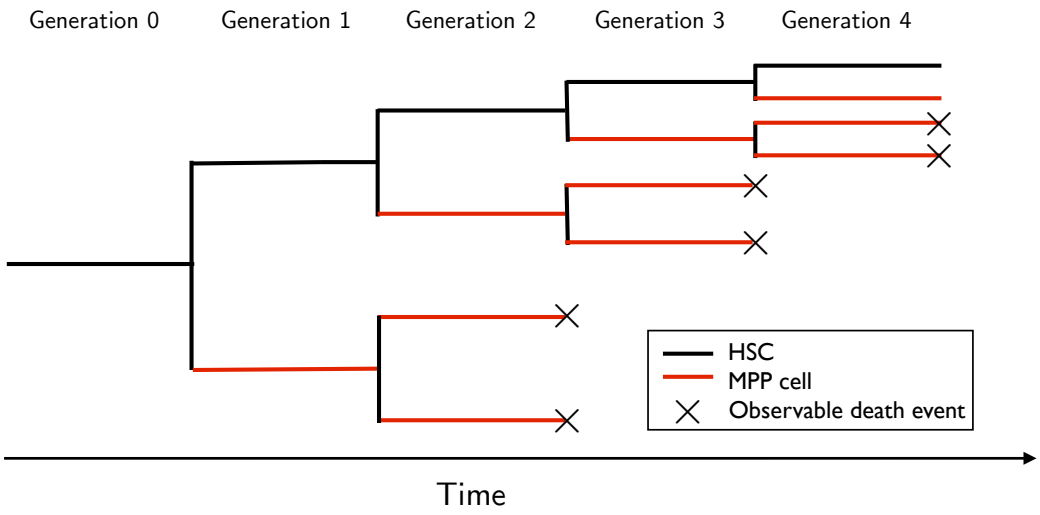


Figure 5.3: Illustration of a process with vertical dependencies : Generations are enumerated relative to the start of observation. As in our applications the self-renewal fate of cells is assumed to be not observable. Death events can be observed. HSC are assumed not to die and divide asymmetrically producing two cells of distinct self-renewal fate i.e. $\theta = (0, 1, 0)$. MPP cells resulting from an asymmetric cell division divide once producing two daughter cells which die before dividing. Analysis with model 1 of a sample based on this process would yield a maximum likelihood estimate for θ of $(0.5, 0, 0.5)^\top$, which is a perfectly symmetric parametrization. Apparently this result of inference would be totally wrong!

components $X^{(1)}, \ldots, X^{(n)}$ are not identically distributed, it follows from theorem2.4.1 that the inference of θ yields systematic overestimation of the corresponding correlation between the binary pairing components, which can easily result in $\theta \in \mathcal{K}_+$. In the following we present two cases, where we have evidence that the assumption of identically distributed components in our samples is wrong.

Heterogenity in the HSC population

The basic assumption for all our models so far is that we can model the self-renewal fate as a dichotomous feature and that the self-renewal capacity of a cell fully characterize a HSC which justifies that we describe the fates of daughter cells of HSC with identical distributed binary pairings. However the assumption of homogeneity within the haematopoietic stem cell population contradicts with experimental results, which have shown, that haematopoietic stem cells differ widely in their behavioral and molecular properties [16]. Therefore, it is questionable to model the fate of the daughter cells of HSC by identical distributed binary pairings. A falsely assumed homogeneity can lead to an overestimation of the correlation between the fates of the daughter cells [2].

Time dependency of the marginal probability π

In order to avoid mixture effects, we already performed the inference of θ in model 1 separately on all generations. We showed that the marginal survival probability π is significantly decreasing with higher generation. However, since the time point at which a cell division occurs and the generation of the corresponding mother cell are highly correlated, we can presume that this decrease might origin in a time dependency of the marginal survival probability π. Since division events in each generation are widely dispersed over time, this would also cause a mixture effect for the inference of θ. In order to test a time dependency of the marginal survival probability we performed for each generation and experiment a logistic regression (Figure 5.4) using the time points of divisions as predictor variables and the number $k = 0, 1, 2$ of corresponding daughter cells which divide as response variables. The results are very different among the experiments, but at least in the regression over the data of all experiments, the decrease of the marginal survival probability π over time is significant in all generations. That means, that in most cases we have a decrease of the survival probability for the daughter cells within one generation i.e. cells origin from cell divisions occurred early in the experiment have a higher probability to survive, than cells in the same generation originating from later division events. We presume, that this kind of mixture effects can be solved with a suitable regression approach where we assume a constant correlation coefficient over time.

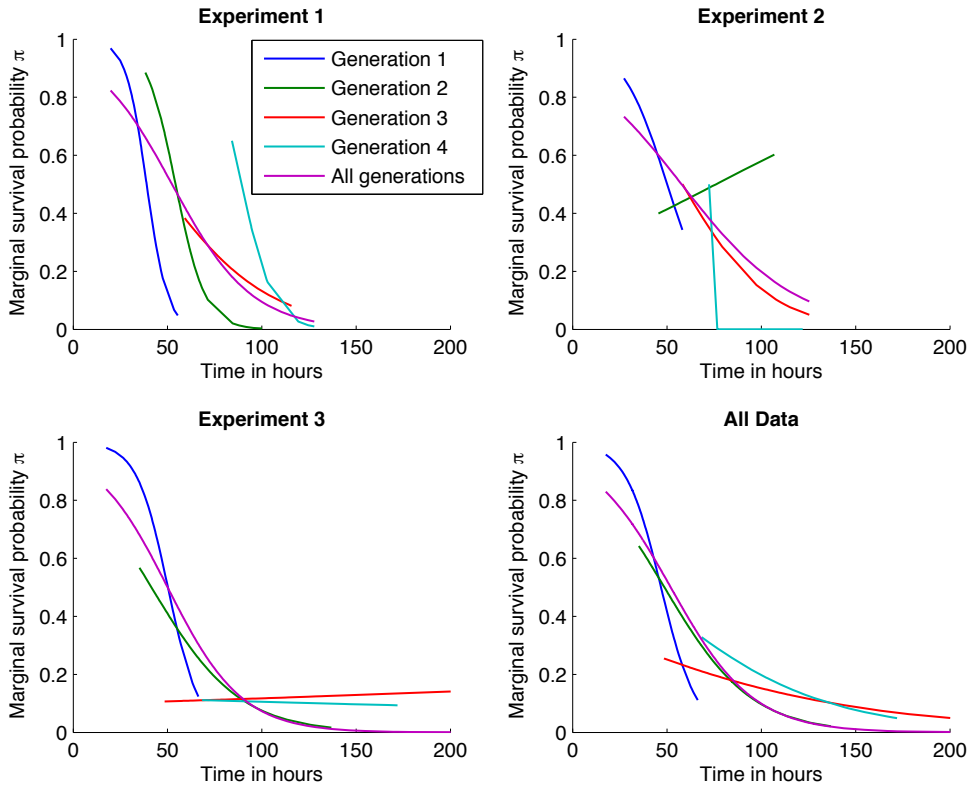


Figure 5.4: Logit fits for the marginal survival probability π: Experiment 1-3, All data. Time points of divisions of cells in generation 1-4 and all generations, respectively are used as predictor variables and the number $k = 0, 1, 2$ of corresponding daughter cells which divide are used as response variables. The result is very different among the experiments. In the regression over the data of all experiments the decrease of the marginal survival probability π over time is significant in all generations. It follows from the corresponding p-values, that in most cases of decreasing fits this behavior is significant.

Chapter 6

Outlook

As we have shown in chapter 5 the attempt to show evidence for asymmetric cell division by a quantitative analysis on cellular genealogies faces many problems. Some of these problems might be solved in the future with better experimental methods (section 5.3.1) and some might be unfeasible, like a strong heterogeneity within the HSC population (section 5.3.3). We showed by a simple example that ignoring vertical dependencies within the genealogy can lead to wrong inference results. In order to cope with such dependencies between cells of distinct generations, we need a more general stochastic description for cellular genealogies. A very general approach is to model the development of a cellular genealogy as a probability measure on the labeling space of an infinite binary labeled tree. In order to be consistently defined, such measures need to be invariant under graph isomorphisms, since sibling cells are not ordered in cellular genealogies. Such probability measures can be defined recursively by a markov random field structure. Based on this idea, one could extend model 1 such that MPP cells do not deterministically die before division but with a certain probability produce two dying daughter cells, therefore delaying the cell death for one generation. Using this approach we account for such delay effects that cause artificial symmetry and might finally find evidence for asymmetric cell division.

Bibliography

[1] P.M. Agapow and A. Purvis. Power of eight tree shape statistics to detect nonrandom diversification: A comparison by simulation of two models of cladogenesis. *Systematic Biology*, 51(6):866–872, 2002.

[2] F.B. Christiansen. The wahlund effect with overlapping generations. *The American Naturalist*, 131(1):149–156, 1988.

[3] T.H. Emigh. A comparison of tests for hardy-weinberg equilibrium. *Biometrics*, pages 627–642, 1980.

[4] ZD Feng and CE McCulloch. On the likelihood ratio test statistic for the number of components in a normal mixture with unequal variances. *Biometrics*, pages 1158–1162, 1994.

[5] I. Glauche, M. Cross, M. Loeffler, and I. Roeder. Lineage specification of hematopoietic stem cells: mathematical modeling and biological implications. *Stem Cells*, 25(7):1791–1799, 2007.

[6] I. Glauche, R. Lorenz, D. Hasenclever, and I. Roeder. A novel view on stem cell development: analysing the shape of cellular genealogies. *Cell proliferation*, 42(2):248–263, 2009.

[7] L.A. Goodman. On simultaneous confidence intervals for multinomial proportions. *Technometrics*, pages 247–254, 1965.

[8] L. Held and D. Sabanés Bové. Methoden der statistischen inferenz. *Likelihood und Bayes. Heidelberg: Spektrum Akad. Verl*, 2008.

[9] L. Hosking, S. Lumsden, K. Lewis, A. Yeo, L. McCarthy, A. Bansal, J. Riley, I. Purvis, and C.F. Xu. Detection of genotyping errors by hardy–weinberg equilibrium testing. *European Journal of Human Genetics*, 12(5):395–399, 2004.

[10] M. Kirkpatrick and M. Slatkin. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution*, pages 1171–1181, 1993.

[11] A.O. Mooers and S.B. Heard. Inferring evolutionary process from phylogenetic tree shape. *Quarterly Review of Biology*, pages 31–54, 1997.

[12] R.A. Neumüller and J.A. Knoblich. Dividing cellular asymmetry: asymmetric cell division and its implications for stem cells and cancer. *Genes & development*, 23(23):2675–2699, 2009.

[13] J. Neyman and E.S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933.

[14] O. Nikodym. Sur une généralisation des intégrales de mj radon. *Fund. Math*, 15:131–179, 1930.

[15] M.A. Rieger, P.S. Hoppe, B.M. Smejkal, A.C. Eitelhuber, and T. Schroeder. Hematopoietic cytokines can instruct lineage choice. *Science's STKE*, 325(5937):217, 2009.

[16] T. Schroeder. Hematopoietic stem cell heterogeneity: subtypes, not unpredictable behavior. *Cell stem cell*, 6(3):203–207, 2010.

[17] S.B. Ting, E. Deneault, K. Hope, S. Cellot, J. Chagraoui, N. Mayotte, J.F. Dorn, J.P. Laverdure, M. Harvey, E.D. Hawkins, et al. Asymmetrical segregation and self-renewal of hematopoietic stem and progenitor cells with endocytic ap2a2. *Blood*, 2011.

[18] EO Wiley. *The theory and practice of phylogenetic systematics*. Wiley Online Library, 1981.

[19] S.S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

[20] M. Wu, H.Y. Kwon, F. Rattis, J. Blum, C. Zhao, R. Ashkenazi, T.L. Jackson, N. Gaiano, T. Oliver, and T. Reya. Imaging hematopoietic precursor division in real time. *Cell Stem Cell*, 1(5):541–554, 2007.