Turun yliopisto
University of Turku

STATISTICAL METHODS FOR THE ANALYSIS OF HIGH-CONTENT ORGANOTYPIC CANCER CELL CULTURE IMAGING DATA

Ilmari Ahonen

TURUN YLIOPISTON JULKAISUJA - ANNALES UNIVERSITATIS TURKUENSIS
Sarja - ser. AI osa - tom. 574 | Astronomica - Chemica - Physica - Mathematica | Turku 2018

Turun yliopisto
University of Turku

STATISTICAL METHODS FOR THE ANALYSIS OF HIGH-CONTENT ORGANOTYPIC CANCER CELL CULTURE IMAGING DATA

Ilmari Ahonen

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
Sarja - ser. A I osa - tom. 574 | Astronomica - Chemica - Physica - Mathematica | Turku 2018

University of Turku

Faculty of Mathematics and Natural Sciences

Department of Mathematics and Statistics

Doctoral Programme in Mathematics and Computer Sciences

Supervised by

Professor Jaakko Nevalainen, PhD
School of Health Sciences
University of Tampere, Tampere, Finland

Adjunct Professor Matthias Nees, PhD
Institute of Biomedicine,
University of Turku, Turku, Finland

Reviewed by

Professor Mu Zhu, PhD
Department of Statistics & Actuarial Science
University of Waterloo, Ontario, Canada

Assistant Professor Antti Honkela, PhD
Helsinki Institute for Information Technology HIIT
Department of Mathematics and Statistics
University of Helsinki, Helsinki, Finland

Opponent

Professor Fred Godtliebsen, PhD
Department of Mathematics and Statistics
Faculty of Science and Technology
University of Tromsø, Tromsø, Norway

ISBN 978-951-29-7111-4 (PRNT)

ISBN 978-951-29-7112-1 (PDF)

ISSN 0082-7002 (Print)

ISSN 2343-3175 (Online)

Painosalama Oy - Turku, Finland 2018

Abstract

Organotypic cancer cell cultures combined with modern imaging technology have greatly expanded the possibilities of *in vitro* cancer research and drug development. In fact, imaging and subsequent image analyses have become a main component for high content screening in early stage drug discovery. The scale of such screening campaigns is rapidly growing, while at the same time, cell cultures become increasingly complex and now also include multicellular organoids in three-dimensional cultures. As a result of these imaging experiments, large amounts of image data are generated, posing ever-increasing demands to the related analysis methodology. In this doctoral thesis, novel and efficient statistical methods are introduced to meet these demands, spanning a variety of research topics in both statistics and machine learning. As a starting point, the preprocessing and segmentation of the image data are described, leading to the statistical analysis of treatment effects through descriptive features of the multicellular structures. A novel flexible finite mixture regression model is introduced in this context to account for the intra-tumor heterogeneity in the cultures. To gain a more direct interpretation for the treatment effects, an unsupervised analysis sequence is proposed leading to the phenotypic grouping of the cell structures. This is achieved by using a selected set of feature principal components as inputs for clustering algorithms. Finally, the problem of global level novelty detection is formulated and tackled with permutation tests. While the feature analysis and clustering approaches deal with very specific applications, the flexible FMR and global level novelty detection methods represent more abstract problems that are inspired by the challenges in image analysis but are not directly motivated by them. The application of all methods is demonstrated with a real cancer culture dataset in the introductory part of this thesis.

Tiivistelmä

Organotyyppiset syöpäsoluviljelmat ja moderni kuvantamisteknologia ovat merkittävästi lisänneet syövän tutkimisen ja lääkekehityksen mahdollisuuksia *in vitro* -ympäristössä. Kuvantamiskokeiden tuotteena syntyy suuria määriä kuvamuotoisia aineistoja, jotka asettavat alati kasvavia vaatimuksia analyysimenetelmille. Tässä väitöskirjassa esitellään uusia ja tehokkaita tilastollisia menetelmiä vastaamaan näihin haasteisiin. Menetelmät koskettavat laajasti erilaisia tutkimusongelmia sekä tilastotieteen että koneoppimisen aloilta. Tarkastelu aloitetaan kuva-aineistojen esikäsittelyyn ja segmentointiin sopivista menetelmistä, jonka jälkeen käsitellään lääkeaineiden vaikutusten mallintamista syöpäsolurakenteita kuvailevien piirteiden avulla. Kasvaimen sisäisen heterogeenisyyden huomioonottamista varten esitellään uusi joustava latenttien luokkien regressiomalli. Suorempi menetelmä lääkkeiden vaikutusten arvioimiseksi johdetaan ryhmittelemällä solurakenteet niiden fenotyyppejä vastaaviin luokkiin ohjaamattomilla menetelmillä.
Tämä saavutetaan käyttämällä rakenteita kuvaavien piirteiden pääkomponentteja muuttujina ryhmittelyalgoritmeille. Lopuksi määritellään globaalin tason puoliohjattu poikkeavien havaintojen tunnistusongelma, jonka ratkaisuksi esitellään permutaatioperiaatteeseen perustuvia testejä. Kuvailevien piirteiden ja niistä johdettujen fenotyyppiryhmien analysointiin esitellyt menetelmät liittyvät erityisesti käsiteltyyn sovellukseen, kun taas joustava latenttien luokkien regressio ja poikkeavien havaintojen tunnistusongelma ovat luonteeltaan yleisempiä ja enemmänkin kuva-analyysin inspiroimia menetelmiä. Riippumatta menetelmien luonteesta niiden kaikkien soveltamista esitellään todellisen syöpäsolujen kuvantamisaineiston kanssa tämän väitöskirjan johdanto-osassa.

Acknowledgements

The work towards this thesis started in 2012 at the Department of Mathematics and Statistics in close collaboration with VTT Technical Research Centre of Finland with the goal of developing innovative statistical methodology for the analysis of the large cancer cell imaging datasets. Many changes to this initial configuration have taken place since then but thankfully the destination of the thesis and our crew of experts remained, for the most part, intact through the rough seas. As is often the case with great journeys, the path was more important than the actual destination. I'm truly lucky to say, that my path was rich with inspirational people and memorable experiences.

First and foremost, I wish to thank my primary supervisor Professor Jaakko Nevalainen for his seemingly infinite pools of encouragement, inspiration, patience and determination. It has truly been a pleasure to sail under the guidance of such a seasoned and able captain. I thank doctor Matthias Nees, whose ideas and vision kept our course directed to uncharted waters. I give my special thanks to Professor Denis Larocque, who became both my guide and friend in a foreign port.

I wish to thank Professor Mu Zhu and Assistant Professor Antti Honkela for their thorough review of the thesis and their encouraging comments.

A successful voyage requires a good vessel with quality equipment and sound rules of conduct. I acknowledge the heads of the Deparment of Mathematics and Statistics, Professor Juhani Karhumäki and Professor Iiro Honkala, for the outstanding research facilities and funding towards the early phases of my thesis. I thank Professor Mervi Eerola for her committed management of the Statistics wing of the Department. I acknowledge Auli Raita for her coordination and care over the years at VTT. Furthermore, I wish to thank the Emil Aaltonen Foundation, Jenny and Antti Wihuri Foundation, the Magnus Ehrnrooth Foundation, the MATTI Doctoral Programme, the Natural Sciences and Engineering Research Council of Canada, the Oscar Öflund Foundation, the Turku University Foundation and, last but not least, my grandpa Lauri for their financial support.

One is truly lucky to have been surrounded by such a crew of colleagues as I have during these past years. I sincerely thank all the people of the Department of Mathematics and Statistics and the late Turku branch of VTT for your friendship and support. I give my thanks to Ville Härmä and Sean Robinson with whom I've had the pleasure of sharing common quarters with. Both joy and frustration tastes better when shared. Furthermore, I wish to thank all my colleagues for the insightful and pleasant coffee break and lunch hour discussions. Finally, I give my special thanks to my dear friend Maiju Pesonen, who bears much of the blame for getting me into this mess to begin with.

My warmest thanks belong to the people in my home port. I thank my

mother Tiina for her never-ending support and encouragement towards education. I thank my partner Camilla for all her love and understanding during the darkest hours of research. Finally, I wish to thank all my friends for showing unwavering interest in my thesis, even when not having the slightest idea of what I was actually doing.

Turku, 4.12.2017

Ilmari Ahonen

Contents

Notation

x	a scalar
$\boldsymbol{x}$	a (column) vector
$\boldsymbol{X}$	a matrix
$\boldsymbol{X}'$	the transpose of $\boldsymbol{X}$
$\log(x)$	the natural logarithm of x
$\sim$	distributed as
$\approx$	approximately
$I(.)$	indicator function

List of publications

This thesis consists of an introduction and the following original publications:

I Ahonen I, Härmä V, Schukov H, Nees M and Nevalainen J (2016): Morphological Clustering of Cell Cultures Based on Size, Shape and Texture Features, *Statistics in Biopharmaceutical Research*, 8(2), 217-228

II Ahonen I, Åkerfelt M, Toriseva M, Oswald E, Schüler J and Nees M (2017): A high-content image analysis approach for quantitative measurements of chemosensitivity in patient-derived tumor microtissues, *Scientific Reports*, 7, article number: 6600

III Ahonen I, Larocque D and Nevalainen J (2015): Global tests for novelty, *Statistical Methods in Medical Research*, 26(4), 1867-1880

IV Ahonen I, Nevalainen J and Larocque D: Prediction with a flexible finite mixture-of-regressions (Under revision)

Part I

Introduction

1 Background

This thesis introduces novel statistical methodology inspired by the analysis of high-content imaging data from cancer cells and – most importantly – the three-dimensional structures such cancer cells form in so-called "organotypic cell cultures". These datasets are the result of cell culturing experiments where the main goal is to either study the efficacy of various drugs or to further the understanding about the disease itself. A wide scope of problems is associated with the analysis of these data, ranging from the initial pre-processing of the images to the final estimation of the treatment effects. While some of the methods in this thesis are directly designed to meet these challenges, the others tackle problems that are more general in nature and have applications beyond the field of cell-based imaging.

This thesis consists of an introduction and four publications. The introduction outlines the main steps in the analysis of the imaging datasets and considers the developed methods in this context. The focus is placed on the main results of the publications and in their application, without delving too deeply into their technical details. A summary of the broad research field is given and the contributions of the developed methodology are discussed in relation to this wider context.

This introductory part is organized as follows. In Chapter 1, a brief introduction to cell culturing and the related methodological challenges is given. The chapter is ended with a description of an example dataset that is utilized in the following chapters for illustrative purposes. Chapter 2 concerns the pre-processing of the image data, including both the segmentation of individual multicellular structures and their morphological characterization with numerical features. Chapter 3 presents various approaches to treatment effect estimation based on the extracted image features. A novel flexible finite mixture regression model is introduced in this context. Chapter 4 considers the novel problem of global level novelty detection and its utility in the cell culture image analysis. Finally, the Introduction is concluded with a brief summary of the main findings of the thesis in Chapter 5.

1.1 A biological model for cancer

Organotypic cancer cell cultures accompanied with efficient imaging technology have become a routine practice in modern laboratories to study the behavior and treatment of cancers *in vitro* [18, 29, 57]. This is now widely

accepted as a superior alternative to the two-dimensional cell cultures on artificial plastic plates. Under latter conditions, cells typically do not form any tissue-like structures but are mainly concerned with proliferation (cell division and growth). In contrast, cells cultured in extracellular matrix (ECM) form complex multicellular structures with distinct morphologies that directly relate to the behavior and characteristics of the disease [41]. For instance, Härmä et al. [25] associate round and consistent structures to benign and non-aggressive tumors. In contrast, structures with invasive outreaching "appendages" and irregular, inconsistent shapes are considered as indicators of malignant and invasive growth. By observing and quantifying these different morphologies, new insight on the mechanics of cancer and on the effects of treatments can be obtained, given that the required analysis methodology is available.

The representativeness of the cell cultures is naturally a critical question: how well do the artificial conditions in the laboratory resemble the true tumor micro-environment in the patient? As in many other fields of technology, huge advances have been made in this regard during the last decade. The majority of cell culturing has been traditionally performed on rigid and flat plastic substrates, in an essentially two-dimensional environment [18]. Now, this setup is widely considered suboptimal, as the cells are able to form multicellular structures but lose many crucial properties such as differentiation and cell-to-cell communication [25]. By the time the research towards this thesis was started, three-dimensional cell models had been established providing greatly improved relevance compared to the traditional monolayer cultures [29]. In these 3D micro-environments, cells are able to differentiate and freely grow into the surrounding ECM forming truly three-dimensional organoids that resemble similar structures that are observed in clinical tumors from real patients. The recent developments in cell models largely build onto this 3D-approach and aim at adding new components to the environment. A real tumor does not grow in a void but is instead surrounded by other cell types such as inflammatory immune cells and endothelial cells which form blood vessels. Arguably most important of these are the so-called stromal tissues, such as cancer-associated fibroblasts (CAFs) [85] that form the connective tissue of tumors and surround the actual tumor cells or organoids. Further improved clinical relevance is thus obtained in 3D co-culture models, where CAFs are introduced to the environment together with the cancer cells. The most recent advances however, relate to the actual cells that are utilized for *in vitro* cultures. Traditionally, the cultures rely on so called cancer cell lines that are originally extracted from real tumors, but are highly adapted by selection *in vitro* to behave particularly well under laboratory conditions [54]. These cell lines have also acquired additional genetic mutations that are not related to the original tumors from which they were isolated, therefore their relevance for

the actual disease is diminished after decades of *in vitro* culture. Examples include the human prostate carcinoma cell lines PC3 [39] from 1979 and DU-145 [73] from 1978, and the prostate adenocarcinoma cell line LNCaP [32] from 1977. Over the years, cell lines have brought enormous insights to the field and have formed a standard for both method and treatment comparisons. One of the promises of these organotypic models however, also relates to personalized medicine: if patient-specific samples could be efficiently and robustly cultured in the laboratory, this could greatly contribute towards more customized and optimal treatment [18, 49, 54]. One of the first steps towards this goal in the 3D-environment have now been taken in Publication II using so called patient-derived xenograft (PDX) cultures that utilize cell material directly extracted from the patient. These are essentially co-cultures, since the samples naturally contain both tumor and fibroblast cells. Publication II presents a multidisciplinary examination of the biological, technical and computational aspects involved in the design and analysis of a non-small cell lung carcinoma PDX experiment in the 3D-environment. Finally, examples of the image data obtained from the various 3D models are illustrated in Figure 1.1 while the increasing number of publications in the field is illustrated in Figure 1.2.

There is no single approach to conducting 3D culturing experiments, but instead, the type of the resulting image data depends on the adopted laboratory practices and the utilized equipment. Next, we give a general outline of the cell culture and imaging practices leading to the specific kind of data that this thesis is focused on. A more detailed description of the procedure is provided by Härmä et al. [26]. The cells are cultured on plastic plates containing multiple wells, ranging generally from 96 to 384 wells per plate. The cells inside a given well are subjected to consistent growth conditions, thus on a 96-well plate one could in theory test 96 different drugs. In the case of three-dimensional cell cultures, this requires also the deposition of equal amounts of extracellular matrix into each well, and seeding identical numbers of cells into the matrix of every well (approximately 750 to 1000). After a growth period of around 10 days (for PC3 cells), the three-dimensional cultures are ready to be imaged with a confocal microscope. Data could be collected also from earlier time points if longitudinal measurements are sought for. Instead of taking a single image of a given spot, the microscope images multiple cross-sectional slices of the culture at varying vertical levels in order to capture as much of the material as possible. This results in an image stack that is, in a sense, a three-dimensional observation of the entire culture at the location of imaging. However, the resolution of the images as well as the number of images forming a stack is usually far too low for a true 3D representation of the cultures, as is also the case with the datasets considered here. A full 3D representation is however not required, since we are primarily interested in the shape, size and texture of the organoids and

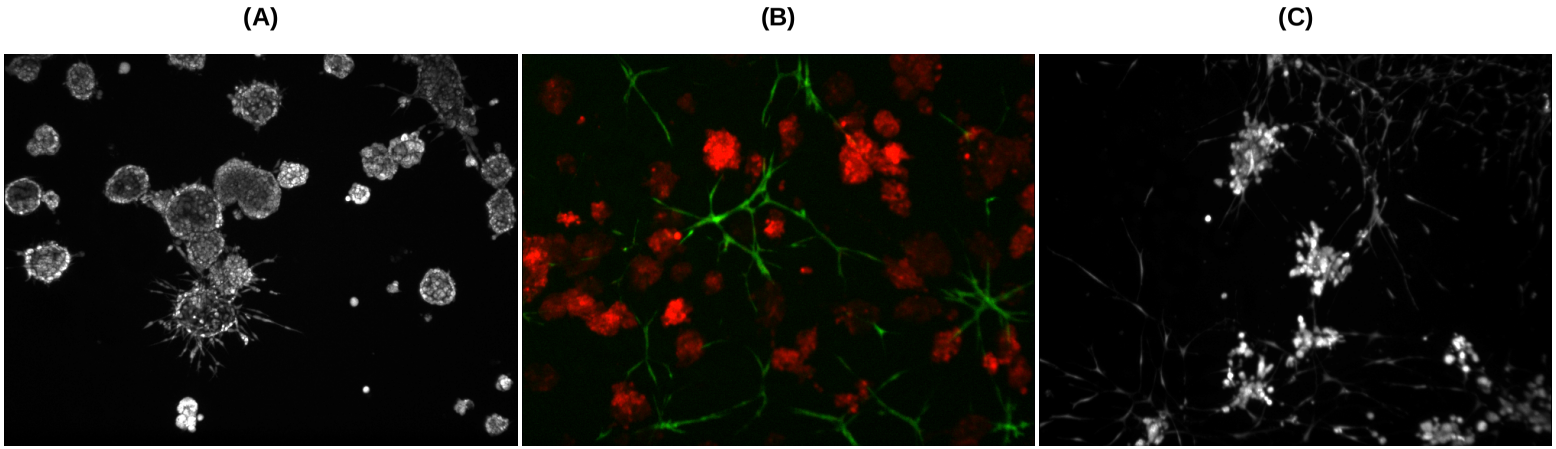

Figure 1.1: ***Examples of images obtained with the various organotypic 3D cancer models.*** ***A:*** *Structures cultured from the PC3 cell line. Both round spheroids and structures with invasive appendages are visible.* ***B:*** *Structures from a LNCaP-CAF co-culture. The cancer cell line (red) and the fibroblasts (green) are labeled with differing colors (fluorescent proteins that are stably expressed in these cells) for easy separation of the cell types.* ***C:*** *Structures formed by patient-derived xenograft cells. Both compact cancer cell structures and web-like fibroblast structures are present.*

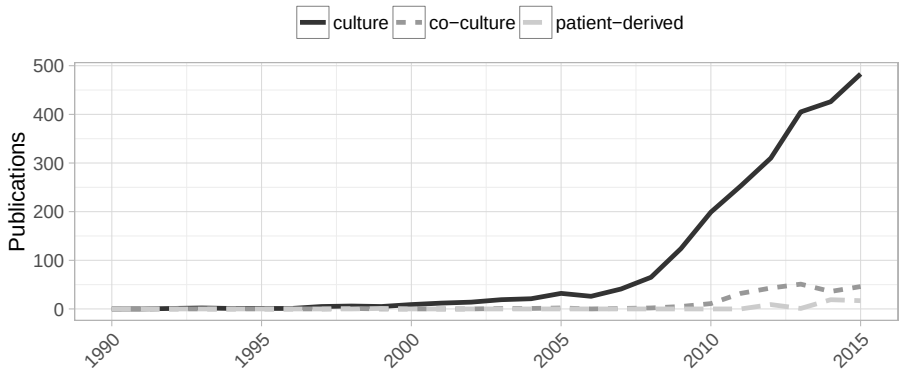

Figure 1.2: ***Yearly number of publications in the cancer 3D culturing field.*** *The data was collected by counting the number of hits in the Web of Science™ publication search. In order to be counted in, the articles needed to include the words "3D", "cancer" and "cell" in their title, abstract or in the listed keywords. Three additional search words "culture", "co-culture" and "patient-derived" were tested to produce the plotted data.*

whether they form invasive structures (appendages) or not. Consequently, the focus of this thesis is limited only to the analysis of two-dimensional projections of the stacks. Specifically, from a set of pixels located on top of each other in the stack, the pixel with the highest brightness is chosen for the final projected image. This is known as *maximum-intensity projection* or simply *max-projection*. The three example images in Figure 1.1 are obtained exactly this way. While it is acknowledged that relying on such a projection risks losing some of the information content in the original stack, it is considered a justifiable sacrifice in the high-throughput setting: the 2D-projections are computationally more manageable, while possible biases resulting from the simplification are considered negligible in large datasets. The problem is further alleviated by the fact that the structures are initially seeded in an approximately 2D-plane embedded between two layers of the extra-cellular matrix (ECM), thus aiming to minimize the chance of possible vertical overlap.

The grayscale images considered here are represented as matrices, where 0 denotes a fully black pixel and 1 a fully white pixel. Any values between 0 and 1 then define the various shades of gray. These values are hereafter simply referred to as *pixel values* or *pixel intensities*.

1.2 Data-analytical challenges

The growing complexity of the cell models has placed ever-increasing demands on the related analysis methodology. While a wound-healing assay on a monolayer 2D culture could be addressed with simple measurements of culture area [27], a 3D co-culture assay requires much more detailed approaches. The sequence of images A-C in Figure 1.1 clearly demonstrate this gradual increase in difficulty. The three-dimensional organoid structures formed by PC3 cells shown in A are more or less coherently shaped and sized, and can be visually distinguished from each other with relative ease. The main interest in the analysis of such data would revolve around the characterization of the structures' morphologies and how these are affected by different drug treatments. Similar goals can also be sought when tackling the data in B, but with the additional considerations that relate to the typical branching fibroblast structures. In addition to individual analyses of the two cell types, extra effort has to be placed on understanding their interaction. As in image B, the patient-derived cells in C form a co-culture of cancer cell structures and cancer-related fibroblasts but with one crucial difference: the two cell types are not labeled with differing colors that would allow for easy separation of the cell classes. Hence, an additional analysis step is needed to achieve the classification of the complex, and often interlaced, multicellular structures.

As will become evident in the later chapters of this thesis, cancer cell imaging has both required and inspired a great deal of research in both statistics and computer sciences. We categorize the methods described in the literature to three key steps: 1) image processing, 2) feature extraction and 3) feature analysis. Image processing in this context means the thresholding and segmentation of the image, that is the separation of image foreground from the background and defining the outlines of individual structures. Segmentation is followed by feature extraction where a set of numeric measurements are extracted from the obtained objects. These features can then be analyzed with conventional statistical learning methods, for instance to estimate treatment effects. The methods discussed in this thesis touch all of these three steps but concentrate mostly on the statistical analysis of the features.

Despite the variability in the culturing methods, imaging tools and study designs, what is common to all kinds of high-content cell culturing imaging experiments are the challenges related to both the quality and quantity of the images. We further elaborate on these two issues before moving on to describing the example dataset. The 3D-cell culture experiments produce large volumes of data. The number of individual images, such as those shown in Figure 1.1, is measured in hundreds if not thousands of (stacks of) images per experiment, each containing tens of individual multicellular

structures to analyze (see for instance [26, 27]). Computational efficiency is required of the related methodology to make the analysis of these large scale datasets feasible, especially if high experimental throughput is desired. Furthermore, due to the highly variable nature of the original image data and possible imperfections in the applied processing algorithms, the large set of cell structures is likely to contain artifacts of various kinds. Some examples of such aberrations are shown in Figure 4.2, the contents of which are described in more detail in the related chapter. Similarly to extreme outliers in least-squares regression, these exceptional structures can distort the later analyses if not correctly addressed.

1.3 Description of the example PC3 dataset

To demonstrate the developed methods in practice, we refer to a PC3 prostate cancer cell line dataset containing 3120 grayscale 512x672 -resolution max-projected images, similar to those in Figure 1.1 A. The cells were cultured on multiple 96-well plates, with only the central 60 wells utilized for the experiment to exclude any possible edge effects. A total of four non-overlapping images were taken from each well, resulting in a total of 240 images per plate. Härmä et al. [25] describe the behavior of the PC3 cell line: When left untreated, the PC3 cells initially form well rounded and differentiated organoids, up until day 8 or 9 of cell cultures. After day 8-9, these increasingly large structures spontaneously transform into highly invasive phenotypes with many multicellular appendages rapidly invading the surrounding matrix. Such multicellular or "collective" invasion is typical for prostate cancers and also observed in other cell lines. This active and invasive behavior results in complex and highly variable data, consisting of multiple distinct cell structure phenotypes. These properties make the PC3 cell line a particularly relevant model for studying cancer invasion and heterogeneity.

The data consist of two independent parts, a longitudinal one gathered over multiple days in homogeneous conditions, and a cross-sectional study containing only endpoint results but with multiple treatments. More details of these two parts are listed in Table 1.1. These data allow us to investigate both the dynamic behavior of the PC3 cultures over time and the potency of anti-cancer treatments in possibly controlling the invasion of the cell line. The cross-sectional part of the dataset has been analyzed previously by Härmä et al. [27] and used as an example in Publications I and III. The longitudinal part is so far unpublished. The cultures were imaged using a Zeiss Axiovert 200M spinning disk confocal microscope with a Yokogawa disc, after which the resulting image stacks were compressed to the 2D images using maximum projection. In the cross-sectional part, each drug-dose-combination was replicated in at least three different wells.

Part	Cross-sectional	Longitudinal
Number of plates	6	1
Number of images	1440	1680
Times of imaging	Day 10	Days 1, 4, 5-8 and 11
Tested treatments	28 drugs and a control in four doses	No treatments
Cell structures	$\approx 30,000$	$\approx 10,000$

Table 1.1: Details of the two parts of the example dataset. The accurate number of individual cell structures to analyze depends on the chosen pre-processing steps.

2 Image data pre-processing

Ljosa and Carpenter [47], and Shamir et al. [69] describe the typical workflow associated with the statistical analysis of cell-based images. Instead of directly analyzing the raw pixel-format data, the images are first processed in order to obtain the regions of interest after which the analysis is based on their extracted features. This chapter considers these preliminary steps in the context of 3D cancer cell cultures.

2.1 Segmentation

Segmentation is a crucial step in image analysis concerning distinct objects, such as individual cells or multicellular structures [17, 70]. The goal of this procedure is to determine which pixels belong into which object (and which to the background). Some segmentation methods can be directly applied to the pixel data, however it is common to first aim at separating the image background (empty space) from the foreground (space occupied by the culture). This is commonly called *thresholding*. The found foreground can then be split into the individual regions of interest with a suitable segmentation method. We will now briefly summarize the thresholding and segmentation approaches adopted in the cell-based imaging literature and propose an efficient candidate for the particular type of data considered here.

As described by Chaki et al. [9], thresholding methods are used for finding a suitable cut-point for the pixel values, either locally for neighborhoods of pixels, or globally for the whole image. Local methods, such as the Bradley [6] and Niblack [55] algorithms, look at each single pixel at a time and decide its label based on its local neighborhood. In contrast, global approaches consider the complete pixel intensity distribution to find a suitable common threshold value for the whole image. Examples of this latter approach include the methods by Li [46], Otsu [56] and Tsai [79]. Both local and global methods have been used in cell-based imaging, for example Niblack by Di et al. [16], Li by Hoque et al. [31], and a modification of Tsai's method by Härmä et al. [26, 27]. Thresholding is sometimes applied directly to the raw images, however it is common to first seek to improve and standardize the image quality with various filters. Typical examples of such operations are the median filter to reduce the possible background noise (see for example [16]) and the rolling ball filter to account for imbalanced luminance in the image (see [31]).

After thresholding, the individual objects are defined from the image foreground. This step is trivial if the objects are disjoint, however this is rarely the case with the multicellular structures that are often either touching or even merging to form larger superstructures. To divide such joined regions in the image is a challenging problem that often lacks an objectively optimal solution. A popular method for the task is the watershed algorithm [82] that seeks to split the foreground along the lowest pixel values. This is a sensible approach, if the structures are bright in the middle but darker around their edges. If these characteristics apply in the original image data, the watershed method can be applied directly, thus skipping the thresholding step entirely (see for instance [48]). To apply watershed to a thresholded image however, we first form the *distance map* of the image foreground, that is, the image where each pixel value represents the distance to the nearest background pixel. Here, the watershed separates objects that are connected only by a small number of pixels. This approach has been widely adopted in the literature, recent examples being the works of Åkerfelt et al. [1], Härmä et al. [25, 26, 27], Park et al. [59] and Wang et al. [83]. This is also the course of action we have taken here based on the findings in Publication II, which we will soon describe in more detail.

As a side note, all previously cited approaches consider only the analysis of 2D images, however a truly 3D analysis of the original stack data could still be desirable in occasions where high accuracy and detail is valued over computational efficiency and high-throughput. Such 3D segmentation approaches have been considered by Bilgin et al. [4] and Robinson et al. [64].

To summarize, the images are typically pre-processed with multiple different methods that are intended to tackle specific tasks. These tasks are highly data-specific and thus it is difficult to propose an all-purpose solution – even to the seemingly limited field of 3D cancer cultures – while maintaining a reasonable level of efficiency and simplicity. Consequently, the amount of proposed approaches is vast. Furthermore, it is not uncommon in the cell culture literature to leave the details of the image segmentation unaddressed and simply cite the utilized software, raising some concerns about the replicability of the research results. Examples of frequently used open-source software include the BioimageXD [40], BisQue [45], CellProfiler [37], Icy [14] and ImageJ [66]. A dedicated solution for high-throughput 3D culture analysis is provided by AMIDA [26]. These tools contain the most common image analysis operations by default and most of them can be extended with additional methods if needed.

To segment the example PC3 dataset, we utilize the method suggested in the original publication II, considering the analysis of patient-derived 3D cultures. Here, the image quality was found to pose considerably more challenge than the traditional cell line experiments due to weakened cell staining. As none of the existing approaches yielded acceptable results, the

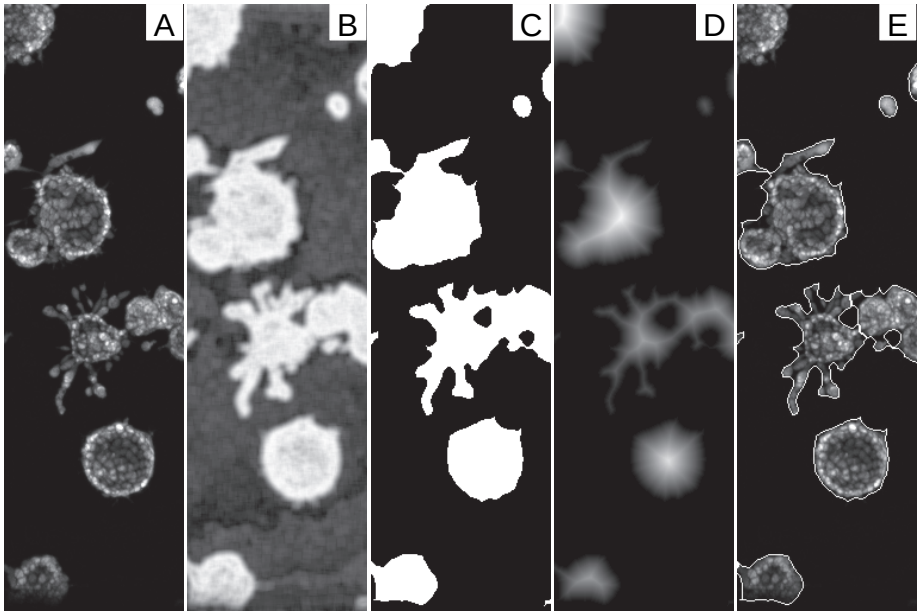

Figure 2.1: ***Segmentation steps for the PC3 dataset.*** ***A:*** *A part of a raw image containing multiple multicellular structures with varying phenotypes* ***B:*** *Local entropy -filtered image* ***C:*** *Thresholded image i.e. the image mask* ***D:*** *Distance map of the image* ***E:*** *Final segmentation result after applying watershed.*

thresholding of these data required novel measures. Inspired by the paper by Robinson et al. [64], local entropy filtering was introduced as a standardizing step into the analysis. The filter was found extremely effective at highlighting the image foreground while standardizing the possible differences in foreground brightness. The filtering was followed by the global Otsu's thresholding [56] to extract the image foreground. In combination, these two steps were found to provide a simple, yet powerful solution to image thresholding. Their superiority was shown by comparing their performance against a set of conventional methods in a manually obtained ground truth dataset. The final segmentation was then obtained with the conventional method of applying the watershed algorithm to the distance map of the foreground. In addition to the PDX data, the method was also found to work well also on the cell line data that was utilized in the paper for validation purposes.

Figure 2.1 illustrates the segmentation of the example PC3 data. Panel A shows a 150x512 pixels detail of one of the images in the dataset. Both round and invasive phenotypes are visible. A few clearly individual structures are present, while the rest are located adjacent to each other and are

possibly in the process of merging. Panel B demonstrates the effect of the local entropy filter on the images. Image foreground is homogeneously illuminated regardless of the brightness differences in the original image, making the actual thresholding task relatively easy. Panel C shows the obtained thresholding result after applying Otsu's method to the local entropy filtered image. In panel D, a distance map has been formed, depicting each pixels distance to the nearest background pixel. Finally, in panel E, the individual structures are segmented by applying the watershed algorithm to the distance map. As can be seen in this final panel, watershed has separated the formation in the middle of the image into two distinct structures, while the structure above it, seemingly consisting of two separate spheroids, has been kept as a single object. This behavior can be influenced by changing the metric in the distance map and by tweaking the parameters of the algorithm. However, the desired end result remains highly subjective and problem-specific.

After applying the described segmentation procedure, a total of 72894 individual objects were identified in the example PC3 dataset. This number is reduced to 42915 after removing both very small objects (less than 100 pixels in size) and objects that lie partly outside the image boundaries (over 25% of the perimeter on the image border). The removed small objects are likely to be individual cells, cell fragments (debris) or simply segmentation artifacts that are not of interest here. Similarly, it is difficult to draw reliable conclusions from structures that are not fully captured in the images.

2.2 Feature extraction

As described by Ljosa and Carpenter [47], and Shamir et al. [69], the analysis of the individual objects of interest is based on set of numeric features, measured from the raw pixel data. The goal of this feature extraction is to represent the otherwise intractable data in a coherent and standardized format without a loss of information. In other words, a fixed set of numeric measurements are made for each object and collected into a data matrix, which can then be analyzed with known statistical learning methods. For example, one could measure the width and height of a multicellular structure to get an idea of its general shape and size. If width and height were the only biologically interesting characteristics of these structures, the resulting feature dataset would then describe the original images without a loss of (biologically relevant) information. In this section, we briefly outline the types of features that are commonly used in the analysis of cell-based images and describe the suggested feature set for the 3D cancer cell culture data considered here.

The previously mentioned height and width are examples of features

that are readily interpretable and meaningful as such. By studying their distributions in the data, one could gain understanding of the general size and shape of the observed structures and estimate the effects different treatments might have on these characteristics. For the purposes of this chapter, we will refer to these types of measures as *descriptive* features. On the other hand, one could consider features that do not allow an immediate interpretation but would still be informative to a learning algorithm, such as a random forest [7, 28] or a support vector machine [28, 81]. As an example, one could calculate some higher order central moments of the pixel intensities to represent the general texture of the structure. Here, we refer to these to types of measurements as *nondescriptive* features.

A vast collection of features has been developed for 2D images, however there are some guidelines that limit the number of sensible choices for characterizing the multicellular cell structures in our example dataset. Firstly, the features should be rotation-invariant, since the alignment of the objects is not informative. This represents a clear difference to many common image analysis problems, such as the recognition of human facial expressions or written characters, where different alignments can relate to different moods or can make the difference between the letters "L" and "V". Secondly, the features should not depend on the general luminosity of the objects as this can vary based on their position in the well and on other non-biologically relevant factors. Furthermore, the same biological characteristics can be shared by a multitude of seemingly differently shaped structures. For instance, two structures, one with three appendages and the other with four, would be considered very similar biologically and should thus be represented with matching feature values. Again, this is a clear difference to optical character recognition, where the letters "X" and "T" should be distinguishable in their features.

For the specific case of analyzing multicellular structures with invasive properties, researchers have considered mainly descriptive features. Härmä et al. [26, 27] and Åkerfelt et al. [1] utilized the feature set of the imaging software AMIDA that contains measurements of for instance the area, roundness, density and the appendages of the spheroids. These features relate to high level characteristics, such as invasion and differentiation. To investigate more detailed lower level qualities, Park et al. [59] used measurements related to the cell nuclei, structure shape and the staining intensity of the structures. In the broader scope of cell-based imaging and high content screening, the variety of research problems and the number of proposed approaches to solve them is naturally greater. However, when it comes to nondescriptive features the publications in the field (for instance [5, 23, 53]) tend to use features of very general nature, instead of focusing on more problem-specific attributes. As implied by Shamir et al. [69] in their review paper on pattern recognition in biological imaging, this is not

necessarily a critical issue, on the condition that the features are intended to be used as inputs for supervised learning algorithms. In fact, the authors state that using a broad set of features can support the generalizability of such a method. Examples of widely used general image features include the Zernike moments [76] and the Haralick texture features [24], both utilized for instance by Boland and Murphy [5], Loo et al. [48] and Wang et al. [83]. While such features have proven powerful in supervised learning, the possible excess, biologically irrelevant information they contain can have adverse effects in unsupervised learning, such as clustering [71]. This was an important consideration when designing the approach proposed in Publication I, which we will next look into in more detail.

The motivation for publication I was to develop a data-driven method to characterize the morphologies of the multicellular structures in a given study. As these morphologies of interest are not always clearly identified and can even change from study to study depending on the questions at hand, an unsupervised approach was preferred that could be easily adapted to various settings. To facilitate this goal, extra care was taken to choose and develop nondescriptive features that would be meaningful in this context. Three general characteristics were named as important to determine the phenotype of a cell structure: its size, shape and texture. Size is easily measured by calculating the number of pixels in the objects while the other two aspects require more attention.

As described by Härmä et al. [25], the PC3 cultures typically follow a sequence of morphological changes that start from round, consistently shaped structures that develop to "stellate" or star-shaped objects and finally transform to completely irregular forms. As these changes are a sign of invasive behavior that an efficient treatment should suppress, it is highly important that they are quantitatively captured in the extracted features. Considering that the structures are initially approximately round, the deviation of the structure's outline from an ellipse is a natural measurement of invasion. A more detailed approach would however be required to detect the development of appendages and general roughness in the outline. As a novel contribution in Publication I, principal curve -based features were introduced. A closed principal curve[2] is fitted on the outline, giving a smooth approximation of the shape. If a large difference between the observed and the fitted shape is not measured, the structure is likely to have a smooth outline. In contrast, rough borders and appendages result in clear deviations between the shapes, which are then captured in the extracted features.

As in shape, the cell structures also display changes in texture as they go through their transition towards the invasive phenotypes. Well differentiated and therefore benign PC3 cell structures initially display a hollow interior that then increasingly becomes filled up by cells when the culture becomes invasive. In the 2D-projections, these stages of change would look

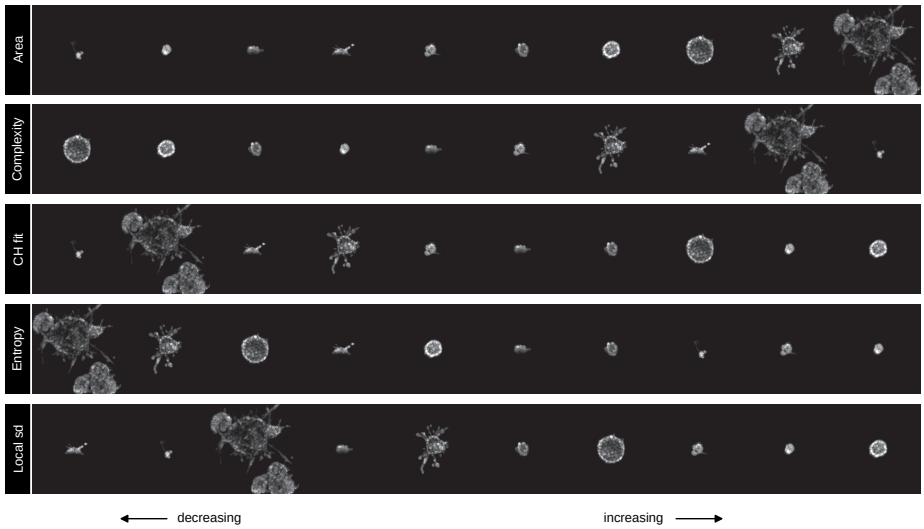

Figure 2.2: ***Cell structures sorted based on descriptive features. Area:*** *Number of pixels in the structure* ***Complexity:*** *Measure of shape's irregularity* ***CH fit:*** *Complex hull fit statistic* ***SD:*** *Standard deviation of the pixel values.*

like initially hollow rings that start filling up with material. In the final stages of this transition, individual cells would become visible in the surface. A set of texture features were proposed, that were based on the local distribution of pixels in the image, namely their first four standardized moments: mean, variance, skewness and kurtosis. These measurements were calculated both in very small pixel neighborhoods and larger regions, to capture texture both in the level of individual cells and in the overall structure.

In addition to the features considered in Publication I, we also extract a set of additional descriptive features from the structures in the PC3 example dataset for demonstration purposes. To aid the interpretation of the these features, we refer to Figure 2.2, containing a set of cell structures ordered by their measured feature values. *Complexity* was shown in the co-authored publication of Härmä et al. [27] to be a powerful measure of shape's general regularity. The feature is defined as the residual of the linear model $\log(O) = \beta_0 + \beta_1 \log(A) + e$, where O is the length of the outline, A is the area of the structure, e is the residual and the parameters β_0 and β_1 are estimated from the data. This feature is closely related to the *compactness* and *roundness* measures often used in the literature (see for instance: [59], [26]). *Convex hull fit* (CH fit) measures convexity of the shape and ranges between 0 and 1. In practice, this measure correlates strongly with the shape's *complexity*, however the two features disagree when considering objects with pronounced appendages but otherwise regular shapes, or vice

versa, structures with no clear appendages but irregular shapes. *Entropy* is the standardized entropy of the pixel intensities and provides a brightness- and size-invariant measure of the shape's texture. Low entropy relates to structures with very few distinct pixel values, while entropies close to 1 are measured from structures with very heterogeneous textures. Finally, *local sd* measures the average of the local variances in the texture. High local variances are observed when the object's surface is very rough, while low values indicate relatively smooth texture patterns. Finally, structure's *area* is simply measured by the number of pixels. These five features alone provide a powerful basis for many analyses, while retaining easy interpretation. For instance, cell structures exposed to an invasion-inhibiting drug would be expected to demonstrate decreased complexity and increased convex hull fit when compared to a control treatment.

3 Estimating treatment effects

The goal of the cell culture experiments is often to assess the effects that various treatments or other interventions have on the tumor cells. This is achieved by analyzing the extracted features. A direct comparison is possible if the features are descriptive in nature, while others could be used as inputs in further analysis such as clustering, classification or dimension reduction. In this section, we discuss the various approaches available for treatment effect analysis. A novel flexible finite mixture regression model is introduced in this context.

3.1 Comparisons of treatments based on location statistics

Given a set of descriptive features, such as those considered in Figure 2.2, one can easily study the marginal effects of treatments of interest against the control. By "marginal" we mean the comparisons of single location estimates, such as the means. In Figure 3.1 such a comparison has been made in the cross-sectional part of the example dataset based on the medians of the descriptive features. The motivation for collecting these data was to test the efficacy of 25 novel betulin-derived compounds extracted from the bark of Finnish birch trees for blocking the invasion of prostate cancer cell cultures. Three already established drugs, Abiraterone, Enzalutamide and Paclitaxel were included as references. The heatmap in Figure 3.1 clearly highlights the Paclitaxel drug and a number of novel compounds, namely the drugs 3-6, 15, 16, 19 and 20, as promising treatment candidates. A clear dose effect is also seen as the difference to the control grows more pronounced in higher concentrations of the compounds.

These type of comparisons of location statistics are frequently used in the literature. For instance, Bilgin et al. [4] measured the changes in mean elongation and edge length of their segmented cell colonies over time, while Åkerfelt et al. [1] considered the average growth and number of extensions in the cancer-associated fibroblasts. While this practice is common, the applied statistical methods often remain rudimentary in nature and limit to multiple tests for location and to estimating standard errors for the mean estimates. However, in addition to the treatment effect, there are multiple other plausible sources of variation, that could be addressed with more sophisticated statistical models. For instance, it is reasonable to assume that

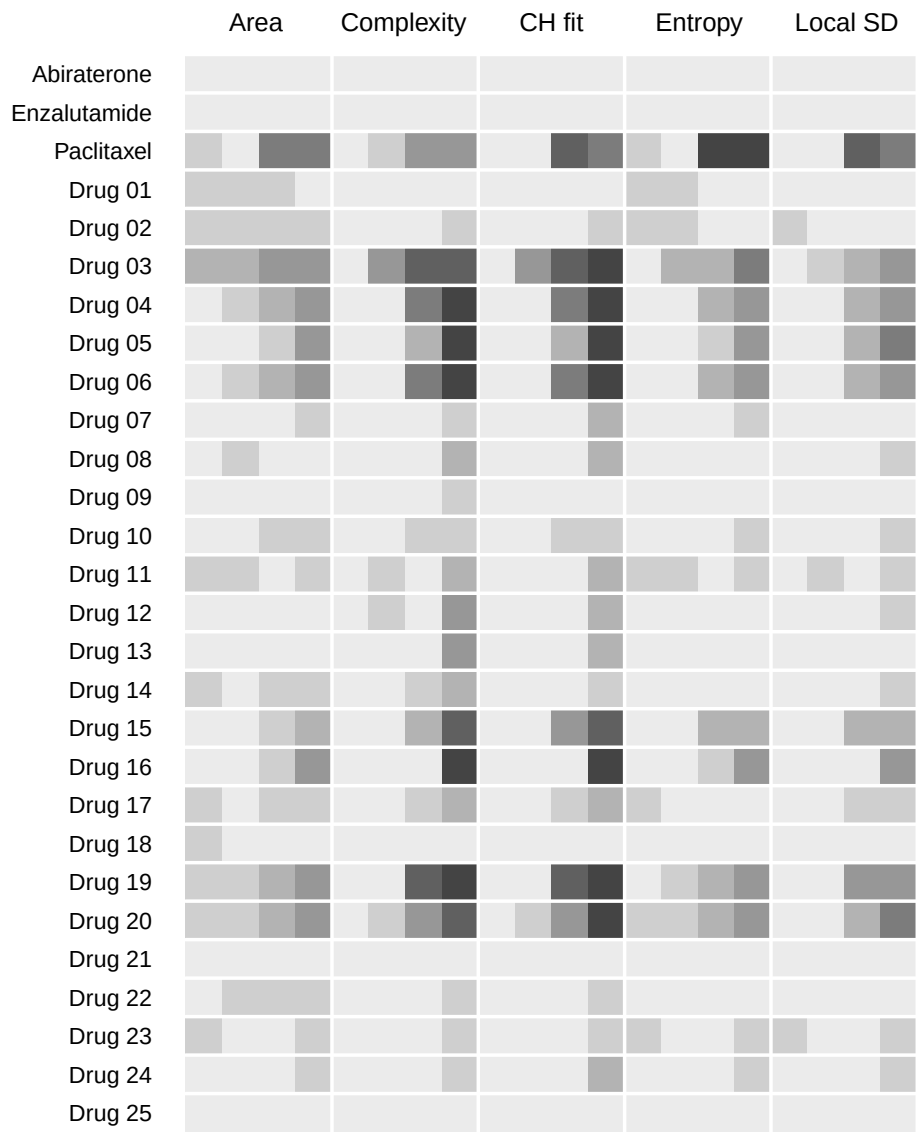

Figure 3.1: ***Estimation of treatment effects based on locations statistics in the PC3 dataset.*** *Median value of each descriptive feature was calculated for each drug and compared against the median of the control treatment. Darker colors indicate improved results compared to the control. Each drug was administered in four concentrations that are arranged here side by side in increasing order.*

the cell structures cultured on the same plate or in the same well could be correlated in their behavior. To this end, a mixed effects model was considered in the co-authored publication by Härmä et al. [27], that included

random effects to address both the plate-to-plate and well-to-well variation in the response. The well number was also included as a continuous covariate in order to model the possible systematic trend that could be induced by the laboratory practices; cell culture experiments include many steps, such as culturing, feeding and imaging, where the wells are processed row by row starting from the top left corner and ending to the opposite corner of the plate. In the end, these additional variables in the model were not found very important, and were mostly overshadowed by the strong treatment effects. This is naturally a desirable outcome since it bolsters our confidence in the biological model and especially in the reproducibility of the results. However, the additional sources of variation should not by all means be neglected in future studies, if nothing more than for the sake of quality control.

As discussed in the first chapter, the cell culture datasets may contain extreme outliers that can adversely affect the analyses unless properly addressed. In the case of the mixed effects model utilized by Härmä et al. [27], such outliers were however not observed. The use of robust methods is nonetheless advocated throughout the analyses.

3.2 The challenge of within-culture-heterogeneity

Marginal analyses of descriptive measures provide a simple and efficient tool for treatment effect assessments, that can, potentially, be extended to address multiple sources of variation. However, regardless of the complexity of the chosen statistical model, a fundamental assumption on the nature of the outcome variable is made, that is, it follows a unimodal distribution around some average value. In the context of cell structure morphology, this assumption proclaims that the phenotypic properties of the structures vary randomly around some specific average phenotype. A biologically sufficient analysis then only needs to look at the changes in this average induced by the treatments and other factors. For many purposes, this assumption is valid and does not oversimplify the biology. However, in the case of cancer cell structures, it should be met with healthy suspicion.

Cancer heterogeneity [3, 19] is currently a highly important topic in the overall field of cancer research because it may explain why drugs that proved effective in the first round of treatment lose their efficacy at recurrence. It has now been acknowledged, that tumors do not consist of a homogeneous population of cells but are instead formed by a mixture of populations, possibly differing in the genetic level. At the risk of oversimplifying, a seemingly efficient treatment could in reality be targeting only a high percentage of the heterogeneous population of cells instead of completely eradicating the tumor. The relapsed tumor then consists of only the surviving cells that are not affected by the previous treatment. Remarkably, this intra-tumor hetero-

geneity is not limited to clinical samples but is observed even in the highly standardized cell line cultures. Indeed, the structures formed by the PC3 cells in Figure 1.1A demonstrate highly variable phenotypes despite having the same clonal origin.

Due to the cancer heterogeneity, an analysis based on the location statistics of the descriptive features might not give an accurate representation of the culture. In an extreme case, a drug reducing the size of a half of the cell structures while leaving the other half grow unaffected, might result in a zero average effect in structure size. More sophisticated approaches would be needed for the analysis of these possibly multi-modal data.

3.3 Flexible finite mixture of regressions

Finite mixture regression (FMR) provides a framework for dealing with data that are suspected to originate from a heterogeneous population, or more precisely, from K latent classes [42, 52]. The simple linear FMR fits a separate linear regression between the outcome y and the covariate vector $\boldsymbol{x} = (x_1, \ldots, x_p)'$ in each latent class, typically with Gaussian errors. In this case, the full density of the outcome is a mixture of Gaussians:

$$f(y|\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \phi(y; \boldsymbol{x}'\boldsymbol{\beta}_k, \sigma_k), \quad \sum_{k=1}^{K} \pi_k = 1, \tag{3.1}$$

where $\phi(y; \mu, \sigma)$ is the density of the normal distribution with mean μ and variance σ^2, $\boldsymbol{\beta}_k$ is the vector of regression coefficients for class k, σ_k is the standard deviation of the error term in class k and $\pi_k > 0$ is the probability of membership in class k. The model parameters are then estimated with the expectation-maximization algorithm [15, 28].

FMR models could prove a viable way to address the within-tumor heterogeneity by naturally accommodating the latent cell subclasses into the analysis. While many problems could be tackled with the linear FMR, more complicated applications including nonlinear relationships and interactions between variables could create difficulties. In these cases, it is not straightforward to identify the right structure for the linear predictor $\boldsymbol{x}'\boldsymbol{\beta}_k, k = 1, \ldots, K$; more flexible models would be preferable. For example, consider the dynamic changes in the *convex hull fit* feature in the longitudinal part of the example dataset displayed in Figure 3.2. The cell structures in this part of the data have not been subjected to treatments and thus undergo their characteristic unaltered growth processes. As seen in the figure, the distribution of the feature changes over time, but not only in location, but also in dispersion. Furthermore, the distribution does not resemble a symmetric Gaussian by being heavily skewed and even hinting at bimodality. An FMR

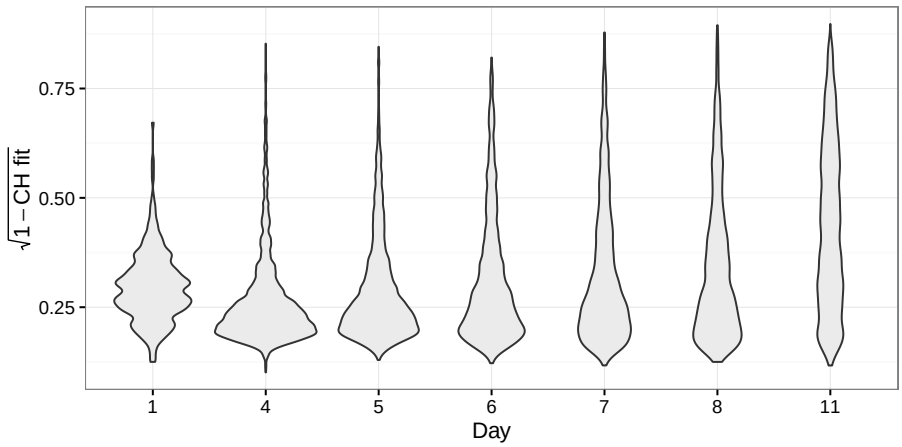

Figure 3.2: ***Change in the distribution of convex hull fit over time in the PC3 dataset*** *The plots show a kernel estimate of the density of the feature. The feature has been transformed to decrease its skewness. Lower values of the transformed feature correspond to increased convexity while larger values are associated with deviation from convexity.*

model could provide a good fit for these data, however, to assume a linear transformation over time might be an error.

In Publication IV, we proposed a flexible alternative for obtaining predictions with a finite mixture of regressions in settings with large number of variables. Instead of assuming a rigid structure for the FMR, such as in equation (3.1), the idea is to find a flexible estimate for the complete mixture $f(y|\boldsymbol{x})$. The goal is therefore not to identify the individual latent component means but the full resulting density. This was achieved by combining two existing ideas: the random forest [7] and a penalized FMR regression model [72]. Similar more flexible approaches to FMR have been proposed by Huang et al. [33] and Xiang [84] however with limited applicability to datasets with multiple covariates due to the curse of dimensionality [28].

Random forest is a powerful learning method that is obtained by averaging an ensemble of binary decision trees, each trained on a bootstrap sample of the data with randomly selected set of covariates at each branch of the tree. The method makes minimal assumptions on the shape of the data and is able to capture arbitrarily complex interactions and nonlinear dependencies between variables. Importantly, each tree in the forest consists of a sequence of binary cuts that can be expressed in the form of dummy variables in the original data. Joly et al. [36] used this representation to compress the forest by supplying the dummy variables as covariates to a LASSO-penalized regression model, with the goal of retaining only the most relevant branches

in the forest. Following this idea, the procedure for flexible FMR is obtained:

1. Fit a random forest on the data and extract the dummy variables corresponding to the leaves (final branches) of the trees.

2. Supply the original covariates and the extracted dummy variables as covariates to a penalized FMR model

The proposed flexible FMR is thus essentially a linear FMR model with the added dummy variables to account for the possible deviations from linearity. The dummy variables are extracted only from the leaves of the trees in order to reduce the computational expense of the model.

While several penalized FMR methods exist [21, 42, 43], the method proposed by Städler et al. [72] was utilized here due to its ability to efficiently handle a large number covariates and its accessible R-implementation. The Bayesian information criterion (BIC)[68] is used for selecting optimal tuning parameters for the random forest and the amount of penalization. The BIC is calculated as $-2l(\hat{\boldsymbol{\theta}})+\log(n)d_e$, where $l(\hat{\boldsymbol{\theta}})$ is the observed maximum log-likelihood, n is the number of observations and d_e is the effective number of unknown parameters [58]. Here, $d_e = K(p+d+1+1)-1-q$, where the summands in $p+d+1+1$ correspond to the dimension of the $\boldsymbol{\beta}_k$ vector (p), the number of dummy variables (d), σ_k (1) and π_k (1). The -1 is added due to the sum constraint $\pi_1,\ldots,\pi_K = 1$ and q is the number of parameters set to zero due to penalization.

In Publication IV, the method was shown to outperform the linear reference models in simulation studies that included nonlinear relationships and interactions, and was found to be able to handle a large number of covariates. Furthermore, the method was not greatly influenced by surplus variables that were not related to the outcome.

To demonstrate the usage of the method, we apply the flexible FMR to the data in Figure 3.2. Four models are fitted with varying number of latent classes: $K = 1,2,3,4$ using the time point t as the single (continuous) covariate. The model with $K = 3$ produced the best BIC, providing evidence for the existence of multiple latent classes. Only one dummy variable, $I(t > 8)$, was retained in the final model while the rest were penalized away in the estimation phase. This seems to indicate that there is a relatively large change in the distribution at the end of the study. Figure 3.3 shows the mixture densities at different time points, as defined by the estimated model. According to the model, the changes in the feature are not fully understood as a simple homogeneous location shift, but instead as multiple separate trends. While perhaps most of the cell structures remain unchanged, a part of them continue progressing towards higher values of the transformed feature.

The main drawbacks of the flexible FMR compared to its linear competitors are the losses in computational efficiency, interpretability and in appli-

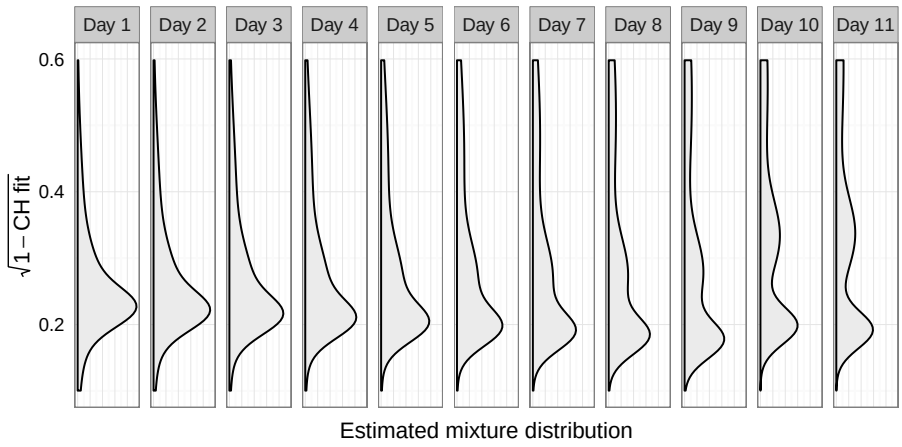

Figure 3.3: ***Mixture distributions as estimated by the flexible FMR.***

cability to model-based clustering. If a large number of dummy variables is retained in the final model, it becomes very difficult to understand the roles of individual covariates in the regression. In a linear FMR, this would be easily achieved by examining the estimated model parameters. Similarly, linear FMR can also be used for clustering purposes, by looking at the latent class memberships estimated by the model. In the flexible FMR however, the introduction of the dummy variables allows for almost arbitrary label switching between the observations, hence making the final clustering results meaningless. Because of these two drawbacks, the utility of the method is reduced to obtaining accurate predictions in the cases where a lot of dummy variables are retained. As shown in the original publication however, we are relieved of these issues if the true underlying model is in fact linear. In this case the flexible FMR model is observed to reduce to a linear FMR on the original covariates due to penalization. The only suffered loss is then the increased computational burden.

3.4 Morphological clustering of the multicellular structures

The previously discussed finite mixture regression approaches provide an elegant framework for modeling individual features while accounting for the heterogeneous population. However, trying to understand the full complexity of the cell cultures through a set of models, each addressing a single descriptive feature, can be a tedious task. Furthermore, it can be difficult to construct accurate and informative descriptive features for the more

complex and perhaps abstract characteristics of the cell structures, such as their invasiveness. A more straightforward approach directly accessing the phenotypes themselves would be preferable. One possible route to such an analysis is provided by clustering algorithms that aim at identifying groups of observations that are close each other in the variable space. By using these methods, we would be able to allocate the cell structures into groups of similar features. In the ideal case, these groups would have direct biological interpretations. The subsequent treatment effect estimation could then be based on the proportions of these groups in the data, instead of the indirect analysis of the descriptive features. Furthermore, this approach would naturally incorporate the observed heterogeneity of the cell structures.

A clustering approach to the cell culture analysis was proposed in Publication I, where first a large set of both descriptive and nondescriptive features were extracted, their dimension was reduced with the principal component analysis (PCA) [35] and finally, the obtained principal components were used as inputs for a hierarchical clustering algorithm. Here, PCA served mainly as a dimension reduction tool that enabled the clustering of the otherwise too high-dimensional data. This well known method projects the original data to a lower dimensional space while minimizing the loss of variance. The approach can therefore be expected to provide sensible results if high variance can be equated with high biological information. A related benefit is that PCA automatically groups together highly correlated variables thus avoiding unnecessary redundancy. In practice, the directions of the components are obtained as the eigenvectors of the covariance matrix of the data, and their associated variances as the eigenvalues of the matrix. Similar ideas have been proposed for various purposes in the literature, the work by Park et al. [59] on clustering breast cancer cell structures being the nearest example. Here, the authors measured mainly descriptive features such as the number of nucleai, compactness and different integrin levels, and used these as inputs for a farthest-neighbour hierarchical clustering. Some dependency between proportions of the obtained phenotypic clusters and the assigned treatments were observed. Other closely related applications are the phenotypic grouping of cell lines presented by Han et al. [23] and the grouping of treatments based on feature principal components by Di et al. [16].

Next, we demonstrate the method proposed in Publication I with the PC3 example dataset. We do this by utilizing the current segmentation approach and feature set, that were not available at the time this Publication I was prepared. The main steps of the analysis are kept essentially the same, however further effort is made to carry the idea of robustness throughout the procedure. When applying unsupervised methods, such as clustering or dimension reduction algorithms, one must pay careful attention to the format and content of the data. Especially, many unsupervised methods are not

independent of the scales of variables and tend to place a large weight on inputs with high variances. Clustering algorithms based on euclidean distances provide prime examples of such behavior. Partly for the same reason, outlying observations can have distorting effects on the results if not properly addressed [22]. Furthermore, surplus variables that are either highly correlated with other measurements or simply unimportant for the question at hand, adversely effect the analysis due to the *curse of dimensionality* [28]. With these considerations in mind, we first apply a set of normalizing operations and aim to reduce the dimension of the PC3 feature data before moving into clustering. The obtained features are first linearly transformed to the approximately same location and scale in order to achieve equal weighting in the subsequent analyses. Specifically, the following transformation is used:

$$\frac{\boldsymbol{x} - \bar{\boldsymbol{x}}_{5\%}}{1.48 \times MAD(\boldsymbol{x})},$$

where $\boldsymbol{x}$ is the vector of observed values of a given feature, $\bar{x}_{5\%}$ is its 5% trimmed mean (90% of data is used) and $MAD(x)$ is its median absolute deviation. These latter two serve as robust estimators of location and scatter.

Next, the dimension of the data is reduced in order to group highly correlated features together. Again a robust alternative is used, specifically the robust variant of PCA proposed by Hubert et al. [34], which is suited for high-dimensional data. The method consists of three main steps that are outlined here in a rough fashion:

1. The data is projected to a lower subspace using singular value decomposition.

2. An outlyingness score is calculate for each observation, after which the innermost 90% of observations are used to estimate the covariance matrix. The data is then projected to the first principal components of this matrix.

3. The location and scatter of the projected data is calculated using the robust MCD estimator [65]. This method aims at selecting that subset of the data that results in a covariance matrix with the smallest determinant. Finally, the robust PCA solution is obtained by finding the eigenvectors of this covariance matrix.

We extract principal components separately from the shape- and texture-related features in order to analyze these two aspects independently. Components associated with eigenvalues greater than 1 are retained, leading to a total of 11 shape-related components and 9 texture-related components. This corresponds to choosing those components that are associated with

more variance than a single original variable (after standardization). In total, the extracted components explain 93.7% and 94.0% of the total variance in the original shape and texture features correspondingly.

Figures 3.4 and 3.5 show a sample of cell structures projected onto the first two principal components of both the shape and texture features. In both cases, a clear gradient is observed, ranging from highly irregular to close to circular forms in shape, and from smooth to heterogeneous in texture. To a biologist, these archetypes are highly meaningful and can be directly related to the cell structure's phenotype. If so desired, one could even use these figures to manually label the individual structures under some predefined categories and carry on with the analysis on that basis. In the general case however, using more than the two first principal components might be necessary for achieving satisfactory results, making such manual assignments difficult to implement. Instead, we utilize a clustering method to automatically divide the data into groups of similar observations using the full set of extracted principal components. The content of these obtained clusters can also be visually inspected with relatively small effort by plotting samples of cell structures from each group. The clusters can then be assigned under some pre-specified labels that are relevant for the goals of the study. Here, we define three such labels for the shapes: "regular", "rough" and "irregular", and two labels for the textures: "smooth" and "variable".

In Publication I, the grouping was obtained using a hierarchical clustering algorithm, which has the visually attractive "tree"-representation. This structural form provides a convenient framework for both observing the cluster relationships and further refining the clustering solution (merging and splitting). Here however, we settle for a simple alternative that allows us to further sustain the theme of robustness, namely the trimmed k-means algorithm [13]. As implied by its name, the method seeks those cluster centers and $1-\alpha$ proportion of observations that minimize the within-cluster-variance. The α observations can then contain arbitrarily extreme outliers without affecting the clustering result. As before, we choose $1-\alpha = 0.9$.

As for the number of clusters to extract, there are no clear guidelines we could follow. It is unlikely that the cell structures would form clearly separated clusters in the variable space, as this would imply that the data consist of clearly distinguishable phenotypes. Instead, a smooth spectrum of structures between different phenotypic extremes is both expected and observed (for instance in Figures 3.4 and 3.5). Indeed, the tested methods for estimating the number of clusters, *prediction strength* by Tibshirani and Walther [77] and the *gap statistic* by Tibshirani et al. [78], both suggest one cluster. In this case, we use the clustering algorithm simply for dissecting the data into more comprehensible parts. A relatively large number of 10 clusters for both the shape and texture components is extracted. These clusters are then manually assigned under the previously specified labels on visual

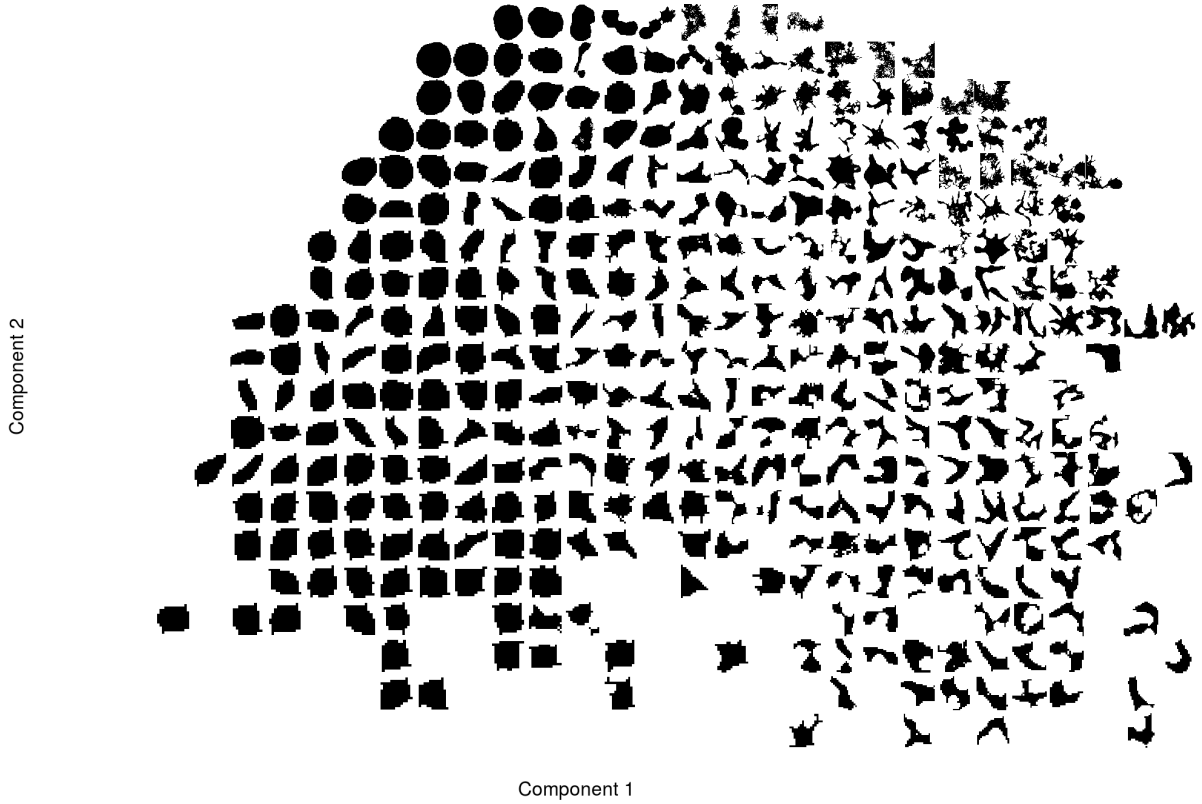

Figure 3.4: ***A sample of cell structure masks projected onto the first two principal components of the shape features.***

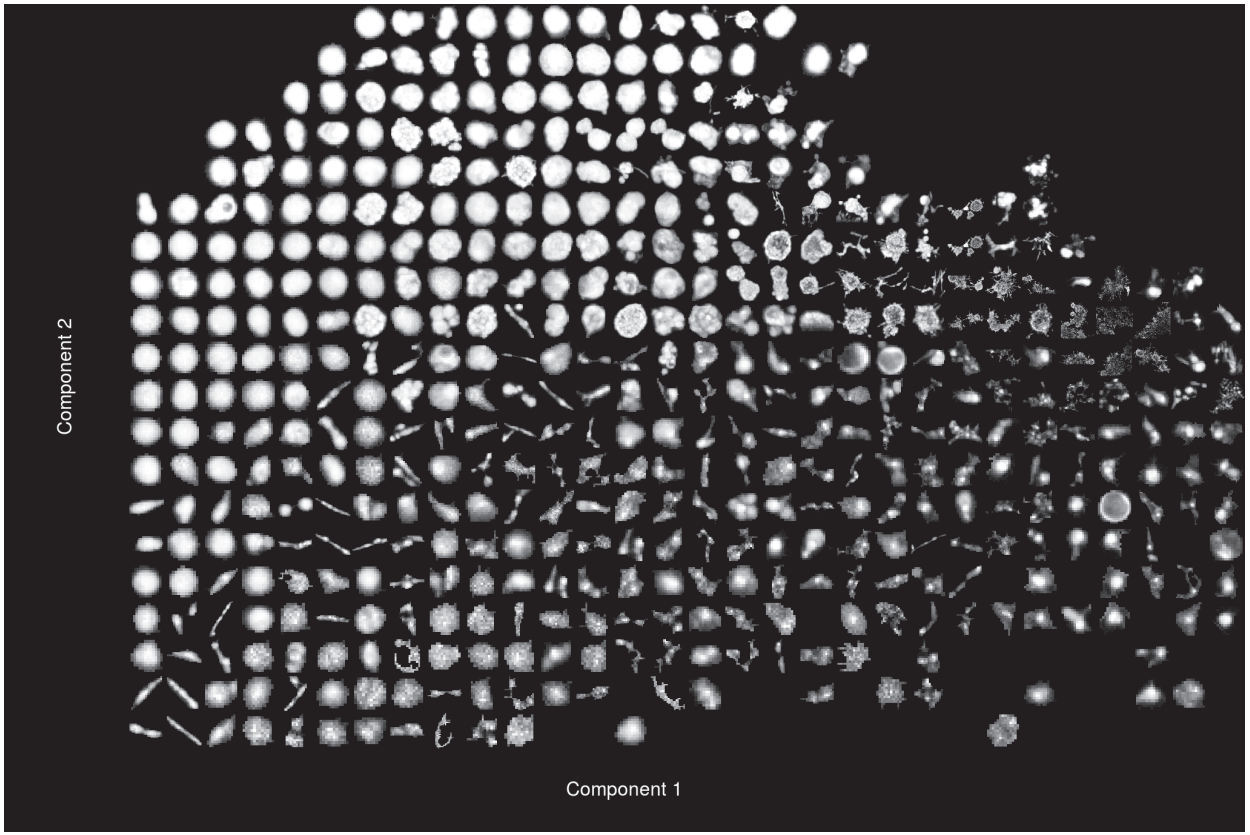

Figure 3.5: ***A sample of cell structures projected onto the first two principal components of the texture features.***

inspection. Size of the structures, measured by the number of pixels, is also split into three categories via the trimmed k-means: "small", "medium" and "large".

The trimmed k-means algorithm is computationally relatively demanding, which is why the clustering is obtained for a reduced set of 1000 randomly sampled cell structures. The remaining data are then allocated to the cluster with the nearest center.

Figure 3.2 displays the observed changes in the longitudinal part of the example dataset in terms of the obtained phenotypic groups. The first images were obtained at the first day of the experiment when most of the observed structures are still very small and few in number. The total of observed structures reaches its maximum at day 4, after which it gradually declines towards the end of the experiment as individual structures merge to form larger objects. In the beginning, most structures are small in size, regularly shaped and have a smooth surface texture. Over time, the structures grow in size and form irregular shapes, as expected due to the invasive tendency of the PC3 cultures. Their surface and inner structure also become more heterogeneous leading to more variable textures. The goal of a successful treatment would be to stop these changes and, in some respect, to stabilize the culture.

Similar to the marginal treatment effects considered in Figure 3.1, we can examine the effects the treatments have on the cluster proportions. These effects are shown in Figure 3.7 in terms of absolute differences compared to the proportions observed with the control treatment. The same set of compounds stand out in this comparison as in the marginal analysis, namely Paclitaxel and the novel compounds 3-6, 15, 16, 19 and 20. On top of these, compounds 7, 11-14 and 24 result in statistically significant effects, albeit of lower magnitude. Out of the three characteristics, only size and shape seem greatly affected by the treatments, while texture remains mostly unchanged. Overall, it can be seen that the effective treatments tend to limit the growth of the cells leading to an increased proportion of small structures. Similarly, the proportion of irregular shapes is reduced. In contrast to the marginal comparisons, some more detailed results are also observable. In particular, the highly toxic Paclitaxel treatment stands out from the other effective compounds when examining its effects on the structure shapes and textures. While the other drugs result in a great increase of regularly-shaped structures, Paclitaxel is also connected to the increase of roughly-shaped structures. Furthermore, the drug seems to have a significant increasing effect on the number of smooth textures. These unique results are explained by the high toxicity of the Paclitaxel treatment: instead of just inhibiting the growth of the cells, the drug is known to directly damage the cultures, leading to broken structures formed by dying cells. These apoptotic cells typically appear as very bright objects in the images resulting

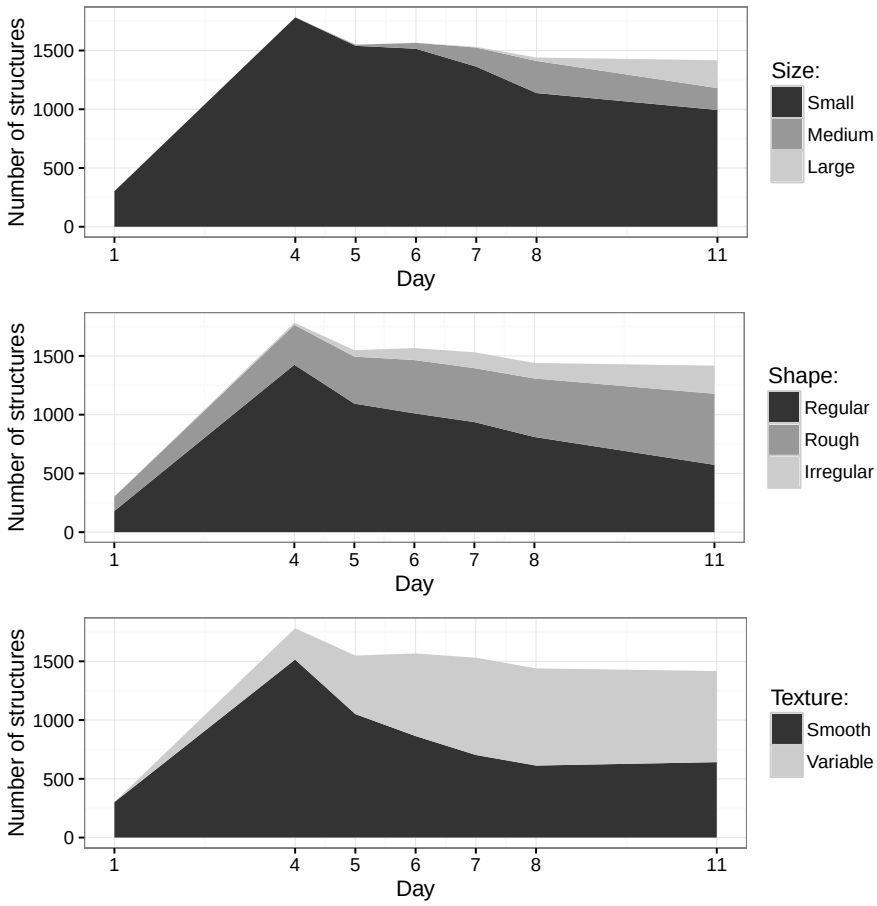

Figure 3.6: ***Temporal changes in cluster sizes in the untreated PC3 culture.***

in a smooth observed texture.

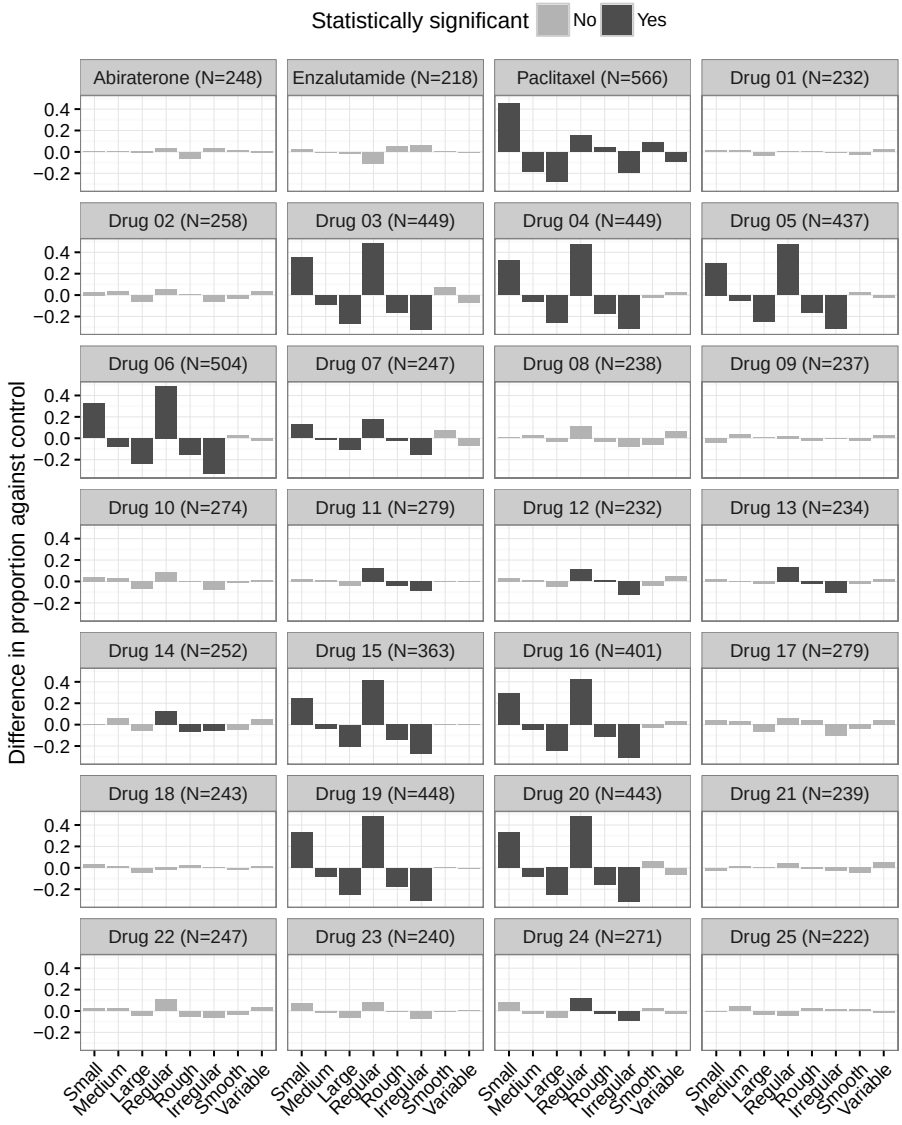

Figure 3.7: ***Treatment effects on the cluster proportions****. The proportion of structures in each cluster was measured and compared against the control treatment. Only the highest dose of each drug was considered. Pearson's chi-squared tests were performed to assess the statistical significance of the differences. Comparisons associated with p-values lower than the Bonferroni-corrected level of* $0.000595\ (= 0.05/(28\times 3))$ *were considered significant and shown here with a darker shade. Each characteristic – size, shape and texture – was analyzed independently.*

4 Detection of novelty among new data

Identification of exceptional observations is a recurring concept in cell-based imaging, where outliers can exist in multiple levels of analysis. For instance, cell structures with exceptional features are typically artifacts resulting from either imaging or segmentation errors. Even before addressing single objects, one might be interested in unusual cases among the original raw image data for quality control purposes; substandard data could be removed prior to segmentation. To go even further, the goal of drug screens is often to find the few outstanding treatments among a large set of candidates. Clearly, there is a strong demand for methods where the goal is to identify the exceptional. In this chapter, we give a brief overview on the existing methods in this field and introduce the novel problem of global level novelty detection.

4.1 Existing approaches

Following the convention adopted in the survey paper by Hodge and Austin [30], the problem of detecting exceptional observations is divided into three different categories depending on the amount of available information. The classical outlier (or anomaly) detection aims at identifying abnormal data without any prior information. The definition of the outliers vary but they might for instance be observations far away from the most of the data or otherwise situated in an exceptional part of the variable space. In contrast, novelty detection aims at detecting unusual observations among new data given a training dataset consisting of only "normal" observations. The goal is therefore to first learn the characteristics of the data that are known to be normal and then use these learned patterns to locate abnormal observations in the future. Consequently, novelty detection is also known as one-class classification or semi-supervised learning. Finally, if the classes of both normal and abnormal observations are known *a priori*, the problem reduces to a supervised classification problem. Further reviews of both anomaly and novelty detection methods are given by Chandola et al. [10], Markou and Singh [50, 51] and Pimentel et al. [61]. Here, we concentrate on the problem of novelty detection and introduce a new type of problem called global level novelty detection.

As described by Pimentel et al. [61], the vast amount of methods developed for novelty detection can be categorized further to probabilistic, distance-based, reconstruction-based, domain-based and information theoretic approaches. The probabilistic approaches aim at estimating the underlying probability distribution of the normal data, against which the novelty of the new data can be assessed. Similarly, distance-based methods base the novelty scoring on a dissimilarity measure, the simplest example being the euclidean distance to nearest normal observation. Reconstruction-based approaches aim at learning the dependencies between the variables in the data. The predictions of the learned model are then compared against the observed values in the new data two determine which observations produce unpredicted results. As an example, the self-organizing map (SOM) [44] can be used in this fashion. Domain-based methods aim at finding a boundary around the normal points that can be used to exclude the novel observations. One-class support vector machine[67] is a prime example of such an approach. Finally, information-based approaches typically construct some global measure of entropy for the whole data. The subset of observations that maximizes the change in the measure are then deemed novel.

To summarize, the vast literature surrounding novelty detection deals exclusively with identifying individual or groups of outlying observations. Each observation in the dataset is typically assigned some *novelty value* that measures its deviation from the learned normal pattern. The novel observations are then identified by selecting some sensible threshold for this value. The methods are associated with applications where the anomalies are relatively rare or otherwise expensive to collect. Examples listed by Pimentel et al. [61] include the detection of computer system intrusions [38], medical diagnostic problems [11, 63, 75] and identifying structural damage in buildings [74]. In Publication III however, we considered a different problem, where the aim is not to focus on individual observations but rather on the existence of novelty in general. In other words, instead of asking the question, “Which observations are novel?”, we ask, “Is there novelty at all?”. The latter question could be more valuable in cases where a quick decision on the quality of a larger batch of data is preferred over studies of individual instances. For example, the existence of novelty in a set of new compounds for cancer treatment could be tested in such a fashion. Furthermore, looking at the collected data as a whole could bring benefits in statistical power. One can imagine situations, where multiple individual observations would be deemed borderline novel, but would not break the set threshold for novelty. A more holistic approach could however decide that such a group of observations would be unlikely given that the data were all normal. As an analogy, global level novelty detection relates to ordinary novelty detection in the same way as analysis of variance relates to individual t-tests. This is a novel problem that has not been previously investigated in such a gen-

eralized and formal manner. Next, we will examine the definition of this problem as a formal hypothesis test and demonstrate its application with the example dataset.

4.2 Definition of global level novelty detection

In Publication III, the problem of global level novelty detection is defined through a mixture of distributions. Let the $n \times p$ matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ and the $m \times p$ matrix $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)'$ denote the multivariate training (normal) and test (new) datasets, respectively. We assume that the observations in both sets are independent and identically distributed (i.i.d.), and that the two samples are independent. Let $1 \leq K < L$ be two integers, and let $f_1, \ldots, f_L$ be unknown but distinct density functions. Assume that the distribution of the training observations can be given as a mixture:

$$\sum_{k=1}^{K} \pi_k f_k \text{ where } \pi_k > 0 \text{ for all } k \text{ and } \pi_1 + \cdots + \pi_K = 1. \tag{4.1}$$

Moreover, assume that the distribution of the test observations follow a mixture distribution:

$$\sum_{l=1}^{L} \lambda_l f_l \text{ where } \lambda_l \geq 0 \text{ for all } l \text{ and } \lambda_1 + \cdots + \lambda_L = 1. \tag{4.2}$$

The null and alternative hypotheses for global level novelty are then written as:

- H_0 : there exist at least one representation (4.1) and (4.2) such that $\lambda_{K+1} = \cdots = \lambda_L = 0$
- H_1: for all representations (4.1) and (4.2), $\max(\lambda_{K+1}, \ldots, \lambda_L) > 0$.

This definition for global level novelty is very versatile and allows for novelty in terms of both location and scale, and in terms of completely new classes. Loosely speaking, the null hypothesis states that the new data is originated from a mixture of the same distributions as the previously observed normal data. Importantly, the proportions of the different underlying distributions in the mixture can vary freely between the two under this hypothesis. Thus, it is not a simple test of differences in distributions, such as the two-sample Kolmogorov-Smirnov test [12]. Furthermore, it is not a test about the number of latent classes, since the same density can have multiple representations as mixtures of different densities. A more thorough explanation of the problem definition with examples is given in the original Publication III.

The problem was approached by deriving various test statistics based on both existing and novel constructs, that assign a novelty value for each observation. The actual test statistics were then taken as the sum of these values. The statistics were designed so, that a large value would indicate a higher belief in global novelty in the data. Due to the complexity of the problem, the sampling distributions of these statistics under the null hypothesis are intractable. This forces us to rely on computational methods, namely the permutation test [20, 62], to measure the statistical significance of the observed statistics.

Two test statistics $\mathscr{A}$ and $\mathscr{B}$ were proposed in Publication III based on local neighborhoods around the observations. Here, we concentrate on the $\mathscr{A}$ statistic, that was found superior in high-dimensional datasets. The aim behind the related score is to detect groups of observations that are not close to the normal data while taking into account both their number and distance from a point of interest. A large score value is obtained when many new observations are tightly clustered together. Conversely, the score equals zero, if the closest observation from the point of interest is in fact a normal observation. This "raw" score is already a powerful measure of novelty, however the related tests statistic does not hold its level under the very general null hypothesis. In particular, varying class proportions were observed to break the considered methods, both existing and novel. To remedy this, a standardization for the statistic was sought in order to obtain a measure that would be approximately invariant to the class proportions. This was achieved by obtaining the expectation and variance of the score under the assumption that the observations are uniformly distributed in their local neighborhoods. Notably, deriving such a standardization for the more complex anomaly scores, such as the local outlier factor [8] or the Tukey's half-space depth [80], is a tedious if not an intractable task.

We will now demonstrate the application of the method in the PC3 example dataset. The score is dependent on one tuning parameter that is the neighborhood radius r, which is set uniquely for each test observation. Based on the findings in Publication III, we choose r such that the neighborhood contains 5% of the new data and at least one normal observation. In Publication III, the method was applied to the cross-sectional part of the PC3 dataset in order to identify treatments with novel effects compared to the control. This was achieved by calculating the proportions of 12 phenotypic groups derived in Publication I in each of the original images, thus representing each image as an 12-dimensional data point. The goal was to see, whether the drugs produced any novel images compared to the control treatment. In Figure 4.1 this analysis has been reproduced with the newly obtained 8 clusters, and by including all concentrations of the treatments. The familiar set of drugs is associated with statistically significant (uncorrected) p-values on the 5% level: Paclitaxel and the compounds 3-6, 15, 16,

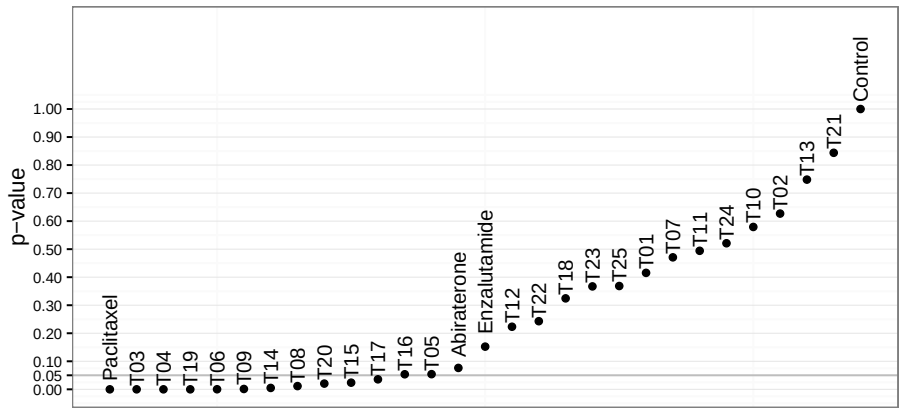

Figure 4.1: ***Results of the tests for global level novelty.***

19 and 20. Three other compounds – 8, 14 and 17 – also hit the 5% mark, however we cannot be sure of their validity without further research on the topic. It is possible that the novelty detection method is able to recover something that was not captured with the previously examined methods, or it could be that these "extra" compounds are simply cases of false discovery that are to be expected with any statistical test. Naturally, these p-values should be corrected to match a false positive rate acceptable in the context of application. For instance, after the very conservative Bonferroni correction, the corrected threshold for statistical significance would drop to 0.00178.

As a side note, the formulation of the permutation test presented in Publication III allows estimates of the p-values to be exactly zero, and is problematic when trying to control for the false positive rate among large number of tests[60]. The authors suggest to solve the issue by biased estimation of the p-value. They propose to add one artificial negative result to the set of realized permutation results i.e. add 1 both to the numerator and to the denominator of the p-value estimator.

It seems that as an early discovery tool, the method is successful in identifying the drugs of interest. The more detailed analyses discussed before, could then be carried out with this reduced set of drugs without losing any relevant candidates. In such framework, the small number of false positive discoveries can be considered acceptable at such a preliminary step of the analysis.

As an additional application of the method, we consider the problem of detecting poorly imaged or segmented structures. To this end, we consider the data from the endpoint (Day 11) of the longitudinal part of the example PC3 dataset. A set of 10 images, containing 47 well segmented

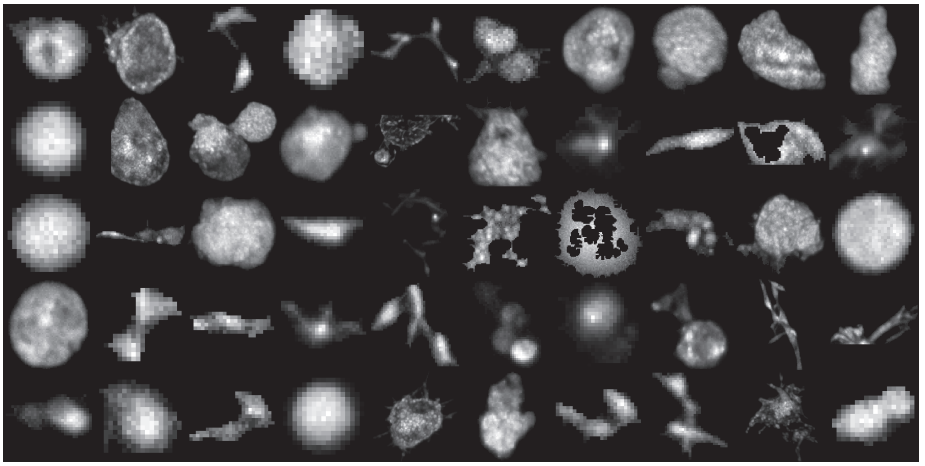

Figure 4.2: ***Structures with unusual phenotypes found using the $\mathscr{A}$ measure for novelty.***

and ordinary-looking structures, were chosen to form the training data. The rest of the ≈ 1300 structures were then compared against these normal observations to detect any groups of outliers in terms of their combined shape and texture principal component scores. The test for novelty gives us a p-value of 0.0382, indicating that there is indeed evidence for novel groups in the complete data. Figure 4.2 displays those 50 structures that are associated with the highest $\mathscr{A}$ scores and are in this sense the most extreme outliers. Based on this visual inspection, we can quickly conclude that the group truly consists of structures with peculiar shapes and textures. Many of the regularly shaped cell structures have a blurry texture indicating shortcomings in the imaging, while the more oddly shaped structures are likely to be segmentation artifacts. While this is not the intended application of the developed method, we see that the $\mathscr{A}$ score is effective in discovering also single novel observations for the purpose of quality control.

5 Concluding remarks

In this thesis, we considered statistical methods for the analysis of image data generated in organotypic 3D cancer culturing experiments. The full analysis arch from the early image pre-processing to forming the final study conclusions was addressed, and demonstrated on an example PC3 cell line dataset. A set of key topics were identified that span a wide scope of research fields in statistics and machine learning.

The main goal of image pre-processing was the identification and characterization of individual cell structures. The literature surrounding biological imaging is filled with various, partly overlapping solutions to both of these goals, and no clear all-encompassing approach could be identified. Therefore, the characteristics and needs of these particular data were carefully addressed. The local entropy filter was considered a key tool in image segmentation that has not been used in such a manner in the reviewed literature. A set of biologically informative features measuring the size, shape and texture of the cell structures were proposed. A key consideration was to obtain features that do not carry significant quantities of excess information, not relevant to the biological phenotype. This would be of less concern if the features were to be used for supervised learning, but was deemed important for the unsupervised approaches.

The estimation of treatment effects was considered from various perspectives. A straightforward approach, that is popular in the literature, was obtained by comparing the location statistics of descriptive features between the drugs. A mixed effects model taking into account various additional sources of variation was suggested as an improvement over the typically elementary statistical methods applied in the literature. Although these outside sources of variation were ultimately deemed negligible in comparison to the treatment effect, they should not be disregarded in future studies, as this might not always be the case.

While acknowledging the attractiveness of the location-based treatment effect assessments, their underlying assumptions were placed under criticism. Specifically, the existence of cancer heterogeneity suggests that modeling the features with an unimodal (and usually symmetric) distribution might not be a realistic approach. Instead, the data likely consist of multiple latent populations of cell structures that can differ in their reaction to the administered drugs. In part, this motivated the alternative approaches through FMR model and unsupervised clustering.

A flexible FMR model was proposed that allows for regression in latent classes while placing minimal assumptions on the form of the actual regression functions. While this general idea is not novel, approaches that are able to handle many covariates are not available in the literature. In the original publication, the method was shown to have superior predictive performance in simulation studies when compared against linear alternatives. However, this increased flexibility and predictive power is gained with the cost of lower interpretability and computational efficiency compared to penalized linear FMR methods. As a redeeming quality however, the method was observed to reduce to a linear FMR if the true underlying model is in fact linear.

A clustering approach was proposed for obtaining a more biologically meaningful mapping for the extracted image features. The unsupervised classification scheme was adopted to reflect the needs of the research environment, where the study questions are not constant, but change depending on the used cell line and the goals of the drug treatments. Clustering was performed separately based on the size, shape and texture features of the cell cultures, giving independent summaries of these three relevant aspects. The subsequent treatment effect estimation was then based on the proportions of these groups. In the original publication, much care was devoted to the selection of appropriate features and to proving the validity of the method with simulation studies. While the grand idea of applying clustering methods to imaging data is not new, an approach with the same level of specialization has not been introduced for the purpose. In this thesis, a relatively simplistic approach was adopted in the actual statistical analysis of the group data, which was considered enough to convey the core message of the chapter. However, a more statistically rigorous approach, similar to the mixed effect models considered for the individual features, could be beneficial also in this context.

As the final contribution of the thesis, the novel problem of global level novelty detection was introduced. The problem was formulated as hypothesis test, constructed on mixtures of unidentified distributions. A key issue in discovering a valid test statistic was obtaining (approximate) invariance towards the changes in proportions of these latent classes of densities. This was achieved with a standardization of the test statistic that would have been tedious if not impossible to obtain for the existing novelty detection methods. In addition to performing drug selection, the method was shown to be capable of identifying abnormal cell structures within the data. As such, the method serves as an opening for discussion in this new problem, rather than being an exhaustive solution. The very general definition of global level novelty encompasses many more types of settings than were not possible to be investigated in the scope of one publication. Furthermore, the generalization of the method to categorical variables, or to any other non-

numeric data, through specialized distance metrics was not touched upon.

Having a specific application in mind, one is forced to consider many practical issues that are not directly related to the statistical properties of the methods. As many of the analyses discussed here are intended to be conducted by non-statistician practitioners, the adopted methods should behave reliably in all imaginable future experiments. This requires robustness both in the computational and statistical sense. While we have demonstrated that the methods are applicable in real datasets, some of these issues could be more thoroughly investigated. As a practical example, the suggested shape and texture features require a sufficiently large cell structure to be computable. While this was not an issue in the datasets considered here, future studies could use lower resolution data, and thus display similar structures with a fewer pixels. Computational efficiency is another critical issue given the large size of the datasets. However, in the applications considered here, the greatest bottlenecks relate mainly to the image processing step, rather than to the following statistical analyses. Thus, the greatest gains in computation time can likely be achieved by optimizing these early stages of the analysis.

Summaries of original publications

I The problem of quantifying treatment effects in heterogeneous 3D cancer cell cultures is tackled by considering the differences in the proportions of phenotypic groups. A set of features are considered that capture the biologically relevant information relating to the size, shape and texture of the multicellular cell structures. The dimension of the features is reduced using robust principal components analysis (PCA) after which the structures are hierarchical clustered into phenotypic groups. A simulation study is performed, where the method is shown to behave well in the presence of mild random modifications in the size, shape and texture of the structures.

II Methods for the analysis of patient-derived xenograft (PDX) 3D cultures are investigated. An efficient segmentation approach based on the local entropy filter is proposed and compared against existing alternatives. The obtained foreground is split into local regions that are classified into either malignant or benign tissues in a supervised manner. The measured *in vitro* effects of the administered drug treatments are compared against their observed *in vivo* efficacy in mouse experiments.

III The problem of global level novelty detection is defined. The problem is tackled with a set of hypothesis tests based on the permutation principle. Two test statistics are presented that operate in local neighborhoods around the observations. Standardizations are sought, that render the scores invariant to differences in cluster proportions when the local densities are close to uniform. Rigorous simulation studies are considered for validating the tests and assessing their power in various settings. The developed methods are applied to three real datasets in a similar fashion and their utility in discovering novel treatments in 3D cancer culturing experiments is demonstrated.

IV A flexible variation of the finite mixture regression (FMR) is presented for prediction purposes. Random forest is utilized for obtaining a set of indicator variables that account for possible nonlinear dependencies and interactions in the data. A penalized FMR model is used for selecting the important variables from the original covariates and from the obtained indicator variables. Bootstrap is applied for correcting the level of obtained prediction intervals. The method is shown to outperform existing linear alternatives in extensive simulations studies, some containing large sets of surplus variables. The method is successfully applied to a wage prediction dataset.

Bibliography

[1] M. Åkerfelt, N. Bayramoglu, S. Robinson, M. Toriseva, H.-P. Schukov, V. Härmä, J. Virtanen, R. Sormunen, M. Kaakinen, J. Kannala, et al. Automated tracking of tumor-stroma morphology in microtissues identifies functional targets within the tumor microenvironment for therapeutic intervention. *Oncotarget*, 6(30):30035–30056, 2015.

[2] J. D. Banfield and A. E. Raftery. Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. *Journal of the American Statistical Association*, 87(417):7–16, 1992.

[3] S. Bhatia, J. V. Frangioni, R. M. Hoffman, A. J. Iafrate, and K. Polyak. The challenges posed by cancer heterogeneity. *Nature Biotechnology*, 30(7):604, 2012.

[4] C. C. Bilgin, S. Kim, E. Leung, H. Chang, and B. Parvin. Integrated profiling of three dimensional cell culture models and 3D microscopy. *Bioinformatics*, 29(23):3087–3093, 2013.

[5] M. V. Boland and R. F. Murphy. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of hela cells. *Bioinformatics*, 17(12):1213–1223, 2001. doi: 10.1093/bioinformatics/17.12.1213.

[6] D. Bradley and G. Roth. Adaptive thresholding using the integral image. *Journal of Graphics, GPU, and Game Tools*, 12(2):13–21, 2007.

[7] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[8] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: Identifying Density-Based Local Outliers. In *ACM Sigmod Record*, volume 29, pages 93–104. ACM, 2000.

[9] N. Chaki, S. H. Shaikh, and K. Saeed. *Exploring Image Binarization Techniques*. Springer, 2014.

[10] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15, 2009.

[11] L. Clifton, D. A. Clifton, P. J. Watkinson, and L. Tarassenko. Identification of patient deterioration in vital-sign data using one-class support

vector machines. In *Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on*, pages 125–131. IEEE, 2011.

[12] W. J. Conover. *Practical Nonparametric Statistics*. Wiley New York, 1980.

[13] J. Cuesta-Albertos, A. Gordaliza, C. Matrán, et al. Trimmed k-means: An attempt to robustify quantizers. *The Annals of Statistics*, 25(2): 553–576, 1997.

[14] F. De Chaumont, S. Dallongeville, N. Chenouard, N. Hervé, S. Pop, T. Provoost, V. Meas-Yedid, P. Pankajakshan, T. Lecomte, Y. Le Montagner, et al. Icy: an open bioimage informatics platform for extended reproducible research. *Nature Methods*, 9(7):690–696, 2012.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38, 1977.

[16] Z. Di, M. J. Klop, V.-M. Rogkoti, S. E. Le Dévédec, B. van de Water, F. J. Verbeek, L. S. Price, and J. H. Meerman. Ultra high content image analysis and phenotype profiling of 3d cultured micro-tissues. *PLoS ONE*, page e109688, 2014. doi: 10.1371/journal.pone.0109688.

[17] A. A. Dima, J. T. Elliott, J. J. Filliben, M. Halter, A. Peskin, J. Bernal, M. Kociolek, M. C. Brady, H. C. Tang, and A. L. Plant. Comparison of segmentation algorithms for fluorescence microscopy images of cells. *Cytometry Part A*, 79(7):545–559, 2011.

[18] R. Edmondson, J. J. Broglie, A. F. Adcock, and L. Yang. Three-dimensional cell culture systems and their applications in drug discovery and cell-based biosensors. *ASSAY and Drug Development Technologies*, 12(4):207–218, 2014.

[19] R. Fisher, L. Pusztai, and C. Swanton. Cancer heterogeneity: implications for targeted therapeutics. *British Journal of Cancer*, 108(3): 479–485, 2013.

[20] R. A. Fisher. *The design of experiments*. Oliver & Boyd, 1935.

[21] G. Galimberti, A. Montanari, and C. Viroli. Penalized factor mixture analysis for variable selection in clustered data. *Computational Statistics & Data Analysis*, 53(12):4301–4310, 2009.

[22] L. A. García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar. A review of robust clustering methods. *Advances in Data Analysis and Classification*, 4(2-3):89–109, 2010.

[23] J. Han, H. Chang, O. Giricz, G. Y. Lee, F. L. Baehner, J. W. Gray, M. J. Bissell, P. A. Kenny, and B. Parvin. Molecular predictors of 3D morphogenesis by breast cancer cell lines in 3D culture. *PLoS Computational Biology*, 6(2):e1000684, 2010.

[24] R. M. Haralick, K. Shanmugam, and I. H. Dinstein. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, (6):610–621, 1973.

[25] V. Härmä, J. Virtanen, R. Mäkelä, A. Happonen, J.-P. Mpindi, M. Knuuttila, P. Kohonen, J. Lötjönen, O. Kallioniemi, and M. Nees. A comprehensive panel of three-dimensional models for studies of prostate cancer growth, invasion and drug responses. *PLoS ONE*, 5(5):e10431, 2010.

[26] V. Härmä, H.-P. Schukov, A. Happonen, I. Ahonen, J. Virtanen, H. Siitari, M. Åkerfelt, J. Lötjönen, and M. Nees. Quantification of Dynamic Morphological Drug Responses in 3D Organotypic Cell Cultures by Automated Image Analysis. *PloS One*, 9(5):e96426, 2014.

[27] V. Härmä, R. Haavikko, J. Virtanen, I. Ahonen, H.-P. Schukov, S. Alakurtti, E. Purev, H. Rischer, J. Yli-Kauhaluoma, V. M. Moreira, et al. Optimization of invasion-specific effects of betulin derivatives on prostate cancer cells through lead development. *PLoS ONE*, 2015.

[28] T. Hastie, R. Tibshirani, J. Friedman, et al. *The Elements of Statistical Learning*. Springer, 2nd edition, 2009.

[29] J. W. Haycock. 3d cell culture: a review of current approaches and techniques. *3D Cell Culture: Methods and Protocols*, pages 1–15, 2011.

[30] V. J. Hodge and J. Austin. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

[31] M. T. Hoque, L. C. Windus, C. J. Lovitt, and V. M. Avery. Pcaanalyser: A 2d-image analysis based module for effective determination of prostate cancer progression in 3d culture. *PLoS One*, 8(11):e79865, 2013.

[32] J. S. Horoszewicz, S. S. Leong, E. Kawinski, J. P. Karr, H. Rosenthal, T. M. Chu, E. A. Mirand, and G. P. Murphy. Lncap model of human prostatic carcinoma. *Cancer Research*, 43(4):1809–1818, 1983.

[33] M. Huang, R. Li, and S. Wang. Nonparametric Mixture of Regression Models. *Journal of the American Statistical Association*, 108(503):929–941, 2013.

[34] M. Hubert, P. J. Rousseeuw, and K. Vanden Branden. Robpca: A new approach to robust principal component analysis. *Technometrics*, 47 (1):64–79, 2005.

[35] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[36] A. Joly, F. Schnitzler, P. Geurts, and L. Wehenkel. L1-based compression of random forest models. In *20th European Symposium on Artificial Neural Networks*, 2012.

[37] T. R. Jones, I. H. Kang, D. B. Wheeler, R. A. Lindquist, A. Papallo, D. M. Sabatini, P. Golland, and A. E. Carpenter. Cellprofiler analyst: data exploration and analysis software for complex image-based screens. *BMC Bioinformatics*, 9(1):482, 2008.

[38] V. Jyothsna, V. R. Prasad, and K. M. Prasad. A review of anomaly based intrusion detection systems. *International Journal of Computer Applications*, 28(7):26–35, 2011.

[39] M. Kaighn, K. S. Narayan, Y. Ohnuki, J. Lechner, L. Jones, et al. Establishment and characterization of a human prostatic carcinoma cell line (PC-3). *Investigative Urology*, 17(1):16, 1979.

[40] P. Kankaanpää, L. Paavolainen, S. Tiitta, M. Karjalainen, J. Päivärinne, J. Nieminen, V. Marjomäki, J. Heino, and D. J. White. Bioimagexd: an open, general-purpose and high-throughput image-processing platform. *Nature Methods*, 9(7):683–689, 2012.

[41] P. A. Kenny, G. Y. Lee, C. A. Myers, R. M. Neve, J. R. Semeiks, P. T. Spellman, K. Lorenz, E. H. Lee, M. H. Barcellos-Hoff, O. W. Petersen, et al. The morphologies of breast cancer cell lines in three-dimensional assays correlate with their profiles of gene expression. *Molecular Oncology*, 1(1):84–96, 2007.

[42] A. Khalili and J. Chen. Variable Selection in Finite Mixture of Regression Models. *Journal of the American Statistical Association*, 102(479), 2007.

[43] A. Khalili, J. Chen, and S. Lin. Feature selection in finite mixture of sparse normal linear models in high-dimensional feature space. *Biostatistics*, 2010. doi: 10.1093/biostatistics/kxq048.

[44] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9): 1464–1480, 1990.

[45] K. Kvilekval, D. Fedorov, B. Obara, A. Singh, and B. Manjunath. Bisque: a platform for bioimage analysis and management. *Bioinformatics*, 26 (4):544–552, 2010.

[46] C. Li and P. K.-S. Tam. An iterative algorithm for minimum cross entropy thresholding. *Pattern Recognition Letters*, 19(8):771–776, 1998.

[47] V. Ljosa and A. E. Carpenter. Introduction to the quantitative analysis of two-dimensional fluorescence microscopy images for cell-based screening. *PLoS Computational Biology*, 5(12):e1000603, 12 2009. doi: 10.1371/journal.pcbi.1000603.

[48] L.-H. Loo, L. F. Wu, and S. J. Altschuler. Image-based multivariate profiling of drug responses from single cells. *Nature Methods*, 4(5): 445–453, 2007.

[49] L. Ma, J. Barker, C. Zhou, W. Li, J. Zhang, B. Lin, G. Foltz, J. Küblbeck, and P. Honkakoski. Towards personalized medicine with a three-dimensional micro-scale perfusion-based two-chamber tissue model system. *Biomaterials*, 33(17):4353–4361, 2012.

[50] M. Markou and S. Singh. Novelty Detection: A Review - Part 1: Statistical Approaches. *Signal Processing*, 83:2003, 2003.

[51] M. Markou and S. Singh. Novelty Detection: A Review - Part 2: Neural Network Based Approaches. *Signal Processing*, 83:2499–2521, 2003.

[52] G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, 2004.

[53] B. Misselwitz, G. Strittmatter, B. Periaswamy, M. Schlumberger, S. Rout, P. Horvath, K. Kozak, and W.-D. Hardt. Enhanced cellclassifier: a multi-class classification tool for microscopy images. *BMC Bioinformatics*, 11(1):30, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-30. URL `http://www.biomedcentral.com/1471-2105/11/30`.

[54] A. Mitra, L. Mishra, and S. Li. Technologies for deriving primary tumor cells for use in personalized cancer therapy. *Trends in Biotechnology*, 31(6):347–354, 2013.

[55] W. Niblack. *An Introduction to Digital Image Processing*. Strandberg Publishing Company, Birkeroed, Denmark, Denmark, 1985. ISBN 87-872-0055-4.

[56] N. Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[57] F. Pampaloni, E. G. Reynaud, and E. H. Stelzer. The third dimension bridges the gap between cell culture and live tissue. *Nature Reviews Molecular Cell Biology*, 8(10):839–845, 2007.

[58] W. Pan and X. Shen. Penalized model-based clustering with application to variable selection. *The Journal of Machine Learning Research*, 8: 1145–1164, 2007.

[59] C. C. Park, W. Georgescu, A. Polyzos, C. Pham, K. M. Ahmed, H. Zhang, and S. V. Costes. Rapid and automated multidimensional fluorescence microscopy profiling of 3D human breast cultures. *Integrative Biology*, 5(4):681–691, 2013.

[60] B. Phipson and G. K. Smyth. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010.

[61] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

[62] E. J. G. Pitman. Significance Tests Which May Be Applied to Samples From Any Populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130, 1937.

[63] J. A. Quinn and C. K. Williams. Known unknowns: Novelty detection in condition monitoring. In *Pattern Recognition and Image Analysis*, pages 1–6. Springer, 2007.

[64] S. Robinson, L. Guyon, J. Nevalainen, M. Toriseva, M. Åkerfelt, and M. Nees. Segmentation of Image Data from Complex Organotypic 3D Models of Cancer Tissues with Markov Random Fields. *PLoS ONE*, 2015. doi: 10.1371/journal.pone.0143798.

[65] P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

[66] C. A. Schneider, W. S. Rasband, K. W. Eliceiri, et al. Nih image to imagej: 25 years of image analysis. *Nature Methods*, 9(7):671–675, 2012.

[67] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New Support Vector Algorithms. *Neural Computation*, 12(5):1207–1245, 2000.

[68] G. Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[69] L. Shamir, J. D. Delaney, N. Orlov, D. M. Eckley, and I. G. Goldberg. Pattern recognition software and techniques for biological image analysis. *PLoS Comput Biol*, 6(11):e1000974, 11 2010. doi: 10.1371/journal.pcbi.1000974.

[70] N. Sharma, L. M. Aggarwal, et al. Automated medical image segmentation techniques. *Journal of Medical Physics*, 35(1):3, 2010.

[71] C. Smyth, D. Coomans, and Y. Everingham. Clustering noisy data in a reduced dimension space via multivariate regression trees. *Pattern Recognition*, 39(3):424–431, 2006.

[72] N. Städler, P. Bühlmann, and S. Van De Geer. L1-penalization for Mixture Regression Models. *Test*, 19(2):209–256, 2010.

[73] K. R. Stone, D. D. Mickey, H. Wunderli, G. H. Mickey, and D. F. Paulson. Isolation of a human prostate carcinoma cell line (du 145). *International Journal of Cancer*, 21(3):274–281, 1978.

[74] C. Surace and K. Worden. Novelty detection in a changing environment: a negative selection approach. *Mechanical Systems and Signal Processing*, 24(4):1114–1128, 2010.

[75] L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady. Novelty detection for the identification of masses in mammograms. In *Artificial Neural Networks, 1995., Fourth International Conference on*, pages 442–447. IET, 1995.

[76] M. R. Teague. Image analysis via the general theory of moments. *Journal of the Optical Society of America*, 70(8):920–930, 1980.

[77] R. Tibshirani and G. Walther. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3):511–528, 2005.

[78] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[79] W.-H. Tsai. Moment-preserving thresholding: A new approach. In *Document Image Analysis*, pages 44–60. IEEE Computer Society Press, 1995.

[80] J. W. Tukey. Mathematics and the Picturing of Data. In *Proceedings of the International Congress of Mathematicians*, volume 2, pages 523–531, 1975.

[81] V. N. Vapnik and V. Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[82] L. Vincent and P. Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):583–598, 1991.

[83] J. Wang, X. Zhou, P. L. Bradley, S.-F. Chang, N. Perrimon, and S. T. Wong. Cellular Phenotype Recognition for High-Content RNA Interference Genome-Wide Screening. *Journal of Biomolecular Screening*, 13(1):29–39, 2008.

[84] S. Xiang. *Semiparametric Mixture Models*. PhD thesis, Kansas State University, 2014.

[85] F. Xing, J. Saidou, and K. Watabe. Cancer associated fibroblasts (CAFs) in tumor microenvironment. *Frontiers in Bioscience*, 15:166, 2010.

Annales Universitatis Turkuensis

Turun yliopisto
University of Turku

Painosalama Oy, Turku, Finland 2018

ISBN 978-951-29-7111-4 (PRNT)
ISBN 978-951-29-7112-1 (PDF)
ISSN 0082-7002 (PRINT) | ISSN 2343-3175 (PDF)