JOHANNES KEPLER
UNIVERSITÄT LINZ | JKU

TNF

Technisch-Naturwissenschaftliche
Fakultät

Analysis of single molecule microscopy images with application to ultra-sensitive microarrays

DISSERTATION

zur Erlangung des akademischen Grades

Doktorin

im Doktoratsstudium der

TECHNISCHEN WISSENSCHAFTEN

Eingereicht von:

Leila Mureşan

Angefertigt am:

Institut für Wissensbasierte Mathematische Systeme

Betreuung:

Univ.-Prof. Dr. Erich Peter Klement

A. Univ.-Prof. Dr. Gerhard Schütz

Linz, March 2010

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Linz, 26.02.2010 Leila Mureşan

Acknowledgment

First of all, I would like to express my deep gratitude to my supervisors, Prof. Erich Peter Klement and Prof. Gerhard Schütz for the opportunity to discover an exciting field of research and for the continuous guiding, support and encouragement they provided during all these years.

I would like to thank the GEN-AU program of the Austrian Federal Ministry of Education, Science and Culture for funding the research. Within this project, I found the close cooperation with Dr. Jan Hesse and Dr. Jaroslaw Jacak very enriching and motivating, and enjoyed the tremendous advantages of teamwork in multidiciplinary research. I thank all the project members, colleagues from the Biophysics Institute, (Johannes Kepler University, Linz), Dr. Alois Sonnleitner and the Center for Biomedical Nanotechnology, Upper Austrian Research GmbH, Prof. Anna-Maria Frischauf and Prof. Fritz Aberger at the Department of Molecular Biology, Division of Genomics, (University of Salzburg) and Prof. Hannes Stockinger and his team at the Department of Molecular Immunology, (Medical University of Vienna) for the interesting discussions and the inspiration they provided.

I am very grateful to Dr. Irene Tiemann-Boege for explaining me the biological background and relevance of microarrays, as well as for all the help and advices she provided, and the careful reviewing of the biology related part of this thesis.

I would like to thank DI. Bettina Heise for the collaboration during the whole project, and all my colleagues and friends for the great time spent together. I have a very special thought for Sabine Lumpi for all the help in (not only) administrative matters.

Finally, I thank Jérôme and my family for everything.

Contents

Abstract

This work presents the analysis of images obtained via a novel ultra-sensitive microarray technique. The practical goal of the new technology is to compare the concentrations of mRNA in cases when only minute amounts of samples are available, amounts that cannot be analyzed by classical methods.

The two main parts of the analysis are related to the detection of single molecules in the images recorded under these special conditions and the analysis of the point patterns represented by the positions of the detected molecules.

For the first part, an adapted wavelet thresholding method was investigated, the thresholding being based on the control of the false discovery ratio (FDR). A study of the influence of the noise models (Gaussian, Poisson and a Gauss-Poisson mixture) on the detection accuracy as well as a possible way to cope with the correlation of the undecimated wavelet coefficients are given.

The intensity of the point patterns representing the positions of the detected single molecules is assumed to be piece-wise constant, typically one concentration characterizing the hybridized molecules inside the microarray spot and another the clutter outside the spot. The shape of the microarray spot is not fixed (although usually circular), permitting the modeling of evaporation effects, spotting errors etc. Three approaches were studied and compared: a method of moments applied to count data, an expectation-maximization on k-nearest neighbor distances and a level set segmentation method on the point densities.

The results of the analysis were validated on simulated and real data. Differently expressed genes were detected via the presented method for multiple myeloma samples, result validated by an independent biological technique (qPCR). The techniques presented in this work can be directly applied to other single molecule imaging experiments.

Die vorliegende Arbeit präsentiert ein Bildanalyseverfahren, das für eine neuartige hochempfindliche Microarray-Technik entwickelt wurde. Das praktische Ziel dieser neuen Technologie ist es, die Konzentration von mRNA in jenen Fällen zu vergleichen, wo nur so geringe Probenmengen zur Verfügung stehen, daß diese nicht durch klassische Methoden analysiert werden können.

Die beiden Hauptaufgaben der Analyse sind einerseits die Erkennung einzelner Moleküle unter den besonderen Bedingungen, die durch die neue Technologie impliziert werden, und andererseits die Analyse der Punktmuster, welche die Positionen der erkannten Moleküle repräsentieren.

Für die erste Aufgabe wurde eine adaptierte Wavelet Thresholding Methode verwendet, wobei die Schwellwerte auf der Kontrolle des False Discovery Ratio (FDR) basierten. Es wurde sowohl der Einfluss von Rausch-Modellen (Gauß, Poisson und ein Gauß-Poisson Kombination) auf die Genauigkeit der Erkennung von einzelnen Molekülen untersucht als auch ein möglicher Weg, um der Korrelation der Undecimated Wavelet Koeffizienten gerecht zu werden.

Die Intensität der Punktmuster, welche die Positionen der erkannten Moleküle repräsentieren, wird als stückweise konstant angenommen, wobei typischerweise eine Konzentration die hybridisierten Moleküle im Microarray-Spot und eine andere Konzentration die Stördaten außerhalb charakterisiert. Die Form des Microarray-Spots ist nicht festgelegt (obwohl sie als kreisförmig angenommen wird), was es unter anderem ermöglicht, Verdunstungseffekte oder Spotting Fehler zu modellieren.

Drei Ansätze wurden untersucht und verglichen: Eine Methode der Momente angewandt auf Häufigkeitsdaten, eine Expectation-Maximization von k-NN Distanzen und eine Level Set Segmentation Methode auf den Punktdichten.

Die Ergebnisse der Analyse wurden anhand simulierter und realer Daten validiert. Das vorgestellte Verfahren erkannte unterschiedliech exprimierte Gene für Proben des Multiplen Myeloms und das Resultat durch eine unabhängige biologische Technik (qPCR) nachgewiesen. Die in dieser Arbeit vorgestellten Techniken können auch direkt bei anderen Einzelmolekülfluoreszenz-Experimente angewendet werden.

List of symbols

$\perp$	Orthogonality $\quad U \perp V : \forall u \in U, v \in V, \quad \langle u, v \rangle = 0$
$\dot{+}$	Direct sum $\quad U \dot{+} V = \{u + v \mid u \in U, v \in V, \quad U \cap V = \{0\}\}$
$\langle f, g \rangle$	Inner product $\quad \langle f, g \rangle := \int f(x)\overline{g(x)}\, dx$
$\mathbb{E}(X)$	Expectation of a random variable X
$\mathcal{F}f(\omega)$	Fourier transform $\quad \mathcal{F}f(\omega) = \widehat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int f(x) e^{-ix\omega}\, dx$
$\mathcal{N}(\mu, \sigma)$	Normal distribution with mean μ and variance σ^2
$\mathcal{P}oi(\mu)$	Poisson distribution of parameter μ
$\mid A \mid$	Cardinality of the set A
$\oplus$	Orthogonal sum $\quad U \oplus V = U \dot{+} V, \quad$ and $\quad U \perp V$
$\otimes$	Tensor product
$\widehat{x}$	Estimator of x
$\widetilde{x}$	Empiric value of x
$f * g$	Convolution $\quad (g * f)(t) := \int g(t - s) f(s)\, ds$
i.i.d.	Independent and identically distributed

List of Figures

Chapter 1

Introduction

The aim of this work is to provide efficient tools for the analysis of single molecules and the patterns they form in fluorescence microscopy images in general and ultra-sensitive microarray scans in particular. The features of interest are fluorescent-tagged single molecules, their image being typically equivalent to diffraction limited point sources. The number, intensity and relative positions of single molecules, and at a different scale, the various pattern these single molecules form represent a source of information that can be exploited in inference on biological processes. This information is not available in classical, lower resolution microscopy imaging, in which case only an aggregated signal intensity is being measured. In order to achieve detection of single molecules, high resolution and high sensitivity images have to be obtained through a set of techniques affecting the resulting two-dimensional signal. These techniques introduce new challenges including the handling of significantly bigger images that correspond to the same scanned sample size and also of a different imaging regime. Single molecule imaging is characterized by different dynamic range and under certain conditions by low signal-to-noise ratios as well as a different impact of the Poisson and Gaussian measurement noise on the recorded signal.

1.1 Context

The work presented in this thesis was performed in the frame of the multidisciplinary project *Ultra-sensitive genomics and proteomics* funded by *Genome Research - Austria* including the following project partners

- Biophysics Institute - Johannes Kepler University of Linz

- Upper Austrian Research GmbH - Linz
- Department of Knowledge-based Mathematical Systems - Johannes Kepler University of Linz
- Department of Molecular Biology - University of Salzburg
- Department of Molecular Immunology - Medical University of Vienna

The biological expertise was provided and the biological samples prepared by the partners at the Department of Molecular Biology, University of Salzburg. The ultra-sensitive microarray technique was developed and perfected by the Biophysics Institute at Johannes Kepler University, Linz and Upper Austrian Research GmbH, Linz. This work was done in close collaboration with Dr. Jan Hesse and Dr. Jaroslaw Jacak and the project coordinator Dr. Gerhard Schütz.

1.2 Motivation

We shall motivate the technique of ultra-sensitive microarrays via the example of the cancer stem cells (CSC) hypothesis [110].

The hypothesis received much attention recently and it states that tumors are initiated by a small population of tumor cells similar to adult stem cells, that have the ability to self-renew as well as give rise to differentiated tissue cells. Knowledge of the gene expression profile of these special CSC is crucial in understanding the biological mechanisms at work and in development of effective therapies. A quite well established way to determine the expression profile is the microarray technique that will be described in Chapter 3. The technique offers the advantage of studying thousands of genes in the frame of a single experiment, thus under identical experimental conditions. However it has several drawbacks, one of the most restrictive in our case being the relatively important quantity of target sample mRNA necessary for a reliable analysis. Since the fraction of CSC might be as low as 1% [46] the classical microarray technique cannot be applied directly to the minute amount of mRNA obtained from CSC.

In order to alleviate the restriction imposed by the small amount of available mRNA, the microarray technology is combined with high resolution imaging. Thanks to the development of the Nanoreader [56], a fast and highly sensitive imaging system, the scanning of areas of 1×0.2 cm^2 at a pixel size of 200 nm

became possible within 50 seconds, with a good ability to discriminate among minute differences.

The resulting high resolution images of microarray spots can be understood as a zoom in the classical, low-resolution microarray images and are formed of single molecules clustered in a circular pattern, representing the hybridized single molecules of interest corresponding to one gene. The analysis of this new kind of images and the estimation of the hybridization signal makes the object of the present work.

1.3 Outline of the thesis

Excluding this introduction, the thesis consists of seven chapters. The outline of the chapters is given in the following.

The principles of fluorescence microscopy with focus on single molecule imaging are presented in Chapter 2. Mathematical models of image formation as well as single molecule image content are proposed in the end of the chapter. The next chapter, Chapter 3 gives an overview of the microarray technology, together with a discussion of the importance of the method, its advantages and the challenges of the classical low-resolution approach.

The following two chapters are related to single molecule detection in the fluorescence microscopy images. First, Chapter 4 sets the framework for denoising and detection via a brief introduction to wavelet transforms, then in Chapter 5 a detailed discussion of wavelet thresholding as an approach to denoising and detection is given, together with several methods for threshold selection, adaptation to specific noise models and effect of signal sparsity on detection. The discussion emphasizes detection approaches and results for the models described in Chapter 2.

After the detection of single molecules, we estimate the concentration of the hybridized molecules inside the microarray spot. This new hybridization measure requires the separation of molecules bound to the microarray spot from unspecific binding, dirt etc. outside of it. The modeling of the problem is based on spatial point patterns and is described in Chapter 6 together with the algorithms that separate two superposed point patterns (signal and clutter) based on the intensities of the two processes.

The validation of the method of ultra-sensitive microarray technique is done through analysis based on simulated images and image series as well as on real data. The real data includes microarray images of oligonucleotide dilution series

and the scans of three competitively hybridized slides pertaining to an experiment on multiple myeloma stem cells. The results are gathered in Chapter 7.

The last chapter of the thesis, Chapter 8, summarizes problems discussed in this work as well as the proposed methods and concludes with possible applications of the algorithms presented and indicates further lines of research.

Chapter 2

Single molecule imaging: techniques and models

The main purpose of microscopic imaging is to detect small structures and visualize their dynamics. Although other techniques can achieve the same or even higher resolution (e.g. electron microscopy, atomic force microscopy etc.), fluorescence microscopy is the most important *in vivo* approach, that can operate in biologically relevant condition, without (or only minimally) disturbing the observed process. Moreover the great specificity and ease of use make fluorescence microscopy the main light microscopy tool in biomedical research.

In 1976, Hirschfeld achieved the first successful detection of single molecules in solution [57] (although marking the molecule with multiple labels) and ever since the field has received increasing attention in the fluorescence microscopy community.

2.1 Fluorescence and fluorophores

The information offered by a typical light microscope is the intensity of light emitted by an object, measured after having passed through an optical system. In the case of fluorescence microscopy, the light is emitted by specific molecules, called *fluorophores*, *fluorochromes* or *fluorescent dyes* that have the property that after excitation with light at a certain wavelength as a response emit light at a longer wavelength. If the molecules of interest are nonfluorescent, they can be tagged with a fluorescent dye to make them visible. Various labeling methods have been developed, like epitope tagging (inserting short DNA sequences of known epitopes into the coding sequences of proteins), fluorescent proteins (the best known being

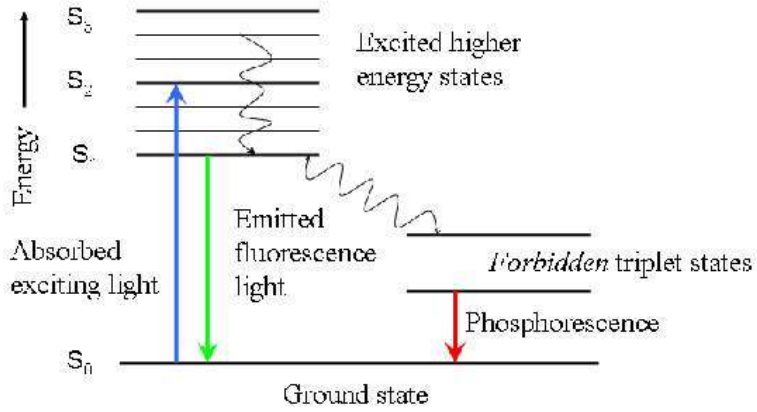


Figure 2.1: Jablonski diagram showing photophysical transitions.

the green fluorescent protein *GFP* used to create fluorescent chimeric proteins that can be expressed in living cells), immunofluorescence microscopy via fluorescent antibodies that label fixed, permeabilized cells etc. Briefly, the principle of fluorescence is explained as a cycle in which an electron of a fluorescent molecule is excited to a higher energy state following the absorption of a photon of the excitation wavelength. Almost immediately (in an interval of $10^{-9} - 10^{-12}$ seconds) the electron returns to its initial ground state, as an effect the molecule possibly releasing the absorbed energy as a fluorescent photon. The emitted fluorescent photon typically exhibits a longer wavelength than the absorbed photon, due to the energy loss through the process, the difference in wavelengths representing the Stokes shift (we shall discuss below). The cycle is depicted graphically in the Jablonski diagram, Fig. 2.1, which shows increasing energy states as a stack of horizontal lines. There are two categories of excited states (characterized by different spin states of the excited electrons): the singlet excited state and the triplet excited state (occurring via spin-flipping). With high probability the electron is excited to a singlet state (straight upward pointing blue arrow), and when it collapses to the ground state, energy can be given up as fluorescence emission (downward pointing green arrows). Alternatively, the electron can return to ground state without photon emission, the energy being released as heat (internal conversion). However, there is a probability that an electron can also enter the triplet excited state through a process called *intersystem crossing*. There is a possibility of emission process from this triplet state after a few seconds or even minutes, through a process called phosphorescence. For more details see [85].

The true dynamic nature of fluorescence emission can be examined at single molecules level, when *individual molecules are seen to rapidly cycle between emis-*

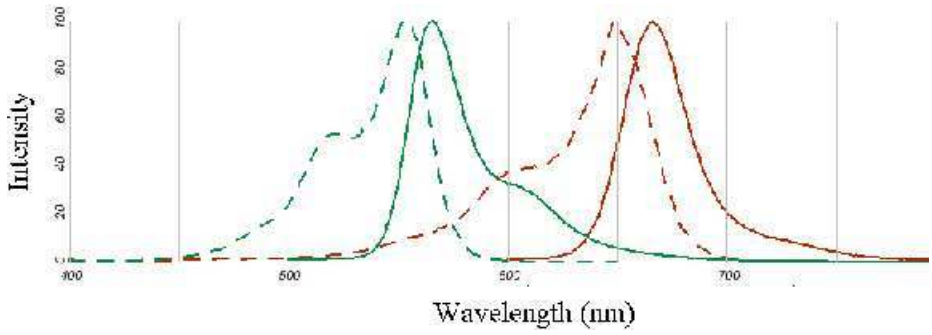


Figure 2.2: Absorbance (dashed line) and emission(solid line) spectra of Cy3 (green) and Cy5 (red). (Figure made with Fluorescence Spectra Viewer, Invitrogen [1])

sive and nonemissive states, displaying a highly complex and fascinating dynamic behavior [32].

The spectrum of wavelengths over which molecules absorb light (in order to re-emit it) is called *excitation spectrum*, while the emitted fluorescent light from excited dye molecules likewise ranges over a broad spectrum of longer wavelengths, called *the emission spectrum*.

Although the excitation and emission spectra of fluorescent molecules overlap, usually there is a distance between their respective maxima known as *Stokes shift*. This separation permits to distinguish between the excitation light and the fluorescent signal with the aid of specially designed filters.

The following fluorescence related characteristics influence the quality of the measured signal (the emitted fluorescence light): *quantum efficiency* - the fraction of absorbed photons quanta that is re-emitted by a fluorochrome, *photobleaching* the —usually permanent— loss of fluorescence due to chemical damage, when the fluorescent molecule transits to a triplet excited state and *autofluorescence*, the part of the signal that is not due to the fluorescent tags. Cells contain autofluorescing metabolites, but dirt can also contribute to the autofluorescence of the image.

Smith *et al.* [102] have proved that cell autofluorescence also increases with cell DNA content. In the case of microarray slides, autofluorescence will play a role due to the fluorescence of dirt particles on the slide that might hamper the true cDNA signal. As for photobleaching at single molecule level in [43] a detailed description is given.

2.2 Microscopy techniques

In order to achieve the goal of imaging single molecules, which emit only a limited number of photons, the imaging technique needs to show high sensitivity and suppress the influence of the background and the structures of no interest (outside of the focal plane).

Single molecules can be observed with simple methods such as wide-field microscopy, illuminating an area of several microns of the specimen. The schematic representation of an epifluorescence microscope is given in Fig. 2.3. The (not collimated) laser beam is reflected by the dichroic mirror toward the microscope objective and illuminates the sample. The fluorescent emission is collected through the same microscope objective and transmitted through the dichroic mirror. Filters eliminate the residual excitation light. It is important in order to maximize the detection of fluorescent light to maximize the *numerical aperture* (NA) of the objective, defined as: $NA = n \sin \varphi$, where n is the refractive index of the medium between the sample and objective and φ the maximum collection angle ([81]).

The sensitivity can be improved by limiting the excitation volume through e.g. two techniques: laser scanning confocal microscopy (LSCM) and total internal reflectance fluorescence (TIRF) [81].

In the case of **LSCM** both illumination and detection are confined to a single, diffraction-limited, spot in the specimen. To obtain an image, the procedure is repeated scanning across the specimen using some form of scanning device. Schematically the system is represented in Fig. 2.3.

The laser beam is reflected by a dichroic mirror and passes through the microscope objective and is focused to a diffraction-limited spot at the focal plane. The objective collects the emitted fluorescent light as well as the backscattered laser light and passes it through the dichroic beamsplitter. Filters help to eliminate residual laser light. In front of the detector a pinhole is inserted which prevents the out-of-focus light to reach the detector. The diameter of the pinhole determines the thickness of the optical section from which fluorescent light is collected. More details can be found in [89, 85, 81]. Finally the image is captured by a point detector (a photodiode, a photomultiplier) or —as in our case — a digital charge coupled device (CCD) camera.

TIRF measurements make use of the evanescent field generated upon total internal reflection. Total internal reflection occurs at the interface between two media: a higher refractive index medium n_1 (the glass or plastic coverslip) and a lower refractive index one, $n_2, n_2 < n_1$ (air, water). If the angle of the incident

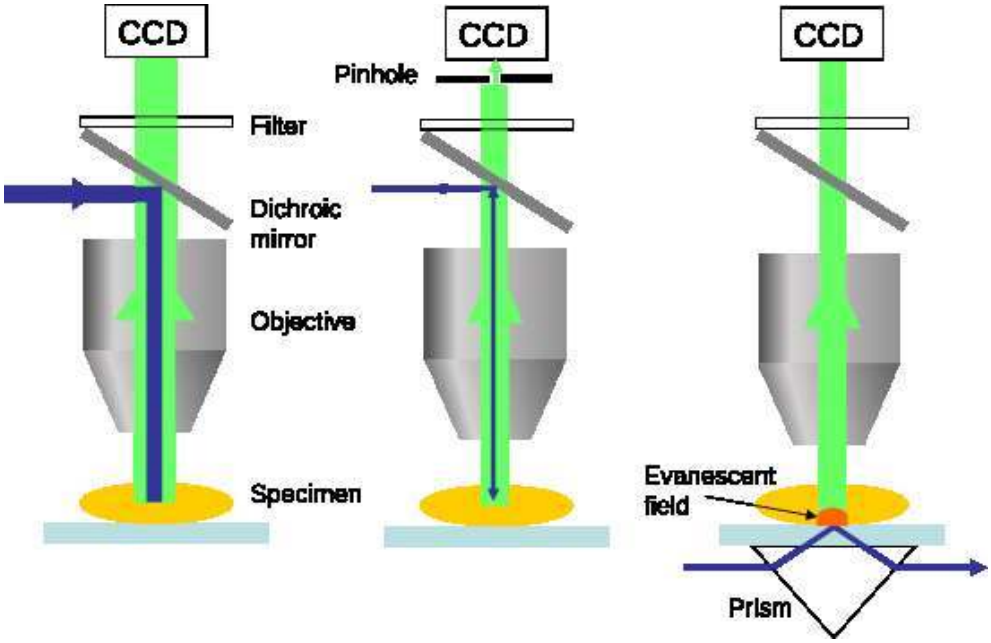


Figure 2.3: Microscope configurations: left - epifluorescence, middle – confocal, right - TIRF

beam with the interface normal is higher than the critical value $\arcsin(n_2/n_1)$ the light is entirely reflected and it no longer passes into the second medium. Nevertheless the reflected light generates an evanescent electromagnetic field in the lower-index medium, in the close proximity of the interface. This evanescent field is identical in frequency to the incident light, and decays exponentially in intensity with distance z from the interface:

$$I(z) = I(0)\exp(-z/d),\ d = \frac{\lambda_0}{2\pi}(n_1^2\sin^2(\theta) - n_2^2)^{-1/2},$$

where the decay distance d is dependent on the wavelength of the excitation light λ_0, the refractive indices n_1, n_2 and the angle of incidence. Thus the field extends at most 100 nanometers normal to the interface into the specimen and only the fluorophores in this limited volume are excited. In this way the background can be kept low and a good contrast is achieved for single molecule imaging.

2.3 Image formation

The dual nature of light, as wave on one hand and as quanta (photons) of electromagnetic radiation on the other can be exploited in order to describe image formation as a linear shift invariant system and model the various sources of noise

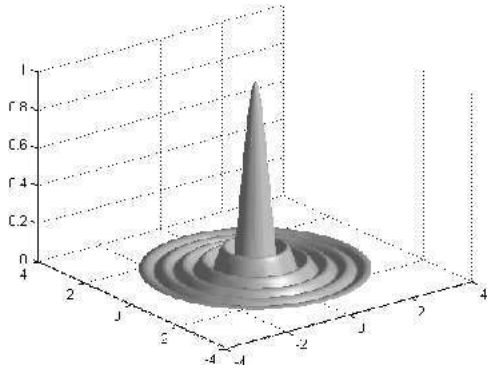


Figure 2.4: The Airy function - image of a point source

that are corrupting this process.

Point spread function (PSF) and spatial resolution

Making use of the wave description of light, the Fraunhofer diffraction caused by a circular aperture of a point source on the optical axis produces an intensity distribution centered at $r = 0$ also known as *Airy pattern* ([117]):

$$\mathrm{PSF}(r) = \left(2\frac{J_1(\pi q_c r)}{\pi q_c r}\right),$$

with J_1 the Bessel function of the first kind of order 1 and $q_c = \frac{2NA}{\lambda}$, such that the shape of the Airy pattern depends on the wavelength of light λ and the numerical aperture NA of the objective lens. The smaller the wavelength and the higher the numerical aperture the smaller the Airy disk with direct implications on the system's resolution properties as we shall see below.

In practice, most often the PSF is approximated by the following functions [58]: Gaussian:

$$G(r) = e^{-\left(\frac{r^2}{2a^2}\right)} \tag{2.3.1}$$

modified Lorentzian:

$$L(r) = \frac{1}{1 + \left(\frac{r^2}{a^2}\right)^b} \tag{2.3.2}$$

or a Moffat function:

$$M(r) = \frac{1}{\left(1 + \frac{r^2}{a^2}\right)^b} \tag{2.3.3}$$

In biological experiments, the PSF depends on the specific refractive properties of the sample [31]. An approximated PSF can be constructed by imaging small beads

or in case this proves too difficult/tedious for the given biological experiment, by extracting the images of small features. A similar procedure is used in astronomy, where stars represent natural point like sources, which can be used to model the system's PSF.

The diffraction phenomenon affects the (spatial or lateral) resolution of an optical system. *Spatial resolution* describes the smallest resolvable distance between two points in an image. There exist several ways to define resolution.

The *Abbe resolution* is given by the FWHM (full width at half maximum) of the Airy disk:

$$\Delta_A \approx \frac{\lambda}{2\mathrm{NA}}.$$

The *Rayleigh limit* (resolution) is the distance between the central maximum and the first minimum of the intensity of the Airy pattern generated by the optical system:

$$\Delta_A \approx 0.61\frac{\lambda}{\mathrm{NA}}.$$

The *Sparrow limit* is defined as the minimum distance of two point objects of equal intensity so that no intensity minimum exists between both images:

$$\Delta_A \approx 0.48\frac{\lambda}{\mathrm{NA}}.$$

Noise sources and signal-to-noise ratio (SNR)

All measurements have an inherent uncertainty referred to as *noise*. A quantification of how much this uncertainty affects the measurement is given by the signal-to-noise ratio (SNR). A high SNR indicates high confidence in the measured value. According to their source different types of noise can be classified as:

- *shot-noise*- the number of photons recorded by the camera over a discrete interval of time can be described due to the quantum nature of light as a stochastic process with a Poisson distribution:

$$N \sim \mathcal{P}oi(\mu), \quad p(N) = \frac{\mu^N \exp(-\mu)}{N!},$$

  where N is the number of detected photons and μ is the expected value of the Poisson process. This kind of noise can not be suppressed.

- *thermal noise or dark current* - thermal electrons generated by the kinetic vibrations of silicon atoms in the CCD that are mistaken for photoelectrons.

The number of thermal electrons follows a Poisson law. The higher the temperature, the more important the contribution of thermal noise. Cooled CCD cameras are used in order to suppress dark current.

- *readout noise* - noise added during readout of charges by the camera chip electronics and it is characterized by a Gaussian (normal) distribution $\mathcal{N}(\mu, \sigma)$, with probability density function (pdf)

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right].$$

(Note that a standard normal distribution has $\mu = 0$ and $\sigma = 1$.) The amount of noise depends on the readout rate. The higher the readout rate the higher the readout noise level due to the on-chip electronics.

- *quantization noise* - the round off error due to the ADC when converting analogue data to integer numbers. It depends on the number of bits used for the digital representation of the data.

- *dead/defect pixels* - pixels with a constant white or black value.

An image is considered to be *photon limited* if the photon noise of the object signal is greater than the camera read noise.

In order to measure the effect of the noise on the quality of signal the notion of signal-to-noise (SNR) ratio is introduced. SNR is the ratio between the sum of components contributing to the signal and the square root of the sums of the variances of the various noise components. If the different types of noise are independent the SNR can be written as:

$$SNR = \frac{S_1 + S_2 + S_3 + \ldots}{\sqrt{N_1^2 + N_2^2 + N_3^2 + \ldots}}.$$

In microscopy, the background signal can be large, sometimes 90% of the total signal representing an object. In most cases, photon noise from the background is the major source of noise, not the read noise of the camera. Thus, the SNR equation includes a term for the background noise. The signal S is equal to the difference between the counts in the target area (T) and those in the background area (B): $S = T - B$. The contrast is defined via normalization with B:

$$C = \frac{T - B}{B},$$

such that the signal in terms of contrast becomes: $S = C \cdot B$. Due to the Poisson statistic model of the image, for the standard deviation of the noise we get $N = \sqrt{B}$, and since T and B have similar values for low contrast, the SNR can be written as:

$$SNR = \frac{C \cdot B}{\sqrt{B}} = C\sqrt{B},$$

or alternatively:

$$SNR = C\sqrt{\Phi \cdot A},$$

where A is the area of the region of interest (the size of the smallest object we wish to detect) and Φ is the photon flux (photons per unit area).

A more complex analytical equation for SNR was given by Newberry [86]:

$$SNR = \frac{\sqrt{C_o}}{\sqrt{[(1/g) + (n\sigma^2/C_o) + (n\sigma^2/pC_o)]}}$$

where

$C_o = T - B$ counts due to the object

n - number of pixels in measured object area

p - number of pixels in measured background area

σ^2 - variance of background pixels

g - gain (the number of electrons recorded by the CCD camera per number of digital units contained in the image).

Usual ways to improve SNR are to increase the amount of light by reducing the scan rate, increasing the recording time or opening the confocal pinhole, by averaging several frames of the same imaged object etc.

Models of image formation

The optical system is locally shift invariant, thus a microscope can be well approximated as a linear and shift-invariant (LSI) system [121], so that each point source in the object plane f is replaced by a scaled and translated PSF in the image plane (Huygens principle)

$$g(x, y) = f(x, y) \otimes \mathrm{PSF} = \iint_{\Omega} \mathrm{PSF}(u, v) f(x - u, y - v)\, du\, dv.$$

A general microscopy image model together with algorithms for image restoration is given in [104]. Imposing restrictions on the image content, Giovanelli and Coulais [47] describe a model consisting of the superposition of two components: one component (PS) is formed by point sources on a dark background, while the other component is formed of spatially extended, smooth objects, called extended sources (ES). These components are treated as two distinct maps, that have to be accurately reconstructed. For the ES component the correlation structure is introduced by a convolution kernel or pixel interactive penalties.

A similar description fits many single molecule microscopy images. The content of images, a sparse set of point-like objects, is modeled in the object space as a sum of delta peaks:

$$f(x,y) = \sum_{k=1}^{N} A_k \delta(x - x_k, y - y_k).$$

As a result of image formation process one obtains in the ideal image space

$$g(x,y) = K * f(x,y) = \sum_{k=1}^{N} A_k K(x - x_k, y - y_k)$$

describing the PS component of the image. The kernel K corresponds to the ideal PSF or any of its approximations like the ones offered by the G, L, M models. Due to the various sources of noise, the measurement can be written as

$$\tilde{g}(x,y) = \alpha\,(b + \lambda) + \varepsilon, \tag{2.3.4}$$

$$\lambda \sim \mathcal{P}oi(g(x,y)),\ \varepsilon \sim \mathcal{N}(0,\sigma), \tag{2.3.5}$$

with b the ES component or background and α the gain of the system.

Chapter 3

Microarrays with single molecule sensitivity

Cells exposed to toxins, pharmacologic agents, human hormones etc. respond to these changes by changes in the expression of particular genes [80]. Change in gene expressions can occur also in normal cellular activity, e.g. cell division. Thus the gene expression levels are highly informative about the cell state, the activity of genes as well as changes in protein abundance in the cell.

The complete set of DNA transcripts and their relative levels of expression pertaining to a cell or tissue as well as a specific condition is called the *transcriptome*. As opposed to the genome, it is highly dynamic, changing rapidly in response to environmental conditions or during certain cellular events [75].

Microarray technology offers an advanced and efficient way to measure gene expression of large sets of different genes at a high throughput. As a consequence, it is an important tool in solving problems such as the detection of differently expressed genes for different conditions, the annotation of gene function — in case the encoded protein's function is unknown, shared regulation patterns with genes whose function is known offers valuable cues on the function of interest—, definition of genetic pathways (the regulation of gene expression by other genes) via temporal profiling, molecular phenotyping for prediction of pharmacological response or evolution of a disease, etc. (see [108, 75]).

3.1 Biological background

Chromosomes are structures found in the cell nucleus, formed by a single molecule of coiled deoxyribonucleic acid (DNA) carrying the individual's hereditary mate-

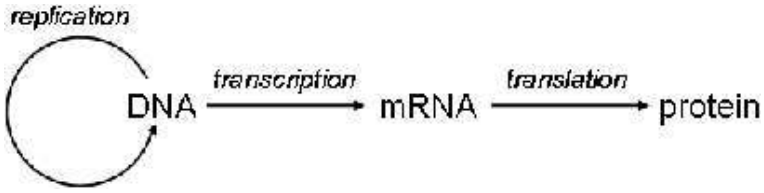


Figure 3.1: Central dogma of molecular biology

rial, and DNA-bound proteins. The DNA molecule is a double-stranded helix, each strand formed of linkages of sugar-phosphate. The strands are bound together by the noncovalent hydrogen bonding between pairs of attached bases.

The basic units of biological inheritance are the *genes*, specific segments of a DNA molecule that contain all coding information for the cell to synthesize a specific product, as an RNA molecule or a protein. They occupy a specific location (locus) on a chromosome and are identified according to their function. Some of the genes, the so-called *housekeeping genes*, encode proteins needed for basic cellular activity, and thus are expressed in all cells. Others have specific tasks, as coding the synthesis of specific antibodies or limit the formation of malignant cells (antioncongene). More details can be found in [80, 69, 116, 33].

The genetic code is communicated from DNA to RNA via *transcription*, that is the synthesis of an RNA strand of complementary bases to the DNA strand. The production of proteins is then guided by the RNA transcribed from the DNA, called *messenger RNA* or *mRNA*). The transcription occurs in the nucleus. The mRNA is transported into the cytoplasm, where the ribosomes read the mRNA sequence and *translate* it into the amino acid sequence of the produced protein. Schematically, the process is presented in Fig. 3.1.

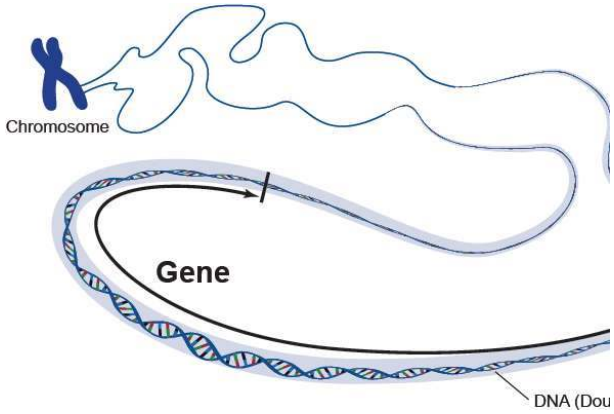


Figure 3.2: Genes, the units of biological inheritance. Figure from [2].

A typical microarray experiment measures (and compares) the mRNA abundance (*expression levels*) of the cell samples in order to infer the type and the function of the proteins produced under certain conditions in the cell. However the mRNA molecule is fragile and can be easily broken down by enzymes from biological solutions. Instead the more stable *complementary DNA* (cDNA) is created from the mRNA sample, through *reverse transcription* and subsequently used in the experiment.

3.2 Classical microarray technology

The high density DNA microarray allows the monitoring via single experiments and on a single experimental medium the interactions among thousand of gene transcripts in an organism. Good overviews of the technique can be found in [119, 69, 80, 103].

Although the first immunoassay technologies were developed already in the 1950s and 1960s, the *Southern blot* developed in 1975 was the first array of genetic material [105]. The miniaturization of the technique, microspotting in the 1980s [40] and the subsequent technological improvements (mechanization of the construction of slides and of microspotting) lead to the industrialization of the technology.

The technology is based on the specific pattern of bonding known as (Watson-Crick) *base pairing*: the base known as adenine (A) specifically bonds with thymine (T) while cytosine (C) specifically bonds with guanine (G). Note that in the case of RNA, adenine bonds to uracil (U). The amine base that will form a bonding pair with another amine base is considered its complementary base, and single DNA or RNA (ribonucleic acid) strands form stable bonds or *hybridize* only with a complementary strand. Hybridization is the fundamental process that constitutes the basis of DNA microarray technology [69]. The hydrogen bonding between bases is weak and breaks at approximately 90°C, through a process called *denaturation*. After cooling, at about 60°C *reassociation* occurs.

Measurement process

Briefly, the experiment consists of a robotic machine laying *spots* or droplets of probes, representing e.g. cDNA sequences, on a glass, silicon or plastic slide in a regular pattern forming a $2d$ grid. The cDNA is immobilized on the array by adherence to the slide coating, air drying and ultraviolet irradiation [80]. The slide

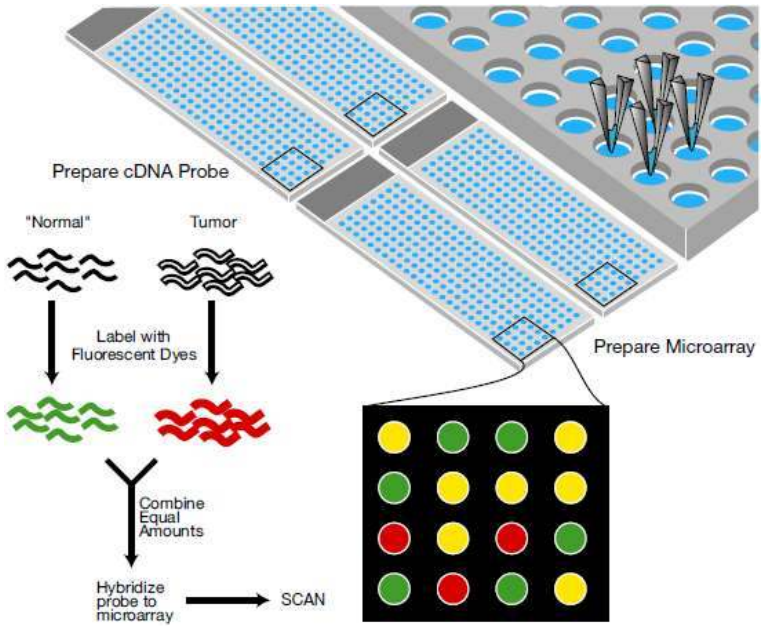


Figure 3.3: Schematic representation of two-color microarray technology. Figure from [2].

with the affixed spots is called *microarray* or *chip* and it represents a dictionary of probes, in which an entry (DNA, oligonucleotide, antibody etc.) is identified by its position in the grid. The process is represented in the upper right part of Fig. 3.3. Via denaturation, double stranded cDNA is broken into single strands preparing for the binding of target samples. For the sake of statistical soundness, each probe on the chip is replicated a certain number of times subject to a trade-off between cost and reliability of results.

Next, the target sample is marked with a fluorescent tag during reverse transcription and washed over the chip. In the case of two different samples, a control sample and an unknown target sample, (e.g. a control from normal cells and a sample from cancerous cells), the samples are marked with distinguishable tags, emitting at different wavelengths (typically red and green, Cyanine 5, Cy5 with emission maximum at about 670nm, and Cyanine 3, Cy3 with maximum approximately at 570nm), then equal amounts of the samples are mixed and finally washed over the microarray slide, where molecular hybridization takes place. If necessary, the quantity of DNA can be amplified via a process called *polymerase chain reaction* (PCR). The amount of bound sample to the probes is an indicator of the strength of interaction intensity with the respective probe.

This hybridization is measured by scanning the microarray after excitation with laser light. The scanning is performed usually with a confocal laser microscope

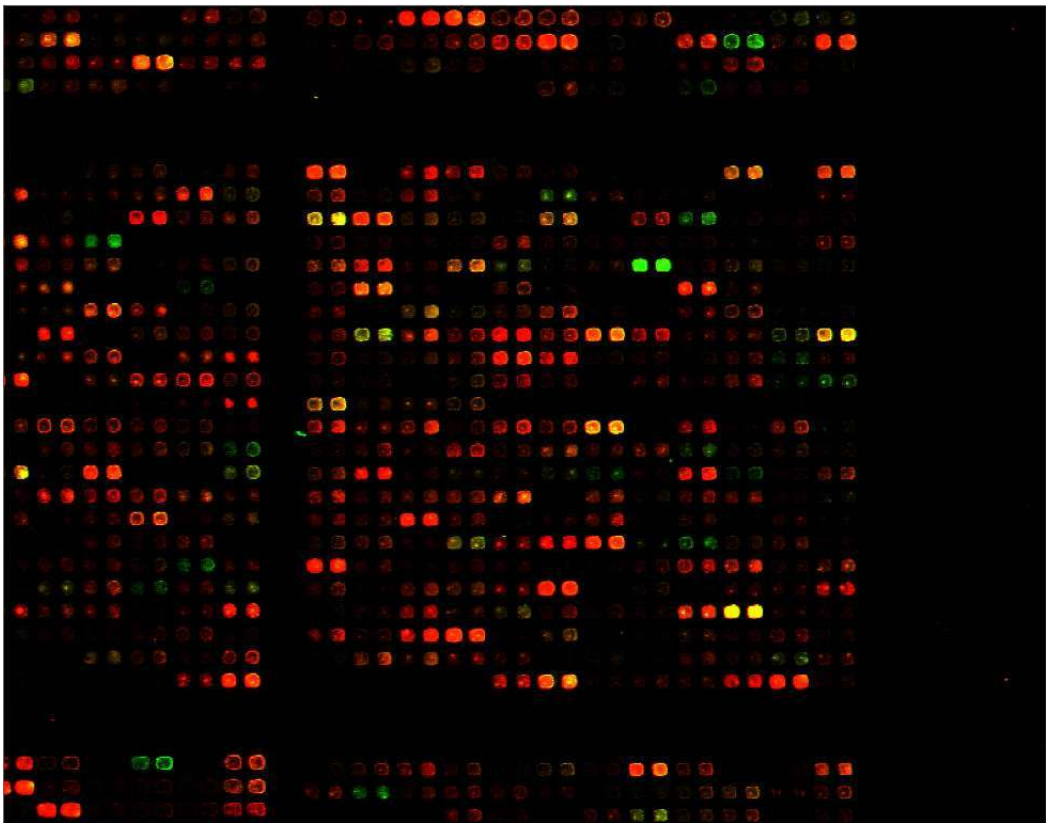


Figure 3.4: Detail of microarray. Pseudocolors indicate: red - high expression in target labeled with Cy5, green - high expression in target labeled with Cy3, yellow - similar expression in both samples. The gene is identified by the position of the respective spot in the grid.

and the fluorescent intensity for each probe spot (and in each color) is recorded. The output of the last step are 16 bit images, a pseudocolored detail of which can be seen in Fig. 3.4. The images have to be analyzed by robust image processing tools to produce reliable intensity estimates that describe the hybridization.

Data analysis

According to [69], the main tasks image processing has to perform are:

Gridding Matching a grid to the image of the printed spot pattern and identifying the approximate position of each spot

Segmentation Separation of foreground(signal) and background pixels for each spot region. If shape priors are included in the process, the simplest assumption is a circular shape with fixed or variable radius.

Intensity extraction Computation of a summary statistic that best describes intensity of the spot (e.g. mean, median etc.)

Background correction If necessary, the signal intensity is corrected with the estimate of the local or global background (most often an additive background model is assumed). The effect of this step might be crucial in the case of weakly expressed genes (dim spots).

The result of the image processing step is a gene expression matrix of (mean) fluorescence intensity of each spot (or of the ratio of the intensities recorded in the two channels). *Normalization* and *data interpretations* are the last steps of the analysis that leads to the gain of biological insight on the role the reporters play in the problem at hand. Normalization adjusts microarray data for effects which arise from variation in the technology (systematic errors) rather than from biological differences between the mRNA samples. It is a sequence of sophisticated statistical methods that takes into account several factors such as unequal quantities of starting RNA, differences in dye incorporation during labeling or detection efficiencies between the fluorescent dyes used, variations in spotting and/or background, and systematic biases in the measured expression levels (see [41, 120, 39, 16, 33] for details). Normalization can be performed on several levels: within a single array (among replicate spots), between a pair of replicate arrays and among multiple arrays (over samples to be compared).

The last step, data interpretation is based on a plethora of tools from statistical hypothesis testing, to clustering, cluster analysis, data mining and it tries to answer biological questions such as the detection of differently expressed genes (for two sample experiments), the detection of co-regulated genes, characterization of expression profiles etc.

Limitations of microarray technology

Each step of the technique is prone to distortions and noise contamination. A thorough description of variation and noise sources can be found in [10, 80, 71]. Most of these factors have to be dealt with by the algorithms used for the analysis.

Results are influenced by several factors related to the surface of the array, the binding efficiency, the typical number of dye attached to each cDNA molecule, differences in PCR amplification, etc. For instance, *reverse transcription bias* occurs due to variation in the degree of efficiency corresponding to different types

of mRNA molecules, while *sequence bias* is due to varying binding affinity of fluorescent dyes to different nucleotides (e.g. cDNA strands containing more guanine appear brighter, due to better binding of the dye to guanine). The differences in labeling are called *dye-bias*, and can be studied and corrected for via a repeat of the experiment on the same targets, but with changed labeling, procedure known in the literature as dye-swap.

One of the most important limitations of microarray technology is the availability of tissue samples in sufficient quantity [80]. The problem is particularly acute in cancer cell experiments, when the amount of mRNA that can be extracted from rare cells is inadequate to perform a microarray experiment.

Moreover, the technique represents an indirect measurement of the abundance of gene transcripts via the fluorescent dyes attached to the hybridized polynucleotides. Fluorescence itself is not a linear phenomena (it is linear only over a limited range). Thus, this method renders only a semi-quantitative measure of differences in gene expression with small differences being obscured within the experimental variation of the array technology.

A bias might be introduced by less-than-perfect denaturation process, resulting in non-denatured strands of (spotted) DNA, as well as, target molecules that stick to the slide and are not washed away, and subsequently contributing to the background noise around the spot. Furthermore, experimental conditions cannot be optimized for all the genes. Due to the fact that under the chosen conditions hybridization simply did not happen for certain genes, the measured expressions in these cases are misleading, since they do not reflect the real biological situation.

The complexity of the resulting data makes data storage, retrieval and sharing difficult, a problem further increased by the lack of standardization among the different existing systems.

Types of arrays

The best known types of arrays are the cDNA and the oligonucleotide arrays. Besides the single-slide cDNA microarray experiments described above, in which one compares transcript abundance in two mRNA samples hybridized to the same slide, there are other approaches such as multiple-slide experiments comparing transcript abundance in two or more types of mRNA samples hybridized to different slides and a variant of this, the time-course experiments, in which transcript abundance is monitored over time for processes such as the cell cycle. The technology is based on clones obtained from cDNA libraries spotted on the support, typically

formed by strands of 500 to 5000 bases of known sequence [80]. The design of the array (the selection of cloned cDNA) can be optimized for the (hypothesis testing) problem at hand.

Oligonucleotides are short sequences of base-pair segments, having a length between 15 and 70 nucleotides. They can be used as probing material on the array (one of the best known chip is the GeneChip produced by Affymetrix). The advantages of this kind of arrays are specificity and efficiency of hybridization (due to uniform length), they are also easier to engineer and easier to use for finding optimal hybridization conditions. The disadvantages of the oligo arrays are cross-hybridization with several genes, due to the short lengths of the oligonucleotides — this can be seen as a strength when the purpose is discovery and not predefined sequence matching— the absence of purification processes, irregularities in the fluorescence signal, etc.

Other types of microarrays involve chromatin immunoprecipitation assays (ChIP-chip)[118] or protein microarrays [114, 51].

3.3 The high resolution technique

A main objectives of our project was to correct some of the limitations described in Section 3.2, such as unspecific binding (due to randomly bound molecules or unattached fluorophores), the varying background intensity profile of the array, the binding efficiency of the sample (which might be gene dependent), the dye distribution per molecule (also gene dependent), varying illumination distortion etc.

Our technology is based on the combination of the classical technology with ultrasensitive fluorescence microscopy for reading the specially designed DNA chip.

Some of the advantages of scanning spots with ultrasensitive microscopy capable of measuring the signal of single dyes equivalent to single hybridization events are illustrated in Fig. 3.5. In the low resolution image on the left, only the intensity can be observed, without any further visual clue if it is due to true signal or artifacts. On the right, in the high resolution image, one can clearly distinguish the bright artifacts (probably dirt) around the microarray spot from the true signal. Artifacts that might distort the low-resolution signal intensity are identified using ultrasensitive fluorescent microscopy.

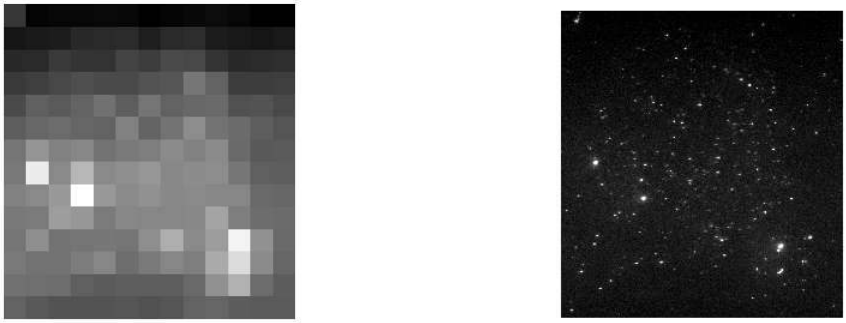


Figure 3.5: Image of a microarray spot with artifacts. (Left) Low resolution image, with high intensity pixels (without visual clues for presence of artifacts) (Right) High resolution image, where the artifact can be distinguished from signal. The image intensity was rescaled for better visibility.

Imaging setup

The novel technique introduced in [56] increases tremendously the resolution of the scanning, it represents a 20 times zoom in the classical microarray. To one pixel imaged in the classical way correspond 400 pixels with the new technique. At this resolution it is possible to detect and count single molecules. Given the size of the fluorophore, the image of a molecule is equivalent to the point spread function of the optical system applied to a point source as described in Chapter 2.

The imaging setup, NanoScout (developed by the SDT group, Institute of Biophysics, Johannes Kepler University and Upper Austrian Research), presented in Fig. 3.6, is based on a modified epi-fluorescence microscope with a 100× oil im-

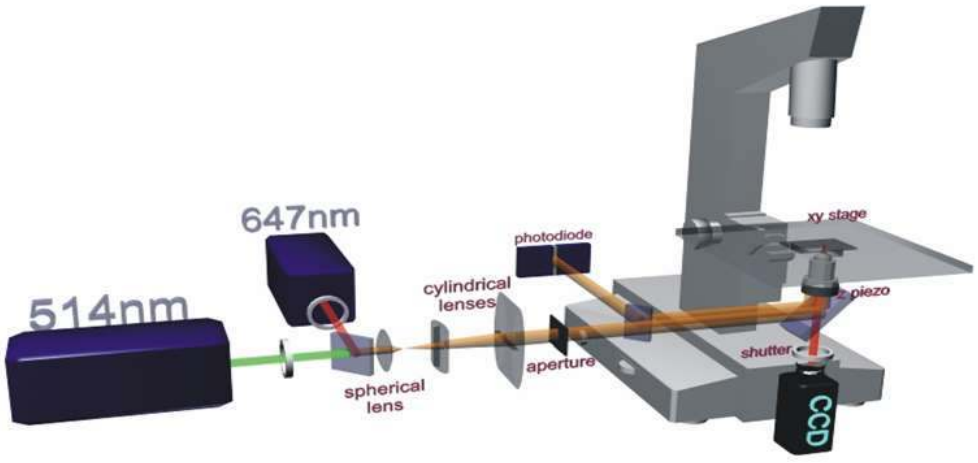


Figure 3.6: Nanoscout - the high resolution setup used in microarray imaging

mersion objective (α-Plan Fluar, $NA = 1.45$, Zeiss). The samples are illuminated in objective-type total internal reflection (TIR) configuration. For the selective excitation of the dyes, Cy3 and Cy5, respectively, Ar^+- and Kr^+- lasers (514nm and 647nm) are used. After appropriate filtering using standard Cy3 and Cy5 filter sets (Chroma Technology Corp., VT), the fluorescence images are taken with a 12-bit back-illuminated CCD camera (chip-size 1300×100 pixel, $20\mu m$ pixel-size). The CCD camera is operated in time delay and integration (TDI)-mode. The samples are mounted on a motorized xy-stage and synchronized to the line-shift of the camera. This allows fast scanning of large areas ($1 \times 0.02\text{cm}^2$ in 58s) with single fluorophore sensitivity and diffraction limited resolution. The setup and the imaging process are described in detail in [56, 55, 61].

For validation of the system, microarrays comprising two full complementary 60mer oligonucleotides were used (Human Genome Oligo Set Version 3 (Operon)). Hybridization of a dilution series of Cy5-labeled target oligonucleotides yielded a quantification limit of 1.3fM corresponding to only 39.000 molecules in $50\mu l$ sample.

The ultra-sensitive microarray technique was used to study the expression profile of putative Multiple Myeloma (MM) stem cells using the human MM cell line NCI-H929 and the results are provided in Chapter 7. This new technology brings important insights in the field of biochips, with several advantages, like the analysis of images with very low sample concentrations, a new way of background suppression, and more refinement in information.

Main steps in single molecule microarray image analysis

The high resolution microarray image analysis preserves the same main tasks as the low resolution technique:

- Addressing/ Gridding - Localization of each spot of the grid pattern in the image
- Estimation of the hybridization measure for each spot via spot identification and concentration estimation
- Gene expression analysis.

However there is a need to redesign and/or adapt the existing algorithms in order to be applicable to the new kind of data. Gridding is performed on a down-sampled

version of the original image, while for the other tasks new algorithms are designed, implemented and validated.

Gridding is performed by registering a predefined grid pattern with the image of detected microarray spots. There is a plethora of methods to find the location (and shape) of bright microarray spots. Among these we mention the method of Angulo and Serra based on mathematical morphology [7], clustering of the pixels into foreground, background and artifact pixels [14], watershed based algorithms etc. Some of these approaches also adjust the shape and position of spots in a subsequent step after an initial grid was found.

However, in our case due to the low amount of mRNA the spots appear much dimmer than in the classical microarray images, making both tasks, spot detection and grid estimation, more difficult. For spot detection we have used the same wavelet thresholding algorithms as for single molecule detection, in this case applied to a downscaled, low resolution image. All the details are described in Chapter 5. The combination of the resulting images from the two samples (corresponding to the red and green channel) improve the registration result. As the result of the gridding step, we assume that the detected rectangular region includes a single microarray spot, potentially surrounded by background pixels (as in Fig. 3.7(a)).

The steps related to the analysis of a single molecule microarray spot are illustrated by Fig. 3.7, where the results of each step are shown for the original spot image presented in Fig. 3.7(a):

1. Detection of the support of single molecule signals, Fig. 3.7(b)
2. Identification of single peaks inside the detected signal support, Fig. 3.7(c)
3. Separation of specifically bound molecules from background clutter, Fig. 3.7(d).

The estimation of hybridization measure is usually complemented with information regarding the local background for each spot as well as various spot quality measures (homogeneity, circularity, etc.).

Each of these steps will be discussed in detail in the following chapters, the detection of single molecules, based on wavelet thresholding in Chapters 4 and 5, while the specific signal detection and concentration estimation in Chapter 6.

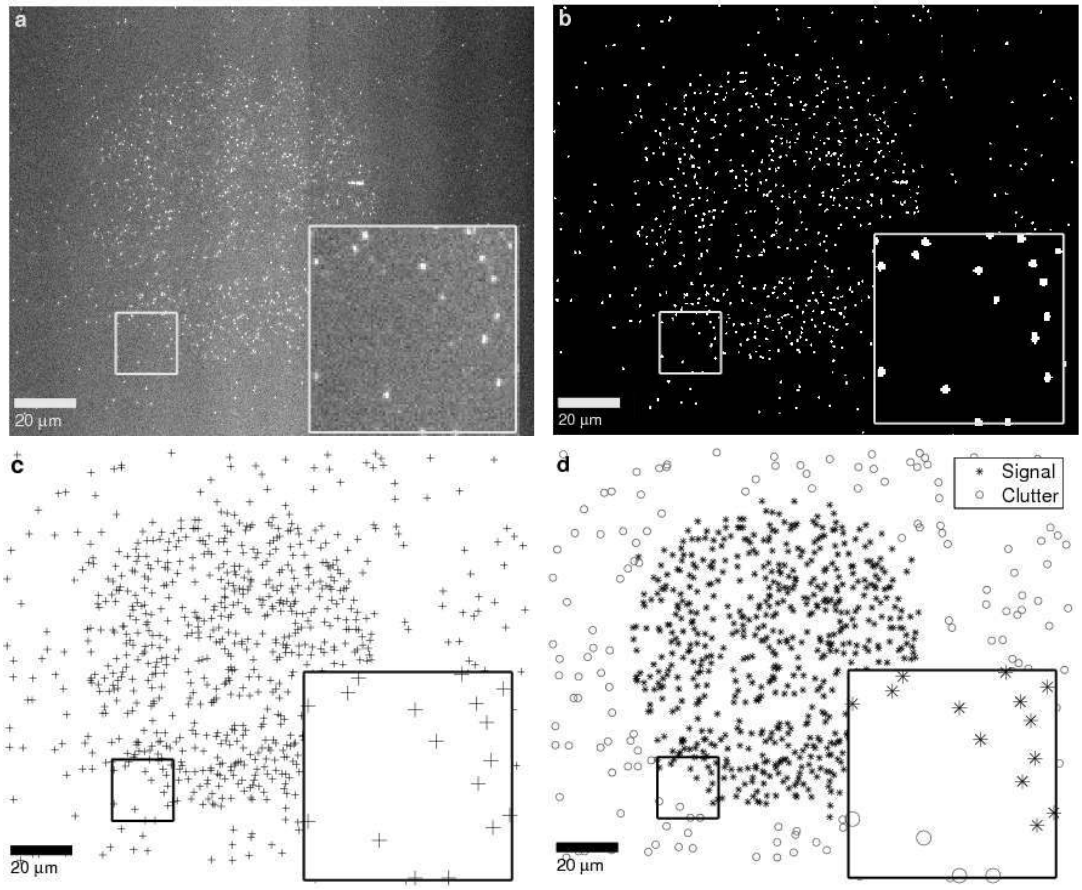


Figure 3.7: Analysis of a spot in a high-resolution microarray image. (a) Original image, bright features correspond to molecules bound to the chip. (b) Detection of single molecules. (c) Selection of single molecule locations (local maxima on denoised image inside the detection support in (b)). (d) Separation of hybridization signal from clutter.

Mathematical model

The differences between the low resolution and high resolution microarray experiments might be best understood by describing an underlying model of fluorescence intensity for each of them.

Our aim is to propose a new measure for hybridization: instead of aggregations of the pixel intensities inside the spot we use as hybridization measure single molecule counts (or concentration of hybridized molecules per area unit). The model we propose as well as the new hybridization measure suggest that the high resolution technique has the following advantages: besides offering a way to analyze very low concentration samples, it removes bias due to background heterogeneity and PCR amplification, making several normalization steps and dye swap (all prone to distortions and errors) unnecessary.

So far, for the classical case of low resolution imaging, the analysis is based on the comparison of appropriately chosen summary statistics of pixels inside the spot. Conventional analysis is based on models of the microarray signal formation, like the ones proposed by [10, 6, 23, 70, 74]. These models include several aspects of the acquired data such as image intensity, spot shapes, noise.

A very simple but frequently used spot intensity model proposed in [6] can be written

$$Y_i = \mu_i \cdot s(i - i_c) + \varepsilon_i,$$

where Y_i is the intensity of the spot at pixel i, μ_i represents the amplitude of the spot as a measure of the hybridization, s is a function describing the shape and texture of the spot, while i_c represents the center of the spot, and ε_i models the local and/or global background noise, usually $\varepsilon_i \sim \mathcal{N}(0, \sigma)$.

A more realistic model, taking into account the sensor properties, includes both additive and multiplicative noise [71]:

$$Y_i = \alpha \mu_i e^{\eta_i} + \varepsilon_i,$$

where α is the gain of the sensor, e^{η_i} and ε_i representing the multiplicative and additive noise, respectively.

Often a lognormal model of pixel values Y_i is assumed. If Y is replaced by G or R for intensities in the green or red channel respectively, and BR_i and BG_i are the respective background estimates used for background correction, the final quantity of interest is:

$$\log(G_i - BG_i) - \log(R_i - BR_i). \tag{3.3.1}$$

The log-transform plays also a variance stabilizing role.

However, gaining access to microarray spot images at single molecule resolution, we understand better the image formation process in both high and low resolutions. For modeling we shall adopt an approach close to the concentration modeling described in [23].

For the classical low-resolution microarray image analysis we propose a compound Poisson process model explained below. The discussion of the high resolution model will be detailed in Chapter 6. We first give the definition of the compound Poisson process and than see how it helps in modeling the microarray pixel intensities.

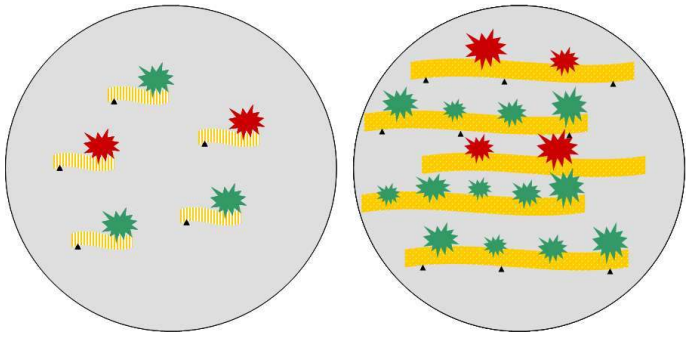


Figure 3.8: Schematically represented microarray spots. Although the red/green ratio is the same in both spots, the intensity ratio in the two channels will differ

Definition 3.3.1. Given a rate $\lambda > 0$ and an arbitrary distribution Q, the random sum

$$Z = \sum_{k=1}^{N} X_k,$$

is distributed according to $\mathcal{CP}(\lambda, Q)$ (*compound Poisson process*), where $N \sim \mathcal{P}oi(\lambda)$ and X_i are independent and identically distributed random variables with distribution Q (*jump distribution*), independent of N.
The tail behaviour of $\mathcal{CP}(\lambda, Q)$ is inherited from the distribution Q, the expectation and variance are given by: $E(Z) = \lambda E(Q)$ and $\mathrm{Var}(Z) = \lambda E(Q^2)$.

Model 1

The intensity of pixel i in the low-resolution microarray spot image is obtained by integrating the fluorescent intensity due to hybridized molecules, unspecific binding, background and artifacts over the area corresponding to one pixel and it can be written as:

$$Y_i = B_i + Z_i = B_i + \sum_{k=1}^{N_i} D_k, \tag{3.3.2}$$

where B_i represents background fluctuation, Z_i is a compound Poisson process such that N_i represents the number of single molecules inside the low resolution pixel area i and D_k are the intensities of single molecules in the same area, $D_k \sim \mathcal{P}oi(\mu)$.

Thus each pixel in the low resolution spot image can be considered a realization of a random variable distributed according to a compound Poisson model $\mathcal{CP}(\lambda, \mu)$. Adding the effect of thermal noise and the gain of the optical system

(as in Section 2.3), a pixel can be modeled as

$$\widetilde{g}_i(x,y) = \alpha\left(Y_i\right) + \varepsilon = \alpha\left(B_i + \sum_{k=1}^{N_i} D_k\right) + \varepsilon_i,$$
$$B_i \sim \mathcal{P}oi(B),\ \varepsilon \sim \mathcal{N}(0,\sigma),$$

Model 2

A more complex model takes into account the variation of the number of fluorophores, M_k, bound to each molecule. In this case the the jump distribution D follows itself a compound Poisson distribution, $\mathcal{CP}(\lambda_1, \mathcal{CP}(\lambda_2, \mu))$. If M_k is the number of dyes bound to a molecule, $M_k \sim \mathcal{P}oi(\lambda_2)$, and D_{jk} is the intensity of a single dye, $D_{jk} \sim \mathcal{P}oi(\mu)$, the model can be written as:

$$Y_i = B_i + Z_i = B_i + \sum_{k=1}^{N_i}\sum_{j=1}^{M_k} D_{jk}. \tag{3.3.3}$$

The compound Poisson process model has applications in gene expression analysis ([73, 113]), but to our knowledge has not been used in the case of microarray image analysis. The compound Poisson model is useful in illustrating the difference between the classical low resolution microarray analysis and the high resolution one.

The technology of high resolution microarrays offers access to previously hidden information: instead of analyzing the low-resolution pixel values Y_i, the inference is based on N_i the number of single molecules in the image. The measure of hybridization is in our case the concentration of single molecules inside the spot of interest. The knowledge of the background values of B_i and photon counts D_k has little or no relevance, being only nuisance parameters.

In microarray analysis the identification of the true signal and the control of the unspecific intensity variation is essential. We have identified the following shortcomings of the low resolution technique, which make this analysis less appropriate than the single molecule one:

1. In case of low concentrations, many spots are rejected from analysis due to low signal-to-noise (SNR) ratio and artifacts, especially in the case of dim spots. Low resolution microarrays cannot discriminate between signal and background. In the high resolution technique, single molecules are as easily detected in low concentration (dim spots) as in high concentration ones,

allowing the analysis of trace amount of biological sample.

2. Even when segmentation is possible, background estimation is difficult in low resolution microarray images and might introduce a bias. In the high resolution case, background subtraction becomes irrelevant, it is implicitly achieved by the single molecule detection procedure.

3. It was shown that the labeling efficiency is gene dependent (see, e.g., [55]) and it can distort the results in the case of low-resolution microarrays. In the high resolution case, the hybridization measure is represented by counts of single molecules and it is not affected by the variability of the number of fluorophores bound the a molecule. Dye swap normalization typically performed for classical arrays in order to compensate for dye related differences also becomes unnecessary.

4. We note that although normally only the count information is relevant in the case of the high resolution technique, still it is possible to use the intensity corresponding to a single molecule, for instance to discriminate it from dirt. Dirt particles show autofluorescence, with a significantly lower (but sometimes not negligible) intensity than the dye marked molecules. The high resolution technique allows discrimination between two or several entities based on total intensity of a single molecule but the same cannot be performed in low resolution images, since the information is not available at this granularity .

Some of the challenges of the new method are related to the huge amount of data that has to be processed (22 GB per slide), the scarcity of the highly expressed spots necessary to perform the gridding, the high variance of the Poisson processes modeling the counting of molecules. Also, in case of high density of the molecules in a spot, when single molecules cannot be discriminated, classical intensity summary based methods have to be used instead of single molecule counting algorithms.

Chapter 4

Multiscale signal decomposition

This chapter gives a brief introduction to wavelets, both the continuous (CWT) and discrete wavelet transform (DWT), as well as to multiresolution analysis, orthogonal wavelet bases and their generalization, the wavelet frames. We focus on particular discretizations of the CWT, the dyadic wavelet transform,the undecimated wavelet transform and the isotropic undecimated wavelet transform (for the two dimensional case), and on a particular wavelet shape, the B-spline wavelet, especially fitted to the analysis of biological images. The isotropic undecimated wavelet transform is the representation of the signal that will be used in the next chapter to detect single molecules.

4.1 Continuous wavelet transform

The choice of wavelet transforms for the problem of single molecule detection can be motivated in a first step by Mallat's heuristic: *Bases of smooth wavelets are the best bases for representing objects composed of singularities, when there may be an arbitrary number of singularities, which may be located in all possible spatial positions* (see for instance [36]).

Since the Fourier basis functions are localized in frequency but not in time, the Fourier transform

$$\mathcal{F}f(\omega) = \widehat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(x) e^{-ix\omega} \, d\, x$$

is the ideal tool to study stationary signals defined by invariance of statistical properties over time. However a small perturbation of frequencies in the Fourier domain will produce changes over the whole time domain.

On the other hand, wavelet transforms, local in both frequency (scale) and time, provide the way to analyze non-stationary signals, characterized by transient events [78, 63]. They offer a compact representation of a signal (e.g. with sharp transitions), which for comparable approximation properties takes significantly fewer wavelet basis functions than Fourier basis functions.

The wavelet transform breaks up a complicated signal in "atoms", called *wavelets*.

A *wavelet* (also called *mother wavelet*) is a function $\psi \in L^2$, which is normalized $\|\psi\| = 1$, has zero average

$$\int_{-\infty}^{+\infty} \psi(t)dt = 0,$$

and is centered in the neighborhood of $t = 0$.

The wavelet ψ is said to have order $m \in \mathbb{N}$ if it has m vanishing moments:

$$\int_{-\infty}^{+\infty} t^n \psi(t)dt = 0, \qquad n = 0, 1, \dots m-1$$

and $\int_{-\infty}^{+\infty} t^m \psi(t)dt \neq 0$. Thus a wavelet of order m is orthogonal to polynomials of degree smaller than m.

The mother wavelet generates a family of wavelets $\psi_{s,u}(x), s > 0, u \in \mathbb{R}$ by change of scale s (i.e. by dilation) and by change of position u (i.e. by translation) of the mother wavelet function $\psi(x)$. Thus each wavelet of the family is obtained from ψ as:

$$\psi_{s,u}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right).$$

Definition 4.1.1. The (basic) wavelet $\psi \in L^2$ satisfies the *admissibility condition* if

$$C_\psi = \int_0^{+\infty} \frac{|(\mathcal{F}\psi)(\omega)|^2}{\omega} d\omega < +\infty. \tag{4.1.1}$$

If ψ satisfies (4.1.1), the *continuous wavelet transform* (CWT) of $f \in L^2(\mathbb{R})$ at position u and scale s is defined by

$$(\mathcal{W}f)(u,s) = \langle f, \psi_{s,u} \rangle = \int_{-\infty}^{+\infty} f(t) \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right) dt = f * \bar{\psi}_s(u), \tag{4.1.2}$$

where $u, s \in \mathbb{R}$, $u \neq 0$, $*$ denotes the convolution operator and

$$\bar{\psi}_s(t) = \frac{1}{\sqrt{s}} \psi^*\left(\frac{-t}{s}\right).$$

Under the assumption (4.1.1), it follows that for reasonable conditions imposed on ψ (i.e. $\psi \in L^1(\mathbb{R})$) the continuous function $\widehat{\psi}$ satisfies $\widehat{\psi}(0) = 0$, or equivalently $\int_{-\infty}^{\infty} \psi(t)\, dt = 0$, justifying the use of condition (4.1.1) as an alternative definition of wavelet functions.

The wavelet transform measures the variation of f in a neighborhood of u of size proportional to s. The regularity of the function f in the neighborhood of u is characterized by the decay of the respective wavelet coefficients, allowing the detection of transients in the signal.

The Fourier transform of $\mathcal{W}f$ with respect to the u variable is

$$(\mathcal{F}\mathcal{W}f)(\omega, s) = (\mathcal{F}f)(\omega) \cdot (\mathcal{F}\psi)(s\omega)$$

and this equation is often used for the implementation of the CWT. The decomposition obtained after the wavelet transform can be further processed for different purposes, such as compression, denoising, pattern recognition etc. These tasks are easier performed in the new wavelet space. Inversion formulas make possible the recovery of the original or a suitably improved function from the expression of the original function in terms of its wavelet transform.

The necessary condition to obtain the inverse of a wavelet transform is the admissibility condition (4.1.1). According to [78], p.81, Calderon in 1964, and Grossman and Morlet in 1984 prove independently that (4.1.1) is the condition for a wavelet transform to be complete (to admit an inverse transform) and conserve energy.

Theorem 4.1.2. *Let the function $\psi \in L^2$ satisfying the admissibility condition (4.1.1). Any function $f \in L^2$ satisfies*

$$f(t) = \frac{1}{C_\psi} \int_0^{+\infty} \int_{-\infty}^{+\infty} \mathcal{W}f(u,s) \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right) du \frac{ds}{s^2}$$

and

$$\int_{-\infty}^{+\infty} |f(t)|^2\, dt = \frac{1}{C_\psi} \int_0^{+\infty} \int_{-\infty}^{+\infty} \left|\mathcal{W}f(u,s)^2\right| du \frac{ds}{s^2}.$$

If $\mathcal{W}f(u,s)$ is known only for $s < s_0$, in order to recover f the information

corresponding to $s \geq s_0$ can be obtained via a *scaling function* φ, satisfying:

$$|(\mathcal{F}\varphi)(\omega)|^2 = \int_1^{+\infty} |(\mathcal{F}\psi)(s\omega)|^2 \frac{ds}{s} = \int_\omega^{+\infty} \frac{|(\mathcal{F}\psi)(\xi)|^2}{\xi} d\xi,$$

and from the admissibility condition results that: $\lim_{\omega \to 0} |\widehat{\varphi}(\omega)|^2 = C_\psi$.

The scaling function φ can be interpreted as a low-pass filter, and the low frequency approximation of function f at scale s is written as:

$$\mathrm{L}f(u,s) = \left\langle f(t), \frac{1}{\sqrt{s}} \varphi\left(\frac{t-u}{s}\right) \right\rangle = f * \bar{\varphi}_s(u).$$

With the help of the scaling function φ, Theorem 4.1.2 can be reformulated as:

$$f(t) = \frac{1}{C_\psi} \int_0^s \mathcal{W}f(\cdot, s) * \psi_s(t) \frac{ds}{s^2} + \frac{1}{C_\psi s_0} \mathrm{L}f(\cdot, s) * \varphi_{s_0}(t). \tag{4.1.3}$$

4.2 Frames

Although very elegant and descriptively rich, the CWT has several drawbacks, such as dimension increase via the introduction of an extra dimension through the scale parameter, high redundancy of the representation as well as the need to adapt the CWT for the case of discrete signals, that are overwhelming in practical applications. In order to overcome these drawbacks, frames are defined as (not necessarily independent) generalizations of bases, leading to redundant signal expansions. Ubiquitous examples of frames are the discrete windowed Fourier transform and the discrete wavelet transforms, we will describe later.

Intuitively, frames are families of functions $\{\varphi_n\}_{n \in \Gamma}$, $\Gamma \subset \mathbb{N}$ that characterize a signal f via its inner products. The study of frames was motivated by the search for the necessary and sufficient conditions that $f \in H$, where H is a Hilbert space, is characterized by the inner products $\langle f, \varphi_n \rangle$ (the frame wavelet coefficients), meaning that: $\langle f_1, \varphi_n \rangle = \langle f_2, \varphi_n \rangle$ implies $f_1 = f_2$ [30]. Moreover, if f can be reconstructed via these inner products, the reconstruction turns out to be of the following simple form:

$$f = \sum_{n \in \Gamma} \langle f, \varphi_n \rangle \, \widetilde{\varphi}_n, \tag{4.2.1}$$

where $\widetilde{\varphi}_n$ is the dual frame of φ_n, and will be defined below.

Definition 4.2.1. The sequence $\{\varphi_n\}_{n \in \Gamma}$ is a *frame* of a Hilbert space H if there

exist two constants $A, B > 0$ such that for any $f \in H$

$$A \|f\|^2 \leq \sum_{n\in\Gamma} |\langle f, \varphi_n\rangle|^2 \leq B \|f\|^2 . \tag{4.2.2}$$

If $A = B$ the frame is called a *tight frame*.

The linear operator from H to $\ell^2(\Gamma) = \left\{x = (x_j)_{j\in\Gamma} : \|x\|^2 = \sum_{j\in\Gamma} |x|_j^2 < \infty\right\}$, defined as $(Ff)_j := \langle f, \varphi_j\rangle$ is called a *frame operator*.

Condition (4.2.2), called *stability condition*, ensures that a function can be recovered from its wavelet frame transform $\langle f, \varphi_n\rangle$ as in (4.2.1).

When the frame is normalized, $\|\varphi_n\| = 1$, A and B give a measure of the redundancy:

- if $\{\varphi_n\}_{n\in\Gamma}$ are linearly independent then $A \leq 1 \leq B$.
- $\{\varphi_n\}_{n\in\Gamma}$ is an orthonormal basis if and only if $A = B = 1$.
- If $A > 1$ the frame is redundant, and A can be seen as a measure of the redundancy factor.

The adjoint of the frame operator F^* can be computed as

$$\langle F^*x, f\rangle = \langle x, Ff\rangle = \sum_{j\in\Gamma} x_j \overline{\langle f, \varphi_j\rangle} = \sum_{j\in\Gamma} x_j \langle \varphi_j, f\rangle ,$$

so that $F^*x = \sum_{j\in\Gamma} x_j\varphi_j$.

Note that F^*F is invertible ([30]), and $\widetilde{F} := (F^*F)^{-1}F^*$ is the left inverse, of minimum sup norm, called pseudo-inverse (see [78]).

Definition 4.2.2. The *dual frame* of $\{\varphi_n\}_{n\in\Gamma}$ is defined as:

$$\widetilde{\varphi}_j = (F^*F)^{-1} \varphi_j.$$

It can be shown that the dual frame is indeed a frame, satisfying

$$\frac{1}{B} \|f\|^2 \leq \sum_{j\in\Gamma} \left|\langle f, \widetilde{\varphi}_j\rangle^2\right| \leq \frac{1}{A} \|f\|^2 .$$

If the frame is tight then $\widetilde{\varphi}_j = A^{-1}\varphi_j$ (for the proof and further details see [30]).

Finally, the reconstruction formula can be now written:

$$f = \widetilde{F}^{-1} F f = \sum_{j \in \Gamma} \langle f, \varphi_j \rangle \, \widetilde{\varphi}_j = \sum_{j \in \Gamma} x_j \, \langle \widetilde{\varphi}_j, f \rangle \, \varphi_j.$$

A *wavelet frame* is constructed from a continuous wavelet family by restricting the parameter u and s in (4.1.2) to discrete values. Since the energy of a wavelet is concentrated around u over a domain proportional to s (in frequency space $1/s$), for good localization properties in both domains, s is sampled along an exponential sequence $\{a^j\}_{j \in \mathbb{Z}}$, while the translation parameter u is sampled uniformly proportional to a^j:

$$\psi_{j,n}(t) = \frac{1}{\sqrt{a^j}} \psi \left(\frac{t - n u_0 a^j}{a^j} \right) \tag{4.2.3}$$

Not every choice of ψ, a and u_0 generates a frame of $L^2(\mathbb{R})$, even if ψ is admissible (the frame condition actually imposes the admissibility of ψ). Necessary and sufficient conditions for ψ, a and u_0 to form a frame are given in [30].

In general, the canonical dual frame $\widetilde{\psi}_{j,n} = (F^* F)^{-1} \psi_{j,n}$ does not have necessarily a wavelet structure, since $F^* F$ does not commute with translation, and neither does $(F^* F)^{-1}$ (however commutativity with dilation is fulfilled).

The function ψ_j is called also *analysis wavelet*, while $\widetilde{\psi}_j$ when it has a wavelet structure is called the *synthesis wavelet*, corresponding to the analysis (decomposition) and the synthesis (reconstruction) steps in the wavelet transform.

Since a frame does not have to be linearly independent (e.g Haar family), the stability condition is weaker than the Riesz basis requirement, defined below.

Definition 4.2.3. A family $\{\varphi_k\}$ is Riesz basis for $L^2(\mathbb{R})$, if

$$L^2(\mathbb{R}) = \overline{\operatorname{span} \{\varphi(\cdot - k) | k \in \mathbb{Z}\}}$$

and

$$A \sum_{k \in \mathbb{Z}} c_k^2 \leq \left\| \sum_{k \in \mathbb{Z}} c_k \varphi(\cdot - k) \right\|_{L^2}^2 \leq B \sum_{k \in \mathbb{Z}} c_k^2$$

for all $\{c_k\}_{k \in \mathbb{Z}} \in l^2(\mathbb{Z})$, where A and B are positive constants.

Every orthonormal basis is a Riesz basis, and every Riesz basis is a frame. The following theorem on the relationship between frames and Riesz bases is proven in [24]:

Theorem 4.2.4. *Given $\psi \in L^2(\mathbb{R})$ the following statements are equivalent:*

- $\{\psi_{j,k}\}$ *is a Riesz basis of* $L^2(\mathbb{R})$
- $\{\psi_{j,k}\}$ *is a frame of* $L^2(\mathbb{R})$ *and it is a* ℓ^2*-linearly independent family, i.e.,*

$$\sum c_{j,k}\psi_{j,k} = 0 \Rightarrow c_{j,k} = 0.$$

Moreover the Riesz bounds coincide with the frame bounds.

For two particular frame classes the duals have an easy characterization, for semi-orthogonal wavelets and the more restrictive orthogonal wavelets both presented in the following definition.

Definition 4.2.5. Given a wavelet $\psi \in L^2((R))$

a) ψ is called an *orthogonal wavelet* if $\{\psi_{j,k}\}$ satisfies:

$$\langle \psi_{j,k}, \psi_{l,m} \rangle = \delta_{j,k} \cdot \delta_{l,m}, \quad \forall j,k,l,m \in \mathbb{Z},$$

b) ψ is called a *semi-orthogonal wavelet* if

$$\langle \psi_{j,k}, \psi_{l,m} \rangle = 0, \quad \forall j \neq l, k, m \in \mathbb{Z}.$$

Note that the orthogonal wavelets are self-dual: $\psi_{j,k} = \widetilde{\psi}_{j,k}$, while the result concerning the dual of a semi-orthogonal wavelet is given in the following theorem (see [24], for details).

Theorem 4.2.6. *If $\psi \in L^2(\mathbb{R})$ is a semi-orthogonal wavelet then its dual can be defined via its Fourier transform as:*

$$(\mathcal{F}\widetilde{\psi})(\omega) := \frac{(\mathcal{F}\psi)}{\sum_{-\infty}^{\infty} |(\mathcal{F}\psi)(\omega + 2\pi k)|^2}$$

meaning that

$$\left\langle \psi_{j,k}, \widetilde{\psi}_{l,m} \right\rangle = \delta_{j,k}\delta_{l,m}, \quad \forall j,k,l,m \in \mathbb{Z}, \tag{4.2.4}$$

where $\widetilde{\psi}_{l,m} := 2^{l/2}\widetilde{\psi}(2^l x - m)$.

In this case the dual has a wavelet structure generated by $\widetilde{\psi}$ and every $f \in L^2(\mathbb{R})$ can be written as:

$$f = \sum_{j,k\in\mathbb{Z}} c_{j,k}\psi_{j,k}(x) = \sum_{j,k\in\mathbb{Z}} d_{j,k}\widetilde{\psi}_{j,k}(x), \tag{4.2.5}$$

where the sums are known as *wavelet series* and from (4.2.4) follows that:

$$c_{j,k} = \left\langle f, \widetilde{\psi}_{j,k} \right\rangle$$
$$d_{j,k} = \langle f, \psi_{j,k} \rangle .$$

It is clear that ψ and $\widetilde{\psi}$ are interchangeable (either one can be used in analysis and the other in synthesis).

4.3 Multi-resolution analysis

For an efficient computation of the wavelet transform and the reconstruction of a signal via the inverse transform the concept of multiresolution analysis is introduced. The multiresolution analysis (MRA) offers a hierarchical framework for interpreting the information in the signal, with a parsimonious representation of important features at different scales in only a few coefficients. It allows a scale invariant analysis of signal features and via a special coarse-to-fine strategy examines the decorrelated components of the signal at different resolutions. It is a convenient tool in pattern recognition, denoising, detection etc.

The computations implied by MRA can be conveniently described through filters and filter banks. For every continuous linear operator $F : \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$ there exist $\{h_k\}_{k\in\mathbb{Z}} \in \ell^2(\mathbb{Z})$ such that

$$(Fx)_k = \sum_{n\in\mathbb{Z}} h_{k-n} x_n.$$

$h = (\ldots h_{k-1}, h_k, h_{k+1} \ldots)$ represents the impulse response of the filter, and the elements are called filter coefficients. A convenient way to describe filters is through their Z transform, expressed as a Laurent polynomial:

$$H(z) = \sum_i h_i z^{-i}.$$

Louis *et al.* [76] give an elegant motivation for multiresolution analysis. A signal $f \in V_{-1}$, where V_{-1} is a subspace of $L^2(\mathbb{R})$, is decomposed in a smooth (low-frequency) component via a projection $P_0 f \in V_0$ and a detail (high-frequency) component, $Q_0 f \in W_0$, where W_0 is the orthogonal complement of V_0, $W_0 =$

$\{x \in V_{-1} : \langle x, y\rangle = 0,\ y \in V_0\}$ denoted $V_{-1} = V_0 \oplus W_0$, such that

$$f = P_0 f + Q_0 f.$$

In the next step, $P_0 f$ is decomposed by projections P_1 and Q_1 on orthogonal subspaces V_1 and W_1, respectively, such that $P_0 f = P_1 f + Q_1 f$ and $f = P_1 f + Q_1 f + Q_0 f$. The process can be repeated recursively.

Definition 4.3.1. A sequence of nested, closed subspaces $V_j \subset L^2(\mathbb{R})$,

$$0 \subset \ldots V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \ldots \subset L^2(\mathbb{R})$$

is called a *multiresolution analysis* (MRA) if

- $\overline{\lim_{j\to-\infty} V_j} = \overline{\bigcup_{j\in\mathbb{Z}}} = L^2(\mathbb{R})$
- $\lim_{j\to\infty} V_j = \bigcap_{j\in\mathbb{Z}} V_j = 0$
- $f(\cdot) \in V_j$ if and only if $f(2^j \cdot) \in V_0$
- $f \in V_j$ if and only if $f(\cdot - 2^j k) \in V_j$,$k \in \mathbb{Z}$
- $\exists \varphi \in L^2(\mathbb{R}) : \varphi(\cdot - k)_{k\in\mathbb{Z}}$ is a Riesz basis for V_0.

It can be shown that $f \in V_0$ if and only if $f(\cdot - k) \in V_0, \forall k \in \mathbb{Z}$ and

$$V_j = \overline{\operatorname{span}\{\varphi_{j,k}(\cdot - k) | k \in \mathbb{Z}\}},$$

where $\varphi_{j,k}(x) := 2^{-j/2}\varphi(2^{-j}x - k)$.

To approximate f in V_j the projection on the scaling basis is considered, such that the inner products

$$c_{j,n} = \langle f, \phi_{j,n}\rangle = \int_{-\infty}^{\infty} f(t) \frac{1}{2^{j/2}} \phi\left(\frac{t - 2^j n}{2^j}\right) dt$$

represent a discrete approximation at scale 2^j.

The following lemma is a consequence of the observation:

$$2^{-1/2}\varphi(x/2) \in V_1 \subset V_0 = \overline{\operatorname{span}\{\varphi(x - k) | k \in \mathbb{Z}\}}.$$

Lemma 4.3.2. *There exists* $\{h_k\}_{k\in\mathbb{Z}} \in \mathbb{R}$ *such that*

$$2^{-1/2}\varphi\left(\frac{x}{2}\right) = \sum_{k\in\mathbb{Z}} h_k \varphi(x - k), \tag{4.3.1}$$

where

$$h_k = \left\langle 2^{-1/2}\varphi\left(\frac{x}{2}\right), \varphi(x-k)\right\rangle.$$

Equation (4.3.1) *is called* two-scale equation *or* refinement equation.

Lemma 4.3.2 relates a dilation of φ by 2 to its integer translations via h, a discrete filter called *conjugate mirror filter.*

On the other hand, the difference of information between the approximation of the function f at resolution 2^{j+1} and 2^j respectively is called detail signal at resolution 2^j. By denoting with W_j the complement of V_j in V_{j+1}, $V_{j-1} = V_j \dot{+} W_j$, one can write the decomposition of $L^2(\mathbb{R})$ as the *direct sum* of the W_j spaces :

$$L^2(\mathbb{R}) = \dot{\sum}_{j\in\mathbb{Z}} W_j = \ldots \dot{+} W_1 \dot{+} W_0 \dot{+} W_1 \ldots.$$

As opposed to the nested subspaces V_j, the subspaces W_j are disjunct: $\forall j \neq l$, $W_j \cap W_l = \{0\}$. In order to have an orthonormal basis one can chose W_j to be the orthogonal complement of V_j in V_{j-1}, $V_j \perp W_j$, as in the motivational example for MRA,:

$$V_{j-1} = V_j \oplus W_j.$$

The following theorem due to Mallat and Meyer [78] p.236 gives the construction of an orthonormal basis of $W_j$by scaling and translations of a wavelet ψ, constructed upon a scaling function φ.

Theorem 4.3.3. *Given φ a scaling function and h the corresponding conjugate mirror filter, denote by ψ the function whose Fourier transform is*

$$(\mathcal{F}\psi)(\omega) = \frac{1}{\sqrt{2}}(\mathcal{F}g)\left(\frac{\omega}{2}\right)\left(\mathcal{F}(\varphi)\left(\frac{\omega}{2}\right)\right),$$

where

$$(\mathcal{F}g)(\omega) = \mathrm{e}^{-i\omega}(\mathcal{F}h^*)(\omega+\pi). \tag{4.3.2}$$

For any scale 2^j, $\{\psi_{j,n}\}_{n\in\mathbb{Z}}$ is an orthonormal basis of W_j, where

$$\psi_{j,n}(t) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{t-2^j n}{2^j}\right).$$

For all scales, $\{\psi_{j,n}\}_{(j,n)\in\mathbb{Z}^2}$ is an orthonormal basis of $L^2(\mathbb{R})$.

From Theorem 4.3.3 it results that

$$2^{-1/2}\psi\left(\frac{x}{2}\right) = \sum_{k\in\mathbb{Z}} g_k \varphi(x-k),$$

and subsequently

$$g_k = \left\langle 2^{-1/2}\psi\left(\frac{x}{2}\right), \varphi(x-k)\right\rangle.$$

Finally, the relation between g and h in (4.3.2) can be rewritten after the inverse Fourier transform as:

$$g_k = (-1)^{1-k} h_{1-k}.$$

4.4 Generalizations of MRA to two dimensions

The previous one dimensional multiresolution analysis setting can be extended to higher dimensions $L^2(\mathbb{R}^d)$ in a straightforward way.

If $\{V_j\}_{j\in\mathbb{Z}}$ is a two-dimensional multiresolution analysis, the two-dimensional signal $f(x,y) \in L^2(\mathbb{R}^2)$ can be approximated by its projection on V_j. Denoting $\varphi_{2^j} := 2^{2^j}\varphi(2^j, 2^j y)$, it can be shown that the family of functions obtained by the dilation and translation of the scaling function $\varphi(x,y)$:

$$\left\{2^{-j}\varphi_{2^j}(x-2^{-j}n, y-2^{-j}m)\right\}_{(n,m)\in\mathbb{Z}^2}$$

forms an orthonormal basis of V_j.

The function $\varphi(x,y)$ is unique with respect to a given multiresolution analysis. A special case of multiresolution approximations is the one of separable MRA, where each vector space V_j can be decomposed as a tensor product of two identical subspaces of $L^2(\mathbb{R})$ (representing MRA of $L^2(\mathbb{R})$)

$$V_j = V_j^1 \otimes V_j^1 = \overline{\operatorname{span}\left\{f(x)g(y) \mid f,g \in V_j^1\right\}}.$$

and subsequently the scaling function can be written as $\varphi(x,y) = \varphi(x)\varphi(y)$, where $\varphi(x)$ is the respective one-dimensional scaling function.

The orthogonal basis of V_j is given by:

$$\left\{2^{-j}\varphi_{2^j}(x-2^{-j}n, y-2^{-j}m)\right\}_{(n,m)\in\mathbb{Z}^2} = \left\{2^{-j}\varphi_{2^j}(x-2^{-j}n)\varphi_{2^j}(y-2^{-j}m)\right\}_{(n,m)\in\mathbb{Z}^2}$$

and the approximation of the signal $f(x,y)$ at resolution 2^{-j} is characterized by

the inner products:

$$\left\{\left\langle f(x,y), \varphi_{2^j}(x-2^{-j}n)\varphi_{2^j}(y-2^{-j}m)\right\rangle\right\}_{(m,n)\in\mathbb{Z}^2}.$$

Just as V_j is defined with the help of V_j^1, also the details W_j of the two-dimensional MRA, are based on the respective W_j^1 the complements of V_j^1 in the one dimensional case:

$$\begin{aligned} V_{j-1} &= V_{j-1}^1 \otimes V_{j-1}^1 \\ &= (V_j^1 \oplus W_j^1) \otimes (V_j^1 \oplus W_j^1) \\ &= (V_j^1 \otimes V_j^1) \oplus ((V_j^1 \otimes W_j^1) \oplus (W_j^1 \otimes V_j^1) \oplus (W_j^1 \otimes W_j^1)) \\ &= V_j \oplus W_j. \end{aligned}$$

Using the notations above, the following result, is proven in [77]:

Theorem 4.4.1. *If W_j is the orthogonal complement of V_j, where $\{V_j\}_{j\in\mathbb{Z}}$ is an MRA of $L^2(\mathbb{R}^2)$, with generating scaling function $\varphi(x,y)=\varphi(x)\varphi(x)$, and $\psi(x)$ is the one-dimensional wavelet associated to $\varphi(x)$ then*

$$\begin{aligned} &\{2^{-j}\psi_{2^j}^1(x-2^{-j}n, 2^j, y-2^{-j}m), \\ &\ 2^{-j}\psi_{2^j}^2(x-2^{-j}n, 2^j, y-2^{-j}m), \\ &\ 2^{-j}\psi_{2^j}^3(x-2^{-j}n, 2^j, y-2^{-j}m)\}_{(n,m)\in\mathbb{Z}^2}, \end{aligned}$$

where

$$\psi^h(x,y)=\varphi(x)\psi(y),\ \psi^v(x,y)=\psi(x)\varphi(y),\ \psi^d(x,y)=\psi(x)\psi(y),$$

is an orthonormal basis of W_j.

The scalar products corresponding to the three kinds of wavelets give the details of the function $f(x,y)$, with respect to three emphasized directions: horizontal, vertical and diagonal. These directions might have a special role in human vision, however are not so important in microscopy images.

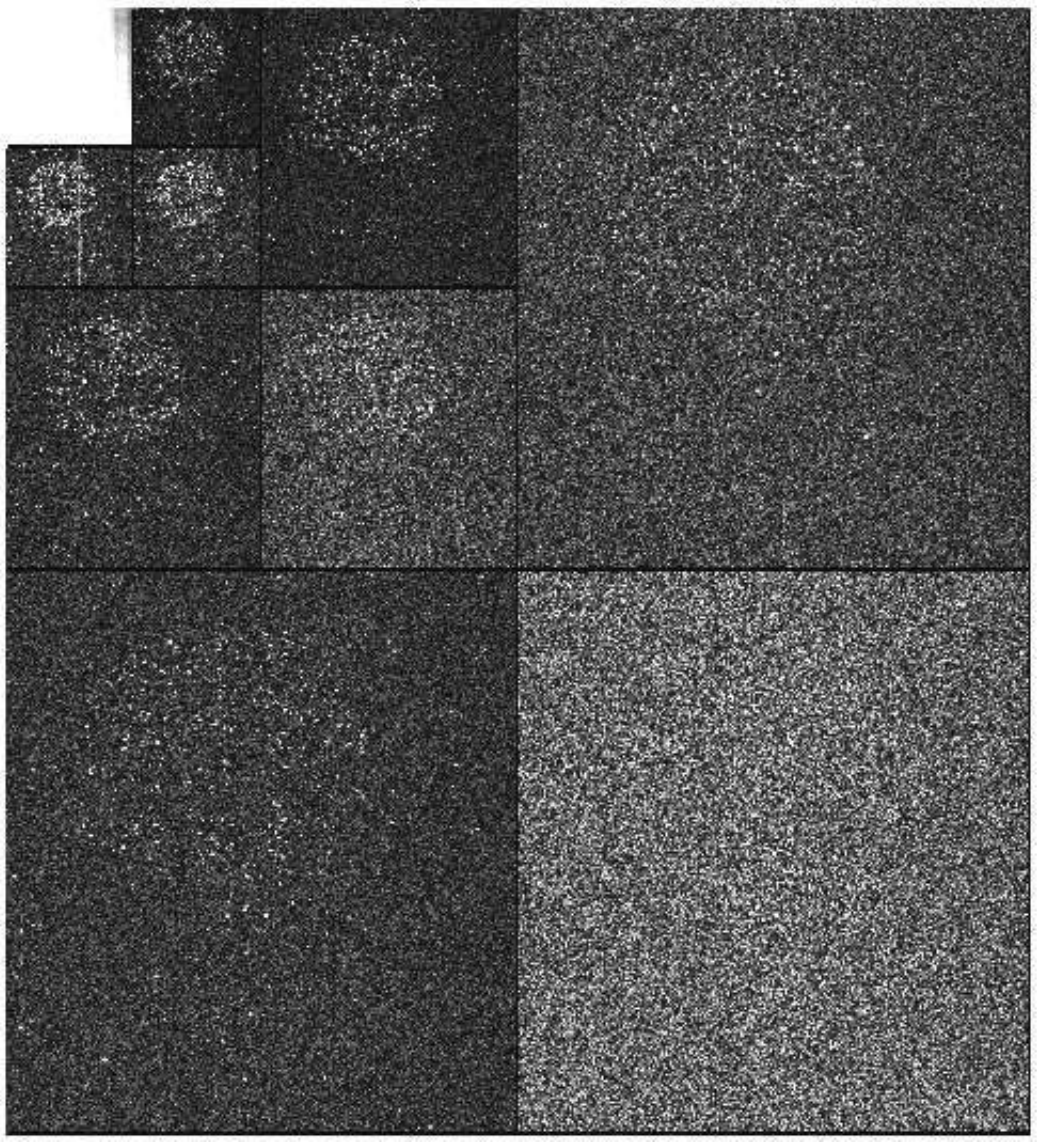


Figure 4.1: Orthogonal wavelet decomposition (Haar wavelets)

4.5 B-spline frames

The first and still very popular wavelet family is the family introduced by Haar in 1910 (see for example [12]), the *Haar wavelets*. Based on the scaling function:

$$\varphi(x) = \begin{cases} 1, & 0 \leq x < 1 \\ 0, & \text{otherwise} \end{cases} \tag{4.5.1}$$

they fulfill a simple scaling equation:

$$\varphi_{j+1,k} = \frac{1}{\sqrt{2}}(\varphi_{j,2k} - \varphi_{j,2k+1})$$

and based on the difference $\frac{1}{\sqrt{2}}(\varphi_{j,2k} - \varphi_{j,2k+1}) = \psi_{j+1,k}$ the Haar wavelet can be written as:

$$\varphi(x) = \begin{cases} 1, & 0 \le x < 1/2 \\ -1, & 1/2 \le x < 1 \\ 0, & \text{otherwise.} \end{cases} \tag{4.5.2}$$

Haar wavelets are a particular case of B-spline wavelets (of order 0). The n-th order B-spline is recursively defined as:

$$B_0(x) = \chi_{[0,1]}(2x)$$
$$B_{n+1}(x) = (B_n * B_0)(x) = \int_{-\infty}^{\infty} B_{m-1}(x-t)B_0(t)\,dt = \int_0^1 B_{m-1}(x-t)\,dt,$$

normalized such that $\int_{\mathbb{R}} B_n(x)\,dx = 1$. The Fourier transform of B_m is

$$(\mathcal{F}B_m)(\omega) = \frac{1}{\sqrt{2\pi}}\left(\frac{\sin(\omega/2)}{\omega/2}\right)^{m+1} = \frac{1}{\sqrt{2\pi}}\text{sinc}^{m+1}(\omega/2).$$

For any j and $m \ge 2$ the family

$$\mathcal{B}_j := \left\{2^{j/2}B_m(2^j x - k), k \in \mathbb{Z}\right\}$$

is a Riesz basis of V_j^m. Since the spline function $B_m(2^j x) \in V_j^m$ and $V_j^m \subset V_{j+1}^m$ it results that it can be written as:

$$B_m(2^j x) = \sum_{k=-\infty}^{\infty} p_{m,k}B_m(2^{j+1}x - k), \tag{4.5.3}$$

where $\{p_{m,k} : k \in \mathbb{Z}\}$ is a ℓ^2-sequence. Solving (4.5.3) for $p_{m,k}$ in Fourier domain yields:

$$p_{m,k} = \begin{cases} 2^{-m+1}\begin{pmatrix} m \\ k \end{pmatrix}, & 0 \le k < m \\ 0, & \text{otherwise} \end{cases} \tag{4.5.4}$$

and the corresponding two-scale relation is then given by:

$$B_m(x) = \sum_{k=0}^{m} 2^{-m+1} \begin{pmatrix} m \\ k \end{pmatrix} B_m(2x-k).$$

The first result concerning spline wavelets associated to B-spline scaling functions of order m are called *cardinal spline wavelets* (see [24] for a detailed description). Their disadvantage is that although the wavelets have exponential decay, their support is not compact thus being tedious to compute.

In order to remedy this shortcoming the orthogonality requirement is dropped. Semi-orthogonal wavelets with compact supports always exist if the two-scale sequence of the scaling function φ has finite support. Chui in [24] describes the unique solution for the compactly supported semi-orthogonal wavelet ψ_m with minimum support that corresponds to the m-th order cardinal B-spline:

$$\begin{cases} \psi_m(x) := \sum_n q_n B_m(2x-n) \\ q_n = \frac{(-1)^n}{2^{m-1}} \sum_{l=0}^{m} \begin{pmatrix} m \\ l \end{pmatrix} B_{2m}(n+1-l), n = 0, \ldots, 3m-2. \end{cases} \tag{4.5.5}$$

Several properties of spline wavelets are discussed in [24, 111, 77, 78].

4.6 Fast wavelet transform algorithms via filter banks

In most of practical cases, one has to deal with discrete signals. A real digital signal is a sequence $f_i = f(i) \in \mathbb{R}$. We shall consider only square summable signals $f \in l_2(\mathbb{Z})$. The simplest discretization of an analog input $f(t)$ can be achieved via natural sampling:

$$f[n] = f(t = n),\ n \in \mathbb{Z}.$$

The discrete signal can be related to $f(t)$ also via a D/A converter:

$$f(t) \approx \sum_n f[n]\chi(t-n).$$

It includes the natural sampling case when $\chi(n) = \delta_{0n}$. Wavelet algorithms for discrete signals as well as their connection to the CWT were thoroughly described

in [100, 77, 90, 87], with a focus on computational details.

Most frequently computations are made over the dyadic grid $u = 2^j, s = k2^j$ for the scale and translation parameter of the wavelet. Mallat in [77] proposed the use of conjugate mirror filters for the computation of the discrete wavelet transform (DWT) in the frame of multiresolution analysis. Two special operators are used in the transition from one resolution to another. The decimation operator $D : \ell^2(\mathbb{Z}) \to \ell^2(2\mathbb{Z})$, $D_{k,m} = \delta_{2k,m}$ retains only the terms with even index in the sequence (alternative notation: $2 \downarrow 1$). The Fourier transform of a decimated signal $y_n = x_{2n}$ is

$$\widehat{y}(2\omega) = \sum_n x_{2n} e^{-i2n\omega} = \frac{1}{2}\left(\widehat{x}(\omega) + \widehat{x}(\omega + \pi)\right).$$

The component $\widehat{x}(\omega + \pi)$ creates a frequency folding called aliasing that must be canceled at the reconstruction of the signal.

The adjoint operator $E = D^* : \ell^2(2\mathbb{Z}) \to \ell^2(\mathbb{Z})$ inserts 0's at the odd indices: $E_{k,m} = \delta_{k,2m}$ resulting in

$$\ldots 0,\ x_{-1},\ 0,\ x_0,\ 0,\ x_1,\ 0 \ldots$$

It is also denoted by $1 \uparrow 2$.

The signal f is decomposed iteratively by the low pass filter h and the high-pass filter g, both filtered results being subsequently decimated (by using the operator D). The procedure is repeated for the low-pass filtered signal. These steps applied recursively represent the analysis part of the algorithm, depicted on the left of Fig. 4.2. The connection between φ, ψ and h and g, respectively was given at the end of Section 4.2.

The computation can be summarized as the iteration of the following steps:

$$c_{j,n}(f) = \sum_k h_{2n-k} c_{j-1,k}(f), \tag{4.6.1}$$

$$w_{j,n}(f) = \sum_k g_{2n-k} c_{j-1,k}(f), \tag{4.6.2}$$

where $c_{0,k} = f(k)$ and

$$g_i = (-1)^i h_{-i+1}, \quad h_i = 2^{1/2} \int \varphi(x - i)\varphi(2x)\, dx.$$

As for reconstruction (synthesis), on the right side in Fig. 4.2, since $c_{j,n}(f)$

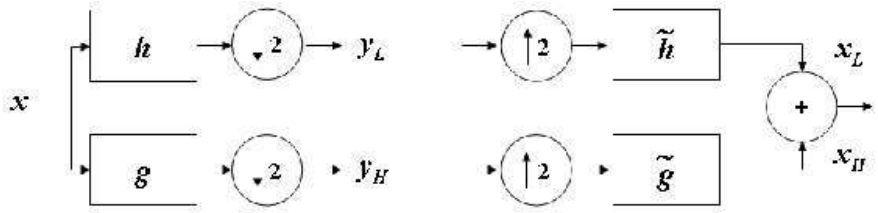


Figure 4.2: Signal decomposition and reconstruction (one level)

represents the projection on the subspace V_j, and $w_{j,n}$ on W_j, the following holds in case of orthogonality:

$$c_{j-1,i}(f) = \sum_n h_{2n-i} c_{j,n}(f) + \sum_n g_{2n-i} w_{j,n}(f).$$

Before the convolutions with the respective filters, note that the signals are upsampled via the operator E. If no further processing is applied after the analysis step, the synthesis will recover the original signal. It is important for the computational efficiency of the algorithm that the filters h and g (equivalently ψ and φ) have compact support.

The algorithm above represents a critically sampled filter bank corresponding to the orthonormal and biorthonormal filter bases (for details see [77, 78]). In the more general class of frames described in Section 4.2 the equivalent algorithm is based on the oversampled filter banks (for details see [28]). The reconstruction formula is given in (4.2.5), where the scaling functions are denoted by φ, wavelets by ψ and their duals $\widetilde{\varphi}$ and $\widetilde{\psi}$. Special reconstruction conditions (see [28, 5]) have to hold in order to obtain a perfect reconstruction of the signal. The perfect reconstruction condition (canceling the aliasing) for the oversampled case (see [28]) is given by

$$\begin{aligned}(\mathcal{F}h^*)(\omega+\pi)(\mathcal{F}\widetilde{h})(\omega) + (\mathcal{F}g^*)(\omega+\pi)(\mathcal{F}\widetilde{g})(\omega) &= 0\\ (\mathcal{F}h^*)(\omega)(\mathcal{F}\widetilde{h})(\omega) + (\mathcal{F}g^*)(\omega)(\mathcal{F}\widetilde{g})(\omega) &= 2.\end{aligned}$$

4.7 Translation invariance

Although computationally and memory-wise very efficient, the sampling over the dyadic grid has drawbacks in pattern recognition, since the description of a pattern should not depend on the position where the pattern appear. The lack of

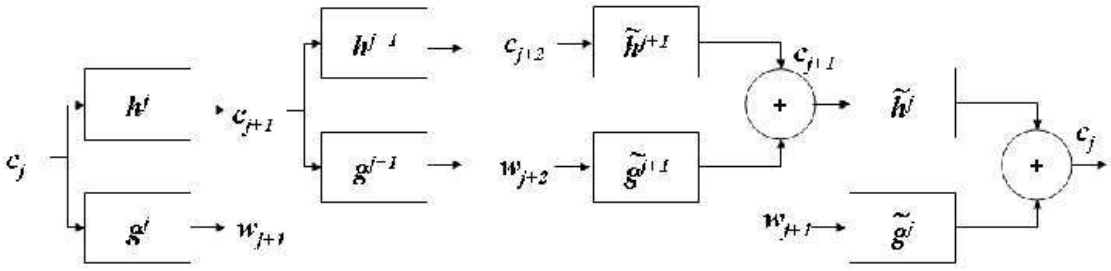


Figure 4.3: Undecimated signal decomposition and reconstruction (two consecutive levels)

translation invariance of the dyadic grid sampling violates the principle of independence of description from the location. We shall see in the next chapter that also a certain amount of redundancy of wavelet coefficients can prove beneficial in applications like denoising and image restoration.

In order to recognize a pattern in a signal, the numerical descriptors of the pattern should be translation invariant. The continuous wavelet transform is translation invariant: $Wf_\tau(u,s) = Wf(u-\tau,s)$, where $f_\tau(t) = f(t-\tau)$, but the uniform sampling of the translation parameter u of a wavelet frame as in (4.2.3) does not preserve this property: $Wf_\tau(ka^ju_0, a^j)$ might be very different from $Wf(ka^ju_0, a^j)$ if τ is not of the form $\tau = ka^ju_0$ and the sampling interval a^ju_0 is large relative to the rate of variation of $f * \psi_{a^j}$ (see [78]).

Undecimated wavelet transform

The obvious solution to preserve translation invariance is to modify the discretization of the translation parameter u. The scale s is still sampled along the dyadic sequence $\{2^j\}_{j\in\mathbb{Z}}$ but u is not subsampled and the wavelet transform is defined as

$$Wf(u,2^j) = \int f(t)\frac{1}{2^j}\psi\left(\frac{t-u}{2^j}\right)dt = f * \psi_j(u). \tag{4.7.1}$$

It can be proven that the suitably normalized dyadic wavelet transform from (4.7.1), $\sqrt{2^{-j/2}}W$, is a frame operator. A fast dyadic transform is schematically represented in Fig. 4.3. Note that all decimations and dilations are dropped. Instead modified filters $h^{(j)}$ and $g^{(j)}$ are used, where $h^{(j)}$ is obtained from $h^{(j-1)}$ by applying the dilation operator E:

$$h^{(j)} = (\ldots 0, h_{-1}^{(j-1)}, 0, h_0^{(j-1)}, 0, h_1^{(j-1)} 0 \ldots),$$

and $g^{(j)}$ is obtained similarly from $g^{(j-1)}$. In order to compute the multiresolution analysis, the operation above performs the dilation of the wavelet ψ and scaling function φ (and also samples these functions in more points) in order to compensate the lack of decimation of the signal, through which multiresolution was achieved in the algorithm of Mallat. The construction based on zero insertion inspired the name of the method (*à trous* is French for 'with holes').

With these new filters the algorithm becomes the iteration of

$$c_{j+1,k} = c_{jk} * h_k^{(j)}, \qquad w_{j+1,k} = c_{jk} * g_k^{(j)} \tag{4.7.2}$$

as decomposition steps and as for reconstruction:

$$c_{j+1,k} = \frac{1}{2}(c_{j+1,k} * \widetilde{h}_k^{(j)} + w_{j+1,k} * \widetilde{g}_k^{(j)}).$$

The extension of the undecimated wavelet transform to two dimensions is achieved (for the sake of simplicity) via separable filters:

$$\begin{aligned}
c_{j+1}(k,l) &= (h^{(j)}h^{(j)} * c_j)(k,l) \\
w^1_{j+1}(k,l) &= (g^{(j)}h^{(j)} * c_j)(k,l) \\
w^2_{j+1}(k,l) &= (h^{(j)}g^{(j)} * c_j)(k,l) \\
w^3_{j+1}(k,l) &= (g^{(j)}g^{(j)} * c_j)(k,l), \quad j = 1, \ldots, J-1.
\end{aligned}$$

Note that each of the transformed images has the same size as the original image, such that the redundancy factor is $3(J-1)+1$.

Isotropic undecimated wavelet transform

In single molecule images, features are usually isotropic and in order to perform a good image analysis the isotropy of the filters is a desirable property. In order to preserve isotropy, the filters h and g as well as the scaling function and the mother wavelet function φ and ψ have to be nearly isotropic. A popular choice is based on one-dimensional B-spline scaling function of order 3, $\varphi(x) = B_3(x)$, to which corresponds the filter $h_k = \left[\frac{1}{16}, \frac{1}{4}, \frac{3}{8}, \frac{1}{4}, \frac{1}{16}\right]$. Based on $\varphi(x)$, a $2d$ separable filter is constructed such that:

$$\varphi(x,y) = B_3(x)B_3(y)$$

and

$$\psi\left(\frac{x}{2}, \frac{y}{2}\right) = 4\varphi(x,y) - \varphi\left(\frac{x}{2}, \frac{y}{2}\right).$$

The related filters are $h_{(k,l)} = h_k \times h_l$ and $g_{(k,l)} = \mathbf{1}_{0,0} - h_{k,l}$, where $\mathbf{1}_{0,0} = 1$ at $(0,0)$ and is null otherwise. It means that the wavelet detail coefficients are given by: $w_{j+1,(k,l)} = c_{j,(k,l)} - c_{j+1,(k,l)}$ [106] and the reconstruction is the sum of all details and the coarsest approximation:

$$f_{(k,l)} = c_{J,(k,l)} + \sum_{j=0}^{J} w_{j+1,(k,l)}. \tag{4.7.3}$$

The detail at each resolution has the same dimension as the original $2d$ signal. When there is no confusion, a single index will be used to denote the $2d$ index (k,l). The index j will be preserved to denote the scale.

The relation to the UWT based on the same h and $g = \mathbf{1} - h = [-1,-4,10,-4,-1]$ is

$$w_j^1 + w_j^2 + w_j^3 = w_j = c_j - c_{j+1}$$

Further details on UWT and IUWT can be found in [106]. The *à trous* scheme was used in [107] for astronomical images, and in [106, 88] for microscopy.

Chapter 5

Wavelet based detection

The previous chapter offered a short introduction in wavelet transforms and focused on the transforms suited to analyze biological structures in general, and single molecules in particular, like the IUWT transform described in Section 4.7.

The multiresolution analysis of a signal makes possible a good localization of interesting areas (e.g. singularities) as well as a sparse representation, by only few important wavelet coefficients, since the transform adapts to the local regularity of the signal. Jaffard and Meyer ([62]) explain the efficiency of the wavelet transform via the following paradigm: *Objects, images, or signals with simple geometrical structures have sparse wavelet expansions. Conversely, functions whose wavelet expansion is sparse have interesting geometrical properties.* In this chapter we shall describe thresholding algorithms that make use of the convenient representation with the final goal of performing the single molecule detection task.

5.1 Statistical applications of wavelet transforms

The solution of several statistical problems (estimation, approximation, model selection, pattern recognition, etc.) can be generically formulated as the recovery of the original (signal) coefficients expressed in a certain basis from the empirical ones (see e.g. [78, 19]). Donoho explored the connection between the estimation problem and the representation problem, and showed that *"efficient representations lead to efficient estimations"* [36, 37, 19].

The efficiency of estimation is motivated by the decorrelating property of a wavelet transform which creates a sparse signal, with most coefficients zero or close to zero. The noise however affects equally all coefficients and if its level is moderate, signal wavelet coefficients can be easily discriminated from coefficients due to noise

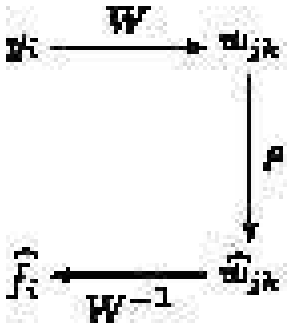


Figure 5.1: Diagram of thresholding algorithms

by simple thresholding techniques. A fundamental oracle inequality relates the performance of thresholding rules to the sparsity of such wavelet representations [19].

A simple setting to illustrate the applications of the wavelet transforms in statistics is the single sequence problem, seen as a non-parametric regression problem and defined as the estimation of a function f (e.g. $f : \mathbb{R} \to \mathbb{R}$) from the data $\{y_i, i = 1, \ldots, n\}$ corrupted by additive noise:

$$y_i = f(t_i) + \varepsilon_i, \;\; i = 1, \ldots, n. \tag{5.1.1}$$

We shall consider f equally sampled at points $t_i = i\Delta$ and denote $f_i = f(t_i)$. Most frequently the noise ε_i is considered to be independent and identically distributed (i.i.d) white noise, following a Gaussian distribution $\varepsilon_i \sim \mathcal{N}(\nu, \sigma)$. Without loss of generality, one can assume $\varepsilon_i \sim \mathcal{N}(0, \sigma)$. Different noise models are described in Section 5.6.

The performance of an estimator of f, denoted $\widehat{f}$, can be measured by the risk based on ℓ_2 loss:

$$R(\widehat{f}, f) = n^{-1}\mathbb{E}\left\|\widehat{f} - f\right\|_{\ell_n^2}^2 = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{f}_i - f_i\right)^2\right]. \tag{5.1.2}$$

Wavelets offer a convenient basis for signal representation, so the entire setting (5.1.1) can be transposed in wavelet space. A typical algorithm to estimate $\widehat{f}$ consists of three major steps: wavelet decomposition, estimation of true wavelet coefficients with the help of an operator ρ and reconstruction of the signal via inverse wavelet transform based on the estimated coefficients. See Fig. 5.1 for a schematic representation. The function f is not known, so the continuous wavelet

coefficients

$$w_{j,k} = \int f(u)\psi_{j,k}(u)\,du$$

cannot be directly computed, and are estimated from the empirical wavelet coefficients. The coefficients $w_{j,k}$ of the discrete wavelet transform of the signal $f = (f_i)_{i\in 1,\dots n}$ at scale j and location k are given by:

$$w_{j,k} = \sum_{i=1}^{n} f_i \psi_{j,i}.$$

Moreover, the empirical coefficients $\widetilde{w}_{j,k}$ of the discrete wavelet transform of the noisy data $y = (y_i)_{i\in 1,\dots n}$ at scale j and location k are given by:

$$\widetilde{w}_{j,k} = \sum_{i=1}^{n} y_i \psi_{j,i}.$$

Given an estimate $\widehat{w}_{i,k}$ of the wavelet coefficient of the signal f and an orthogonal wavelet transform, one can get an estimate of the signal by applying the following inverse transform:

$$\widehat{f}_i = \sum_{j=1}^{J} \sum_{k=0}^{2^j-1} \widehat{w}_{j,k}\psi_{j,k} + \sum_{k=0}^{2^J} \widehat{c}_{J,k}\varphi_{J,k}.$$

The approximation $n^{1/2}\psi_{j,k}(i) \approx 2^{j/2}\psi(2^j u - k),\ u = i/n$ holds for the discrete case. Furthermore, to each wavelet level corresponds an estimation problem of type (5.1.1), and we have:

$$\widetilde{w}_{j,k} = w_{j,k} + \varepsilon_{j,k} \quad \text{for all } j = 1,\dots,J \quad \text{and} \quad k = 0,\dots,2^j - 1, \tag{5.1.3}$$

where $\mathbb{E}\left[\varepsilon_{j,k}\right] = 0$ and $\mathbb{E}\left[\varepsilon_{j,k}^2\right] = \sigma_j^2$.

Finally, due to the orthonormality we have also $\mathbb{E}\left\{\varepsilon_{i_1}\epsilon_{i_2}\right\} = \sigma^2\delta_{i_1-i_2}$ since:

$$\begin{aligned}\mathbb{E}\left(\sum_{i_1}\psi_{j_1,k_1}\varepsilon_{i_1}\right)\left(\sum_{i_2}\psi_{j_2,k_2}\varepsilon_{i_2}\right) &= \sum_{i_1}\sum_{i_2}\mathbb{E}\left\{\varepsilon_{i_1}\varepsilon_{i_2}\right\}\psi_{j_1,k_1}\psi_{j_2,k_2} \\ &= \sigma^2\delta_{(j_1,k_1),(j_2,k_2)}. \end{aligned}\tag{5.1.4}$$

Thus, the orthogonal wavelet transformed noise coefficients are also white noise.

5.2 Wavelet coefficient estimation

The problem described in (5.1.1) can be transposed in wavelet space, as a wavelet coefficient estimation problem. The risk equivalent to (5.1.2) becomes:

$$R(\widehat{f}, f) = \mathbb{E}\left[\sum_{j}\sum_{k}\left(\widehat{w}_{j,k} - w_{j,k}\right)^2\right]. \tag{5.2.1}$$

The minimization of the risk is achieved via the algorithms described in Fig. 5.1, more precisely the modification of wavelet coefficients via the operator ρ. In general, these algorithms operate with two simplifications. First, only diagonal estimators are considered, and second, these estimators have a special form (linear or thresholding functions, called also shrinkage functions).

A *diagonal estimator* in the discrete wavelet basis estimates independently each $w_{j,k}$ from $\widetilde{w}_{j,k}$ via a function ρ: $\widehat{w}_{j,k} = \rho(\widetilde{w}_{j,k})$. The algorithm can be thus summarized as:

$$\widehat{f} = \sum_{j=1}^{J}\sum_{k=0}^{2^j-1}\rho\left(\langle y, \psi_{j,k}\rangle\right)\psi_{j,k} + \sum_{k=0}^{2^J}\rho\left(\langle y, \varphi_{J,k}\rangle\right)\varphi_{J,k}.$$

Additionally as a simple scenario we can assume ρ to be a linear function $\rho(\widetilde{w}_{j,k}) = a_{j,k}\widetilde{w}_{j,k}$ that minimizes the risk (5.2.1). Level-wise, considering (5.1.3) it follows that

$$\mathbb{E}\left\|w_{j,k} - a_{j,k}\widetilde{w}_{j,k}\right\|^2 = \left|w_{j,k}\right|^2\left(1 - a_{j,k}\right)^2 + a_{j,k}\sigma_j^2. \tag{5.2.2}$$

The minimum of (5.2.2) is attained for

$$a_{j,k} = \frac{\left|w_{j,k}\right|^2}{\left|w_{j,k}\right|^2 + \sigma_j^2},$$

yielding

$$R_L(\widehat{f}, f) = \sum_{j,k}\frac{\left|w_{j,k}\right|^2\sigma_j^2}{\left|w_{j,k}\right|^2 + \sigma_j^2}.$$

We note that the linear factor $a_{j,k}$ cannot be computed in practice, since it depends on the unknown coefficient $w_{j,k}$. Therefore the risk is not reachable, it is an ideal risk that can be attained only with information on ideal smoothing of f provided by an *oracle*, (an information which practically is not available). The ideal risk has

theoretical importance, being a term of comparison for the optimality of estimation algorithms.

Since the signal is assumed sparse, there are only a few important features at detail level, the rest of the wavelet coefficients being zero or very close to zero. This motivates the introduction of a further restriction: $a_{j,k} \in \{0, 1\}$.

The nonlinear projector that minimizes (5.2.2) consists of the selection of the bases that correlate best with the signal, thus having the most extreme valued wavelet coefficients $|w_{j,k}| = |\langle f, \psi_j, k\rangle|$. This selection shows the signal dependent aspect of the procedure. It can be described for instance via a thresholding defined by:

$$a_{j,k} = \begin{cases} 1, & |w_{j,k}| \geq \sigma \\ 0, & |w_{j,k}| < \sigma. \end{cases} \tag{5.2.3}$$

Basically the thresholding includes the coefficients that exceed the noise level. Again the estimator via $a_{j,k}$ depends on the unknown signal, such that an oracle is necessary to decide which coefficients are above the noise level.

The risk of the oracle projector is

$$R_{NL} = \sum_{j,k} \min\left(w_{j,k}^2, \sigma^2\right)$$

which is of the same order as R_L:

$$R_{NL} \geq R_L \geq \frac{1}{2} R_{NL}$$

since $\min(x, y) \geq \frac{xy}{(x+y)} \geq \frac{1}{2}\min(x, y)$.

Donoho proved that simple estimates of $w_{j,k}$ are remarkably close to R_{NL}. Some of these simple estimates are the subject of Section 5.3 [36].

5.3 Wavelet thresholding

Thresholding means setting to zero the small wavelet coefficients $w_{j,k} \leq T$, for a given threshold T. It is equivalent to ignore the details of the signal at the respective scale and location and describe the signal based only on the coarser (smoother) approximation level. It can be seen as a local smoothing of the signal, when the signal is regular. At detail level only a few coefficients are kept, that are exceeding the amplitude of (the transformed) noise. These coefficients correspond to the sharp signal transitions, assumed to be limited in number and that translate

in high wavelet coefficients in the neighborhood of the transition. Thus the wavelet thresholding represents an *adaptive smoothing* approach, with adaptive bandwidth with respect to the local regularity of the signal. Several problems arise such as

- the exact form of the thresholding function ρ_T essential for the estimation of the wavelet coefficients,
- the automatic choice of the threshold value T,
- if the value T should be unique over all levels or level dependent (T_j),
- if the threshold should be fixed or data-dependent.

In general, thresholding functions ρ are supposed to fulfill the following properties [65]:

- $\rho_T(-y) = \rho_T(y)$
- $\rho_T(y) \leq x$ if $y \geq 0$ (shrinking)
- $x - \rho_T(y) \leq T + b$ if $y \geq 0$ (boundedness)
- $\rho_T(y) = 0$ if $|y| \leq T$.

The two most well-known thresholding functions are the hard and the soft thresholdings, represented in Fig. 5.2. The hard thresholding ρ_H of wavelet coefficients is a "keep-or-kill" rule, according to which only those w_{jk} coefficients are kept that exceed a certain threshold T, while the others are set to 0:

$$\rho_H(y) = \begin{cases} y, & |y| \geq T \\ 0, & |y| < T. \end{cases} \tag{5.3.1}$$

In order to avoid visual artifacts, hard thresholding can be replaced by a continuous function, known as the soft thresholding function, and defined as:

$$\rho_S(y) = \begin{cases} y - T, & y \geq T \\ 0, & |y| < T. \\ y + T, & y \leq -T \end{cases} \tag{5.3.2}$$

The hard thresholding rule solves the following penalized least square problem:

$$\begin{aligned} \rho_H(y) = \rho_H(y, T) &= \text{argmin}_{\widehat{f}}(y - \widehat{f})^2 + T^2 \left\| \widehat{f} \right\|_0 \\ &= \text{argmin}_{\widehat{f}}(y - \widehat{f})^2 + T^2 I_{\widehat{f} \neq 0} \end{aligned} \tag{5.3.3}$$

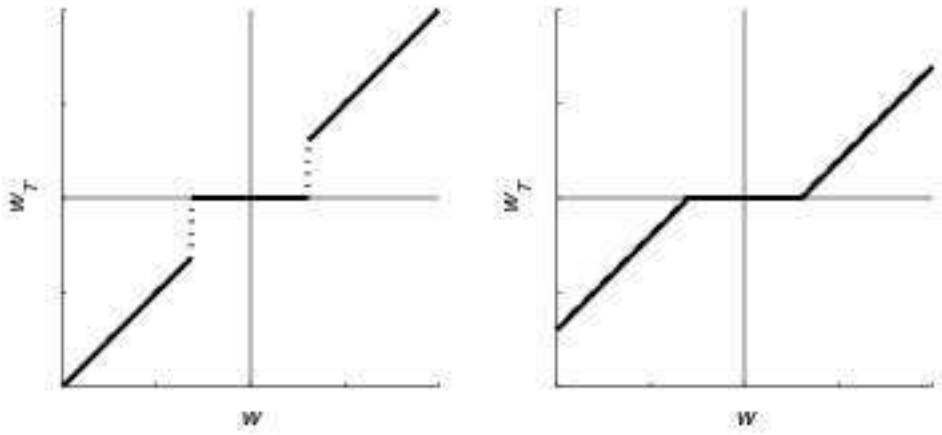


Figure 5.2: Hard (left) and soft(right) thresholding functions.

while the soft thresholding rule is the solution of:

$$\begin{aligned}\rho_S(y) = \rho_S(y,T) &= \operatorname{argmin}_{\widehat{f}}(y-\widehat{f})^2 + 2T\left\|\widehat{f}\right\|_1 \\ &= \operatorname{argmin}_{\widehat{f}}(y-\widehat{f})^2 + 2T|\widehat{f}|. \end{aligned} \tag{5.3.4}$$

Several other thresholding functions were proposed in the literature, (see for example [45, 44, 9]). Their common characteristic is the interval $[-T, T]$ where the value of the wavelet coefficients is set to 0. This is related the property of the wavelet transform to concentrate the energy of a signal in just a few high wavelet coefficients (high in this context meaning exceeding to threshold T in absolute value).

Besides the questions related to the choice of threshold, one has to take into account the adaptation to different noise models, given that the image formation is not always best described by Gaussian noise as in the model (5.1.1) (see the description of image formation and microscopy noise models in Chapter 2). Another factor affecting the behaviour of thresholding algorithms is the amount of

signal (sparsity) discussed in the following.

5.4 Signal sparsity

We have described how the signal can be approximated by a few significant wavelet coefficients, referring to this property as sparsity in wavelet bases. We shall describe in the subsequent the notion of sparsity in more detail.

The notion is related in the case of a wavelet transformed signal to the number of significant wavelet coefficients in the wavelet representation.

More generally, sparsity implies that most of the signal strength is concentrated in a few of the coefficients. The most straightforward examples are the *spike* $(1, 0, ..., 0)$ and the *comb* vectors $(n^{1/2}, ..., n^{1/2})$ [19]. They have the same ℓ_2 norm, however the spike is considerably more sparse than the comb signal. If the two vectors are two representations of the same signal in two different bases and knowing that the noise would normally not correlate with any basis (is not sparse in any basis) the *spike* representation would be easier *denoised*, (the significant coefficients easier found) then the *comb* one.

According to [4], an intuitive definition of sparsity of a signal μ is that it implies a relatively small proportion of nonzero coefficients. For a fixed proportion η and the l_0 quasi-norm $\|x\|_0 = |\{i : x_i \neq 0\}|$ the set with a proportion η of nonzero entries is

$$B_0(\eta) = \{\mu \in \mathbb{R}^n : \|\mu\|_0 \leq \eta n\}.$$

Alternatively, sparsity can be defined on the decreasing rearrangement of the amplitudes:

$$\left|\mu_{(1)}\right| \geq \left|\mu_{(2)}\right| \ldots \geq \left|\mu_{(n)}\right|$$

as a term-wise power-law bound:

$$\left|\mu_{(k)}\right| \leq C \cdot k^{-\beta},\ k = 1, 2 \ldots.$$

Finally, sparsity can be measured using l_p norms, with p small:

$$\|\mu\|_p = \left(\sum_{i=1}^{n} |\mu_i|^p\right)^{1/p}.$$

Strong-l_p balls of small average radius η are defined as:

$$B_p(\eta) = \left\{ \mu \in \mathbb{R}^n : 1/n \sum_{i=1}^{n} |\mu_i|^p \leq \eta^p \right\}$$

In the light of the sparsity definitions above, the sparsity of signal representation in wavelet space is thus related to the l_p norm of the wavelet coefficient vectors, for p small. Thus the regularizations terms (ℓ_0 and ℓ_1 penalties in (5.3.3) and (5.3.4)) corresponding to the soft and hard thresholding justify the preference of nonlinear thresholding estimators over the linear one.

5.5 Threshold selection

Strongly connected to the sparsity of a signal is the selection of the threshold T in the thresholding function $\rho = \rho(\cdot, T)$. For a sparse signal only a few coefficients should be kept, while if the signal is not as sparse more coefficients are expected to appear in the representation. Therefore, the threshold T has to be adjusted according to the (unknown) sparsity of the signal.

In order to perform the selection one can think of at least two general classes of algorithms, generated by two ways of modeling the wavelet coefficients. Both ways are based on a signal-background separation problem formulation based on the assumption that a given data sample $\{x_i,\ i = 1, \dots n\}$ was generated partly by a (possibly very general) signal distribution $P_{\mathcal{S}}\left(\theta_{P_{\mathcal{S}}}\right)$ while the rest of the samples pertain to the background distribution $P_{\mathcal{B}}\left(\theta_{P_{\mathcal{B}}}\right)$, $\theta_{P_{\mathcal{B}}}, \theta_{P_{\mathcal{S}}}$ representing the parameters of $P_{\mathcal{B}}$ and $P_{\mathcal{S}}$. Different views on the modeling of this problem lead to different approaches in setting the right threshold T and detecting the significant coefficients.

Model 1

One possible approach to model the data x_i is by considering it a realization of random variables X_i distributed according to the following two-component mixture:

$$X_i \sim \pi P_{\mathcal{S}}\left(\theta_{P_{\mathcal{S}}}\right) + (1-\pi) P_{\mathcal{B}}\left(\theta_{P_{\mathcal{B}}}\right),$$

where π represents the proportion of signal in the data (and thus reflects the sparsity of the signal). Note that $P_{\mathcal{S}}$ might represent an arbitrary complex distribution, for example it can be at its turn a mixture distribution.

Model 2

The drawback of the previous approach is that a model of the signal has to be specified. An alternative view, especially appropriate when the proportion of the background, $1 - \pi$, is much higher than that of the signal samples, is to consider the signal samples as outliers with respect to the distribution $P_{\mathcal{B}}$, i.e. as data deviating from the data described by $P_{\mathcal{B}}$.

The problem of outlier-detection is closely related to the robust estimation of $\theta_{P_{\mathcal{B}}}$, the parameters of $P_{\mathcal{B}}$, which are usually unknown. Consequently, the two tasks are intertwined: if the parameters are known, the outliers can be identified based on the deviation from $P_{\mathcal{B}}(\widehat{\theta}_{P_{\mathcal{B}}})$, on the other hand, if we would know (and remove) the outliers, the parameters could be correctly estimated based on the remaining data.

Universal threshold

One approach to model the outliers is related to their link to the extremal value of the background distribution. A simple view in one dimension would be that all elements higher than the maximum (smaller than the minimum) of a sequence of i.i.d. random variables distributed according to $P_{\mathcal{B}}$ are considered outliers. It is an approach based on Model 2 described above, since it does not make any assumption neither on the signal nor the signal distribution and was proposed by Donoho and Johnstone in [36] under the name *universal thresholding*.

The universal threshold is defined as

$$T_U = \sqrt{2 \log n}\sigma,$$

where σ is the standard deviation of the distribution of wavelet coefficients (or an appropriate estimate of it) and n is the number of coefficients. It originates from modeling the wavelet coefficients assumed normally distributed, $w_{j,k} \sim \mathcal{N}(0, \sigma)$, given that most coefficients are due to Gaussian noise and only a few are generated by signal.

The threshold T_U is an extreme value of the $w_{j,k}$ and is based on the following property([72]): given a sequence of i.i.d. $Z_1, \ldots, Z_n \sim \mathcal{N}(0, 1)$

$$P\left(\left\{\max_{1 \le i \le n} |Z_i| > \sqrt{2 \log n}\right\}\right) \le \frac{1}{\sqrt{4\pi \log n}} \to 0.$$

However the convergence is slow and the threshold often proves to be too high.

Instead of the unknown parameter σ a robust estimation of the standard deviation is used. More details on the estimate can be found in Appendix A.

Finally, Donoho proved the following optimality result [36].

Theorem 5.5.1. *The risk $R_T(\widehat{f}, f)$ of a hard or soft thresholding estimator with threshold $T_U = \sqrt{2\log n}\sigma$ satisfies for $n \geq 4$*

$$R_T(\widehat{f}, f) \geq (2\log n + 1)\left(\sigma^2 + R_{NL}(f)\right),$$

and is optimal among diagonal estimators D:

$$\lim_{n\to\infty} \inf_{D} \sup_{f\in\mathcal{F}} \frac{\mathbb{E}\left\|f - \widehat{f}\right\|}{\sigma^2 + R_{NL}} \frac{1}{2\log n} = 1.$$

A similar result holds as well in the case of colored noise.

SURE

A data driven, sub-band adaptive threshold is SURE (Stein Unbiased Risk Estimation) proposed in [38].

The equivalent problem to wavelet coefficients estimation can be defined for each wavelet level as a mean vector estimation $(w_1, \ldots, w_n)^T$ (we omit the scale index j to simplify the notation) based on the data $\widetilde{w}_k \sim \mathcal{N}(w_k, 1), \quad k = 1, \ldots, n$ with minimum risk. In case of a sparse data sample, most w_k are thought to be zero. The result of Stein [87] states that $\widehat{\mu}(x) = x + g(x)$ is an unbiased estimator, where $g = (g_1, g_2, \ldots, g_n) : \mathbb{R}^n \to \mathbb{R}^n$ is weakly differentiable:

$$\mathbb{E}\left\|\widehat{\mu}(x) - \mu\right\|^2 = n + \mathbb{E}_\mu\left\{\|g(x)\|^2 + 2\nabla \cdot g(x)\right\}.$$

Using the soft thresholding function one gets

$$g_k(w_k) = \rho_S(w_k) - w_k = \begin{cases} -T, & w_k \geq T \\ -w_k, & -T \leq w_k < T \\ t, & w_k < -T \end{cases}$$

and $\|g(w)\|^2 = \sum_{k=1}^n \min^2(|w_k|, T)$. Also

$$\nabla \cdot g \equiv \sum_{k=1}^n \frac{\partial}{\partial w_k} g_k(w) = -\sum_{k=1}^n \mathbf{1}_{[-T,T]}(w_k)$$

so that Stein's estimate of risk can be written as:

$$\text{SURE}(T, w) = -n + 2 \cdot |\{k : |w_k| > T\}| + \sum_{k=1}^{n} \min^2(|w_k|, T).$$

The threshold is chosen to minimize the estimate of risk:

$$T_0 = \text{argmin}_{T \geq 0} \text{SURE}(T, w).$$

It has been noted that the threshold computed by the SURE criterion is often too low. Donoho and Johnson also proposed a correction for the very sparse signals, when the method is known to have a weaker performance. The correction is based on the use of the universal threshold if the signal is sparse and of the SURE criterion otherwise. Heuristically, the signal is considered sparse if its variance (the mean being assumed 0) satisfies:

$$\sum_{k=1}^{n} w_k^2 \leq 1 + \frac{(\log_2 n)^{3/2}}{\sqrt{n}}.$$

Bayesian thresholding

In a Bayes setting, according to Model 1 above, the wavelet coefficients corresponding to problem (5.1.1) (or more precisely after the DWT to (5.1.3)) follow a prior distributions defined usually as a mixture

$$w_{j,k} \sim (1 - \pi_j)\delta(0) + \pi_j \gamma,$$

where γ is assumed a fixed unimodal symmetric density, e.g. the normal distribution $\mathcal{N}(0, \tau_j)$ as in [25] or a heavy tail distribution, like a double exponential as in [66, 67]. The parameter π_j gives a measure of the sparsity of the signal. The prior can be chosen directly based on available information or can be constructed partially or totally on the data itself, as in the case of the empirical Bayes procedure (EBayes), described below [67].

The noise is considered independent of the wavelet coefficients and normally distributed. Using the notation $g = \gamma * \Phi$, Φ being the standard normal density, the maximum likelihood estimator $\widehat{\pi}_j$ of π_j is obtained as the maximizer of the

	Declared non-significant	Declared significant	Total
True H_0	U	V	m_0
False H_0	T	S	$m - m_0$
	m-R	R	m

Table 5.1: Cases in multiple hypothesis testing

marginal log-likelihood:

$$\ell(\pi) = \sum_{i=1}^{n} \log \left\{ (1 - \pi_j)\Phi(\widetilde{w_{ji}}) + \pi g(\widetilde{w_{ji}}) \right\},$$

subject to the constraint $\pi_j \leq \sqrt{2 \log n}$, the universal threshold representing the largest value obtained from a zero signal and it corresponds to the limit case $\pi_j \to 0$. The estimated value $\widehat{\pi}_j$ is plugged back into the prior and the wavelet coefficients are estimated via a Bayesian procedure as the mean or median of the posterior distribution (see [66, 67] for computational details).

False Discovery Rate

The task of significant wavelet coefficient selection can be reformulated from a multiple hypothesis testing point of view: to each wavelet coefficient of the true, unknown function f corresponds the null hypothesis $H_{jk} : w_{jk} = 0$, representing the case that the wavelet coefficients corresponds exclusively to noise, to the no signal case (m such hypotheses in total).

We expect that most of the empirical wavelet coefficient values are small, the variation around 0 due only to noise. Ideally, only the signal coefficients should be kept in the reconstruction meaning that only in these cases H_{jk} should be rejected. This setting corresponds to Model 2 of wavelet coefficients described above.

Suppose that H_{jk} is rejected or accepted based on the value of a test statistic $\mathcal{T}_{jk}(w_{jk})$, more specifically H_{jk} is rejected if $\mathcal{T}_{jk}(w_{jk}) < c_{jk}$. The p-value related to a test is defined as the probability of having a larger value than a fixed p given that the null hypothesis H_0 is true: $P(t > p|H_0)$.

Let R denote the number of coefficients kept in the representation out of which S are correctly and V are erroneously kept, $R = V + S$. V represent the false positives or false discoveries (also known as type I errors). The false negatives (type II errors) are denoted by T. These quantities are summarized in Table 5.1. Of all the variables that appear in Table 5.1, only m and R are observed.

The importance of controlling the false positives is convincingly motivated

through the humorous but noteworthy *False Positive Rules* of Young [115]:

- With enough testing, false positives will occur
- Internal evidence will not contradict a false positive result
- Good investigators will come up with a possible explanation
- It only happens to the other person.

Definition 5.5.2. The *Family-Wise Error Rate* (FWER) is the probability of making any false positive call $P(V \leq 1)$, at the desired significance level α.

An approach based on controlling the FWER is the Bonferroni approach and consists of adjusting the p-values by the number of comparisons (or tests) m: $\widetilde{p}_j = \min(1, m \cdot p_j)$ and reject the hypothesis for which the adjusted p-values are smaller than desired significance level α.

The approach is often too conservative and lacks enough power to detect significant differences (produces too many false negatives). A better power can be obtained by the Sidak adjusted p-values $\widetilde{p}_j = 1 - (1 - p_j)^m$ that control FWER when the comparisons are independent.

Definition 5.5.3. The *False Discovery Rate* (FDR) is defined as the expectation of Q, the expected proportion of erroneously kept coefficients among all the coefficients kept in the representation:

$$\text{FDR} = \mathbb{E}(Q), \quad Q = V/R.$$

The random variable Q, represents the proportion of (false positives) coefficients, that were kept although they should have been dropped, and it is an error measure of this procedure. Naturally, when $R = 0$ also Q is set to 0.

The Benjamini-Hochberg method described in [11] was applied to wavelet thresholding (see [3, 4]) and maximizes the number of kept coefficients, controlling meanwhile the FDR to a level q.

Benjamini-Hochberg algorithm

For each wavelet detail level j:

1. For each $\widetilde{w}_{jk}$ calculate the two-sided p-value:

$$p_{jk} = 2\left(1 - \Phi\left(\left|\widetilde{w}_{jk}\right| / \sigma\right)\right),$$

where Φ is the standard normal distribution function.

2. Order ascendingly the computed p_{jk}-s,

$$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}.$$

3. Find the largest i such that $p_{(i)} \leq \frac{i}{m} q$ and denote it i_0.
Compute $\lambda_{i_0} = \sigma \Phi^{-1}(1 - p_{(i_0)}/2)$. The true value σ can be replaced by an estimate of the wavelet coefficients' variance.

4. The estimates $\widehat{w}_{jk}$ are obtained by thresholding all coefficients $\widetilde{w}_{jk}$ at level λ_{i_0}.

Note that if the data is pure noise the control of the FDR is equivalent to the Bonferroni approach. If many true coefficients are present, R tends to be large, and FDR will be smaller, so that the error rate is adaptive with respect to the sparsity of the estimated function.

It was proved in [11] that the above procedure controls the FDR at the (unknown) level $\frac{m_0}{m} q \leq q$. In the case of non-Gaussian noise one should replace Φ with F the c.d.f. of the noise model.

Variance - covariance estimation

As we have seen most of the wavelet threshold estimates rely on the variance of the noise distribution (or wavelet transformed noise at the respective scale). In the FDR approach the variance appears in the formulation of the null-hypotheses and the computation of p-values. Most conveniently, we shall use an estimate of the variance, computed from the data. However, since the data is a mixture of noise and true signal, we have to strive to compute the estimate of the noise variance based only on true noise values, a typical problem from the field of robust statistics [52]. For the robust estimation of scale we adopt, as description of the data, the Model 2 above, assuming that the true signal coefficients are a small fraction of all the coefficients and can be regarded as outliers.

For the one dimensional case the most frequently used scale estimator is $\text{MAD}(x) = \text{median}\,|x - \text{median}(x)|\,/0.674$ (the estimator is described in Appendix A). For higher dimensional problems, a state of the art methods is the Minimum Covariance Determinant (MCD) estimate. Based on the MCD estimate of the covariance, the *Mahalanobis distances* (MD) of wavelet coefficients (seen as

tuples) are computed, and used in an alternative algorithm for detection of significant wavelet coefficients as proposed in Section 5.8 below. Several distribution can be used to approximate the distribution of MD in order to construct the H_0 hypothesis of the FDR approach, such as $\chi^2, \mathcal{F}$ and $\mathcal{B}eta$ (all details can be found in Appendix A).

We shall incorporate the variance and covariance estimates described in the Appendix A found at the end of this work in the wavelet thresholding algorithms used for signal reconstruction in general, and in our particular case, for single molecule detection.

5.6 Other noise models

Wavelet methods are typically designed for additive Gaussian noise: $X_i = \mu_i + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma)$. In the case of microscopy imaging, the three most frequent models based on image formation are- in order of model accuracy but also increasing difficulty/complexity:

- **M1: Gaussian noise**

$$y_i = \mu_i + \varepsilon_i,\ \varepsilon \sim \mathcal{N}(0, \sigma)$$

- **M2: Poisson noise**

$$y_i = x_i,\ x_i \sim \mathcal{P}oi(\mu_i)$$

- **M3: Poisson and Gaussian noise**

$$y_i = x_i + \varepsilon_i,\ x_i \sim \mathcal{P}oi(\theta_i), \varepsilon_i \sim \mathcal{N}(\mu, \sigma) \tag{5.6.1}$$

Especially low intensities (small photon counts) collected by the sensor are not well modeled by Gaussian noise (M1).

A combination of Poisson (shot-noise) and Gaussian noise is more appropriate to describe photon count variations and read-out noise. The main difference is the heteroscedasticity of models M2 and M3 as opposed to M1, the variance of the noise being dependent on the signal.

Several methods were devised to handle models M2 and M3. The most straightforward is to consider their simple approximation via a Gaussian model, given that

asymptotic normality in terms of probabilities of large deviations often holds:

$$\frac{P\left(\pm(w_{j,k}-\theta_{j,k})/\sigma_{j,k}\geq x\right)}{1-\Phi(x)}\to 1, \text{as } n\to\infty.$$

A different way to take into account the heteroscedasticity of the noise are variance stabilizing transforms, applied prior to wavelet detection to the input image which transform the heteroscedastic noise into Gaussian noise of variance approximately equal to one.

In case of a Poisson noise model M2 (suitable to describe the photon count model) the best known variance stabilization is the *Anscombe transform* [8]:

$$t(X_i)=2\sqrt{X_i+3/8}.$$

However it underestimates the low intensity values, especially those with pixel intensity under 30. Modeling both the photon count noise as well as the read-out noise, one obtains the mixed Poisson-Gaussian image model M3. $X_i=\alpha\cdot N_i+\varepsilon_i$, where $\alpha>0$ represents the gain of the detector, $N_i\sim Poi(\mu_i)$ and $\varepsilon_i\sim\mathcal{N}(0,\sigma)$.

A more elaborate transform is the *Generalized Anscombe transform (GAT)* introduced in [107], that can stabilize the Poisson - Gaussian noise: $t_G(x)=\frac{2}{\alpha}\sqrt{\alpha x+\frac{3}{8}\alpha^2+\sigma^2-\alpha\mu}$. The parameters α represent the gain of the detector, μ and σ are the parameters of the Gaussian noise as in (5.6.1) and all three parameters can be determined from the image itself via robust fitting as described in [17].

5.7 Single molecule detection and evaluation of the detection method

As opposed to the denoising and restoration problems to which wavelet thresholding usually is applied, we are not interested in the original true value of each pixel in the image, but on the detection of single molecules. We shall define the task as the detection of significant pixels in the first J scales of the wavelet transform.

Through simulations and experiments on real images we have found that the choice $J=3$ performs well for a wide range of SNR, signal intensity and sparsity. The significant pixels were found via FDR thresholding, and for the null hypothesis the Gaussian model was assumed, with expected value 0 with the MAD estimation of its variance.

The significant pixels, those belonging to true signal are those that have non-zero coefficients in all the J detail levels (except the finest, which usually is pure noise). The binary image obtained from the J detail coefficient levels after level-wise wavelet thresholding and combination of each level via a logical AND operation is an indicator image for the support of the detected single molecules:

$$B = \prod_{j=j_0}^{J} \mathbf{1}_{\{w_{jk} > \lambda_{i_0}(j)\}}.$$

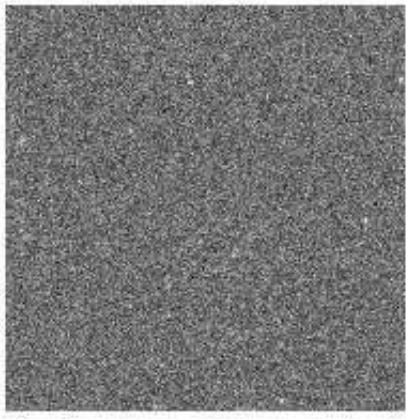
If the first detail level is too noisy, one can set $j_0 = 2$.

At very low concentrations the single molecules are mostly well separated of each other. At higher concentrations, the support of imaged single molecules can be touching each other, so that the number of connected components in the binary image underestimates (strongly in case of high concentration) the number of single molecules in the image.

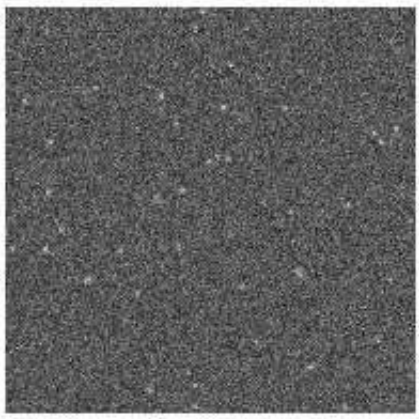
In order to alleviate the problem, a denoised image is additionally obtained after applying the reconstruction step (4.7.3) with thresholded detail coefficients. Since the wavelet detection algorithm has the "resolution" described above, two molecules that are spatially close together will be detected as one. The correction of the estimation of the number of molecules consists in combining the binary image obtained after the detection step with the denoised image (via a simple product operation), and all the local maxima of the denoised image inside the support of the binary mask are considered distinct single molecules (see for instance Fig. 3.7, c).

The detection algorithm was tested on a set of simulation images with varied image quality parameters, as measured by the signal-to-noise ratio (SNR), as well as several molecule concentrations. Each image is of dimension 512×512 pixels and contains $10, 50, 100, 500$ or 1000 randomly placed molecules.

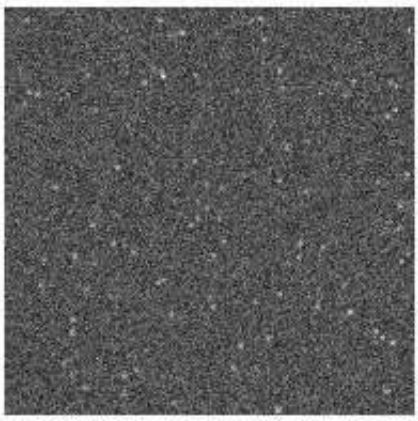
To each single molecule corresponds a diffraction limited spot, approximated by a two-dimensional Gaussian shape, with width s corresponding to the point spread function of the optical system (1.1 in our simulations). Both the constant background intensity and the peak intensity S were chosen on a logarithmic scale between 10 and 100. Noise is generated for each pixel as described in (5.6): the photon count noise was modeled by draws from Poisson distributions, and finally Gaussian noise is added to each pixel from $\mathcal{N}(0, \sigma)$, where σ was chosen as $0, 5, 10, 15$ and 20% of the maximum peak intensity. For this special case of Poisson-Gaussian model described, we use the following signal-to-noise (SNR) def-



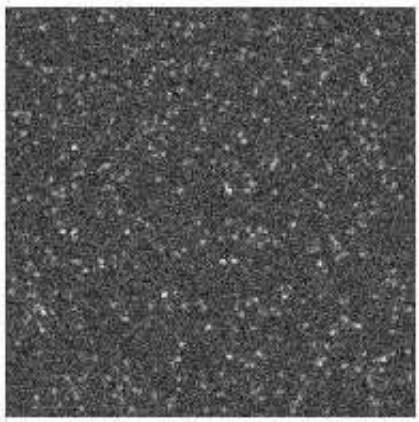
(a) $N = 10$



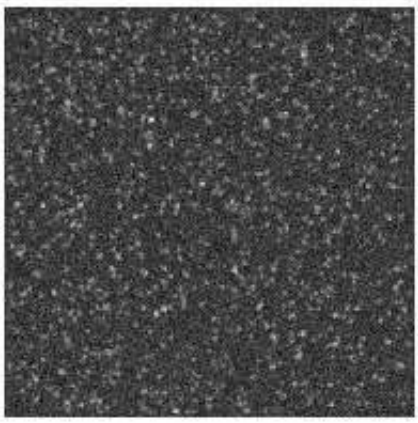
(b) $N = 50$



(c) $N = 100$



(d) $N = 500$



(e) $N = 1000$

Figure 5.3: A set of simulation images is shown in (a)-(e) for different concentrations: N represents the number of peaks in the 512×512 pixel image (corresponding to $102.4\mu m \times 102.4\mu m$). SNR = 5.02 (additional Gaussian noise with $\sigma = 2.2$). The images are scaled for better visibility. The same pixel intensity scale is used for the five images.

inition as described in Chapter 2:

$$\text{SNR} = \frac{S}{\sqrt{B + \sigma^2}},$$

where S represents the single molecule intensity, B the (local) background of the image and σ the standard deviation of the read-out noise. (For simplicity S is the maximum intensity of the single molecule profile).

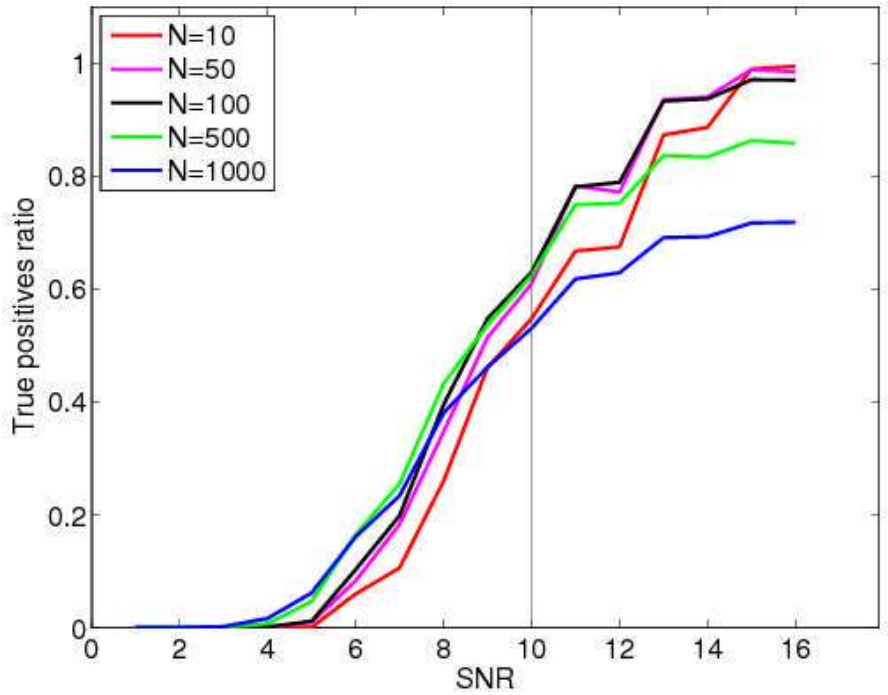
The SNR for our simulations is between 0.9 and 31.6.. For each set of parameters 10 images were generated and analyzed. One set of simulations is presented in Fig. 5.3 $(a) - (e)$.

The results are summarized in Fig. 5.4 (a) and (b). For SNR above 10 the detection for all five concentrations levels exceeds 80% true positives for less than 500 molecules, but is only above 60% for high concentrations ($N = 1000$). However, at high concentrations the spots can be analyzed also with conventional methods designed for low resolution microarrays. The SNR typical for our system, is usually at least 15, (at this level true positives are higher than 85% for $N = 500$, and 70% for$N = 1000$). The rate of false positives is under 9% even for very low concentrations ($N = 10$) and substantially lower for $N > 100$. The detection performance was similar for simulation with the same SNR, independent of the weight of Poisson and Gaussian noise in the generation of the simualtion image.

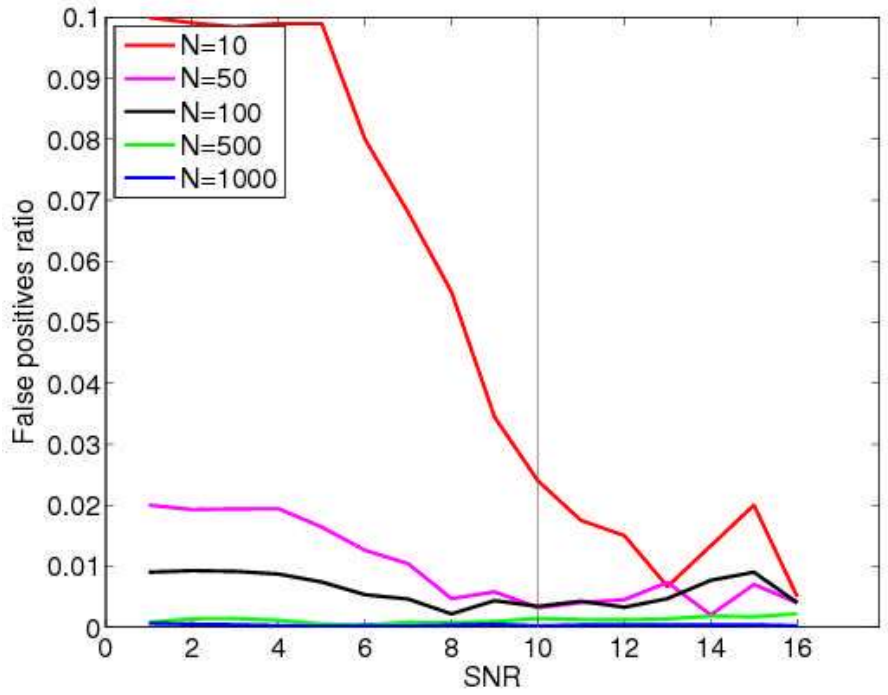
5.8 Signal detection via robust distance thresholding

In the case of orthogonal wavelet transforms, the independent thresholding of wavelet coefficients at different scales is the natural approach. However, in the case of the redundant *à trous transform* the wavelet coefficients are still correlated across scales, the gain of translation invariance being made at the price of losing orthogonality. We shall investigate the effect of the correlation between wavelet scales on the detection task. As alternative to the scale-wise thresholding of the wavelet coefficients we suggest the modeling of the null-hypothesis in a higher dimensional space. In case we are using J wavelet detail levels, H_0 will be modeled in a J-dimensional space, taking into account the correlations between detail levels.

If $\{w_{ij},\ j = 1, \dots J\}$ are the wavelet coefficients obtained via the wavelet transform of the image, we assume that H_0 is described by a J-dimensional Gaussian $\mathcal{N}(T, S)$, where T and S are the MCD robust estimates described in Appendix A.



(a) True positive ratios (5% Gaussian noise added)



(b) False positive ratios (5% Gaussian noise added)

Figure 5.4: The results of detection on the simulations are summarized in figures: (a) ratio of true positives and (b)ratio of false negatives with respect to the true number of simulated single molecules

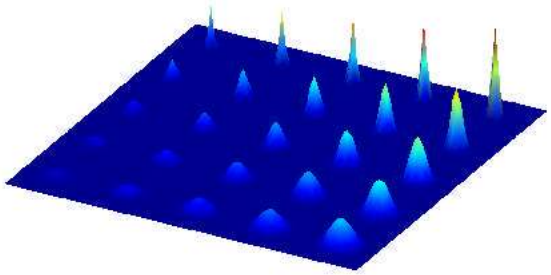


Figure 5.5: Test pattern for detection (testing several intensities and several stricture sizes)

The signal detection is performed in one step in the following way: the one dimensional data representing the p-values of the robust distances, r_i, are computed and subsequently thresholded, via the FDR approach at a fixed significance level q assuming that r_i are distributed according to $r_i \sim \chi^2_d$, as described in Section 5.5.

The dependence of wavelet coefficients was discussed also in [27], in particular the dependence between two successive wavelet levels. The solution proposed in [27] is based on Bayesian modeling, MAP estimators and bivariate shrinkage functions.

As test case an image of 25 Gaussian shapes was used, the Gaussian having total intensities ranging on a logarithmic scale from 100 to 1000 and widths $\sigma \in \{0.5, 1, 1.5, 2, 2.5\}$. The original test pattern without noise is shown in Fig. 5.5. The intensity of the shapes increases from left to right and each line is characterized by the same spread σ of the Gaussian shape. The correlation of successive wavelet coefficients can be observed in Fig. 5.6. The first scale is mostly noise, and is uncorrelated with the higher scales, however, one can notice a correlation between successive scales, starting from the second level (second line and/or column) onwards.

The Fig. 5.7 presents the detection results in case of the test image corrupted by Poisson noise and additive Gaussian noise of parameter $\mathcal{N}(10, 3)$ (as described in Section 5.7). The detection by classical scale-wise thresholding (upper right image) detects 20 out of 25 structures, and shows little difference in size of the detected structures (all detected blobs have similar areas, although five different

σ parameter values were used in the simulation). However the performance of the method both in accuracy and computation time is remarkable.

The Mahalanobis distance based detection with standard estimates has a very low detection performance, only 13 structures being detected (lower left image in Fig. 5.7).

The robust Mahalanobis distance thresholding, RMDT, (lower right image), detects 23 of the 25 structures and shows relatively good sensitivity to the parameter σ of the detected structures. Its drawback consists of detecting one small false positive structure as opposed to none in the scale-wise thresholding case.

A further advantage of RMDT is that it is more robust to the number of levels J selected in the wavelet transform than the scale-wise thresholding detection. The strong constraint that a pixel is considered signal only if the corresponding wavelet coefficients are outliers in all wavelet scales results in several pixels not being detected as signal.

For the Poisson-Gaussian model with no background there are no significant difference between several models for the distance: $\chi^2, \mathcal{B}eta, \mathcal{F}$ as can be seen in

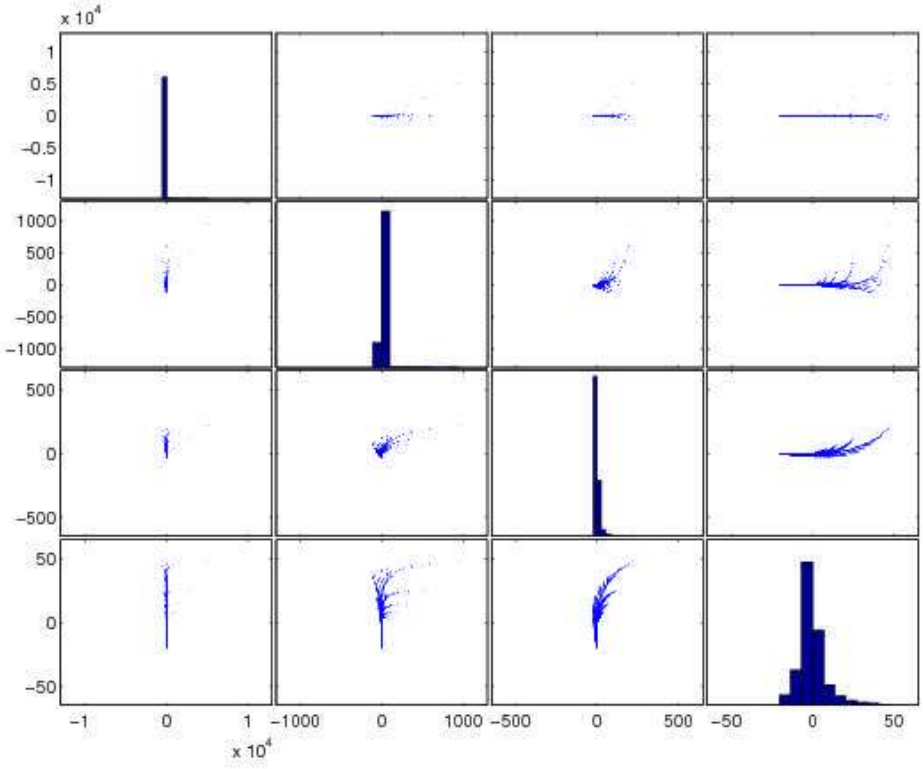


Figure 5.6: Pairwise wavelet coefficients correlated across the scales.

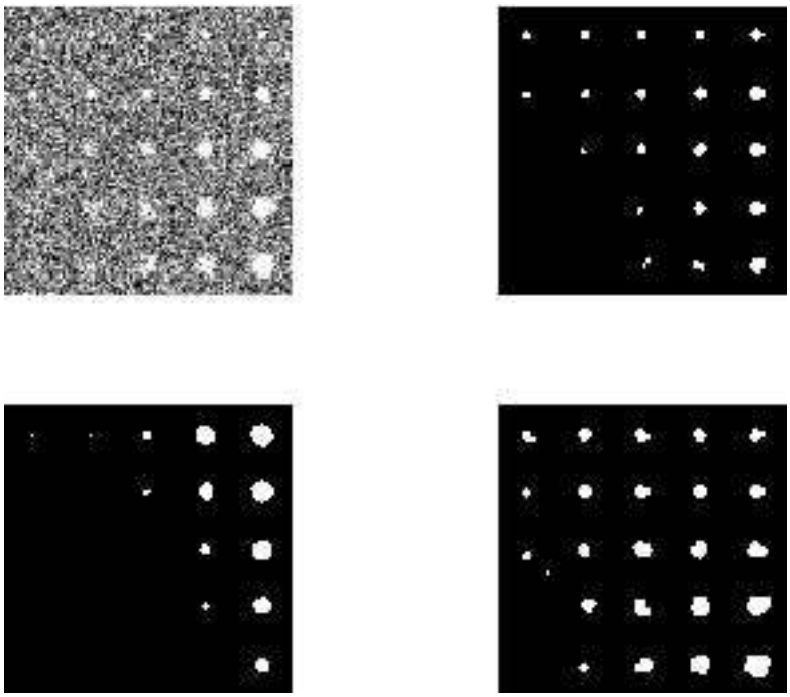


Figure 5.7: Wavelet based detection. Upper left: original noisy test image, scaled for better visibility. Upper right: the support of the signal detected via scale-wise thresholding. Lower left: thresholding based on Mahalanobis distance with standard estimates. Lower Right: thresholding based on Mahalanobis distance with MCD estimates assuming a χ^2 distribution. (All thresholds are based on control of FDR).

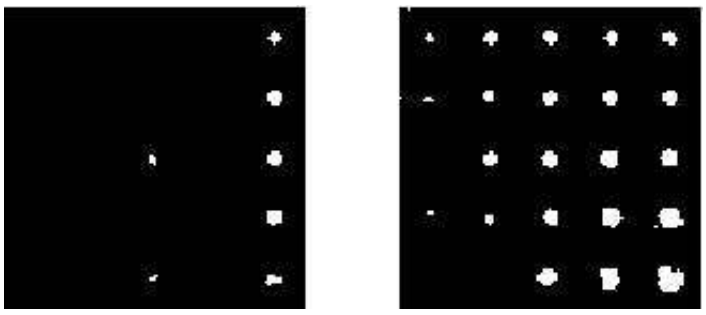


Figure 5.8: Detection based on 5 scales. Left: scale-wise thresholding. Right: RMDT.

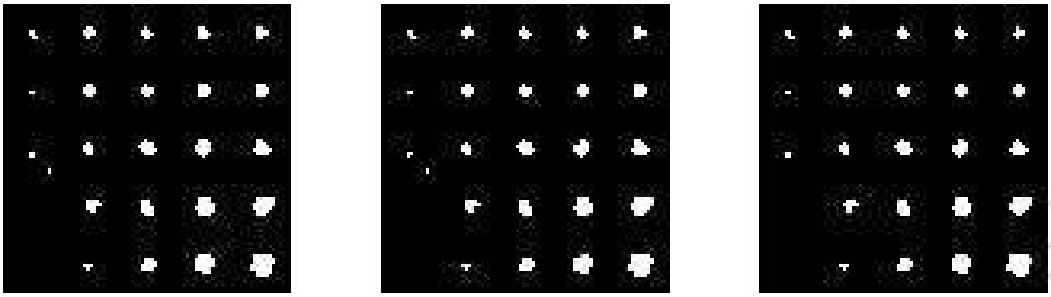


Figure 5.9: Wavelet based detection. The robust Mahalanobis distance distribution is modeled as (left) χ^2 distributed, (middle) $\mathcal{B}eta$ distributed (right) $\mathcal{F}$ distributed. No significant difference in detection is observed.

Fig. 5.9.

Chapter 6

Spatial patterns

After detecting the approximate position of spots and localizing the single molecule peaks in the detected region it is necessary to refine the analysis in order to remove single peaks outside the printed spot. Typically peaks representing non-specific binding, artifacts, dirt, etc. outside the spot area have to be separated from hybridized peaks. This step corresponds to the background/foreground pixel separation in the classical microarray image analysis.

The main assumption for this step of the analysis is that the concentration of false peaks is different - typically lower - from the concentration of the peaks inside the spot. For this reason, the signal and the background are jointly modeled as a mixture of two (spatial) point processes.

This chapter presents a brief introduction in the field of spatial point patterns, in order to offer a perspective of modeling single molecules as point patterns. We will focus on two models, the two-dimensional stationary Poisson process and the mixture of stationary Poisson processes. The latter will be used in separating hybridized single molecules from clutter.

Furthermore we describe and compare solutions for two problems: the complete randomness test and concentration estimation. We apply the described algorithms to prove the soundness of randomness assumption in the case of single molecule microarray images, as well as concentration estimation at the single molecule level, which will give the measure of hybridization for each spot in the high resolution microarray.

6.1 Introduction to spatial point processes

Intuitively, *point processes* are stochastic models for finite or countable collection of points X lying in some set $S \subseteq \mathbb{R}^d, d \geq 1$. A *point pattern* is typically interpreted as a realization of a point process.

The area where points of the pattern can possibly be observed is called *observation window*, denoted usually by W. When the point pattern is restricted to W ($W = S$), we talk about *finite point patterns*. More frequently though W is strictly contained in S ($W \subset S$). Since there is no information for the unobserved region $\mathbb{R}^d \backslash W$, in the case of certain statistical summaries one has to handle the problems raised by the missing data that could affect these summaries. This problem is known under the name of *boundary or edge effects*.

We use the notations: $X_A = X \cap A$ and $X(A)$ is the cardinality of X_A, where $A \subset \mathbb{R}^d$ is a bounded Borel set.

Definition 6.1.1. A point process defined on $S \subset \mathbb{R}^d$ is a random variable X taking values in a measurable space $[\mathbf{N}, \mathcal{N}]$, where $\mathbf{N}$ is the family of all locally finite point configurations in S:

$$\mathbf{N} = \{x \subset S : x(A) < \infty, \forall A \in \mathcal{B}_0\}$$

and $\mathcal{B}_0$ denotes the class of bounded Borel sets in S and such that the points have single multiplicity ($x_i \neq x_j$ if $i \neq j$.) $\mathcal{N}$ is the smallest σ-algebra generated by all sets of the form $\{x \in \mathbf{N} : x(A) = m\}$, with $A \in \mathcal{B}_0$ and $m \in \mathbb{N}_0$. X can be regarded as a measure on a probability field $(\Omega, \mathcal{F}, \mathbb{P})$ that makes all mappings $X(A) : \Omega \to \mathbb{N}$ measurable.

Intuitively, a point process, denoted by X has two interpretations:

1. a random counting measure: $X(B) = n$ meaning that B contains n points of N.

2. a random closed set $X = \{x_1, x_2, \ldots\} = \{x_i\}$, or $x_i \in N$

The point process X generates a distribution P on $[\mathbf{N}, \mathcal{N}]$. The distribution P is determined by the probabilities $P(Y) = P(X \in Y)$. Several good introductions in the theory of points process are available, e.g. [109, 29, 82, 60].

Two important properties of point processes are stationarity and isotropy.

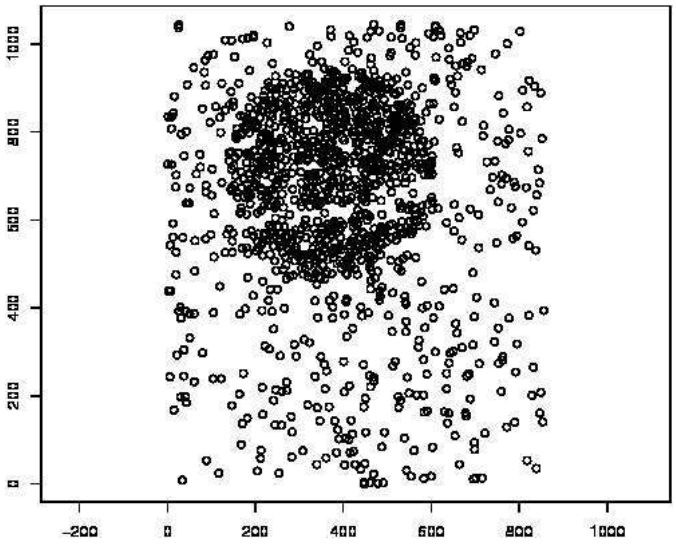


Figure 6.1: Point pattern corresponding to a high resolution microarray area

Definition 6.1.2. A point process X is said to be *stationary* if its statistics are invariant under translation, i.e. the processes $X = x_i$ and $X_x = x_i + x$ have the same distribution for all $x \in \mathbb{R}^d$.

Definition 6.1.3. A point process X is said to be *isotropic* if its statistics are invariant under rotation, i.e. the processes X and rX have the same distribution for all rotations r around the origin.

Moments of point processes

The *intensity measure* of a point process N is the analogue of the first moment of a real-valued random variable and is defined as:

$$\Lambda(B) = \mathbb{E}N(B),\ B \subseteq \mathbb{R}^d.$$

The *intensity function* $\lambda(x)$ models the mean structure of the process, such that if x is an infinitesimally small set of $\mathbb{R}^d$ having area dx then $\mathbb{E}\left(N(dx)\right) \approx P\left(N(dx) = 1\right) \approx \lambda(x)\nu_d(\,dx)$. It follows that

$$\mu(B) = \Lambda(B) = \int_B \lambda(x)\,dx.$$

The *(Papangelou) conditional intensity* is a function $\lambda(u|\mathcal{X})$ of a spatial location $u \in W$ and the entire pattern $\mathcal{X}$. Considering an infinitesimal region of area du around u, $\lambda(u|\mathcal{X})\,du$ is the conditional probability that the point process

contains a point in this region, given the position of all the points in the process outside of the infinitesimal region:

$$\mathbb{E}\left(N(\,dx) = 1 | N \cap\, dx^C = \mathcal{X}\right) \approx \lambda(x|\mathcal{X})\nu(\,dx).$$

The stationary Poisson process discussed below has the conditional intensity $\lambda(u|\mathcal{X}) = \beta$, where β is the intensity of the process.

Higher order moments

Computations of expected values, usually are simplified when the *Campbell theorem* is applied (see [109]).

Theorem 6.1.4. *If f is a non-negative measurable function and $S = \sum_{x\in X} f(x)$ then*

$$\mathbb{E}(S) = E\left(\sum_{x\in X} f(x)\right) = \int_{\mathbb{R}^d} f(x)\lambda(x)\,dx. \tag{6.1.1}$$

Generalizing the first moment measure, one obtains the nth-order moment measures $\mu^{(n)}$ of the point process X defined by

$$\int f(x_1,\ldots,x_n)\mu^{(n)}(\,d(x_1,\ldots,x_n)) = \mathbb{E}\left(\sum_{x_1,\ldots,x_n\in X} f(x_1,\ldots,x_n)\right) \tag{6.1.2}$$

where $f(x_1,\ldots,x_n)$ is any non-negative measurable function on $\mathbb{R}^{nd}$.

The following equalities hold

$$\mu^{(n)}(B_1\times\ldots\times B_n) = \mathbb{E}\left(X(B_1)\ldots X(B_n)\right))$$

and in case $B_1 = \ldots = B_n = B$, $\mu^{(n)}(B^n) = \mathbb{E}X(B)^n$.

In particular:

$$\begin{aligned} &n = 1:\ \mu^{(1)}(B) = \mathbb{E}X(B) = \Lambda(B)\\ &n = 2:\ \mu^{(2)}(B_1\times B_2) = \mathbb{E}X(B_1)X(B_2)\end{aligned}$$

$$\begin{aligned} &\operatorname{var}\left(X(B)\right) = \mu^{(2)}(B\times B) - \left(\Lambda(B)\right)^2\\ &\operatorname{cov}\left(X(B_1), X(B_2)\right) = \mu^{(2)}(B_1\times B_2) - \Lambda(B_1)\Lambda(B_2)\end{aligned}$$

Factorial moments

The nth order *factorial moment* measure $\alpha^{(n)}$ of the point process X is defined by:

$$\int f(x_1, \ldots, x_n)\alpha^{(n)}(\, d(x_1, \ldots, x_n)) = \mathbb{E} \sum_{x_1 \neq x_2 \neq \ldots \neq x_n \in X} f\,(x_1, \ldots, x_n)\,,$$

where f is a non-negative measurable function on $\mathbb{R}^{nd}$.

For $n = 2$, $\mu^{(2)}(B_1 \times B_2) = \Lambda(B_1 \cap B_2) + \alpha^{(2)}(B_1 \times B_2)$. If $B_1, \ldots, B_n$ are pairwise disjoint, then

$$\mu^{(n)}(B_1 \times \ldots \times B_n) = \alpha^{(n)}(B_1 \times \ldots \times B_n).$$

The name *factorial moment* is justified by the following equality:

$$\alpha^{(n)}(B^n) = \mathbb{E}\left[X(B)(X(B) - 1)\ldots(X(B) - n + 1)\right],$$

such that $\alpha^n(B^n)$ is the factorial moment of $X(B)$.

Point process operations

Many point process models can be derived from simpler models by applying one or several operations, such as thinning, superposition, clustering.

A *thinning* operation consists of removal of certain points of a point process N_0, yielding the thinned point process $N \in N_0$. The removal of a point x can be done according to a constant probability p, a location dependent probability $p(x)$ or a random probability. We talk of independent thinnings if removal of points is done independently of other points, and of dependent thinnings, when the removal depends on the configuration of N_0.

The intensity function of a $p(x)$-thinned process with original process intensity λ_0 is

$$\lambda(x) = p(x)\lambda_0(x),\ x \in \mathbb{R}^d.$$

In a *superposition* operation two or more point processes are superimposed onto each other, such that the resulting process is the set-theoretic union of the operand processes: $N = N_1 \cup N_2 \ldots \cup N_k$. The intensity of the resulting process is the sum of the intensities of component processes $\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_k$.

Finally, *clustering* is the operation through which points x of the process N_0

are replaced by clusters N^x of points. The union of the clusters defines the cluster point process

$$N = \cup_{x \in N_0} N^x.$$

A special case of clustering is when all clusters have the same distribution, with $\bar{c}$ the mean number of points in each cluster. In this case the intensity of the cluster process is

$$\lambda = \bar{c}\lambda_0. \tag{6.1.3}$$

Point process models

With a few simple generating mechanisms and the operation described above several point process models can be described. The simplest and best studied point process is the (homogeneous) Poisson point process.

Spatial Poisson process

Definition 6.1.5. A *Poisson process* N defined on S and with intensity measure Λ and intensity function λ satisfies for any bounded region $B \subseteq S$ with $\Lambda(B) > 0$:

Poisson distribution of point counts: for all bounded Borel sets B:

$$P\left(N(B) = m\right) = \left[\Lambda(B)\right]^m \cdot \exp\left(-\Lambda(B)\right)/m!,\ m = 0, 1, 2, \ldots$$

Independent scattering property: the number of points of N in k disjoint sets form k independent random variables.

The homogeneous Poisson process is a special case, such that $\Lambda(B) = \lambda\nu(B)$, for a constant $\lambda > 0$, which also implies that the process is isotropic. The homogeneous Poisson process represents a model for *no interaction* or *complete spatial randomness* (CSR) since for any disjoint subsets $A, B \subseteq S$, $N(A)$ and $N(B)$ are independent.

In the case of an inhomogeneous Poisson process, the intensity is not proportional to Lebesgue measure, but is a deterministic function of spatial location. If a inhomogeneous Poisson process has random mean measures we talk of doubly stochastic processes or *Cox processes*. For instance a stationary Poisson process with randomized intensity parameter is a particular Cox process, called *mixed Poisson process*.

As an exemplification of clustering processes we shall define Neymann-Scott point processes and in order to illustrate processes that exhibit regularity we introduce the hard-core models.

Neymann-Scott processes are the result of homogeneous independent clustering applied to stationary point processes. The points of the parent process form a stationary Poisson process of intensity λ_0 and are not observable. Each parent point x generates a cluster N^x of average size $\bar{c}$, independently and identically distributed around the parent point. The process has the intensity given in (6.1.3). If N^x is a cluster of independently and uniformly distributed points in a ball $b(x, R)$ the point process is called *Matérn cluster process*, while if the daughter points follow a normal distribution around the parent point the process is called (modified) *Thomas process* (see [60]).

The hard core models belong to the class of point processes that model inhibition or regularity and are characterized by a minimal distance r_0 between their points. No points may be located at distance smaller than r_0 of each other.

The simplest model is the *Matérn hard-core process I*, N_I, obtained from a homogeneous Poisson process N_0 with intensity λ_0 by deleting all pairs of events of N_0 that are separated by a distance smaller than r_0. The probability that a point is retained in N_I is $\exp\left\{-\lambda_0 V_d r_0^d\right\}$ and the intensity of N_I is

$$\lambda = \lambda_0 \exp\left\{-\lambda_0 V_d r_0^d\right\},$$

where $V_d = \pi^{d/2}/\Gamma(1 + d/2)$ is the volume of the unit sphere. If the initial process N_0 is endowed with independent marks $Z(x)$, the *model II of Matérn hard-core process* N_{II} is defined as the process obtained after the deletion of points u, such that $\|u - x\| < r_0$ and $Z(u) < Z(x)$. The probability that an arbitrary point is retained in N_{II} is $\left\{1 - \exp\left(-\lambda_0 V_d r_0^d\right)\right\} / \left\{\lambda_0 V_d r_0^d\right\}$ and the intensity of N_{II} is

$$\lambda = \frac{1 - \exp\left(-\lambda_0 V_d r_0^d\right)}{V_d r_0^d}.$$

Several other point process models are known in the literature, many of which are described in [91, 109, 26, 60].

6.2 Summary characteristics for (stationary) point processes

In order to make inferences on the random mechanisms that generate the point processes or test the assumed models, summary characteristics of the point patterns are computed and analyzed. In the following we shall discuss first and second order summary characteristics of point patterns (as well as their estimation). Further details on the properties and estimation of these characteristics as well as other (e.g. directional) characteristics can be found in [26, 60, 109].

The *spherical contact distribution function* (empty-space distribution) is based on the probability that a disk or (hyper)sphere of radius r does not contain any point of N. It is defined as a function of r:

$$H_{s,x}(r) = 1 - \mathrm{P}\left(N\left(b(x,r)\right) = 0\right),\ r \geq 0.$$

If the process is stationary: $H_s(r) = 1 - \mathrm{P}\left(N\left(b(0,r)\right) = 0\right),\ r \geq 0.$

The nearest-neighbour distance distribution function $D(r)$ (denoted also by $G(r)$ in the literature) is the random distance from the typical point to its nearest neighbour. In the context of the Palm distribution (conditioning on the typical point of the process):

$$D(r) = \mathrm{P}_o\left(N\left(B(o,r)\backslash\{o\}\right)\right) > 0,\ r \geq 0.$$

The nearest-neighbour captures the behaviour of the process only at relatively small range. Generalizations such as kth nearest-neighbour distance distributions may be seen as a remedy to this shortcoming:

$$D_K(r) = \mathrm{P}_o\left(N\left(B(o,r)\backslash\{o\}\right) \geq K\right),\ r \geq 0.$$

In the case of Poisson processes $D(r) = H_s(r),\ r \geq 0$. If the process is regular, $D(r) \leq H_s(r),\ r \geq 0$, while for cluster processes the reverse inequality holds, $D(r) \geq H_s(r),\ r \geq 0$, since nearest inter-point distances tend to be distances between points in the same cluster.

Based on the observation above, the following function provides information on the nature of the studied process:

$$J(r) = \frac{1 - D(r)}{1 - H_s(r)},\ r \geq 0 \text{ and } H_s(r) < 1.$$

For Poisson processes $J(r) = 1$ holds. However the inverse implication does not hold, there exist other processes different from Poisson, with $J(r) = 1$.

The estimation of $J(r)$ poses some difficulties, especially for small values of $1-H_s(r)$ (i.e. for large r). Furthermore, the numerator and denominator are point- and location-related summaries, respectively, each with different fluctuations that do not cancel out.

The summaries discussed so far are restricted to describe the point pattern in small neighborhoods of a typical point. For longer-range spatial correlation analysis a better approach is offered by second order characteristics.

The *inhomogeneous reduced second moment measure* is defined as:

$$\mathcal{K}(B) = \frac{1}{|A|}\mathbb{E}\sum_{u\in N(A)}\sum_{v\in N\setminus\{u\}}\frac{\mathbf{1}_{u-v\in B}}{\lambda(u)\cdot\lambda(v)}, \tag{6.2.1}$$

where the point process N is second-order reweighted stationary, meaning that the right hand side of (6.2.1) does not depend on the choice of $A \subseteq \mathbb{R}$, where $0 < |A| < \infty$. When the process is stationary and isotropic, the reduced moment function becomes Ripley's $K-$function (see [91]):

$$K(r) = \mathcal{K}\left(b(o,r)\right),\ r \geq 0.$$

The quantity $\lambda K(r)$ represents the mean number of points of N within a sphere of radius r centered on a 'typical point', which itself is excluded from the count. The function K is monotonically non-decreasing on $r > 0$ and converges to 0 as $r \to 0$. In the case of a stationary Poisson process: $K(r) = b_d r^d$, $r \geq 0$. A variance stabilized transformation of K (K estimated by non-parametric methods) is the L-function defined as:

$$L(r) = \sqrt{\frac{K(r)}{\pi}}, \tag{6.2.2}$$

in the planar case, and as

$$L(r) = \sqrt[d]{\frac{K(r)}{b_d}},\ r \geq 0, \tag{6.2.3}$$

in the general d-dimensional case.

For a Poisson process, $L(r) = r$. In general, $L(r) > r$ indicates aggregation and $L(r) < r$ indicates regularity.

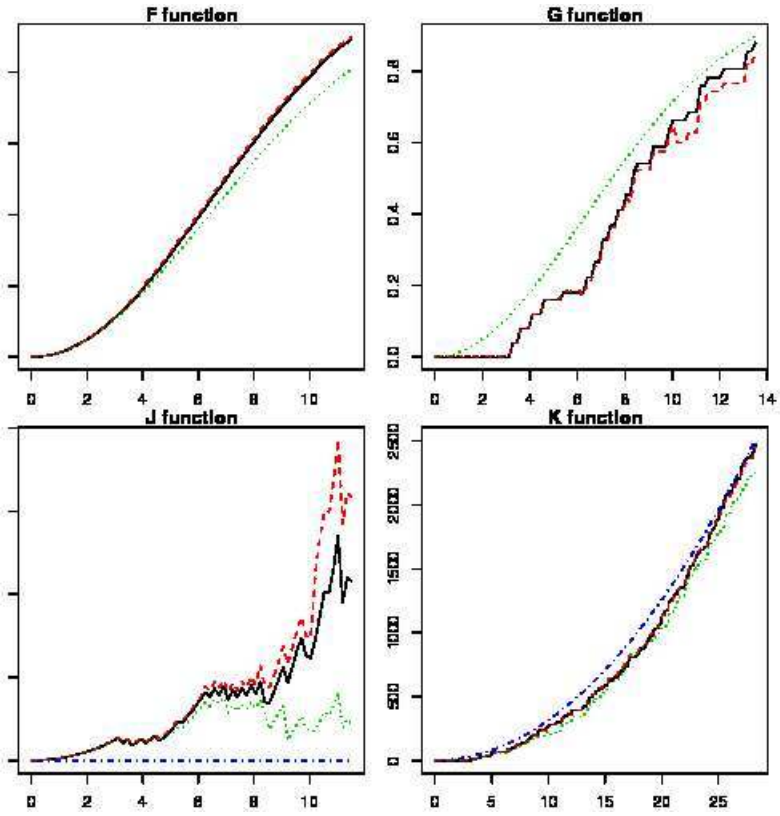


Figure 6.2: Summaries for the point pattern in Fig. 6.1 (different colors correspond to different edge correction methods)

6.3 Testing in the framework of point patterns

Based on the summary characteristics described in the previous section, several statistical tests are designed in order to gain information on the studied point process, its properties and generating mechanism.

In the frame of microarray analysis, we would like to test the randomness of the single molecule distributions inside the spots of the high resolution microarray data. Therefore the positions of the single molecules, obtained by the wavelet detection methods will form the input data we test against the CSR hypothesis.

Since the homogeneous Poisson process is the mathematical model that describes CSR, the "null model" implying complete lack of structure, the CSR tests consist of estimating a summary statistic (scalar or functional) from the data and compare it with the relevant summary characteristic for a Poisson process. If the difference between the two characteristics is large, the Poisson hypothesis is rejected.

We describe two CSR tests, one based on point numbers and the other on

inter-point distances of the point patterns, which we apply in order to test if the assumption of randomness of single molecule positions inside the microarray spot should be rejected or not.

The index-of-dispersion test

The sampling window W can be divided into subregions of equal size, called *quadrats*, usually for practical reasons in squares or rectangular grid cells. Under the homogeneous Poisson process hypothesis, the number of points in each quadrat follows a Poisson distribution of mean $\lambda \cdot \nu(Q)$ and the counts in disjoints quadrats are independent.

The test is defined based on the *index of dispersion* defined as:

$$I = \frac{(k-1)s^2}{\bar{x}},$$

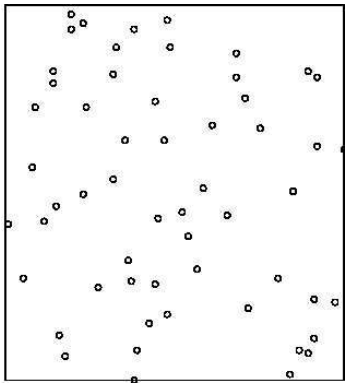
where k is the number of quadrats, $\bar{x}$ the mean of the number of points per quadrat, and s^2 is the sample variance of the number of points per quadrat. It represents the χ^2 goodness-of-fit test of the hypothesis that the n points are independently and uniformly distributed in W. It follows approximately a χ^2 distribution with $k-1$ degrees of freedom, for $k > 6$ and $\lambda \cdot \nu(Q) > 1$.

The CSR hypothesis is rejected if $I > \chi^2_{k-1;\alpha}$ (the alternative being that there is aggregation in the process), or if $I < \chi^2_{k-1;1-\alpha}$ (alternative: regularity of point pattern).

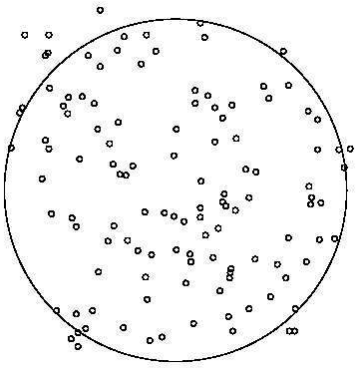
An extension of the above approach is the Greig-Smith method, which basically applies the previous test to a quadtree structure. It also gives information at which scale clustering occurs, if at all.

L test

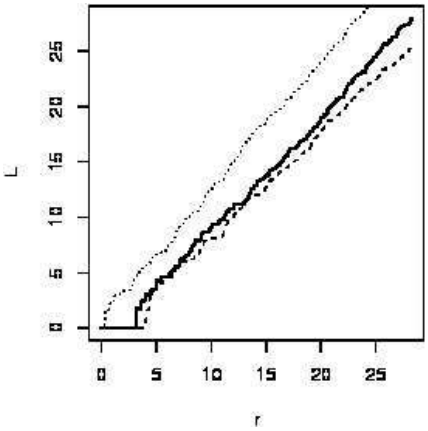
In order to test the hypothesis that the single molecules are randomly distributed we compute the empirical $L-$function based on the location of single molecules, obtained via the wavelet transform approach. Knowing that the $L-$function of a homogeneous Poisson process has the simple linear form: $L(r) = r$ for $r \geq 0$, the large deviation of the empirical $L-$function indicates rejection of the CSR hypothesis. The square root transformation in the computation of the $L-$function: $L(r) = \sqrt{K(R)/\pi}$ has a variance stabilization effect, that makes the $L-$function a better candidate for the test than the $K-$function.



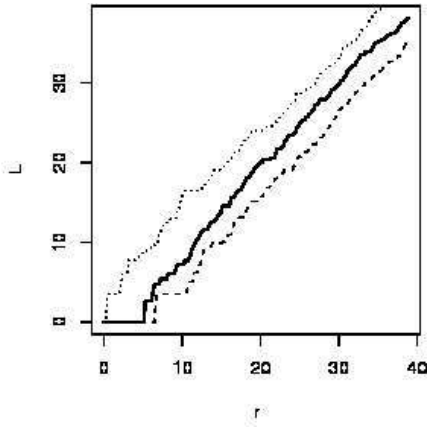
(a) Positions of single molecules in a PALM image



(b) Positions of single molecules in a high resolution microarray spot



(c) Estimated $\hat{L}$ and envelopes of 49 simulations for the point pattern in (a)



(d) Estimated $\hat{L}$ and envelopes of 49 simulations for the point pattern in (b)

Figure 6.3: Testing the CSR hypothesis

In order to compare the empirical $L-$function, $\hat{L}$, with the theoretical one, the following two methods were used. First, the empirical L-function is compared to the envelope of m Poisson process simulations with the same intensity $\hat{\lambda}$ as the original process. Denoting the simulated processes with $S_P^1, S_P^2, \ldots, S_P^m$, the lower and upper envelopes are defined as:

$$S_{P,min}(r) = \min\left\{S_P^1, S_P^2, \ldots, S_P^m\right\}$$
$$S_{P,max}(r) = \max\left\{S_P^1, S_P^2, \ldots, S_P^m\right\}$$

In Fig. 6.3, (a) and (b) the two point patterns corresponding to a Photo Activated Light Microscopy (PALM) image and to the high resolution microarray spot respectively are shown, together with the observation window. As can be seen in

sub-figures 6.3 (c) and (d), for both the point patterns the CSR hypothesis cannot be rejected: the empirical $L-$functions show great similarity with the theoretical $L-$function and are consistently located between the lower and upper envelopes computed from 49 simulations.

Moreover, as an additional test, the maximum deviance, computed as:

$$\tau = max_{r \leq s} |\hat{L}(r) - r|$$

can be approximated by a normal distribution if nus^3/a^2 small, where $a = \nu(W)$, u is the length of the boundary of W and n is the number of observed points [34]. The critical value of τ for the significance level $\alpha = 0.01$ is approximately $\tau_{0.01} = 1.75\sqrt{a}/n$. In the case of the PALM image, pattern (a) in Fig. 6.3, $\tau_a = 3.14 < \tau_{0.01} = 3.71$, as for pattern (b), the high resolution microarray spot, $\tau_b = 5.07 < \tau_{0.01} = 5.90$.

Although the CSR test does not prove the correctness of our model assumption, we shall consider that the stationary Poisson model is a reasonable description for the spatial distribution of single molecules.

Note however a small hard core in both processes (the region where the L function is null). It is due to the fact that in case two molecules are close to each other, only one is detected. The "resolution" of the wavelet detection algorithm applied to single molecule images is approximately 4 pixels.

6.4 Estimation of hybridization signal

In the case of high resolution microarray analysis, the hybridization signal is the intensity of the point process corresponding to the position of single molecules inside the microarray spot. We are interested especially in two models for intensity estimation:

- the homogeneous Poisson process
- the mixture of two homogeneous Poisson processes.

We have shown that inside the spot the CSR hypothesis of the single molecule locations cannot be rejected and we adopt this model due to its simplicity and the good agreement with the physics of the hybridization process. If the exact location of the spot is known, we consider the observation window W coinciding

with the spot and the positions of single molecules inside the spot represent an homogeneous Poisson process, with the only parameter the intensity λ.

A short overview of estimation methods for this case is given in [60].

- If the number of points in the sampling window W is $N(W)$ and the area (hyper-volume) of W is denoted $\nu(W)$ then:

$$\hat{\lambda} = \frac{N(W)}{\nu(W)} \tag{6.4.1}$$

  is an unbiased estimator with variance $var(\hat{\lambda}) = \frac{\lambda}{\nu(W)}$. This is also the maximum likelihood estimator.

- Distance methods are based on "nearest neighbour"-type distances, having a known distribution for homogeneous Poisson process model.

- Confidence intervals for λ are based on the normal approximation of the Poisson distributed $N(W) = \hat{\lambda}\nu(W)$:

$$\left(\frac{z_{\alpha/2}}{2} - \sqrt{N(W)}\right)^2 \leq \lambda\nu(W) \leq \left(\frac{z_{\alpha/2}}{2} + \sqrt{N(W)+1}\right)^2,$$

  where $z_{\alpha/2}$ are quantiles of the standard normal distribution. This method is most useful in determining the window size required for a given estimation accuracy: if δ is the desired width of the confidence interval and α is the required confidence level from

$$\delta\nu(W) \approx \left(\frac{z_{\alpha/2}}{2} + \sqrt{\lambda\nu(W)+1}\right)^2 - \left(\frac{z_{\alpha/2}}{2} + \sqrt{\lambda\nu(W)}\right)^2$$

  results that the required window size is

$$\nu(W) \approx \frac{4z_{\alpha/2}\lambda}{\delta^2}. \tag{6.4.2}$$

As we have mentioned in the beginning of the chapter, not all the peaks detected in the subimage corresponding to a grid element belong to the spot of interest (see Fig 3.7, c)). The background might be heavily contaminated by unspecifically bound signal, impurities, etc. which we shall call generally clutter, which when unaccounted for could seriously distort the hybridization results. Also in practice the exact spot location inside the subimage is not known, small distortion might be due to printing tip motion, scanning artifacts etc.

Each rectangular subimage of the microarray, obtained after the gridding step, contains the circular spot of mRNA hybridized to the printed cDNA (foreground) surrounded by a background region with clutter. Therefore, peaks detected in the subimages have to be assigned either to foreground or background. In order to make a correct concentration estimation (or hybridization estimation), we have to model both the spot and the background concentration. A straightforward model turns out to be the superposition of spatial Poisson point processes. However a more precise model is that of a homogeneous Poisson process (modeling the clutter) on which is superposed another homogeneous Poisson process restricted to a subdomain of the initial process. The new process has piecewise-constant intensity function. The estimation of the intensity of the process can be done either by parametric methods, typically based on mixture models on one hand, and on nonparametric methods, on the other hand.

In the case of parametric models, in order to distinguish between peaks within the spot, representing true hybridization signal and those representing clutter we use spatial information, counts of peaks in subregions or relative distances of the peaks to each other and design a spatial mixture model based on this information. One component of the mixture will represent the true signal, while the other the clutter.

In the case of non-parametric intensity estimation approaches first the intensity is estimated and subsequently the location of the spot is found. We will use the Nadaraya-Watson kernel estimate or a wavelet estimation method, and in both cases the spot position is found after intensity estimation via segmentations methods.

In the classical microarray case, this step corresponds to foreground/background segmentation and in [14] it is based on a Gaussian mixture model for pixel intensity values, where each pixel is assumed having a mixture distribution of two (signal/background) or three (signal/background/artifacts) components.

Analysis of count data via the method of moments

A first parametric approach for concentration estimation is to consider the count y_i, of the detected peaks inside non-overlapping systematic quadrats of size $s \times s$ pixels covering the spot subimage, where i is an index over the quadrat covering, $i \in \{1, \ldots, N\}$. Let Y be the random variables whose realizations are the counts y_i. It is supposed to follow a mixture distribution with K components and we would like to determine the parameters of this mixture distribution. The basic idea is

to apply the method of moments to the data y_i. In the general case, the method consists of selecting a set of moments $\{\mathbb{E}(H_j(Y)|\theta)\}_j$ of the random variable Y and determine the unknown parameters θ, by matching the theoretical moments with their empirical counterparts $\widetilde{H}_j$.

It results in solving a usually nonlinear equation system:

$$\sum_{k=1}^{K} \eta_k \mathbb{E}(H_j(Y)|\theta_k) = \widetilde{H}_j \tag{6.4.3}$$

for the unknown parameters θ_k of each mixture component and η_k representing the proportion of each component in the mixture.

For the case of Poisson mixture distribution,

$$P(Y = y) = \sum_{k=1}^{K} \eta_k \frac{e^{-\lambda_k} \lambda_k^y}{y!},$$

where $\sum_{k=1}^{K} \eta_k = 1 >$, and $\eta_k \neq 0, \quad k = 1, \quad \ldots, K$, following the method of Everitt described in [42], we use the factorial moments :

$$H_j(Y) = \frac{Y!}{(Y-j)!}, \quad j = 1, \ldots 2K - 1.$$

For data distributed according to a Poisson mixture distribution the factorial moments are equal to the moments about the origin of the mixing distribution, $\sum_{k=1}^{K} \eta_k \lambda_k^j$.

In the case of a microarray spot subimage, the quadrat values describe a mixture of two Poisson processes: one describing counts inside the spot and the other outside of the microarray spot. The counts are modeled as a mixture of two Poisson distributions, with constant concentrations λ_1, λ_2 (expressed in counts per quadrat): $p(y_i|\lambda_1, \lambda_2, \eta_1) = \eta_1 \mathcal{P}oi\,(\lambda_1) + (1 - \eta_1)\mathcal{P}oi\,(\lambda_2)\,$, where η_1 denotes the weight of the first component. This simple model does not account for correlations between neighbouring quadrats. For this mixture of two components one has to determine three parameters: λ_1, λ_2 the intensity of the two Poisson distributions and η_1 the proportion of the first component in the whole population.

The first three factorial moments $\mathbb{E}(H_j(Y)|\lambda_1, \lambda_2, \eta_1)$ of the random variable Y, are matched with empirical moments obtained from y_i. The equation system

for λ_1, λ_2 and η_1 becomes:

$$\begin{aligned}\eta_1(\lambda_1 - \lambda_2) + \lambda_2 &= v_1\\ \eta_1(\lambda_1^2 - \lambda_2^2) + \lambda_2^2 &= v_2\\ \eta_1(\lambda_1^3 - \lambda_2^3) + \lambda_2^3 &= v_3\end{aligned}$$

with $v_j = \frac{1}{N}\sum_{i:y_i \geq j} y_i(y_i - 1)\cdots(y_i - (j-1))$.

Although inefficient compared to other estimators (see, e.g., [68]), this simple method offers a closed form solution in the case of mixture of two Poisson distributions, which is crucial for the speed of the analysis for such a large quantity of data.

We tested several quadrat sizes, but the results obtained on real images were robust for quadrats of size 20×20 pixels and above. However further study is necessary to select the optimal quadrat size.

Expectation maximization (EM) based on Kth nearest neighbor distances

As a second approach we adopt the method of Byers and Raftery, used in minefield detection [18]. The detection results are not pooled together, but are considered as the realization of a spatial Poisson process. The location of the detected peaks are treated as a mixture of two spatial Poisson processes with different rates for foreground and background.

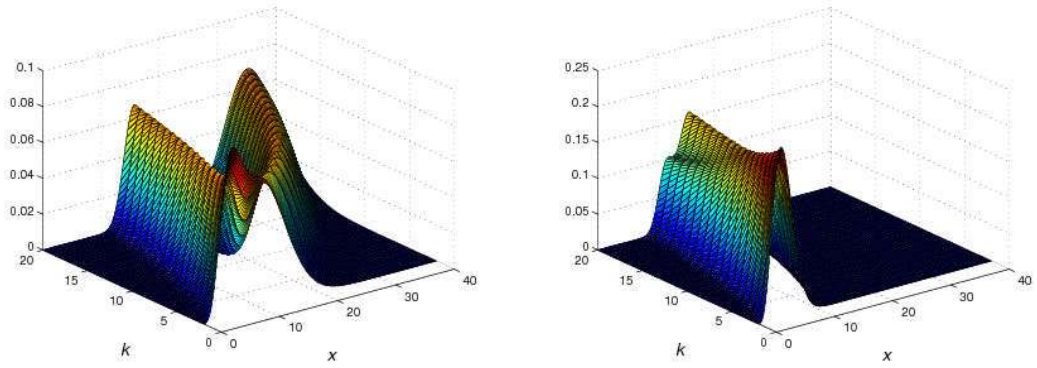
In the case of a single spatial Poisson processes with constant rate λ the distribution D_K of the distance from a point of the Poisson process to its K-th nearest neighbor can be written as

$$P(D_K \geq x) = \sum_{k=0}^{K-1} \frac{e^{-\lambda\pi x^2}\left(\lambda\pi x^2\right)^k}{k!} = 1 - F_{D_K}(x). \tag{6.4.4}$$

This leads to the density function:

$$f_{D_K}(x) = \frac{dF_{D_K}(x)}{dx} = \frac{e^{-\lambda\pi x^2} 2\left(\lambda\pi\right)^K x^{2K-1}}{(K-1)!}$$

meaning that the D_K^2 follows a transformed Gamma distribution, $D_K^2 \sim \Gamma(K, \lambda\pi)$.



(a) Easily separable components of the mixture
$\lambda_1 = 0.025, \lambda_2 = 0.005$

(b) A case with mixture components difficult to separate
$\lambda_1 = 0.065, \lambda_2 = 0.045$

Figure 6.4: The density function of kNN distance $D_k(\lambda_1, \lambda_2)$ for the spatial Poisson mixture with concentrations λ_1 inside and λ_2 outside the spot, respectively.

The maximum likelihood estimate of the rate of the Poisson process is:

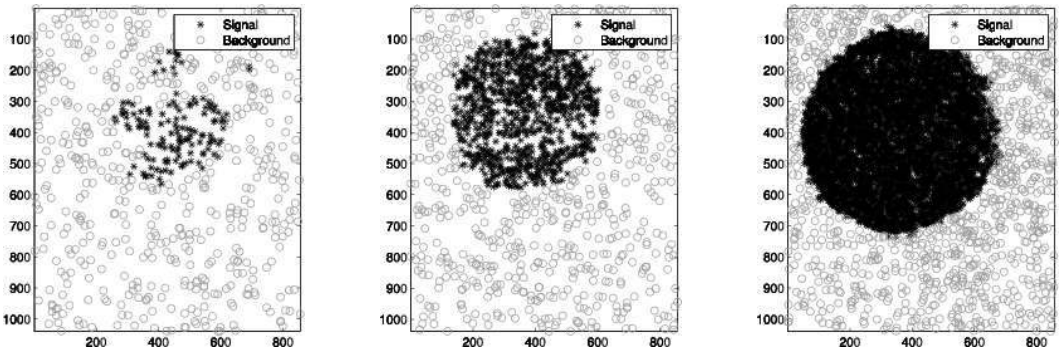
$$\widehat{\lambda} = \frac{K}{\pi \sum_{i=1}^{n} d_i^2} \tag{6.4.5}$$

where $d_i, i = 1, \ldots, n$ are the realizations of the K-th nearest neighbour distances.

In the case of a mixture of two Poisson processes with two intensity rates λ_1 and λ_2, the model for D_K can be written as

$$D_K \sim p\, \Gamma^{\frac{1}{2}}(K, \lambda_1 \pi) + (1-p)\Gamma^{\frac{1}{2}}(K, \lambda_2 \pi)\,. \tag{6.4.6}$$

The pdf of the mixture is plotted in Fig. 6.4. The Figure 6.4 is a pseudo-surface representing the distribution function of the k-th nearest neighbour $D_k(\lambda_1, \lambda_2)$ in the spatial mixture model for each fixed value K on the axis kNN, and for fixed rates λ_1 and λ_2. The bigger the difference between the signal and clutter concentration the easier is to separate the two components of the mixture. Also, as one can see in Fig. 6.4 for high concentrations the task is more challenging than for lower concentrations. As opposed to the method of moments' η, p represents a proportion of the samples D_K. The three unknown parameters that describe the distribution D_K: p, λ_1, λ_2, are estimated via the EM algorithm, together with the assignments to components ("missing data") $\delta_i \in \{0, 1\}$, where $\delta_i = 1$ if the $i-$th point belongs to the first component (signal), and $\delta_i = 0$ otherwise.



(a) 0.08 amol/80μl (b) 0.8 amol/80μl (c) 8 amol/80μl

Figure 6.5: Background/foreground separation of peaks for three different concentrations via the EM method applied to the Kth nearest neighbour distances.

The expectation step becomes:

$$E\left(\widehat{\delta}_i^{(t+1)}\right) = \frac{p^{(t)} f_{D_K}\left(d_i, \widehat{\lambda}_1^{(t)}\right)}{p^{(t)} f_{D_K}\left(d_i, \widehat{\lambda}_1^{(t)}\right) + (1-p^{(t)}) f_{D_K}\left(d_i, \widehat{\lambda}_2^{(t)}\right)}$$

and the maximization:

$$\begin{aligned}
\widehat{\lambda}_1^{(t+1)} &= \frac{K \sum_{i=1}^n \widehat{\delta}_i^{(t+1)}}{\pi \sum_{i=1}^n d_i^2 \widehat{\delta}_i^(t+1)} \\
\widehat{\lambda}_2^{(t+1)} &= \frac{K \sum_{i=1}^n \left(1-\widehat{\delta}_i^{(t+1)}\right)}{\pi \sum_{i=1}^n d_i^2 \left(1-\widehat{\delta}_i^{(t+1)}\right)} \\
p^{(t+1)} &= \sum_{i=1}^n \frac{\widehat{\delta}_i^{(t+1)}}{n}.
\end{aligned}$$

As initial values for the three parameters one can use the results obtained through the method of moments. In order to gain insight in the challenges of microarray image analysis, we present the results of the classical method and the high resolution adapted peak counting method for three different concentration: $0.08, 0.08$, and 8 amol/80μl. In Fig. 6.5 the result of the EM method applied to real oligonucleotide data is shown. In the leftmost image, characterized by low concentration of molecules, one can see that only for a part of the spot the hybridized signal was correctly identified. For higher concentrations (middle and right image) the spot shape can clearly be recognized. The EM method is insensitive to the shape that has to be detected, allowing for the analysis of anomalous spots due

to uneven drying, spotting errors etc. [99]. An example of a possible shape that could be analyzed is shown in Fig. 6.6.

Segmentation of point processes based on a level set approach

Given that we deal with homogeneous Poisson point processes in distinct regions of the image (the inside respective outside region of the microarray spot), the intensity of the respective point process can be seen as a piece-wise constant function over disjunct regions of the image.

The level set approach proposed in [20, 21] is designed to segment images that can be well approximated by piecewise constant or piecewise smooth functions and for which the separating contours is not necessarily gradient based. Thus it is a reasonable choice to partition the microarray spot image and estimate the two concentrations (or intensities) via this level set algorithm.

The algorithm is a particular case of the Mumford-Shah segmentation problem [84], that finds a decomposition of the image domain $\Omega \subset \mathbb{R}^2$ into disjunct components Ω_i, defined as the connected components of $\Omega \backslash C$, where C is a closed subset in Ω, consisting of a finite set of smooth curves. Thus $\Omega = \bigcup_i \Omega_i \cup C$. In order to solve this problem, Mumford and Shah minimize the following energy functional

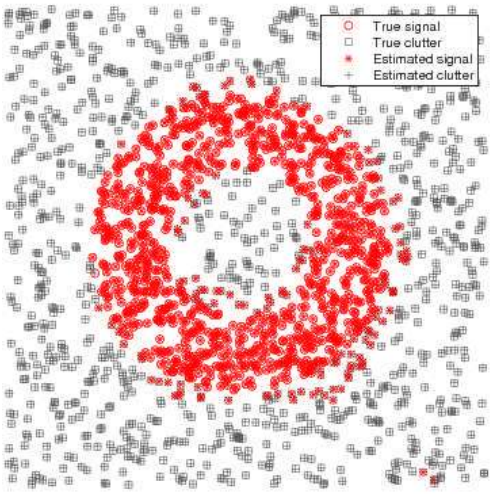


Figure 6.6: Anomalous shape detection. A high concentration donut shape was simulated on a background formed of low concentration clutter. The proposed approach is able to separate signal from clutter.

with respect to u and C

$$F^{MS} = \int_{\Omega} (u - u_0)^2 \, dx \, dy + \mu \int_{\Omega \setminus C} |\nabla u|^2 \, dx \, dy + \nu |C|,$$

where $u_0 : \Omega \to \mathbb{R}$ is the bounded image function, $|C|$ denotes the length of C and ν and μ are fixed weight parameters. In particular, if u_0 is formed of regions of approximately piecewise-constant intensities, approximately c_1 inside C and approximately c_2 outside this region and setting $\mu = 0$ (for constant regions having anyway $\nabla u_0 = 0$ fulfilled) the minimizing problem becomes:

$$\inf_{u,C} F = \lambda_1 \int_{\text{int}(C)} (u(x,y) - c_1)^2 \, dx \, dy + \lambda_2 \int_{\text{out}(C)} (u(x,y) - c_2)^2 \, dx \, dy + \nu |C|. \quad (6.4.7)$$

As a matter of notation if $\Omega^* \subset \Omega$ an open subset then C is the boundary of the subset $C = \partial\Omega^*$, $\text{int}(C) = \Omega^*$ and $\text{out}(C) = \Omega \setminus \Omega^*$. An ubiquitous choice of parameters is $\lambda_1 = \lambda_2 = 1$.

The algorithms makes use of the implicit representation for the evolving curves C as the zero level set of a scalar Lipschitz continuous function $\phi : \Omega \to \mathbb{R}$ such that $\phi(x, y) > 0$ in $\text{int}(C)$, $\phi(x, y) < 0$ in $\text{out}(C)$ and $\phi(x, y) = 0$ on C and with the help of the Heaviside function H

$$H(z) = \begin{cases} 1, & \text{if} \quad z \geq 0 \\ 0, & \text{if} \quad z < 0 \end{cases}$$

and the Dirac measure δ_a given by

$$\delta_a(A) = \begin{cases} 1, & \text{if} \quad a \in A \\ 0, & \text{if} \quad a \notin A \end{cases}$$

By replacing H and δ with their regularized versions H_ε and δ_ε (e.g. obtained via convolution with a molifier), $\delta_\varepsilon(z) = \frac{d}{dz} H_\varepsilon(z)$, the length of C can be expressed as

$$|C| = \lim_{\varepsilon \leftarrow 0} \int_{\Omega} |\nabla H_\varepsilon(\phi(x,y))| \, dx \, dy = \lim_{\varepsilon \leftarrow 0} \int_{\Omega} \delta_\varepsilon(\phi(x,y)) |\nabla \phi(x,y)| \, dx \, dy.$$

Chan and Vese rewrite (6.4.7) as

$$\inf_{u,C} F = \int_\Omega (u(x,y) - c_1)^2 H_\varepsilon(\phi)\, dx\, dy + \int_\Omega (u(x,y) - c_2)^2 (1 - H_\varepsilon(\phi))\, dx\, dy + \nu \int_\Omega |\nabla H_\varepsilon(\phi)|. \tag{6.4.8}$$

Now the solution can be written as: $u(x,y) = c_1 H(\phi(x,y)) + c_2(1 - H(\phi(x,y)))$ and the constants c_1, c_2 minimizing (6.4.8) for fixed ϕ are

$$c_1(\phi) = \frac{\int_\Omega u_0(x,y) H_\epsilon(\phi(x,y))\, dx\, dy}{\int_\Omega H_\epsilon(\phi(x,y))\, dx\, dy}$$
$$c_2(\phi) = \frac{\int_\Omega u_0(x,y)(1 - H_\epsilon(\phi(x,y)))\, dx\, dy}{\int_\Omega (1 - H_\epsilon(\phi(x,y)))\, dx\, dy}$$

while for fixed constants c_1, c_2 the minimizer with respect to ϕ is

$$\frac{\partial \phi}{\partial t} = \delta_\epsilon \left[\nu \mathrm{div}\left(\frac{\nabla \phi}{|\nabla \phi|} \right) - (u_0 - c_1)^2 + (u_0 - c_2)^2 \right]. \tag{6.4.9}$$

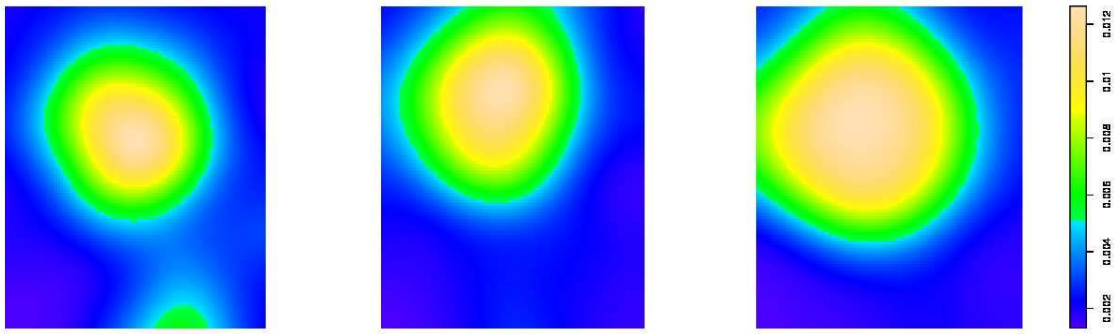
The authors summarize the advantages of this method as opposed to other active contour methods as being able to segment images in which the contour is not necessarily based on the gradient, (as in our case, where the two different point Processes are separated by a cognitive contour and not a gradient based one), it detects interior contours and is not dependent on the initial starting contour. See [20, 21] for more details [1].

The level set approach applied directly to our point process (or even to the microarray spot image) however, does not detect the two regions of different point concentrations, but instead detects single peaks (for a very wide range of the penalization terms - we have tested a range over 10 orders of magnitude). Instead the approach is applied to a kernel density estimate of the point process described below.

Nadaraya-Watson kernel density estimator The best known non-parametric intensity estimator is the Nadaraya-Watson or simply kernel estimator [101]. The *kernel estimator* with kernel K is defined by

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h} \right),$$

[1] The implementation of the level set segmentation method was kindly provided by Bettina Heise (Dept. of Knowledge-based Mathematical Systems, Johannes Kepler University, Linz).



(a) 0.008 amol/80μl (b) 0.08 amol/80μl (c) 8 amol/80μl

Figure 6.7: Nadaraya-Watson kernel smoothing

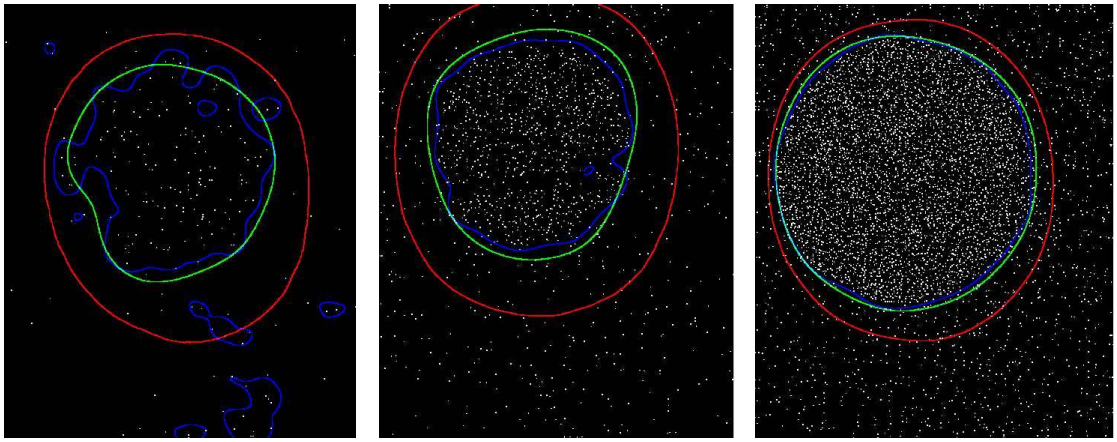
where where K is a function satisfying

$$\int_{-\infty}^{\infty} K(x)\,dx = 1$$

and h is the *kernel width*, called also *smoothing parameter* or *bandwidth*. The two most important problems that arise in kernel density estimation is the choice of the kernel function K and of the kernel width h. From these two problems, the crucial one is the choice of the parameter h (see [101]). As kernel function we shall use a 2-dimensional Gaussian shape.

Many approaches have been developed for bandwidth selection (see e.g. [101]) however a simple heuristic to chose the width of the kernel is to use the interquartile distance. The kernel estimator is illustrated on the single molecules detected in three oligo-nucleotide spots with three different concentration (the heuristically chosen h is the interquartile range bandwidth $h = h_{iq}$) and the results are presented in Fig. 6.7. The kernel smoothed images were subsequently segmented via the level set approach described above. Three separation contours are presented in Fig. 6.8 together with the input point patterns for three bandwidth choices: in red $h = h_{iq}$, in green $h = 1/4h_{iq}$ and in blue $h = 1/10h_{iq}$.

Although the segmentation is only slightly affected, the signal concentration estimate is more biased in the case of larger h (when a higher area is contributes to the smoothed value at one point). Instead we propose a simple re-estimation of the single molecule concentration inside the spot int(C) as in (6.4.1), where $W = \text{int}(C)$.



(a) 0.008 amol/80μl (b) 0.08 amol/80μl (c) 8 amol/80μl

Figure 6.8: Segmentation of the point patterns via level sets for three choices of smoothing bandwidths (red: $h = h_{iq}$, green $h = 1/4h_{iq}$ and blue $h = 1/10h_{iq}$)

6.5 Evaluation of concentration estimation

The concentration estimation algorithms were tested on simulated data representing the position of single molecules and clutter, respectively.

We assumed that signal (molecules' position) has a higher concentration than clutter. For each data set, two spatial Poisson processes are simulated: one of intensity λ_1 inside a disk of radius R (150 pixels in our case) and a second one, of intensity λ_2, independent of the first one, in a rectangle excluding this area. The parameter values were chosen such that $(\lambda_1, \lambda_2) \in [0.01, 0.05] \times [0.005, 0.05]$, $\lambda_2 < \lambda_1$. For each (λ_1, λ_2) parameter pair, ten data sets were generated. The results of the estimation of the signal concentration λ_1 are presented in Fig. 6.9. In order to evaluate the performance of the presented algorithms we have in mind three criteria:

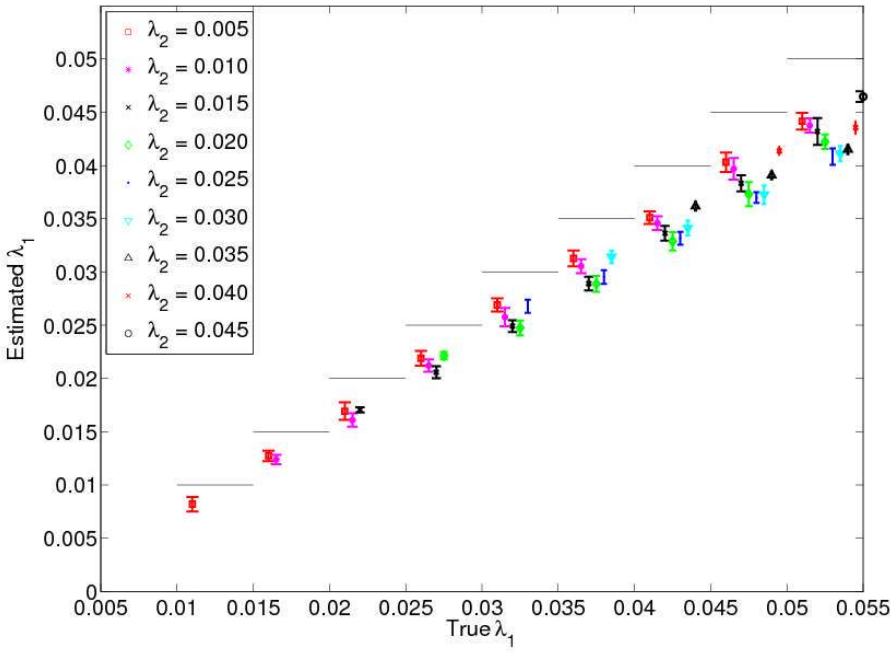
- accuracy
- precision
- computational cost.

Accuracy can be seen as a measure of how close the estimate approaches the true simulation parameter, while precision describes repeatability or more precisely how tight estimates cluster for given parameter. As for the last criterion, given the very different nature of the algorithms to be compared we have considered the execution time to be the measure of computational cost. In Fig. 6.9 and

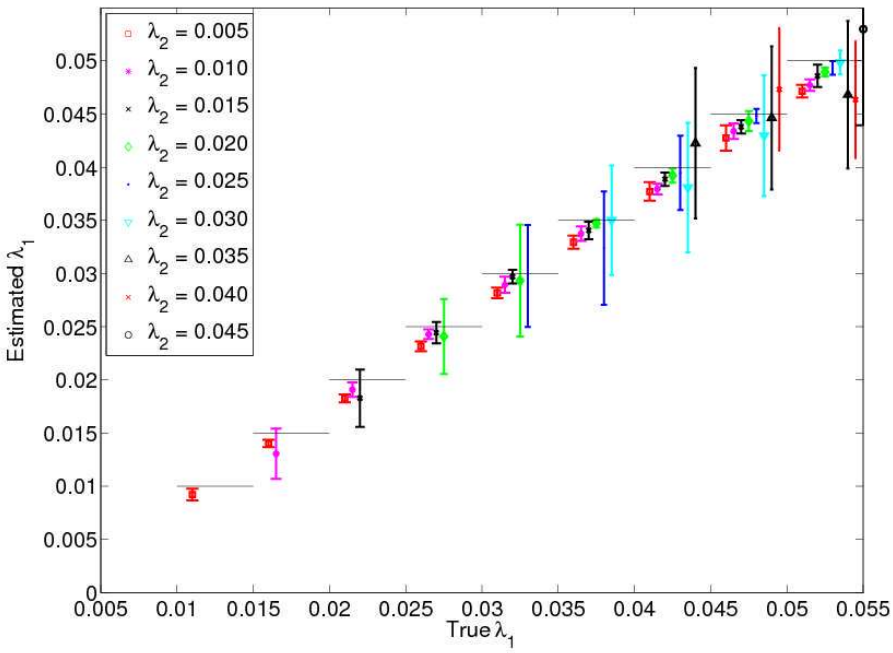
6.10 the accuracy of the estimate is represented by the distance to the true value, while the precision is reflected by the size of the error bars. The results in the case of MOM estimation are biased to lower intensities λ_1 than the true value, represented as a stair-case function on the figure. When the two concentrations are close together one can see a stronger bias, due to the failure of separating the two components of the mixture. The same behaviour can be noticed in the case of level set segmentation (with and without re-estimation of the concentration) in Fig. 6.10. In the case of re-estimated concentration the phenomenon is significantly stronger than in the estimation obtained implicitly by the level set method. The EM-based algorithm produces almost no bias. The true parameter value belongs to the estimated confidence interval. The precision of the method tends to increase with the concentration λ_2 of the clutter around the spot of interest (which appears as color code in the figure).

The highest precision among the three methods is achieved by the level set approach at the cost of higher computational cost. Although more computationally expensive than the MOM, the increased accuracy of the EM method makes it preferable to the MOM and the level set approach. Finally, MOM offers a good trade-off in case all three criteria: accuracy, precision and computational cost are taken into account.

The mean squared error (MSE) for the three methods over the range of signal and clutter concentration is summarized in Fig. 6.11. Only the parameters in the region $\lambda_2 < \lambda_1$ were taken into account in the simulations. The smallest mean squared error is achieved by the EM algorithm, while the MOM method and the level set method with reestimated concentrations have comparable behaviour.

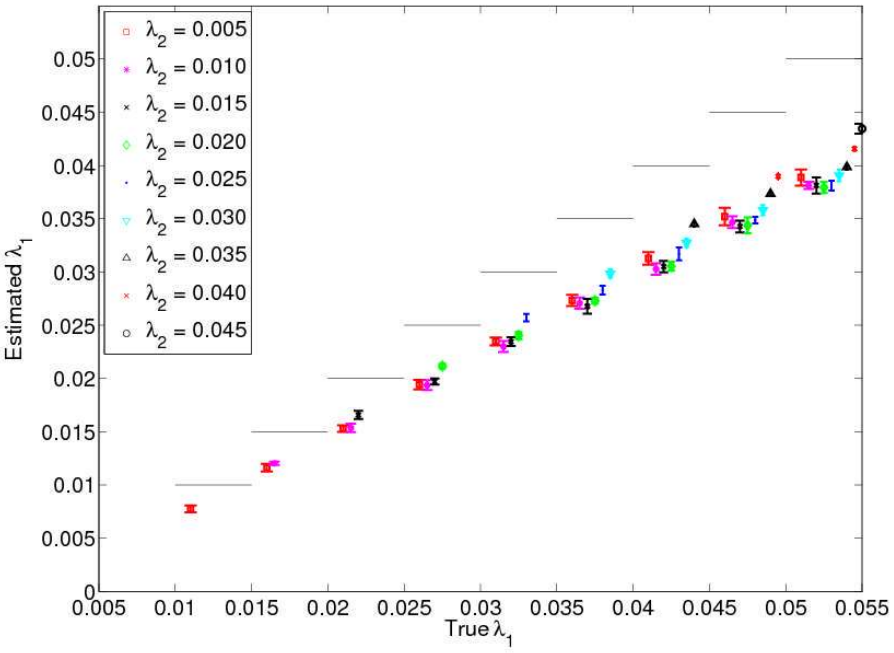


(a) MOM

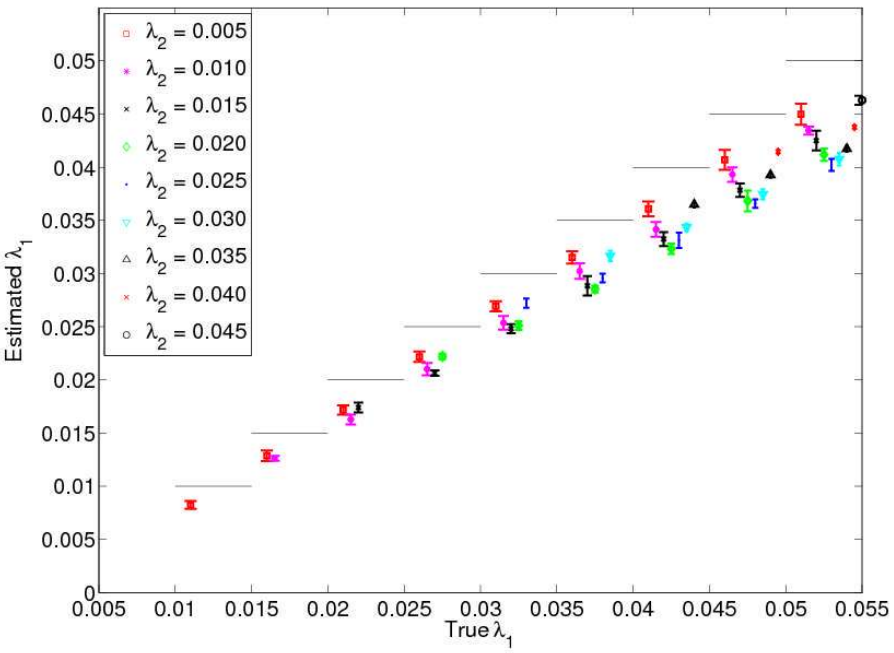


(b) EM

Figure 6.9: Concentration estimation results for the MOM and EM method on simulated data. The true λ values are represented as a stair-case function and for better visibility, the estimation results were slightly shifted on the abscissa.



(a) Concentration estimation after level set segmentation



(b) Reestimated concentration after level set segmentation

Figure 6.10: Concentration estimation results for the level set based method on simulated data. The true λ values are represented as a stair-case function and for better visibility, the estimation results were slightly shifted on the abscissa.

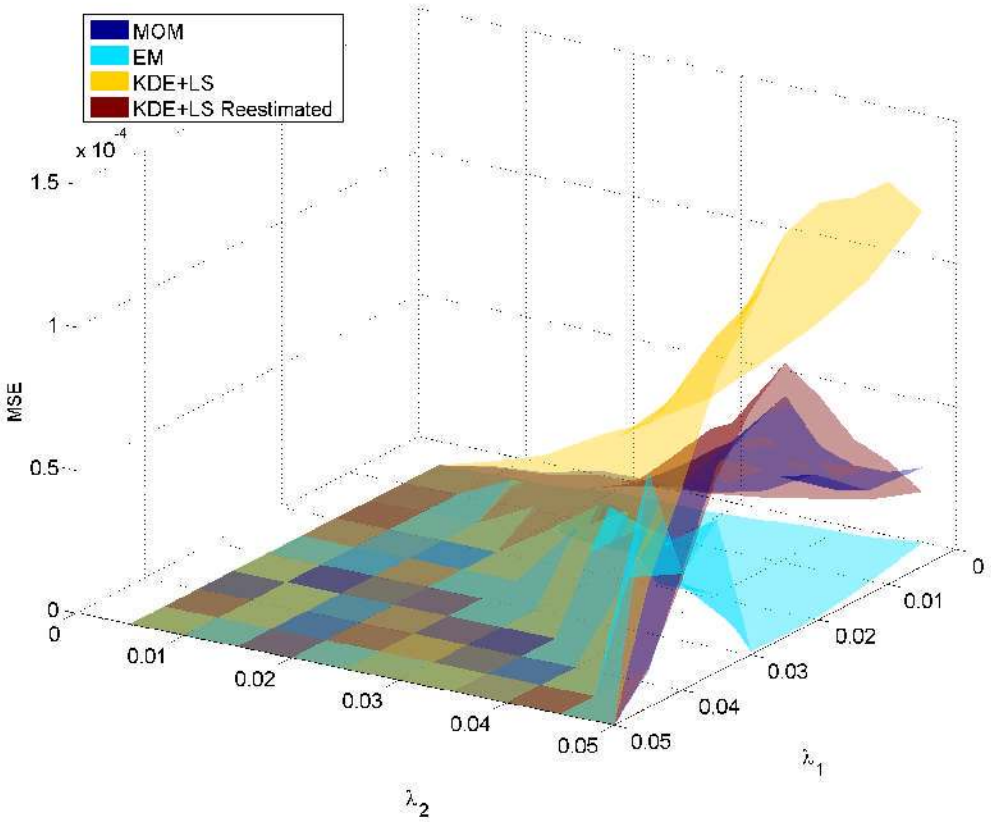


Figure 6.11: Mean squared error of the proposed algorithms over the range of signal λ_1 and clutter λ_2 concentrations. The EM algorithm has the best performance, followed by MOM and level sets with reestimated concentrations.

Chapter 7

Results of microarray analysis with single molecule sensitivity

Given the huge amount of data that needs to be analyzed the single molecule detection and signal estimation algorithms have to be fast, and in the same time accurate and robust to noise and model misspecification. Moreover, the range of interest of mRNA and cDNA concentration in the sample is much below the required concentration for classical low-resolution microarray analysis.

The proposed single molecule analysis tackles all the above described difficulties. We have separately evaluated the single molecule detection algorithms in Chapter 5 and the concentration estimation in Chapter 6. Now we validate our entire approach on simulations that check the correlation of the estimated parameters with the simulation parameters, as well as on real high-resolution microarray images. Since the ground truth in the case of real images is not known, special dilution series experiments for oligonucleotide arrays are used. Finally, we describe the results of a full analysis of cDNA microarray slides for a Multiple Myeloma sample. The list of differently expressed genes obtained via this analysis was confirmed by independent PCR analysis.

7.1 Validation on simulated data

Simulation images

Simulation images represent a square domain of 512×512 pixels with a spot of radius 150 pixels at its center (the proportions correspond to the real case situation). The (optional) background of the image, due in practice mainly to the

laser profile, is constructed as a $2d$ image of a ridge. It is a vertical repeat of a $1d$ horizontal Gaussian signal with $\sigma = 500$.

The positions of single peaks are generated according to a mixture of two spatial Poisson processes with parameters λ_S representing the peak concentration inside the spot and λ_B representing peaks outside the spot of interest (noise). At each position generated as above, an appropriately discretized $2d$ Gaussian approximation of the PSF is added to the background image (see the details of the model in Section 2.3. The total intensity of each Gaussian peak is randomly drawn from a Poisson distribution $\mathcal{P}(\mu)$ or a compound Poisson distribution $\mathcal{CP}(n, \mu)$ (in the case of multiple dye model), where μ is the expected number of photons emitted by a single dye and n is the expected number of dyes bound to a single molecule. The width of the Gaussian can be chosen constant (e.g. $\sigma = 1$) or independent for each peak, as an instance of a Gaussian random variable $\mathcal{N}(m, s)$ (e.g. $m = 1, s = 0.01$). Finally, Poisson and Gaussian noise are added to the simulation model.

Figure 7.1 shows the images simulated as described above, and for the sake of visual comparison also real oligonucleotide spots are presented in Fig. 7.2.

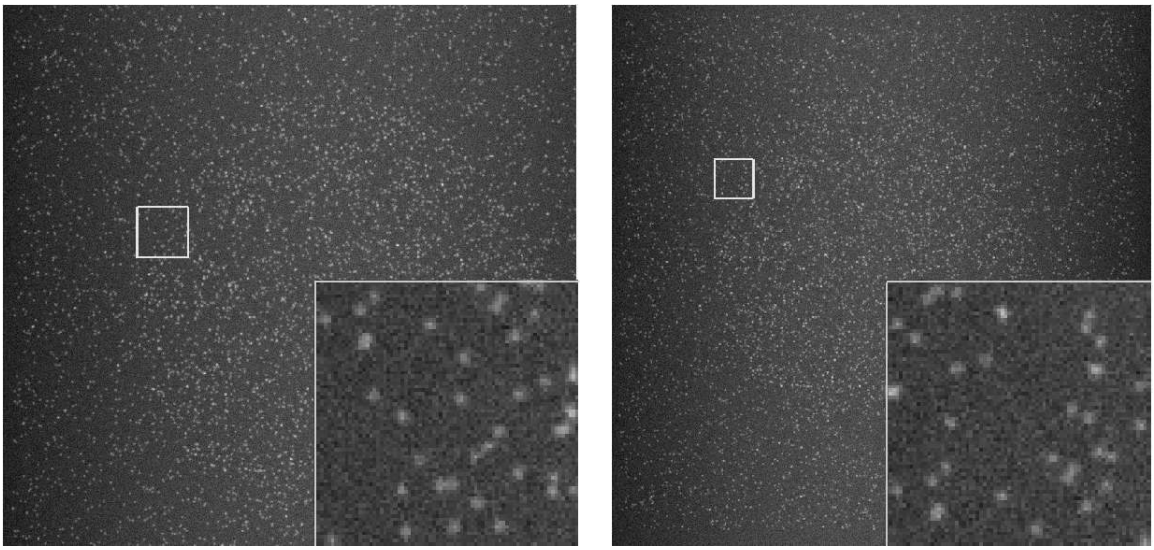


Figure 7.1: Simulation of microarray spots. Peak concentrations: *left:* 0.005 peaks/pixel inside the spot and 0.0005 peaks/pixel outside the spot, *right:* 0.01 peaks/pixel inside the spot and 0.0025 peaks/pixel outside the spot

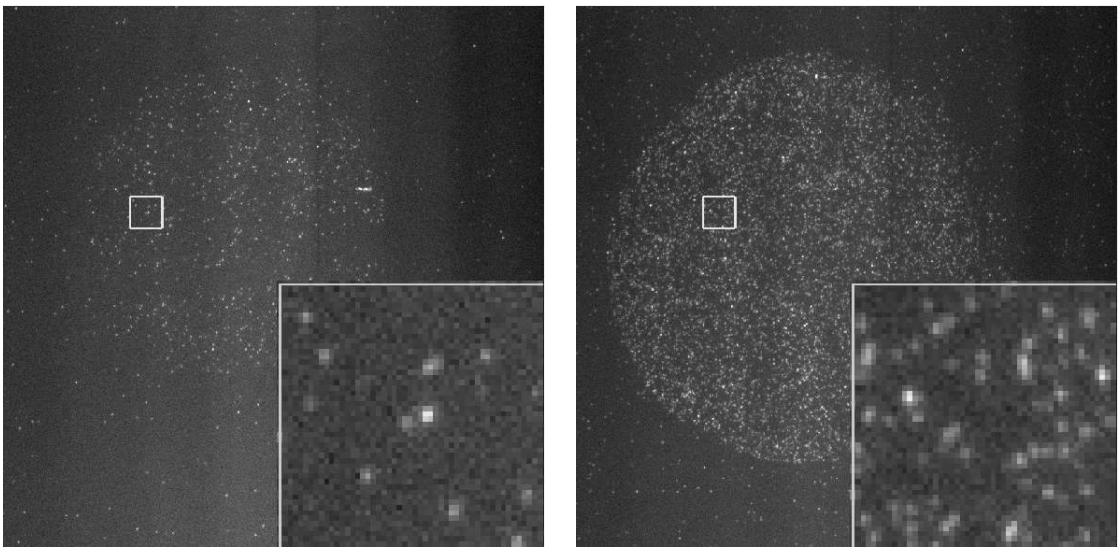


Figure 7.2: Scanned oligonucleotide spots (dilution: 0.8 and 8 amol/80μl)

Classical microarray methods applied to downsampled simulation data

In order to check the effect of the novel high resolution information on the estimation of hybridization intensity we have compared simulation results for state-of-the-art microarray spot segmentation methods. The algorithms selected for the comparison, particularly suited for low SNR images, are: maximum likelihood estimation segmentation (MLE) [14], a segmentation based on the Mann-Whitney test [22] and a Markov Chain Monte Carlo (MCMC) approach [49]. Briefly we describe the three algorithms.

MLE segmentation

The pixel values of the subimage corresponding to a grid element are assumed distributed according to a Gaussian mixture model (GMM):

$$f = \sum_{j=1}^{K} \pi_j \mathcal{N}(\mu_j, \sigma_j).$$

Each component of the mixture describes one of the following classes: signal S and background B if $K = 2$, or signal S, background B and artifacts A if $K = 3$.

The parameters of the mixture $\Psi = \{\pi_k, \mu_j, \sigma_j, j = 1, \ldots, K\}$ are estimated via

maximization of the log-likelihood:

$$\ell(x) = \sum_{i=1}^{N} \log f(x_i|\Psi) = \sum_{i=1}^{N} \log \left\{ \sum_{j=1}^{K} \pi_j \mathcal{N}(\mu_j, \sigma_j) \right\}.$$

The maximization of the log-likelihood can be obtained e.g. via an expectation-maximization approach similar to the one described in Section 6.4.

Segmentation based on Mann-Whitney test

The iterative algorithm starts with an initial partition of the region of interest in foreground (signal) $S_i, i = 1, \ldots, m$ and background pixels $B_j, j = 1, \ldots, n$.

The iteration step consists of the (non-parametric and distribution-free) Mann-Whitney test applied to eight random background pixels B_j^8 and the eight lowest intensity foreground pixels S_j^8. For each pixel value a testing problem is defined, based on the accept/reject of the null hypothesis

$$H_0 : \mu_S - \mu_B = 0$$

the alternative hypothesis being

$$H_1 : \mu_S - \mu_B > 0,$$

where μ_S and μ_B are the mean values of signal, respective background. The test is based on the rank-sum statistic R, the sum of the ranks of all the samples S_j^8 in the ordered sequence of $S_j^8 \cup B_j^8$. The hypothesis H_0 is rejected if $R \geq r_\alpha$, the critical value corresponding to the significance level α, knowing that under the null hypothesis

$$\frac{R - \frac{n(n+1)}{2} - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} \propto \mathcal{N}(0,1),$$

for $m, n \geq 8$ (and can be directly computed for smaller values).

If H_0 is not rejected, the smallest values (possibly only one) of $S_i, i = 1, \ldots, m$ are discarded and the lowest eight samples from the remaining S_i values are selected as S_j^8 and the procedure is iterated until no potential signal pixels remain or the H_0 is rejected. If the second situation occurs, all the remaining pixels in S_i are signal, the rest belonging to the background of the spot. Further details on the approach can be found in [22].

MCMC segmentation

The approach is based on a hierarchical Bayes model of the spot image and uses Markov Chain Monte Carlo algorithms as computational tool to sample the posterior distribution and estimate the parameters of the model. The algorithm was fully described in [35], and reimplemented and applied to low-resolution microarray images in [99]. The parameters describe the spot shape model and the signal intensity model. We assume a simplified circular spot shape $\mathcal{R}(\tau) = \{z = (x, y) : (x - x_0)^2 + (y - y_0)^2 \leq r^2\}$, where $\tau = [x_0, y_0, r]$ represents the shape parameter vector, but it is possible to consider more complex shapes, like ellipse or annulus. The subimage corresponding to a grid element can be written as the union of the spot region and the complement of the spot region: $G = \mathcal{R}(\tau) \cup \mathcal{R}^C(\tau)$.

Each pixel r is assumed normally distributed, and corrupted by additive measurement noise ε_r: $u(r) = u_0(r) + \varepsilon_r, \quad \varepsilon_r \sim \mathcal{N}(0, \sigma_B)$. The pixels inside the spot are normally distributed around an average hybridization level μ with hybridization signal variability σ, while those outside the spot have a mean value 0:

$$p(u(r)|\phi) = \begin{cases} \mathcal{N}(\mu, \sqrt{\sigma^2 + \sigma_B^2}) & \text{if } r \in \mathcal{R}(\tau) \\ \mathcal{N}(0, \sigma_B) & \text{if } r \in \mathcal{R}^C(\tau). \end{cases} \tag{7.1.1}$$

The likelihood of the measurement u given the parameters ϕ is

$$\begin{aligned} L(u|\phi) &= \prod_{z \in \mathcal{R}(\tau)} \mathcal{N}(u(\cdot); \mu, \sqrt{\sigma^2 + \sigma_B^2}) \cdot \prod_{z \in \mathcal{R}^C(\tau)} \mathcal{N}(u(\cdot); 0, \sigma_B) \\ &\propto \left(1 + \frac{\sigma^2}{\sigma_B^2}\right)^{-\frac{N(\tau)}{2}} \exp\left\{-\frac{1}{2} \sum_{r \in \mathcal{R}(\tau)} \left[\frac{(u(\cdot) - \mu)^2}{\sigma^2 + \sigma_B^2} - \frac{(u(\cdot))^2}{\sigma_B^2}\right]\right\} \end{aligned} \tag{7.1.2}$$

where $N(\tau) = \sum_z 1_{z \in \mathcal{R}(\tau)}$ is the number of signal measurements in the sub-image.

The shape τ and the signal $\theta = [\mu, \sigma]$ parameters are considered independent *a priori*. For each parameter a uniform prior is assumed

$$\begin{aligned} P_{x_0}(x_0) &\sim U[x_{\text{MIN}}, x_{\text{MAX}}], \\ P_{y_0}(y_0) &\sim U[y_{\text{MIN}}, y_{\text{MAX}}], \\ P_r(r) &\sim U[r_{\text{MIN}}, r_{\text{MAX}}], \\ P_\mu(\mu) &\sim U[0, \sigma_{\text{MAX}}], \\ P_\sigma(\sigma) &\sim U[0, \sigma_{\text{MAX}}], \end{aligned}$$

and given the independence of the parameters, the joint prior distribution is given by:

$$P_\phi(\phi) = P_\tau(\tau)P_\theta(\theta) = P_{x_0}(x_0) \cdot P_{y_0}(y_0) \cdot P_r(r) \cdot P_\mu(\mu)P_\sigma(\sigma).$$

In order to estimate the parameters of the model, samples are drawn via Gibbs sampling from the posterior probability density function

$$p(\phi|u) \propto P_\phi(\phi)L(u|\phi).$$

The estimator based on the samples $\phi^{(t)}$ is defined as the average of a part of the samples:

$$\widehat{\phi} = \frac{1}{T - t_0} \left(\sum_{t=t_0+1}^{T} \phi^{(t)} \right),$$

where t_0 is the burn-in period, and the first t_0 samples are discarded. Note that the estimator minimizes the mean square error based on the samples $\phi^{(t)}, t = t_0, \ldots T$. The details of the algorithm can be found in [35, 99].

However classical methods are not directly applicable to high resolution images. Instead we apply them to downsampled versions of the original high resolution images. The downsampling can be done in several ways:

1. integration over a quadratic region of size $q \times q$
2. integration over a quadratic region of size $q \times q$ over the denoised image via wavelet thresholding
3. integration over a quadratic region of size $q \times q$ over the counts of single molecules after detection
4. kernel smoothed counts of single molecules
5. adaptive kernel smoothed counts of single molecules.
6. non parametric wavelet estimation applied to counts of single molecules.

Except for the first case, all other downsampling methods are based on information available only via the high resolution technique. A first issue is the selection of the downsampling region size ($q \times q$) for some of the downsamplings. As we have seen in the MOM estimation approach the size of $q \times q$ influences the estimation of hybridization intensity: for small q the bias is small and variance is high, wile for large q the reverse holds (high bias, but small variance). We use

in our computations $q = 20$. In future we plan implementation of the algorithms based on adaptive kernel smoothed data and wavelet smoothed images.

Correlation tests

In Chapter 5 we have validated the detection results on simulation images, while in Chapter 6 we have tested the separation of point patterns on simulated data. In this section, we perform correlation tests in order to compare the results of our analysis to those obtained by classical, low-resolution, ensemble analysis on the downscaled images of the same data. Since our method results in peaks/pixel concentrations, while the ensemble approach gives mean pixel intensity values we compare the correlation coefficients of the estimated hybridization measures computed by each method with the ground truth concentrations used in the simulations.

For this purpose, 60 sets of images were generated with SNR between 2.85 and 31.6. Each set of images, characterized by a SNR value, contains 15 images. In each image, single molecules were simulated with concentrations $\lambda_1 \in \{0.003, 0.005, 0.007, 0.009, 0.01\}$ peaks/pixels inside a disk of radius 150 pixels. The concentration λ_2 of peaks corresponding to clutter outside this disk was also varied from 0.001 to $\lambda_1 - 0.002$ by steps of 0.002.

In Fig. 7.3 are shown the images corresponding to background value $\lambda = 0.003$, $SNR = 12.52$. The single molecule simulation procedure is the one described in Section 7.1.

In Fig. 7.4 we present the correlation results for the three high resolution algorithms: MOM, EM and level set method. Each point in Fig. 7.4 corresponds to a correlation coefficient computed between the hybridization measure (the estimated λ_1 values) and the true λ_1 values used in simulations. MOM performs slightly worse than the other two algorithms, while EM and the level set method have similarly good performance, with a small advantage for the level set method.

In order to compare the algorithms with state of the art microarray segmentation methods we have created the corresponding low resolution images, by integrating the intensities of the high resolution images over 20×20 pixel patches. The pixel intensities of the downscaled images were modeled as a Gaussian mixture of foreground (signal) and background pixels respectively and the parameters of the mixture were estimated via a maximum likelihood (ML) approach as described in the previous section. Moreover, we have applied the same downsampling and ML estimating procedure to the original high-resolution image after a wavelet denois-

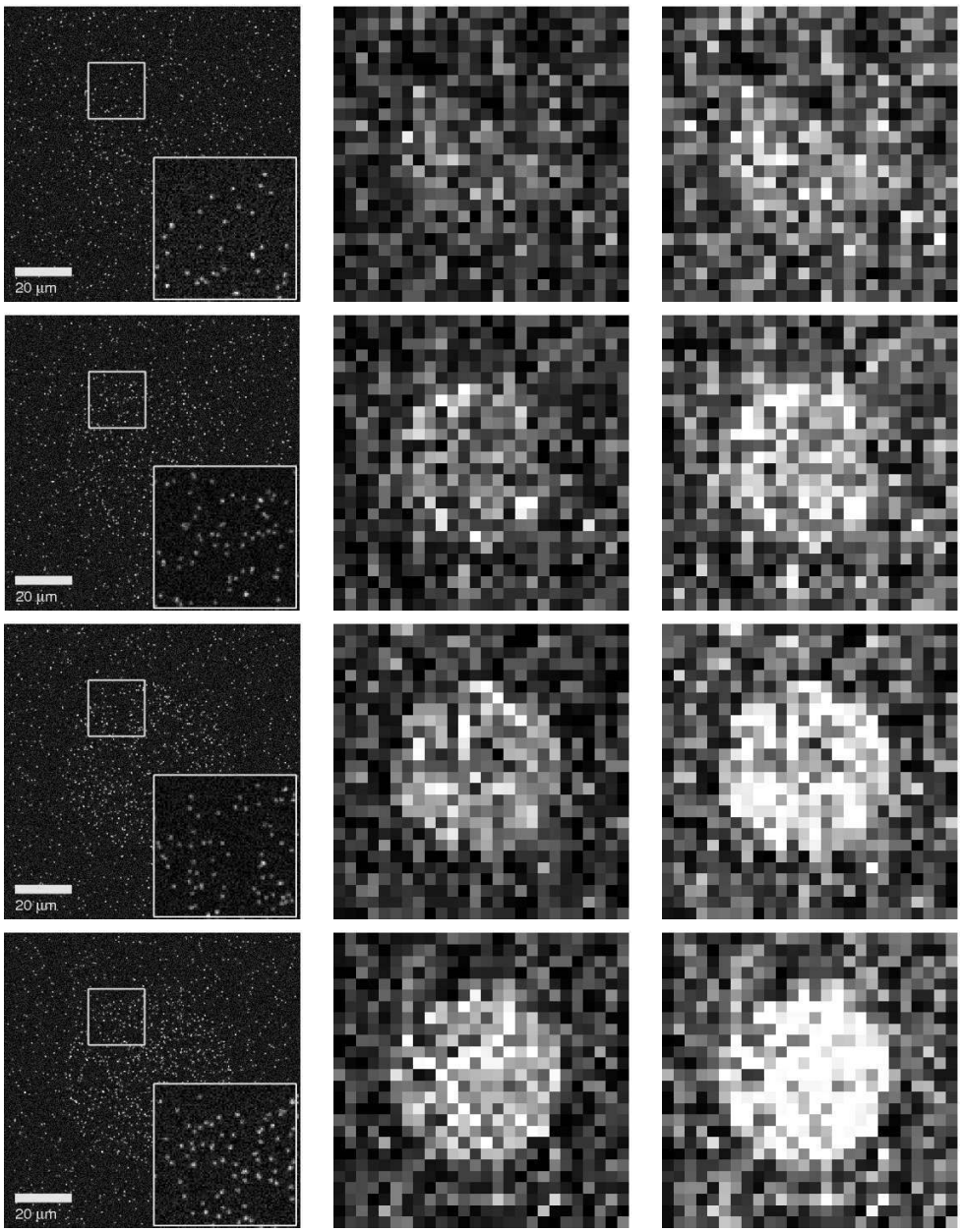


Figure 7.3: Simulation of microarrays spots. Left column: spots at single molecule resolution (200nm pixel size) with different peak concentrations (λ) inside each spot. Starting from the first row up till the fourth down: $\lambda = 0.005, 0.007, 0.009, 011$ peaks per pixel. (Background concentration representing dirt, unspecific binding etc.: 0.003 peaks per pixel). Middle column: the same spots downsampled to $4\mu m$, the size used by existing commercial microarray systems. Right column: The original spots, denoised via wavelet thresholding and then downsampled to $4\mu m$.

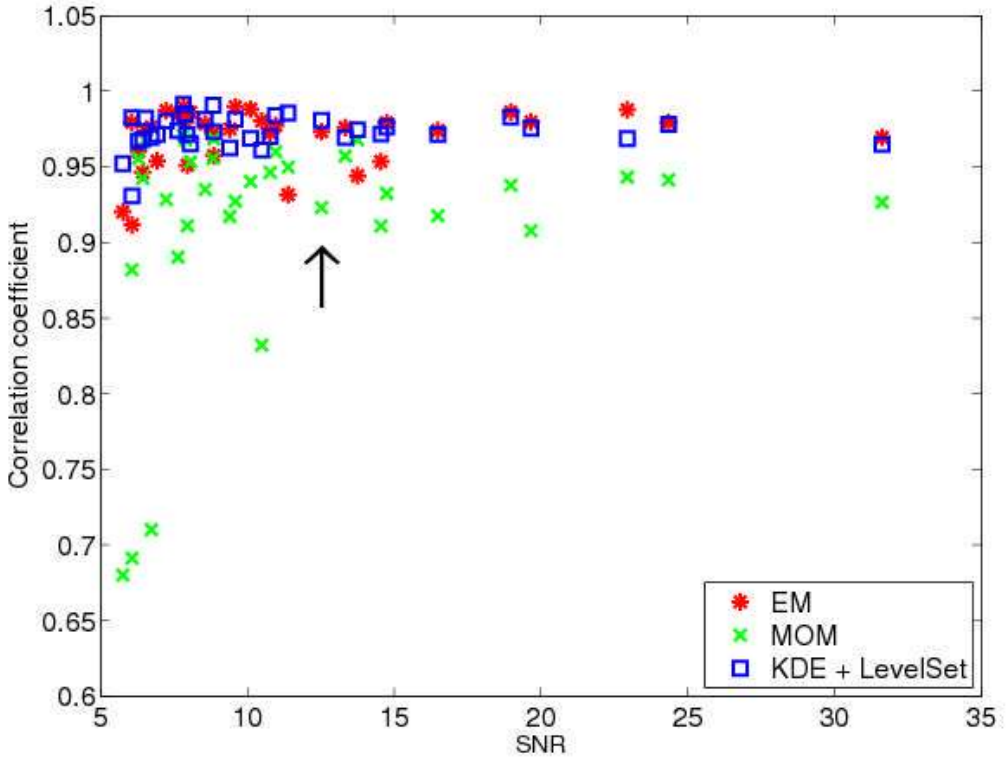


Figure 7.4: Correlations between the estimated and the true signal concentrations for the three high resolution algorithms: MOM, EM and level sets. The arrow indicates the correlation coefficient corresponding to the images in Fig. 7.3.

ing step. Thus we eliminated the effect of the background on the hybridization measure.

We have compared our high resolution analysis with two more state-of-the-art microarray spot segmentation and intensity estimation methods: the first based on the Mann-Whitney rank-sum statistic (M-W) and the MCMC algorithm also presented in the previous section.

The results are visualized in Fig. 7.5. In this case, out of the three high resolution algorithms only the EM method's results are plotted, in order not to overload the figure, together with four low-resolution results, all in different colours. In their case the correlation coefficients are computed based on the mean foreground pixel intensities and the true λ_1 values used in simulations. For the whole range of different SNR corresponding to the test data, our approach (with EM concentration estimation) has higher correlation coefficients than any of the alternative four algorithms. The lowest correlation value for single molecule analysis, 0.77, was obtained at SNR = 3.99. Since some of the low-resolution spot analysis have failed

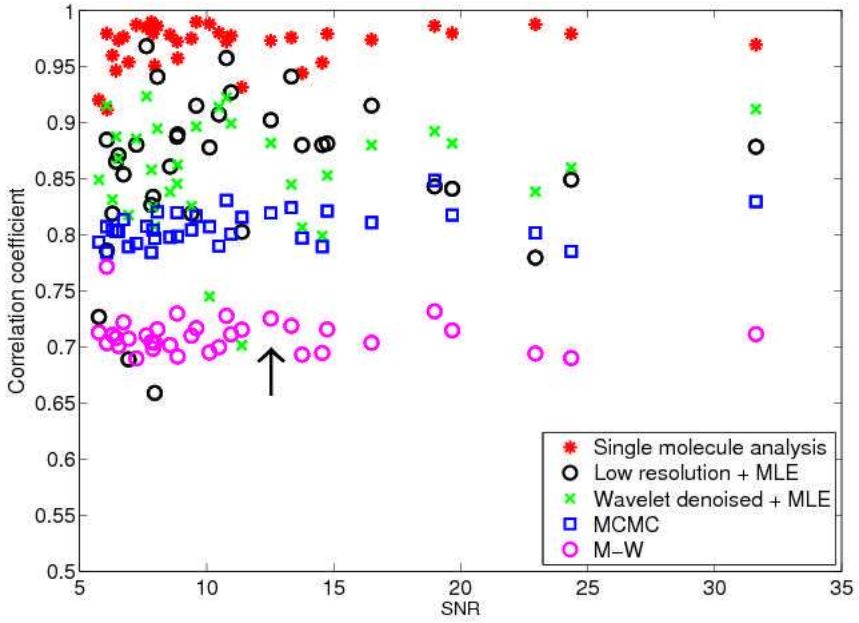


Figure 7.5: Correlations between the estimated and the true signal concentrations. The single molecule analysis performs better than the analysis on the downscaled data (original and denoised via wavelet thresholding). The arrow indicates the correlation coefficient corresponding to the images in Fig. 7.3.

for low SNR image data, we show the results for the data sets with $SNR \geq 5$. We mention that the analysis of a low-resolution microarray spot via the MCMC segmentation method takes approximately 10 minutes [99] for low SNR, this time complexity being a prohibitive factor for use on chips with several thousand spots. Our EM approach takes approximately 90 seconds/spot for average concentration spots, and significantly less for low concentration spots. The mean of the correlation coefficients for the compared algorithms on these simulation datasets are as follows:

- MOM approach: 0.889
- High resolution EM approach: 0.969
- KDE and level set segmentation: 0.973
- Low resolution MLE: 0.858
- MLE on wavelet denoised images: 0.809. However on datasets with $SNR >$ 16 the method outperforms the low resolution MLE.

- MCMC segmentation: 0.809
- M-W segmentation: 0.709.

The first three high resolution approaches perform better than any of the classical low resolution methods.

7.2 Oligonucleotide dilution series

The performance of the algorithms is best measured on real data. However in the case of real images the ground truth is missing and the accuracy of the results is difficult to evaluate. A special design of experiments based on dilution series of oligonucleotide arrays (see Section 3.3 for technology details) provides reliable control for the quality of the results on real images. We have used images of oligonucleotide arrays with four different concentrations: $0.01, 0.1, 1$ and 10amol/μl, and 20 replicated spots for each concentration. The main visual difference between oligonucleotide arrays and cDNA arrays is the uniformity of peak intensities, since in the oligonucleotide array each peak is due to a single dye molecule, (basically meaning that Model 1 in 3.3 is sufficient to describe the imaging process).

Single molecule approaches cannot be applied on the images corresponding to 10amol/μl concentration, since at this concentration peaks due to single molecules overlap and cannot be separated. At this concentration one can successfully use methods typical for ensemble analysis. We shall concentrate in the following on the concentration range of $0.01 - 1$amol/μl.

The results of the high resolution algorithms MOM, EM and KDE for the dilution series data are shown in Fig. 7.7, while the low resolution analysis on the software downsampled images are presented in Fig. 7.6. Each point corresponds to the concentration estimate of one spot. One expects a linear relation on the log-log plot for concentrations up to 1amol/μl. The line in each figure unites the mean values over the replicas of estimated concentrations for the three concentrations. The high resolution algorithms MOM and EM show a close to linear relationship, while the KDE approach is the best in this respect. Among the low resolution methods, the best linearity yields the MLE algorithm applied to the count of the peaks after wavelet based detection, followed by the MW and K-means algorithms. Note that the MLE approach applied to counts of wavelet detected peaks is in fact a high resolution technique, since it is based on information available only at high resolution (the peaks representing single molecules). The mean value and

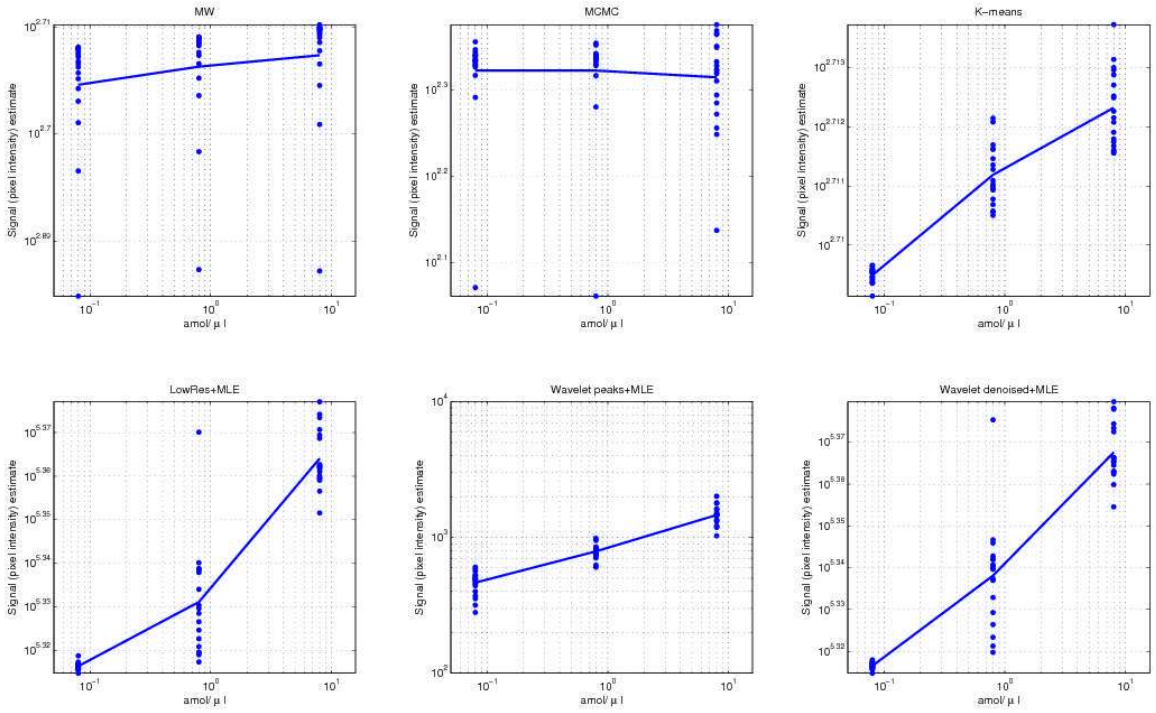


Figure 7.6: Comparison of the concentration estimates for the dilution series in case of six low resolution algorithms

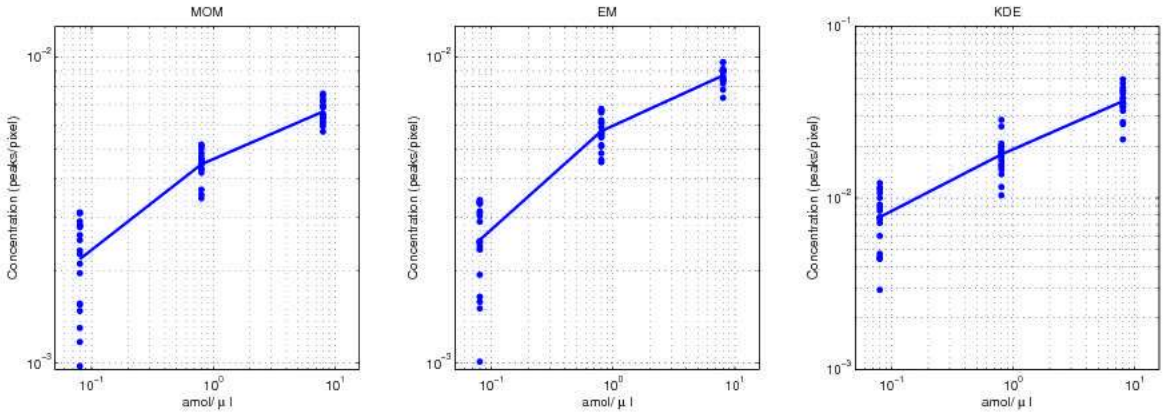


Figure 7.7: Comparison of the concentration estimates for the dilution series in case of the three high resolution algorithms

standard deviation of the signal estimates are summarized in Table 7.1. From the point of view of the separability of the three classes of spot concentrations the high resolution techniques are definitely superior, with the best results achieved by the

Table 7.1: Results of signal concentration estimation on dilution series for several algorithms (concentrations given in amol/μ l)

Algorithm	**c=0.01**		**c=0.1**		**c=1**	
	mean	std	mean	std	mean	std
MW	506.5310	6.6084	508.5122	6.2745	509.7575	6.4127
MCMC	210.4264	24.0670	210.3824	24.9649	206.6088	25.2465
K-means	512.2587	0.1529	514.2674	0.5970	515.6143	0.7614
LowRes+MLE	207211.85	409.39	214352.35	6335.14	231223.28	3600.98
W. peaks+MLE	464.85	90.17	790.43	94.14	1463.10	244.33
W. denoised+MLE	207254.94	372.30	217812.76	6394.59	233116.49	3726.15
MOM	0.0022	0.0007	0.0045	0.0005	0.0066	0.0005
EM	0.0025	0.0007	0.0057	0.0007	0.0087	0.0006
KDE	0.0077	0.0028	0.0180	0.0045	0.0367	0.0070

EM algorithm (see Fig. 7.7).

7.3 Gene expression in multiple myeloma data

A typical application of the ultra-sensitive microarray system is the analysis of cancer stem cells(CSC) in order to characterize their gene and protein expression profile. The problem was briefly presented in Chapter 1. Given the small fraction of stem cells (fraction as low as $< 1\%$ according to [46]) the analysis is extremely difficult if not impossible via the classical ensemble microarray techniques. The ultra-sensitive microarray analysis was used to study the expression profile of putative Multiple Myeloma (MM) stem cells using the human MM cell line NCI-H929.

The sample of 600 ng of total RNA from a CD138− Multiple Myeloma NCIH929 cell line was extracted at the Salzburg Department of Genetics. RNA from CD138− and CD138+ side population was labeled with Alexa-647 dye or with Alexa-555, respectively, and competitively hybridized on three replica slides. Spotting, labeling and imaging were adapted to the conditions imposed by the small amount of mRNA. Selecting 400 genes for which all (image) quality criteria were satisfied on all the replicas, we have found 64 differently expressed genes. The expression levels of several of these genes were tested and confirmed by employing a different technique — qPCR. The qPCR tests were performed at the Department of Genetics of Salzburg University. The results of the qPCR approach are in good correlation with the gene expression estimated from the microarray data. The results are summarized in Fig. 7.3. For each gene the bar is proportional to the ratio for CD138± of the respective gene expressions. The colored stars mark genes

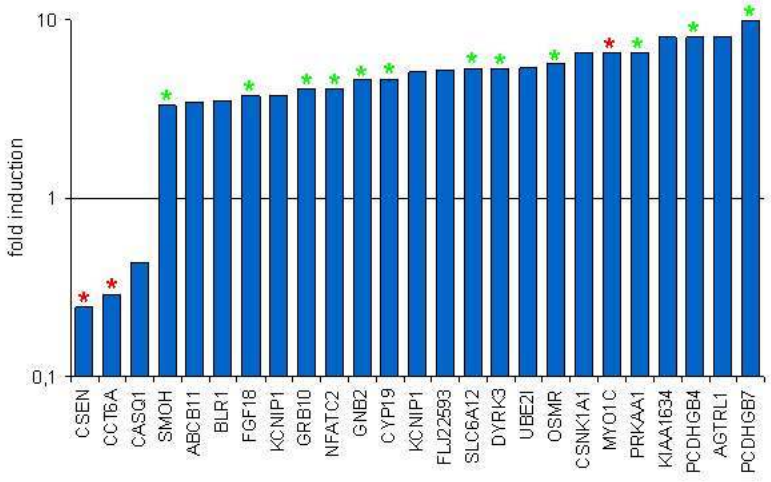


Figure 7.8: The expression profiles of several side population genes showing repressors and over-expressers. The repressed genes like CSEN, CCT6A and CASQ1 were either not analyzed with the qPCR method or showed not interpretable results. Rest of the presented genes show a higher expression level, and are in a good agreement with the microarray results

compared with PCR. The gene ratios confirmed in the validation are marked with a green star, while the three red stars show the ratios diverging in their expression profile from the qPCR results.

Chapter 8

Conclusion

In the concluding chapter of this thesis, we are giving a brief overview of the image processing tasks implied by the new technique of ultra-sensitive microarrays as well as the results discussed in detail in the previous chapters. We shall also point out possible applications of our study to other problems in microscopy as well as other fields. Further ideas and lines of study will be pointed out in the second section of the chapter that concludes this thesis.

8.1 Summary

This work was motivated by the new technology of high-resolution microarray analysis, representing a combination between fast single molecule imaging and the microarray technique. Each step of the microarray technique had to be evaluated and adapted to the novel challenges, the most crucial of which being the small amount of target mRNA typically used in the new technology.

Spotting, labeling and imaging were adapted to the new conditions and the technical innovation in the biological and physical aspects of the problem resulted in a new type of images to be analyzed. The challenges of the new technique from an image processing and analysis point of view are related to:

The modified content of the microarray image. Instead of a microarray circular spot representing the signal over a (uniform) background we have to deal with information at two resolution levels: at high resolution the single molecule peaks can be detected, and at a larger scale these peaks form the hybridized spot imposed over a background of clutter. The detection of single molecule peaks was implemented based on a wavelet thresholding algorithm. The position of the single peaks were modeled as a superposition of two

homogeneous Poisson processes corresponding to hybridization signal and clutter. The intensity of the signal process was proposed as a new measure of hybridization (Chapter 6).

Wide range of SNR and combined Poisson-Gaussian noise models. Single molecule imaging poses new challenges related to SNR of the image and the contrast of signal. Typically detection algorithms have to be very robust over a wide range of SNR and contrast parameters. The support of the signal is very small (of the order of a few pixels) which makes the detection problem difficult. Moreover, the concentration of the peaks (the sparsity of the signal) is influencing the detection, and therefore was also taken into account in our examination. We studied in Chapter 5 the behaviour of the detection algorithms and summarized their results on systematic simulations of images, over a range of simulation parameters and implicitly SNRs, as well as on real images. For the conditions imposed by single molecule imaging the wavelet thresholding methods proved reliable and accurate. Although the detection is based on the control of false positive cases, for the evaluation of the algorithms we measured both the false positive as well as the false negative detections.

The size of the images and complexity of algorithms. To each microarray slide correspond approximately 22GB of image data (partitioned in several subimages). Therefore the time and space complexities of the chosen algorithms bear a crucial importance. Wavelet thresholding methods show good performance with respect to time and memory, thanks to the one step approach and to the special form of operators, (diagonal operators) used. The concentration estimation is very fast in the case of MOM approach, and slightly more computationally intensive in case of the EM and the level set algorithms, with the advantage of increased accuracy. Each additional step, such as improved noise variance estimation or variance stabilization transforms further increase the computation time. Nevertheless the proposed algorithms are on average 10 times faster than some of the state-of-the art algorithms analyzing low-resolution microarray images, such as [99] (see Section 7.1). However, in order to target a wider public, we plan improvements of both memory and time performance.

Low mRNA concentration. The low fluorophore concentration results in few single molecule peaks, which in turn, at low resolution, translate to dim

spots. The wavelet-based detection algorithm used for single molecule detection was successfully applied also to the detection of microarray spots in downsampled (originally high-resolution) microarray images (Section 7.3). The gridding algorithm has also been adapted so that the grid pattern matching works robustly even in the case of missing (or dim) column and/or lines of microarray spots.

Understanding the advantages and disadvantages of the new technique. The access to high resolution information permitted modeling at single molecule and single dye level. It also helped in gaining understanding on image formation, both in the low-resolution and high-resolution cases. Modeling both cases allowed a better identification of the sources of bias and errors in each case, as well as to perform a better comparison of the two techniques (as described in Chapters 2 and 7).

Validation of the results. In order to validate the results the algorithms were tested on simulated and real data. For the real data the ground truth is not available, however with the help of specially designed experiments, such as the dilution series of oligonucleotide arrays the correctness of the results could be evaluated (see Section 7.2). The results of our approach on a real data set based on multiple myeloma were confirmed by an independent technique (qPCR) as described in Section 7.3.

We have presented in this work a comprehensive image processing solution for a new microarray technique, analyzing the similarities and differences to the state-of-the art approaches used in the standard microarray applications. Besides the signal processing and image analysis approach this work attempted to accommodate the point of view of the biologists and biophysicists involved, since it is the result of a close and continuous collaboration in the framework of a multidisciplinary project.

8.2 Outlook

Beyond applications to the ultra-sensitive microarray technology, the study of algorithms analyzing single molecule images has an interest in its own. All the present and future applications of single molecule imaging might benefit of the algorithms detecting single molecules, with special concern to low level SNR, the specific noise model and single molecule concentrations.

A reliable method, based on wavelet thresholding, of finding significant pixels in a microscopy image was described in Chapter 5. The special focus was the detection of pixels due to single molecules. Further analysis of the values of these pixels might provide insight in the temporal behaviour of fluorophores, characterizing the phenomena known as *blinking* and *bleaching* (Section 2.1). We envisage to address the problem of subpixel accuracy, which classically is based on fitting a $2d$ Gaussian shape to the pixel values corresponding to the single molecule image. Bobroff [15] discussed the importance of the signal support over which this fitting is performed. We will investigate the effect of signal support selection via wavelet thresholding on the fitting procedure. Finally, we will extend the analysis of molecule intensities, which for the moment only discriminates between two classes of signal intensity: due to single dyes and due to dirt, to more general stoichiometry problems, as e.g. the one described in [83].

We plan to study different models (e.g. truncated and censored distributions) for the description of single molecule positions and counts, knowing that the wavelet detection methods have a certain "resolution" (see Section 6.4), thus introducing a truncation of the data. We would like to look into the benefits of wavelet methods as density estimators in case of spatial point patterns in general and in microscopy in particular.

A possible line of study in the future is the utility of spatial point pattern approaches to microscopy images of single molecules. There are already a few attempts of using spatial patterns in co-localization [122], tracking [64] and at a higher scale, in the study of the differences between the patterns of nuclei of malignant epithelial tumor (adenocarcenomas) cells as opposed to normal cells [79]. We are interested in characterizing videomicroscopy image sequences of single molecules previous to tracking, in order to assess their quality, make inferences about the dynamics and finally, facilitate the tracking. We believe that point pattern analysis — both as hypothesis testing and parameter estimation — will prove to be a useful tool in PALM imaging [13].

As previously mentioned, optimized versions of the algorithms and their implementations, more efficient both in run-time and memory, are planned to be developed in order to improve the use of the proposed approach.

Appendix A

Outliers and variance-covariance estimators

Several intuitive descriptions of outliers have been proposed in the literature. We adopt the one offered in [98], according to which outliers are "*observations that do not follow the pattern of the majority of the data*".

The deviation of the outliers from the majority of the data is measured by a convenient measure, such as the quadratic distance measure between data x and a location y, given by the (squared) Mahalanobis distance:

$$MD^2(x, y) = (x - y)^T S^{-1}(x - y), \tag{A.0.1}$$

where y is a d-dimensional vector and S is a positive definite symmetric $d \times d$ dimensional matrix. The Mahalanobis distance represents the distance from the point x to the "center" of the data cloud, taking into account the shape of the point cloud. Typically if this distance is larger than a predetermined threshold, x is considered an outlier.

The importance of the Mahalanobis distance is justified by the aim to find methods that are affine equivariant, so that "*measurement scale changes and other linear transformations do not alter the behavior of analysis method*" (see [93]).

Usually, y and S are replaced by estimations of the location and scatter, e.g. the standard estimators, average of the sample $\bar{x} = 1/n \sum_{i=1}^{n} x_i$ for location and for scatter the sample covariance matrix: $\hat{S} = n^{-1}(x - \bar{x})^T(x - \bar{x})$.

For multivariate normally distributed d-dimensional data the values are approximately χ^2 distributed with d degrees of freedom ([48]).

The standard sample location and shape parameters are not robust estimators

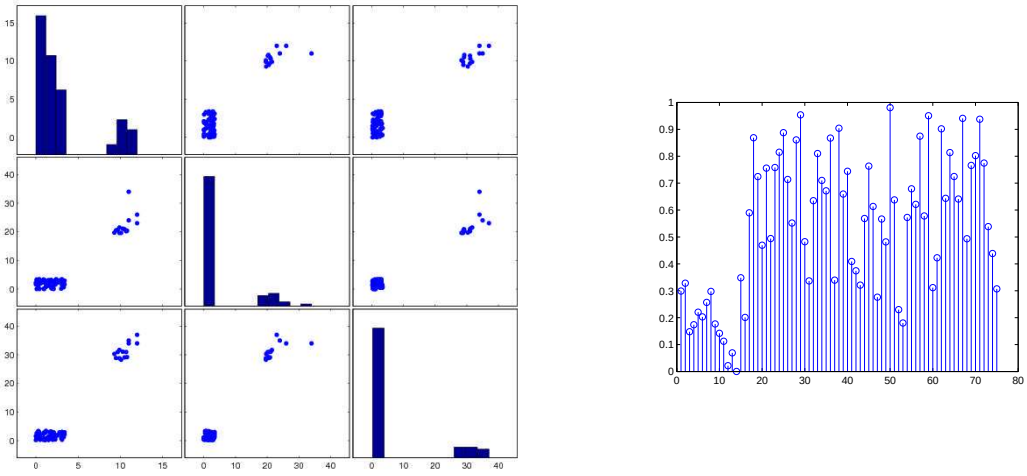


Figure A.1: (Left) Hawkins-Bradu-Kass 3d data. (Right) p-values of the data plotted against the data index. The first 14 points represent outliers.

(the higher the moment the higher the influence of the outlier), leading to masking and swamping phenomena, in the case of multiple outliers or clusters of outliers corrupting the data.

The *masking effect* means that an outlier or an outlying subset is undetected because of the presence of another, usually adjacent subset, while *swamping effects* occur when good observations are incorrectly identified as outliers, usually due to the presence of a remote subset of observations([50]).

These phenomena prevent outliers having a large Mahalanobis distance value comparatively to the rest of the data cloud. The dataset due to Hawkins, Bradu, and Kass, described also in [96] (p. 94), illustrates the failure of the Mahalanobis distance with standard parameter estimation to uncover outliers. Figure shows the Hawkins-Bradu-Kass data, consisting of 75 three-dimensional points and their respective p-values assuming a χ^2 distribution with three degrees of freedom. At least ten p-vales out of the first 14 values corresponding to outliers are not significantly different from the rest of p-values of the clean data.

A possible solution to the aforementioned problem consists of replacing the standard location (T) and scatter (S) estimates by robust counterparts, since the underlying idea in robust estimation is to eliminate or drastically reduce the influence of outliers on the estimates. Such an approach would preserve the nice intuitive idea of using Mahalanobis distance for detecting outliers.

A measure of robustness of an estimator T is its *breakdown point* ε^*, the smallest

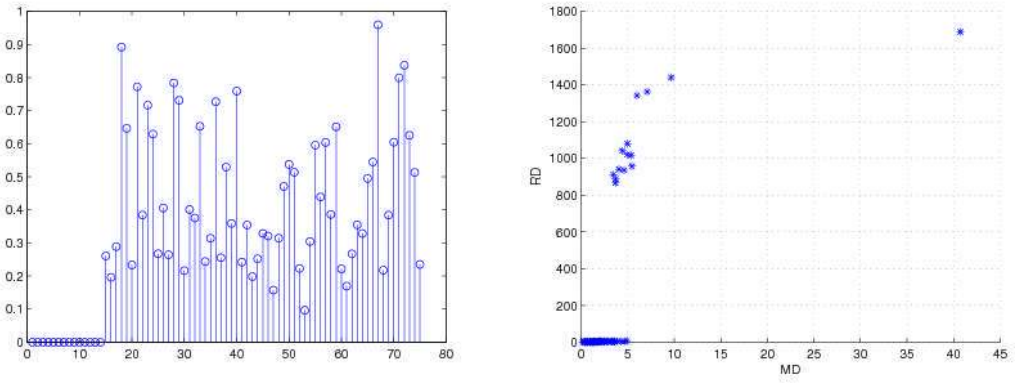


Figure A.2: (Left) p-values of the robust Mahalanobis distances for the HBK data. (Right) The Mahalanobis distances for the HBK data computed according to (A.0.1) with the standard location and scatter estimates plotted against the robust Mahalanobis distance (MCD estimates). The 14 representing outliers are easily identifiable in the case of robust distances (y axis), and blend in the rest of the data for the standard estimates (x axis).

fraction of contamination that can cause T to take arbitrarily large values:

$$\varepsilon^*(T,X) = \min\left\{\frac{m}{n} : sup_{X'} \|T(X') - T(X)\| = \infty, |X'| = m, |X| = n, |X' \cap X| = n - m\right\}$$

Illustrative examples in case of location estimation, are the arithmetic mean with the breakdown point 0, and the median whose breakdown point is 50%.

A.1 One-dimensional robust estimators of location and scale

For the univariate case, $d = 1$, several robust estimates were proposed (see for example [52]). The most popular one-dimensional location and scale estimators are the (one-step) M-estimators, also called *generalized maximum likelihood*.

The classical maximum likelihood estimators $T_n = T(x_1, \ldots, x_n)$ maximizing $\prod_{i=1}^n f_{T_n}(x_i)$ or equivalently minimizing:

$$\sum_{i=1}^{n} (-\ln f_{T_n}(x_i)) =^! \min_{T_n} \tag{A.1.1}$$

are replaced by the following minimization problem

$$\sum_{i=1}^{n} \rho\left(x_i, T_n\right) =^{!} \min_{T_n}, \tag{A.1.2}$$

where ρ is a function defined on $\mathcal{X} \times \Theta$, (where $\mathcal{X}$ is the range of the variable and Θ is the parameter space), with derivative $\psi(x, \theta) = \frac{\partial}{\partial \theta}\rho(x, \theta)$, such that (A.1.2) is replaced by:

$$\sum_{i=1}^{n} \psi\left(x_i, T_n\right) = 0.$$

One of the simplest robust location estimator is the median of the sample. As for scale, the median absolute deviation (MAD) is a popular choice:

$$\psi_{\mathrm{MAD}}(x) = \operatorname{sign}\left(|x| - \Phi^{-1}(3/4)\right)$$

resulting in $\mathrm{MAD}(x) = \operatorname{median}|x - \operatorname{median}(x)|/0.674$. This is the estimate of scale for the distribution of wavelet coefficients proposed in [36, 37] and used in wavelet thresholding.

We have implemented two alternatives to the MAD robust scale approximation are described in [95]:

$$S_n = c_1 \cdot \operatorname{med}_i \left\{\operatorname{med}_j |x_i - x_j|\right\},$$

where $\operatorname{med}_i x_i$ denotes the median of the sample $\{x_1, x_2, \ldots x_n\}$ and

$$Q_n = c_2 \cdot \{|x_i - x_j| : i < j\}\,(k),$$

Their advantage is greater (Gaussian) efficiency and no underlying assumption of a symmetric distribution, with the drawback of higher computation time (which was especially limiting for the large microarray images).

A.2 Robust covariance matrix estimation

The robust estimation in multivariate case, and especially $d > 2$, is much more challenging than in the univariate one. Outliers are difficult or impossible to detect even by visual inspection and sample covariance matrices are very sensitive to outliers. Moreover, Gnanadesikan and Kettenring ([48]) identify several other causes that make outlier detection particularly challenging in high dimensions:

outliers may corrupt not only location and scale, but also orientation parameters (i.e. correlations), the view of outlier as an extremal value does not hold in high dimensions, an outlier might be characterized by a gross error in one dimension, or milder errors in several (or all) dimensions (relating the problem to that of sparsity of the error vector).

A presentation of the difficulties of the task as well as a comparison of some of the solutions is given in [93]. The robust approaches are classified in [59] in three categories:

- Robust estimation of individual matrix elements of the covariance matrix
- Robust estimation of variances in sufficiently many selected directions, to which a quadratic form is then fitted
- Direct estimation of shape matrix for some elliptical distribution.

The breakdown point of all affine equivariant M-estimators is at most $\frac{1}{d+1}$ or even lower, if the outliers belong to a lower dimensional space. This makes the multivariate location and shape problem a very difficult problem to tackle. However a popular choice from the combinatorial category of location and shape estimators in the multivariate case, the Minimum Covariance Determinant (MCD) estimates, have been proved to work well over a wide range of situations.

The objective of the MCD estimator, first introduced in [94] is to find h observation out of the total n whose classical covariance matrix has the lowest determinant. The location and scatter are the standard estimates based on the chosen h observations.

Formally, given n data points, $\{x_1, \ldots, x_n\}$, the MCD is based on the subsample $\{x_i, i \in H\}$, where $H \subset \{1, 2, \cdots, n\}$ of size $h, h \leq n$, that minimizes the determinant of the covariance matrix, more precisely:

$$MCD = (T_H, S_H), \tag{A.2.1}$$

where

$$H : \{ \{i \in H\}, |H| = h : \det(S_H) \leq \det(S_K), \forall K, \det(K) = h\}$$
$$T_H = \frac{1}{h} \sum_{i \in H} x_i$$
$$S_H = \frac{1}{h} \sum_{i \in H} (x_i - T_H)(x_i - T_H)^T.$$

The estimator is affine equivariant because the determinant of the covariance matrix of the transformed data satisfies:

$$\det(A^T C A) = (\det(A))^2 \det(C)$$

The main idea of the MCD algorithm is that of obtaining a sample subset $H_2 \subset \{1, \ldots, n\}$ starting from the subset H_1 with the same cardinality, $|H_1| = |H_2| = h$, such that the new subset has a smaller covariance matrix determinant than the initial one. The process of obtaining a new subset H_2 is called a C-step in [97] and it is summarized in the following theorem.

We denote with $d_k(i) = MD(x_i, T_k, S_k)$ the Mahalanobis distance of the ith point, $i \in H_k$, with respect to the standard estimators of location and shape, T_k and S_k, of the subset H_k.

Theorem A.2.1. *Given $X = \{x_1, \ldots, x_n\}$ p-variate observations and the subset $H_1 \subset X$, such that $|H_1| = h$, if $\det(S_1) \neq 0$ and*

$$H_2 = \{\pi(i), i = 1, \ldots h : d_1(\pi(1)) \leq d_1(\pi(2)) \leq \ldots \leq d_1(\pi(n))\},$$

where $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ is the permutation corresponding to the increasing ordering of the data then

$$\det(S_2) \leq \det(S_1).$$

The equality holds if and only if $T_2 = T_1$ and $S_2 = S_1$.

A proof of the theorem is given in [97].

The exact MCD solution can be found only in very simple cases, instead approximations based on random search [96], steepest descent [54] and heuristic search optimization [92, 93] were proposed. We use a computationally efficient version of

the random search MCD algorithm, introduced in [97] and implemented in LIBRA [112].

The properties of the MCD estimators were extensively studied, for example in order to choose the parameter h the maximum breakdown point of the MCD estimator is achieved for $h = \left\lfloor \frac{n+p+1}{2} \right\rfloor$. However for a good trade-off between robustness and efficiency the authors recommend $0.75 \cdot n$ (see [96, 53]).

A.3 Distributions of Mahalanobis distances

The Mahalanobis squared distances give a one-dimensional measure of how far a point is from a location with respect to a shape. Several approximations of the distribution of the Mahalanobis distance are known in the literature. Some results are given in [48, 53] and summarized as follows.

Consider n multivariate data points in $\mathbb{R}^p$, $X_i \sim \mathcal{N}(\mu, \Sigma)$, and S an estimate of Σ. The three distributional results for distances based on multivariate normal data are given below.

1. Considering the true parameters μ and Σ known, if the data are normal distributed, the Mahalanobis distances have an exact χ^2_d distribution:

$$MD(x_i, \mu, \Sigma) \sim \chi^2_d,$$

with

$$\begin{aligned} \mathbb{E}\left[\mathrm{MD}(x_i, \mu, \Sigma)\right] &= d \\ \mathrm{var}\left[\mathrm{MD}(x_i, \mu, \Sigma)\right] &= 2d. \end{aligned}$$

2. The Mahalanobis distances based on the standard estimates have an exact $\mathcal{B}eta$ distribution:

$$\frac{n}{(n-1)^2}\mathrm{MD}(x_i, T, S) \sim \mathcal{B}eta\left(\frac{d}{2}, \frac{n-d-1}{2}\right),$$

and

$$\begin{aligned} \mathbb{E}\left[\frac{n}{(n-1)}\mathrm{MD}(x_i, T, S)\right] &= d \\ \mathrm{var}\left[\frac{n}{(n-1)}\mathrm{MD}(x_i, T, S)\right] &= 2d\frac{n-d-1}{n+1}. \end{aligned}$$

3. The Mahalanobis distances based on an estimate S of Σ, independent of x_i have an exact $\mathcal{F}$ distribution when μ is the location argument, and an approximate $\mathcal{F}$ distribution when the standard location estimate is used.

$$\frac{n-d}{d(n-1)}\mathrm{MD}(x_i, T, S) \sim \mathcal{F}\left(d, n-d\right),$$

and

$$\begin{aligned}\mathbb{E}\left[\frac{n-d-2}{(n-1)}\mathrm{MD}(x_i, T, S)\right] &= d \\ \mathrm{var}\left[\frac{n-d-2}{(n-1)}\mathrm{MD}(x_i, T, S)\right] &= 2d\frac{n-2}{n-d-4}.\end{aligned}$$

The above distributions can be incorporated in outlier detecting algorithms. Quantiles, especially χ^2 quantiles are a straightforward way to identify outliers, although the results are often spoiled by the high amount of false positives (see [98]).

Bibliography

[1] http://www.invitrogen.com/site/us/en/home/support/Research-Tools/Fluorescence-SpectraViewer.html.

[2] http://www.genome.gov.

[3] F. Abramovich and Y. Benjamini. Adaptive thresholding of wavelet coefficients. *Computational Statistics and Data Analysis*, 22:351–361, 1996.

[4] F. Abramovich, Y. Benjamini, D. Donoho, and I.M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34(2):584–653, 2006.

[5] V.T. Acharya and A.K. Ray. *Image processing: Principles and Applications*. John Wiley & Sons, 2005.

[6] J. Angulo. Polar modelling and segmentation of genomic microarray spots using mathematical morphology. *Image Anal. Stereol.*, 27:107–124, 2008.

[7] Jesus Angulo and Jean Serra. Automatic analysis of DNA microarray images using mathematical morphology. *Bioinformatics*, 19(5):553–562, 2003.

[8] F.J. Anscombe. The transformation of Poisson, binomial and negative-binomial data. *Biometrika*, 35(3/4):246–254, 1948.

[9] A. Antoniadis. Wavelet methods in statistics: Some recent developments and their applications. *Statistics Surveys*, 1:16–55, 2007.

[10] Y. Balagurunathan, N. Wang, E. R. Dougherty, D. Nguyen, Y. Chen, M.L. Bittner, J. Tent, and R. Carroll. Noise factor analysis for cdna microarrays. *Journal of Biomedical Optics*, 9(4):663–678, 2004.

[11] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57(1):289–300, 1995.

[12] T. Berger and J.O. Strömberg. Exact reconstruction algorithms for the discrete wavelet transform using spline wavelets. *Applied and Computational Harmonic Analysis*, 2:392–397, 1995.

[13] E. Betzig, G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J. S. Bonifacino, M. W. Davidson, J. Lippincott-Schwartz, and H. F. Hess. Imaging intracellular fluorescent proteins at nanometer resolution. *Science*, 313(5793):1642–1645, 2006.

[14] K. Blekas, N. P. Galatsanos, A. Likas, and I. E. Lagaris. Mixture model analysis of DNA microarray images. *IEEE Transactions on Medical Imaging*, 24:901–909, 2005.

[15] N. Bobroff. Position measurement with a resolution and noise-limited instrument. *Rev. Sci. Instrum.*, 57(6), 1986.

[16] B.M. Bolstad, R.A. Irizarry, M. Astrand, and T.P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[17] J. Boulanger, J.-B. Sibarita, Ch. Kervrann, and P. Bouthemy. Non-parametric regression for patch-based fluorescence microscopy image sequence denoising. In *Proc. IEEE Int. Symp. on Biomedical Imaging: from nano to macro (ISBI)*, 2008.

[18] S. Byers and A. E. Raftery. Nearest-neighbor clutter removal for estimating features in spatial point processes. *Journal of the American Staistical Association*, 93:577–584, 1998.

[19] E.J. Candès. Modern statistical estimation via oracle inequalities. *Acta Numerica*, 15:257–325, 2006.

[20] T. F. Chan and L. A. Vese. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277, 2001.

[21] T. F. Chan and L. A. Vese. A multiphase level set framework for image segmentation using the mumford and shah model. *International Journal of Computer Vision*, 50(3):271–293, 2002.

[22] Y. Chen, E.R. Dougherty, and M.L. Bittner. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*, 2(4):364–374, 1997.

[23] E. Chudin, S. Kruglyak, S.C. Baker, S. Oeser, D. Barker, and T.K. McDaniel. A model of technical variation of microarray signals. *Journal of Computational Biology*, 13(4):996–1003, 2006.

[24] C.K. Chui. *An introduction to wavelets*. Academic Press, New York, 1992.

[25] M. Clyde, G. Parmigiani, and B. Vidakovic. Multiple shrinkage and subset selection in wavelets. *Biometrika*, 85:391–401, 1998.

[26] N. Cressie. *Statistics for spatial data*. Wiley series in probability and mathematical statistics. Applied probability and statistics section, 1991.

[27] L. Şendur and I.W. Selesnick. Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency. *IEEE Transactions on Signal Processing*, 50(11):2744–2756, 2002.

[28] Z. Cvetkovic and M. Vetterli. Oversampled filter banks. *IEEE Transactions on Signal Processing*, 46(5), 1998.

[29] D.J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes*, volume I. Elementary Theory and methods. Springer, 2003.

[30] I. Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[31] J. Boutel de Monvel, E. Scarfone, S. Le Calvez, and M. Ulfendahl. Image-adaptive deconvolution for three-dimensional deep biological imaging. *Biophysical Journal*, 85:3991–4001, December 2003.

[32] P. Dedecker, J. Hofkens, and J. Hotta. Diffraction-unlimited optical microscopy. *Materials Today*, pages 12–21, 2008.

[33] R.C. Deonier, S. Tavaré, and M.S. Waterman. *Computational Genome Analysis. An introduction*. Springer, 2005.

[34] P.J. Diggle. *Statistical Analysis of Spatial Point Patterns*. Edward Arnold, 2003.

[35] A. Dogandžić and B. Zhang. Bayesian NDE defect signal analysis. *IEEE Transactions on Signal Processing*, 55(1), 2007.

[36] D. Donoho and I. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[37] D.L. Donoho. De-noising by soft-thresholidng. *IEEE Transactions on Information Theory*, 41(3):613–627, 1995.

[38] D.L. Donoho and I.M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90:1200–1224, 1995.

[39] S. Dudoit, Y.H. Yang, M.J. Callow, and T.P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12:111–139, 2002.

[40] R.P. Ekins and F.W. Chu. Multianalyte microspot immunoassay-microanalytical "compact disc" of the future. *Clinical Chemistry*, 37(1955-1967), 1991.

[41] J. Fan, P. Tam, G. Vande Woude, and Y. Ren. Normalization and analysis of cDNA microarrays using within-array replications applied to neuroblastoma cell response to a cytokine. *PNAS*, 101(5):1135–1140, 2004.

[42] S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer New York, 2006.

[43] E. Füreder-Kitzmüler, J. Hesse, A. Ebner, H. J. Gruber, and G. J. Schütz. Non-exponential bleaching of single bioconjugated Cy5 molecules. *Chemical Physics Letters*, 404:13–18, 2005.

[44] H. Y. Gao. Wavelet shrinkage denoising using the non-negative garrote. *J. Comp.Graph.Statist.*, 7:469–488, 1998.

[45] H. Y. Gao and A.G. Bruce. Waveshrink with firm shrinkage. *Stat.Sinica*, 7:855–874, 1997.

[46] N. Ghosh and W. Matsui. Cancer stem cells in multiple myeloma. *Cancer Letters*, 277(1):1–7, 2009.

[47] J.-F. Giovanelli and A. Coulais. Positive deconvolution for superimposed extended source and point sources. *Astronomy and Astrophysics*, 439:401–412, 2005.

[48] R. Gnanadesikan and J. R. Kettenring. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28(1):81–124, 1972.

[49] R. Gottardo, J. Besag, A. Murua, and M. Stephens. Probabilistic segmentation and intensity estimation for microarray images. *Biostatistics*, 7(1):85–99, 2006.

[50] A.S. Hadi and J.S. Simonoff. Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, 88(424):1264–1272, 1993.

[51] D.A. Hall, J. Ptacek, and M. Snyder. Protein microarray technology. *Mechanisms of Ageing and Development*, 128:161–167, 2007.

[52] F. R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust Statistics. The Approach based on Influence Functions*. Wiley & Sons, 1986.

[53] J. Hardin and D. M. Rocke. The distribution of robust distances. *Journal of Computational and Graphical Statistics*, 14(4):928–946, 2005.

[54] D. M. Hawkins. The feasible solution algorithm for the minimum covariance determinant estimator. *Computational Statistics and Data Analysis*, 17(2):197–210, 1994.

[55] J. Hesse, J. Jacak, M. Kasper, G. Regl, T. Eichberger, M. Winklmayr, F. Aberger, M. Sonnleitner, R. Schlapak, S. Howorka, L. Muresan, A. Frischauf, and G. J. Schütz. RNA expression profiling at the single molecule level. *Genome Research*, 16:1041–1045, 2006.

[56] J. Hesse, M. Sonnleitner, A. Sonnleitner, G. Freudenthaler, J. Jacak, O. Höglinger, H. Schindler, and G.J Schütz. Single-molecule reader for high-throughput bioanalysis. *Anal. Chem.*, 76:5960–5964, 2004.

[57] T. Hisrchfeld. Optical microscopic observation of single small molecules. *Applied Optics*, 15(12):2965–2966, 1976.

[58] S. B. Howell. *Handbook of CCD Astronomy*. Cambridge University Press, 2000.

[59] P.J. Huber. *Robust Statistics*. Wiley & Sons, 2nd edition, 2004.

[60] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan. *Statistical Analysis and Modelling of Spatial Point Patterns*. Wiley & Sons, Ltd., 2008.

[61] J. Jacak, J. Hesse, C. Hesch, M. Kasper, F. Aberger, A. Frischauf, G. Freudenthaler, S. Howorka, and G.J. Sch utz. Ultrasensitive dna detection on microarrays. volume 5699, pages 442–449, 2005.

[62] S. Jaffard and Y. Meyer. On the pointwise regularity of functions in critical besov spaces. *Journal of Functional Analysis*, 175(2):415–434, 2000.

[63] S. Jaffard, Y. Meyer, and R. D. Ryan. *Wavelets. Tools for Science & Technology*. SIAM, 2001.

[64] K. Jaqaman, D. Loerke, M. M. Mettlen, H. Kuwata, S. Grinstein, S.and Schmid, and G. Danuser. Robust single particle tracking in live cell time-lapse sequences. *Nature Methods*, 5:695–702, 2008.

[65] I. Johnstone. Threshold selection in transform shrinkage. In E.D. Feigelson and G.J. Babu, editors, *Statistical challenges in astronomy.Third Statistical Challenges in Modern Astronomy (SCMA III) Conference*, pages 343 – 364. Springer, 2003.

[66] I.M Johnstone and B.W. Silverman. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649, 2004.

[67] I.M Johnstone and B.W. Silverman. Empirical Bayes selection of wavelet thresholds. *The Annals of Statistics*, 33(4):1700–1752, 2005.

[68] D. Karlis and E. Xekalaki. A zero frequency alternative method to the moment method of estimation in finite Poisson mixtures. *Journal of Statistical Computation and Simulation*, 73(6):409–427, 2003.

[69] S. Knudsen. *Guide to analysis of DNA microarray data*. Wiley-Liss, second edition edition, 2004.

[70] E.L. Korn, J.K. Habermann, M.B. Upender, T. Ried, and L.M. McShane. Objective method of comparing DNA microarray image analysis systems. *Bioimaging*, 36(6):960–967, 2004.

[71] D.P Kreil and R. R. Russel. There is no silver bullet - a guide to low-level data transforms and normalisation methods for microarray data. *Brief. Bioinform.*, 6(1):86–97, 2005.

[72] M.R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and Related Properties of Random Sequences and Processes*. Springer Series in Statistics. Springer, 1983.

[73] A.M. Levin, D. Ghosh, K.R. Cho, and S.L.R. Kardia. A model-based scan statistic for identifying extreme chromosomal regions of gene expression in human tumors. *Bioinformatics*, 21:2867–2874, 2005.

[74] Q. Li, C. Fraley, R.E. Bumgarner, K.Y. Yeung, and A.E. Raftery. Donuts, scratches and blanks: robust model-based segmentation of microarray images. *Bioinformatics*, 21(12):2875–2882, 2005.

[75] D.J. Lockhart and E.A. Winzeler. Genomics, gene expression and DNA arrays. *Nature*, 405:829–836, 2000.

[76] A.K. Louis, P. Maas, and A. Rieder. *Wavelets*. Teubner Studienbücher, 1998.

[77] Mallat. A theory of multiresolution signal decomposition: The wavelet represntation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7), 1989.

[78] S. Mallat. *A wavelet tour of signal processing*. Academic Press, 1999.

[79] T. Mattfeldt, S. Eckel, and V. Fleischer, F. andSchmidt. Statistical analysis of labelling patterns of mammary carcinoma cell nuclei on histological sections. *Journal of Microscopy*, 235(1):106–118, 2009.

[80] G.J. McLachlan, K.A. Do, and C. Ambroise. *Analyzing microarray gene expression data*. Wiley-Interscience, 2004.

[81] W. E. Moerner and D. P. Fromm. Methods of single-molecule fluorescence spectroscopy and microscopy. *Review of Scientific Instruments*, 74(8):3597–3619, 2003.

[82] J. Møller and R. P. Waagepetersen. Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, 34(4), 2007.

[83] M. Mörtelmaier, M. Brameshuber, M. Linimeier, G. Schütz, and H. Stockinger. Thinning out clusters while conserving stoichiometry of labeling. *Appl. Phys. Lett.*, 87(26):263903, 2005.

[84] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematic*, 42(5):577–685, 1989.

[85] D. B. Murphy. *Fundamentals of light microscopy and electronic imaging*. Willey-LISS, 2001.

[86] M. V. Newberry. Signal-to-noise considerations for sky-subtracted ccd data. *Astronomical Society of the Pacific*, 103:122–130, 1991.

[87] R.T. Ogden. *Essential Wavelets for Statistical Applications and Data Analysis*. Birkhäuser, 1997.

[88] J.-C. Olivo-Marin. Extraction of spots in biological images using multiscale products. *Pattern Recognition*, 35:1989–1996, 2002.

[89] S.W. Paddock. Principles and practices of laser scanning confocal microscopy. *Molecular Biotechnology*, 16(127-149), 2000.

[90] O. Rioul and P. Duhamel. Fast algorithms for discrete and continuous wavelet transforms. *IEEE Transcations on Information Theory*, 38(2), 1992.

[91] B. D. Ripley. *Spatial statistics*. John Wiley & Sons, 1981.

[92] D. M. Rocke and D. L. Woodruff. Computation of robust estimates of multivariate location and shape. *Statistica Neerlandica*, 47(1):27–42, 1993.

[93] D. M. Rocke and D. L. Woodruff. Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91(435):1047–1061, 1996.

[94] P.J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79:871–881, 1984.

[95] P.J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993.

[96] P.J. Rousseeuw and A.M. Leroy. *Robust regression and outlier detection*. John Wiley & Sons, 1987.

[97] P.J. Rousseeuw and K. van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

[98] P.J. Rousseeuw and B.C. van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639, 1990.

[99] P. Sarder, A. Nehorai, P. H. Davis, and S. L. Jr. Stanley. Estimating gene signals from noisy microarray images. *IEEE Transactions on Nanobioscience*, 7(2), 2008.

[100] M. Shensa. The discrete wavelet transform: Wedding the à trous and Mallat algorithms. *IEEE Trans. on Signal Processing*, 40(10):2464–2482, 1992.

[101] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapmann and Hall, 1986.

[102] Ch. A. Smith, A. Pollice, D. Emlet, and S.E. Shackney. A simple correction for cell autofluorescence for multiparameter cell-based analysis of human solid tumors. *Cytometry Part B (Clinical Cytometry)*, 70B:91–103, 2006.

[103] Gordon K. Smyth, Yee Hwa Yang, and Terry Speed. Statistical issues in cDNA microarray data analysis. In M.J. Brownstein and A.B. Khodursky, editors, *Functional Genomics: Methods and Protocols*. Humana Press, 2002.

[104] Donald L. Snyder, Abed M. Hammoud, and Richard L. White. Image recovery from data acquired with a charge-coupled-device camera. *J. Opt. Soc. Am. A*, 10:1014–1023, 1993.

[105] E.M. Southern. Detection of specific sequences among dna fragments separated by gel electrophoresis. *Journal of Molecular Biology*, 98:503–517, 1975.

[106] J. L. Starck, J. Fadili, and F. Murtagh. The undecimated wavelet decomposition and its reconstruction. *IEEE Trans. on Image Processing*, 16(2), 2007.

[107] J.-L. Starck, F. Murtagh, and A. Bijaoui. *Image and Data Analysis: The Multiscale Approach*. Cambridge University Press, 1998.

[108] R. B. Stoughton. Applications of DNA microarrays in biology. *Annual Review of Biochemistry*, 74(53-82), 2005.

[109] D. Stoyan, W.S. Kendall, and J. Mecke. *Stochastic geometry and its applications*. Wiley & Sons, 1987.

[110] B.T. Tan, C. Y. Park, L.E. Ailles, and I.L. Weissman. The cancer stem cell hypothesis: a work in progress. *Lab Invest*, 86(12):1203–1207, 2006.

[111] M. Unser. Ten good reasons for using spline wavelets. In *Proc. SPIE, Wavelets Applications in Signal and Image Processing V*, pages 422–431, 1997.

[112] S. Verboven and M. Hubert. Libra: a matlab library for robust analysis. *Chemometrics and Intelligent Laboratory Systems*, 75:127–136.

[113] J. Z. Wang, B.G.Lindsay, L. Cui, P.K. Wall, J. Marion, J. Zhang, and C. W. dePamphilis. Gene capture prediction and overlap estimation in EST sequencing from one or multiple libraries. *BMC Bioinformatics*, 6(300), 2005.

[114] C.L. Warren, N.C.S. Kratochvil, K.E. Hauschild, S. Foister, M.L. Brezinski, P.B. Dervan, G.N. Phillips, and A.Z. Ansari. Defining the sequence-recognition profile of DNA-binding molecules. *PNAS*, 103(4):867–872, 2006.

[115] P.H. Westfall and S.S. Young. *Resampling-based multiple testing: examples and methods for P-value adjustment*. John Wiley & Sons, 1993.

[116] E. Wit and J. McClure. *Statistics for Microarrays*. Wiley & Sons, 2004.

[117] D.E. Wolf. The optics of microscope image formation. *Methods in Cell Biology*, 81:11–42, 2007.

[118] J. Wu, L.T Smith, C. Plass, and T. H.-M. Huang. ChIP-chip comes of age for genome-wide functional analysis. *Cancer Research*, 66(14):6899–6902, 2006.

[119] Y. H. Yang, J.M. Buckley, S. Dudoit, and T. P Speed. Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics*, 11:108–136, 2002.

[120] Y.H. Yang, S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, and T.P. Speed. Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acid Research*, 30(4), 2002.

[121] I.T. Young. *Methods in Cell Biology*, chapter Image fidelity: Characterizing the imaging transfer function, pages 1–45. Academic Press, 1989.

[122] B. Zhang, N. Chenouard, J.-C. Olivo-Marin, and V. Meas-Yedid. Statistical colocalization in biological imaging with false discovery control. In *Proc. IEEE Int. Symp. on Biomedical Imaging: from nano to macro (ISBI)*, 2008.

Curriculum Vitae of Leila Mureşan

Affiliation and Address

Leila Mureşan

Department of Knowledge-based Mathematical Systems - FLLL
Johannes Kepler University, Linz
Altenbergerstr. 69, 4040 Linz
Phone: 0043 732/2468-9195
Email: leila.muresan@jku.at
URL: http://www.flll.jku.at

Personal Data

Date and place of birth: 31.05.1974, Cluj
Nationality: Romanian

Education

2004-:	PhD studies, Johannes Kepler University, Linz
1996-1997:	M.Sc. degree, Faculty of Mathematics and Computer Science, "Babeş-Bolyai" University, Cluj, Romania
1992-1996:	B.Sc. degree, Faculty of Mathematics and Computer Science, "Babeş-Bolyai" University, Cluj, Romania

Career History

since 2002:	Research assistant at Department of Knowledge-based Mathematical Systems/FLLL, Johannes Kepler University, Linz
1999-2002:	Department of Telecommunication and Telematics and Automobile Department of Road Vehicles, Technical University of Budapest, Hungary
1997-1998:	Software Engineer, Réseaux et Systèmes Informatiques SARL.
1996-1997:	Teaching assistant at Faculty of Mathematics and Computer Science, "Babeş-Bolyai" University, Cluj, Romania

Career Related Activities

GEN-AU I - II - II, 2002 - : *Ultrasensitive proteomics and genomics*: (Investigator). Topics: Single dye microscopy image analysis, ultra-sensitive microarray analysis.

Scholarship at the Complex Systems Summer School (1 month), Santa Fe Institute, Santa Fe, NM, June 2004

Ceepus (Central European Exchange Programme for University Studies) scholarships at:

- Czech Technical University, Prague - 2004
- J. Kepler University, Linz (Fuzzy Logic Laboratory) (4 months)
- Ostrava University of Science (1 month) - 2001
- Slovakian Technical University, Bratislava (2 months) - 1999, 2002

Chair of the IEEE EMBS Student Club, Johannes Kepler University, 2005-2007.

Selected Publications

1. Mureşan,L., Jacak,J., Klement, E.P., Hesse, J., Schütz, G. J.
   Microarray analysis at single molecule resolution
   IEEE Transcations on Nanobioscience, 2010, in print

2. Vlad, A., Yakunin, S., Kolmhofer, E., Kolotovska, V., Mureşan, L., Sonnleitner, A., Bäuerle,D., Pedarnig,J.D.
   Deposition, characterization and biological application of epitaxial Li:ZnO/Al:ZnO double-layers
   In *Thin Solid Films*, vol. 518, 2009, pp. 1350-1354

3. Hesse, J. , Jacak, J. , Regl, G., Eichberger, T. , Aberger, F. , Schlapak, R., Howorka , S., Mureşan, L. , Frischauf, A.-M. , Schütz , G.J.
   Single molecule fluorescence microscopy for ultra-sensitive RNA expression profiling
   In *Electronic Imaging*, SPIE / 6444, 2007

4. Mureşan, L., Heise, B., Klement, E.P.
   Tracking fluorescent spots in wide-field microscopy images
   In *Electronic Imaging*, San Jose, SPIE / IST 2006

5. Mureşan, L., Klement, E.P.
   Denoising Microscopy Image Sequences with Fine Structure Preservation
   In *Workshop of the Austrian Association for Pattern Recognition* , ÖAGM/AAPR 2006

6. Hesse, J., Jacak, J., Kasper, M., Regl, G., Eichberger, T., Winklmayr, M., Aberger, F., Sonnleitner, M., Schlapak, R., Howorka, S., Mureşan, L., Frischauf, A-M., Schütz, G.J.
   RNA expression profiling at the single molecule level
   In *Genome Res.* , vol. 16, pp. 1041-1045, 2006

7. Mureşan L., Heise B., Kybic J., Klement E.P.,
   Quantitative microarray Analysis of Microarray Images,
   In *International Conference on Image Processing*, ICIP 2005, Genoa, Italy, pp. 1274-1277

8. Mureşan, L., Tracking in microscopy images In *Uncertainty Modelling*, Bratislava, 2003

9. Koczy, L.T., Mureşan, L.,
Interpolation in hierarchical rule bases with normal conclusion
In *Lecture Notes in Computer Science*, Springer, 3-6, 2002, no. 2275, pp. 34-39.

10. Mureşan, L., Koczy, L.T.,
Similarity in hierarchical fuzzy rule-base systems
In *IEEE Int. Conf. on Fuzzy Systems* FUZZ-IEEE'02, Hawaii, U.S.A., vol. 1, 2002, pp. 746-750.

11. Mureşan, L., Vida, G.,
An approach to the detection of possible vehicle insurance frauds
In *The transport of the 21st century, International Scientific Conference*, Warsaw, 2001, pp. 237–242.

12. Tikk, D., Baranyi, P., Gedeon, T.D., Mureşan, L.,
Generalization of the rule interpolation method resulting always in acceptable conclusion
In *Tatra Mt. Math. Publ.* 21, 2001, pp. 73–91.

13. Koczy, L.T., Mureşan, L.,
Fuzzy systems with interpolation. An overview
In *Joint 9th IFSA World Congress and 20th NAFIPS International Conference*, 2001, vol. 5, 2001, pp. 2494-2498.

14. Tikk, D., Baranyi, P., Gedeon, T.D., Mureşan, L.,
Generalization and properties of the modified rule interpolation method
In *Proceedings of 6th International Conference on Soft Computing*, Iizuka, 2000, pp. 769-776.

15. Koczy, L.T., Hirota, K.,Mureşan, L.,
Interpolation in hierarchical fuzzy rule bases
International Journal of Fuzzy Systems, vol. 1, no. 2, December, 1999, pp. 77-84.

16. M.Sc. Thesis:Transactions in Object Oriented Database Management Systems
Supervisor: Dr. Pop Dragoş, *Babeş-Bolyai* University, Cluj, Romania, 1997

17. B.Sc. Thesis: Fuzzy Logic and Approximate Reasoning
Supervisor:Prof. Dr. Dan Dumitrescu, *Babeş-Bolyai* University, Cluj, Romania, 1996