Beiträge zur Biomedizinischen Bildgebung mit einem Seitenblick auf Molecular Imaging

G. Winkler
(Herausgeber)

Institut für Biomathematik und Biometrie

GSF- Bericht 03/04

GSF – Forschungszentrum
für Umwelt und Gesundheit
in der Helmholtz-Gemeinschaft

Herausgeber:

GSF - Forschungszentrum
für Umwelt und Gesundheit, GmbH

Ingolstädter Landstraße 1
D-85764 Neuherberg

Telefon 089/3187 - 0
Telefax 089/3187 - 3372
Internet http://www.gsf.de

Mitglied der Hermann von Helmholtz-Gemeinschaft
Deutscher Forschungszentren (HGF)

ISSN 0721 - 1694

Gedruckt auf umweltfreundlichem, chlorfrei gebleichtem Papier

Vorwort

In diesem Bändchen sind einige Beiträge zum Gebiet 'Bildgebende Verfahren in Biologie und Medizin' zusammengestellt. Sie stammen sämtlich aus dem 'Institut für Biomathemathematik und Biometrie', IBB, am 'Forschungszentrum für Umwelt und Gesundheit', GSF, in München/Neuherberg, und seinem engeren Umfeld. Mein Ziel war es, zu sichten, was in und um diesen Themenkreis herum an Wissen und sonstiger Kompetenz in meiner näheren Umgebung vorhanden ist.

Einige am IBB etablierte Gebiete wie Röntgen-Mammographie oder funktionelle Magnetresonanztomographie, wurden ausgeblendet. Der Grund ist die Fokussierung auf ein nicht exakt definierbares, neues Gebiet der Bildgebung, das unter dem Namen 'Molekular Imaging' kursiert und derzeit Furore macht. Eine gewisse Eingrenzung aus Sicht der Radiologie erlaubt folgendes Zitat aus der Zusammenfassung des Artikels R. WEISSLEDER & U. MAHMOOD (2001):

> The term molecular imaging can be broadly defined as the in vivo characterization and measurement of biological processes at the cellular and molecular level. *In contradistinction to 'classical' diagnostic imaging, it sets forth to probe the molecular anomalities that are the basis of disease rather than to image the end effects of these molecular alterations.* While the underlying biology represents a new arena for many radiologists, concomitant efforts such as development of novel agents, signal amplification strategies, and imaging technologies clearly dovetail with prior research efforts of our speciality. Radiologists will play a leading role in directing developments of this embryonic but burgeoning field.

Daraus erhellt unmittelbar einer der Gründe für unser Interesse an dieser Materie: Wo Radiologie und Bildgebung sind, sollten Mathematische Modellierung, Statistische Inferenz und Signal- bzw. Bildanalyse nicht weit sein. Am IBB wurde eine gewisse Kompetenz auf diesen Gebieten aufgebaut. Vor

dem Hintergrund eines exzellenten Umfeldes in den Gebieten der Genomik, Proteomik, Metabolomik, Cellomik und Molekularbiologie im allgemeinen an der GSF ist damit klar, daß sich das IBB dieser Aufgabe einfach nicht entziehen kann.

Unterstrichen wird dies durch eine Reihe weiterer Aktivitäten. Eine Vortragsreihe mit Kollegen aus der Radiologie vom Klinikum Rechts der Isar, unter Leitung von Prof. Schwaiger und Prof. Lasser, hat stattgefunden, erste wissenschaftliche Zusammenarbeiten haben begonnen. Am IBB läuft eine Seminarreihe zum Thema, aus der einige der hier präsentierten Beiträge hervorgegangen sind. Sondierende Gespräche und eine Sichtung der an der GSF vorhandenen Geräte findet zusammen mit Vertretern der Genetik gegenwärtig statt. Schließlich wird vom IBB im Juni 2004 ein international und hochkarätig besetzter Workshop zum Molecular Imaging ausgerichtet.

Wir hoffen, mit dieser Sammlung einen Beitrag zu dem skizzierten Projekt leisten zu können. Ich danke allen beteiligten Kolleginnen und Kollegen für ihre Zu- und Zusammenarbeit. Besonders danke ich Frau Annemarie Helmer für die Hilfe bei der Edition.

München/Neuherberg, 1. Mai 2004 Gerhard Winkler

Literatur

R. Weissleder & U. Mahmood (2001): Molecular Imaging. *Radiology*, 219: 316–333

Inhaltsverzeichnis

Teil I

Übersichtsartikel

Biomedical Imaging

Brigitte Forster*, **Rupert Lasser**†

Abstract

This report is intended as a brief introduction to the emerging scientific field of *Biomedical Imaging*. Focus is on emphasizing the breadth of the subject, on the one hand, and to indicate future fields of research, on the other hand. Hopefully, it serves as a guide to the identification of starting points for the research in 'Biomedical-and /or Molecular Imaging' at the GSF-National Research Center for Environment and Health. This may help to reveal research chances and opportunities in this context, in order to invigorate the future foundations of research groups, for example young researcher's groups. The report starts with a brief sketch of the history. Then a - necessarily incomplete - list of research topics is presented. It is organized in two parts: the first one addresses medical imaging, and the second one is concerned with biological point aspects of the matter.

*Biomedical Imaging Group EPFL LIB, Bat. BM 4.134
Swiss Federal Institute of Technology Lausanne
CH-1015 Lausanne, Switzerland
brigitte.forster@epfl.ch , www.brigitte-forster.de
†Institute of Biomathematics and Biometry
GSF - National Research Center for Environment and Health,
Postfach 1129, D-85758 Oberschleißheim, Germany,
lasser@gsf.de, http://ibb.gsf.de

Before we start...

...it is evident that twenty-first-century medicine will present a culmination of continuing evolutionary developments in multidimensional image visualization and analysis.

Richard A. Robb
Head of the Biomedical Imaging Resource
at Mayo Clinic/Foundation [49]

Contents

Chapter 1

Introduction

Biomedical imaging is an extremely active and quickly developing field. About five to ten years ago the main topics were image enhancement such as denoising and coding/compression of images. Today feature detection, 3-D visualization and real time processing with classical deterministic methods but more and more with soft computing are areas of modern research. In the following we will describe these areas in detail.

Today, biomedical imaging can be divided into two main parts: The medical one, where modeling and visualization of the body for diagnosis, surgery and training is emphasized, and the biological part, where the task lies in measurement, visualization, and simulation of chemical and physical reactions on the micro- and even the nanoscale. On small scale, genomics, proteomics, metabolomics, and cellomics, are the regions of interest; on a larger scale features like shape and evolution of organic matter are of interest, as in order to study the relation between the genome, and its function. At the GSF - for example - focus is on the mouse model.

We will start with medical imaging, the various modalities and its positive and critical sides. Then we mention the problems and tasks of classical image processing and give information on the new advances in soft computing.

We continue with further modalities for biological imaging and the nanoscale, and proceed as for medical imaging with the questions of modern research in this area.

Chapter 2

On the history of biomedical imaging

The beginning of biomedical imaging lies in 1890s, when Wilhelm Roentgen accidentally discovered X-rays. He observed that if he aimed a cathode-ray tube's X-rays through a persons hand at a chemically coated screen, he could see the bones in the hand clearly on the screen.

Over the next few decades, X-rays grew into a widely used diagnostic tool. The great strength of X-rays are their high resolution and their ease of use. On the other hand, X-rays do not distinguish well between tissues of similar densities, capture only a slice of the body's three-dimensional structure and expose the patient to potentially harmful radiation.

For more than half a century, the science of medical imaging grew steadily but slowly, as incremental improvements were made in the X-ray technique. In the early 1970s that growth kicked into high gear with the appearance of a new imaging technique: computerized tomography (CT). By taking a series of X-rays from various angles and then combining them with a computer, CT made it possible to build up a three-dimensional image of any part of the body.

This development lead to a flood of new medical imaging tools working according to various physical principles. But they have one thing in common: The new machines all depend on computers to construct the images from a mass of data that is collected electronically instead of on film.

In the early 1980s, CT was joined by a completely different way of taking pictures of the body's interior: magnetic resonance imaging, or MRI. Instead of passing X-rays through the body, MRI relies on a strong magnetic field and a radio signal that together trigger atoms in the body to send out signals of their own. By collecting and analyzing these signals, it is possible to compute

a three-dimensional image which, like a CT image, is normally displayed in two-dimensional slices.

Although MRI avoids harmful radiation by using a magnetic field and radio signals instead of X-rays, it has the disadvantage of less resolution than CT. What made MRI nonetheless a valuable complement to CT is that it provides very different information about the body, since it responds to the prevalence of particular types of atoms, such as hydrogen and thus offers information about tissue.

In the past several years, two other imaging techniques have joined CT and MRI. In positron emission tomography, or PET, a patient ingests or is injected with a slightly radioactive substance that emits positrons, which can be monitored as the substance moves through the body. Where CT and MRI are mainly valuable in viewing the body's internal structures, PET is also useful in tracking its metabolism.

Closely related to PET is single-photon emission computed tomography, or SPECT. The major difference between the two is that instead of a positron-emitting substance, SPECT uses a radioactive tracer that emits high-energy photons.

Another techniques is ultrasound. Ultrasound makes an image by bouncing sound waves that are then converted into a picture. It is known by many for its pictures of babies in the womb, but is also widely used to measure blood flow.

This array of tools is changing the way medicine is practiced, giving doctors much more information than ever before diagnose disease or injury. Not only doctors but medical researchers are putting the new tools to use. After all, although X-rays are over 100 years old and CT scans over 25, most of the tools are still relatively new, with much of their potential still to be tapped.

Many of the advances will come in data processing. Because most of the images are constructed by computer from a collection of electronic data, image quality depends on the power of the computer and the caliber of its software and there is plenty of room for improvement of both. PET and SPECT images could be sharper, for example. Both rely on mathematical calculations to determine the location from which a positron or photon was emitted. Imprecise calculations show up as fuzz in the picture. For many applications scientists know how to get better images from the data than are produced by current devices, but the necessary calculations take so long that they are impractical.

Many of today's medical images are processed digitally. This opens up many new possibilities. Data can be manipulated to enhance image quality. The

information for images can also be transmitted to other locations over telephone lines, and various images even those from different types of machines can be compared and fused into composites. All of this requires specialized methods, computer software and hardware, which researchers are developing and refining.

To improve data processing, many researchers are looking to enhance the ways data are collected, for example to avoid contrast media in the patients body or to reduce X-ray dose.

Recent exiting advances were reached by the ability to use MRI not just to image static internal structures but to monitor body activities, such as blood flow and brain activity. So-called functional MRI can be performed with existing MRI machines, but will be much more effective with the next generation devices.

Another approach is to combine imaging techniques. In many cases the imaging techniques by itself are not completely adequate, but they provide complementary information and the whole may be more than the sum of its parts.

There are many other important and promising research topics in biomedical imaging. A detailed description can be found in the following chapter.

(Source: Adapted from [59].)

Chapter 3

Medical Imaging

As recently as 25 years ago, a physician or a surgeon who suspected the presence of a tumor in a patient had few options. Order x-ray studies to define and localize the tumor as accurately as the pictures would permit. Schedule the patient for surgery and examine the tumor directly, excise a portion of the unhealthy tissue for biopsy, remove the tumor if possible, and explore surrounding tissues to determine whether the cancer had spread.

Over the last quarter century, refinements in imaging technology have substantially broadened the range of medical options. Current imaging tests now provide much clearer and more detailed pictures of organs and tissues than were possible previously. imaging already has had a lifesaving effect in detecting some early cancers. X-ray mammography, for example, has saved the lives of many women by revealing the presence of very small cancers before they could be detected by physical examination. Computed tomography (CT) and ultrasound permit physicians to guide long, thin needles deep within the body to biopsy organs, often eliminating the need for an open surgical procedure. CT can reveal whether a tumor has invaded vital tissue, grown around blood vessels, or spread to distant organs; important information that can help guide treatment choices.

The potential of imaging to improve cancer treatment extends well beyond using imaging information to help select effective treatments or preventives. For example, combining precise imaging techniques with radiation sources and high performance computing is significantly improving our ability to shape radiation treatment to the tumor's three-dimensional contours. In principle, imaging techniques can be interfaced with other tumor-killing approaches — toxic chemicals, gene therapy, heat, and cold — to more precisely guide tissue destruction at the tumor site.

Being able to distinguish between cancerous and normal tissue and deliver

treatments only to diseased tissue in a minimally invasive way will potentially minimize surgical trauma, short recovery time, and reduce health costs.

Although it sees clear that better imaging tools will improve patient care, we need better ways of measuring that improvement. Improved methodologies to assess the ultimate value of diagnostic tests are needed.

The goals for further research work are

- Develop and validate imaging technology and agents.
- Develop imaging techniques that identify the biological properties of precancerous or cancerous cells.
- Develop minimally invasive imaging technologies.
- Foster interaction and collaboration among imaging scientists and basic biologists, chemists and physicists to help to advance imaging research.

(Source: Adapted from [43]. Many more resources to this section can be found in [17].)

3.1 Modalities

X-ray

For imaging purposes, X-rays are generated by the interaction of accelerated electrons with a target material — usually tungsten. A heated cathode tube is the source of the electrons which are accelerated by applying a voltage of about 100 keV between the cathode and the tungsten anode. The X-ray tube is placed one or more meters from the patient. As the emitted X-rays pass through the patient, they are deflected by the bodily tissues and recorded by a detector on the opposite side. The most common detector is silver halide film, which is darkened by the interaction of the transmitted photons. Thus, darker areas indicate high transmission intensities while brighter areas show places of low exposure.

X-rays are deflected and absorbed to different degrees by the various tissues and bones in the patients body. The amount of absorption depends on the tissue composition. For example, dense bone matter will absorb many more X-rays than soft tissues, such as muscle, fat and blood. The amount of deflection depends on the density of electrons in the tissues. Tissues with high

electron densities cause more X-ray scattering than those of lower density. Thus, since less photons reach the X-ray film after encountering bone or metal rather than tissue, the X-ray will look brighter for bone or metal.

A problem arises when using X-rays to distinguish between blood vessels and the surrounding tissues because they are very close in composition. They do not contrast enough to distinguish on the detector film. To increase image contrast, a dense fluid with elements of high atomic number ("heavy atoms" such as iodine or barium) can be injected or swallowed during the X-ray exposures. The movement through the body vasculature of a "contrast agent" such as an iodinated compound can be visualized by acquiring a sequence of X-rays. The iodine or barium agent absorbs photons more than blood and tissue because the density is higher and the elements iodine and barium have a high atomic number which gives rise to more photoelectric absorption.

As recording media the target material in a cassette was used for long. Over the years, innovation has brought better imaging but fundamental limitations imposed by cassettes have remained. Alignment issues, dust artifacts, light leaks, and quality control delays comprise image quality. Direct Digital radiography captures the radiographic image directly into digital format. This process is cassette-free, saves money and time spent for handling cassettes and gives high resolution and good imaging quality.

(Source: [8], [33].)

Computed Tomography CT

Computed tomography (CT) became generally available in the mid 1970's and is considered one of the major technological advances of medical science. X-ray CT gives anatomical information on the positions of air, soft tissues, and bone. Three dimensional imaging is achieved by rotating an X-ray emitter around the patient, and measuring the intensity of transmitted rays from different angles.

Unlike conventional X-ray transmission, X-ray CT does not use film to detect the transmitted gamma rays. Instead, the photons are collected by an electronic device which converts the X-ray photons into an electric current. Either a solid state device or a crystal scintillator is used to make visible light from the X-rays. Then the light photons are converted to electrons in a photoelectron multiplier tube (PMT). The result is an electric current whose magnitude is proportional to the initial energy of the X-ray photon.

X-ray CT techniques offer excellent spatial resolution, but have very limited soft tissue contrast. Contrast in CT is determined primarily by the electron

density of the materials imaged. Bone is the only tissue in the body with dramatically different electron density; soft tissue generally have very similar electron densities and atomic number.

(Source: [8], [29].)

Magnetic Resonance Imaging MRI

Magnetic Resonance Imaging (MRI) has evolved into one of the most powerful non-invasive techniques in diagnostic clinical medicine and biomedical research. The technique is an application of nuclear magnetic resonance (NMR), a well known analytical method of chemistry, physics and molecular structural biology. MRI is primarily used as a technique for producing anatomical images, but as described below, MRI also gives information on the physical-chemical state of tissues, flow diffusion and motion information. Magnetic Resonance Spectroscopy (MRS) gives chemical/composition information.

Most elements have at least one reasonably abundant isotope whose nucleus is magnetic. In biological materials, the magnetic nuclei of ^{1}H, ^{13}C, ^{23}Na, ^{31}P, and ^{39}K are all abundant. The hydrogen nucleus (a single proton) is abundant in the body due to the high water content of non-bony tissues. When the body is immersed in a static magnetic field, slightly more protons become aligned with the magnetic field than against the static field. At 0.25 Tesla and 25°C the difference between these aligned populations of about one proton in a million produces a net magnetization. A rapidly alternating magnetic field at an appropriate resonant frequency in the radio frequency range, applied by a coil near the subject or specimen in the static magnetic field, changes the orientation of the nuclear spins relative to the direction of the static magnetic field.

These changes are accompanied by the absorption of energy (from the alternating magnetic field) by nuclei which undergo the transition from a lower energy state to a higher one. When the alternating field is turned off, the nuclei return to the equilibrium state, emitting energy at the same frequency as was previously absorbed. The nuclei of different elements, and even of different isotopes of the same element, have very different resonance frequencies. For a field of 0.1 Tesla, the resonance frequency of protons is 4.2 MHz and that of phosphorus is 1.7 MHz. Thus, the magnetic nuclei in the body, when placed in a static magnetic field, can be thought of as tuned receivers and transmitters of radio frequency energy.

The principal components of the MRI machine are the magnet, radio fre-

quency coils and the gradient coils. Magnet types in current use are of the superconducting, resistive and permanent magnet designs ranging in strength from 0.08 to 4 Tesla. The majority of MR systems use superconducting magnets which provide fields of high strength and stability. Most currently produced magnets are based on niobium-titanium (NbTi) alloys, which are remarkably reliable, but require a liquid helium cryogenic system to keep the conductors at approximately 4.2 Kelvin (−268.8°C). The radio frequency coils used to excite the nuclei usually are quadrature coils which surround the head or body, but small (e.g., 6–10 cm) flat coils placed on the surface of the head or body are also used. Besides being the essential element for spatial encoding, the gradient-coil sub-system of the MRI scanner is responsible for the encoding of specialized contrast such as flow information, diffusion information, and modulation of magnetization for spatial tagging.

MRI contrast is determined by the relaxation parameters of the water in tissue; the intrinsic contrast between soft tissues in MRI is many times what it is for CT. For example, the white matter of the brain has 12 % lower CT number than the gray matter, while white matter has a 140 % higher signal than gray matter in an MRI examination. Another important advantage of MRI is the ability to obtain views at arbitrary positions and orientations. On the other hand, MRI is sometimes limited by its spatial resolution properties and long imaging times.

(Source: [8], [29].)

Spin-Echo Magnetic Resonance Imaging

In spin-echo MRI, gradients and Fourier analysis are used to perform three-dimensional imaging. Other techniques of MRI, such as gradient-echo, are slight variations of spin-echo imaging. The component of the imaging system which allows the spatial localization of the protons is a set of magnetic field gradients, set up by magnetic coils which are turned on and off at appropriate times.

When hydrogen nuclei relax, the frequency that they transmit is positively correlated with the strength of the magnetic field surrounding them. A magnetic field gradient along the z-axis, called the "slice select gradient", is set up when the radio frequency pulse is applied, and is shut off when the radio frequency pulse is turned off. This gradient causes the hydrogen nuclei at the high end of the gradient (where the magnetic field is strong) to precess at a high frequency (e.g., 65 MHz), and those at the low end (weak field) to precess at a lower frequency (e.g., 63 MHz). When the radio frequency pulse,

of a single frequency, is applied, only those nuclei which precess at that frequency will be tilted, to later relax and emit a radio transmission (i.e., the nuclei "resonate" to that frequency). For example, if the magnetic gradient caused hydrogen nuclei to precess at rates from 63 MHz at the low end of the gradient to 65 MHz at the high end, and the gradient were set up such that the high end was located at the patient's head and the bottom part at the patient's feet, then a 63 MHz radio frequency pulse would excite the hydrogen nuclei in a slice near the feet, and a 65 MHz pulse would excite them in a slice near the head. Thus a single "slice" along the z-axis is selected; only the protons in this slice are excited to a higher energy level, to later relax to a lower energy level and emit a radio transmission.

The second dimension of the image is extracted with the help of a phase encoding gradient. Immediately after the radio frequency pulse ceases, all of the nuclei in the activated slice are "in phase", that is, their magnetic vectors all point in the same direction. Left to their own devices, these vectors would relax. In MRI, however, the phase encoding gradient (in the y-dimension) is briefly applied, in order to cause the magnetic vectors of nuclei along different portions of the gradient to point in different directions.

After the radio frequency pulse, slice select gradient, and phase encoding gradient have been turned off, the MRI instrument sets up a third magnetic field gradient, along the x-axis, called the "frequency gradient" or "read-out gradient". This gradient causes the relaxing protons to be differentially re-excited, so that the nuclei near the low end of the gradient begin to precess at a faster rate, and those at the high end pick up even more speed. When these nuclei relax again, the fastest ones (those which were at the high end of the gradient) will emit the highest frequency of radio waves. The frequency gradient is applied only when the signal is measured.

The second and third dimensions of the image are extracted by means of Fourier analysis. The entire procedure must be repeated multiple times in order to form an image with a good signal-to-noise ratio.

Finally, in spin-echo imaging, there is the problem that the inhomogeneity of the main magnetic field induces variations in the rate of precession of nuclei. To fix this problem, a 180-degree radio frequency pulse is inserted into the cycle, at a time point halfway between the 90-degree pulse and the measurement of the radio transmission signal given off by the relaxing nuclei.

(Source: [60].)

Positron Emission Tomography PET

The history of positron emission tomography (PET) can be traced to the early 1950's, when workers in Boston first realized the medical imaging possibilities of a particular class of radioactive isotopes. Whereas most radioactive isotopes decay by release of a gamma ray and electrons, some decay by the release of a positron. A positron can be thought of as a positive electron. Widespread interest and an acceleration in PET technology was stimulated by development of reconstruction algorithms associated with X-ray CT and improvements in nuclear detector technologies. By the mid-1980s, PET had become a tool for medical diagnosis, for dynamic studies of human metabolism and for studies of brain activation.

PET has a million fold sensitivity advantage over other techniques used to study regional metabolism and neuroreceptor activity in the brain and other body tissues. In contrast, magnetic resonance has exquisite resolution for anatomic studies and for flow or angiographic studies. In addition, magnetic resonance spectroscopy has the unique attribute of evaluating chemical composition of tissue but in the millimolar range rather than the nanomolar range. Since the nanomolar range is the concentration range of most receptor proteins in the body, positron emission tomography is ideal for this type of imaging.

PET imaging begins with the injection of a metabolically active tracer-a biological molecule that carries with it a positron-emitting isotope (for example, ^{11}C, ^{13}N, ^{15}O, or ^{18}F). Within minutes, the isotope accumulates in an area of the body for which the molecule has an affinity. As an example, glucose labeled with ^{11}C (half-life, 20 min), or a glucose analog labeled with ^{18}F (half-life, 1.8 h), accumulates in the brain, where glucose is used as the primary source of energy. The radioactive nuclei then decay by positron emission. The emitted positron collides with a free electron usually within less than 1 mm from the point of emission. The interaction of the two subatomic particles results in a conversion of matter to energy in the form of two gamma rays. These high-energy gamma rays emerge from the collision point in opposite directions, and are detected by an array of detectors which surround the patient.

When the two photons are recorded simultaneously by a pair of detectors, the collision that gave rise to them must have occurred somewhere along the line connecting the detectors. Of course, if one of the photons is scattered then the line of coincidence will be incorrect. After 500,000 or more annihilation events are detected, the distribution of the positron emitting tracer is calculated by tomographic reconstruction procedures. PET then reconstructs a two-

dimensional image. Three dimensional reconstructions can also be done using 2-D projections from multiple angles.

The major clinical applications of PET have been in cancer detection of the brain, breast, heart, lung and colorectal tumors. Another application is the evaluation of coronary artery disease by imaging the metabolism of heart muscle. Contemporary instruments can scan 15 cm segments of the body. Thus, with the combination of eight full scans, a whole body image of the distribution of ^{18}F-deoxyglucose (or other common radioisotopes) can be acquired in 40 minutes; as is commonly done for evaluation of breast cancer.

Applications to the study of epilepsy, brain tumors, stroke and Alzheimer's disease have occupied the interest of neurologists for nearly 20 years. A major attribute of PET is its ability to show activity of neuroreceptors such as the dopamine, serotonin and noradrenergic receptor systems. The concentrations of neurochemical sites in the body are generally too low (e.g. μm) for MRI or MRS studies.

(Source: [8]. A review on the history of PET can be found in [45].)

Functional Magnetic Resonance Imaging fMRI

The term "functional MR" can include the technique of co-registering PET and MRI scans, but it is usually used to denote techniques involving fast MRI scans, which can allow imaging of a complete brain slice in 20 ms. The first fMRI of the brain (a perfusion MRI) was done in 1991 by Belliveau and co-workers, who injected a chemical that increases MRI contrast into a patient and imaged the brain using echo-planar techniques, to show that the perception of visual stimuli increases blood volume in primary visual cortex. The same group later used gradient echo and spin-echo inversion recovery fMRI to examine blood oxygenation levels and blood flow rates, respectively, in brain.

FMRI is a technique for determining which parts of the brain are activated by different types of physical sensation or activity, such as sight, sound or the movement of a subject's fingers. This "brain mapping" is achieved by setting up an advanced MRI scanner in a special way so that the increased blood flow to the activated areas of the brain shows up on functional MRI scans. The whole fMRI process will now be briefly described.

The subject in a typical experiment will lie in the magnet and a particular form of stimulation will be set up. For example, the subject may wear special glasses so that pictures can be shown during the experiment. Then, MRI images of the subject's brain are taken. Firstly, a high resolution single scan

is taken. This is used later as a background for highlighting the brain areas which were activated by the stimulus. Next, a series of low resolution scans are taken over time, for example, 150 scans, one every five seconds. For some of these scans, the stimulus (in this case the moving picture) will be presented, and for some of the scans, the stimulus will be absent. The low resolution brain images in the two cases can be compared, to see which parts of the brain were activated by the stimulus.

After the experiment has finished, the set of images is analyzed. Firstly, the raw input images from the MRI scanner require mathematical transformation (Fourier transformation, a kind of spatial "inversion") to reconstruct the images into "real space", so that the images look like brains. The rest of the analysis is done using a series of tools which correct for distortions in the images, remove the effect of the subject moving their head during the experiment, and compare the low resolution images taken when the stimulus was off with those taken when it was on. The final statistical image shows up bright in those parts of the brain which were activated by this experiment. These activated areas are then shown as colored blobs on top of the original high resolution scan, for interpretation of the experiment. This combined activation image can be rendered in 3-D, and the rendering can be calculated from any angle.

(Sources: [57],[60].)

Single-Photon Computed Tomography SPECT

Single-photon emission computed tomography (SPECT), like PET, acquires information on the concentration of radionuclides introduced to the patient's body. SPECT dates from the early 1960's, when the idea of emission traverse section tomography was introduced by D. E. Kuhl and R. Q. Edwards prior to either PET, X-ray CT, or MRI.

As in X-ray CT, SPECT imaging involves the rotation of a photon detector array around the body to acquire data from multiple angles. Using this technique, we seek the position and concentration of radionuclide distribution. Because the emission sources (injected radionuclides) are inside the body cavity, this task is far more difficult than for X-ray CT, where the source position and strength (outside the body) are known at all times. That is, in x-ray CT the attenuation is measured, not the transmission source. To compensate for the attenuation experienced by emission photons from injected tracers in the body, contemporary SPECT machines use mathematical reconstruction algorithms to increase resolution.

SPECT imaging is inferior to PET because of attainable resolution and sensitivity. Different radionuclides are used for SPECT imaging that emit a single photon (usually about 140 keV), rather than positron emission (resulting in two high energy 511 keV photons) as in PET. Because only a single photon is emitted from the radionuclides used for SPECT, a special lens known as a collimator is used to acquire the image data from multiple views around the body. The use of a collimator results in a tremendous decrease in detection efficiency as compared to PET. For positron emission tomographs, collimation is achieved naturally by the fact that a pair of detected photons (gamma rays) can be traced back to their origin since they travel along the same line after being produced. In PET, there might be as many as 500 detectors that could "see" a PET isotope at any one time where as in SPECT, there may be only one or three collimators. Higher resolutions require better sensitivity and the resolution of SPECT is many times less than that of PET. The resulting useful resolution (about 7 mm) for SPECT is inferior to PET resolutions by a factor of 3 or 4.

Although SPECT imaging resolution is not that of PET, the availability of new SPECT radiopharmaceuticals, particularly for the brain and head, and the practical and economic aspects of SPECT instrumentation make this mode of emission tomography attractive for clinical studies of the brain. The cost of SPECT imaging is around $ 700 while that of PET is $ 2000.

(Source: [8].)

Ultrasound

Ultrasound, as currently practiced in medicine, is a real-time tomographic imaging modality. Not only does it produce real-time tomograms of the position of reflecting surfaces (internal organs and structures), but it can be used to produce real-time images of tissue and blood motion.

Ultrasound denotes the use of acoustical (sound) waves at frequencies greater than 20 kHz. Generally, medical ultrasound is performed at frequencies in the range of 1 MHz. The technique is used to determine the location of surfaces within tissues by measuring the time interval between the production of an ultrasonic pulse and the detection of its echo resulting from the pulse reflected from those surfaces. By measuring the time interval between the transmitted and detected pulse, one can calculate he distance between the transmitter and the object. The ultrasound pulses are both produced and detected by a piezoelectric crystal. The crystal has the property of changing its physical dimensions in response to an electric field, and can produce an electric field

if its physical shape is changed mechanically. Thus, ultrasonic compression waves (vibrations) are produced by applying an oscillating potential across the crystal. The reflected ultrasound imposes a distortion on the crystal, which in turn produces an oscillating voltage in the crystal. The same crystal is used for both transmission and reception.

If a structure is stationary, the frequency of the reflected wave will be identical to that of the impinging wave. A moving structure will cause a back-scattered signal frequency shifted higher or lower depending on the structure's velocity toward or away from the sound generator (called a transducer).

For example, when an impinging sound pulse passes through a blood vessel, scattering and reflection occurs from the moving red cells. In this process, small amounts of sound energy are absorbed by each red cell, then re-radiated in all directions. If the cell is moving with respect to the source, the back scattered energy returning to the source will be shifted in frequency, with the magnitude and direction proportional to the velocity of the respective blood cell. Thus, if we use ultrasound to image the cross-sectional area of the blood vessel, the volume of blood flow can be calculated from the area of the vessel and the average velocities of the blood cells.

The major use of Doppler ultrasound is the study of the heart and human cartoid artery disease wherein imaging and frequency shift are combined to produce images of artery and ventricle lumens. The frequency shift data is used to color the image, showing direction of flow (e.g. cartoid arteries in red and veins in blue). Obstructions to blood flow are readily evaluated by this method using hand held scanning devices.

In addition to imaging heart valves and blood vessels, ultrasound is the most convenient and inexpensive method for medical evaluations such as fetal gender and gallbladder stones. Ultrasound imaging is also being used for monitoring therapy methods such as hyperthermia, cryosurgery, drug injections, and as a guide during biopsies and catheter placements.

Ultrasound imaging is inexpensive and is done in real time. However, ultrasound can not be used for structures that are deep in the body or where there is air or bone around the structure of interest. Therefore, with few exceptions, ultrasound is mainly used in the abdomen and heart.

(Source: [8], [29].)

Electrical Impedance Tomography EIT

Electrical impedance tomography (EIT) — also called applied potential tomography — is a novel imaging technique with applications in medicine and

process control. Compared with techniques like computerized X-ray tomography and positron emission tomography, EIT is about a thousand times cheaper, a thousand times smaller and requires no ionizing radiation. Further, EIT can in principle produce thousands of images per second. Its major limitations are its low spatial resolution, and — in the medical field — large variability of images between subjects. Recordings are typically made by applying current to the body or system under test using a set of electrodes, and measuring the voltage developed between other electrodes. To obtain reasonable images, at least one hundred, and preferably several thousand, such measurements must be made.

In the medical field, the most studied applications for EIT are measurement of gastric emptying and lung function. In the industrial field typical applications are imaging the distribution of oil and water in a pipeline and imaging the flow of substances in a mixing vessel. In some ways industrial applications are more favorable for EIT because it is usually possible to use a rigid, fixed array of electrodes. The fixing of electrodes on the human body is one of the residual problems facing medical EIT.

EIT produces images of the distribution of impedivity (or, more commonly, resistivity), or its variation with time or frequency, within the tissue. There is a large resistivity contrast (up to about 200:1) between a wide range of tissue types in the body. It ought therefore to be possible to use resistivity to form anatomical images. Furthermore, there is often a significant contrast between normal and pathological tissue. For example, at 1 kHz, cerebral gliomas have a resistivity about half that of normal tissue. To measure resistivity or impedivity, a current must flow in the tissue and the resulting voltages be measured. This applied current will be referred to as the excitation current. In practice almost all EIT systems use constant current sources, and measure voltage differences between adjacent pairs of electrodes. To obtain an image with good spatial resolution, a number of such measurements is required. This can be achieved by applying different current distributions to the body, and repeating the voltage measurements. From the set of measurements, an image reconstruction technique generates the tomographic image. Mathematically, the known quantities are the voltages and currents at certain points on the body; the unknown is the impedivity or resistivity within the body. At low frequencies, these quantities are related by Laplace's equation. In practice, the solution of Laplace's equation is very sensitive to noise in the measurements, and normalization techniques must be used. Most in-vivo images have been produced using linearized, approximating techniques. These attempt to find a solution for a small change in resistivity from a known starting value. Until recently, the change in resistivity was measured over time, and EIT

images were inherently of physiological function. It is now possible to produce anatomical images using the same reconstruction technique, by imaging changes with frequency.

EIT is expected to have relatively poor resolution compared to MRI, X-ray CT, PET, SPECT, and ultrasound imaging. Resolution is largely controlled by the number of electrodes that can be reasonably attached simultaneously to a patient. Schemes used to date have normally used EEG-style electrodes, relatively few in number and large in size. It is not yet known whether special belts or head bands having large numbers of small electrodes designed especially for EIT applications might facilitate a major improvement.

At the present time, EIT is the only method known that images the electrical conductivity, although MRI and electromagnetic methods also have some potential to measure conductivity. One approach to EIT is called the applied potential tomography (APT) system which uses a ring of 16 electrodes. Current is passed through two adjacent electrodes, then voltage on all the other electrodes is measured. Next, current is generated between another pair of adjacent electrodes, and so on, until all adjacent pairs have served as "driver pairs". This approach has the advantage of simplicity of design, but its resolution is intrinsically limited, and is sensitive to small errors in electrode placement.

Electrical Impedance Tomography has some very attractive features for clinical applications such as monitoring lung fluid. The technology for doing electrical impedance imaging is safe and inexpensive. At the low current levels needed for this imaging technique, the method is not known to cause any long-term harm to the patient, and therefore could be used to do continuous monitoring of bedridden patients. Technology for acquiring data, and algorithms for inverting those data to produce images of conductivity/resistivity, have been developed to the point that real-time imaging could become routine today. However, the physiological interpretation of the image changes are not yet well defined, nor is the mathematical problem of multiple non-straight line current paths.

(Source: Adapted from [7] and [8].)

Electrical Source Imaging ESI

Electrical source imaging (ESI) is an emerging technique for reconstructing electrical activity in the brain or heart from electric potentials measured on the scalp or torso. Standard electroencephalographic (EEG), electrocardiographic (ECG) and vectorcardiographic (VCG) techniques are limited in

their ability to provide information on regional electrical activity or localize bioelectrical events within the brain and heart. Noninvasive ESI of the brain requires simultaneous electric potential recordings from 20 or more electrodes for the brain and 100 to 250 torso electrode sites to map the body surface potential from the heart.

Interest in current source localization in the brain from detection of surface potentials (using EEG), has largely been supplanted by interest in magnetic field measurement approaches (MSI). For the heart however, clinical needs have encouraged research to produce practical solutions to ESI. Abnormal electrical activity of the heart (cardiac arrhythmias) cause at least 400 000 sudden deaths each year in the United States alone, more than two-thirds of the nation's heart-related fatalities. Body surface potential maps reflective of heart electrical activity (obtained using ECG) allow physicians to localize the abnormal electrical pathways of arrhythmias. The syndrome can then be treated by surgical resection or catheter ablation.

(Source: [8].)

Magnetic Source Imaging MSI

Ion currents arising in the neurons of the heart and the brain produce magnetic fields outside the body. These fields can be measured by arrays of SQUID (Superconducting QUantum Interference Device), detectors that are placed on or near the head or chest. The recording of magnetic fields of the head is known as magnetoencephalography (MEG) while that of the heart is called magnetocardiography (MCG). Magnetic Source Imaging (MSI) is the general term for the reconstruction of current sources in the heart or brain from the measurements of external magnetic fields.

MSI seeks to determine the location, orientation, and magnitude of electric current flow regions within the body. A major strength of MSI is that it can resolve events separated by milliseconds, whereas other methods such as functional magnetic resonance imaging (fMRI), magnetic resonance spectroscopy (MRS), PET, and SPECT have time resolutions of seconds to many minutes, depending on the information sought. The weakness of magnetic source imaging is that any magnetic field distribution on the surface of the body can be explained by an infinite number of current distributions inside the head or heart. Thus, a successful source analysis is dependent on the availability of additional information that can constrain the problem to be solved. A current line of research is to improve the spatial resolution of MSI by using prior knowledge of the anatomy, obtainable by MRI or X-ray CT imaging.

An advantage of MSI over ESI (electrical source imaging) is that body tissues are magnetically transparent. ESI methods such as the EEG and EKG give surface electric potentials; the combination of electric potentials at the source (e.g. heart) and the complex electrical properties between the source and the electrodes on the body surface. Because different tissues in the body have relatively close conductivities, the electric potential measurements in ESI can be seriously distorted. However, body tissues have a neglible effect on magnetic forces, and thus are negligible in MSI measurements.

Biomagnetism offers a tool to study processes where electric function is important. Promising results have been obtained in the fields of cardiology and epilepsy. Sites of origin for heart arrhythmias (irregular beats) can be identified by electrical activity in abnormal anatomical locations. Noninvasive localization of the problem by MSI allows treatment by guiding a catheter directly to the correct site. MSI can also be used in the surgical treatment of intractable epilepsy to locate an epileptic focus, or can detect functional areas of the brain that must be conserved during surgery. A potential use in neuroscience is the spatial and temporal (time) study of functional processing areas in the brain in response to auditory, visual, and physical (somatosensory) stimuli.

(Source: [8].)

Laser Optical Imaging

Scientists have speculated about shining light through the human body for 100 years or more. For example, quantitative light absorption at specific wavelengths has been used since the 1930's for determining the oxygen content of blood, and now this work has been extended to imaging. In the late 1980's, a concentrated effort was directed toward imaging the transmission of light through tissue. Light in the near infrared range (wavelengths from 700 to 1200 nm) penetrates tissue and interacts with it in complicated ways, with the predominant effects being absorption and scattering. Laser optical tomography involves reconstruction of the amount of transmitted laser light through an object along multiple paths.

Laser optical tomography has a spatial resolution of about 10 mm, thus it cannot give pictures of the resolution quality of X-ray CT scans. However, the method does have a number of practical applications even at low resolution. These include the measurement of tissue oxygenation for the study of muscular dystrophy, tissue perfusion in the extremities for diabetic disease, the detection of brain hemorrhaging, monitoring stroke patients, the

study of brain activity during specific tasks and possibly the study of glucose concentration changes. The clinical potentials of determining the oxygenation level in the brain of young children has been demonstrated. Substances which play a crucial role in the bodies metabolic (energy making) processes, such as NAD/NADH (nicotinamide adenosine diphosphate), exhibit fluorescent properties which allow detection after being excited by light. Their assessment by indirect measurements has important potentials for medical applications.

(Source: [8].)

Thermography

Thermography is a technique for sensing and recording on film hot and cold areas of the body by means of an infrared detector that reacts to blood flow. Disease states that manifest increased or decreased blood flow present thermographic patterns that may possibly be distinguished from normal areas.

Thermography measures the slightest variations in temperature of soft tissue in the body using infrared heat sensors. Thermography for use in cancer screening is a recent invention based on the concept that cancer gives off more heat than normal tissue. It was originally a much-heralded technique, since it does not involve radiation, or putting anything else in the body. Unfortunately, this technique has not proved accurate, there are too many false positives and false negatives. Not all cancers give off heat, and of those that do, some are too deep, or located under wedges of fat, and the heat does not register on the device. Though thermography does not detect cancer, some physicians believe that it can define the aggressiveness of a cancer known to exist. The more aggressive a cancer is, the more heat it gives off. This has not been substantiated and is currently under research.

There may be uses for thermography in other fields, including dentistry and the diagnosis of headaches. In dentistry, facial skin temperature can be measured in a clinical setting, without direct skin contact, by monitoring the emitted infrared radiation. This is the basis of static area telethermography (SAT) and dynamic area telethermography (DAT). SAT has recently been shown to be of help to the dentist in the diagnosis of chronic orofacial pain, as a unique tool in assessment of TMJ (temporomandibular joint) disorders, as an aid in assessment of inferior alveolar nerve deficit, and as a promising research tool.

(Source: [40].)

Moiré topographic imaging

Moiré fringes are formed when one line or grid pattern is superimposed upon a similar line or grid pattern. The use of moiré fringes to acquire 3-D surface shape information is well established. Their application to the measurement of areas of the human body began with the work of Hiroshi Takasaki as early as 1973 when he successfully applied moiré topography to the measurement of the human body for medical purposes. Early work in this field, however, employed a wholly manual fringe analysis method to evaluate contours from photographs produced by the shadow moiré method. This form of analysis is slow and heavily reliant on prior knowledge of the object to resolve the ambiguities present in complex moiré patterns. More recent work made use of commercially available semi-automatic moiré fringe analysis equipment to produce cross-sections of parts of the face. This analysis method requires some operator intervention and the analysis time is lengthy at around two minutes per cross-section. It is really only since the early 1980s that advances in video and computing technology have facilitated the development of a fully automated, moiré-based imaging system.

A grating is projected onto the object and an image of the object formed in the plane of a reference grating. The image is detected by a CCD array camera, linked via a video frame digitizer to a desktop computer. The interaction of the superimposed projection grating lines with the reference grating causes moiré fringes to be produced which appear superimposed on the surface of the object being measured. As the projected grating is distorted by the irregularities in the shape of the object's surface, the resulting fringe pattern describes surface contours. In moiré contouring systems, computer software detects and analyses the fringe pattern, and from this produces 3-D coordinates for a number of points on the surface of the object.

For example, spinal deformity is a serious problem mainly for teenagers and medical doctors inspect moiré topographic images of their backs visually for the screening. If a subject is normal, the moiré image is almost symmetric with respect to the middle line of the subject's back, otherwise it shows asymmetric shape.

(Source: [1], [35].)

Digital Radiography

Digital Radiography refers to the application of digital image processing techniques to projection radiography. A wide range of digital image acquisition techniques are possible in digital radiography. These include:

- Film Digitization involves using a video camera or LASER scanner to digitize previously-generated radiographs.

- Digital Fluoroscopy involves digitization of the output of the video camera of a fluoroscopy system, operated using low, continuous radiation exposures.

- Digital Fluorography is exactly the same as Digital Fluoroscopy, but with the imaging system operated using intense, pulsed radiation exposures.

- Computed Radiography involves the digitization of images acquired using photostimulable phosphor technology.

In addition, a large number of radiological applications have been developed for the Digital Fluoroscopy/Fluorography image acquisition modes, such as:

- Digital Subtraction Angiography involves subtraction of images of opacified and non-opacified blood vessels.

- Digital Cardiac Imaging relates to the digitization of images produced in cardiac angiography.

- Digital Spot Imaging relates to the digitization of images produced in, for example, studies of the alimentary system.

It should be noted that the term Digital Radiography could also be used to encompass X-Ray Computed Tomography in that CT also uses an imaging system which is interfaced to a digital computer. From the same viewpoint, Digital Radiography could also encompass techniques which are used for bone densitometry and which rely on acquiring images at two different x-ray energies, e. g. Dual-Energy X-Ray Absorptiometry (DEXA).

Digital radiography is a major subject of image processing research, since it offers high quality pictures with different noise models than for classical radiography. In the last years digital radiography was evaluated in many clinical studies and found its way to hospitals and practices. It is expected that the performance of such systems will highly increase in the next decade. Therefore image processing methods for high resolution images is needed.

(Source: Adapted from [37].)

3.2 Classical Image Manipulation

Computerized Imaging

The objective of computerized imaging is to obtain high quality images from data which may be noisy or incomplete. Denoising and interpolation are the main operations. For denoising today statistical and wavelet methods play the main role, for interpolation also wavelets and in addition splines are important tools.

Noise Models: In spite of the continuing sophistication of medical image acquisition hardware, postprocessing to reduce noise can still be very useful. The noise influence for some different medical imaging techniques can be summarized as follows.

Noise in X-ray CT is due to the Poisson statistics of the X-ray photons, to beam hardening, and to photon scatter, in addition to blur by motion and partial volume effects. The raw measurements are transformed into tomographic images via one of the possible reconstruction methods. The particular method evidently influences how the noise of the measurement results in noise in the images. The various phenomena are often studied by reconstruction from simulated projection data, in particular from projections of mathematically described objects called phantoms.

The noise in emission tomography measurements such as PET or SPECT has Poisson distribution. Reconstruction from these measurements involves filtering and other convolutions, and for PET, also corrections, followed by back-projection, which can be iterative. These steps again determine the resulting noise in the images. Simulation is therefore also for PET and SPECT an appropriate way to study noise influences.

Noise in MR images is uncorrelated, i.e., white. Its distribution depends on the signal-to-noise ratio (SNR) of the image. For SNRs larger than 10 to 15 dB, the distribution is Gaussian. A Rayleigh distribution is an appropriate model for lower SNRs.

Images produced with ultrasound techniques suffer from speckle noise. It is caused by interference of reflected ultrasonepulses at the transductor surface. Noise suppression through image postprocessing seems, at this point, not effective. Current research efforts concern mainly the improvement of ultrasone transducers and models for the reflections in tissue.

Similar noise influences are also involved in the acquisition of biological images. For instance, noise in electron micrographs as well as in gel electropherograms is often modeled as additive Gaussian white noise.

Denoising: Today wavelet based methods outperform the best comparable

earlier methods, such as the adaptive Wiener filter. This can be demonstrated by quantitative results, such as SNR gain and by the qualitative appearance of the images. Especially probabilistic wavelet methods are fast and easy to use.

It is not straightforward to remove all noise in biomedical images at the time of image acquisition. Evidently, all depends on the requirements of the application. In biomedical imaging it seems commendable to spend some more effort on the method in return for the best denoising quality.

(Source: [38].)

Image Coding and Compression

Compression in general is intended to provide efficient representations of data while preserving the essential information contained in the data. The main idea is to find efficient digital representations and compressions of digital information into the fewest possible bits. Such digital information can be contained in speech, audio, image and video signals.

Both operations — image coding and compression — should yield the highest possible reconstruction fidelity subject to constraints on the bit rate and implementation complexity. The conversion of signals into such efficient digital representations has several goals:

- to minimize the communication capacity required for transmission of high quality signals such as speech and images or, equivalently, to get the best possible fidelity over an available digital communication channel,
- to minimize the storage capacity required for saving such information in fast storage media and in archival data bases or, equivalently, to get the best possible quality for the largest amount of information stored in a given medium.
- to provide the simplest possible accurate descriptions of a signal so as to minimize the subsequent complexity of signal processing algorithms such as classification, transformation, and encryption.

In addition to these common goals of communication, storage, and signal processing systems, efficient coding of both analog and digital information is intimately connected to a variety of other fields including pattern recognition, image classification, speech recognition, cluster analysis, regression, and

decision tree design. Thus techniques from each field can often be extended to another and combined signal processing operations can take advantage of the similar algorithm structures and designs.

(Source: Adapted from [25].)

Image Coding: In the 1940s and the 1950s Claude Shannon developed a theory of source coding in order to quantify the optimal achievable performance trade-offs in analog-to-digital conversion (A/D) and data compression systems. Two of his fundamental ideas lead to a variety of coder design techniques over time: The first idea was that purely digital signals could be compressed by assigning shorter codewords to more probable signals and that the maximum achievable compression could be determined from a statistical description of the signal. This led to the idea of noiseless coding, which is often called entropy coding. The second idea was that coding systems can perform better if they operate on vectors or on groups of symbols (such as speech samples or pixels in images) rather than on individual symbols or samples. Although the first idea led rapidly to a variety of specialized coder design techniques, the second idea took many years before yielding useful coding schemes.

In the 1980s vector coding or vector quantization has come of age and made an impact on the technology of signal compression. Several commercial products for speech and video coding have emerged which are based on vector coding ideas.

During the past ten years, vector quantization has proved a valuable coding technique in a variety of applications, especially in voice and image coding. In recent articles new developments like fast fractal image coding and hierarchical representations play the most important role.

(Source: Adapted from [25].)

Image Compression: Approaches are ranging from wavelets and fractals, to ideas from pattern recognition and computer vision. Today there are high standards for lossless as well as lossy coding and compression, as for example the well known JPEG 2000 standard, which is a lossy wavelet compression.

Therefore research concentrates on coding and compression for special classes of pictures, e.g. the compression of radiographies or ultrasound-images and films. For applications as teleimaging a lossy compression would be favored because of its high performance, but for medical images a lossless compression to ensure high quality diagnosis is needed.

A transform coder decomposes a signal in orthogonal basis and quantizes the decomposition coefficients. The distortion of the restored signal is minimized by optimizing the quantization, the basis and the bit allocation. For signals that are realizations of a Gaussian random vector, a high resolution quantization yields a mean-square distortion that minimizes in the Karhunen–Loève basis.

For non-Gaussian signals such as images, whose coefficients are coarsely quantized, the distortion depends on the precision of non-linear approximations in the basis. The image compression with wavelet bases or cosine block bases for JPEG can be improved by embedding strategies that use a partial ordering of the coefficients' amplitude. Embedded code can transmit a coarse image approximation quickly, and then progressively refine the quality by adding more bits.

In a grey-level image each pixel is typically coded with 8 bits. Images include many types of structures that are difficult to model. Currently, the best image compression algorithms are transform codes, with cosine or wavelet bases. The efficiency of these bases comes from their ability to construct precise non-linear image approximations with few non-zero vectors. With fewer than 1 bit/pixel, visually perfect images are reconstructed. At 0.25 bit/pixel, the image remains of good quality.

Applications of digital video range from low quality videophones and teleconferencing to high resolution television. The most performant compression algorithms remove the redundancy with a motion compensation. Local image displacements are measured from one frame to the next, and are coarsely approximated with a few motion vectors. Each frame is predicted from previous ones by compensation the motion that is encoded. An error image is calculated and compressed with a transform code. The MPEG standards implement such motion-compensated video compression.

The High Definition Television (HDTV) format has color images of 1280 by 720 pixels, and 60 images per second. The resulting bit rate is of the order of 10^3 Mb/s. To transmit the HDTV through channels used by current television technology, the challenge is to reduce the bit rate to 20 Mb/s, without any loss of quality.

(Source: [39].)

Interpolation and Approximation

The development of orthonormal wavelet bases has opened a new bridge between approximation theory and signal and image processing. This exchange

is not quite new since the fundamental sampling theorem comes from an interpolation theory result proved in 1935 by Whittaker. However, the state of the art of approximation theory has changed since 1935. In particular, the properties of non-linear approximation schemes are much better understood, and give a firm foundation for analyzing the performance of many non-linear signal-processing algorithms.

A further degree of freedom can be introduced by choosing the basis adaptively, depending on the signal properties. From families of wavelet packet and local cosine bases, a fast dynamical programming algorithm is used to select a "best" basis that minimizes a Schur concave cost function. The approximation vectors chosen among this "best" basis outline the important signal structures, and characterize their time-frequency properties. Pursuit algorithms generalize these adaptive approximations by selecting the approximation vectors among redundant dictionaries of time-frequence atoms, with no orthogonality constraint. These procedures are sufficiently flexible to build compact representations of complex signals.

(Source: [39].)

Pattern classification, feature detection, and scene analysis

Pattern classification is the assignment of a physical object or event to one of several prespecified categories. Extensive study of classification problems has led to an abstract mathematical model that provides the theoretical basis for classifier design. Of course, in any specific application one ultimately must come to grips with the special characteristics of the problem at hand. Of the various problem area, the domain of pictorial problems has received by far the most attention.

The classification model contains tree parts: a transductor, a feature detector and a classifier. The transductor senses the input and converts it into a form suitable for machine processing. The feature detector extracts presumably relevant information from the input data. The classifier uses this information to assign the input data to one of a finite number of categories. The problem of feature detection is much more a problem dependent than the problem of classification, which is basically one of partitioning the feature space into regions, one region for each category. Ideally, one would like to arrange this partitioning so that none of the decisions is ever wrong. Since this is not possible in general, the aim is to minimize the probability of error or the average cost of error. The problem of classification becomes a problem in statistical decision theory, a subject that has many applications to pattern classification. Methods like Fisher's linear discriminant, perception and

relaxation procedures, minimum-squared-error methods, stochastic approximation, potential functions and linear programming techniques are widely used. New techniques include unsupervised learning and clustering.

Scene analysis uses simplification techniques for pictures for suppressing irrelevant detail characterizing shapes and sizes of objects in a picture, integrating parts of a picture into meaningful entities and in general reducing the complexity of the data. Spacial differentiation and smoothing, template matching, region analysis and contour following are used for initial simplifications.

(Source: Adapted from [14].)

Feature detection in mammograms: One of the most important feature detection research areas is mass detection in mammograms. Recent approaches where done by a sector-form model in the template matching process and by analyzing oriented flow-like textural information along with features in adaptive ribbons of pixels along the margins.

Especially applications of feature detection in cancer prevention and treatment show that there is huge need on optimization of existing methods for fast, sensitive and specific feature detection.

Motion Estimation: Ultrasound is an effective imaging modality that enables the clinician to study shape, size, and dynamics of organs. To estimate the motion extended optical flow algorithms based on wavelets or B-splines are adapted to the respective medical application.

Motion estimation over a block of pixels is a standard approach for estimating motion in a moving image sequence. In block motion estimation, a rectangular block of pixels is sequentialy compared to rectangular blocks of pixels within a search range in a neighboring frame. A distortion measure is applied to each different possible displacement in order to find the best match, which is then chosen as the motion vector.

(Source: [56] and [46].)

Cortical imaging: Intrinsic signals within the brain are often observed using cameras capable of measuring small signal changes. Cortical imaging involves measuring scattered infrared light from the surface of the cerebral cortex due to enhanced electrical activity. Illuminating light levels are set to approach the saturation level of the camera so that the greatest signal-to-noise ratio can be obtained. A controlled stimulus (such as a change in a visual cue presented to the animal or the movement of a whisker) triggers a sequence of

images. Changes in the intensity of the scattered light indicates processing within certain regions of the brain. This leads to segmentation and classifying problem.

(Source: Adapted from [50].)

Neonatal seizures: Neonatal seizures are paroxysmal alterations in neonatal behavior and/or motor or autonomic function, initiated by hypersynchronous activity of neurons in the brain. Recently there are many publications on feature detection algorithms to detected neonatal seizures for avoiding child death.

3.3 Soft computing

Soft computing differs from conventional (hard) computing in that, unlike hard computing, it is tolerant of imprecision, uncertainty and partial truth. In effect, the role model for soft computing is the human mind. The guiding principle of soft computing is: Exploit the tolerance for imprecision, uncertainty and partial truth to achieve tractability, robustness and low solution cost.

The basic ideas underlying soft computing in its current incarnation have links to many earlier influences, among them fuzzy sets in the 1960s. The inclusion of neural network theory in soft computing came at a later point. At this juncture, the principal constituents of soft computing are fuzzy logic, neural network theory and probabilistic reasoning, with the latter subsuming belief networks, genetic algorithms, chaos theory and parts of learning theory. What is important to note is that soft computing is not a melange of fuzzy logic, neural network theory and probabilistic reasoning. Rather, it is a partnership in which each of the partners contributes a distinct methodology for addressing problems in its domain. In this perspective, the principal contributions of fuzzy logic, neural network theory and probabilistic reasoning are complementary rather than competitive.

The complementarity of fuzzy logic, neural network theory and probabilistic reasoning has an important consequence: in many cases a problem can be solved most effectively by using fuzzy logic, neural network theory and probabilistic reasoning in combination rather than exclusively. A striking example of a particularly effective combination is what has come to be known as neurofuzzy systems. Such systems are becoming increasingly visible as consumer products ranging from air conditioners and washing machines to

photocopiers and camcorders. Less visible but perhaps even more important are neurofuzzy systems in industrial applications. What is particularly significant is that in both consumer products and industrial systems, the employment of soft computing techniques leads to systems which have high Machine Intelligence Quotient. In large measure, it is the high Machine Intelligence Quotient of soft computing-based systems that accounts for the rapid growth in the number and variety of applications of soft computing - and especially fuzzy logic.

In the last years soft computing got more and more important in biomedical imaging. In the context of medical imaging, soft computing technique appears as a power framework since it provides tools adapted to this task. Its main properties are:

1. It provides a way to represent and manipulate imprecise and uncertain information.
2. It will be able to demonstrate knowledge of Medical doctors.
3. It is well adapted to image processing since the natural spatial interpretation of soft computing leads to efficient representations of imprecise or implicit structures or classes in pictures.

Especially, a mixture of pattern recognition and soft computing-aided expert system techniques lead to successful tasks to segmentation of regions of interests and to give enhanced anatomical and functional resolution. This explosion of new techniques also emphasizes the need to integrate and focus the efforts of scientists and clinicians to facilitate communication, establish standards, and develop training programs.

The conceptual structure of soft computing suggests to apply not just in neural network theory or fuzzy logic or probabilistic reasoning but all of the associated methodologies, though not necessarily to the same degree. In the last years there began to appear journals and books with soft computing in their title. A similar trend is visible in the titles of conferences. Soft computing seems to be a promising research field.

(Sources: Adapted from [65] and [6].)

3.4 Real time processing

Real-time imaging is concerned with various imaging techniques, technologies and systems where timing constraints are as critical as being logically correct.

Today there are many good and effective algorithms for denoising, interpolation, compression, feature detection, etc. for medical images, but many of them need too much processing time. Therefore they are not usable for real time applications, as it is needed for clinical studies. Many researchers are working on speeding up existing methods. It is major work field in mathematics, informatics and electrical engineering.

Recent advances have for example been made for angiographic operations using catheters. A new real time technique in NMR tomography gives information on the location of the catheter as well as a 3D image of the arteries.

(Source: Adapted from [55].)

3.5 Modelling and Visualization

Recent developments in the computerized analysis of medical images are expected to aid radiologists and other healthcare professionals in various diagnostic tasks of medical image interpretation. In medical imaging, the accurate diagnosis and/or assessment of disease depends on both image acquisition and image interpretation. The role and contribution of radiology to medical diagnosis has expanded tremendously due to advances in image quality compliance regulations, image detector systems, and computer technology.

The image interpretation process has only recently begun to benefit from computer technology. Most interpretations of medical images are performed by radiologists; however, image interpretation by humans is limited due to the nonsystematic search patterns of humans, the presence of structure noise (camouflaging normal anatomical background) in the image, and the presentation of complex disease states requiring the integration of vast amounts of image data and clinical information.

(Source: Adapted from [26].)

Computer Aided Diagnosis

CAD, defined as a diagnosis made my a radiologist who uses the output from a computerized analysis of medical images as a second opinion in detecting lesions, assessing extent of disease, and making diagnostic decisions, is expected to improve the interpretation component of medical imaging. With CAD, the final diagnosis is made by the radiologist.

With the advances of CAD also the training of health care professionals becomes possible. Data bases and computer simulations of diseases and injuries help to improve the surgeons expertise, avoid mistakes in real operations and thus lead to high quality in health care at relatively low cost.

Screening: The screening a of asymptotic people involves radiologists visually scanning the images of mostly healthy subjects for a specific abnormality. The interpretation of screening images lends itself to CAD since it is a repetitive, burdensome task involving mostly normal images—a situation prime for oversight errors. Many computerized analysis methods have been developed for screening mammography, which are nowadays approved for clinical use.

The role of CT in screening programs is also rapidly growing especially in the thorax for lung cancer screening and in the colon (colonography) for the detection of suspect polyps for colon cancer.

Many CAD papers during the last two decades have involved either mammograms or chest radiographs. This early research was performed on digitized radiographs. While the computerized analysis of mammographs is mainly focused on one disease, breast cancer, the computerized analysis of chest radiographs ultimately requires the diagnosis of a multitude of diseases (e.g., lung cancer, pneumothorax, interstitial diseases).

It is expected that the development and implementation of computer techniques for projection radiography of the chest will advance rapidly with the advent and acceptance of digital chest imaging units.

For the screening of lung cancer, single-projection chest radiographs and thoracic CT scans have been considered. The existence of 3-D image data from CT removes much structure noise. These 3-D data sets greatly increase the number of images that must be reviewed by a radiologist in a screening program—leading to an overwhelming task for the human search process. Accordingly, image interpretation may greatly benefit from a computer search aid.

Malign/Benign decisions: Once a lesion is detected, characterization is necessary to determine the status of the lesion, e.g., the likelihood that the lesion is cancerous. Complex anatomy, variation in the presentation of malignant and benign states, and varying abilities of the radiologist can lead to interpretation errors. Methods for the computerizes analysis of a medical image contain many stages. Improvement at one stage may influence performance at later stages and, subsequently, the overall performance. It is interesting to note that many changes in an image that lead to improvements in human interpretation performance also lead to improvements in computer image interpretation performance.

Computerized image analysis has been applied mainly to medical imaging

techniques such as X-ray, sonography, and MRI. A role for computer-aided diagnosis is emerging for applications involving less well-known modalities such as thermography and Moiré imaging.

Evaluation: With the new development in computerized medical image analysis, investigators, journal readers, and end-users are anxious to ask the question, "How good is the new technique compared with other techniques?" It is difficult to compare the various computerized methods under development due to the use of different databases and the varying criteria for reporting and evaluating computer results. While an independent test site with an independent database for the evaluation of each new development would be quite beneficial, it is not yet always practical. However, some guidelines may assist the communication of the merits of a new technique through publication in the scientific literature. It is a future challenge to all investigators who aim to publish to incorporate these suggestions.

Performance gain: The ultimate acceptance of CAD will depend not only on the performance of the computerized method alone, but also on how well the human performs the task when the computer output is used as an aid and on the ability to integrate the computerized analysis method into routine clinical practice. Observer studies have shown that radiologists' performance increased when using a computer output as an aid. It is important to note, that observer experience may influence the reported effect of CAD. Thus, it is important to determine and report the amount of experience of the observers and their current reading load in the relevant diagnostic gain. It should be noticed that a computerized method will be useful even at a less-than-perfect sensitivity, especially if the lesions detected by the computer do not overlap completely with those detected by a radiologist.

Clinical Practice: Integration of CAD into clinical practice has been shown for screening mammography, and the introduction of other CAD methods to the clinical area is awaited. In the clinical setting, CAD methods might be used routinely as part of a screening protocol or used only when requested by a radiologist interpreting a particular case. Ultimately, a CAD workstation would be configured for each radiologist to allow individual control over the sensitivity and specificity of the computer output with adjustment depending on the nature of the case material and personal preference.

The practice of interpreting medical images is being modified by the information technology revolution and it appears that both the medical profession and the public (patients) welcome enthusiastically such advances. In the future, it is quite likely that all medical images will undergo some form of computer analysis in order to benefit the diagnosis.

(Source: Adapted from [26].)

Computer Aided Surgery

Computer Aided Surgery or Computer Assisted Surgery is concerned with supporting the physician with complementing visualizations of the human body. This gives new information to the physician and helps him to make the right medical decisions while treating the patient.
The research fields in this area can be categorized into

- Visualization of medical images

  for organ detection and movement, visualization of FNMR data, enhancement of important features in ultrasound images,...

- Simulation and virtual reality
- Surgical planning using simulation techniques
- Surgical navigation

  e.g., in endoscopic surgery

- Automation and control for patients under surveillance

  helping to detect newborn's seizures, loss of blood pressure, and others

- Education and training for new surgery techniques

Computer Aided Surgery is the final step at the long image processing ladder. It gives the practical application of image denoising, interpolation, compression, analysis etc. in real time systems and needs to be evaluated hand in hand with physicians. High quality and particularity are a must owing respect to the patients.

(Source: See also [34].)

Chapter 4

Biological Imaging

The power of imaging technology allows to do more than simply view structures anatomically, such as s visualizing bones, organs, and tumors in the body. Functional imaging — the visualization of physiological, cellular, or molecular processes in living tissue — allows to see such things as blood flow, oxygen consumption, or glucose metabolism in real time, as they take place in living cells of the body. In gaining a better understanding or the fundamental nature of cancer and other diseases, cellular and molecular imaging will be a key tool in translating this knowledge into better ways of diagnosis, treating, and preventing these diseases.

Imaging can identify the kinds of molecular structures/receptors that cover the surface of a tumor, information that potentially can predict how it may behave and respond to certain treatments. And seeing how the processes and pathways inside a cell change as the cell transforms from normal to cancerous will allow to detect this change in people earlier in the cancer process, perhaps before a tumor has even had the chance to become fully malignant.

Parallel developments in image enhancement agents are improving the ability to capture changes in the biochemical makeup of cells and other living structures. These

This section introduces common small scale modalities for biological imaging, gives fields of application and shows new approaches in recent research. They allow anatomic and receptor localization as well as activate certain imageable biophysical processes.

(Source: Adapted from [43].)

4.1 Modalities

As in medical imaging modalities as X-ray, ultrasound, magnetic resonance, emission tomography, electrical source imaging, impedance tomography are common modalities in biological imaging. In addition for imaging the very small scales electron microscopy, atomic force microscopy, bioluminescence and biofluorescence and others are used.

Light Microscopy

In the light microscope or optical microscope light from the microscope lamp passes through the condenser and then through the specimen. Some of the light passes both around and through the specimen undisturbed in its path (undeviated light). Some of the light passing through the specimen is deviated when it encounters parts of the specimen. Such deviated light is rendered one-half wavelength out of phase with the direct light. This leads to causes destructive interferences with the direct light when both arrive at the intermediate image plane located at the fixed diaphragm of the eyepiece. The eye lens of the eyepiece further magnifies this image which finally is projected onto the retina, the film plane of a camera, or the surface of a light sensitive computer chip.

The undeviated light is projected by the objective and spread evenly across the entire image plane at the diaphragm of the eyepiece. The light diffracted by the specimen is brought to focus at various localized places on the same image plane, where the diffracted light causes destructive interference, and reduces intensity resulting in more or less dark areas. These patterns of light and dark are what one recognizes as an image of the specimen. Because our eyes are sensitive to variations in brightness, the image becomes a more or less faithful reconstruction of the original specimen.

Reflected light microscopy is often referred as to incident light, epi-illumination, or metallurgical microscopy, and is the method of choice for fluorescence and for imaging specimens that remain opaque even when ground to a thickness of 30 microns. The range of specimens falling into this category is enormous and includes most metals, ores, ceramics, many polymers, semiconductors, slag, coal, plastics, paint, paper, wood, leather, glass inclusions, and a wide variety of specialized materials. Because light is unable to pass through these specimens, it must be directed onto the surface and eventually returned to the microscope objective by either specular or diffused reflection.

There are many specialized methods such as darkfield and phase contrast microscopy, as well as illumination and contrast techniques. Although a lot

of achievements where gained in hard ware techniques the development of digital imaging devices needs good software algorithms for image enhancement.

(Source: Adapted from [13].)

Fluorescence microscopy

Fluorescence microscopy is an excellent tool for studying material which can be made to fluoresce, either in its natural form (primary or autofluorescence) or when treated with chemicals capable of fluorescing (secondary fluorescence). This form of optical microscopy is rapidly reaching maturity and is now one of the fastest growing areas of investigation using the microscope.

The basic task of the fluorescence microscope is to permit excitation light to irridiate the specimen and then to separate the much weaker re-radiating fluorescent light from the brighter excitation light.

Fluorescence microscopy has advantages based upon attributes not as readily available in other optical microscopy techniques. The use of fluorochromes has made it possible to identify cells and sub-microscopic cellular components and entities with a high degree of specificity amidst non-fluorescing material. What is more, the fluorescence microscope can reveal the presence of fluorescing material with exquisite sensitivity. An extremas small number of fluorescing molecules (as few as 50 molecules per cubic micron) can be detected. Although the fluorescence microscope cannot provide spatial resolution below the diffraction limit of the respective objectives, the presence of fluorescing molecules below such limits is made visible.

Techniques of fluorescence microscopy can be applied to organic material, formerly living material or to living material with the use of in vivo or in vitro fluorochromes, but also to inorganic material, e.g. for the investigation of contaminants on semiconductor wavers. There are also a burgeoning number of studies using fluorescent probes to monitor rapidly changing physiological ion concentrations and pH values in living cells.

(Source: Adapted from [13].)

Bioluminescence optical imaging

Bioluminescence is light produced by a chemical reaction which originates in an organism. Bioluminescence is not the same as fluorescence. In fluorescence, energy from a source of light is absorbed and reemitted as another photon.

In bioluminescence or chemiluminescence the excitation energy is supplied by a chemical reaction rather than from a source of light.

Bioluminescence is primarily a marine phenomenon. It is the predominant source of light in the largest fraction of the habitable volume of the earth, the deep ocean. In contrast, bioluminescence is essentially absent (with a few exceptions) in fresh water, even in Lake Baikal. On land it is most commonly seen as glowing fungus on wood, or in the few families of luminous insects. Bioluminescence has evolved many times in the sea as evidenced by the several distinct chemical mechanisms by which light is emitted and the large number of only distantly related taxonomic groups that have many bioluminescent members. Bioluminescent bacteria occur nearly everywhere, and probably most spectacularly as the rare "milky sea" phenomenon, particularly in the Indian Ocean where mariners report steaming for hours through a sea glowing with a soft white light as far as the eye can see.

In the sea, bioluminescent light is concentrated in the blue window of greatest optical transparency of seawater. Most organisms emit between 440 nm and 479 nm. Some have green fluorescent proteins that absorb an initially blue emission and emit it shifted towards the green (approx. 505 nm). One remarkable fish has a similar mechanism to shift the initial emission into the red for use in viewing prey in the near infrared with its red-sensitive eyes. Measurements in situ at various depths confirm emission clustering in the blue to green region of the spectrum.

The luminescence of a single dinoflagellate is readily visible to the dark adapted human eye, as the demonstration will show. Most dinoflagellates emit about $6 \cdot 10^8$ photons in a flash lasting only about 0.1 second. Much larger organisms such as jellyfish emit about $2 \cdot 10^{11}$ photons per second for sometimes tens of seconds. The intensity of luminescence by photosynthetic dinoflagellates is strongly influenced by the intensity of sunlight the previous day. The brighter the sunlight the brighter the flash.

Some organisms emit light continuously, but most emit flashes of durations ranging from about 0.1 s to 10 s. Some dinoflagellates can respond repetitively to excitation over a short period. In most multicellular species luminescence is neurally controlled. Thus in some fish the sympathetic nervous system controls luminescence by way of the neurotransmitter nor-adrenaline. In fireflies the transmitter is glutamate. In most marine invertebrates the transmitters are unknown. In such forms the "trigger" to luminescence is some detected behaviorally significant event.

In single cell organisms like dinoflagellates or radiolarians luminescence is triggered by deformation of the cell surface by minute forces (1 dyne per square cm). Mechanical deformation causes an action potential sweeping over

the vacuole membrane and this is thought to induce light emission by admitting protons from the acidic vacuole into contact with the cellular elements that contain the light emission chemistry.

In a some instances in marine invertebrates with eyes or other light receptors, light emission can be induced by photic excitation, even by another luminescing organism. Called "empathetic" luminescence, this phenomenon has as yet undemonstrated potential to enhance the luminescence generated by a moving source by photic transfer from the luminescent organisms mechanically triggered by the moving source.

Because so many marine organisms are bioluminescent, measurements of stimulated bioluminescence in the oceans are valuable for determining the distribution patterns of different populations. Such measurements are useful over a wide range of scales, from kilometers down to centimeters. Coarse-scale measurements have been made using airborne image intensifying TV cameras to detect schools of fish that are made visible at night by the bioluminescent plankton, which they stimulate as they swim. Fine-scale measurements of stimulated bioluminescence in the ocean are made with bathyphotometers. These instruments pull sea water, containing organisms, through a pipe into a light-tight chamber. There, a light detector measures the bioluminescence, stimulated by some turbulence-generating device like a paddle wheel. The light is measured in photons per second, or watts or sometimes in numbers of flashes stimulated.

For micro-scale studies an intensified video transect technique is used to identify and map bioluminescent organisms based on the spatial and temporal properties of their stimulated bioluminescent displays. Video data is collected with an intensified camera.

Combining micro-scale and fine-scale measurements of bioluminescence gives a rapid means of assessing plankton distribution patterns relative to the physical and chemical variables in the environment.

Image recognition algorithms identify organisms based on their bioluminescent displays and map their locations in three-dimensional space. Then spatial point analyses of these data can be performed.

(Source: Adapted from [27] and [28].)

Confocal microscopy

Confocal laser scanning microscopy (CLSM) is a relatively new light microscopical imaging technique. It was introduced around 1980 by M. Petran and A. Boyde and has found wide applications in the biological sciences. The

primary value of the CLSM to the biologist is its ability to produce optical sections through a 3-dimensional specimen — e.g., an entire cell or a piece of tissue — that, to a good approximation, contain information from only one focal plane. Therefore, by moving the focal plane of the instrument step by step through the depth of the specimen, a series of optical sections can be recorded. This property of the CLSM is fundamental for solving 3-D biological problems where information from regions distant from the plane of focus can obscure the image, for example in thick objects. With biological specimens, either the epi-fluorescence or the epi-reflection mode is generally employed. As a valuable by-product, the computer-controlled CLSM produces digital images which are amenable to image analysis and processing, and can also be used to compute surface- or volume-rendered 3-D reconstructions of the specimen.

To image the specimen point by point with a confocal microscope, a collimated, polarized laser beam is deflected stepwise in the x- and y-direction by a scanning unit before it is reflected by a dichroic mirror (beam splitter) so as to pass through the objective lens of the microscope, and focused onto the specimen. The emitted, longer-wavelength fluorescent light collected by the objective lens passes through the dichroic mirror (transparent for the longer wavelength) and is focused into a small pinhole (i.e., the confocal aperture) to eliminate all the out-of-focus light, i.e., all light coming from regions of the specimen above or below the plane of focus.

Therefore, the confocal microscope does not only provide excellent resolution within the plane of section (0.25 mm in x- and y-direction), but also yields similarly good resolution between section planes (0.3 mm in z-direction). The in-focus information of each specimen point is recorded by a light-sensitive detector positioned behind the confocal aperture, and the analog output signal is digitized and fed into a computer. At the same time, the analog photo-multiplier signal can be used to generate a TV-like image on a video monitor. The obvious advantage of having a stack of serial optical sections through the specimen pixel by pixel in digital form is that either a composite projection image can be computed, or a volume-rendered 3-D representation of the specimen can be generated on a graphics computer.

The confocal part of a CLSM consists of an elaborate, highly folded optical bench on which the laser, all the filters, an oscillating-mirror or acousto-optic scanning device, and the detector are mounted. When working in the epi-fluorescence mode, the laser beam is filtered to select the 488 nm, 568 nm or 647 nm wavelength line from an Argon/Krypton laser, and a triple dichroic mirror is used to transmit — rather than reflect — the longer-wavelength fluorescence signal to the detector. For the epi-reflection mode no wavelength

filters are needed. Instead, a semi-transparent mirror reflects 50% of the incident laser beam through the objective lens and to reach the specimen, and it transmits 50% of the light reflected by the specimen and collected by the objective lens to the detector. To suppress light reflected by the various optical elements of the microscope, a 1/4-wavelength plate and a polarizing filter are put into the beam path.

Usually, objective lenses with a high numerical aperture are used to provide good resolution in x-, y- and z-direction. Often, good specimen areas are sparse, so to find them the specimen is screened in the conventional transmission or epi-fluorescence mode. Once a good region has been located, the CLSM mode is activated, and serial optical sections are recorded at user-selectable depth increments which can be as small as 0.02 mm. In digital form, each image element yields an 8-bit intensity value in the range 0-255. Typically, the image frame size is 512 × 512 pixels, and in the best case images are recorded at video rate, i.e., at 25 frames per second in PAL or 30 frames per second in NTSC norm. To improve the signal-to-noise ratio of individual image frames, several of them may be recorded in series and averaged. In the case of epi-fluorescence imaging, the total number of scans is generally adjusted to limit specimen bleaching to an acceptable level.

In general, thick and opaque specimens that can barely be observed in a conventional light microscope are excellent specimens when it comes to demonstrate the power of a CLSM. For example, 20-25 mm thick sections of bone, cartilage or muscle are ideally suited for 3-D imaging in the CLSM. Independent of the thickness and surface quality of such tissue sections, individual confocal planes readily reveal a lateral resolution of 0.3 mm. By recording its auto-fluorescence, even a piece of wood can be optically sectioned to a depth of about 100 mm.

Cultured fibroblasts grown as monolayers can be multiple-labeled with fluorochrome-tagged antibodies against different cytoskeletal components: e.g., actin labeled with rhodamine for red fluorescence or tubulin labeled with fluorescein for green fluorescence. In such cells, actin is primarily in the form of stress fibers (red fluorescence), actin filament bundles adhering to the plasma membrane attaching the cell to the coverslip on which it has been grown. In contrast, the microtubules (green fluorescence) usually radiate outward from the two perinuclear centrioles and reach all parts of the cell. During cell division the microtubules form the spindle apparatus that separates the condensed chromatin, i.e., the chromosomes.

The CLSM allows the 3-D distribution and relative spatial relationship of these two filament systems to be visualized directly.

Exploring the growth and differentiation of cultured cells in 3-D collagen ma-

trices — a condition which more closely resembles the natural environment of cells — has been another application of the CLSM. While difficult to assess by conventional fluorescence microscopy, the spatial relationship of the various cytoskeletal components relative to the different sub-cellular compartments in these matrix-embedded cells is readily determined by CLSM.

(Source: Adapted from [15].)

Polarizing Microscopy

The simplest tool to study three-dimensional arrangements is polarizing microscopy (PM). PM tests the orientation of optical axes of the liquid crystal specimen; these optical axes are closely related to the molecular arrangements in the medium. Unfortunately, PM yields only two-dimensional textures in the so-called plane of observation which is perpendicular to the optical axis of the microscope. This 2D image integrates the true 3D configuration of optical birefringence over the path of light. As the result of such an integration, the director profile along the direction of observation (i.e., the "vertical cross section" of the specimen) is hard to decipher. Regrettably, it is precisely the director configuration in the vertical cross-section that is often the most valuable and desirable.

The fluorescence confocal polarizing microscopy (FCPM) allows one to recover the missing information and to obtain a truly 3D image of the liquid crystal director, both in the plane of observation and along the direction of observation. The principle of imaging is different from the traditional PM. The FCPM maps the intensity of polarized fluorescent light emitted by the liquid crystal sample, rather than the pattern of integrated birefringence as the PM texture does. This feature allows one to avoid the ambiguity of the in-plane PM textures that do not distinguish between two mutually perpendicular director configurations. More importantly, the confocal scheme allows one to collect the fluorescent light from a very small region of the sample and thus to optically slice the specimen by scanning the focused laser beam. The obtained map of fluorescence intensity is the 3D image of orientation of the fluorescent probe.

Fluorescence Confocal Polarizing Microscopy is very similar to Fluorescence Confocal Microscopy, with two basic distinctive features:

(a) the incident light is polarized and

(b) the fluorescent dye molecules are aligned by the "host" material.

Consider a nematic liquid crystal cell doped with a fluorescent dye. For simplicity, assume that the nematic is of a calamitic type (elongated molecules) and that the fluorescent molecules are also elongated. The transition dipoles of both excitation and fluorescence are along the long axis of the dye molecule. As it is well known, from the earlier studies of the so-called "guest-host" display modes, the anisometric "guest" molecules are aligned by the nematic "host".

The three requirements to the fluorescent probe for FCPM are anisometric shape, transition dipole moment along the long axis, and uniformly distributed and well aligned in the liquid crystal.

(Source: Adapted from [21] and [19].)

Fluorescent confocal microscopy

Although conventional light and fluorescence microscopy allow the examination of both living and fixed specimens, and thus the observation of dynamic processes as they actually occur, certain problems do exist with these techniques. One of the main difficulties faced is out-of-focus blur degrading the image. De-focused information often obscures important structures of interest, particularly in thick specimens.

In a conventional microscope set-up, not only is the plane of focus illuminated, but much of the specimen above and below this point is also illuminated at the same time. This results in out-of-focus blur from these areas above and below the plane of interest. This out-of-focus light leads to a reduction in image contrast and a decrease in resolution.

The illumination in a confocal microscope system however, is sequential in nature. The specimen is not uniformly illuminated throughout its depth, the light being focused on a spot on one volume element of the specimen at a time. The dimensions of this spot will vary from one system to another, and are also dependent on the specific illumination wavelength and the way in which the confocal microscope is set up. It is possible, however, to obtain illumination spots as small as 0.25 μm in diameter and 0.5 μm deep.

The design of the confocal system is such that as the beam of illuminating light diverges above and below the plane of focus, volume elements receive less light as one moves away from he focal plane. This results in a reduction of some of the out-of-focus information.

Today, a laser is used to provide the excitation light in order to get very high intensities. The laser light reflects off a dichroic mirror. From there, the laser hits two mirrors which are mounted on motors; these mirrors scan

the laser across the sample. Dye in the sample fluoresces, and the emitted light green gets descanned by the same mirrors that are used to scan the excitation light blue from the laser. The emitted light passes through the dichroic and is focused onto the pinhole. The light that passes through the pinhole is measured by a detector, i. e., a photomultiplier tube.

There never is a complete image of the sample – at any given instant, only one point of the sample is observed. The detector is attached to a computer which builds up the image, one pixel at a time. In practice, this can be done perhaps 3 times a second, for a 512 × 512 pixel image. The limitation is in the scanning mirrors.

In recent years confocal imaging has grown rapidly in popularity as a method for optical sectioning fluorescence microscopy. The technique allows direct visualization of optical sections within thick, fluorescently labelled tissue, and is finding increasing use within the study of living cells and tissues. The main advantage of confocal microscopy over regular light microscopy is that it enables the worker to produce high quality images even in the presence of a large background of out of focus fluorescence flare, produced by fluorescently labelled structures above and below the plane of focus.

One common problem with both conventional fluorescence and confocal microscopy is that fluorophores may bleach when excited. This can limit the maximum time available for image collection, and can be particularly frustrating when collecting a series of images throughout the thickness of a sample. This problem also places limitations on the maximum time for which living samples may be observed, since during photobleaching, toxic by-products are produced.

(Source: Adapted from [61] and [62]. See also [42].)

Multi-Photon Microscopy

Recently, Multi-Photon Fluorescence Microscopy has emerged as a new optical sectioning technique. In this type of microscopy excitation is confined to the optical section being observed, generally by the process of 2-photon absorption. Illuminating light of a wavelength approximately twice that of the absorption peak of the fluorophore being used is employed. So, for example, if a fluorescein isothiocyanate (FITC) labelled sample is being observed, excitation in a 2-photon system can be achieved at approximately 1000 nm (FITC has an absorption peak at around 500 nm). This means that, essentially, excitation of the fluorophore will not be achieved at this wavelength, thus eliminating photobleaching in the bulk of the sample. If, however, a

high-powered pulsed laser is used as the source of illumination, excitation at the point of focus can be achieved. If the laser employed has a peak power of greater than 2 kilowatts, and if this output is delivered to the specimen in pulses of sub-picosecond duration 2-photon events will occur at the focal plane. The very short life of the laser pulses ensure that the mean power levels of the output are only moderate and damage to the specimen does not occur.

During the very brief laser pulse, photon density is sufficiently high at the point of focus for two photons to be absorbed (essentially simultaneously) by the fluorophore. This absorption of two photons of long wavelength is equivalent to the absorption of single photon, with a shorter wavelength, and results in fluorescence excitation.

As well as the reduction in total photobleaching and photodamage, the non-linear optical absorption property of two-photon excitation that limits the fluorochrome excitation to the point of focus, has several other advantages over confocal microscopy. With two-photon microscopy, deeper optical sections can be obtained within a specimen. A detection pinhole is no longer necessary, thus not limiting the number of photons being detected. However, a pinhole may still be used to slightly improve the resolution of two-photon excitation. In addition, the use of UV fluorophores is no longer limited to UV corrected objectives. Two-photon excitation utilizes visible wavelengths to excite UV fluorescent stains and indicators, therefore the objectives do not have to pass UV light.

(Source: Adapted from [61].)

Electron Microscopy

Electron Microscopes are scientific instruments that use a beam of highly energetic electrons to examine objects on a very fine scale. This examination can yield the following information:

- Topography: The surface features of an object and its texture.
- Morphology: The shape and size of the particles making up the object.
- Composition: The elements and compounds that the object is composed of and the relative amounts of them.
- Crystallographic Information: Arrangement of the atoms in the object.

Electron Microscopes were developed due to the limitations of Light Microscopes which are limited by the physics of light to 500 × or 1000 × magnification and a resolution of 0.2 micrometers. In the early 1930's this theoretical limit had been reached and there was a scientific desire to see the fine details of the interior structures of organic cells (nucleus, mitochondria, etc.). This required 10000 × plus magnification which was just not possible using Light Microscopes.

The Transmission Electron Microscope (TEM) was the first type of Electron Microscope to be developed and is patterned exactly on the Light Transmission Microscope except that a focused beam of electrons is used instead of light to "see through" the specimen. It was developed by Max Knoll and Ernst Ruska in Germany in 1931.

The first Scanning Electron Microscope (SEM) debuted in 1942 with the first commercial instruments around 1965. Its late development was due to the electronics involved in "scanning" the beam of electrons across the sample.

Electron Microscopes function exactly as their optical counterparts except that they use a focused beam of electrons instead of light to image the specimen and gain information as to its structure and composition. First a stream of electrons is formed by the electron source and accelerated toward the specimen using a positive electrical potential. Then this stream is confined and focused using metal apertures and magnetic lenses into a thin, focused, monochromatic beam. This beam is focused onto the sample using a magnetic lens. Interactions occur inside the irradiated sample, affecting the electron beam. These interactions and effects are detected and transformed into an image These steps are carried out in all Electron Microscopes regardless of type.

(Source: Adapted from [9].)

Auger Imaging

In Scanning Auger Microscopy (SAM), a high energy (3 – 10 keV) primary electron beam is focused on the sample. This bombardment results in the emission of secondary, backscattered and Auger electrons that can be detected and analyzed. The secondary and the backscattered electrons are used for imaging purposes similar to that in a scanning electron microscope (SEM). The Auger electrons are emitted at discrete energies, that are characteristic of the elements present on the sample surface. All elements in the periodic table, except hydrogen and helium, can be detected, and the depth of analysis is in the range of 3 – 5 nm. As the electron beams can be focused to a

very small probe size, SAM has excellent spatial resolution (0.5 mm).

Using present instrumentation, microanalysis of a volume 100 × 100 × 2 nm can be obtained; this is significantly smaller than the analytical volume excited in SEM/EDX analysis. This spatial resolution is important for the analysis of integrated electronic circuits, as well as small phases in metallurgical and corrosion studies.

The scanning Auger microprobe has an exciting electron beam with a diameter of less than 500 nm. The cylindrical mirror energy analyzer is coaxial with the electron beam. This geometry is considered to be crucial for accurate representation of Auger intensity distributions on irregular surfaces. The specimen is held on a stage capable of four independent motions. Entry and exit into the high vacuum chamber is by a specimen exchange air lock.

Ion bombardment for depth profiling is available in the SAM. The angle of incidence of the ion beam can be changed easily by altering the sample tilt angle. An interfaced microcomputer controls the electron and ion beams. In this way, depth profiles can be run automatically and maps and line scans of Auger electron distributions can be generated.

Two interchangeable stages are available. One of these is equipped with a resistance heater for heating samples in-situ to temperatures of 600° Celsius.

When comparing the surface sensitive techniques of Auger and X-ray photoelectron spectroscopy (XPS), it should be emphasized that both have certain advantages and disadvantages. The Auger spot size is much smaller than the XPS and has the capability of identifying fine features on the surface. The XPS has the capability of determining surface chemical structure and bonding through the use of chemical shifts. Although Auger lines also exhibit chemical shifts, these are not generally as large or as well-documented as those obtained by XPS. Also, X-radiation used in XPS imparts less damage to the sample surface than does the electron beam used in SAM. As mentioned above, the spatial analysis and imaging capabilities of the scanning Auger microprobe make it a very useful and complementary technique to XPS.

Several modes of analysis can be performed in the SAM. These modes are Survey Scan, High-Resolution, Depth Profile, Imaging, Mapping, and Point Analysis. Survey scans of the entire range of Auger electron energies, carried out by detecting and counting the number of Auger electrons, could reveal the presence of contaminants on the sample surface. By taking into account the sensitivity factors of the elements detected, quantification is possible. This is useful in identifying the unknown elements and estimating their concentration on the surface. With argon ion bombardment, the surface layers can be removed gradually, and analysis carried out on new layers exposed

after each sputtering cycle. This is known as depth profiling, and it provides the relative concentrations of elements of interest as a function of depth. Finally, the Auger elemental maps display the presence and the distribution of elements of interest within the area analyzed.

The Auger instrument is capable of handling specimens ranging in size from submillimeter (e.g., integrated circuits) to 35 × 25 × 5 mm. A typical analysis of an unknown surface region (survey scan on the as received surface and after three different sputter intervals) would take 1 - d- 2 hours.

(Source: Adapted from [53].)

Scanning Tunnelling Microscopy STM

Optical microscopes reach their limit at a resolution of approximately 250 nm. This corresponds roughly to half the wavelength of visible light. With an electron microscope, this limit can be overcome, since electrons have shorter wavelengths than light. In certain materials, a modern electron microscope is even able to visualize atomic structures of subnanometer dimensions. Under certain conditions it is also possible to depict surfaces at the atomic scale by employing methods that use X-rays or electron diffraction.

However, none of these methods is as simple and as non-invasive as scanning tunnelling microscopy. The scanning tunnelling microscope does not "see" the atoms, but "feels" them. An ultra-fine tip scans a surface at a constant interval of just a few atomic widths. The control of the distance between the tip and the surface is constantly checked by means of the so-called tunnelling current. This flows between the tip and the target when a voltage is applied between them. The tunnelling current can only be explained by means of quantum mechanics.

If the distance changes by a tenth of a nanometer, the tunnelling current typically changes by a factor of ten. This phenomenon can be employed to control piezoelectric crystals via an electronic feedback loop. A quartz crystal has piezoelectric properties. For example, if a varying voltage is applied, the length of the crystal changes. Thus it is possible to stabilize the distance between the scanning tip and the surface by a series of tiny movements to an accuracy of about one hundredth of a nanometer. This makes it possible to investigate spatial surface structures in the order of atomic dimensions with a precision never achieved before. By means of appropriate computer programs it is also possible to obtain an optical image of the surface from the scan data of the scanning tunnelling microscope.

When the tip approaches the surface of a specimen up to a distance of few atomic widths and a voltage is applied, a so-called tunnelling current flows between the tip and the specimen. While the surface is being scanned, the distance of the tip from the specimen is kept constant via a feedback loop by keeping the tunnelling current constant. The movement of the tip produces the elevation profile of atoms lined up in series. By shifting the tip sideways a field of scanned lines is produced. Computer processing is used to turn this into a three-dimensional image of the surface.

On 18th March 1981, it was at last possible to prove that the tunnelling current depends on the distance between an extremely fine tip and the surface. A short time afterwards, atomic gradations on a surface could be shown for the first time. Finally, in 1985, the capacity of the scanning tunnelling microscope for detailed atomic resolution was recognized unanimously. One year later, Gerd Binning and Heinrich Rohrer of IBM's Zurich Research Laboratory were awarded the Nobel Prize in physics for their pioneering work.

Another great advantage of the scanning tunnelling microscope is its wide field of applications. It can be used in a vacuum, in the natural environment in air and even in liquids. This characteristic has led to a variety of applications in different fields such as metallurgy, electrochemistry and molecular biology. It allows engineers to obtain an insight into the miniaturization of electronic components, biologists to investigate the basic components of life under almost natural conditions, and it also allows chemists to gain a better understanding of batteries, by directly observing chemical surface reactions in an electrolytic solution at a molecular level.

(Source: Adapted from [5].)

Atomic Force Microscopy AFM

A rich variety of forces can be sensed by atomic force microscopy. In the non-contact mode of distances greater than 10 Å between the tip and the sample surface, Van der Waals, electrostatic, magnetic or capillary forces produce images of topography, whereas in the contact mode, ionic repulsion forces take the leading role. Because its operation does not require a current between the sample surface and the tip, the AFM can move into potential regions inaccessible to the Scanning Tunnelling Microscope or image fragile samples which would be damaged irreparably by the Scanning Tunnelling Microscope tunnelling current. Insulators, organic materials, biological macromolecules, polymers, ceramics and glasses are some of the many materials which can be imaged in different environments, such as liquids, vacuum, and low temperatures.

The basic objective of the operation of the AFM is to measure the forces at the atomic level between a sharp probing tip, which is attached to a cantilever spring, and a sample surface. Images are taken by scanning the sample relative to the probing tip and measuring the deflection of the cantilever as a function of lateral position. Typical spring constants are between 0.001 to 100 N/m and motions from microns to approx. 0.1 Å are measured by the deflection sensor. Typical forces between tip and sample range from 10^{-11} to 10^{-6} N. For comparison the interaction between two covalently bonded atoms is of the order of 10^{-9} N at separations of approx. 1 Å. Therefore, non-destructive imaging is possible with these small forces.

The electromagnetic wavefield in solids and their surfaces plays an equal and complementary role to the electron wavefield, a role emphasized by recent experimental developments. Inverse photoemission observations from Scanning Tunnelling Microscope tips show detailed structure in the visible region of the spectrum having its origin in electromagnetic resonances between tip and the surface. In more conventional inverse photoemission experiments, structure is dominated by surface electronic band structure. The same electromagnetic fields are responsible for forces acting at large distances between an AFM tip and the surface in the non-contact mode. The electromagnetic field comes into its own in nanoscale structures. For atomic scale materials it remains firmly pinned to its origins in the electron charge but, given more space, develops a dynamics of its own, more properly described by Maxwell's equations than by the Schrödinger equation.

The force can be thought to arise from changes in the zero-point energy of the electromagnetic wavefield, which are caused by bringing the tip close to the surface. When no observed surface is present, these waves are singly scattered from the tip and escape to infinity. At the proximity of the observed surface, however, the waves are multiply scattered between the tip and the surface, thus modifying the net field, which implies a change in the field energy, i.e., a force. To calculate it, the reflection coefficient of the tip and of the surface to incident electromagnetic waves is needed.

(Source: Adapted from [11].)

Photoemission and Inverse Photoemission

Angle-Resolved photoemission has shed a great light on band structures, both of metals and nonmetals. The first work is due to W. E. Spicer in 1958. This was followed, in 1983, by inverse photoemission. The former explores occupied states, the latter unoccupied states. Photoemission and inverse photoemission have been used to study bulk bands and surface bands.

In bulk photoemission a photon is absorbed by an electron in an occupied state, which makes a transition to an unoccupied state. The external momenta of those final electrons that reach the dipole barrier give direct information about the initial momenta and energies of the electrons. In this way occupied energy bands of many materials have been directly determined by photoemission.

In inverse photoemission external electrons may penetrate the surface, occupy unoccupied Bloch states and then emit a photon. From such measurements direct information about unoccupied states can be obtained.

(Source: Adapted from [36].)

Raman Spectroscopy

Raman microscopy is based on interaction of laser light with lattice vibrations (phonons). The light can either gain a small extra energy from absorbing a phonon or loose some energy by creating a phonon. The population of phonons depends on the temperature, so the temperature is measured by the ratio between the Raman peaks at the phonon loss and gain sides. Since visible light is used, a spatial resolution below 1 micron can readily be achieved using a conventional microscope, with a temperature resolution of a few degrees. Semiconductors have very short optical penetration depth since visible light has energy above the material electronic bandgap. Hence, Raman scattering provides information from the near surface region.

Raman spectroscopy could be a powerful tool for the diagnosis of arteriosclerosis, the build-up of plaques within arteries that can lead to heart attacks and strokes. Raman can distinguish the spectra of particular molecules in the plaque, enabling doctors to verify, for instance, how much calcium it contains.

(Source: Adapted from [58] and [52].)

Reflectance Spectroscopy

Reflectance spectroscopy is the study of light as a function of wavelength that has been reflected or scattered from a solid, liquid, or gas.

As photons enter a mineral, some are reflected from grain surfaces, some pass through the grain, and some are absorbed. Those photons that are reflected from grain surfaces or refracted through a particle are said to be scattered.

Scattered photons may encounter another grain or be scattered away from the surface so they may be detected and measured.

Photons are absorbed in minerals by several processes. The variety of absorption processes and their wavelength dependence allows us to derive information about the chemistry of a mineral from its reflected light. The human eye is a crude reflectance spectrometer: we can look at a surface and see color. Our eyes and brain are processing the wavelength-dependent scattering of visible-light photons to reveal something about what we are observing, like the red color of hematite or the green color of olivine. A modern spectrometer, however, can measure finer details over a broader wavelength range and with greater precision. Thus, a spectrometer can measure absorptions due to more processes than can be seen with the eye.

When photons enter an absorbing medium, they are absorbed according to Beers Law.

Isolated atoms and ions have discrete energy states. Absorption of photons of a specific wavelength causes a change from one energy state to a higher one. Emission of a photon occurs as a result of a change in an energy state to a lower one. When a photon is absorbed it is usually not remitted at the same wavelength. For example, it can cause heating of the material, resulting in grey-body emission at longer wavelengths.

In a solid, electrons may be shared between individual atoms. The energy level of shared electrons may become smeared over a range of values called "energy bands". However, bound electrons will still have quantized energy states.

The most common electronic process revealed in the spectra of minerals is due to unfilled electron shells of transition elements and iron is the most common transition element in minerals. For all transition elements, unfilled d orbitals have identical energies in an isolated ion, but the energy levels split when the atom is located in a crystal field. This splitting of the orbital energy states enables an electron to be moved from a lower level into a higher one by absorption of a photon having an energy matching the energy difference between the states. The energy levels are determined by the valence state of the atom, its coordination number, and the symmetry of the site it occupies. The levels are also influenced by the type of ligands formed, the extent of distortion of the site, and the value of the metal-ligand interatomic distance. The crystal field varies with crystal structure from mineral to mineral, thus the amount of splitting varies and the same ion produces obviously different absorptions, making specific mineral identification possible from spectroscopy.

The unfilled shells of rare earth ions involve deep-lying electrons which are

well shielded from crystal fields so the energy levels remain largely unchanged. Thus, absorption bands due to rare earth elements are not diagnostic of mineralogy but to the presence of the ions in the mineral.

The bonds in a molecule or crystal lattice are like springs with attached weights: the whole system can vibrate. The frequency of vibration depends on the strength of each spring and their masses. For a molecule with N atoms, there are $3N$-6 normal modes of vibrations called fundamentals. Each vibration can also occur at roughly multiples of the original fundamental frequency. The additional vibrations are called overtones when involving multiples of a single fundamental, and combinations when involving different types of vibrations.

A vibrational absorption will be seen in the infrared spectrum only if the molecule responsible shows a dipole moment.

Reflectance spectroscopy shows a wealth of information about mineralogy. Why, then, is spectroscopy not a more widely used technique? In many cases spectroscopy is overly sensitive to subtle changes in crystal structure or chemistry. This has resulted in confusion in the past. More recently, this sensitivity has been recognized as a powerful means of studying the structure and composition of minerals. Additional problems occur with reflectance spectra due to scattering and will be discussed below.

Because spectroscopy is sensitive to so many processes, the spectra can be very complex and there is still much to learn. However, it is because of this sensitivity that spectroscopy has great potential as a diagnostic tool.

Reflectance spectroscopy can be used without sample preparation, and it is non-destructive. This makes mapping of minerals from aircraft possible, including detailed clay mineralogy.

Scattering is the process that makes reflectance spectroscopy possible: photons enter a surface, are scattered one or more times, and while some are absorbed, others are scattered from the surface so we may see and detect them. Scattering can also be thought of as scrambling information. The information is made more complex, and because scattering is a non-linear process, recovery of quantitative information is difficult.

The amount of light scattered and absorbed by a grain is dependent on grain size. A larger grain has a larger internal path where photons may be absorbed according to Beers Law. It is the reflection from the surfaces and internal imperfections that control scattering. In a smaller grain there are proportionally more surface reflections compared to internal photon path length, or in other words, the surface-to-volume ratio is a function of grain size. As the grain size increases, the reflectance decreases.

Absorptions in a spectrum have two components: continuum and individual features. The continuum is the "background absorption" onto which other absorption features are superimposed. The depth of an absorption is related to the abundance of the absorber and the grain size of the mineral.

Reflectance spectroscopy is a rapidly growing science that can be used to derive significant information about mineralogy and with little or no sample preparation. It may be used in applications when other methods would be too time consuming. For example, imaging spectrometers are already acquiring millions of spatially gridded spectra over an area from which mineralogical maps are being made. It is possible to set up real-time monitoring of processes using spectroscopy, such as monitoring the mineralogy of drill cores at the drilling site. Research is still needed to better understand the subtle changes in absorption features before reflectance spectroscopy will reach its full potential.

(Source: Adapted from [10].)

Optical charge coupled device (optical CCD) camera

Like many technologies, the Charge-Coupled Device (CCD) started out as one kind of creature and wound up as something completely different. Invented in the late 1960's by researches at Bell Labs, it was initially conceived as anew type of computer memory circuit. It soon became apparent that the CCD had many other potential applications, including signal processing and imaging, the latter because of silicon's light sensitivity that responds to wavelengths less than 1.1 μm (the visible spectrum falls between 0.4 – 0.7 μm). The CCD's early promise as a memory element has since disappeared, but its superb ability to detect light has turned the CCD into the premier image sensor technology.

Similarly to integrated circuits, CCDs begin on thin silicon wafers which are precessed with a series of elaborate steps which define the various functions within the circuit. Each wafer contains several identical devices called chips, each capable of yielding a functional device. Selected chips, based on a variety of preliminary screening tests, are then cut from the wafer and packed into a carrier for use in a system.

Image sensing with CCDs is performed using line scanning and area scanning techniques. CCD imaging is based in a three step process:

(1) Exposure which converts light into an electronic charge at concrete sites called pixels;

(2) Charge transfer which moves the packets of charge within the silicon substrate; and,

(3) Charge to voltage conversion and amplification.

An image is acquired when incident light, in the form of photons, fall on the array of pixels. The energy associated with each photon is absorbed by the silicon and causes a reaction to take place. This reaction yields the creation of an electron-hole pair. The number of electrons collected at each pixel is linearly dependent on light level and exposure time and non-linearly dependent on wavelength.

Silicon based CCDs are monochrome in nature. Techniques like color sequential ore three-chip color allow the construction of color images.

In vivo imaging of cells tagged with light-emitting probes is a powerful new technology that enables a wide range of biological studies in small research animals. The use of light-emitting probes, such as firely luciferase or fluorescent proteins, as a reporter of gene expression in living cells is a well-established technique for the study of biological activity. Reporters of gene expression with emission in the red to infrared (¿ 600 nm) are preferred due to the low absorption in tissue at these wavelengths. Such probes have been widely used in vitro where light detection is easily accomplished using standard photomultiplier tubes or inexpensive CCD arrays for imaging applications. Detection of light-emitting probes in vivo within small living animals is also possible due to the semi-transparent nature of mammalian tissue, but improved instrumentation is required, consisting of high sensitivity low-noise detectors, a more advanced imaging system, and sophisticated software tools for interpreting images.

The ability to track light-emitting cells in small laboratory animals such as mice or rats opens up a wide range of applications in pharmaceutical and toxicological research. These include in vivo monitoring of infectious diseases, tumor growth and metastases, transgene expression, compound toxicity, and viral infection or delivery systems for gene therapy. The ability to detect signals in real time in living animals without the traditional need to sacrifice for each data point. This results in higher quality data using fewer animals and ultimately will speed the process of screening compounds leading to more rapid drug discovery.

(Source: Adapted from [16] and [48].)

Digital autoradiography

Autoradiography is a technique that enables scientists to locate the distribution of radioactive compounds in a sample.

Biological autoradiography takes advantage of the fact that animals and plants cannot distinguish between stable and radioactive isotopes of the same elements in their physiological reactions. This enables the path of labelled compounds to be traced in an organism using nuclear emulsions. This technique is invaluable in the studies of drugs, pesticides and hormones. Autoradiography has provided a useful tool in metallurgy and the study of chemically reactive surfaces.

Isotopic labelling is by far the most sensitive and quantitative technique for the measurement of the distribution of molecules of interest in biological tissues or samples. Drug development has very heavily relied on this technique for biodistribution and pharmacokinetic studies, and makes now a growing use of it in the expanding fields of pharmacogenomics and proteomics.

Most beta label autoradiography techniques are based on a two step process in which a storage screen (film, phosphor screen) stores the energy deposited by Beta disintegrations, and is consequently scanned to obtain a twodimensional image. Besides being a blind process, it also is a very inefficient process for low energy Beta labels, and a significant amount of information is lost in this energy accumulation method, such as quantitation accuracy or energy discrimination.

Autoradiography instruments based on particle counting techniques, such as the Beta Imager developed from Georges Charpak gaseous chambers, or the Micro Imager issued from the IPN in Orsay, offer in comparison a number of powerful features : real time acquisition, high sensitivity to 3H labels leading to very high gain in experiment time, and quantitation accuracy.

Recent developments now allow the simultaneous acquisition of dual labelled samples and consequent filtering of the distribution of each label. This technique has proven to be very powerful in genomics, both for microarray gene expression measurements and for in situ hybridization experiments. It is also an additional tool for the pharmacokinetics and biodistribution of two metabolites or of two simultaneously administered molecules.

(Source: Adapted from [31] and [12].)

Acoustic Microscopy

Scanning acoustic microscopy (SAM) uses acoustic impedance to produce high resolution images of a sample's interior structure to detect "difficult-

to-find" defects, such as interfacial separation. This kind of microscope uses focused beams to detect acoustic properties. It naturally has a limit of resolution because of Rayleigh diffraction limits of focused beams.

To overcome this resolution limit, the so-called near-field acoustic microscope is developed. The principle of this kind of microscopes is almost the same: an acoustic wave is produced in a tiny area just in the vicinity of surface (near-field area) through different interaction mechanisms and by detecting this acoustic wave, one can gain the acoustic properties of materials at a high resolution which is not dependent on wavelength. Scanning Electron Acoustic Microscope (SEAM) and Scanning Probe Acoustic Microscope (SPAM), which are developed from the commercial Scanning Electron Microscope (SEM) and Scanning Probe Microscope (SPM), are two typical microscopes of this kind with highest resolution. With these techniques it is possible to obtain the acoustic properties, which are different from topography, at the resolution of micro- and nanometer range.

(Source: Adapted from [44] and [2].)

Photon induced X-ray emissions — PIXE Imaging

Recent advances in technique now permit routine non-destructive analysis and imaging of fluid inclusion in minerals using Proton Induced X-ray Emission (PIXE) and beams of 3 MeV protons. Development is continuing to augment this capability with Proton Induced Gamma-ray Emission (PIGE) for light element detection and 3D fluid mapping using elastic recoils. Using PIGE it is now possible to provide images of light elements such as Na and F.

Energetic protons pass easily through minerals like quartz to depths of more than 80 μm, and excite X-rays from elements within trapped fluid inclusions. The X-rays are counted using a cooled germanium or lithium-drifted silicon detector. By focussing the proton beam to about 1.3 μm in diameter, and raster scanning the beam over each inclusion, individual fluid inclusions can be imaged and analyzed non-destructively.

In order to relate the detected X-rays to concentration of an element in an inclusion, a model has been developed for calculating X-ray yields from specific 3D inclusion geometries, e.g. inclusion size, shape, density, orientation and depth. Now by raster-scanning the beam over a fluid inclusion to provide a uniform dose, and using the 3D model, the method permits the extraction of the concentration of all detected elements in the inclusion. If there has

been no loss from the inclusion, this provides the composition of the original homogeneous trapped fluid.

The high resolution also provides a tool for imaging the internal contents of individual fluid inclusions.

The sensitivity of the Proton Microprobe enables the analysis of individual fluid inclusions down to approx. 5 μm in diameter. In typical inclusions (10-15 μm) ore-related elements (such as Cu, Zn, Au and Pb) can be detected and analyzed down to concentrations of approx. 20 ppm.

Test performed using synthetic fluid inclusions in quartz have shown the method to be accurate to 10-15% for undersaturated solutions (no daughter minerals present). With the presence of daughter minerals, and due to the uncertainties associated with the positions of these minerals, uncertainties typically grow to approx. 30%.

The PIXE technique for the non-destructive analysis of fluid inclusions is in routine use for the analysis of ore-fluids associated with porphyry Cu-Au and VMS deposits worldwide.

(Source: Adapted from [51].)

Secondary Ion Mass Spectroscopy — SIMS Imaging

Time of Flight Secondary Ion Mass Spectrometry (ToF-SIMS) is a surface analytical technique used for obtaining elemental and molecular chemical information about surfaces (static SIMS), as well as, detecting part per billion (ppb) concentrations of impurities in semiconductors and metals (dynamic SIMS). All elements, including hydrogen, are detectable by SIMS.

In ToF-SIMS analysis, the sample is placed in an ultrahigh vacuum environment where primary ions bombard and sputter atoms, molecules, and molecular fragments from the sample surface. The mass of the ejected particles (i.e. secondary ions) are analyzed via time-of-flight mass spectrometry.

In the ToF analyzer, ejected ions are accelerated into the analyzer with a common energy (but different velocities depending on the particle mass). Due to the differences in velocities, smaller ions move through the analyzer more rapidly than the larger ions. The mass of the secondary ions are determined by their travel time through the analyzer. SIMS is a surface sensitive analysis method since only the secondary ions generated in the outermost 10 to 20 Å region of a sample have sufficient energy to overcome the surface binding energy and escape the sample surface for detection and analysis.

The mass spectrum identifies the elemental and ion composition of the uppermost 10 to 20 Å of analyzed surface from positive and negative mass spectra.

The high resolution of the ToF analyzer can distinguish species whose masses differ by only a few millimass units.

During SIMS analysis, the sample surface is slowly sputtered away. Continuous analysis obtains composition information as a function of depth. Depth resolution of a few Å is possible. High sensitivity mass spectra can be recorded or reconstructed at any depth of the depth profile.

The secondary ion mapping measures the lateral distribution of elements and molecules on the surface. To obtain a SIMS map, a highly focused primary ion beam is rastered across the sample surface, and the secondary ions are collected at specific points. Image brightness at each point is a function of the relative concentration of the mapped element or molecule. Lateral resolution is less than 100 nm for elements and 0.5 μm for large molecules.

Typical applications of SIMS are

- Identifying lubricants on magnetic hard discs
- Measuring dopant distributions in semiconductors
- Profiling thickness of insulating films on glasses
- Mapping elemental and molecular patterned surfaces
- Identifying compounds in thin organic films
- Extent of crosslinking in polymers

The sample size cannot exceed 85 mm in any lateral direction, height should not exceed 20 mm and the sample must be compatible with high vacuum environment.

(Source: Adapted from [54].)

Capacitive Probe Imaging

The capacitive sensor used does not work on images, but rather on the physical probe itself. In the chip are 90,000 plates connected to the circuitry that measures capacitance. Over these plates is a dielectric passivation where the probe is placed. The probe itself forms the second plate of the capacitor above each of the 90,000 embedded plates. The ridges and valleys are different distances from the underlying plate, so the capacitance is different. The chip measures this capacitance and converts it into an image.

(Source: Adapted from [20].)

Compound Imaging

Compound imaging uses several ultrasound beams that strike the tissue from different angles. This method significantly reduces speckle noise and improves the contrast of the image.

(Source: Adapted from [32].)

4.2 Molecular Imaging

The term molecular imaging can be broadly defined as the in vivo characterization and measurement of biological processes at the cellular and molecular level. In contradistinction to "classical" diagnostic imaging, it sets forth to probe the molecular abnormalities that are the basis of disease rather than to image the end effects of these molecular alterations. While the underlying biology represents a new arena for many radiologists, concomitant efforts such as development of novel agents, signal amplification strategies, and imaging technologies clearly devotail with prior research efforts of radiology.

Advances in molecular and cell biology techniques, the ability to decode entire genomes, the continuous search for new targets, and the unraveling of the molecular pathways of many diseases have had a marked effect on the way we practice medicine today. Much research attention has been rightfully directed towards understanding the cellular and molecular mechanisms of diseases, but efforts have also been directed toward the development of noninvasive, high-resolution, in vivo imaging technology. Specially, over the past few years, in vivo molecular imaging has been identified by the national Cancer Institute as an extraordinary opportunity for studying diseases noninvasively and, in many cases, quantitavely at a molecular level. Although there still remains a scientific gulf between the basic scientists who discover new genes and their function and the imaging scientists who could transform these discoveries into noninvasive imaging methods, this gap is rapidly closing.

Molecular imaging is a growing research discipline aimed at developing and testing novel tools, reagents, and methods to image specific molecular pathways in vivo, particularly those that are key targets in disease processes. while certain imaging techniques that may be defined as "molecular" were developed decades ago (e. g., imaging with monoclonal antibodies or receptor imaging with nuclear techniques), it is only recently that a host of needed adjunct basic research tools have become routinely available. Some of these tools that can now be build on include molecular cloning, microfabrication,

chip arrays, robots, X-ray crystallography, fast mass spectrometry, and sophisticated computer analysis. These more recently available basic science tools now allow to answer basic biologic questions in vivo and do this in a high-throughput fashion. It is expected that the fruits of today's molecular imaging research will have direct effect on patient care within the next 5–15 years. The current assessment of disease is based on anatomic changes or, more recently in specialized cases, physiologic changes that are a late manifestation of the molecular changes that truly underlie disease. Direct imaging of these molecular changes will directly affect patient care by allowing much earlier detection of disease. Potentially it may be possible to image molecular changes that are currently defined as "predisease states", which would allow intervention at a time when the outcome is most likely to be affected. In addition, by directly imaging the underlying alterations of disease, it will potentially be possible to directly image the effects of therapy.

To image specific molecules in vivo, several key criteria must generally be met:

(a) availability of high-affinity probes with reasonable pharmacodynamics;

(b) the ability of these probes to overcome biologic delivery bassiers (vascular, interstitial, cell membrane);

(c) use of amplification strategies (chemical and biologic), and

(d) availability of sensitive, fast, high-resolution techniques.

In a typical scenario, all four prerequisites must be met for successful in vivo imaging at the molecular level.

Molecular information can be obtained with some but not all of the presently used "high-end" imaging technology. Despite recent advances in imaging technology, further improvements in current imaging modalities and exploration of new modalities are still at the center stage of molecular imaging research. For example, the development of optical imaging technology (including diffuse optical tomography, phase-array detection, photon counting, near-infrared fluorescence imaging), high-spatial-resolution MNR and nuclear imaging techniques (e. g., micro-MR, micro-PET) play an important role in the field. Improving spatial resolution now allows imaging of mouse models of human disease, and imaging findings and concepts can thus be directly translated into a clinical context.

The following subsections will give some applications of molecular imaging.

(Source: Adapted from [63].)

The Human Genome Project

The entire body of genetic information required to form and sustain life is contained in the DNA molecule. Most living entities, including viruses and bacteria, fungi, plants, animals, and humans use this universal instruction language. The Human Genome Project is an effort to decipher the specific sequences of all human genes. The impact of the Human Genome Project is expected to revolutionize medical practice and biologic research well into the 21st century. For the first time, we will have the specific codes that govern life.

The role of imaging in these efforts and its applications in gene therapy have been largely unexplored until recently. Yet, the influence of the Human Genome Project on diagnostic imaging will be widespread, both in experimental research and in clinical applications.

(Source: Adapted from [63].)

Clinical Gene Therapy

Gene therapy has been heralded as a potential revolution in medicine because, for the first time, a therapy is aimed at correcting the cause of disease rather than treating phenotypic symptoms. Gene therapy also represents a platform technology that can be applied to a wide range of diseases and targets. In an ex vivo gene therapy approach, cells are removed from the patient, treated, and then introduces again. In the in vivo approach, the gene (in the form of a vector) is directly administered to the patient. It is clear that recent improvements in the manufacturing process for adenoviruses, retroviruses and others have generated the need for in vivo imaging of both vector and target distribution, as well as in vivo imaging of the resultant transgene expression.

(Source: Adapted from [63].)

Gene Delivery

For gene therapy vectors to work, they must be delivered efficiently to the intended target. Although systemic intravenous administration of a vector is occasionally performed, more common application strategies include stereotactic and focal image-guided applications. Localization and quantification

of viral accumulation in vivo will enable detailed analysis of viral and organ interactions critical for advancing the therapeutic use of ever increasing number of developed vector systems.

(Source: Adapted from [63].)

Exogenous marker genes

Given the availability of animal models and vectors, ease of molecular cloning techniques, and relevance to clinical applications, the field of gene expression imaging has recently boomed. A moderate number of imaging marker genes (i. e., gene products that can be detected with different imaging modalities) have been described. The choice of a specific system depends on the imaging requirements (single or repeated), spatial resolution and several other factors.

(Source: Adapted from [63].)

Angiogenesis, Apoptosis and other imaging fields

Angiogenesis: Angiogenesis occurs physiologically during embryonic development, the female reproductive cycle, wound healing, and hair growth and describes the growth and remodeling of primitive vascular network into a complex one. Up regulation and/or down regulation of the angiogenic process are central to approximately 30 diseases, including cancer, cardiovascular disease, immunologic disease, and diabetes. Understanding the molecular basis of angiogenic control is slowly emerging.

Despite the continued development of angiogenesis inhibitors, a notable problem has been the need for surrogate markers to monitor drug effects. It has become clear that imaging of physiologic or molecular markers may be the most fruitful avenue yet to aid in the evaluation of response. A number of imaging techniques, in particular fMRI and nuclear techniques that rely on first-pass or equilibrium contrast material enhancement, are available for studying microvascular circulation.

The attractiveness of imaging and treating the same molecular target is obvious and may allow the earliest target assessment possible. More experimentation is needed to validate new imaging strategies and translate them into the clinic. Once validated, the new approaches can be used to answer questions such as the following: Is administration of a single antiangiogenic agent sufficient? Do we deliver sufficient inhibitors to their targets? Do the inhibitors truly affect their target in vivo? While these questions are highly

relevant in tumor treatments, there are many other applications that will also ultimately benefit from this type of imaging research.

Apoptosis: Apoptosis is a physiologic form of programmed cell death critical for organ development, tissue homeostasis, and removal of defective cells in vivo without causing a concomitant inflammatory response. Apoptosis depends on the recognition of multiple extracellular and intracellular signals, integration and amplification of signals, and activation of a family of effector proteases called caspases. Defects in control of apoptotoc pathways may contribute to a variety of diseases, such as AIDS, ischemia, stroke and others.

For the development and evaluation of treatment drugs it is critical to assess drug efficacy at the molecular level either by imaging caspase activity directly or by imaging other downstream effects.

Other topics: Through the specific examples of angiogenesis and apoptosis it becomes apparent how molecular imaging techniques will devotail with other developments. They have and will have great impact on topics like signal transduction, stem cell biology and cell-based therapy, antisense technology, cell cycle control, immunotherapy, drug resistance and many others.

(Source: Adapted from [63].)

4.3 Imaging of small animals

While the examples in the previous subsection highlight recent advances in molecular imaging, there has also been a recent interest in phenotypic imaging. Molecular geneticists looking for ways to model human diseases and companies testing new drugs are creating an unprecedented demand for transgenic or knockout mice (which have one or more of their 80,000 genes disabled). This demand is due for one, to the ability to fine tune genetic alterations of model mice. The low maintenance cost, fecundity, and genetic similarity of mice to humans are other factors.

With regard to imaging of small animals, the size of the imaged object, the total volume that must be evaluated, the spatial resolution (voxel size) necessary for meaningful in vivo data, and the total time that may be dedicated to acquiring an image are quite different for a 20 g mouse than for a 70 kg person. Instrumentation requirements reflect these differences and can be exploited to maximize the information that can be obtained from small-animal models of disease states.

(Source: Adapted from [63]. See also [64].)

4.4 Imaging techniques

Classical image manipulation methods as described in subsection 3.2 are developed and used for biological imaging with similar intentions to those in medical imaging. However, it is important to find an adequate noise model for each bio-imaging modality. Compression does not yet play an important role in biological imaging, since there is no sensitive patient data to be transmitted in real time. Approximation and interpolation problems in biological imaging leads to similar models and solutions as in medical imaging.

There is not much literature yet on soft computing in biological image processing and analysis. With respect to the impressive results in genomic analysis with statistical methods and neural networks, soft computing could be an interesting new modelling approach.

In these days, real time processing is not important in biological imaging, yet. But this could change quickly with the continuous understanding of biochemical mechanisms in cells. As these discoveries can be used for the treatment of human diseases, real time imaging of biochemical processes in the body will apparently be relevant.

The main topic today is modelling and visualization of biological and chemical processes on the nano-scale. These questions we will consider in the following subsection.

4.5 Modelling and Visualization

Protein structure and genomic sequences

For research in problems in protein folding and genomics computational and mathematical techniques play an important role. Mathematical models of virus shell assembly, for protein motive recognition, for beta-helix fold in protein sequence data and others help to analyze these structures and processes and lead to visualization models.

Proteins are organic molecules which have interesting geometric features, but they don't easily fit the type of structures favored by mathematicians. They don't quite fit on lattices, for example, although lattice models have been proposed. They do form discrete models of helices and can be analyzed using a discrete version of the classical Frenet frame plus some techniques from the theory of space groups. Most of the information known about the structure of proteins has been obtained through the science of x-ray crystallography. NMR is a tool which is being used to find the structure of proteins which cannot

by crystallized in their native environment. In solid state NMR, orientational constraints are used to find structures.

(Source: Adapted from [3] and [47]. See also [18].)

Cell Activity

To analyze cell activity such as bursting oscillations in excitable nerve cells, the electrical activity of pancreatic cells or synaptic transmitter release mathematical models involving nonlinear ordinary differential equations, dynamical system theory, bifurcation analysis, and computer simulation.

(Source: See [4].)

4.6 Hyperspectral imaging

Hyperspectral imaging is often referred to as multispectral imaging, ultraspectral imaging or just spectral imaging.

Spectral image cubes are analogous to a stack of pictures of an object, a sample, or a scene, where each image is acquired at a narrow spectral band. Each pixel in the image cube, therefore, represents the spectrum of the scene at that point. The nature of imagery data is typically multidimensional, spanning three spatial, one spectral, and one temporal dimensions. Each point in this multidimensional space is described by the intensity of the radiance which is emitted, reflected, or a combination of both depending on the phenomenology under investigation. Since detector arrays in image capture devices are two dimensional at most, they can only capture two dimensions of the data at one time, and another dimension displaced in time. In mobile applications (e.g., air- or space-borne, or a moving web or a conveyor belt) a sensor builds an image cube (x-y-l dimensions) in a pushbroom fashion, by capturing typically one spatial and the spectral dimensions in each camera frame, while the second spatial dimension is captured displaced in time. In stationary applications (e.g., a sample under a microscope) it is possible to capture two spatial dimensions in each camera frame, while the spectrum axis is displaced in time. For rapid time-varying events, other techniques allow the construction of sensors that capture the three x-y-l dimensions in a single camera frame. Capturing image cubes in stationary applications can be accomplished by swapping narrow band pass filters in front of the camera lens, but a more elegant, convenient, and versatile solution is afforded by the use of electronically tunable filters (ETF).

Agriculture: Agriculture is moving today towards "precision farming" techniques in which crop management is performed on a local basis, rather than field wide. This requires the ability to detect and identify spatial distribution of crop stress in monoculture plots. Once identified, using multi- or hyperspectral imaging techniques, local treatment may be applied (e.g., irrigation, fertilization, insecticide or herbicide). The approach has broad implications on production costs and the environment management. Present efforts are directed towards remote sensing applications. These Images allow careful study of the various spatial features of objects that can be then used to support remote sensing of similar objects. Such detailed spatial data distribution are lost when conducting field measurements with a non imaging (e.g., fiber optics) spectrometer, that integrates a significant extent of the scene.

Medical spectral imaging: Biological tissue exhibits unique spectra in induced or auto fluorescence, and in transmission or reflection. Spectral differences in tissue pathology may be spatially resolved using imaging spectroscopy. Hyperspectral imaging can work with a variety of spectroscopic techniques: fluorescence, reflectance, light-scattering, and Raman. Some techniques give structural information. Light-scattering spectroscopy, for instance, can measure the distribution of the sizes of cell nuclei or the density of chromatin in genetic material. Fluorescence can take the advantage of naturally occurring fluorescence or be enhanced with fluorescing dyes that bind to specific materials in the body. Most such dyes are well understood and well accepted in medicine.

Other applications of imaging spectroscopy include human color vision, color perception and image understanding, forensics, pharmaceutical, manufacturing, archaeology and art, inspection, military target detection based on spectral and polarization properties and true color night vision.

(Source: Adapted from [24] and [52]. See also [30], [22], and [23].)

4.7 Synthetic Aperture Radar

Environmental monitoring, earth-resource mapping, and military systems require broad-area imaging at high resolutions. Many times the imagery must be acquired in inclement weather or during night as well as day. Synthetic Aperture Radar (SAR) provides such a capability. SAR systems take advantage of the long-range propagation characteristics of radar signals and the complex information processing capability of modern digital electronics to provide high resolution imagery. Synthetic aperture radar complements

photographic and other optical imaging capabilities because of the minimum constraints on time-of-day and atmospheric conditions and because of the unique responses of terrain and cultural targets to radar frequencies.

Synthetic aperture radar technology has provided terrain structural information to geologists for mineral exploration, oil spill boundaries on water to environmentalists, sea state and ice hazard maps to navigators, and reconnaissance and targeting information to military operations. There are many other applications or potential applications. Some of these, particularly civilian, have not yet been adequately explored because lower cost electronics are just beginning to make SAR technology economical for smaller scale uses.

Consider an airborne SAR imaging perpendicular to the aircraft velocity. Typically, SARs produce a two-dimensional (2-D) image. One dimension in the image is called range (or cross track) and is a measure of the "line-of-sight" distance from the radar to the target. Range measurement and resolution are achieved in synthetic aperture radar in the same manner as most other radars: Range is determined by precisely measuring the time from transmission of a pulse to receiving the echo from a target and, in the simplest SAR, range resolution is determined by the transmitted pulse width, i.e., narrow pulses yield fine range resolution.

The other dimension is called azimuth (or along track) and is perpendicular to range. It is the ability of SAR to produce relatively fine azimuth resolution that differentiates it from other radars. To obtain fine azimuth resolution, a physically large antenna is needed to focus the transmitted and received energy into a sharp beam. The sharpness of the beam defines the azimuth resolution. Similarly, optical systems, such as telescopes, require large apertures (mirrors or lenses which are analogous to the radar antenna) to obtain fine imaging resolution. Since SARs are much lower in frequency than optical systems, even moderate SAR resolutions require an antenna physically larger than can be practically carried by an airborne platform: antenna lengths several hundred meters long are often required. However, an airborne radar could collect data while flying this distance and then process the data as if it came from a physically long antenna. The distance the aircraft flies in synthesizing the antenna is known as the synthetic aperture. A narrow synthetic beamwidth results from the relatively long synthetic aperture, which yields finer resolution than is possible from a smaller physical antenna.

Achieving fine azimuth resolution may also be described from a doppler processing viewpoint. A target's position along the flight path determines the doppler frequency of its echoes: Targets ahead of the aircraft produce a positive doppler offset; targets behind the aircraft produce a negative offset. As the aircraft flies a distance (the synthetic aperture), echoes are resolved into

a number of doppler frequencies. The target's doppler frequency determines its azimuth position.

Transmitting short pulses to provide range resolution is generally not practical. Typically, longer pulses with wide-bandwidth modulation are transmitted which complicates the range processing but decreases the peak power requirements on the transmitter. For even moderate azimuth resolutions, a target's range to each location on the synthetic aperture changes along the synthetic aperture. The energy reflected from the target must be "mathematically focused" to compensate for the range dependence across the aperture prior to image formation. Additionally, for fine-resolution systems, the range and azimuth processing is coupled (dependent on each other) which also greatly increases the computational processing.

(Source: Adapted from [41].)

Bibliography

[1] Moir Fringe Contouring. 3D-MATIC Research Laboratory, University of Glasgow.
www.faraday.gla.ac.uk/moire.htm.

[2] A small introduction into Scanning Acoustic Microcscopy. University of Wuppertal, Faculty of Electrical Engineering and Information Technology, Department of Electronics.
www.electronics.uni-wuppertal.de/english-/forschung/intro/snam_intro.html.

[3] Bonnie Berger. Dept. of Mathematics, Massachusetts Institute of Technology, Cambridge, MA.
theory.lcs.mit.edu/~bab/.

[4] Richard Bertram. Institute of Molecular Biophysics, Florida State University.
www.sb.fsu.edu/~bertram/.

[5] Gerd Binnig and Heinrich Rohrer. The scanning tunnelling microscope. Deutsches Museum, München.
www.deutsches-museum.de/ausstell/meister/e_rtm.htm.

[6] BISC - Special Interest Group: Medical Imaging. Information Systems Laboratory, Himeji Institute of Technology.
wwwj3.comp.eng.himeji-tech.ac.jp/med/.

[7] Kevin Boone. Electrical Impedance Tomography. Department of Clinical Neurophysiology run by David Holder at the Middlesex Hospital, London.
www.eit.org.uk.

[8] Thomas B. Budinger. Medical Imaging Techniques. Berkeley Lab.
imasun.lbl.gov/~budinger/medhome.html.

[9] Electron Microscopy. Center for Materials Research and Analysis, University of Nebraska-Lincoln.
www.unl.edu/CMRAcfem/em.htm.

[10] Roger N. Clark. About Reflectence Spectroscopy. U.S. Geological Survey, a bureau of the U.S. Department of the Interior. Derived from: Clark, R. N., 1995, Reflectance Spectra, AGU Handbook of Physical Constants, p. 178–188, 1998.
speclab.cr.usgs.gov/aboutrefl.html.

[11] The Atomic Force Microscope (AFM). Condensed Matter Theory Group, Blackett Laboratory, Imperial College, London.
www.sst.ph.ic.ac.uk/photonics/intro/AFM.html.

[12] Philippe Coulon, Serge Matrejean, and Marie Meynadier. New developments in digital autoradiography. Biospace Mesures, Paris.
www.medim.net/Toulouse/Meynadier.htm.

[13] Michael W. Davidson and Mortimer Abramowitz. Optical Microscopy. Technical report, Olympus.
http://www.olympusmicro.com/primer/microscopy.pdf.

[14] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, 1973.

[15] Markus Dürrenberger and Rosmarie Sütterlin. Looking inside cells and tissues by optical sectioning with a confocal laser scanning microscope. M. E. Müller-Institute for Microscopy, Biozentrum, Basel, Switzerland.
www.mih.unibas.ch/Booklet/Booklet96/Chapter1/Chapter1.html.

[16] Charge-Coupled Device (CCD) Image Sensors. Technical report, Eastman Kodak Company — Image Sensor Solutions, 2001.

[17] Nick Efford. Medical Imaging Resources: Content Listing. Centre of Medical Imaging Research, University of Leeds.
www.comp.leeds.ac.uk/comir/resources/links_c.html.

[18] Phil Bradley et al. Betawrap: Successful prediction of parallel β-helices from primary sequence reveals an association with many microbial pathogens. *PNAS*, 98(26):14819–14824, 2001.
www.pnas.org/cgi/doi/10.1073/pnas.251267298.

[19] FCPM versus PM. Laboratory of Prof. O. D. Lavrentovich, LCI, Kent State University.
www.lci.kent.edu/Lavrentovich/FCPMweb_site/fcpm_pm.html.

[20] Fingerprint Bioscan Terminal. Legiant, Austin, Texas, 2000. `www.legiant.com/Products/bioscan.htm`.

[21] Fluorescence Confocal Polarizing Microscopy. Laboratory of Prof. O. D. Lavrentovich, LCI, Kent State University. `www.lci.kent.edu/Lavrentovich/FCPMweb_site/fcpm_mth.html`.

[22] Nahmum Gat. Imaging spectroscopy: links to hyperspectral & multispectral (sensor, algorithms, & data processing applications) resources on the web. OKSI Opto-Knowledge Systems, Inc., Torrence, CA, 2000. `www.techexpo.com/WWW/opto-knowledge/IS_resources.html`.

[23] Nahum Gat. Directions in environmental spectroscopy. *Spectroscopy Showcase*, March 1999. `www.techexpo.com/WWW/opto-knowledge/March99.pdf`.

[24] Nahum Gat. Imaging spectroscopy using tunable filters: A review. In *Proc. SPIE, Wavelet Applications VII*, volume 4056, pages 50–64, April 2000. `www.techexpo.com/WWW/opto-knowledge/E-tunable-filters.pdf`.

[25] Allen Gersho and Robert M. Gray. *Vector Quantization and Signal Compression.* Kluwer Academic Publishers, 1991.

[26] Maryellen L. Giger, Nico Karssemeijer, and Samuel G. Armato, III. Guest Editorial. Computer-Aided Diagnosis in Medical Imaging. *IEEE Transactions on Medical Imaging*, 20(12):1205–1208, December 2001.

[27] S. H. D. Haddock, C. M. McDougall, and J. F. Case. The Bioluminescence Web Page. Biological Sciences at UC Santa Barbara, 2000. `lifesci.ucsb.edu/~biolum/`.

[28] HBOI - More about Bioluminescence. Harbor Branch Oceanographic Institution, 2000. `www.biolum.org/`.

[29] Dennis M. Healy, Jr. and John B. Weaver. Adapted Wavelet Techniques for Encoding Magnetic Resonance Images. In *Wavelets in Medicine and Biology.* CRC Press, 1996.

[30] Hyperspectrum News Letter. OKSI and TechExpo. `www.techexpo.com/WWW/opto-knowledge/hyperspectrum/`.

[31] Frequently asked questions about autoradiography. Ilford Products. `www.ilford.com/html/us_english/prod_html/nuclear/faq1.html`.

[32] Imaging Technologies Being Developed for Detection of Breast Cancer. National Academy of Sciences, Washington, DC, 2002. www4.nas.edu/onpi/webextra.nsf/44bf87db309563a0852566f2006d63bb/1301d1 641f585256a80007160f3?OpenDocument.

[33] INSIGHT The Swissray Magazine, November 2001.

[34] Internet Resources of Computer Aided Surgery. homepage2.nifty.com/cas/.

[35] HyoungSeop Kim. KIT T&R: HyoungSeop KIM. Faculty of Engineering, Kyushu Institute of Technology. www.kyutech.ac.jp/teachres/mce/e-kim-hyoungseop.html.

[36] W. Kohn. An essay on condensed matter physics in the twentieth century. *Reviews of Modern Physics*, 71(2):59–76, 1999. Centenary.

[37] Kieran Maher. Digital Radiography. Division of Medical Radiations, School of Medical Sciences, RMIT University, 2002. www.bh.rmit.edu.au/mrs/DigitalRadiography/.

[38] Maurits Malfait. Using Wavelets to Suppress Noise in Biomedical Images. In *Wavelets in Medicine and Biology.* CRC Press, 1996.

[39] Stphane Mallat. *A wavelet tour of signal processing.* Academic Press, San Diego, 1997.

[40] Thermography. MetroWest Medical Center. www.mwmc.com/Apps/Library/.

[41] Brain Mileshosky. What is Synthetic Aperture Radar? Sandia National Laboratories, California. www.sandia.gov/RADAR/whatis.html.

[42] Marvin Minsky. Memoir on Inventing the Confocal Scanning Microscope. *Scanning*, 10:128–138, 1988.

[43] The Nation's Investment in Cancer Research. Report, The National Cancer Institute, National Institutes of Health, U.S.A., 2001.

[44] Non-Destructive Evaluation: Scanning Acoustic Microscopy (SAM). CALCE and the University of Maryland, 2001. www.calce.umd.edu/general/Facilities/sam.htm.

[45] Ronald Nutt. The history of positron emission tomography. *Molecular Imaging and Biology*, 4(1):11–26, 2002.

[46] Yusuf Ozturk. Block Motion Estimation and Block Combination. `ozturk.sdsu.edu/multimedia/me.pdf`.

[47] Jack Quine. Homepage. Dept. of Mathematics, Florida State University. `web.math.fsu.edu/~quine/`.

[48] B. W. Rice, M. D. Cable, and M. B. Nelson. In vivo imaging of light-emitting probes. Preprint.

[49] Richard A. Robb. Guest Editorial. The Biomedical Imaging Resource At Mayo Clinic. *IEEE Transactions on Medical Imaging*, 20(9):854–867, September 2001.

[50] Cortical Imaging. Roper Scientific. `www.photomet.com/library_app-brief_cort.shtml`.

[51] Chris Ryan. Fluid Inclusion Microanalysis using PIXE. CSIRO Exploration and Mining, North Ryde, Australia, 2001. `www.syd.dem.csiro.au/research/hydrothermal/chris/FI_PIXE.html`.

[52] Neil Savage. Hyperspectral imaging. *OPN Trends*, pages 5–7, July 2001.

[53] Scanning Auger Microscopy (SAM). Surface Science Western, Western Science Center, The University of Western Ontario, London, Ontario, Canada. `www.uwo.ca/ssw/services/auger.html`.

[54] Secondary Ion Mass Spectroscopy. Handbook of Analyical Methods, Materials Evaluation and Engineering, Inc., Plymouth, MN, 2000. `www.mee-inc.com/sims.html`.

[55] Stefanie Seltmann. Innenansichten. *einblick — Zeitschrift des Deutschen Krebsforschungszentrums*, 2:2–5, 2002.

[56] Michael Shling, Muthuvel Arigovindan, Patrick Hunziker, and Michael Unser. Motion Analysis of Echocardiograms using a local-affine, spatio-temporal model. In *Proceesings of the First 2002 IEEE International Symposium on Biomedical Imaging: Macro to Nano*, 2002. In Press.

[57] Steve Smith. Brief Introduction to FMRI. Oxford Centre for Functional Magnetic Resonance Imaging of the Brain. `www.fmrib.ox.ac.uk/fmri_intro/`.

[58] Anna Swan, Mark Lande, Selim Ünlü, and Bennett Goldberg. High Resolution Raman Microscopy. Ultra, Boston University, 2000. ultra.bu.edu/projpages/tram/.

[59] Biomedical Imaging. The Whitaker Foundation. www.whitaker.org/94_annual_report/over.html.

[60] Thomas R. Gregg. Use of Functional Magnetic Resonance Imaging to Investigate Brain Function. Neurosience on the internet by Neil A. Busis, M.D. www.neuroguide.com/gregg.html.

[61] Tony Wade. Confocal and Multi-Photon Microscopy. School of Biological Sciences, The University of Manchester. www.biomed2.man.ac.uk/ireland/.

[62] Eric R. Weeks. How does a confocal microscope work? Department of Physics, Emory University. glinda.lrsm.upenn.edu/~weeks/confocal/.

[63] Ralph Weissleder and Umar Mahmood. Molecular Imaging. *Radiology*, 219:316–333, 2001.

[64] Thomas Willke. Zu Risiken und Nebenwirkungen: Fragen Sie Ihren Bioinformatiker. *bild der wissenschaft*, 2:30–33, 2002.

[65] Lotfi A. Zadeh. What is BISC? www.cs.berkeley.edu/projects/Bisc/bisc.memo.html.

Bildgebende Verfahren der medizinischen Diagnostik

Gerhard Winkler*

Zusammenfassung

Dieser Beitrag ist als knappe Einführung in bildgebende Verfahren gedacht. Diese sind seit geraumer Zeit in der medizinischen Diagnostik und Forschung etabliert und gewinnen immer noch an Bedeutung. In der Zukunft werden sie auch in anderen Bereichen der Bio- und Lebenswissenschaften eine wichtige Rolle spielen. So finden die meisten Eingang in das sich rapide entwickelnde Feld des Molecular Imaging.
Es ergeben sich Überschneidungen mit dem Essay von B. Forster und R. Lasser, [2]. Andererseits stellt er auch eine Einleitung zu dem Artikel über die Radontransformation [12] von G. Winkler in diesem Band dar.
Es wurden keine medizinischen Abbildungen aufgenommen, da hierzu exzellente Literatur in Fülle existiert. In Fußnoten wird auf Illustrationen in Standardwerken hingewiesen.

*Institute of Biomathematics and Biometry
GSF - National Research Center for Environment and Health,
Postfach 1129, D-85758 Oberschleißheim, Germany,
gwinkler@gsf.de, http://ibb.gsf.de

Inhaltsverzeichnis

Kapitel 1

Vorbemerkungen

In dieser Einleitung werden einige wichtige Begriffe geklärt, die wichtigsten Anwendungsbereiche der Bild- und Signalanalyse im medizinischen Bereich erwähnt und die Prinzipien einiger bildgebender Verfahren skizziert. Weder Vollständigkeit noch mathematische oder physikalische Präzision werden beansprucht. Wir beschränken uns auf die 'klassischen' Methoden, die im Prinzip - modulo wesentlich verbesserter Gerätetechnik - immer noch das Arsenal der (klinischen) Technologie umfassen. Dramatisch geändert hat sich allerdings der Grad der Anwendung der modernen bildgebenden Verfahren, was mit einer Kostenexplosion einhergeht, zugunsten wesentlich verbesserter Diagnostik.

Dieses Sammelwerkchen ist im Hinblick auf den Begriff des 'Molecular Imaging' zu verstehen, unter dem sowohl klassische, als auch in der Entwicklung befindliche, und sogar zukünftige Verfahren der Bildgebung zusammengefaßt werden. Er ist qualitativ ähnlich zu dem immer häufiger verwendete Begriff der 'Life sciences', der eine Zusammenfassung klassischer Disziplinen wie Mikro- und Molekularbiologie, Teilen der Medizinischen Forschung, Biophysik, insbesondere Genomik, Proteomik und Metabolomik, unter Einbeziehung endogener und in letzter Zeit verstärkt, auch exogener Einflußgrößen meint. Insofern ist 'Molecular Imaging' eng mit 'Life Sciences' verflochten, wenn nicht ein Teilgebiet des letzteren. Die absehbare zukünftige Entwicklung scheint einerseits in der Digitalisierung, z.B. der Röntgendiagnostik, und andererseits in der Zusammenführung physikalisch und technisch verschiedener Ansätze, deren Matching, zu bestehen. Letzteres hat das Ziel, einen Informationsgewinn durch Zusammenführung verschiedener Meßdaten aus unterschiedlichen Quellen zu erlangen. Hier sind mathematische und statistische Modellierung und Interferenz besonders gefragt.

1.1 Signale und Bilder

Wir werden nun spezieller. Zunächst diskutieren wir grundlegende Begriffe und geben einen Ausblick auf die abzuhandelnden Themen.
Die Gebiete *Signal-* und *Bildverarbeitung* bzw. *Analyse* lassen sich inhaltlich streng genommen nicht trennen. 'Bilder' sind ja mehrdimensionale Signale, insbesondere wenn man unter den Begriff beliebige Intensitätsmuster subsummiert.
Die Unterscheidung zwischen Signal und Bild rührt eher von den eingesetzten theoretischen und technischen Methoden her. Eindimensionale Größen, die also nur von einem Parameter - meist als Zeit interpretiert - sind natürlich leichter zu handhaben als mehrdimensionale, bei denen i.a. Korrelationen zwischen den Koordinaten vorliegen. Außerdem wächst die zur Auswertung benötigte Rechenleistung mit der Dimension beträchtlich. Deshalb wurden früher vor allem eindimensionale Signale bearbeitet, während jetzt mehrdimensionale zunehmend auch in der Praxis an Bedeutung gewinnen.
Im Gegensatz dazu sollte man die Begriffe 'Verarbeitung' und 'Analyse' von Signalen auseinanderhalten. Während 'Verarbeitung' im wesentlichen eine Transformation der Daten (z.B. Filtern) meint, die meist linearer Natur ist, bedeutet 'Analyse' die Interpretation der Daten ('dieses Bild zeigt einen Hund', 'dieser Abschnitt des EEG weist auf Epilepsie hin'), was notwendig nichtlineare Verfahren erfordert.
Schließlich noch ein Wort zum Begriff 'medizinisch'. Selbstverständlich sind die Methoden der Bildverarbeitung und -Analyse allgemein anwendbar. Fast alle in der Medizin angewandten Technologien sind in der Biologie und zum Teil auch in der Technik einsetzbar, zumindest wenn man sie in gewisser Allgemeinheit betrachtet. Die zerstörungsfreie Inspektion eines Körpers z.B. kann bei (menschlichen) Patienten, Pflanzen oder Tieren, sonstigen Geweben oder auch teuren Werkstücken nützlich sein. Verwandte Methoden greifen auch in der Astronomie und der Mikroskopie. Wollte man dem Attribut 'medizinisch' wirklich Geltung verschaffen, so müßte man sich auf spezifisch medizinische, also sehr spezielle Gesichtspunkte einschränken. Dies ist aufgrund des unterstellten Vorwissens nicht möglich und auch nicht erstrebenswert. Somit werden wir uns mit einigen grundlegenden Prinzipien der Bildverarbeitung auseinandersetzen und diese (unter anderem) durch ihre Anwendungen in der Medizin motivieren.
Abschließend sei noch das Wort 'digital' kommentiert. Natürlich ist die konkrete rechnerische Verarbeitung der Signale letztlich digital. Andererseits sind die zu untersuchenden Objekt praktisch immer kontinuierlich. Eine systematische Theorie muß sich also auch mit kontinuierlichen Größen, die anschließend digitalisiert werden, beschäftigen.

Ein Bild im landläufigen Sinne wäre ein zweidimensionales Muster von Grauwerten oder Farben, d.h. ein Intensitätsmuster sichtbarer Strahlung. Solche Bilder liegen unmittelbar z.B. bei der optischen Mikroskopie vor. Sie können nach Digitalisierung im üblichen Sinne be- und verarbeitet oder analysiert werden. Die Einschränkung auf solche Muster wäre allerdings zu eng, da gerade in der Medizin verschiedenartigste Strahlungen oder Wellen eingesetzt werden. Neben der optischen Mikroskopie ist z.B. die Elektronenmikroskopie von entscheidender Bedeutung. Man denke auch an Röntgenaufnahmen, Ultraschall, Tomographie in ihren unterschiedlichsten Varianten, etc. In diesen Fällen geht es erst ein Mal um die *Erzeugung* eines beschaubaren Bildes aus den aufgenommenen Signalen.

1.2 Versuch einer Systematik

Um uns einen groben Überblick über die gängigen Methoden zu verschaffen, ziehen wir [13], Band I, Kapitel 26, zu Rate. Unter 'Biomedical Image Analysis' findet man in Kapitel 26:

1. Mikroskopische Bildanalyse: Biologie und Pathologie
   (a) Überblick
      i. Optische Mikroskopie
      ii. Elektronenmikroskopie
      iii. Autoradiographie
      iv. Gel-Elektrophorese und Chromatographie
      v. Digitalisierung in der Biologie
   (b) Analyse optischer medizinischer Bilder
      i. Vorverarbeitung
      ii. Morphometrie
      iii. Bildanalyse in der Pathologie
   (c) Makroskopische Bildanalyse: Radiologie, Psychiatrie, Kardiologie, und Ophtalmologie[1]
      i. Computergestützte Analyse medizinischer Bilder
      ii. zeitabhängige Bilder
      iii. Stationäre nichtinvasive Bilder
      iv. tomographische Rekonstruktion

[1] Augenheilkunde

v. Röntgen-Computertomographie
vi. Positron-Emissionstomographie
vii. Magnetische Resonanztomographie
viii. Nuklearmedizin
ix. Ultraschallbildanalyse
x. Dianographie
xi. Ophtalmographie
xii. Elektroenzephalographische Tomographie

Dies ist der Stand von 1986, der heute sicher nur noch eine Teilmenge der verwendeten Techniken darstellt, von der Praxis her gesehen der Wirklichkeit aber noch recht nahe kommt. Eine genauere Übersicht wird in [2], dieser Band, gegeben.

Kapitel 2

Bildgebende Verfahren

Wir skizzieren die Wirkungsweise einiger bildgebender Verfahren, die vor allem im medizinischen Bereich eingesetzt werden. Die Skizzen der Wirkungsweisen sind grob und lassen viele Aspekte außer betracht.

2.1 Optische Mikroskopie

Methoden der Bildverarbeitung können natürlich sowohl im Forschungs- als auch im klinischen Labor eingesetzt werden. In der Praxis handelt es sich oft um Standardverfahren zur Bildverbesserung wie Histogrammtransformationen oder Filterung, Extraktion quantitativer morphometrischer Daten, 3-D Rekonstruktionen aus 2-D Schichtbildern usw. Standardisierte Methoden für diagnostische Tests für Blutzellen, Gewebeproben, Abstriche, Chromosomenpräparate etc. werden angewandt. Im wesentlichen sollen die Verfahren die Fachkräfte unterstützen und entlasten.
Den praktischen Einsatz sollte man allerdings nicht überschätzen. Dies liegt zum einen an den hohen Gerätekosten und zum anderen an den oft sehr speziellen Problemstellungen, für die Software zu vernünftigen Preisen nicht am Markt ist. Häufig werden hausinterne Verfahren handgestrickt. Hier ist der Einsatz der Bildverarbeitung praktisch in den Anfängen. Anders verhält es sich im Bereich der Großforschung, wo die Chance zur integrierten und übergreifenden Implementierung besteht.
Trotzdem liegen interessante Problemstellungen vor. Software zur Zählung von Zellen gewissen Typs in einem Schnittbild z.B. benutzen ein vielfältiges Arsenal an klassischen Techniken wie Histogrammegalisierung, Hoch- und Tiefpaßfilterung, Thresholding, Erosion etc. Andere Probleme treten z.B. bei der Pathologie auf. Aus 2-D Schnitten soll ein 3-D Objekt rekonstruiert werden. Die Probleme liegen im Detail. Hier wird das Objekt in Paraffin

gegossen und dann in feine Streifen geschnitten. Die Schichten werden dabei gefältelt. In einer Flüssigkeit wird das Paraffin wieder abgelöst, wobei sich die Schichten glätten. Unglücklicherweise treten dabei gleich mehrere Störfaktoren auf: Die Schichten quellen unterschiedlich und nehmen dabei unterschiedliche Gestalt an, Stücke brechen heraus und neue Falten können entstehen. Das Matchen von sich in verschiedenen Schichten entsprechenden Punkten wird also erheblich erschwert.
Das verbreitetste Instrument ist das *Brightfield-Mikroskop*. Im durchscheinenden Licht sind mehr oder weniger lichtundurchlässige Körper erkennbar. Pigmentarme Objekte müssen eingefärbt werden, was eine eigene Kunst ist. Lebende Organismen können damit nur eingeschränkt untersucht werden.
Das *Phasen-Kontrast-Mikroskop* wird üblicherweise zur Untersuchung lebenden Materials ohne vorhergehende Färbung eingesetzt. Das entstehende Bild ist eine Funktion der Variation des Brechungsindexes im Material. Somit werden vor allem Ränder hervorgehoben. Hier werden Fourier-Filtertechniken eingesetzt.
Polarisations-Mikroskopie (POL) und *differentielle Interferenz-Kontrast-Mikroskopie* (DIC) werden zu ähnlichen Zwecken eingesetzt. Erstere sind Modifikationen des Brightfield-Mikroskops, wobei verschiedene Polarisationsfilter eingesetzt werden. Bei letzterem werden spezielle Prismen verwendet. Es gibt den Gradienten des Brechungsindexes wieder.
POL und DIC Bilder sind oft sehr wenig mit menschlichen Sehgewohnheiten verträglich. Bildverabeitung wird deshalb herangezogen.

2.2 Röntgendiagnostik

Hier wird die Probe von Röntgenstrahlen durchquert. Aufgenommen wird die auf ein Medium wie z.B. eine Filmfolie auftreffende Strahlung. Die Betrachtung erfolgt im durchscheinenden Licht, an Monitoren etc. Wir werden nicht näher auf die Physik dieser Vorgänge eingehen. Trotzdem erwähnen wir einige grundlegende Fakten (vgl. [7]).[1]
Röntgen- und Gammastrahlen sind elektromagnetische Wellen sehr kurzer Wellenlänge. Gewisse Eigenschaften können aber nur durch ihren Quantencharakter erklärt werden. Unter Röntgenstrahlung im engeren Sinne versteht man die Strahlung, die durch Abbremsung von Ladungsträgern hoher Geschwindigkeit entsteht. (Energiereichere Strahlung entsteht zum Beispiel bei radioaktivem Zerfall, d.h. sie hat ihre Ursache nicht allein in der Elektronenhülle, sondern im Atomkern.)

[1] Bilder: [7], [5], S. 1-57

Ein Elektron mit Elementarladung $e = 1,6 \cdot 10^{-19}$ As kommt bei freier Strecke von Kathode[2] zu Anode (dem Target) [3] unter einer Potentialdifferenz U mit der Energie

$$E = e \cdot U$$

an der Anode an. Durch Abbremsung entsteht Strahlung der Wellenlänge λ bzw. der Frequenz ν, wobei $\nu = c/\lambda$ mit der Lichtgeschwindigkeit $c = 2,998 \cdot 10^8\ m/s$. Andererseits gilt $E = h \cdot \nu$ mit dem Planckschen Wirkungsquantum $h = 6,625 \cdot 10^{-34}\ Js$. Somit ist

$$e \cdot U = h \cdot c/\lambda, \quad \lambda = \frac{h \cdot c}{e \cdot U}.$$

Einsetzen und Umrechnung der Einheiten ergibt dann maximale Größen

$$\lambda_{grenz}(m) = \frac{1,24 \cdot 10^{-9}}{U\ (kV)}, \quad \nu_{grenz}\ (s^{-1}) = 2,42 \cdot 10^{17}\ U\ (kV).$$

Es kommen aber auch niedrigerfrequente Strahlungsanteile vor.
Solch kurzwellige Strahlung (hoher Quantenenergie) kann Materie gut durchdringen. Ein Teil geht dem primären Strahlenbündel durch Wechselwirkung verloren. Dies nutzt man zur Erzeugung von Abbildungen des Körperinneren.
Wir verwenden den in der Physik üblichen Formalismus.
Die Anzahl dN der mit einer Schicht der (infinitesimalen) Dicke dx wechselwirkenden Quanten ist proportional zur Schichtdicke und der Zahl der auftreffenden Quanten. Die Proportionalitätskonstante μ ist der 'totale Schwächungskoeffizient' (total attenuation coefficient):

$$dN = -\mu\, N\, dx,$$

d.h.

$$N = N_0\, \mathrm{e}^{-\mu d},$$

mit der Schichtdicke d.
Dies gilt nur für dünne homogene Schichten und konstante Energie $h\nu$ der einfallenden Quanten. Bei Vorliegen eines Spektrums ist über dieses zu integrieren. μ hängt von Dichte und Ordnungszahl der Materie, sowie $h\nu$ ab.
Die Schwächung beruht auf drei Effekten.
(a) Photoeffekt. Trifft ein Quant auf ein Hüllenelektron geringerer Bindungsenergie, so fliegt dieses als freies Photoelektron heraus. Seine Energie ist die um die Bindungsenergie verringerte Energie des Quants. Das entstehende

[2]- Pol
[3]+ Pol

Elektronenloch wird durch ein Elektron aus äußeren Hüllen gefüllt. Die hierbei frei werdende Energie wird in Form eines elektromagnetischen Quants emittiert.
(b) Streuung. Das Quant kann auch direkt mit einem Hüllenelektron wechselwirken. Es verläßt die Hülle dann in neuer Richtung. Behält es dabei seine volle Energie, so liegt kohärente oder Rayleigh-Strahlung vor. Sonst heißt der Effekt inkohärente Strahlung oder Compton-Effekt.
(c) Paarbildung. Ist die Energie des Quants höher als 1,022 MeV, so kann das Quant mit dem Coulombfeld des *Kerns* ein Elektron-Positron-Paar bilden. Dieser Effekt ist im vorliegenden Fall nicht von Bedeutung.
Der inverse Effekt ist die Vereinigung eines Positrons mit einem Elektron

$$e^+ + e^- \longrightarrow 2\gamma$$

unter Bildung zweier 511-keV-Gammaquanten. Diese fliegen in entgegengesetzter Richtung auseinander und sind Grundlage für die *Positronenemissionstomographie* (PET), mit der wir uns später beschäftigen werden.
Die entstehende Streustrahlung bildet einen Streustrahlenschleier, der in erster Näherung einer homogenen Zusatzbelichtung entspricht. Dies führt zu erheblicher Kontrastminderung. Für Beckenaufnahmen beträgt der Anteil der Streustrahlung ca. 85% ![4]. Dem versucht man u.a. durch Kollimatoren zu begegnen, die nur Strahlen spezieller Richtungen durchlassen. Damit verlängert sich wiederum die Belichtungszeit, also auch die Strahlenbelastung.
Es werden mehr oder weniger klassische Bildverarbeitungstechniken eingesetzt. Eine witzige Art der 'Filterung' ist z.B. das 'Verwischen' oder 'Veratmen'. Für Aufnahmen der Halswirbel wählt man etwa die Belichtung so, daß der Patient den Mund öffnen und schließen kann. So wird der störend überlagernde Unterkiefer weggewischt. Alternativ können auch R-Quelle und Target scherend so bewegt werden, daß das interessierende Objekt scharf abgebildet und der Rest verwischt wird ('konventionelle Tomographie')[5]. Ähnlich funktioniert die transaxiale Tomographie, bei der die Röntgenquelle fest bleibt, aber Patient und Film rotieren[6]. Die Rotation wird so ausgeführt, daß die interessante Region im Körper stets derselben Region auf dem Film entspricht, der Rest aber auf dem Film wandert und so verwischt wird. Diese Methoden verlieren durch neue Tomographieverfahren an Bedeutung.
Die *digitale Subtraktionsangiographie* (DSA) dient der scharfen Darstellung von mit Kontrastmittel markiertem Gewebe. Hier kommt eine der primitivsten Bildverabeitungstechniken zum Einsatz: Das Maskenbild, d.h. eine

[4]Bild [7], S. 251
[5]Bild [5], S. 2
[6]Bild [4], S. 29

Aufnahme ohne Kontrastmittel wird einfach vom Füllungsbild mit Kontrastmittel subtrahiert. Dies wurde ursprünglich durch gleichzeitige Sicht durch das Füllungsbild und ein Positiv der Maske bewerkstelligt![7] Fensterung der Grauwerte und deren anschließend Aufspreizung dienen der Kontrasthebung, Hochpaßfilterung der Bildverbesserung[8].
Ein wesentliches Problem ist die Strahlenbelastung, die bis vor kurzem dramatisch unterschätzt wurde. Diese Problematik spielt vor allem bei Reihenuntersuchungen wie den Mamma Screenings eine wesentliche Rolle. Von Seiten des IBB liegen Untersuchungen in Zusammenarbeit mit dem Institut für Strahlenschutz zum Thema der Niedrigdosismammographie vor, siehe z.B. [10] und [3].

Abbildung 2.1: : Wilhelm Conrad Röntgen, geboren am 27. März 1845 in Lennep (Remscheid), gestorben am 10. Februar 1923 in München. Physiker, der die später nach ihm benannten Röntgenstrahlen entdeckte und damit bedeutende Grundlagen für die moderne Physik lieferte. Röntgen war der erste Physiknobelpreisträger.

Im Jahre 1995 jährte sich übrigens die Endeckung der Röntgenstrahlen (X-Rays) durch WILHELM CONRAD RÖNTGEN (1845-1923) zum hundertsten Mal. Er bekam dafür 1901 den Nobelpreis, siehe Fig. 2.1.

[7]Bild [7], S. 351, S. 348(a), [5], S. 66
[8]Bild [7], S. 348

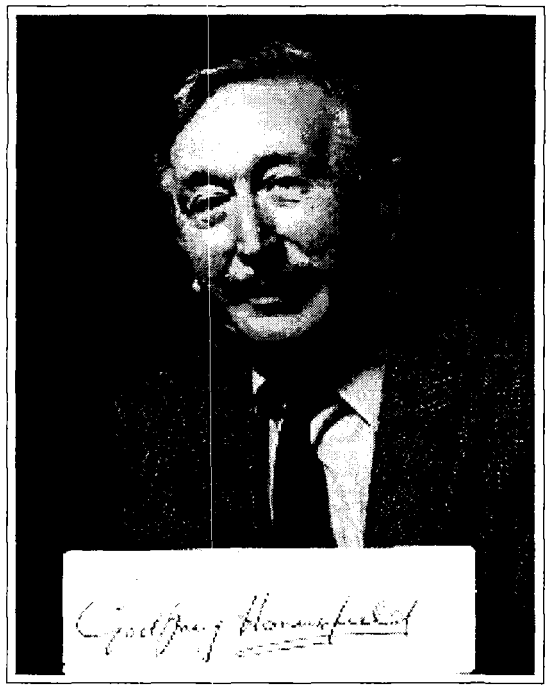


Abbildung 2.2: : Sir Godfrey Newbold Hounsfield, geboren am 28. August 1919 in Newark, Nottinghamshire, England. Englischer Elektroingenieur, der den Nobelpreis für Physiologie und Medizin mit Allan Cormack telte, für seinen Anteil an der Entwicklung der Axialen Computertomographie (CAT).

2.3 Röntgen-Computertomographie

In der klassischen Röntgendiagnostik beobachtet man den 'Schatten' eines Körpers bei 'Beleuchtung' durch Röntgenstrahlung. Dies läßt bei guter Ausbildung und Übung Rückschlüsse auf gewisse Eigenschaften des Körperinneren zu. Eine genauere Rekonstruktion des inneren Aufbaus ist nicht möglich.

Aus einem Schattenwurf läßt sich ein (etwa 2-D) Objekt nicht rekonstruieren. So wirft etwa eine Kreisscheibe mit Radius r denselben Schatten, wie ein achsenparallel beleuchtetes Quadrat mit Seitenlänge r. Schatten aus verschiedenen Richtungen liefern allerdings Zusatzinformation. Es stellt sich die Frage:

> Kann man die Gestalt eines Körpers oder eine Massenverteilung aus mehreren Projektionen rekonstruieren?

Falls diese Frage positiv beantwortet wird, bietet sich damit die Möglichkeit,

Details des Körperinneren aus Röntgenaufnahmen unter mehreren Richtungen zu rekonstruieren. Diese Idee liegt der Computertomographie zugrunde. Im Jahre 1972 stellten G.N. HOUNSFIELD, siehe Abb. 2.2, und J.A. AMBROSE der Fachwelt die ersten Röntgenschichtaufnahmen vor.
Diese Aufnahmen waren zwar sehr grob gerastert, zeigten aber innere Organe mit bis dahin nicht gekannter Deutlichkeit[9]. Graue und weiße Hirnsubstanz waren unterscheidbar, viele Verletzungen waren deutlich erkennbar. Dies leitete eine revolutionäre Erneuerung (Verbesserung und Verteuerung) der Röntgendiagnostik ein. Sie gilt als der größte Fortschritt der Röntgendiagnostik seit der Endeckung der Röngenstrahlen. Im Jahre 1979 wurde G.N. HOUNSFIELD von Central Research Laboratories of EMI und A.M. CORMACK, einem Physiker an der Tufts University, der Nobelpreis in Physiologie/Medizin zuerkannt. Welch ungemeinen Fortschritt die neue Methode brachte, zeigt sich im Vergleich[10].
Die 3-D Rekonstruktion beruht auf der Überlagerung von Schichtaufnahmen, die idealerweise als 2-D Bilder aufgefaßt werden können. Jede Schicht ist in der Praxis zwischen wenigen mm und 1 cm dick. Gemessen wird der Schattenwurf über approximative 1-D Strahlen. Eine Projektion erstreckt sich dabei entweder über viele parallele Strahlen, oder über einen Fächer von Strahlen. Solche Projektionen werden dann unter mehreren Winkeln gewonnen. Das Schema ergibt sich aus den Abbildungen 2.3. Natürlich kann man nicht direkt auf die eigentlichen Gewebeeigenschaften schließen, sondern nur auf den (Röntgenstrahlungs-) Schwächungskoeffizienten $\mu(x, y)$, wobei x, y

[9]Bild [7], S. 109
[10]Bild [4], S.30, 31

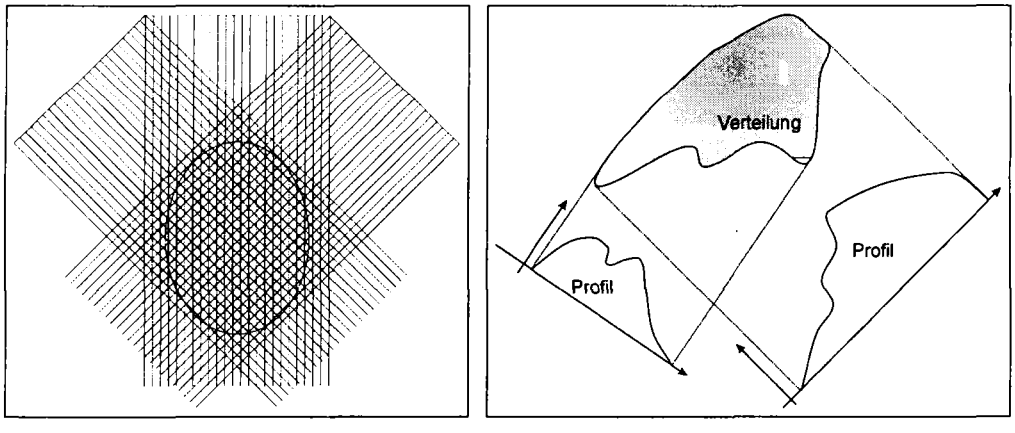


Abbildung 2.3: Projektionen einer 2-D Intensitätsverteilung

die ebenen Koordinaten bezeichnen (attenuation coefficient, siehe oben). Ist $o(s) = p + s\eta$, $s \geq 0$, eine Parametrisierung des Strahls von der Quelle am Ort p durch den Körper bis zum Detektor am Ort $p + s_d\eta$, so gilt für die Intensität des Strahls am Ort $p + s\eta$:

$$dI(s) = -\mu(s)I(s)ds.$$

Ist die Anfangsintensität I_0, so ergibt sich als Lösung der entsprechenden Differentialgleichung

$$\frac{dI(s)}{ds} = \mu(s)I(s)$$

die Lösung

$$I(s) = I_0 \exp\left(-\int_0^s \mu(t)\,dt\right),$$

was durch Differentiation sofort zu verifizieren ist. Also hat man für die Intensität I_d am Detektor

$$-\ln\frac{I_d}{I_0} = \int_L \mu(s)\,ds.$$

Die meisten Detektoren geben als Ausgangssignal in der Tat den Logarithmus der kumulierten Intensität wieder; Logarithmieren wäre andererseits auch kein zu großer Aufwand.
Bezeichnet man die Projektion mit $p = -\ln(I_d/I_0)$, so erhält man eine einzelne *Projektion* entlang einer Linie als

$$p = \int_L \mu.$$

Die Gesamtheit dieser Projektionen bezeichnet man (in 2-D) als *Radon-Transformation* $\mathcal{R}(\mu)$. Die *mathematische* Fragestellung ist, ob diese Transformation invertierbar ist, d.h. ob man aus der Gesamtheit der Projektionen die Funktion μ an jeder Stelle rekonstruieren kann und ob diese Rekonstruktion mittels eines praktikablen Algorithmus durchgeführt werden kann. Damit ist der eigentliche Kern dieser Vorlesung umrissen.
Wir werden sehen, daß die Radon-Transformation in engem Zusammenhang zur Fouriertransformation steht und die mehr oder weniger bekannten Fourieralgorithmen deshalb zur Lösung des Problems beitragen.
Es gibt eine Fülle von Problemen, die mit der groben Vereinfachung bei obigem Modell zusammenhängen. Eines davon ist das sogenannte 'beam hardening': Die Funktion μ hängt in Wirklichkeit auch von der Energie E der Strahlen ab. Ein genaueres Modell wäre

$$\frac{I}{I_0} = \int T(E)\exp\left(-\int_L \mu(u,v,E)\,dudv\right)\,dE,$$

wobei $T(E)$ das Energespektrum bezeichnet. Verwendung des vereinfachten Modells verursacht Artefakte (z.B. Streifen) die häufig nach Erfahrung beseitigt werden. Quantenrauschen führt zu radialen Rauschstrukturen. Reduktion des Rauschpegels erfordert Dosiserhöhung (oder längere belichtung). Streueffekte müssen bei manchen Systemen rechnerisch beseitigt werden.
Moderne Geräte tasten den Patienten nicht mehr Schicht für Schicht, sondern spiralförmig ab. Dies bedingt eine Modifikation der geschilderten Methoden.Um den Kern der Algorithmen einfach darstellen zu können, bleiben wir beider Schichtbild CT.
Den technischen Forschritt im Lauf von 20 Jahren illustriert die Tabelle in [14], S.24. Die Schwächung wird übrigens in Hounsfield-Einheiten (HU) gemessen, wobei -1000 HU keiner Schwächung (Luft) und 0 HU der Schwächung in Wasser entspricht.
Ohne diese Anwendungen zu ahnen, löste JOHANN RADON das mathematische Rekonstruktionsproblem in $\mathbb{R}^2$ und $\mathbb{R}^3$ im Jahre 1917. Dem gingen Ideen von H. MINKOVSKI (1904-1906), P. FUNK (1913, 1916) und Diskussionen mit G. HERGLOTZ voraus. Anfänglich waren bei der Entwicklung der RCT diese Arbeiten nicht bekannt und vieles wurde (teilweise) wiedererarbeitet und wiederentdeckt. Dies hat damit zu tun, daß die Entwickler der Tomographie englischsprachig waren, während die erwähnten Mathematiker in deutscher Sprache publizierten. Radons Arbeit erschien in *Berichte Sächsische Akademie der Wissenschaften. Leipzig, Math.-Phys.Kl.*, **69**, 262-267, [9].

2.4 SPECT

In der 'single photon emission tomography' (SPECT) wird dem Patienten ein radioaktives Kontrastmittel verabreicht, das sich im interessierenden Bereich konzentriert. Beim radioaktiven Zerfall werden Photonen entlassen und mit einer um den Patienten rotierenden γ-Kamera gezählt. SPECT liefert Schnittbilder der Organfunktionen, des regionalen Blutflusses und des regionalen Stoffwechsels. Bereits 1968 wurde ein Digitalrechner zur Rekonstruktion des Schnittbildes eingesetzt. Idealerweise fängt ein Detektor genau die Elektronen auf, die entlang eines idealen Strahls L durch den Körper in seine Richtung fliegen. Ist f die Konzentration des Radiopharmazeutikums, so ist die Intensität am Kollektor durch das Linienintegral

$$I = \int_L f$$

gegeben, d.h. man mißt approximativ die Radontransformierte von f. Die Schwächung der Strahlung im Körper gemäß einem Schwächungskoeffizienten

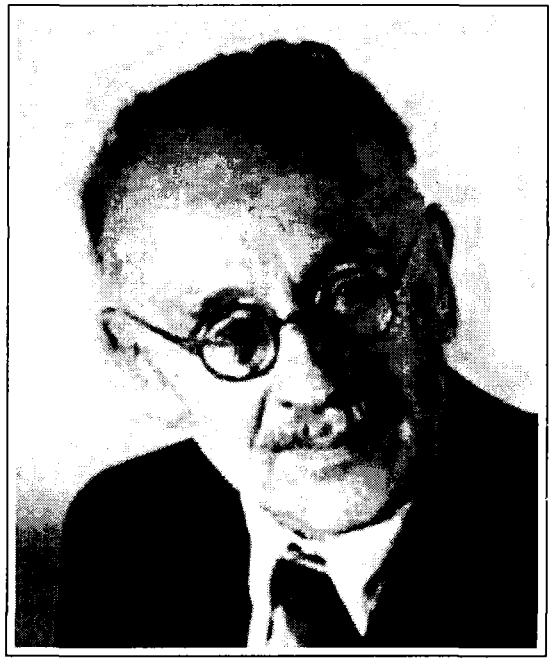


Abbildung 2.4: Johann Radon: Geboren 16. Dezember 1887 in Tetschen, Böhmen (jetzt Decin, Tschechische Republik), Gestorben: 25. Mai 1956 in Wien, Österreich

$\mu(x,y)$ ist allerdings nicht vernachlässigbar, was in Analogie zu obigem zur Intensität

$$I = \int_L f(x,y) \exp\left(-\int_{L(x,y)} \mu(u,v)\, dudv\right) dxdy$$

führt, wobei $L(x,y)$ der Abschnitt von L zwischen (x,y) und dem Detektor ist.Die Gesamtheit dieser Integrale stellt eine Verallgemeinerung der Radontransformation, die abgeschwächte Radontransformation oder 'attenuated Radon transformation' (ART) dar. Deren Theorie ist natürlich noch schwieriger als die übliche. Die (Streu- und) Absorbtionseffektehaben vergleichsweise schlechte Bildqualität und niedrige Auflösung zur Folge. Wegen des starken poissonschen Quantenrauschens können hier(wie oben geschehen) die eigentlich statistischen Vorgänge nur unzureichend durch das deterministische Modell (Mittelwerte!) beschrieben werden. Deshalb werden statistische Verfahren zur besseren Auswertung der Daten entwickelt. Diese sind allerdings vergleichsweise zeitaufwendig.

2.5 PET

Die Positronen-Emissionstomographie (PET) nutzt die besonderen Eigenschaften der Positronenstrahler und der Positronenannihilation aus, um quantitativ die Funktion von Organen oder Zellbereichen zu bestimmen. Beim Zerfall des Positronenstrahlers wird ein Proton umgewandelt in ein Positron, ein Neutron und ein Neutrino:

$$p \longrightarrow e^+ + n + \nu.$$

Die Reichweite des Positrons beträgt nur wenige Millimeter. Im Gewebe wird das Positron abgebremst und von einem Hüllenelektron eingefangen. In einem Annihilation genannten Prozeß wird die Masse dieses Paares in zwei γ-Quanten umgewandelt, die in entgegengesetzter Richtung auseinander fliegen:

$$e^+ + e^- \longrightarrow 2\gamma.$$

Die Energie der beiden γ-Quanten beträgt aufgrund des Energie-und Impulserhaltungsgesetzes jeweils 511 keV.
Die beiden γ-Quanten werden von zwei gegenüberliegenden Detektoren aufgefangen[11]. Die Detektoren sind kreisförmig um die zu messende Aktivitätsverteilung (z.B. den Patienten) angeordnet. Die gemessene Intensität ist dann

$$\begin{aligned} I &= \int_L f(x,y) \exp\left(- \int_{L_+(x,y)} \mu(u,v)\, dudv - \int_{L_-(x,y)} \mu(u,v)\, dudv \right) dxdy \\ &= \int_L f(x,y) \exp\left(- \int_{L(x,y)} \mu(u,v)\, dudv \right) dxdy \end{aligned}$$

wobei L_+ und L_- die beiden Teile von L links und rechts von (x,y) sind. Es ergibt sich also keine neue Transformation.

2.6 Verwandte Verfahren

Ein weiteres modernes Verfahren ist die Magnetresonanztomographie (MRT). Die Darstellung der zugrundeliegenden Physik würde unseren Rahmen sprengen. Schließlich treten ähnliche Rekonstruktionsprobleme in der Sonographie, Ultraschall CT und Elektronenmikroskopie auf (vgl. [7], [1], [8]).

[11] Bild [7], S.492

Literatur

[1] S.R. DEANS (1983): *The Radon Transform and Some of its Applications.* John Wiley & Sons: New York etc.

[2] B. FORSTER & R. LASSER (2004): *Biomedical Imaging: An Overview.* Dieser Band

[3] H. FÜHR, O.TREIBER & F. Wanninger (2003): Cluster-oriented Detection of Microcalcifications in Simulated Low-Dose Mammography. *Proceedings of 'Bildverarbeitung für die Medizin'*, Springer Verlag, 96-100

[4] G.T. HERMAN (1980): *Image Reconstruction from Projections.* Academic Press: New York etc.

[5] D.A. LISLE (1996): *Imaging for Students.* Arnold: London etc. und Oxford University Press: New York

[6] A. MACOVSKI (1983): *Medical Imaging Systems.* Prentice-Hall Inc.: Englewood Cliffs, New Jersey 07632

[7] MORNEBURG H., SIEMENS AG (Hrsg.) (1995): *Bildgebende Systeme für die medizinische Diagnostik.* 3. Auflage, Publicis-MCD-Verlag

[8] F. NATTERER (1986): *The Mathematics of Computerized Tomography.* John Wiley & Sons: Chichester etc.

[9] J. RADON (1917): Über die Bestimmung von Funktionen durch ihre Ingetralwerte langs gewisser Mannigfaltigkeiten. *Berichte Sächsische Akademie der Wissenschaften*, 69:262–279,

[10] O. TREIBER, F. WANNINGER, H. FÜHR, W. PANZER, D. REGULLA & G. WINKLER (2003): An adaptive algorithm for the detection of microcalcifications in simulated low-dose mammography. *Physics in Medicine and Biology*, 48(3):449–466, .

[14] (1996) *Mathematics and Physics of Emerging Biomedical Imaging.* National Academic Press: Washington, D.C.

[12] G. WINKLER(2004): *Mathematische Grundlagen der Radontransformation.* Dieser Band

[13] TZAY Y. YOUNG & KING-SUN FU (Eds.) (1986): *Handbook of Pattern Recognition and Image Processing*, Academic Press: Inc:San Diego etc.

[14] (1996) *Mathematics and Physics of Emerging Biomedical Imaging.* National Academic Press: Washington, D.C.

Teil II

Mathematische Grundlagen der Radontransformation

Mathematische Grundlagen der Radontransformation

Gerhard Winkler*

Zusammenfassung

Dieser Essay ist als Einführung in die mathematischen Grundlagen einiger bildgebender Verfahren, wie sie in der medizinischen Diagnostik und Forschung seit geraumer Zeit etabliert sind, und zunehmend in den Bio- und Lebenswissenschaften eine wichtige Rolle spielen, gedacht. Er beschränkt sich auf die Darstellung der Elemente der Radontransformation. Einige wichtige bildgebende Verfahren werden in [5] und [17], beides dieser Band, sowie im letzten Teil dieses Bandes besprochen.
Es wird am Anfang begonnen, d.h. die Mathematik wird vom Grunde her aufgebaut. Deshalb ähnelt der Essay in seinen vorderen Teilen einer Vorlesung in Analysis. Dies wird ganz bewußt so getan, denn ohne diese - heute im Studium nicht selbstverständlichen Grundlagen - ist ein tieferes Verständnis der Materie aus mathematischer Sicht nicht denkbar.

*Institute of Biomathematics and Biometry
GSF - National Research Center for Environment and Health,
Postfach 1129, D-85758 Oberschleißheim, Germany,
gwinkler@gsf.de, http://ibb.gsf.de

Inhaltsverzeichnis

Einleitung

Ein Essay dieses Titels könnte von einem Mathematiker, Informatiker, Physiker, Techniker oder Mediziner mit jeweils völlig verschiedenen Inhalten gefüllt werden. Dies wird nach einem Blick in [13], [12] und [10] unmittelbar klar. So wird zum Beispiel ein Mediziner Aspekte wie Patientenvor- und Nachsorge für die Untersuchung, Verträglichkeit der Strahlung oder Kontrastmittel, praktische Ruhigstellung bei langen Belichtungszeiten oder mehrfachen Aufnahmen hervorheben. Ein Techniker wird die Beschreibung der zahlreichen hintereinandergeschalteten Apparate und ihrer Systemgrößen in den Vordergrund stellen und ein Mathematiker wird die Grundlagen der verwendeten Algorithmen und deren Analyse betonen. Neue, anspruchsvolle, und wohl für unser Institut, das IBB, zukünftig prägende Herausforderungen kommen von Seiten der Mikro- und Molekularbiologie.
Auch aus der engeren Sicht der Mathematik, Statistik können verschiedene Schwerpunkte gesetzt werden. Dies geschieht vorzugsweise nach Methoden und Modellen und weniger nach konkreten Anwendungen. So stehen z.B. hinter Elektronenmikroskopie und Computertomographie ähnliche mathematische Modelle (wie auch z.B. hinter der Radioastronomie), während man aus Sicht des Anwenders erstere vielleicht nur als Verbesserung der optischen Mikroskopie oder anderer Verfahren auffassen würde.
In der medizinischen Signalverarbeitung spielen Zeitreihenmethoden eine dominante Rolle. Diese beruhen zum großen Teil auf der Theorie stationärer stochastischer Prozesse. Die meisten der in [17] genannten bildgebenden Verfahren hingegen beruhen auf der Lösung inverser Probleme. Ein zentrales Beispiel für solche mathematische Verfahren ist die 'Rekonstruktion aus Projektionen'. Der Kern sind Transformationen vom Typ der 'Radontransformation', die eng mit der bekannten Fouriertransformation zusammenhängt. Dies ist eine rein analytische Theorie.
In diesem Text konzentrieren wir uns auf die Radontransformation. Dabei sind wir uns bewußt, daß in modernen Geräten auch andere mathematische Methoden zum Einsatz kommen, etwa basierend auf dem EM-Algorithmus und anderen nicht leicht zu evaluierenden numerischen Methoden.

Kapitel 1

The Fanciful Engineer: Systemtheorie

Wir versuchen in diesem Abschnitt, wie ein Ingenieur zu denken. Wir unterstellen, daß alle Operationen mathematisch gerechtfertigt sind. Ein 'Filter' empfängt Signale als Input, verändert sie und gibt das veränderte Signal als Output wieder heraus.

1.1 Lineare shiftinvariante Filter

Im Idealfall kann man hoffen, daß ein Filter linear und shiftinvariant arbeitet[1]; letzteres bedeutet, daß ein verschobenes Signal in den ebenso verschobenen Output des ursprünglichen Signals transformiert wird.

Beispiel 1.1 *Glättungsfilter.* Bezeichnen f den In- und g den Output, so entsteht $g(x)$ als Mittelung der Werte von f über ein Fenster $[x-\varepsilon, x+\varepsilon]$:

$$\begin{aligned} g(x) &= \frac{1}{2\varepsilon}\int_{x-\varepsilon}^{x+\varepsilon} f(y)\,dy = \frac{1}{2\varepsilon}\int f(y)\mathbf{1}_{[x-\varepsilon,x+\varepsilon]}(y)\,dy \\ &= \frac{1}{2\varepsilon}\int f(y)\mathbf{1}_{[-\varepsilon,\varepsilon]}(y-x)\,dy = \int f(y)\left\{\frac{1}{2\varepsilon}\cdot \mathbf{1}_{[-\varepsilon,\varepsilon]}(x-y)\right\} dy \\ &=: \int f(y)h(x-y)\,dy =: f*h(x) \end{aligned}$$

In mehreren (d) Dimensionen wird aus dem Integral ein Mehrfachintegral und aus der *Filtermaske* $(1/2\varepsilon)\mathbf{1}_{I_\varepsilon}$ mit $I_\varepsilon = [-\varepsilon,\varepsilon]$ die Maske $(1/(2^d\varepsilon^d))\mathbf{1}_{I_\varepsilon^d}$.

[1] diese Annahme *kann* in der Realität nur annähernd erfüllt sein, da ein realer Filter Signale nur in einer endlichen Skala verarbeiten kann, Linearität aber (durch Multiplikation mit beliebig großen Zahlen) unbeschränkte Skalen erfordert.

Um den Einfluß weiter entfernter Orte oder Zeiten zu dämpfen, hätten wir z.B. die Maske h^ε gemäß[2]

$$h(x) = \begin{cases} \text{const} \cdot \exp\left(-\frac{1}{1-|x|^2}\right) & \text{falls } x \leq 1 \\ 0 & \text{sonst} \end{cases}, \quad \int h(y)\,dy = 1,$$

$$h^\varepsilon(x) = \frac{1}{\varepsilon^d} \cdot h\left(\frac{x}{\varepsilon}\right)$$

verwenden können. An obiger Rechnung hätte sich dadurch nichts geändert.

1.2 Die Faltung

Motiviert durch das Beispiel 1.1 definieren wir auch für allgemeine Filtermasken h die *Faltung*

$$g(x) = f * h(x) = \int f(y)h(x-y)\,dy.$$

Wir halten h fest. Die Faltung ist
(a) linear:

$$(\alpha f + \beta g) * h = \alpha(f * h) + \beta(g * h),$$

(b) kommutativ:

$$f * h = h * f,$$

Beweis

$$f * h(x) = \int f(y)h(x-y)\,dy \underbrace{=}_{z=x-y} \int f(x-z)h(z)\,dz = h * f(x).$$

□

Hierbei wurde die Substitutionsregel benutzt. Die Rechnung bleibt auch im mehrdimensionalen richtig. Anstelle der Substitutionsregel muß dann die Integraltransformationsformel verwendet werden[3].

[2] mit $x = (x_1, \ldots, x_d)$ und $|x| = \sqrt{x_1^2 + \cdots + x_d^2}$. Wir verwenden die Betragstriche für die Norm, weil wir später noch weitere Normen verwenden werden und zu viele Striche vermeiden wollen.

[3] **Die Transformationsformel:** Seien $D_1, D_2 \subset \mathbb{R}^d$ offen und $\varphi : D_1 \longrightarrow D_2$ bijektiv, sowie φ und φ^{-1} stetig differenzierbar und für die Jakobimatrix J der ersten partiellen Ableitungen von φ sei $|\det J(x)| > 0$ auf D_1. Dann gilt

$$\int_{D_2} f(x)\,dx = \int_{D_1} f(\varphi(x))|\{\det J(x)|\,dx.$$

(c) shiftinvariant: Für $a \in \mathbb{R}$ sei

$$(\tau_a f)(x) = f(x-a)$$

der (Rechts-) *Shift* um a. Dann gilt:

$$(\tau_a f * h)(x) = \tau_a (f * h)(x).$$

Beweis

$$(\tau_a f) * h(x) = h * (\tau_a f)(x) = \int h(y) f(\{x-a\}-y)\, dy = h * f(x-a) = \tau_a (h * f)(x).$$

□

(d) assoziativ:

$$(f * g) * h = f * (g * h) =: f * g * h.$$

Diese Eigenschaft werden wir leicht aus der Fouriertransformation ablesen können (siehe (1.5)).

1.3 Die Punktantwort

Die Funktion h beschrieb den Filter nur bezüglich seiner *Wirkung* auf das Signal f über die Faltung. Können wir h durch diese Wirkung auf Signale bestimmen?
In einer Dimension sei

$$f^\varepsilon(x) = \frac{1}{2\varepsilon} \cdot \mathbf{1}_{[-\varepsilon,\varepsilon]}(x).$$

Dann gilt nach Beispiel 1.1 (mit f und h in vertauschten Rollen)

$$f^\varepsilon * h(x) = \frac{1}{2\varepsilon} \cdot \int_{x-\varepsilon}^{x+\varepsilon} h(y)\, dy \longrightarrow h(x), \quad \varepsilon \to 0,$$

nach dem Mittelwertsatz und des weiteren

$$\int f^\varepsilon(y)\, dy = 1.$$

Wir werden für diesen Sachverhalt die *formale Schreibweise*

$$\begin{aligned} \delta(x) &= \lim_{\varepsilon \to 0} f^\varepsilon(x) \\ \delta * h(x) &= h(x) \end{aligned} \tag{1.1}$$

verwenden und $\delta(y)$ den 'scharfen' oder 'Einheitsimpuls' in $x = 0$ nennen. In dieser Sprache wäre $h(x)$ die (Punkt-) Antwort (in der Bildverarbeitung *Point-Spread-Funktion*, abgekürzt PSF) des Filters auf den scharfen Impuls in 0. Entsprechend wäre der verschobene Impuls

$$\delta_a(x) = \delta(x - a)$$

der scharfe Impuls in a mit der Antwort $h(x - a)$ in x.

1.4 Die Delta- oder Diracfunction

Würden wir die Schreibweise in (1.1) nicht nur formal, sondern wörtlich nehmen, so wäre δ eine Funktion mit

$$\delta(0) = \infty, \ \delta(x) = 0 \text{ falls } x \neq 0, \ \int \delta(y)\, dy = 1.$$

Eine solche Funktion gibt es natürlich nicht. In trivialer Weise können wir aber das *Funktional* der Auswertung von h definieren:

$$'\delta * h'(x) := h(x).$$

Bemerkung 1.1 Das Funktional

$$h \longmapsto \delta * h(x), \ \delta * h(x) = h(x)$$

kann als Integral bezüglich des Punkt- oder Diracmaßes ε_x dargestellt werden, denn es gilt

$$\int h(y)\, d\varepsilon_x = h(x) = \delta * h(x) = \int h(y)\delta(x - y)\, dy = \int h(y)\{\delta_x(y)\, dy\}.$$

In der Sprache der Maßtheorie suggeriert die eingeführte Schreibweise (man vergleiche die äußeren Ausdrücke), daß das Diracmaß ε_x die 'Dichte'

$$\delta_x(y) = \begin{cases} \infty & \text{für } y = x \\ 0 & \text{sonst} \end{cases}$$

hat.

Das Symbol '$*$' hat jetzt eine andere Bedeutung als oben für klassische Funktionen. Dies führt aber zu keinerlei Verwechslungen und wir benutzen es unverdrossen.
Bis jetzt haben wir nur Symbole eingeführt. Wir möchten aber die für Funktionen gängigen Operationen wie z.B. die Differentiation so erweitern, daß sie auch für solche Funktionale in der bekannten Weise gelten. Gelänge uns das, so wäre die Schreibweise

$$\int \delta(y)h(x-y)\,dy := \delta * h(x)$$

weiterhin sinnvoll. In der Tat haben Physiker und Ingenieure die Mathematiker gezwungen, einen solchen Kalkül zu entwickeln, so daß sie weiter wie mit Funktionen rechnen durften. Dies ist der Kalkül der *verallgemeinerten Funktionen* oder *Distributionen*; das Funktional δ ist eine solche verallgemeinerte Funktion.

1.5 Die Transferfunktion

Die Auswertung der Faltung kann das Berechnen komplizierter Integrale bedeuten. Die Multiplikation ist eine elementare Operation. Können wir die Faltung auf eine einfache Multiplikation zurückführen?
Besonders schöne Signale sind z.B.

$$f(x) = \cos \lambda x, \quad f(x) = \sin \lambda x, \quad \lambda, x \in \mathbb{R}.$$

So schön sie auch sind, das Rechnen mit ihnen ist ein Horror (vgl. Additionstheoreme). Deshalb arbeitet man lieber mit dem komplexwertigen Signal

$$f_\lambda(x) = \mathrm{e}^{\imath\lambda x} = \cos \lambda x + \imath \sin \lambda x$$

mit der imaginären Einheit $\imath$. Die reellwertigen harmonischen Schwingungen erhält man daraus zurück durch

$$\cos \lambda x = \frac{1}{2}\left(\mathrm{e}^{\imath\lambda x} + \mathrm{e}^{-\imath\lambda x}\right), \quad \sin \lambda x = \frac{1}{2\imath}\left(\mathrm{e}^{\imath\lambda x} - \mathrm{e}^{-\imath\lambda x}\right). \tag{1.2}$$

Bemerkung 1.2 Die Frequenz ν einer solchen Schwingung ist die Anzahl der vollständigen Perioden in einem Intervall der Länge 2π. Also ist

$$\nu = \frac{\lambda}{2 \cdot \pi}.$$

Physikalisch entspricht also λ der *Kreisfrequenz*. In Ingenieurbüchern wird gerne mit der Form $\cos(2\pi\nu x)$ statt mit $\cos(\lambda x)$ etc. gearbeitet.

Bemerkung 1.3 In mehreren Dimensionen entspricht das[4]

$$e^{\imath\langle\lambda,x\rangle} = \cos\langle\lambda,x\rangle + \imath\sin\langle\lambda,x\rangle.$$

Zur Veranschaulichung in der Ebene etwa des Realteiles $\cos(\lambda_1 x_1 + \lambda_2 x_2)$ beachten wir, daß der Cosinus sein Maximum 1 an den Stellen

$$\lambda_1 x_1 + \lambda_2 x_2 = k\pi, \quad k = 0, 1, \ldots$$

annimmt. Das sind Linien senkrecht zum Vektor λ. Der Graph erinnert also an ein Wellblech mit den Rippen senkrecht zu λ.

Wie wirkt der Filter, beschrieben durch eine Faltung, auf solch ein elementares Signal? So:

$$\begin{aligned} f_\lambda * h(x) &= \int f_\lambda(x-y)h(y)\,dy = \int h(y)e^{\imath\langle\lambda,x-y\rangle}\,dy \\ &= e^{\imath\langle\lambda,x\rangle}\int h(y)e^{-\imath\langle\lambda,y\rangle}\,dy =: f_\lambda(x)\cdot H(\lambda). \end{aligned}$$

In Kurzform haben wir

$$f_\lambda * h = H(\lambda)\cdot f_\lambda.$$

f_λ passiert also den Filter bis auf eine Multiplikation mit der (eventuell komplexen) Zahl $H(\lambda)$ ungeschoren. Wie in der linearen Algebra sagen wir: f_λ ist eine Eigenfunktion zum Eigenwert $H(\lambda)$ der linearen Abbildung

$$f \longmapsto f * h.$$

Die Schwingung bleibt also unverändert, außer daß die Amplitude $H(\lambda)$ wird. Die Zahl $H(\lambda)$ charakterisiert die Antwort des Filters auf die Schwingung vollständig, genauso wie h die Antwort des Filters auf einen scharfen Impuls vollständig charakterisierte.
Sei nun f eine gewichtete Mischung

$$f(x) = \sum_{i=1}^{m} \alpha_i f_{\lambda_i}(x)$$

[4]wobei $\lambda = (\lambda_1, \ldots, \lambda_d)$, $x = (x_1, \ldots, x_d)$ und

$$\langle\lambda, x\rangle = \sum_{i}^{d} \lambda_i x_i$$

das euklidische Skalarprodukt im $\mathbb{R}^d$ bezeichnet

der elementaren Schwingungen f_{λ_i}. Dann gilt wegen der Linearität:

$$f * h(x) = \sum_{i=1}^{m} (\alpha_i H(\lambda_i)) f_{\lambda_i}(x).$$

Sei allgemeiner f eine kontinuierliche gewichtete Mischung

$$f(x) = \int F(\lambda) f_\lambda(x)\, d\lambda$$

der f_λ mit der Gewichtsfunktion $F(\lambda)$; dann ist plausibel, daß analog gilt:

$$f * h(x) = \int \{F(\lambda)H(\lambda))\} f_\lambda(x)\, d\lambda. \tag{1.3}$$

D.h. aus dem Filter kommt wieder eine gewichtete Mischung der elementaren Schwingungen f_λ heraus; die Gewichtsfunktion ist gegeben durch

$$F(\lambda) \cdot H(\lambda).$$

Die Wirkung des Filters ist also durch die Funktion

$$H : \lambda \longmapsto H(\lambda)$$

vollständig festgelegt, denn $f*h$ bekommt man über das Integral (1.3) zurück. Wir nennen H die *Modulationstransferfunktion* (MTF) oder auch *optische Transferfunktion* (OTF) des Filters. In der Mathematik heißt

$$H(\lambda) = \hat{f}(\lambda) = \mathcal{F}h(\lambda) = \int h(y) \mathrm{e}^{-\imath\langle\lambda, y\rangle}\, dy \tag{1.4}$$

die *Fouriertransformierte* von h. Die Abbildung

$$h \longmapsto \hat{f}$$

heißt Fouriertransformation[5].
Mehrfache Faltung $g * (f * h)$ hat die Transferfunktion $H_g \cdot (H_f \cdot H_h)$. Da die Multiplikation assoziativ ist, d.h.

$$H_g \cdot (H_f \cdot H_h) = (H_g \cdot H_f) \cdot H_h = H_g \cdot H_f \cdot H_h \tag{1.5}$$

ist auch die Faltung assoziativ, wie oben behauptet (natürlich nur, wenn wir die eineindeutige Zuordnung von Faltung der Signale zu Multiplikation der Transferfunktionen sichergestellt haben). Hier deutet sich die hilfreiche Rolle der Fouriertransformation auch bei theoretischen Untersuchungen an.

[5]Leider wird die Fouriertransformation in der Literatur nicht einheitlich definiert. Wir werden später eine multiplikative Konstante hinzufügen.

1.6 Mehr zu Filtern und Fouriertransformation

Was ist die Fouriertransformierte des scharfen Impulses?

$$H(\lambda) = \hat{\delta}(\lambda) = \int \delta(y) \mathrm{e}^{-i\langle \lambda, y\rangle}\, dy = \mathrm{e}^{i\langle \lambda, 0\rangle} \equiv 1.$$

Das ist klar, denn Faltung mit dem scharfen Impuls ändert das Signal ja nicht, d.h. F darf multiplikativ nicht verändert werden, was einer Multiplikation mit 1 entspricht.
Ist das eine Spielerei? Nein: Entspreche h einem Filter, der das Eingangssignal verrauscht. Die Transferfunktion sei mit H bezeichnet. Es soll ein Filter entwickelt werden, der das Rauschen vollständig beseitigt. Was muß er erfüllen? Wie kann er geplant werden? Klar: Multiplikation von F mit 1 bedeutet, daß f erhalten bleibt. Also suchen wir einen Filter mit

$$H(\lambda) \cdot \tilde{H}(\lambda) = 1.$$

d.h.

$$\tilde{H} = \frac{1}{H}.$$

Fertig![6] Mit Hilfe der verallgemeinerten Funktionen kann folgendes praktisch nützliche Resultat zeigen:
'Jeder lineare, shiftinvariante Filter läßt sich durch eine Faltung mit einer (verallgemeinerten) Funktion beschreiben.'

Beispiel 1.2 (Kantendetektoren) Um Kanten in Bildern oder abrupte Änderungen von Signalen zu finden, benutzt man häufig die Differentiation. Ist die Ableitung größer als eine gewisse Schranke ('Threshold'), so vermutet man eine Kante. Zweifellos ist die Differentiation ein linearer, shiftinvarianter Filter. Läßt er sich als Faltung darstellen? Das zu verifizieren, ist mit Hilfe der verallgemeinerten Funktion δ kein Problem (wir betrachten zunächst eine Dimension):

$$f'(x) = \lim_{\varepsilon \to 0} \frac{f(x+\varepsilon) - f(x-\varepsilon)}{2\varepsilon}.$$

Wir formulieren das mit Hilfe von δ:

$$f'(x) = \lim_{\varepsilon \to 0} f * \left\{ \frac{1}{2\varepsilon} \cdot (\delta(x+\varepsilon) - \delta(x-\varepsilon)) \right\}.$$

[6] so einfach ist es in der Praxis leider nicht. Wir müssen ja noch $\tilde{h}$ aus $\tilde{H}$ bestimmen!

Fassen wir den Begriff der verallgemeinerten Funktionen so, daß dieser Limes enthalten ist, so haben wir die Differentiation als Faltung

$$f'(x) = f * \lim_{\varepsilon \to 0} \left\{ \frac{1}{2\varepsilon} \cdot (\delta(x+\varepsilon) - \delta(x-\varepsilon)) \right\} =: f * \delta'(x).$$

dargestellt. Wie sieht die Transferfunktion der Differentiation aus?

$$f'(x) = \int F(\lambda) \frac{\partial}{\partial x_i} \mathrm{e}^{\imath\langle\lambda,x\rangle}\, d\lambda = \int \{F(\lambda)\imath\lambda\} \mathrm{e}^{\imath\langle\lambda,x\rangle}\, d\lambda.$$

Vergleich mit (1.3) zeigt

$$H(\lambda) = \imath \cdot \lambda.$$

Dies kann leicht auf d Dimensionen erweitert werden. Sei $x = (x_1, \ldots, x_d)$. Dannn gilt

$$\frac{\partial}{\partial x_i} f(x) = \frac{\partial}{\partial x_i} \int F(\lambda) \mathrm{e}^{\imath\langle,\lambda,x\rangle}\, d\lambda = \int \{F(\lambda)\imath\lambda_i\} \mathrm{e}^{\imath\langle\lambda,x\rangle}\, d\lambda.$$

Also ist die Transferfunktion des Filters

$$f \longmapsto \frac{\partial}{\partial x_i}$$

gegeben durch

$$H(\lambda) = \imath\lambda_i.$$

Differenzieren wir mehrfach, z.B. zwei mal, so ergibt sich

$$\frac{\partial^2}{\partial x_i^2} f(x) = \int \{F(\lambda)\imath^2\lambda_i^2\} \mathrm{e}^{\langle\lambda,x\rangle}\, d\lambda.$$

Die Transferfunktion der k-maligen Differentiation nach der i-ten Koordinate ist also

$$H(\lambda) = (\imath\lambda_i)^k.$$

Wir können die Linearität zur Bestimmung der Transferfunktion beliebiger Operatoren der partiellen Differentiation ausnutzen. Als Beispiel betrachten wir den *Laplace -Operator*

$$\triangle = \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_d^2}.$$

Er hat die Transferfunktion

$$H(\lambda) = \left(\imath^2\lambda_1^2 + \cdots \imath^2\lambda_d^2\right) = -|\lambda|^2.$$

Die Rotationsinvarianz von H legt nahe, daß auch der Filter selbst diese Eigenschaft hat. Rotationsvarianz ist eine sehr wünschenswerte Eigenschaft von Kantenfindern, da man ja Kanten in beliebiger Richtung detektieren will!

Mit diesen Illustrationen beenden wir vorläufig die Diskussion von verallgemeinerten Funktionen und der Fouriertransformation und gehen auf die Grundlagen zurück.

Kapitel 2

Das Lebesgue-Integral

Integrale sind die Grundlage des Kalküls. Da wir Anwendungen in der Bildverarbeitung im Auge haben, benötigen wir mehrdimensionale Integration. Wir gehen in mehreren Schritten vor, wobei jedesmal mehr integrierbare Funktionen hinzu kommen.

2.1 Stetige Funktionen mit kompaktem Träger

Eine Funktion

$$f : \mathbb{R} \longrightarrow \mathbb{R},\ x \longmapsto f(x).$$

hat kompakten Träger, wenn sie außerhalb eines endlichen Intervalls verschwindet. Die Funktion f sei nun sogar stetig. Dann existiert das bekannte eindimensionale Riemann-Integral $\int f\,dx = \int f(x)\,dx$. In Verallgemeinerung auf den $\mathbb{R}^d$ entsprechen dem stetige Funktionen

$$f : \mathbb{R}^d \longrightarrow \mathbb{R},\ x = (x_1, \ldots, x_d) \longmapsto f(x).$$

welche außerhalb eines endlichen Quaders

$$Q = I \times \cdots \times I = I^d,\ \ I = [a, b] \subset \mathbb{R}$$

verschwinden. Hält man $(x_2, \ldots x_d)$ fest, so ist

$$x_1 \longmapsto f(x_1, x_2, \ldots, x_d)$$

stetig mit kompaktem Träger in I, so daß

$$F_1(x_2, \ldots, x_d) = \int f(x_1, x_2, \ldots, x_d)\,dx_1$$

in obigem Sinne existiert. Nach dem Satz von der stetigen Abhängigkeit des Integrals von Parametern ist bei festem $(x_3, \ldots, x_d)$ die Funktion

$$x_2 \longmapsto \int f(x_1, x_2, x_3, \ldots, x_d)\, dx_1$$

stetig mit kompaktem Träger in I und somit existentem Riemannintegral

$$F_2(x_3, \ldots, x_d) = \int \left\{ \int f(x_1, x_2, x_3, \ldots, x_d)\, dx_1 \right\} dx_2.$$

Wir fahren bis x_d fort und erhalten das d-dimensionale Riemann-Integral für stetige Funktionen mit kompaktem Träger. Diese Funktionenklasse bezeichnen wir mit $\mathcal{K} = \mathcal{K}(\mathbb{R}^d)$ oder $\mathcal{C}_c(\mathbb{R}^d)$.

2.2 Monotone Limiten

Sei $(f_n)_n$ eine punktweise *aufsteigende* Folge in $\mathcal{K}$ und $f = \sup_n f_n$ der punktweise Limes (möglicherweise mit Werten ∞). Dann existiert

$$\int f\, dx = \sup_n \int f_n\, dx = \lim_n \int f_n\, dx \in \mathbb{R} \cup \{\infty\}.$$

Die Klasse dieser f bezeichnen wir mit $\mathcal{K}^\uparrow$; darauf haben wir das Integral nun erweitert.
Analog definieren wir das Integral für die Klasse $\mathcal{K}^\downarrow$ von Infima *absteigender* Folgen in $\mathcal{K}$. Diese Integrale nehmen Werte in $\mathbb{R} \cup \{-\infty\}$ an.

2.3 Das Lebesgue-Integral

Müssen wir dieses Spiel ad Infinitum treiben? Nein: Wir sind fertig:
Sei f irgendeine Funktion auf $\mathbb{R}^d$ und seien

$$\begin{aligned} \int^* f\, dx &= \inf \left\{ \int \varphi\, dx : \varphi \in \mathcal{K}^\uparrow, \varphi \geq f \right\}, \\ \int_* f\, dx &= \sup \left\{ \int \varphi\, dx : \varphi \in \mathcal{K}^\downarrow, \varphi \leq f \right\}. \end{aligned}$$

$f : \mathbb{R}^d \to \mathbb{R}$ heißt *Lebesgue-integrierbar*, wenn

$$-\infty < \int_* f\, dx =: \int f\, dx = \int^* f\, dx < \infty.$$

Die Menge der integrierbaren Funktionen bezeichnen wir mit $\mathcal{L}^1(\mathbb{R}^d)$. Wir werden sehen, daß weitere Hüllenbildung nichts neues bringt (vgl. Theorem 2.2).

2.4 Konsistenz von Lebesgue-und Riemann-Integral

Das Riemann-Integral wurde in der Analysis für eine größere Klasse als $\mathcal{K}(\mathbb{R}^1)$ eingeführt. Viele Integrale wurden berechnet. Dieses Wissen wollen wir nicht verlieren:

Satz 2.1 *Die Funktion $f : \mathbb{R} \to \mathbb{R}_+$ sei Borel-meßbar*[1] *und über jedes kompakte Intervall Riemann-integrierbar. Dann ist f genau dann Lebesgue-integrierbar, wenn das uneigentliche Riemann-Integral existiert. Lebesgue- und Riemann-Integral stimmen dann überein.*

Bis auf pathologische Fälle stimmen also Riemann-Integral (falls definiert) und Lebesgue-Integral überein. Praktisch alle Berechnungen von Riemann-Integralen übertragen sich also auf das Lebesgue-Integral.

2.5 Nullmengen

Wenn für eine Menge $N \subset \mathbb{R}^d$ gilt

$$\int \mathbf{1}_N \, dx = 0,$$

so heißt N eine *Lebesguesche Nullmenge.* Eine Eigenschaft, die außerhalb einer Nullmenge gilt, gilt *fast überall* oder *fast sicher* (f.s.).
Wir werden im folgenden Funktionen, welche sich nur auf Nullmengen unterscheiden, identifizieren: $f \sim g$, falls $\{x : f(x) \neq g(x)\}$ eine Nullmenge ist. In der Tat werden sehen, daß

$$f = g \text{ f.s.} \implies \int f \, dx = \int g \, dx. \tag{2.1}$$

Beispiel 2.1 Punktmengen sind Nullmengen. Der Beweis ist klar: Zu jedem Punkt x finden wir eine absteigende Folge in $\mathcal{K}$, die gegen $\mathbf{1}_{\{x\}}$ konvergiert und deren Integrale gegen Null konvergieren.

[1] was das heißt, lernt man in der Wahrscheinlichkeitstheorie. Im Moment wäre uns das nur lästig. Wir halten aber fest: Alle stetigen, stückweise stetigen, halbstetigen Funktionen haben diese Eigenschaft.

2.6 Vertauschung von Limiten

Viele Probleme der Analysis können auf die Frage nach der Vertauschung von Limiten reduziert werden. Das Integral ist ein Limes. Also stellt sich die Frage, wann Limes und Integral vertauscht werden können.
Die Stärke des Lebesgue-Integrals liegt im wesentlichen darin, daß diese Vertauschung in allen sinnvollen Situationen erlaubt ist. Aufgrund der Definition des Integrals über monotone Limiten ist folgender Satz sofort einleuchtend:

Satz 2.2 (Satz von der monotonen Konvergenz, von Beppo Levi) *Sei $(f_n)_n$ eine monotone Folge von Funktionen in $\mathcal{L}^1(\mathbb{R}^n)$ mit $f = \lim_n f_n$. Dann gilt*

$$\lim_n \int f_n \, dx = \int \lim_n f_n \, dx = \int f \, dx.$$

Ist $\sup_n \int f_n \, dx < \infty$, so ist $f \in \mathcal{L}^1$.

Beispiel 2.2 Sei $(N_i)_i$ eine Folge von paarweise disjunkten Nullmengen. Dann ist auch

$$N = \bigcup_i N_i$$

eine Nullmenge. Zum Beweis setzen wir $f_n = \mathbf{1}_{N_1} + \cdots \mathbf{1}_{N_n}$. Die Folge dieser f_n steigt monoton gegen $f = \mathbf{1}_N$, nach dem Satz von der monotonen Konvergenz also auch die Folge $(\int f_n \, dx)_n$ der verschwindenden Integrale. Das Supremum verschwindet dann auch, d.h. $\int f \, dx = 0$ und N ist eine Nullmenge. Wegen Beispiel 2.1 sind also insbesondere abzählbare Mengen Nullmengen.
Die Menge Q der rationalen Zahlen ist abzählbar. Also ist

$$\int \mathbf{1}_{\mathbb{Q}} \, dx = 0.$$

Dies ist ein Beispiel einer Lebesgue-integrierbaren Funktion, die nicht Riemann-integrierbar ist.

Eine noch praktischere Version ist

Satz 2.3 (Satz von der majorisiertenKonvergenz) (Satz von der dominierten Konvergenz, Lebesguescher Konvergenzsatz) *Sei $(f_n)_n$ eine Folge in $\mathcal{L}^1$, die fast überall konvergiert und*

$$|f_n| \leq g$$

für ein $g \in \mathcal{L}^1$. Dann gibt es $f \in \mathcal{L}^1$, wogegen f_n fast überall konvergiert, dieses f ist integrierbar und

$$\lim_n \int f_n \, dx = \int \lim_n f_n \, dx = \int f \, dx.$$

Wenn man bedenkt, daß Stetigkeit über Limiten von Funktionswerten und Differenzierbarkeit über Limiten von Differenzenquotienten definiert sind, erkennt man die Bedeutung dieser Sätze für die stetige Abhängigkeit des Integrals von Parametern und die Vertauschbarkeit von Integration und Differentiation.

Satz 2.4 (Stetige Abhängigkeit des Integrals von Parametern) *Sei* $U \subset \mathbb{R}^m$ *offen und*

$$f : \mathbb{R}^d \times U \longrightarrow \mathbb{R}, (x, \xi) \longmapsto f(x, \xi)$$

eine Funktion mit
(i) für jedes ξ *ist* $x \mapsto f(x, \xi)$ *integrierbar,*
(ii) für jedes x *ist* $\xi \mapsto f(x, \xi)$ *stetig,*
(iii) Es gibt ein integrierbares g *mit*

$$|f(x, \xi)| \leq g(x) \text{ für alle } x, \xi.$$

Dann ist die Funktion

$$\xi \longmapsto \int f(x, \xi)\, dx$$

stetig.

Beweis Betrachte eine Folge $\xi_n \to \xi$. Wegen *(iii)* und dem Satz von der dominierten Konvergenz dürfen wir für die Funktionenfolge

$$f_n = f(\cdot, \xi_n)$$

Limes und Integral vertauschen. Das liefert die Stetigkeit des Integrals im Parameter ξ. □

Praktisch noch wichtiger ist:

Satz 2.5 (Vertauschbarkeit von Differentiation und Integration) *Sei* $I \subset \mathbb{R}$ *ein Intervall und*

$$f : \mathbb{R}^d \times I \longrightarrow \mathbb{R}$$

eine Funktion mit
(i) für jedes x *ist* $\xi \mapsto f(x, \xi)$ *differenzierbar,*
(ii) es gibt ein integrierbares g *mit*

$$\left| \frac{\partial f}{d\xi}(x, \xi) \right| \leq g(x).$$

Dann ist

$$I_f(\xi) : \xi \longmapsto \int f(x,\xi)\,dx$$

differenzierbar und es gilt

$$I_f'(\xi) = \frac{\partial}{\partial \xi} \int f(x,\xi)\,dx = \int \frac{\partial f}{\partial \xi}(x,\xi)\,dx.$$

Beweis Die Ableitung von f ist Limes von Differenzenquotienten, z.B.

$$f_n(x,\xi) = \frac{f\,(x,\xi+h_n)) - f(x,\xi)}{h_n},\ h_n \to 0,\ n \to \infty.$$

Die Behauptung erfordert, daß man diesen Limes unter das Integral ziehen darf. Wenn wir

$$|f_n(x,\xi)| \leq g(x)$$

zeigen können, läßt Theorem 2.3 dies zu. Nach dem Mittelwertsatz der Differentialrechnung gibt es aber $\vartheta \in [0,1]$ mit

$$|f_n(x,\xi)| = \left|\frac{\partial f}{\partial \xi}(x,\xi+\vartheta h_n)\right| \leq g(x).$$

Damit ist der Beweis vollständig. □

Merkregel 2.1 *Diese zwei Sätze zusammen mit der partiellen Integration bilden den harten Kern des gesamten zu entwickelnden Kalküls.*

2.7 Vertauschung der Integrationsreihenfolge

Eine Funktion $f(x,y)$ mit zwei Argumenten kann möglicherweise über x und dann über y integriert werden oder umgekehrt; vieleicht auch gleich über $z = (x,y)$. Wir wollen natürlich, daß immer das Gleiche herauskommt.

Satz 2.6 (Satz von Fubini) *Sei $f : \mathbb{R}^k \times \mathbb{R}^m \to \mathbb{R}$ eine integrierbare Funktion. Dann ist die Funktion*

$$\mathbb{R}^k \longrightarrow \mathbb{R}, x \longmapsto f(x,y)$$

für fast alle y integrierbar. Setzt man

$$F(y) = \int f(x,y)\,dx$$

wo möglich und 0 sonst, dann ist F integrierbar und

$$\int f(x,y)\,d(x,y) = \int F(y)\,dy.$$

Man hat also – landläufig formuliert –

$$\int f(x,y)\,d(x,y) = \int \left(\int f(x,y)\,dx \right) dy = \int \left(f(x,y)\,dy \right) dx.$$

Merkregel 2.2 *Die Integrationsreihenfolge spielt keine Rolle.*

Als Illustration definieren wir die *Faltung*. Seien $f, h \in \mathcal{L}^1(\mathbb{R}^d)$. Dann ist $(y,z) \mapsto f(y)h(z) \in \mathcal{L}^1(f, h \in \mathcal{L}^1(\mathbb{R}^d)$. Dann ist $(y,z) \mapsto f(y)h(z) \in \mathcal{L}^1(\mathbb{R}^{2d})$. Nach dem Transformationssatz ist

$$(x,y) \longmapsto f(y)h(z-y)$$

ebenfalls integrierbar und

$$f * h(z) = F(z) = \int f(y)h(z-y)\,dy$$

existiert für fast alle z nach dem Satz von Fubini. Auf der Restmenge definieren wir die Faltung beliebig.

2.8 $\mathcal{L}^p$-Räume

Bei der Konstruktion des Integrals haben wir die Linearität

$$\int \alpha f + \beta g\,dx = \alpha \int f\,dx + \beta \int g\,dx$$

nirgends zerstört. Also ist $\mathcal{L}^1$ ein linearer Raum. Ferner ist f integrierbar genau dann wenn $|f|$ integrierbar ist. Man definiert

$$\|f\|_1 = \int |f|\,dx,$$

die $\mathcal{L}^1$-*Halbnorm*. Es gilt
(1) $\|f\| = 0$ genau dann, wenn $f = 0$ f.s.,
(2) $\|\alpha f\| = |\alpha| \|f\|$,
(3) $\|f + g\| \leq \|f\| + \|g\|$,

d.h. $\|\cdot\|$ ist eine *Halbnorm.* Bildet man den Quotienten L^1 bezüglich des Nullideals der Halbnorm, d.h. identifiziert Funktionen, die sich höchstens auf Nullmengen unterscheiden, so erhält man sogar einen normierten Raum.
Wichtig sind noch die Monotonieeigenschaften (vgl. Seite 129):

$$\begin{aligned} f, g \in \mathcal{L}^1, f \leq g &\implies \int f\,dx \leq \int g\,dx, \\ g \in \mathcal{L}^1, f \text{ meßbar}, |f| \leq |g| &\implies f \in \mathcal{L}^1. \end{aligned}$$

Dasselbe können wir im Hinblick auf die Fouriertransformation auch mit komplexwertigen Funktionen g machen:

$$\int g\,dx = \int \Re g\,dx + \imath \int \Im g\,dx.$$

Eine komplexwertige Funktion ist somit in diesem Sinne integrierbar, wenn Real- und Imaginärteil integrierbar sind. Weil für jede komplexe Zahl $z = u + \imath v$ die elementaren Ungleichungen

$$|u|, |v| \leq \sqrt{u^2 + v^2} \leq \sqrt{2\,(|u| + |v|)^2} = \sqrt{2}(|u| + |v|)$$

gelten, existieren beide Integrale genau dann, wenn $|g| = \sqrt{(\Re g)^2 + (\Im g)^2}$ integrierbar ist. Die Menge dieser Funktionen bezeichnet man auch mit $\mathcal{L}^1(\mathbb{R}^d, \mathbb{C})$. Die Halbnorm definiert man wie oben. Der Quotient L^2 bezüglich des Nullideales ist wieder ein normierter Raum.
Ein zweiter Raum wird uns noch begegnen: Seien

$$\begin{aligned} \mathcal{L}^2 &= \{f : |f|^2 \text{ integrierbar}\} \\ \|f\|_2 &= \sqrt{\int |f|^2\,dx} = \sqrt{\int f\bar{f}\,dx}. \end{aligned}$$

Auch hier liegt wieder ein halbnormierter linearer Raum vor. Die Räume L^1 und L^2 sind sogar vollständig, d.h. daß jede Cauchy-Folge (definiert wie in $\mathbb{R}$ nur mit $\|\cdot\|$ statt $|\cdot|$) einen Limes hat. L^p-Räume sind also *Banachräume.* Ihre Elemente sind Äquivalenzklassen $[f]$, d.h. Mengen von Funktionen in $\mathcal{L}^p$, die sich von f nur auf Nullmengen unterscheiden. Jedes Element $g \in [f]$ ist ein Repräsentant der Äquivalenzklasse. Wir schreiben in Zukunft wieder f für $[f]$.
L^2 ist von besonderer Bedeutung, weil er ein *Skalarprodukt*

$$\langle f, g \rangle = \int f\bar{g}\,dx$$

trägt, d.h.

$$\begin{aligned} \langle \alpha f + \beta g, h\rangle &= \alpha\langle f, h\rangle + \beta\langle g, h\rangle, \ \alpha, \beta \in \mathbb{C} \\ \overline{\langle f, g\rangle} &= \langle g, f\rangle, \\ \|f\|_2 \neq 0 &\Longrightarrow \langle f, f\rangle > 0 \end{aligned}$$

Es ist

$$\|f\|_2^2 = \langle f, f\rangle$$

und es gelten wie im euklidischen Raum

$$\begin{aligned} \langle f, 0\rangle &= 0 \\ |\langle f, g\rangle| &\leq \|f\|_2 \|g\|_2. \end{aligned}$$

Ist die Norm eines Banachraumes so durch ein Skalarprodukt gegeben, so ist er ein *Hilbertraum*. L^2 ist der wichtigste Hilbertraum. Er dient z.B. als wichtigstes Hilfsmittel in der Quantenmechanik.

Kapitel 3

Die Fouriertransformation

Wir führen die Fouriertransformation auf $\mathcal{L}^1(\mathbb{R}^d, \mathbb{C})$, die mathematische Grundlage der Transferfunktionen, ein.
Wir arbeiten mit komplexwertigen Funktionen

$$f : \mathbb{R}^d \longrightarrow \mathbb{C},\ f(x) = \Re f(x) + \imath \Im f(x).$$

Weil

$$|\exp(\imath\lambda)| = \sqrt{\cos^2\lambda + \sin^2\lambda} \equiv 1,$$

existiert zu $f \in \mathcal{L}^1$ das Integral

$$\hat{f}(\lambda) = \frac{1}{(2\pi)^{d/2}} \int f(x) \mathrm{e}^{-\imath\langle x,\lambda\rangle}\, dx.$$

Die Funktion

$$\hat{f} : \mathbb{R}^d \longrightarrow \mathbb{C}, \lambda \longmapsto \hat{f}(\lambda)$$

heißt *Fouriertransformierte* von f.

Bemerkung 3.1 Der Faktor vor dem Integral ist in der Mathematik üblich, weil er eine symmetrische Form der 'Rücktransformation' nach sich zieht. Abhängig vom Anwendungsgebiet sind verschiedene Definitionen im Schwange: z.B.

$$\int f(x) \exp\left(-\imath\langle x,\lambda\rangle\right) dx, \int f(x) \exp\left(-\imath 2\pi\langle x,\lambda\rangle\right) dx,$$

usw.. Dies ist extrem lästig, wenn man Formeln aus den entsprechenden Büchern übernehmen will. Die Umrechnung geht so:

$$\begin{aligned} \tilde{f}(\lambda) &= A \int f(x) \exp\left(-\imath B\langle x,\lambda\rangle\right) dx \Longrightarrow \\ \hat{f}(\lambda) &= \frac{1}{(2\pi)^{d/2} \cdot A} \cdot \tilde{f}\left(\frac{\lambda}{B}\right),\ \tilde{f}(\lambda) = (2\pi)^{d/2} \cdot A \cdot \hat{f}\left(B \cdot \lambda\right). \end{aligned}$$

3.1 Eigenschaften

Wir notieren erste und grundlegende Eigenschaften der Fouriertransformation:

Satz 3.1 (Linearität) *Die Abbildung $f \mapsto \hat{f}$ ist linear.*

Beweis Das ist klar. □

Satz 3.2 (Beschränktheit) *Es gilt*

$$\left|\hat{f}\right| \leq \frac{1}{(2\pi)^{d/2}} \cdot \|f\|_1.$$

Beweis Sei $c = (2\pi)^{-d/2}$. Dann gilt:

$$|\hat{f}(\lambda)| \leq c \cdot \int |f(x)| \cdot |\exp(-\imath\langle x, \lambda\rangle| \, dx = c \cdot \int |f(x)| \cdot 1 \, dx = c \cdot \|f\|_1,$$

und damit die postulierte Ungleichung. □

Satz 3.3 (Stetigkeit) *$\hat{f}$ ist stetig.*

Beweis Weil $|\exp(-\imath\langle x, \lambda\rangle)| = 1$ ist der Satz von der stetigen Abhängigkeit anwendbar. □

Satz 3.4 ($\hat{f}$ bei 0) *Es gilt*

$$\hat{f}(0) = \int f \, dx.$$

Beweis Man werte $\hat{f}$ bei 0 aus. □

Satz 3.5 (Ähnlichkeit) *Mit der Festsetzung*

$$Dx = \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_d \end{pmatrix} x = (d_1 x_1, \ldots, d_d x_d), \quad d_i \neq 0$$

gilt

$$(f \circ D)\hat{}(\lambda) = \frac{1}{|d_1 \cdot \cdots \cdot d_d|} \hat{f}\left(\frac{\lambda_1}{d_1}, \ldots, \frac{\lambda_d}{d_d}\right).$$

Beweis Dies folgt direkt aus dem Integraltransformationssatz. □

Eine Streckung der Ortsskala entspricht also einer Stauchung der Frequenzskala verbunden mit einer Reskalierung der Amplitude.

3.2 Verträglichkeit mit linearen Operationen

Die Fouriertransformation ist maßgeschneidert für den Umgang mit Shiftoperatoren und allen Operatoren, die mit diesen vertauschen. Wir betrachten ihre Reaktion auf Shifts, Faltung und Differentiation. Bezüglich des Shifts gilt

Satz 3.6 (Shifttheorem) *Es gilt*

$$\tau_a f(x) = f(x-a),\ a \in \mathbb{R}^d,\ \textit{so}\ (\tau_a f)^\wedge(\lambda) = \exp(-\imath\langle a,\lambda\rangle)\cdot \hat{f}(\lambda).$$

Beweis

$$\int f(x-a)\exp(-\imath\langle x,\lambda\rangle)\,dx = \int f(z)\exp(-\imath\langle z+a,\lambda\rangle)\,dz$$

$$= \exp(-\imath\langle a,\lambda\rangle)\int f(z)\exp(-\imath\langle z,\lambda\rangle)\,dz.$$

□

Die Faltung von Funktionen geht in die Multiplikation der Fouriertransformierten über:

Satz 3.7 (Faltung) *Sind $f,h \in \mathcal{L}^1$, so gilt*

$$(f*h)^\wedge = (2\pi)^{d/2}\hat{f}\cdot\hat{h}.$$

Beweis Das rechnet man einfach nach

$$(f*h)^\wedge(\lambda) = \frac{1}{(2\pi)^{d/2}}\int\left\{\int f(y)h(x-y)\,dy\right\}\exp(-\imath\langle x,\lambda\rangle)\,dx$$

$$= \frac{1}{(2\pi)^{d/2}}$$

$$\int\left\{\int h(x-y)\exp(-\imath\langle x-y,\lambda\rangle)\,dx\right\}f(y))\exp(-\imath\langle y,\lambda\rangle)\,dy$$

$$= \int \hat{h}(\lambda)f(y)\exp(-\imath\langle y,\lambda\rangle)\,dy = \hat{h}(\lambda)\cdot(2\pi)^{d/2}\hat{f}(\lambda),$$

und gelangt so zum Ergebnis. □

Die nächsten beiden wichtigen Resultate betreffen die Differentiation.

Satz 3.8 (Differentiation im Ortsbereich) *Sei $f \in \mathcal{C}_c^1(\mathbb{R}^d)$. Dann gilt*

$$\left(\frac{\partial}{\partial x_i}f\right)^\wedge(\lambda) = \imath\lambda_i\hat{f}(\lambda).$$

Beweis Wir betrachten zunächst ein reellwertiges f. Für jedes

$$(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$$

ist

$$g(x_i) = \frac{\partial}{\partial x_i} f(x)$$

eine stetige Funktion, die außerhalb eines kompakten Intervalls $[a, b]$ verschwindet. Also existiert stets das Integral $\int g\, dx_i = \int_a^b g\, dx_i$ sowohl im Lebesgueschen wie auch im Riemannschen Sinne. Partielle Integration liefert

$$\begin{aligned} & \int_a^b \frac{\partial}{\partial x_i} f(x) \exp(-\imath\langle x, \lambda\rangle)\, dx_i \\ = \; & f(x) \exp(-\imath\langle x, \lambda\rangle)|_{x_i=a}^b - \int_a^b f(x) \left\{ \frac{\partial}{\partial x_i} \exp(-\imath\langle x, \lambda\rangle) \right\} dx_i \quad (3.1) \\ = \; & \imath\lambda_i \int f(x) \exp(-\imath\langle x, \lambda\rangle)\, dx_i. \end{aligned}$$

Nach dem Satz von Fubini liefert nun Integration über die restlichen Variablen gefolgt von einer Multiplikation der Gleichung mit $(2\pi)^{-d/2}$ das gewünschte Ergebnis. □

Satz 3.9 (Differentiation im Frequenzbereich) Falls $x \mapsto x_i f(x)$ *integrierbar ist, so ist* $\hat{f}$ *nach* λ_i *stetig (partiell) differenzierbar und es gilt*

$$(x_i f)^\wedge = \imath \frac{\partial}{\partial \lambda_i} \hat{f}.$$

Beweis Nach dem Satz von der differenzierbaren Abhängigkeit dürfen wir frisch rechnen:

$$\begin{aligned} \imath \frac{\partial}{\partial \lambda_i} \hat{f}(\lambda) &= (2\pi)^{-d/2} \int f(x) \frac{\partial}{\partial \lambda_i} \exp(-\imath\langle x, \lambda\rangle)\, dx \\ &= (2\pi)^{-d/2} \int x_i f(x) \exp(-\imath\langle x, \lambda\rangle)\, dx = (x_i f)^\wedge(\lambda). \end{aligned}$$

□

Die Integrierbarkeitsbedingung ist lästig und sollte beseitigt werden! Das tun wir später.

3.3 Die Inversion

Führen wir Operationen wie Faltung oder Differentiation mit Hilfe der Fouriertransformation aus, so muß anschließend das Ergebnis 'zurücktransformiert' werden.

Satz 3.10 (Inversionsformel) *Seien $f, \hat{f} \in \mathcal{L}^1$. Dann gilt*

$$f(x) = (2\pi)^{-d/2} \int \hat{f}(\lambda) \exp(\imath\langle x, \lambda\rangle \, d\lambda.$$

Dies löst gleichzeitig unser früheres Problem nach der Darstellung von f als gewichtete Mischung der $f_\lambda(y) = \exp(\imath\langle y, \lambda\rangle)$. Insbesondere ist in diesem Fall eine Funktion durch ihre Fouriertransformierte eindeutig bestimmt.
Oft schreibt man $\mathcal{F}f = \hat{f}$. Definiert man die *inverse Fouriertransformation* (von $g(\lambda)$) als

$$\bar{\mathcal{F}}g(x) = (2\pi)^{-d/2} \int g(\lambda) \exp(\imath\langle x, \lambda\rangle) \, d\lambda,$$

so gilt mit $c = (2\pi)^{-d/2}$, daß

$$c \int g(\lambda) \exp(\imath\langle x, \lambda\rangle) \, d\lambda = c \overline{\int \overline{g(\lambda)} \exp(-\imath\langle x, \lambda\rangle) \, d\lambda},$$

d.h.

$$\bar{\mathcal{F}}g = \overline{\mathcal{F}\bar{g}},$$

wobei Überstreichen den Übergang zum komplex Konjugierten bedeutet. Die Inversionsformel bedeutet in dieser Schreibweise für $f, g \in \mathcal{L}^1$:

$$\bar{\mathcal{F}}\mathcal{F}f = f, \quad \mathcal{F}\bar{\mathcal{F}}g = g.$$

Die Funktionen f und $g = \hat{f}$ bilden in diesem Falle ein *Fouriertransformationspaar*. Wir sehen jetzt, warum die Mathematiker den Faktor $(2\pi)^{-d/2}$ wählen: Transformation und Rücktransformation bekommen denselben Vorfaktor.
Unglücklicherweise ist $\hat{f}$ nicht für jedes integrierbare f wieder integrierbar, so daß dieses schöne Wechselspiel zerstört wird!

DER INGENIEUR WILL, DASS ES IMMER GEHT!

Deshalb brauchen wir Distributionen.

3.4 Spezielle Fouriertransformierte

Als Übung und Beispiel rechnen wir einfache Fälle:

Beispiel 3.1 Sei $h = \mathbf{1}_{[-1,1]}$ die elementare eindimensionale Glättungsmaske. Dann gilt

$$\hat{h}(\lambda) = (2\pi)^{-1/2} \int_{-1}^{1} \mathrm{e}^{-\imath x\lambda}\, dx = (2\pi)^{-1/2} \cdot \left. \frac{\mathrm{e}^{-\imath x\lambda}}{-\imath x} \right|_{x=-1}^{x=1}$$

$$= (2\pi)^{-1/2} \cdot \frac{\exp(\imath\lambda) - \exp(-\imath\lambda)}{\imath\lambda} = \sqrt{\frac{2}{\pi}} \cdot \frac{\sin\lambda}{\lambda},$$

wobei $\sin(\lambda)/\lambda$ für $\lambda = 0$ als 1 zu interpretieren ist.
Wir sehen also:

1. Faltung mit h ist wegen der Überschwinger keine gute Glättung.
2. $\hat{h}$ ist nicht (Lebesgue-)integierbar[1]. (Man beachte aber die Fußnote[2].) Es liegt also kein Transformationspaar vor.

Masken $h_\varepsilon = \mathbf{1}_{[-\varepsilon,\varepsilon]}$ lassen sich als

$$h_\varepsilon(x) = \mathbf{1}_{[-1,1]}(\varepsilon x)$$

darstellen und nach der Ähnlichkeitsformel gilt

$$\hat{h}_\varepsilon(\lambda) = \sqrt{\frac{2}{\pi}} \cdot \frac{1}{\varepsilon} \cdot \frac{\sin\left(\frac{\lambda}{\varepsilon}\right)}{\frac{\lambda}{\varepsilon}}.$$

---

[1] Es ist

$$\int_{(k-1)\pi}^{k\pi} \left|\frac{\sin\lambda}{\lambda}\right| d\lambda \geq \frac{1}{k} \int_{(k-1)\pi}^{k\pi} |\sin\lambda|\, d\lambda = \frac{2}{k}$$

und somit die Summe dieser Teilintegrale unendlich.

[2] $\hat{h}$ ist uneigentlich Riemann-integrierbar, d.h. es existiert der Grenzwert

$$\lim_{R\to\infty} \int_{-R}^{R} \frac{\sin\lambda}{\lambda}\, d\lambda =: \int \frac{\sin\lambda}{\lambda}\, d\lambda \in \mathbb{R}.$$

Dies sieht man im Vergleich mit der (konvergenten) Leibnitzschen Reihe

$$\sum_{i=1}^{\infty} (-1)^{k-1} \frac{1}{k}.$$

Hier haben wir also eine Funktion, die uneigentlich Riemann-integrierbar, aber nicht Lebesgue-integrierbar ist.

Oft benutzt man die Schreibweise

$$\operatorname{sinc}(\lambda) = \frac{\sin(\pi\lambda)}{\pi\lambda},\ \lambda \in \mathbb{R}.$$

Wir schauen (im $\mathbb{R}^1$) nach, wie zweimaliges Glätten wirkt:

$$(h * h)\hat{}(\lambda) = \sqrt{2\pi} \cdot \frac{2}{\pi} \frac{\sin^2 \lambda}{\lambda^2} = 2 \cdot \sqrt{\frac{2}{\pi}} \cdot \frac{\sin^2 \lambda}{\lambda^2}.$$

Wir stellen fest:

1. Die Überschwinger sind beseitigt, aber der Filter ist nicht monoton in den Frequenzen.
2. Die Fouriertransformierte ist integrierbar und es liegt ein Transformationspaar vor.

Wegen

$$h * h(x) = \max(0, 2 - |x|)$$

resultiert zweimaliges Glätten mit h in der Anwendung der symmetrischen Dreiecksmaske über $[-2, 2]$ mit Scheitelhöhe 2.

Um ein Beispiel in mehreren Dimensionen zu haben, rechnet man:

Beispiel 3.2 Sei $h = \mathbf{1}_Q$ mit $Q = [-1, 1]^d$. Dann ist nach dem Satz von Fubini

$$\begin{aligned} \hat{h}(\lambda) &= (2\pi)^{-d/2} \int \cdots \int \prod_{i=1}^{d} \left\{ \mathbf{1}_{[-1,1]}(x_i) \exp(-\imath x_i \lambda_i) \right\} dx_1 \ldots dx_d \\ &= (2/\pi)^{d/2} \prod_{i=1}^{d} \frac{\sin \lambda_i}{\lambda_i}. \end{aligned}$$

Hätte man statt Q eine Kugel genommen, sähe das viel übler aus.

Schließlich betrachten wir noch den schönsten Fall:

Beispiel 3.3 Sei

$$h : \mathbb{R}^d \longrightarrow \mathbb{R},\ h(x) = \exp\left(-\frac{|x|^2}{2}\right).$$

Dann ist

$$\hat{h}(\lambda) = \exp\left(\frac{-|\lambda|^2}{2}\right).$$

h ist also seine eigene Fouriertransformierte. Das ist das natürlichste Transformationspaar, das man sich vorstellen kann.

Kapitel 4

Verallgemeinerte Funktionen oder Distributionen

Wir wollen einen Kalkül, bei dem wichtige Operationen wie Differentiation, Faltung, Limesbildungen oder (Hin- und Rück-) Fouriertransformation ohne weitere Vorsichtsmaßnahmen problemlos ausführbar sind.
Insbesondere sollen nur Fouriertransformationspaare auftreten, 'Funktionen' wie die Dirac-Funktion enthalten sein und Funktionen wie die Heaviside-Funktion $H(x) = \mathbf{1}_{[0,\infty)}(x)$ differenzierbar sein[1]. Darüberhinaus wollen wir die Fouriertransformation nichtintegrierbarer Funktionen wie sin und cos bilden können[2]. Ein solcher Kalkül wird nun schrittweise entwickelt.

4.1 Der Raum $\mathcal{S}$ der schnell fallenden Funktionen.

Wir möchten die technischen Probleme bei Differentiation und Rücktransformation beheben. Z.B. möchten wir beliebig oft differenzieren können und beliebig oft Transformieren können (egal in welcher Richtung). Deshalb ziehen

[1] Wir ahnen es:

$$H' = \text{const} \cdot \delta.$$

[2] Wegen

$$\cos(x) = \frac{1}{2}\left(e^{ix\cdot 1} + e^{ix\cdot(-1)}\right)$$

sollte doch gelten:

$$\cos\hat{}(\lambda) = \text{const} \cdot (\delta(\lambda - 1) + \delta(\lambda + 1)).$$

wir uns auf den Teilraum von $\mathcal{L}^1$ zurück, in dem *alle Operationen problemlos durchführbar sind.*
Tiefere mathematische Überlegungen zeigen, daß dies der Raum $\mathcal{S}$ der *Schwartzschen* oder *schnell fallenden Funktionen* ist. Dazu gehört jede Funktion, die unendlich oft partiell differenzierbar ist und die samt allen möglichen partiellen Ableitungen nach außen, d.h. für $|x| \to \infty$, schneller fällt als jedes $1/|x|^k$.
Um Schreibarbeit zu sparen, führt man ein:

$$D^\alpha = \frac{\partial^{|\alpha|}}{\partial^{\alpha_1}\partial^{\alpha_1}\dots\partial^{\alpha_d}}.$$

Dabei ist $\alpha = (\alpha_1, \dots, \alpha_d)$ mit ganzen Zahlen $\alpha_i \geq 0$ und $|\alpha| = \sum_{i=1}^d \alpha_i$. Der Vektor α gibt also einfach an, wie oft man nach jeder Variablen partiell differenziert. Entsprechend bedeutet

$$x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \cdot \dots \cdot x_d^{\alpha_d}.$$

Die formale Definition lautet:

Definition 4.1 *Der Raum $\mathcal{S}(\mathbb{R}^d)$ besteht aus allen komplexwertigen, im $\mathbb{R}^d$ beliebig oft differenzierbaren Funktionen φ, so daß*

$$|\varphi|_{\alpha,\beta} = \sup_{x\in\mathbb{R}^d} \left|x^\beta D^\alpha \varphi(x)\right| < \infty$$

für alle Multiindizes $\alpha, \beta \in \mathbb{Z}_+$.

Lemma 4.1 *Es gelten:*
(a) $\mathcal{S}$ ist ein linearer Raum (bezüglich der punktweisen Operationen).
(b) Jedes $|\cdot|_{\alpha,\beta}$ ist eine Halbnorm auf $\mathcal{S}$, d.h. es gilt

$$|a\varphi|_{\alpha,\beta} = |a||\varphi|_{\alpha,\beta},\ |\varphi+\psi|_{\alpha,\beta} \leq |\varphi|_{\alpha,\beta} + |\psi|_{\alpha,\beta}.$$

Wir müssen noch festlegen, wann eine Folge in $\mathcal{S}$ konvergieren soll:

Definition 4.2 *Eine Folge (φ_n) in $\mathcal{S}$ konvergiert gegen $\varphi \in \mathcal{S}$, wenn*

$$|\varphi_n - \varphi|_{\alpha,\beta} \longrightarrow 0,\ n \to \infty, \text{ für alle Multiindizes } \alpha \text{ und } \beta.$$

Wir schreiben dafür

$$\varphi_n \xrightarrow{\mathcal{S}} \varphi.$$

Es ist klar, daß jede (gemischte) Ableitung $D^\alpha\varphi$ von $\varphi \in \mathcal{S}$ wieder in $\mathcal{S}$ (und somit in $\mathcal{L}^1$) liegt.
Da $\varphi \in \mathcal{S}$ stärker als jedes Polynom fällt, ist es integrierbar:

Lemma 4.2 *Es gilt:*

$$\mathcal{S}(\mathbb{R}^d) \subset \mathcal{L}^1(\mathbb{R}^d, \mathbb{C}).$$

Insbesondere ist also für jedes $\varphi \in \mathcal{S}$ die Fouriertransformierte $\mathcal{F}\varphi = \hat{\varphi}$ definiert. Der Raum $\mathcal{S}$ verträgt sich ideal damit:

Satz 4.1 *Die Fouriertransformation $\mathcal{F}$ bildet $\mathcal{S}$ linear und eineindeutig auf sich selbst ab; es gilt $\mathcal{F}^{-1} = \bar{\mathcal{F}}$; ferner sind $\mathcal{F}$ und $\mathcal{F}^{-1}$ stetig, d.h. wenn $\varphi_n \xrightarrow{\mathcal{S}} \varphi$, so*

$$\mathcal{F}\varphi_n \xrightarrow{\mathcal{S}} \mathcal{F}\varphi,\ \mathcal{F}^{-1}\varphi_n \xrightarrow{\mathcal{S}} \mathcal{F}^{-1}\varphi.$$

Die Eigenschaften 3.1 bis 3.7 der Fouriertransformation bleiben natürlich erhalten. Mit der Differentiation sieht es jetzt viel freundlicher aus als damals:

Satz 4.2 *Für $\varphi \in \mathcal{S}$ sind* stets *richtig:*

$$\left(\frac{\partial}{\partial x_i}\varphi\right)^{\wedge}(\lambda) = \imath\lambda_i\hat{\varphi}(\lambda),\ \ \imath\frac{\partial}{\partial \lambda_i}\hat{\varphi}(\lambda) = (x_i\varphi)^{\wedge}(\lambda).$$

Allgemeiner formuliert gilt also:

$$(-\imath)^{|\alpha|}\mathcal{F}(D^\alpha\varphi) = \lambda^\alpha\mathcal{F}\varphi,\ \ D^\alpha(\mathcal{F}\varphi) = (-\imath)^{|\alpha|}\mathcal{F}(x^\alpha\varphi).$$

Beweis Wir begründen nur die elementare Formel. Die allgemeine ergibt sich daraus durch Iteration.
Der zweite Teil geht genau wie früher, weil $x_i\varphi(x)$ integrierbar ist. Für den ersten Teil liefert partielle Integration:

$$\begin{aligned}
&\int \frac{\partial}{\partial x_i}\varphi(x)\exp(-\imath\langle x,\lambda\rangle)\,dx_i \\
&= \lim_{\substack{a\to-\infty\\ b\to\infty}} \left(\varphi(x)\exp(-\imath\langle x,\lambda\rangle)|_{x_i=a}^{b} - \int_a^b \varphi(x)\left\{\frac{\partial}{\partial x_i}\exp(-\imath\langle x,\lambda\rangle)\right\}dx_i\right) \\
&= \imath\lambda_i\int \varphi(x)\exp(-\imath\langle x,\lambda\rangle)\,dx_i.
\end{aligned}$$

Jetzt sind die beiden Seiten der Gleichung nur noch über die restlichen Variablen zu integrieren und mit $(2\pi)^{-d/2}$ zu multiplizieren, um die behauptete Gleichung zu verifizieren. □

4.2 Temperierte Distributionen

Natürlich können wir unsere hohen Ansprüche an den Kalkül (oder unsere Bequemlichkeit) mit gewöhnlichen Funktionen nicht befriedigen. Es gibt einfach keine Funktion, welche die Ableitung der Heaviside-Funktion ist. Aber: sind wir denn wirklich an Funktionen als solchen interessiert? Betrachten wir die Wirkung eines Filters mit Punktantwort h auf ein Signal $f \in \mathcal{S}$:

$$(f * h)(x) = \int f(y)h(x-y)\,dx = \int f(y)h_x(y)\,dy,$$

wobei $h_x(y) = h(x-y)$. Es interessiert also nur, wie h_x auf f einwirkt, wenn man die Funktionen erst multipliziert und anschließend integriert. Setzen wir $g = h_x$, so interessiert also nur die Funktion

$$T_g : \mathcal{S} \longrightarrow \mathbb{C}, f \longmapsto \langle f, g\rangle = \int fg\,dy. \tag{4.1}$$

So etwas nennt man *Funktional* und weil

$$T_g(af + b\tilde{f}) = aT_g(f) + bT_g(\tilde{f})$$

gilt, ein *lineares Funktional.* Jeder vernünftigen Funktion g (siehe[3]) entspricht ein lineares Funktional T_g wie in (4.1). Wir kennen aber schon andere Funktionale:

$$T_\delta : \mathcal{S} \longrightarrow \mathbb{C}, f \longmapsto f(0),$$

oder allgemeiner

$$T_{\delta_a} : \mathcal{S} \longrightarrow \mathbb{C}, f \longmapsto f(a).$$

Solche Funktionale kennen wir, wenn wir sie auf ihre Wirkung auf Funktionen testen. Wir definieren:

Definition 4.3 *Ein stetiges lineares Funktional*

$$T : \mathcal{S} \longrightarrow \mathbb{C}$$

heißt verallgemeinerte Funktion *oder* (temperierte) Distribution. *Der Raum dieser Funktionale wird mit $\mathcal{S}'$ bezeichnet. Eine Funktion $\varphi \in \mathcal{S}$ nennen wir* Testfunktion.

[3] 'vernünftig' heißt, daß

$$\int |g(x)|(1+|x|)^q\,dx < \infty$$

für ein q. Insbesondere reicht es, daß g integierbar ist.

Wir akzeptieren unbewiesen:

Satz 4.3 *Für jedes $g \in \mathcal{L}^1$ (oder mit der Integierbarkeitseigenschaft unten) ist die Abbildung*

$$T_g : \mathcal{S} \longrightarrow \mathbb{C}, \varphi \longmapsto T_g\varphi = \int \varphi g \, dx$$

ein stetiges lineares Funktional auf $\mathcal{S}$, d.h. eine verallgemeinerte Funktion.

Jede 'vernünftige' Funktion ist also eine verallgemeinerte Funktion. Eine verallgemeinerte Funktion T_h, die sich mittels einer Funktion h repräsentieren läßt, nennt man *regulär* und bezeichnet sie einfach wieder mit h.
Jetzt wollen wir natürlich alle wichtigen Operationen auf den verallgemeinerten Funktionen ausführen.

Differentiation.

Wir beginnen mit der Differentiation.
Für $h, \varphi \in \mathcal{S}(\mathbb{R})$ gilt vermöge partieller Integration:

$$\begin{aligned} \langle h', \varphi \rangle &= \int h'\varphi \, dx = \lim_{a \to -\infty, b \to \infty} h\varphi|_a^b - \lim_{a \to -\infty, b \to \infty} \int_a^b h\varphi' \, dx \\ &= -\int h\varphi \, dx = \langle h, -\varphi' \rangle. \end{aligned}$$

Betrachten wir also nur die Wirkung von h' auf φ, so kommt es auf dasselbe heraus wie die Wirkung von h auf φ'. Wir *definieren* also für eine verallgemeinerte Funktion T die Ableitung T' als lineares Funktional durch

$$\langle T', \varphi \rangle = -\langle T, \varphi' \rangle.$$

So schieben wir alle Operationen auf die Testfunktionen ab.

Beispiel 4.1 Wir berechnen die Ableitung der Sprungfunktion $h(x) = \mathbf{1}_{[0,\infty)}$:

$$\begin{aligned} \langle T_h', \varphi \rangle &= \langle T_h, -\varphi' \rangle = \langle h, -\varphi' \rangle = -\int_0^\infty \varphi'(y) \, dy = -(\varphi(\infty) - \varphi(0)) \\ &= -(0 - \varphi(0) = \varphi(0) = \langle T_\delta, \varphi \rangle \left(=: \int \delta(y)\varphi(y) \, dy \right). \end{aligned}$$

Also ist die Dirac-Funktion die Ableitung der Sprungfunktion, was uns natürlich sehr gelegen kommt.

Ist h stückweise differenzierbar mit einem Sprung in z, so ist die Ableitung gleich $f'(x)$ in $x \neq z$ und $\pm\delta(x-z) = \pm\delta_z(x)$ in z.
Durch Iteration ergibt sich

$$\langle D^\alpha T, \varphi\rangle = (-1)^{|\alpha|}\langle T, D^\alpha\varphi\rangle.$$

Translation

Sei wieder $\tau_a\varphi(x) = \varphi(x-a)$. Wir definieren:

$$\langle \tau_a T, \varphi\rangle = \langle T, \tau_{-a}\varphi\rangle.$$

Beispiel 4.2 Wir prüfen, ob für Funktionen das Richtige herauskommt:

$$\langle \tau_a T_h, \varphi\rangle = \langle T_h, \tau_{-a}\varphi\rangle = \int h(y)\varphi(y+a)\,dy = \int h(y-a)\varphi(y)\,dy = \langle T_{\tau_a h}, \varphi\rangle.$$

Also haben wir den Shift richtig definiert.

Multiplikation mit Funktionen

Sei f beschränkt. Wir definieren fT durch

$$\langle fT, \varphi\rangle = \langle T, f\cdot\varphi\rangle.$$

Beispiel 4.3 Sei $f \in \mathcal{L}^1$. Dann gilt:

$$\langle fT_h, \varphi\rangle = \langle T_h, f\varphi\rangle = \int h(f\varphi)\,dx = \int (hf)\varphi\,dx = \langle T_{hf}, \varphi\rangle.$$

4.3 Die Fouriertransformation auf $\mathcal{S}'$

Sie ist besonders wichtig und bekommt deshalb einen eigenen Unterabschnitt.
Es gilt:

Satz 4.4 *Für $f, g \in \mathcal{L}^1$ gilt*

$$\int f\hat{g}\,dx = \int \hat{f}g\,dx.$$

Beweis Weil $\hat{f}$ und $\hat{g}$ beschränkt sind, sind $f\hat{g}$ und $\hat{f}g$ integrierbar und nach dem Satz von Fubini gilt:

$$\begin{aligned}\int f\hat{g}\,dx &= c\int f(x)\left\{\int g(y)\exp(-\imath\langle y,x\rangle)\,dy\right\}dx\\ &= c\int\int\{f(x)g(y)\exp(-\imath\langle y,x\rangle)\}\,dydx\\ &= c\int\int\{f(x)g(y)\exp(-\imath\langle y,x\rangle)\}\,dxdy\\ &= c\int\left\{\int f(x)\exp(-\imath\langle y,x\rangle)\,dx\right\}g(y)\,dy = \int\hat{f}g\,dy.\end{aligned}$$

□

Beispiel 4.4 Sei $h\in\mathcal{L}^1$. Dann gilt

$$\langle T_{\hat{h}},\varphi\rangle = \langle\hat{h},\varphi\rangle = \int\hat{h}\varphi\,dx = \int h\hat{\varphi}\,dx = \langle h,\hat{\varphi}\rangle = \langle T_h,\hat{\varphi}\rangle.$$

Für $T\in\mathcal{S}'$ definieren wir also:

$$\langle\hat{T},\varphi\rangle = \langle T,\hat{\varphi}\rangle, \mathcal{F}T = \hat{T}.$$

Die tiefere Rechtfertigung für dieses Vorgehen liefert folgender Satz:

Satz 4.5 *Die Fouriertransformation $\mathcal{F}:\mathcal{S}'\to\mathcal{S}'$ ist linear und bijektiv. Außerdem sind $\mathcal{F}$ und $\mathcal{F}^{-1}$ stetig, d.h.*

$$\varphi_n\xrightarrow{\mathcal{S}}\varphi \implies \mathcal{F}\varphi_n\longrightarrow\mathcal{F}\varphi,\ \mathcal{F}^{-1}\varphi_n\longrightarrow\mathcal{F}^{-1}\varphi.$$

Diese Fouriertransformation ist also nichts anderes, als eine Ausdehnung der üblichen Fouriertransformation auf verallgemeinerte Funktionen.
Um Vertrauen in die Richtigkeit dieser Begriffsbildung zu fassen, betrachten wir einige Beispiele:

Beispiel 4.5 Sei $\delta_a(x)=\delta(x-a)$. Dann gilt

$$\langle\hat{\delta}_a,\varphi\rangle = \langle\delta_a,\hat{\varphi}\rangle = \hat{\varphi}(a).$$

Kann man $\hat{\delta}_a$ als Wirkung einer Funktion darstellen? Wir beginnen mit dem einfachsten Fall:

$$\begin{aligned}\langle\hat{\delta},\varphi\rangle &= \hat{\varphi}(0) = (2\pi)^{-d/2}\int\varphi(x)\exp(-\imath\langle x,0\rangle)\,dx\\ &= (2\pi)^{-d/2}\int\varphi\cdot 1\,dx = \langle(2\pi)^{-d/2},\varphi\rangle.\end{aligned}$$

Also ist

$$\hat{\delta} \equiv (2\pi)^{-d/2}$$

Intuitiv ist δ_a ein verschoben es δ, nämlich $\tau_a\delta$, was natürlich stimmt:

$$\langle \delta_a, \varphi \rangle = \varphi(a) = \varphi(0+a) = \langle \delta, \tau_{-a}\varphi \rangle = \langle \tau_a\delta, \varphi \rangle.$$

Nach der Formel für den Shift gilt also:

$$\begin{aligned} \langle \hat{\delta}_a, \varphi \rangle &= \langle (\tau_a\delta)^\wedge, \varphi \rangle = \langle \tau_a\delta, \hat{\varphi} \rangle = \langle \delta, \tau_{-a}\hat{\varphi} \rangle \\ &= c \cdot \int \varphi(x) \exp(-\imath\langle x, 0+a \rangle)\, dx = \langle c \cdot \exp(-\imath\langle \cdot\,, a \rangle), \varphi \rangle. \end{aligned}$$

Also wird $\hat{\delta}_a$ durch die Funktion $\lambda \mapsto (2\pi)^{-d/2} \exp(-\imath\langle \lambda, a \rangle)$ repräsentiert. Diese Funktion ist nicht integrierbar, aber als verallgemeinerte Funktion eine Fouriertransformierte.

Wir betrachten nun Wellen als Signale.

Beispiel 4.6 (a) Wenn wir uns erinnern, daß $\hat{f}$ die Gewichtung der elementaren 'Wellen' $f_\lambda(x) = \exp(\imath\langle x, \lambda \rangle)$ bei der Darstellung von f war, müßte eigentlich $\hat{f}_\lambda$ ein scharfer Impuls in λ sein. So ist es auch:

$$\langle \hat{f}_\lambda, \varphi \rangle = \langle f, \hat{\varphi} \rangle = \int \exp(\imath\langle x, \lambda \rangle)\hat{\varphi}(x)\, dx = (2\pi)^{d/2}\varphi(\lambda) = (2\pi)^{d/2}\langle \delta_\lambda, \varphi \rangle,$$

wobei beim dritten Gleichheitszeichen die Umkehrformel benutzt wurde. Dies bedeutet

$$\hat{f}_\lambda = (2\pi)^{d/2}\delta_\lambda,$$

was (bis auf die lästige Konstante) genau das erwartete Resultat ist.
(b) Jetzt können wir auch die nicht integrierbaren harmonischen Schwingungen transformieren:
Wegen

$$\cos(\alpha x) = \frac{1}{2}\left(\exp(\imath x\alpha) + \exp(-\imath x\alpha)\right)$$

gilt

$$g(x) = \cos(\alpha x);\ \hat{g}(\lambda) = \frac{(2\pi)^{d/2}}{2}(\delta_\alpha(\lambda) + \delta_{-\alpha}(\lambda)).$$

Analog gilt:

$$g(x) = \sin(\alpha x);\ \hat{g}(\lambda) = \frac{(2\pi)^{d/2}}{2\imath}(\delta_\alpha(\lambda) - \delta_{-\alpha}(\lambda)).$$

Zusammenfassend haben wir folgende Transformationspaare $f \leftrightarrow \hat{f}$ gefunden:

$$\begin{aligned}
\delta_a(x) &\leftrightarrow (2\pi)^{-d/2}\exp(-\imath\langle\lambda, a\rangle) \\
\exp(\imath\langle x,\xi\rangle) &\leftrightarrow (2\pi)^{d/2}\delta_\xi(\lambda) \\
\cos(\alpha x) &\leftrightarrow \frac{(2\pi)^{d/2}}{2}\left(\delta_\alpha(\lambda)+\delta_{-\alpha}(\lambda)\right) \\
\sin(\alpha x) &\leftrightarrow \frac{(2\pi)^{d/2}}{2\imath}\left(\delta_\alpha(\lambda)-\delta_{-\alpha}(\lambda)\right)
\end{aligned}$$

4.4 Mehr zum scharfen Impuls

Die intuitive Idee bei der Delta- oder Diracfunktion war, daß sie einen idealen scharfen Impuls modelliert. Also sollte sie ein Grenzfall von zwar unscharfen aber immer mehr um Null konzentrierten Impulsen sein. Dies präzisiert folgende Aussage:

Satz 4.6 *Sei f integrierbar mit $\int f\,dx = 1$. Seien*

$$f_\varepsilon(x) = \frac{1}{\varepsilon^d} f\left(\frac{x}{\varepsilon}\right),\ \varepsilon > 0.$$

Dann gilt für jedes $\varphi \in \mathcal{S}$, daß

$$\int \varphi f_\varepsilon\,dx \longrightarrow \varphi(0)\left(= \int \varphi\delta\,dx\right),\ \varepsilon \to 0.$$

Dies ist äquivalent zu

$$T_{f_\varepsilon} \xrightarrow{\mathcal{S}'} \delta.$$

Beweis Mit der Substitution $x = \varepsilon y$ erhalten wir:

$$\int \varphi(x) f_\varepsilon(x)\,dx = \int \varphi(\varepsilon x)\frac{1}{\varepsilon^d} f\left(\frac{x}{\varepsilon}\right)dx = \int f(y)\varphi(\varepsilon x)\,dx.$$

Dann gelten

$$|\varphi(\varepsilon y)f(y)| \le \sup_z |\varphi(z)||f(y)| \text{ für alle } \varepsilon > 0 \text{ und } y \in \mathbb{R}^d \qquad (4.2)$$

und

$$\varphi(\varepsilon y)f(y) \longrightarrow \varphi(0)f(y),\ \varepsilon \to 0.$$

Da (4.2) eine integrierbare Majorante für alle Funktionen $y \mapsto \varphi(\varepsilon y) f(y)$ liefert, gilt nach dem Satz von der majorisierten Konvergenz, daß

$$\int \varphi(y) f_\varepsilon(y)\, dy \longrightarrow \int \varphi(0) f(y)\, dy = \varphi(0) \int f\, dy = \varphi(0).$$

□

Beispiel 4.7 Man denkt hier vor allem an Funktionen wie

$$f_\varepsilon(y) = \frac{1}{(2\pi)^{d/2}\varepsilon^d} \exp\left(-\frac{|y|^2}{2\varepsilon^2}\right).$$

Kapitel 5

Die Radontransformation

In der Einleitung wurde dargestellt, daß es bei den tomographischen Verfahren darauf ankommt, eine Verteilung – z.B. den ortsabhängigen Schwächungskoeffizienten der Strahlung – aus den Integralen über Linien zu rekonstruieren. Für den Schwächungskoeffizienten μ etwa waren die Integrale gegeben durch

$$\int_L \mu(s)\,ds = -\ln \frac{I_d}{I_0},$$

wobei L eine Gerade bezeichnete, $\mu(s)$ den Schwächungskoeffizienten an dem durch s parametrisierten Ort auf der Geraden L, $I(s)$ die Strahlungsintensität an diesem Ort, und I_0 die eingesetzte Strahlung.
Bevor wir Verfahren zur Rekonstruktion studieren, müssen wir erst klären, was mit diesen Linienintegralen gemeint ist und nützliche Eigenschaften der Gesamtheit dieser Linienintegrale ableiten. Letzteres geschieht im nächsten Abschnitt, ersteres wird der Gegenstand der weiteren Abschnitte sein.
Es ist vom mathematischen Standpunkt aus sinnvoll, für die zu rekonstruierenden Dichten die Annahme $f \in \mathcal{S}$ zu machen (obwohl die zu untersuchenden Objekte – etwa eine Niere – natürlich durch Funktionen mit kompaktem Träger beschreibbar sind). Ist L eine Gerade im $\mathbb{R}^2$, so definieren wir intuitiv für $f \in \mathcal{S}$ die *Radontransformierte* $\mathcal{R}f$ durch

$$\mathcal{R}f(L) = \int_L f(x)\,dx.$$

In höherer Dimension wird die Gerade durch eine Hyperebene ersetzt.

5.1 Definitionen

Natürlich müssen wir uns erst klar machen, was diese Integrale bedeuten.

5.1.1 Die Radontransformation in der Ebene

Wir betrachten zunächst den Fall $\mathbb{R}^2$. Wir können eine Linie im $\mathbb{R}^2$ z.B. durch die Gleichung $x_2 = ax_1 + b$ darstellen. Geeigneter für unsere Zwecke ist die folgende Beschreibung: Wir beschreiben die Gerade durch

$$x = p\xi + t\xi^1, t \in \mathbb{R},$$

wobei $\xi \in \mathbb{R}^2$ mit $|\xi| = 1$ ein fester (Einheits-) Vektor ist, $p \in \mathbb{R}$ der (vorzeichenbehaftete) Abstand der Geraden (in Richtung ξ) vom Nullpunkt und ξ^1 der auf ξ senkrecht stehende Einheitsvektor. Das Integral von f über L definieren wir als

$$\int_L f\,dx = \int f(p\xi + t\xi^1)\,dt\,(= \mathcal{R}f(p,\xi)),$$

wobei die Gerade L durch den Normalenvektor ξ und den Abstand p zum Ursprung repräsentiert wird. Bezeichnet $\xi^\perp$ den (1-dimensionalen) linearen Unterraum von $\mathbb{R}^2$ senkrecht zu ξ, so können wir auch schreiben:

$$\mathcal{R}f(p,\xi) = \int_{\xi^\perp} f(p\xi + x)\,dx.$$

Nun gilt nach Satz 4.6 für den Integranden, daß (wobei wir f_ε wie in Beispiel 4.7 wählen):

$$\int f(\rho\xi + t\xi^1) f_\varepsilon(p-\rho)\,d\rho \longrightarrow f(p\xi + t\xi^1) = \int f(\rho\xi + t\xi^1)\delta(p-\rho)\,d\rho,\ \varepsilon \to 0. \tag{5.1}$$

Wir führen eine (orthonormale) Variablentransformation gemäß

$$x(\rho,t) = \rho\xi + t\xi^1 = (\xi,\xi^1)\begin{pmatrix}\rho\\ t\end{pmatrix} = A\begin{pmatrix}\rho\\ t\end{pmatrix}$$

durch, wobei ξ, ξ^1 als Spaltenvektoren aufgefaßt werden. Bezeichnet J die Jakobimatrix dieser Transformation, so gilt $J(x) = A$ für alle x und somit $|\det J(x)| = |\det A| = 1$. Außerdem gilt

$$\begin{pmatrix}\rho\\ t\end{pmatrix} = A^{-1}x = A^T x = \begin{pmatrix}\xi^T\\ (\xi^1)^T\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix} = \langle x,\xi\rangle.$$

Der Integraltransformationssatz [1]. liefert somit:

$$\int\int f(\rho\xi + t\xi^1) f_\varepsilon(p-\rho)\,d\rho\,dt = \int f(x) f_\varepsilon(p - \langle x,\xi\rangle)\,dx.$$

[1]Noch mal:
Die Transformationsformel: Seien $D_1, D_2 \subset \mathbb{R}^d$ offen und $\varphi : D_1 \longrightarrow D_2$ bijektiv,

Wegen (5.1) konvergiert die linke Seite gegen

$$\int\int f(\underbrace{\rho\xi + t\xi^1}_{x})\delta(p - \underbrace{\rho}_{\langle x,\xi\rangle})\underbrace{d\rho dt}_{dx} = \int f(p\xi + t\xi^1)\,dt = \mathcal{R}f(p,\xi)$$

Die rechte Seite konvergiert ebenfalls dagegen und es ergibt sich

$$\mathcal{R}f(p,\xi) = \int f(x)\delta(p - \langle x,\xi\rangle)\,dx.$$

Dies ist nicht nur die Ingenieurschreibweise, sondern ermöglicht manchmal übersichtliches Rechnen.

5.1.2 Die Radontransformation im $\mathbb{R}^d$.

Wir gehen genau wie im eindimensionalen vor. Hier sei L eine Hyperebene, beschrieben z.B. durch

$$x(p,t_1,\ldots,t_{d-1}) = p\xi + t_1\xi^1 + \cdots + t_{d-1}\xi^{d-1}$$

mit einen Orthonormalsystem $\{\xi,\xi^1,\ldots,\xi^{d-1}\}$, festem $p \in \mathbb{R}$ und Variablen $t_i \in \mathbb{R}$. Bezeichnet wieder

$$\xi^\perp = \{t_1\xi^1 + \cdots + t_{d-1}\xi^{d-1} : t_i \in \mathbb{R}\}$$

den $(d-1)$-dimensionalen linearen Unterraum von $\mathbb{R}^d$ senkrecht zu ξ, so nimmt die Radontransformation folgende Gestalt an:

$$\mathcal{R}f(p,\xi) = \int_{R^{d-1}} f(p\xi + t_1\xi^1 + \cdots + t_{d-1}\xi^{d-1})\,dt = \int_{\xi^\perp} f(p\xi + x)\,dx.$$

Wie in (5.1) gilt die Approximation

$$\begin{aligned}\int\int f(\rho\xi + t_1\xi^1 + \cdots + t_{d-1}\xi^{d-1})f_\varepsilon(p-\rho)\,d\rho dt \\ \longrightarrow \int \quad f(p\xi + t_1\xi^1 + \cdots + t_{d-1}\xi^{d-1})\,dt \qquad (5.2)\\ = \quad \int f(\underbrace{\rho\xi + t_1\xi^1 + \cdots + t_{d-1}\xi^{d-1}}_{x})\delta(p - \underbrace{\rho}_{\langle x,\xi\rangle})\underbrace{d\rho dt}_{dx},\ \varepsilon \to 0.\end{aligned}$$

---

sowie φ und φ^{-1} stetig differenzierbar und für die Jakobimatrix J der ersten partiellen Ableitungen von φ sei $|\det J(x)| > 0$ auf D_1. Dann gilt

$$\int_{D_2} f(x)\,dx = \int_{D_1} f(\varphi(x))|\det J(x)|\,dx.$$

Formal können wir die nachfolgende Approximation und die Variablentransformation übernehmen, wobei

$$A = (\xi, \xi^1, \ldots, \xi^{d-1})$$

ist. Dies führt wieder zur Formel

$$\mathcal{R}f(p,\xi) = \int_{\xi^\perp} f(p\xi + x)\, dx = \int f(x)\delta(p - \langle x, \xi\rangle)\, dx.$$

5.2 Elementare Eigenschaften

Die Radontransformation hat folgende elementare Eigenschaften:

Satz 5.1 (Symmetrie) *Es gilt*

$$\mathcal{R}f(-p,-\xi) = \mathcal{R}f(p,\xi).$$

Beweis Dies ist offensichtlich, da $-p, -\xi$ und p, ξ dieselbe Gerade repräsentieren. □

Satz 5.2 (Linearität) *die Radontransformation ist linear, d.h.*

$$\mathcal{R}(af + bg) = a\mathcal{R}f + b\mathcal{R}g.$$

Beweis Das ist klar. □

Satz 5.3 (Streckung) *Seien $a \in \mathbb{R}$ und $g(x) = f(ax)$. Dann gilt*

$$\mathcal{R}g(p,\xi) = \frac{1}{a^{d-1}}\mathcal{R}f(ap,\xi).$$

Beweis Nach dem Integraltransformationssatz gilt

$$\begin{aligned}\mathcal{R}g(a,\xi) &= \int f(ap + a(t_1\xi^1 + \cdots + t_{d-1}\xi^{d-1}))\, dt \\ &= \frac{1}{|a|^{d-1}} \int f(ap + t_1\xi^1 + \cdots + t_{d-1}\xi^{d-1})\, dt = \frac{1}{|a|^{d-1}}\mathcal{R}(ap,\xi).\end{aligned}$$

□

Satz 5.4 (Drehung) *Für jede orthonormale Matrix A gilt eine .*

$$\mathcal{R}f \circ A(p,\xi) = \mathcal{R}(p, A^{-1}\xi).$$

Beweis Aus einer Zeichnung liest man das für $g(x) = f(Ax)$ sofort ab. □

Satz 5.5 (Translation) *Sei $g(x) = f(x - z)$. Dann gilt*

$$\mathcal{R}g(p,\xi) = \mathcal{R}(p - \langle z,\xi\rangle, \xi).$$

Beweis Das macht man sich ebenfalls sofort an einer Zeichnung klar. □

5.3 Elementare Beispiele.

Um eine bessere Vorstellung zu bekommen, berechnen wir die Transformierten einiger einfacher Funktionen:

Beispiel 5.1 Sei

$$f(x) = \exp(-|x|^2),\ x \in \mathbb{R}^2.$$

Mit der orthonormalen Transformation

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} \xi_1 & \xi_2 \\ -\xi_2 & \xi_1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

wobei $\xi^T = (\xi_1, \xi_2)$, gilt:

$$\begin{aligned} \mathcal{R}f(p,\xi) &= \int \exp(-(x_1^2 + x_2^2))\delta(p - \langle x,\xi\rangle)\, dx_1 dx_2 \\ &= \int \exp(-(u_1^2 + u_2^2))\delta(p - u_1)\, du_1 du_2 \\ &= \exp(-p^2) \int \exp(-u_2^2)\, du_2 = \sqrt{\pi}\exp(-p^2). \end{aligned}$$

In höherer Dimension führt man eine ähnliche Drehung aus (so daß $\langle x,\xi\rangle$ durch u_1 ersetzt wird) und erhält

$$\mathcal{R}\exp(-|\cdot|^2)(p,\xi) = (2\pi)^{(d-1)/2}\exp(-p^2).$$

Um die Streckungsformel zu illustrieren setzen wir dort $a = \sqrt{\pi}$ und erhalten

$$\mathcal{R}\exp(-\pi|\cdot|^2)(p,\xi) = \exp(-\pi p^2).$$

Schließlich betrachten wir

$$g(x) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{|x-a|^2}{2\sigma^2}\right).$$

Wir strecken also um $1/\sqrt{2\sigma^2}$ und translatieren um $a \in \mathbb{R}^2$. Dies ergibt

$$\mathcal{R}g(p,\xi) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(p-\langle a,\xi\rangle)^2}{2\sigma^2}\right). \tag{5.3}$$

Was ist die Radontransformation einer in einem Punkt konzentrierten Massenverteilung?

Beispiel 5.2 Geht man in (5.3) zum Limes $\sigma \to 0$ über, so ergibt sich

$$\mathcal{R}\delta_a(p,\xi) = \delta_{\langle a,\xi\rangle}(p). \tag{5.4}$$

Dies ist also ein 'scharfer Impuls konzentriert auf der Menge'

$$K = \{(p,\xi) \in \mathbb{R} \times S^1 : \langle a,\xi\rangle = p\}.$$

Bei der bildlichen Darstellung dieser Menge kann man in eine üble Falle tappen (die in der Ingenieurliteratur of in zweifelhafter Weise behandelt wird, z.B. im Buch von DEANS): Identifiziert man (p,ξ) mit dem Punkt $x = p\xi$ in der (x_1,x_2)-Ebene, so wird die K entsprechende Menge

$$\tilde{K} = \{p\xi : p \in \mathbb{R}, \xi \in S^1, \langle a,\xi\rangle = p\}$$

ein Kreis mit Mittelpunkt $m = a/2$ durch den Ursprung, d.h. mit Radius $r = |a|/2$.

Beweis Gegeben ξ ist der Vektor $x = (p,\xi)$, welcher (5.4) erfüllt, die Projektion von a auf die durch ξ bestimmte Gerade. Dies bedeutet, daß das Dreieck mit Basis $\overline{0,a}$ und Scheitel x bei x einen rechten Winkel hat. Die Scheitel aller solcher Dreiecke bilden bekanntlich den *Thaleskreis*, d.h. den Kreis um m mit dem Durchmesser $|a|$. Letzteres ist sehr einfach zu sehen: Für die in der Skizze Abb. 5.1 bezeichneten Winkel α,β,γ,ϕ gilt in Grad:

$$\begin{aligned} 180-\gamma &= 2\phi \\ 180-\beta &= 2\alpha \\ \beta+\gamma-180 &= 0. \end{aligned}$$

Addition der drei Gleichungen ergibt

$$\phi + \alpha = 90,$$

und das ist schon die Behauptung. □

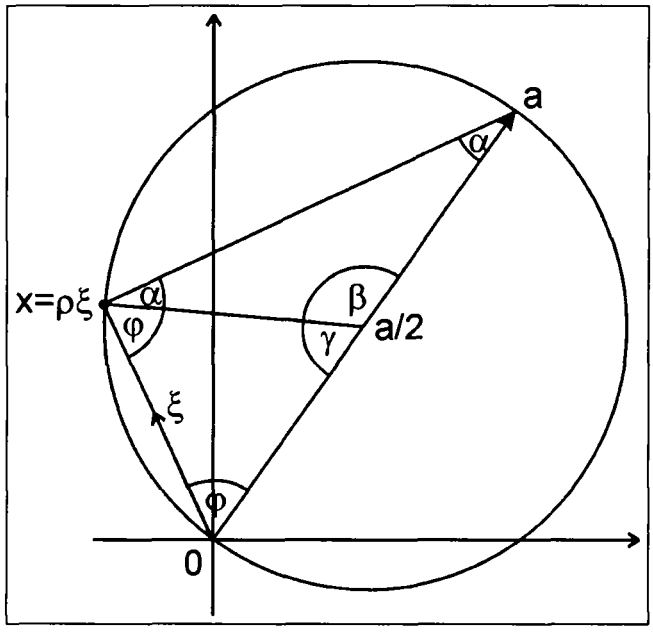


Abbildung 5.1: Zum Beweis

Dieses Vorgehen *ist falsch*, da alle Punkte $(0, \xi)$ auf $x = 0\xi = 0$ abgebildet werden; m.a.W. man hat keine zulässige Parametrisierung von $\mathbb{R} \times S^1$. Dagegen ist die Parametrisierung

$$(p, \xi) \leftrightarrow (p, \phi), \phi = \arccos\left(\left\langle \xi, \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\rangle\right), \phi \in [0, \pi)$$

mit dem Winkel ϕ zwischen ξ und der x_1-Achse sinnvoll. Der Menge K entspricht dann die Menge

$$\hat{K} = \{(p, \phi) : p \in \mathbb{R}, 0 \leq \phi < 2\pi,\ p = a_1 \cos\phi + a_2 \sin\phi\}.$$

Jeder Punktmasse δ_a entspricht also in der (p, ϕ)-Ebene die sinoide Kurve

$$p = a_1 \cos\phi + a_2 \sin\phi,\ \ 0 \leq \phi < 2\pi.$$

Jeder Geraden in der (x_1, x_2)-Ebene entspricht ein Punkt in der (p, ϕ)-Ebene.

5.4 Die Hough-Transformation

Sie wurde 1962 von P.V.C. HOUGH als Methode zur Geradendetektion in zweidimensionalen Bildern vorgeschlagen. Sie leitet sich in einfacher Weise aus den Beobachtungen in Beispiel 5.2 ab.

Seien a und b Punkte in der Ebene und L eine Gerade, parametrisiert durch (q, ψ). Liegen a und b auf L, so schneiden sich also die beiden Sinoiden

$$p = a_1 \cos\phi + a_2 \sin\phi, \quad p = b_1 \cos\phi + b_2 \sin\phi$$

im Punkt (q, ψ). Liegen im Originalbild viele Punkte auf dieser Geraden, so schneiden sich alle zugehörigen Sinoide in (q, ψ). Würde man die Sinoiden durch eine Intensitätsstufe markieren und Schnittpunkten die entsprechend aufsummierten Intensitäten zuordnen, so erschiene in (q, ψ) ein heller Fleck. Man könnte diesen durch Thresholding selektieren und hätte die Gerade im Ursprungsbild detektiert. In der Praxis unterteilt man die (p, ϕ)-Ebene in sogenannte Akkumulatorzellen und schreibt in jede die Anzahl der Sinoiden, welche die Zelle trifft. Diese Häufigkeitsverteilung nennt man *Hough-Transformation*, siehe z.B. R.C. Gonzalez, P. Wintz (1987), [7].
Der Inhalt jeder Zelle gibt an, wieviele Punkte im Ursprungsbild approximativ auf der entsprechenden Geraden liegen. Man setzt dann eine Schranke und nimmt alle Geraden, auf denen mehr Punkte liegen.
Meist liegen im Bild aber nicht Geraden, sondern nur Geradenstücke vor. Diese versucht man anschließend durch ein 'Tracking-Verfahren' zu finden, vgl. z.B. [1].

5.5 Verwandte Transformationen

Man betrachtet auch die *Röntgenstrahl-Transformation*

$$\mathcal{P}(\eta, x) = \int_{-\infty}^{\infty} f(x + t\eta)\, dt, \quad \eta \in S^{d-1}, x \in \mathbb{R}^d,$$

also (auch in höherer Dimension) das Integral von $f \in \mathcal{S}$ entlang der Geraden durch x mit Richtung η. Offensichtlich ändert man diese Integrale nicht, wenn x auf der Geraden bewegt wird und man schränkt sich meist auf $\eta^\perp$ ein. Für $d = 2$ stimmt diese Transformation mit der Radontransformation überein.
Schließlich ist auch noch die *divergent beam*-Transformation von Interesse (in 2 Dimensionen auch *fan*[2]*-beam*-Transformation genannt):

$$\mathcal{D}f(x, \eta) = \int_0^{\infty} f(x + t\eta)\, dt,$$

bei der nur über den Strahl von x aus in Richtung η integriert wird.
Wir werden die Beweise i.a. nur für die Radontransformation führen und die Ergebnisse für die übrigen Transformationen nur vermerken, wo dies sinnvoll ist.

[2]Fächer, Ventilator

Kapitel 6

Der Projektionssatz und Konsequenzen

Wir stellen eine wichtige Verbindung zwischen Fourier- und Radontransformation her und leiten Rechenregeln ab.

6.1 Der Projektionssatz oder das Fourier-Slice-Theorem

Für Funktionen $h(\rho, \eta)$, $\rho \in \mathbb{R}, \eta \in \mathbb{R}^d$ sei

$$h_\eta : \mathbb{R} \longrightarrow \mathbb{R}, \rho \longmapsto h(\rho, \eta).$$

Wir schreiben

$$\hat{h}_\eta = (\mathcal{F}_1 h = \hat{h})$$

für die Fouriertransformation bei festgehaltenem η. Wir führen also eine gewöhnliche Fouriertransformation einer Funktion einer Variablen ρ durch. Die Notation in Klammern hat sich ebenfalls eingebürgert. Der Hauptsatz lautet:

Satz 6.1 *Für $f \in \mathcal{S}(\mathbb{R}^d)$ gilt:*

$$(\mathcal{R}_\xi f)\hat{}(\sigma) = (2\pi)^{(d-1)/2} \hat{f}(\sigma\xi), \sigma \in \mathbb{R}.$$

In Worten: 'Die Radontransformierte ergibt sich aus f durch eine Fouriertransformation gefolgt von einer inversen Fouriertransformation in radialer Richtung.' Der Name erklärt sich wie folgt:

Bemerkung 6.1 Aufgrund der Eineindeutigkeit der Fouriertransformation enthält die linke Seite für festes ξ (und für alle σ) die selbe Information wie die *Projektion* von f senkrecht zu ξ, die sich ja aus den Werten $\mathcal{R}f(p,\xi), p \in \mathbb{R}$, zusammensetzt. Die rechte Seite enthält die Werte von $\hat{f} = \mathcal{F}f$ entlang des *Schnittes* ('slice') $\{p\xi : p \in \mathbb{R}\}$. Der Satz sagt also, daß die Kenntnis der Projektion gleichwertig ist mit der Kenntnis des Schnittes (der Fouriertransformation)

Der Satz hat weitere Konsequenzen:

Bemerkung 6.2 Man schreibt dies oft in der Form

$$\mathcal{F}_1\mathcal{R}f = (2\pi)^{(d-1)/2}\mathcal{F}f.$$

Anwendung der eindimensionalen Umkehrtransformation $\mathcal{F}_1^{-1}$ liefert:

$$\mathcal{R}f(\rho,\xi) = (2\pi)^{(d-1)/2}(2\pi)^{-1}\int \hat{f}(s\xi)\exp(\imath s\rho)\,ds.$$

Somit haben wir bereits eine Methode zur Rekonstruktion der Funktion aus seiner Radontransformation. Insbesondere ist f durch $\mathcal{R}f$ eindeutig bestimmt. Der Teufel liegt im Detail bei der praktischen Umsetzung mittels schneller Fouriertransformation, genauer in der zugehörigen Diskretisierung. Dies führt zu schwer beseitigbaren Artefakten und die *direkte Fourierrekonstruktion* ist anderen Verfahren weit unterlegen. Mehr dazu lernen wir später. Der Satz wurde 1956 von BRACEWELL im Zusammenhang mit der Radioastronomie bewiesen. Er war allerdings in Wahrscheinlichkeitstheorie und Differentialgleichungen schon vorher bekannt (1934, 1936). Er konnte nicht zur Rekonstruktion eingesetzt werden, weil damals die numerischen Methoden zur Inversion der rechten Seite fehlten. Sie wurden erst später entwickelt.

Der Beweis ist leicht zu führen:

Beweis Es gilt

$$\begin{aligned}(\mathcal{R}_\xi f)^\wedge(\sigma) &= \frac{1}{\sqrt{2\pi}}\int_R \exp(-\imath\sigma s)\int_{R^{d-1}} f(s\xi + t_1\xi^1 + \cdots + t_{d-1}\xi^{d-1})\,dt\,ds \\ &= \frac{1}{\sqrt{2\pi}}\int_R\int_{R^{d-1}} \exp(-\imath\sigma s) f(s\xi + t_1\xi^1 + \cdots t_{d-1}\xi^{d-1})\,dt\,ds\end{aligned}$$

Wir führen die Variablentransformation

$$\varphi(s,t) = x(s,t) = s\xi + t_1\xi^1 + \cdots + t_{d-1}\xi^{d-1} =: A\begin{pmatrix} s \\ t \end{pmatrix}$$

durch, wobei

$$A = (\xi, \xi^1, \ldots, \xi^{d-1})$$

eine orthonormale Matrix ist und somit $|\det J\varphi(s,t)| = |\det A| = 1$ gilt. Ferner ist wieder $s = \langle x, \xi \rangle$. Nach Anwendung des Integraltransformationssatzes können wir also fortfahren mit

$$= \frac{1}{\sqrt{2\pi}} \int_{R^d} \exp(-\imath\sigma\langle \xi, x\rangle) f(x)\, dx = (2\pi)^{(d-1)/2} \hat{f}(\sigma\xi).$$

□

6.2 Die Radontransformation unter linearen Operationen

Ähnlich wie bei der Fouriertransformation leiten wir nützliche Hilfsmittel ab, d.h. wir leiten Formeln bezüglich des Verhaltens der Radontransformation unter wichtigen linearen Operationen ab. dabei ist der Projektionssatz das wichtigste Hilfsmittel.
Als erstes betrachten wir die Differentiation.

Satz 6.2 (Differentiation der Funktionen) *Sei $f \in \mathcal{S}$. Dann gilt:*

$$\mathcal{R}_\xi(D^\alpha f) = \xi^\alpha D^{|\alpha|} \mathcal{R}_\xi f.$$

Beweis Wir rechnen unter Verwendung von (3.8) auf Seite 139 und von Satz 6.1:

$$\begin{aligned}
&(\mathcal{R}_\xi(D^\alpha f))^\wedge(\sigma) \\
&\underbrace{=}_{\text{Satz}} (2\pi)^{(d-1)/2} (D^\alpha f)^\wedge(\sigma\xi) \underbrace{=}_{\text{Regel 3.8}} (2\pi)^{(d-1)/2} \cdot \imath^{|\alpha|} \cdot \sigma^{|\alpha|} \cdot \xi^\alpha \cdot \hat{f}(\sigma\xi) \\
&\underbrace{=}_{\text{Satz}} \imath^{|\alpha|} \cdot \sigma^{|\alpha|} \cdot \xi^\alpha \cdot (\mathcal{R}_\xi f)^\wedge(\sigma) \underbrace{=}_{\text{Regel 3.8}} \xi^\alpha (D^{|\alpha|} \mathcal{R}_\xi f)^\wedge(\sigma).
\end{aligned}$$

Rücktransformation liefert nun das Ergebnis. □

Weiter gilt

Satz 6.3 (Differentiation im 'Radonbereich') *Für $f \in \mathcal{S}$ gilt bezüglich der Differentiation nach der zweiten Variablen ξ, daß*

$$D_\xi^\alpha \mathcal{R} f = (-1)^{|\alpha|} \frac{\partial^{|\alpha|}}{\partial p^{|\alpha|}} \mathcal{R}(x^\alpha f).$$

Diese Formel beweisen wir nicht (man approximiert die Linienintegrale mit Hilfe der f_ε auf Seite 156, beweist dafür die Formel und geht zum Limes über, vgl. [13], Seite 12.
Für die Faltung ergibt sich eine besonders einfache Formel:

Proposition 6.1 *Für $f, g \in \mathcal{S}$ gilt:*

$$\mathcal{R}_\xi(f * g) = \mathcal{R}_\xi f * \mathcal{R}_\xi g.$$

Man beachte den Unterschied zur entsprechenden Formel für die Fouriertransformation.

Beweis Wir verwenden den Projektionssatz 6.1 und rechnen:

$$\begin{aligned}
& (\mathcal{R}_\xi(f * g))\hat{}(\sigma) = (2\pi)^{(d-1)/2} (f * g)\hat{}(\sigma\xi) \\
= \; & (2\pi)^{(d-1)/2} (2\pi)^{d/2} \hat{f}(\sigma\xi)\hat{g}(\sigma\xi) \\
= \; & (2\pi)^{1/2} \cdot \left\{(2\pi)^{(d-1)/2} \hat{f}(\sigma\xi)\right\} \left\{(2\pi)^{(d-1)/2} \hat{g}(\sigma\xi)\right\} \\
= \; & \sqrt{2\pi}\{\mathcal{R}_\xi f)\hat{}(\sigma)\}\{\mathcal{R}_\xi g)\hat{}(\sigma)\} = (\mathcal{R}_\xi f * \mathcal{R}_\xi g)\hat{}(\sigma).
\end{aligned}$$

Man beachte, daß es bei der letzten Identität um eine eindimensionale Fouriertransformation geht. Nun wenden wir die inverse Fouriertransformation auf beide Seiten an und erhalten die gewünschte Formel. □

Kapitel 7

Inversion der Radontransformation

Da ja aus Messungen der Radontransformierten auf die ursprüngliche Dichte zurückgeschlossen werden soll, sind natürlich die Inversionsmethoden der springende Punkt.

7.1 Adjungierte Operatoren

Diese benötigen wir zur Rekonstruktion der Verteilungen aus ihren Radontransformierten. Sie sind als Integrale über Sphären definiert und somit müssen wir einen Nachtrag zur Analysis vorausschicken.

7.1.1 Integrale über Sphären.

Wie üblich sei

$$S^{d-1} = \{x \in \mathbb{R}^d : |x| = 1\}$$

die Einheitssphäre im $\mathbb{R}^d$ und

$$rS^{d-1} = \{rx : x \in \mathbb{R}^d, |x| = r\}$$

die Oberfläche der Kugel im $\mathbb{R}^d$ mit Radius r.
Ähnlich wie Integrale $\int_L f\,dx$ von Funktionen f über Hyperebenen L im $\mathbb{R}^d$ benötigen wir Integrale $\int_{rS^{d-1}} f\,dx$ über Sphären. Ist f stetig auf $\mathbb{R}^d$, so ist die einfachste Art, das Integral analog zum Integral über Hyperebenen zu erklären:

$$\int_{rS^{d-1}} f\,dx := \lim_{\varepsilon \to 0} \int f(x) c_\varepsilon f\left(\frac{r - |x|}{\varepsilon}\right) dx,$$

wobei

$$\int c_\varepsilon f\left(\frac{r-|x|}{\varepsilon}\right)\,dx = 1,\ \varepsilon > 0.$$

Im einfachsten Fall wählt man

$$c_\varepsilon f\left(\frac{r-|x|}{\varepsilon}\right) = \frac{1}{V(r,\varepsilon)}\mathbf{1}_{[-1,1]}\left(\frac{r-|x|}{\varepsilon}\right) = \frac{1}{V(r,\varepsilon)}\mathbf{1}_{[-\varepsilon,\varepsilon]}(r-|x|)$$
$$= \frac{1}{V(r,\varepsilon)}\mathbf{1}_{[-\varepsilon,\varepsilon]}(r-|x|) = \frac{1}{V(r,\varepsilon)}\mathbf{1}_{\mathrm{Sch}(r,\varepsilon)}(x),$$

wobei $V(r,\varepsilon)$ der Rauminhalt der Schale

$$\mathrm{Sch}(r,\varepsilon) = \{x \in \mathbb{R}^d : r-\varepsilon \le |x| \le r+\varepsilon\}$$

ist. Damit ergibt sich

$$\int_{rS^{d-1}} f\,dx = \lim_{\varepsilon\to 0}\frac{1}{V(r,\varepsilon)}\int_{r-\varepsilon<|x|<r+\varepsilon} f(x)\,dx \left(= \int f(x)\delta(r-|x|)\,dx\right),$$

wobei letzteres wieder die Ingenieurschreibweise ist. In jedem Fall ergibt sich die Darstellung, die sich in jedem einführenden Text über klassische Analysis findet, z.B. in [4]: Sei $t = (t_1,\ldots,t_{d-1})$; dann gilt:

$$\int_{rS^{d-1}} f\,dx = \int_{|t|<r} f\left(t,\sqrt{r^2-|t|^2}\right)\frac{r}{\sqrt{r^2-|t|^2}}\,dt$$
$$= \int_{|t|<1} f\left(rt, r\sqrt{1-|t|^2}\right)\frac{r^{d-1}}{\sqrt{1-|t|^2}}\,dt.$$

Wichtig ist der folgende Satz vom Fubini-Typ:

Satz 7.1 *Sei $f : \mathbb{R}^d \to \mathbb{R}^d$ integrierbar. Dann ist für fast alle $r \ge 0$ die Funktion f über die Sphäre $S_r^{d-1} = \{x \in \mathbb{R}^d : |x| = r\}$ integrierbar und es gilt*

$$\int f(x)\,dx = \int_0^\infty\left(\int_{|x|=r} f\,dx\right)dr$$
$$= \int_0^\infty\left(\int_{|\xi|=1} f(r\xi)\,d\xi\right)r^{d-1}\,dr = \int_{|\xi|=1}\int_0^\infty f(r\xi)\,r^{d-1}\,dr d\xi.$$

Das einfachste Beispiel ist folgendes:

Beispiel 7.1 Sei

$$K_d = \{x \in \mathbb{R}^d : |x| \leq 1\}$$

die d-dimensionale Einheitskugel. Bezeichnet man mit τ_d ihr Volumen und mit ω_d ihre Oberfläche, so gilt nach dem Satz

$$\tau_d = \int_{|x|\leq 1} \mathbf{1}\, dx = \int_0^1 \left(\int_{|x|=r} \mathbf{1}\, dx \right) dr = \int_0^1 \omega_d r^{d-1}\, dr = \frac{\omega_d}{d}.$$

Im Klartext heißt das

$$\omega_d = d\tau_d$$

Im Fall $d = 2$ ist das die wohlbekannte Beziehung

$$\text{Kreisumfang} = 2\pi = 2(\pi) = 2 \cdot \text{Kreisfläche}.$$

Im Fall $d = 3$ gilt analog

$$\text{Kugeloberfläche} = 4\pi = 3\left(\frac{4}{3}\pi\right) = 3 \cdot \text{Kugelinhalt}.$$

7.1.2 Die Adjungierten.

Die Radontransformation verwandelt Funktionen $f(x)$ auf $\mathbb{R}^d$ in Funktionen $\mathcal{R}(p, \xi)$ auf $\mathbb{R} \times S^{d-1}$. Wir möchten manchmal Integrale über die Variablen p und ξ in einfache Integrale über die Variable x verwandeln. Zu diesem Zweck definieren wir für Funktionen $h(p, \xi)$ die Funktion

$$\mathcal{R}^\sharp h(x) = \int_{S^{d-1}} h(\langle x, \xi \rangle, \xi)\, d\xi.$$

Der Operator $\mathcal{R}^\sharp$ heißt zu $\mathcal{R}$ *adjungierter Operator*. Die Rechtfertigung für diesen Namen liefert folgender wichtige Satz. Um mathematisch präzise zu sein, definieren wir den *Einheitszylinder*

$$Z = \mathbb{R} \times S^{d-1}$$

und den Raum $\mathcal{S}(Z)$ als Menge der Einschränkungen von Funktionen aus $\mathcal{S}(\mathbb{R}^{d+1})$ auf Z.

Satz 7.2 *Für $f \in \mathcal{S}(\mathbb{R}^d)$ und $g \in \mathcal{S}(Z)$ gilt*

$$\int_Z \mathcal{R}f(s, \xi) g(s, \xi)\, ds d\xi = \int_{R^d} f(x) \mathcal{R}^\sharp g(x)\, dx.$$

Beweis Zunächst gilt nach Definition und obigem Satz 7.1, daß

$$\begin{aligned} &\int_{R^d} f(x)\mathcal{R}^\sharp g(x)\,dx \left(= \int_{R^d} f(x) \int_{S^{d-1}} g(\langle x,\xi\rangle,\xi)\,d\xi\,dx\right) \\ = &\int_{S^{d-1}} \int_{R^d} f(x)g(\langle x,\xi\rangle,\xi)\,dx\,d\xi. \end{aligned} \tag{7.1}$$

Nun führen wir (wie schon so oft) die (orthonormale) Variablentransformation

$$x(s,t) = s\xi + Bt,\ t^T = (t_1,\cdots,t_{d-1}), B = (\xi^1,\ldots,\xi^{d-1}),$$

durch, wobei die Spalten von B eine Orthonormalbasis von $\xi^\perp$ bilden.
Der Integraltransformationssatz – angewandt auf das innere Integral – erlaubt uns fortzufahren mit:

$$\begin{aligned} &= \int_{S^{d-1}} \int_R \left\{\int_{R^{d-1}} f(s\xi + Bt)\,dt\right\} g(s,\xi)\ ds\,d\xi \\ &= \int_{S^{d-1}} \int_R \mathcal{R}f(s,\xi)g(s,\xi)\,dsd\xi = \int_Z \mathcal{R}f(s,\xi)g(s,\xi)\,dsd\xi. \end{aligned}$$

□

Interpretiert man das erste Integral als Skalarprodukt $(f,\mathcal{R}^\sharp g)$ von Funktionen über dem $\mathbb{R}^d$ und das letzte Integral als Skalarprodukt $[\mathcal{R}f,g]$ von Funktionen über Z, so besagt der Satz, daß

$$[\mathcal{R}f,g] = (f,\mathcal{R}^\sharp g),$$

was – in Analogie zur Sprechweise bei Matrizen – die Bezeichnung 'Adjungierte' erklärt.

7.2 Die Inversionsformel

In diesem Abschnitt beweisen und diskutieren wir die Inversionsformel. Sie gibt an, wie man (in der Theorie) die Funktion f aus der Radontransformierten $\mathcal{R}f$ zurückgewinnt.

7.2.1 Die abstrakte Version.

Wir beweisen jetzt die Inversionsformel in ihrer abstrakten Form. Wir wählen diese, weil sie weit einfacher zu behandeln ist, als die analytischen Versionen (zumindest, wenn man diese direkt angeht.

Zunächst benötigen wir einen zusätzlichen Operator. Konkret ist er schwer zu handhaben. Deshalb definieren wir ihn über seine Transferfunktion. Für $\alpha \in \mathbb{R}$ ist das *Riesz-Potential*

$$\mathcal{I}^\alpha : \mathcal{S}(\mathbb{R}^d) \longrightarrow L^1(\mathbb{R}^d), f \longmapsto \mathcal{I}f,$$

definiert über

$$(\mathcal{I}^\alpha f)^\wedge(\lambda) = \frac{1}{|\lambda|^\alpha}\hat{f}(\lambda).$$

Zur Einübung beweisen wir

Lemma 7.1 *Für $f \in \mathcal{S}$ gilt*

$$\mathcal{I}^{-\alpha}\mathcal{I}^\alpha f = f.$$

Beweis Es gilt:

$$(\mathcal{I}^{-\alpha}\{\mathcal{I}^\alpha f\})^\wedge(\lambda) = |\lambda|^\alpha(\mathcal{I}^\alpha f)^\wedge(\lambda) = \frac{|\lambda|^\alpha}{|\lambda|^\alpha}\hat{f}(\lambda) = \hat{f}(\lambda).$$

Anwendung der inversen Fouriertransformation liefert nun die gewünschte Formel. □

Die Inversionsformel für die Radontransformation lautet:

Satz 7.3 *Für $f \in \mathcal{S}(\mathbb{R}^d)$ gilt*

$$f(x) = \frac{1}{2 \cdot (2\pi)^{d-1}}\mathcal{I}_x^{-\alpha}\mathcal{R}^\natural\mathcal{I}_s^{\alpha-d+1}\mathcal{R}f(x).$$

Zum Verständnis der Formel bemerken wir folgendes:

- f ist eine Funktion von $x \in \mathbb{R}^d$.
- $\mathcal{R}f$ ist eine Funktion von $(s, \xi) \in Z$.
- $\mathcal{I}_s$ wirkt für jedes feste ξ auf die Funktion $\mathcal{R}_\xi f(s)$ einer Variablen $s \in \mathbb{R}$. Somit ist $\mathcal{I}^{\alpha-d+1}\mathcal{R}f$ eine Funktion von $(s, \xi) \in Z$.
- Der Operator $\mathcal{R}^\natural$ darauf angewandt liefert eine Funktion von $x \in \mathbb{R}^d$.
- Der Operator $\mathcal{I}_x^\alpha$ wirkt auf diese Funktion von $x \in \mathbb{R}^d$ und erzeugt daraus wiederum eine Funktion von $x \in \mathbb{R}^d$.

Wir werden die Formel anschließend von allen Seiten beleuchten. Jetzt beweisen wir sie erst einmal, wobei der Beweis erstaunlich einfach ist.

Beweis Die Fourierinversionsformel gibt:

$$\mathcal{I}_x^\alpha f(x) = (2\pi)^{-d/2} \int |\lambda|^{-\alpha} \hat{f}(\lambda) \exp(\imath\langle x, \lambda\rangle)\, d\lambda.$$

Wir gehen nun zu Polarkoordinaten über, d.h. wir wenden die Transformation

$$\lambda = \sigma\xi, \sigma \geq 0, \xi \in S^{d-1}$$

an. Mit Satz 7.1 erhalten wir

$$\begin{aligned} &= (2\pi)^{-d/2} \int_0^\infty \int_{S^{d-1}} |\sigma|^{-\alpha} \hat{f}(\sigma\xi) \exp(\imath\sigma\langle x, \xi\rangle) |\sigma|^{d-1}\, d\sigma d\xi \\ &= (2\pi)^{-d/2} \int_{S^{d-1}} \int_0^\infty |\sigma|^{d-1-\alpha} \hat{f}(\sigma\xi) \exp(\imath\sigma\langle x, \xi\rangle)\, d\sigma d\xi. \end{aligned}$$

Dieselbe Rechnung mit $-\xi$ statt ξ und $-\sigma$ statt σ liefert dieselbe Formel, allerdings mit $\int_{-\infty}^0$ statt $\int_0^\infty$. Insgesamt erhalten wir

$$\mathcal{I}_x^\alpha f(x) = 2^{-1}(2\pi)^{-d/2} \int_{S^{d-1}} \int_{-\infty}^\infty |\sigma|^{d-1-\alpha} \hat{f}(\sigma\xi) \exp(\imath\sigma\langle x, \xi\rangle)\, d\sigma d\xi.$$

Nach dem Projektionssatz folgt

$$\begin{aligned} \mathcal{I}_x^\alpha f(x) &= \left\{2^{-1}(2\pi)^{-d/2}\right\} (2\pi)^{-(d-1)/2} \\ &\qquad \int_{S^{d-1}} \int_{-\infty}^\infty |\sigma|^{d-1-\alpha} \left\{(2\pi)^{(d-1)/2} \hat{f}(\sigma\xi)\right\} \exp(\imath\sigma\langle x, \xi\rangle)\, d\sigma d\xi \\ &= 2^{-1}(2\pi)^{-d+1/2} \int_{S^{d-1}} \int |\sigma|^{d-1-\alpha} (\mathcal{R}_\xi f)^\wedge(\sigma) \exp(\imath\sigma\langle x, \xi\rangle)\, d\sigma d\xi. \end{aligned}$$

Der Integrand kann durch das Riesz-Potential ausgedrückt werden:

$$\begin{aligned} \mathcal{I}_x^\alpha f(x) &= 2^{-1}(2\pi)^{-d+1/2}(2\pi)^{1/2} \\ &\qquad \int_{S^{d-1}} \left\{(2\pi)^{-1/2} \int (\mathcal{I}_\sigma^{\alpha+1-d} \mathcal{R}_\xi f)^\wedge(\sigma) \exp(\imath\sigma\langle x, \xi\rangle\, d\sigma\right\} d\xi. \end{aligned}$$

Man beachte, daß hier $\mathcal{I}_\sigma^{\alpha+1-d}$ auf die univariate Funktion $\mathcal{R}_\xi f(\sigma)$ der Variablen σ wirkt. Schließlich ergibt eindimensionale Fourierinversion

$$\begin{aligned} \mathcal{I}_x^\alpha f(x) &= 2^{-1}(2\pi)^{1-d} \\ &\qquad \int_{S^{d-1}} \mathcal{I}_\sigma^{\alpha+1-d} \mathcal{R} f(\langle x, \xi\rangle, \xi)\, d\xi = 2^{-1}(2\pi)^{1-d} \mathcal{R}^\natural \mathcal{I}_\sigma^{\alpha+1-d} \mathcal{R} f(x). \end{aligned}$$

Anwendung von $\mathcal{I}^{-\alpha}$ und Beachtung des Lemmas liefert nun die Formel. □

Wir interpretieren diese Formel nun in diversen Spezialfällen.

7.2.2 Inversion bei ungerader Dimension.

Wir setzen in der Inversionsformel $\alpha = 0$. Dann nimmt sie mit

$$c_d = \frac{1}{2 \cdot (2\pi)^{d-1}}$$

die Gestalt an:

$$f(x) = c_d \mathcal{R}^\natural \mathcal{I}_s^{1-d} \mathcal{R} f(x). \tag{7.2}$$

Wir untersuchen das Riesz-Potential etwas genauer. Für eine Funktion $h(s)$ gilt

$$(\mathcal{I}_s^{1-d} h)\hat{}(\sigma) = |\sigma|^{d-1} \hat{h}(\sigma) = (\mathrm{sgn}\sigma)^{d-1} \sigma^{d-1} \hat{h}(\sigma).$$

Nach dem Satz 4.2 über die Differentiation der Fouriertransformierten[1] gilt

$$(\mathrm{sgn}\sigma)^{d-1} (\mathcal{I}_s^{1-d} h)\hat{}(\sigma) = (-\imath)^{d-1} \left(\frac{d^{d-1}}{ds^{d-1}} h \right)\hat{}\,(\sigma).$$

Sei nun d ungerade und somit $d-1$ gerade. Dann kann man fortfahren mit

$$(\mathcal{I}_s^{1-d} h)\hat{}(\sigma) = (-1)^{(d-1)/2} \left(\frac{d^{d-1}}{ds^{d-1}} h \right)\hat{}\,(\sigma), \quad d \text{ ungerade.}$$

Angewandt auf

$$h(s) = \mathcal{R}_\xi(s)$$

ergibt das

$$\mathcal{I}_s^{1-d} \mathcal{R} f(s, \xi) = (-1)^{(d-1)/2} \frac{\partial^{d-1}}{\partial s^{d-1}} \mathcal{R}(s, \xi).$$

Einsetzen in die Inversionsformel (7.2) ergibt

$$\begin{aligned} f(x) &= c_d (-1)^{(d-1)/2} \mathcal{R}^\natural \frac{\partial^{d-1}}{\partial s^{d-1}} \mathcal{R} f(x) \\ &= (-1)^{(d-1)/2} \frac{1}{2(2\pi)^{d-1}} \int_{S^{d-1}} \frac{\partial^{d-1}}{\partial s^{d-1}} \mathcal{R} f(\langle x, \xi \rangle, \xi)\, d\xi. \end{aligned}$$

Im wichtigen Spezialfall $d = 3$ heißt das

$$f(x) = -\frac{1}{8\pi^2} \int_{S^2} \frac{\partial^2}{\partial s^2} \mathcal{R} f(\langle x, \xi \rangle, \xi)\, d\xi, \quad d = 3. \tag{7.3}$$

[1] nämlich:

$$\sigma^k \hat{h} = (-\imath)^k \left(\frac{d^k}{ds^k} h \right)\hat{}\,(\sigma).$$

Mit etwas Analysis sieht man hier den (rotationsinvarianten) Laplace-Operator am Werk. Wir könnten dies sofort ausrechnen, wenn wir kartesische mit Polarkoordinaten in Beziehung brächten. Um Rechnungen zu sparen, gehen wir aber direkt vor:

Beispiel 7.2 Sei Δ_x der Laplace-Operator, gegeben durch seine Wirkung auf eine Funktion $g(x), x \in \mathbb{R}^d$:

$$\Delta_x g = \left(\sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2} \right) g.$$

Dann gilt für die Fouriertransformierte (siehe Fußnote):

$$(-\Delta_x g)\hat{}(\lambda) = -(\Delta_x g)\hat{}(\lambda) = (-i)^2 (\Delta_x g)\hat{}(\lambda) = \left(\sum_{i=1}^{d} \lambda_i^2 \right) \hat{g}(\lambda) = |\lambda|^2 \hat{g}(\lambda);$$

der negative Laplace-Operator hat also die Transferfunktion $|\lambda|^2$.

Bei $d \geq 3$, d ungerade, ergibt die $(d-1)/2$-malige Anwendung von $-\Delta_x$:

$$\left(\left\{ (-\Delta_x)^{(d-1)/2} \right\} g \right)\hat{}\,(\lambda) = (|\lambda|^2)^{(d-1)/2} \hat{g}(\lambda) = |\lambda|^{d-1} \hat{g}(\lambda) = (\mathcal{I}_x^{1-d} g)\hat{}(\lambda),$$

d.h.

$$\mathcal{I}_x^{1-d} = (-\Delta_x)^{(d-1)/2}.$$

Für $\alpha = d - 1$ bekommt die Inversionsformel die Gestalt[2]:

$$f(x) = \frac{1}{2(2\pi)^{d-1}} \mathcal{I}_x^{1-d} \mathcal{R}^\natural \mathcal{R} f(x). \tag{7.4}$$

Mit dem Beispiel und der Definition von $\mathcal{R}^\natural$ erhalten wir schließlich für $d \geq 3$ ungerade:

$$f(x) = (-1)^{(d-1)/2} \frac{1}{2(2\pi)^{d-1}} \Delta_x^{(d-1)/2} \int_{S^{d-1}} \mathcal{R} f(\langle x, \xi \rangle, \xi)\, d\xi, \quad d \geq 3 \text{ ungerade.}$$

Im Spezialfall $d = 3$ bekommt man die zu (7.3) äquivalente Formel

$$f(x) = -\frac{1}{8\pi^2} \Delta_x \int_{S^2} \mathcal{R} f(\langle x, \xi \rangle\, d\xi, \quad d = 3,$$

die schon RADON bekannt war.

[2]Diese Form dient als Basis für den sogenannten ρ-filtered-layer-Algorithmus (BATES, PETER 1971), vgl. NATTERER. Dazu schreibt man sie um in

$$f(x) = c_d \mathcal{F}^{-1} \left(|\lambda|^{d-1} \mathcal{F} \mathcal{R}^\natural \mathcal{R} f \right).$$

Dies bedeutet 'Rückprojektion' (kommt bald), gefolgt von Fouriertransformation, Multiplikation und inverser Fouriertransformation.

7.2.3 Inversion bei gerader Dimension.

Wir setzen in der Inversionsformel wieder $\alpha = 0$, was zur bekannten Formel (7.2) führt:

$$f(x) = \frac{1}{2(2\pi)^{d-1}} \mathcal{R}^\natural \mathcal{I}_s^{1-d} \mathcal{R} f(x).$$

Diese Formel wollen wir nun näher untersuchen.
Sei h eine Funktion von $s \in \mathbb{R}$. Dann gilt

$$(\mathcal{I}_s^{1-d} h)^\wedge(\sigma) = |\sigma|^{d-1} \hat{h}(\sigma) = (\mathrm{sgn}\sigma)^{d-1} \sigma^{d-1} \hat{h}(\sigma).$$

Sei nun $\mathcal{H}$ der Operator mit

$$(\mathcal{H}h)^\wedge(\sigma) = -\imath \cdot \mathrm{sgn}\sigma \cdot \hat{h}(\sigma).$$

Dann gilt

$$\begin{aligned} & (\mathcal{I}_s^{1-d} h)^\wedge(\sigma) = (\mathrm{sgn}\sigma)^{d-1} \sigma^{d-1} \hat{h}(\sigma) \\ = \; & (\mathrm{sgn}\sigma)^{d-1} (-\imath)^{d-1} \left(\frac{d^{d-1}}{ds^{d-1}} h \right)^\wedge (\sigma) = \left(\mathcal{H}^{d-1} \frac{d^{d-1}}{ds^{d-1}} h \right)^\wedge (\sigma) \end{aligned}$$

und somit

$$\mathcal{I}_s^{1-d} h = \mathcal{H}^{d-1} \frac{d^{d-1}}{ds^{d-1}} h.$$

Damit wird aus der obigen Identität (7.2) die Formel

$$f = \frac{1}{2(2\pi)^{d-1}} \mathcal{R}^\natural \mathcal{H}^{d-1} \frac{\partial^{d-1}}{\partial s^{d-1}} \mathcal{R} f.$$

Somit wird es Zeit, den Operator $\mathcal{H}$ näher zu untersuchen. Dazu benötigen wir folgenden wichtigen Begriff aus der Analysis:

Definition 7.1 *Für $h \in \mathcal{S}(\mathbb{R})$ sei*

$$CH \int \frac{h(t)}{s-t}\, dt = \lim_{\varepsilon \to 0} \left(\int_{-\infty}^{s-\varepsilon} \frac{h(t)}{s-t}\, dt + \int_{s+\varepsilon}^{\infty} \frac{h(s)}{s-t}\, dt \right)$$

der Cauchysche Hauptwert[3].

[3]Dieser ist deutlich zu unterscheiden vom uneigentlichen Integral

$$\int \frac{h(t)}{s-t}\, dt = \lim_{\varepsilon \to 0} \int_{-\infty}^{a-\varepsilon} \frac{h(t)}{s-t}\, dt + \lim_{\varepsilon \to 0} \int_{s+\varepsilon}^{\infty} \frac{h(t)}{s-t}\, dt.$$

Der Cauchysche Hauptwert kann existieren, obwohl letzteres uneigentliche Integral nicht

Lemma 7.2 *Für $h \in \mathcal{S}(\mathbb{R})$ gilt*

$$\mathcal{H}h(s) = \frac{1}{\pi} CH \int \frac{h(t)}{s-t}\, dt,$$

wobei der Cauchysche Hauptwert existiert. Ferner gilt:

$$\mathcal{H}^{d-1}h = \begin{cases} (-1)^{(d-2)/2}\mathcal{H}h, & d \text{ gerade} \\ (-1)^{(d-1)/2}h, & d \text{ ungerade} \end{cases}$$

Der Cauchysche Hauptwert läßt sich als gewöhnliches Integral schreiben:

$$CH \int \frac{h(t)}{t}\, dt = \int_{-\infty}^{\infty} \frac{h(t) - h(-t)}{2t}\, dt.$$

Beweis Die erste Behauptung gehört in die Analysis und wird hier nicht bewiesen. Die explizite Gestalt von $\mathcal{H}^{d-1}$ erhält man aus:

$$(\mathcal{H}^{d-1}h)\hat{}(\sigma) = (-\imath)^{d-1}(\mathrm{sgn}\sigma)^{d-1}\hat{h}(\sigma).$$

Dies ergibt nämlich

$$(\mathcal{H}^{d-1}h)\hat{}(\sigma) = \begin{cases} (-1)^{(d-1)/2}\hat{h}(\sigma), & \text{falls } d \text{ ungerade} \\ (-1)(-1)^{(d-2)/2} \cdot \imath \cdot \mathrm{sgn}\sigma \cdot \hat{h}(\sigma) & \\ = (-1)^{(d-2)/2}(\mathcal{H}h)\hat{}(\sigma), & \text{falls } d \text{ gerade.} \end{cases}$$

Fourierinversion liefert nun die Behauptung. □

Mit diesem Lemma wird aus der letzten Umkehrformel:

$$f(x) = \frac{1}{2} \cdot (2\pi)^{1-d} \begin{cases} (-1)^{(d-2)/2} \int_{S^{d-1}} \mathcal{H}\frac{\partial^{d-1}}{\partial s^{d-1}}\mathcal{R}f(\langle x, \xi\rangle, \xi)\, d\xi, & d \text{ gerade,} \\ (-1)^{(d-1)/2} \int_{S^{d-1}} \frac{\partial^{d-1}}{\partial s^{d-1}}\mathcal{R}f(\langle x, \xi\rangle, \xi)\, d\xi, & d \text{ ungerade.} \end{cases}$$

---

existiert.

Beispiel. Das uneigentliche Integral an der Stelle s, $\int_a^b \frac{dt}{s-t}$, $a \le s \le t$, existiert bekanntlich nicht. Der Cauchysche Hauptwert jedoch existiert:

$$\begin{aligned} CH \int_a^b \frac{dt}{s-t} &= -\lim_{\varepsilon \to 0} \left(\int_a^{s-\varepsilon} \frac{dt}{t-s} + \int_{s+\varepsilon}^b \frac{dt}{t-s} \right) \\ &= -\lim_{\varepsilon \to 0} \left(\ln(t-s)|_a^{s-\varepsilon} + \ln(t-s)|_{s+\varepsilon}^b \right) \\ &= -\ln(-\varepsilon) + \ln(a-s) - \ln(b-s) + \ln(\varepsilon) = \ln\left(\frac{a-s}{b-s}\right). \end{aligned}$$

Genau so existiert das uneigentliche Integral $\int \frac{h(t)}{s-t}$ für eine Funktion $h \in \mathcal{S}(\mathbb{R})$, $h \neq 0$ in einer Umgebung von s, nicht, der Cauchysche Hauptwert jedoch schon.

Die bereits abgeleitete Formel für ungerades d finden wir hier übrigens wieder. Für gerades d ergibt sich mit $g = \mathcal{R}f$:

$$f(x) = \frac{1}{2} \cdot (2\pi)^{1-d}(-1)^{d/2-1}\frac{1}{\pi}\int_{S^{d-1}} CH \int \frac{g^{(d-1)}(t,\xi)}{\langle x,\xi\rangle - t}\, dt d\xi.$$

Substitution von $t = q + \langle x,\xi\rangle$ ergibt

$$f(x) = (2\pi)^{-d}(-1)^{d/2}\int_{S^{d-1}} CH \int \frac{g^{(d-1)}(q+\langle x,\xi\rangle,\xi)}{q}\, dq d\xi.$$

Wir verwenden die äquivalente Darstellung des Cauchyschen Hauptwertes und erhalten:

$$\begin{aligned} f(x) &= (2\pi)^{-d}(-1)^{d/2} \\ &\qquad \int_{S^{d-1}} \int \frac{1}{2q}\left(g^{(d-1}(\langle x,\xi\rangle + q,\xi) - g^{(d-1)}(\langle x,\xi\rangle - q,\xi)\right) dq d\xi \\ &= (-1)^{d/2}(2\pi)^{-d}\cdot 2^{-1} \\ &\qquad \int \frac{1}{q}\int_{S^{d-1}} \left(g^{(d-1)}(\langle x,\xi\rangle + q,\xi) - g^{(d-1)}(\langle x,\xi\rangle - q,\xi)\right) d\xi dq. \end{aligned}$$

Nun sind g gerade und $d-1$ ungerade und somit $g^{(d-1)}$ ungerade. Ersetzt man im zweiten Term also ξ durch $-\xi$, so sind beide Teilintegrale gleich und man erhält

$$f(x) = (-1)^{d/2}(2\pi)^{-d}\int \frac{1}{q}\int_{S^{d-1}} g^{(d-1)}(\langle x,\xi\rangle + q,\xi)\, d\xi\, dq.$$

Dies ist die erwünschte Umkehrformel. Es ist in der Literatur üblich, eine etwas andere Form zu benutzen: man schreibt

$$F_x(q) = \frac{1}{|S^{d-1}|}\int_{S^{d-1}} g(\langle x,\xi\rangle + q,\xi)\, d\xi,$$

wobei $|S^{d-1}|$ die Oberfläche der Einheitssphäre ist. Einsetzen liefert die bekannte Formel:

$$f(x) = c(n)\int_{-\infty}^{\infty} \frac{F_x^{(d-1)}(q)}{q}\, dq; \quad c(n) = (-1)^{d/2}(2\pi)^{-d}|S^{d-1}|.$$

Im wichtigen Fall $d = 2$ der Ebene wird daraus

$$f(x) = -\frac{1}{4\pi^2}\int \frac{1}{q}\int_{S^1} (\mathcal{R}f)'(\langle x,\xi\rangle + q,\xi)\, d\xi\, dq,$$

wobei sich die Differentiation auf das erste Argument von $\mathcal{R}f$ bezieht. Damit beenden wir die Diskussion der Umkehrformel.

Kapitel 8

Rekonstruktionsalgorithmen

Wir leiten nun Formeln ab, die den üblichen Rekonstruktionsalgorithmen zugrunde liegen. Im Mittelpunkt wird die sogenannte 'gefilterte Rückprojektion' stehen, ein Standardalgorithmus der Computertomographie.

8.1 Die Adjungierte als Rückprojektion

Die Anwendung des adjungierten Operators auf die Radontransformierte war erklärt als

$$\mathcal{R}^\natural \mathcal{R} f(x) = \int_{S^{d-1}} \mathcal{R} f(\langle x, \xi \rangle, \xi)\, d\xi.$$

Um zu verstehen, was dabei herauskommt, interpretieren wir den Integranden für festes x und ξ anhand der Graphik Abb. 8.1. Wir sehen also, daß bei dieser Operation die Intensität des 'Schattens' von f entlang der Geraden durch x senkrecht zu ξ mit infinitesimalem Gewicht $d\xi$ inkrementiert wird. Die 'Summe' dieser Schatten bildet den Wert in x. Im Prinzip handelt es sich also um eine 'Rückprojektion'. Natürlich wird bei obigem Integral jede Gerade durch x doppelt verwendet, nämlich einmal für ξ und einmal für $-\xi$. Deshalb wird in der Literatur oft die Rückprojektion $\mathcal{B}\mathcal{R} f = \mathcal{R}^\natural \mathcal{R} f/2$ betrachtet. Für uns ist der Vorfaktor unerheblich und wir interpretieren den Operator $\mathcal{R}^\natural$ als Rückprojektion. Somit haben wir den adjungierten Operator als Rückprojektion entlarvt.
Das Bild, welches man durch die Rückprojektion einer einzelnen Projektion senkrecht zu ξ erhält ist

$$\mathcal{R}_\xi^\natural \mathcal{R}_\xi f(x) = \mathcal{R} f(\langle x, \xi \rangle, \xi),$$

das sogenannte *Streifenbild.* Das Bild $\mathcal{R}^\natural \mathcal{R} f$ nennt man auch *Layergram.*

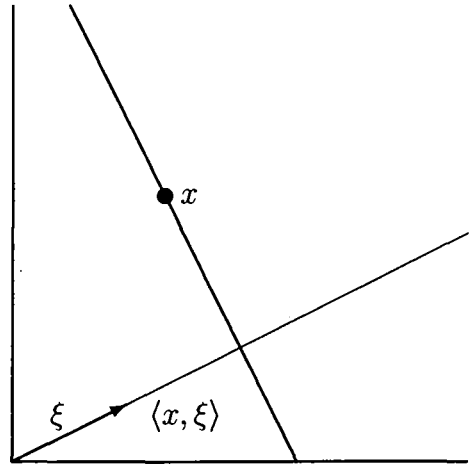


Abbildung 8.1: Interpretation des Integranden

Es ist offensichtlich, daß man untersucht, wie diese Rückprojektion mit f in Beziehung steht und versucht, das Ergebnis gegebenfalls durch geeignete Filterung im Sinne einer Rekonstruktion von f aus $\mathcal{R}f$ zu verbessern. Damit ist die Grundidee der gefilterten Rückprojektion umrissen, mit der wir uns im folgenden beschäftigen werden.

8.2 Rückprojektion und Filterung

So abstrakt er auch aussieht, ist der folgende Satz der Schlüssel zum Erfolg:

Satz 8.1 *Für $f \in \mathcal{S}(\mathbb{R}^d)$ und $h \in \mathcal{S}(Z)$ gilt:*

$$(\mathcal{R}^\natural h) * f = \mathcal{R}^\natural (h * \mathcal{R}f).$$

Man beachte, daß sich die Faltung auf der rechten Seite nur auf die eindimensionale Variable (meist s genannt) bezieht.

Bemerkung 8.1 Auf der linken Seite wird $\mathcal{R}^\natural h$ mit f gefaltet. Wir wissen, daß $\delta * f = f$ ist. Die Kunst ist es nun, h so zu wählen, daß $\mathcal{R}^\natural h \approx \delta$. Dann könnte man f durch Faltung mit h, d.h. Filterung, und anschließende Rückprojektion gut approximieren. Damit sind wir der Idee der gefilterten Rückprojektion wieder ein Stückchen näher gekommen.

Beweis Wir schreiben die linke Seite aus:

$$\mathcal{R}^\natural h * f(x) = \int_{R^d} \mathcal{R}^\natural h(x-y) f(y)\, dy = \int_{R^d} \int_{S^{d-1}} h(\langle x-y, \xi\rangle, \xi)\, d\xi\, f(y)\, dy$$

$$= \int_{S^{d-1}} \int_{R^d} h(\langle x-y,\xi\rangle,\xi) f(y)\, dy d\xi.$$

Wir substituieren im inneren Integral $y = s\xi + z$, wobei $z \in \xi^\perp$ (wie schon öfter) und erhalten

$$\begin{aligned}
\mathcal{R}^\sharp h * f(x) &= \int_{S^{d-1}} \int_R \int_{\xi^\perp} h(\langle x,\xi\rangle - s,\xi) f(s\xi+z)\, dz ds d\xi \\
&= \int_{S^{d-1}} \int_R h(\langle x,\xi\rangle - s,\xi) \mathcal{R}f(s,\xi)\, ds d\xi \\
&= \int_{S^{d-1}} (h * \mathcal{R}f)(\langle x,\xi\rangle,\xi)\, d\xi = \mathcal{R}^\sharp(h * \mathcal{R}f)(x).
\end{aligned}$$

Dies ist die erwünschte Formel. □

Der folgende Satz besagt, daß sich die Faltung mit $\mathcal{R}^\sharp g$ als Faltung mit einer gewöhnlichen Funktion darstellen läßt. Dadurch wird das Filterdesign erleichtert.

Satz 8.2 *Für $h \in \mathcal{S}(Z)$ gilt*

$$(\mathcal{R}^\sharp h)^\wedge(\lambda) = (2\pi)^{(d-1)/2} \frac{1}{|\lambda|^{d-1}} \left(\hat{h}\left(\frac{\lambda}{|\lambda|}, |\lambda|\right) + \hat{h}\left(-\frac{\lambda}{|\lambda|}, -|\lambda|\right)\right).$$

Beweis Zunächst erinnern wir uns an die Definition der Fouriertransformierten über ihre Wirkung auf Testfunktionen $w \in \mathcal{S}(\mathbb{R}^d)$:

$$\int (\mathcal{R}^\sharp g)^\wedge(x) w(x)\, dx = \int \mathcal{R}^\sharp g(x) \hat{w}(x)\, dx.$$

Aus der Charakterisierung von $\mathcal{R}^\sharp$ als Adjungierter von $\mathcal{R}$ folgern wir

$$\int \mathcal{R}^\sharp g(x) \hat{w}(x)\, dx = \int_{S^{d-1}} \int_R g(s,\xi) \mathcal{R}\hat{w}(s,\xi)\, dx d\xi.$$

Nun benutzen wir den Satz von Plancherel, der in der konkreten Situation ergibt:

$$\int_{S^{d-1}} \int_R g(s,\xi) \mathcal{R}\hat{w}(s,\xi)\, dx d\xi = \int_{S^{d-1}} \int_R \hat{g}(s,\xi) (\mathcal{R}\hat{w})^\sim(s,\xi)\, dx d\xi,$$

wobei für eine Funktion h

$$0\tilde{h}(s) = (2\pi)^{-d/2} \int h(\sigma) \exp(\imath s\sigma)\, d\sigma$$

die (Fourier-) Rücktransformation bedeutet. Das Fourier-Slice Theorem wurde zwar nur für die Fouriertransformation bewiesen. Ersetzt man im Beweis aber σ durch $-\sigma$, so sieht man, daß der Satz analog auch für die Rücktransformation gilt, d.h.

$$(\mathcal{R}\hat{w})^{\sim}(s,\xi) = (2\pi)^{(d-1)/2}(\hat{w})^{\sim}(s\cdot\xi).$$

Kombination der zwei letzten Identitäten liefert

$$\int \mathcal{R}^{\natural}g(x)\hat{w}(x)\,dx = (2\pi)^{(d-1)/2}\int_{S^{d-1}}\int_{R}\hat{g}(\sigma,\xi)w(\sigma\xi)\,d\sigma d\xi.$$

Wir verwenden jetzt den Satz von Fubini für die Zerlegung des $\mathbb{R}^d$ in Sphären. Dabei müssen wir die Fälle $\sigma > 0$ und $\sigma < 0$ separat behandeln. Dies ergibt

$$\begin{aligned}&\int (\mathcal{R}^{\natural}g)^{\wedge}(x)w(x)\,dx\\ &= (2\pi)^{(d-1)/2}\int_{R^d}\left(\hat{g}\left(|\lambda|,\frac{\lambda}{|\lambda|}\right)+\hat{g}\left(-|\lambda|,-\frac{\lambda}{|\lambda|}\right)\right)|\lambda|^{1-d}w(\lambda)\,d\lambda.\end{aligned}$$

Damit sehen wir, daß $(\mathcal{R}^{\natural}g)^{\wedge}$ auf Testfunktionen genau so wirkt wie die rechte Seite im Satz. □

Wir fragen uns schon lange, was denn die bloße Rückprojektion von $\mathcal{R}f$ für ein Bild liefert. Für $d = 2$ können wir das durch folgendes (Ingenieuren geläufiges) heuristische Argument aus der letzten Formel ableiten:
Setzen wir in Satz 8.1 formal $h = \delta$, so erhalten wir

$$(\mathcal{R}^{\natural}\delta) * f = \mathcal{R}^{\natural}\mathcal{R}f.$$

Um $\mathcal{R}^{\natural}\delta$ zu bestimmen, setzen wir $h = \delta$ formal in die Formel aus Satz 8.2 ein, was ergibt

$$(\mathcal{R}\delta)^{\wedge}(\lambda) = (2\pi)^{1/2}\frac{1}{|\lambda|}((2\pi)^{-1/2}+(2\pi)^{-1/2}) = \frac{2}{|\lambda|}.$$

Wir entnehmen der Literatur

$$\mathcal{F}^{-1}\left(|\lambda|^{-1}\right)(x) = |x|^{-1}$$

und kommen zum Ergebnis

$$\mathcal{R}^{\natural}\mathcal{R}f = 2\cdot\frac{1}{|x|} * f.$$

Die alleinige Rückprojektion ergibt also eine verschmierte Version von f, nämlich die Faltung mit der Filtermaske $2/|x|$. Benutzt man die oben erwähnte Rückprojektion $\mathcal{B} = \mathcal{R}^\natural/2$, so hat man

$$\mathcal{B}\mathcal{R}f = \frac{1}{|x|} * f.$$

Allgemein gilt:

Satz 8.3 *Für $f \in \mathcal{S}(\mathbb{R}^d)$ gilt*

$$\mathcal{R}^\natural\mathcal{R}f = |S^{d-2}| \cdot \frac{1}{|x|} * f.$$

Zur Einordnung des obigen Spezialfalls beachte man

$$|S^2| = 4\pi \quad |S^1| = 2\pi, \quad , |S^0| = |\{-1, 1\}| = 2$$

(als normiertes Oberflächenmaß).

Beweis Ausschreiben der Definitionen ergibt

$$\mathcal{R}^\natural\mathcal{R}f = \int_{S^{d-1}} \{\mathcal{R}f(\langle x, \xi\rangle, \xi)\}\, d\xi = \int_{S^{d-1}} \left\{ \int_{\xi^\perp} f(\langle x, \xi\rangle\xi + z)\, dz \right\} d\xi$$

Es ist

$$x - \langle x, \xi\rangle\xi \in \xi^\perp.$$

Substituieren wir im inneren Integral

$$z = y + (x - \langle x, \xi\rangle\xi),$$

so bekommen wir

$$\begin{aligned} &\mathcal{R}^\natural\mathcal{R}f(x) \\ &= \int_{S^{d-1}} \int_{\xi^\perp} f(\langle x, \xi\rangle\xi + (x - \langle x, \xi\rangle\xi) + y)\, dy d\xi = \int_{S^{d-1}} \int_{\xi^\perp} f(x + y)\, dy d\xi. \end{aligned}$$

Wir verwenden jetzt unbewiesen die allgemeine Identität

$$\int_{S^{d-1}} \int_{\xi^\perp} g(y)\, dy d\xi = |S^{d-2}| \int_{R^d} |y|^{-1} g(y)\, dy$$

(NATTERER, VII.2(2.8), Seite 190). Angewandt auf $g(y) = f(x + y)$ liefert sie

$$\mathcal{R}^\natural\mathcal{R}f(x) = |S^{d-2}| \int_{R^d} |y|^{-1} f(x + y)\, dy = |S^{d-2}| \int_{R^d} |x - y|^{-1} f(y)\, dy$$

und somit die behauptete Gleichung. □

8.3 Die gefilterte Rückprojektion

Dies ist der gebräuchlichste Rekonstruktionsalgorithmus in der Computertomographie.

Konstruktion des Filters.

Im letzten Abschnitt sahen wir, daß die Punktantwort (auf einen scharfen Impuls in 0) der Rücktransformation die Dichte $1/|x|$ ist. Um diese Unschärfe zu verbessern, wird ein zusätzlicher Filter nachgeschaltet. Die Konstruktion läßt sich am besten anhand der Formel aus Satz 8.1 erläutern, nämlich

$$(\mathcal{R}^\natural h) * f = \mathcal{R}^\natural(h * \mathcal{R}f).$$

Ideal wäre ein h_∞, mit $\mathcal{R}^\natural h_\infty = \delta$, denn dann hätten wir

$$f = \delta * f = (\mathcal{R}^\natural h_\infty) * f = \mathcal{R}^\natural(h_\infty * \mathcal{R}f).$$

Da dies nicht erreichbar ist, begnügen wir uns mit einer Approximation

$$\mathcal{R}^\natural w_b = W_b \approx \delta.$$

Im Hinblick auf die Approximation von δ erinnern wir uns an den Satz 4.6, der besagte

$$\Psi_b(x) = |b|^d \Psi(bx) \longrightarrow \delta(x),\ b \to \infty, \quad b > 0.$$

Nach der Ähnlichkeitsformel gilt für die Fouriertransformierte

$$\hat{\Psi}(\lambda) = \hat{\Psi}\left(\frac{\lambda}{b}\right)$$

Wir starten also mit einem geeigneten Tiefpaßfilter Ψ und wählen b hinreichend groß. Konkret gehen wir aus von der Gleichung

$$\hat{W}_b(\lambda) = (2\pi)^{-d/2}\hat{\Phi}\left(\frac{|\lambda|}{b}\right). \tag{8.1}$$

Der Betrag im Argument zielt auf Rotationsinvarianz. Für den Ausgangsfilter fordern wir

$$0 \le \hat{\Phi} \le 1,\ \hat{\Phi}(\sigma) = 0 \text{ für } \sigma \ge 1$$

oder wenigstens ein gutes Abklingverhalten.

Beispiel 8.1 Der ideale Tiefpaßfilter

$$\hat{\Phi}(\sigma) = \mathbf{1}_{[0,1]}(\sigma)$$

hat

$$W_b(x) = (2\pi)^{-d/2} b^n \cdot \frac{J_{n/2}(b|x|)}{(b|x|)^{d/2}},$$

wobei 'J' für 'Besselfunktion' steht.

Satz 8.2 gibt uns den Zusammenhang zwischen W_b und w_b über die Fouriertransformierten:

$$\hat{W}_b(\lambda) = (\mathcal{R}^\natural w_b)\hat{}(\lambda) = (2\pi)^{(d-1)/2}|\lambda|^{1-d}\left(\hat{w}_b\left(|\lambda|, \frac{\lambda}{|\lambda|}\right) + \hat{w}_b\left(-|\lambda|, -\frac{\lambda}{|\lambda|}\right)\right).$$

Da wir nur rotationsinvariante $\hat{W}_b$ zulassen, die nur von $|\lambda|$ abhängen, können wir das erste Argument rechts weglassen. Ferner wollen wir nur gerade $\hat{w}_b$. Unter diesen Voraussetzungen vereinfacht sich die Formel:

$$(2\pi)^{-d/2}\hat{\Phi}\left(\frac{|\lambda|}{b}\right) = 2 \cdot (2\pi)^{(d-1)/2}|\lambda|^{1-d}\hat{w}_b(|\lambda|).$$

Schreiben wir nun σ statt $|\lambda|$, so ergibt sich nach Vereinfachung

$$\hat{w}_b(\sigma) = \frac{1}{2} \cdot (2\pi)^{1/2-d}\sigma^{d-1}\hat{\Phi}\left(\frac{\sigma}{b}\right) \quad (\hat{w}_b \text{ gerade}), \quad \sigma > 0. \tag{8.2}$$

Diskretisierung.

Der eigentliche Algorithmus ist eine diskrete Version des eben geschilderten Verfahrens. Wir messen $g(s,\xi) = \mathcal{R}f(s,\xi)$ an den Punkten

$$(s_i, \xi_j), \quad i = -q, \ldots, q; \quad j = 1, \ldots, p$$
$$s_i = h \cdot i, \quad h = 1/q, \quad \xi_j \in S^{d-1}.$$

Die Faltung $w_b * g$ wird durch die diskrete Faltung

$$v_j(s) = w_b \overset{h}{*} g(s, \xi_j) = h \sum_{i=-q}^{q} w_b(s - s_i) g(s_i, \xi_j) \tag{8.3}$$

ersetzt (man erinnere sich, daß die Faltung nur im ersten, eindimensionalen Argument ausgeführt wird). Für die Rückprojektion brauchen wir eine Quadraturformel auf S^{d-1} mit den Knoten ξ_j, also eine gewichtete Summe

$$\int_{S^{d-1}} h(\xi)\, d\xi \sim \sum_{j=1}^{p} \alpha_{pj} h(\xi_j). \tag{8.4}$$

Beispiel 8.2 Für $n = 2$ und gerade Funktionen bietet sich die trapezoide Regel an:

$$\int_{S^1} h(\xi)\, d\xi \approx \frac{2\pi}{m+1} \sum_{l=0}^{m} h(\xi_j),\ \xi_j = \begin{pmatrix} \cos\phi_j \\ \sin\phi_j \end{pmatrix},\ \phi_j = \frac{\pi_j}{m+1},\ j = 0, \ldots, m. \tag{8.5}$$

Idealerweise soll diese exakt sein (d.h. gleich dem Integral) für eine möglichst große Klasse günstiger Funktionen (etwa die geraden sphärischen harmonischen Polynome vom Grad $2m$).
Die Rückprojektion hat dann die Gestalt

$$\mathcal{R}_p^\natural v(x) = \sum_{j=1}^{p} \alpha_{pj} v(\langle x, \xi_j \rangle, \xi_j),$$

und die gefilterte Rückprojektion ist

$$f_{FB} = \mathcal{R}_p^\natural w_b \overset{h}{*} (\mathcal{R} f).$$

Beispiel 8.3 (Die parallele Geometrie in der Ebene) Hier hat man üblicherweise:

$$\xi_j = \begin{pmatrix} \cos\phi_j \\ \sin\phi_j \end{pmatrix},\ \phi_j = \frac{j-1}{p}\pi,\ s_i = h \cdot i,\ h = \frac{1}{q}.$$

Für die diskrete Faltung (8.3) erhält man:

$$v_{k,j} = h \sum_{i=-q}^{q} w_b(s_k - s_i) g(s_i, \xi_j),\ k = -q, \ldots, q;\ j = 1, \ldots, p.$$

Für jeden Rekonstruktionspunkt x berechnet man nun die diskrete Rückprojektion mittels (8.4).

Soweit ist alles 'straightforward'. Diskutiert wurde – und wird immer noch – die Wahl der Funktion w_b. Wir schreiben ein klassisches Beispiel nieder:

Beispiel 8.4 Sei im Beispiel $d = 2$. Die Funktion w_b ergibt sich aus der Transferfunktion $\hat{\Phi}$. Ein einfacher Fall ist die Dreiecksfunktion (im Frequenzbereich)

$$\hat{\Phi}(\sigma) = \begin{cases} 1 - \varepsilon \cdot \sigma, & \sigma \leq 1 \\ 0, & \sigma > 1. \end{cases},$$

wobei $\varepsilon \in [0,1]$. Hierin ist mit $\varepsilon = 0$ der ideale Tiefpaßfilter enthalten. Wir erhalten aus der Identität (8.2)

$$\hat{w}_b(\sigma) = \frac{1}{2}(2\pi)^{1/2-d}\sigma^{d-1}\hat{\Phi}\left(\frac{\sigma}{b}\right),$$

und somit

$$\begin{aligned} w_b(s) &= \frac{1}{2}(2\pi)^{-3/2}(2\pi)^{-1/2}\int |\sigma|\left(1-\varepsilon\frac{|\sigma|}{b}\right)\exp(\imath s\sigma)\,d\sigma \\ &= \frac{1}{8\pi^2}\frac{1}{2}\int \sigma\left(1-\varepsilon\cdot\frac{\sigma}{b}\right)\cos(s\sigma)\,d\sigma. \end{aligned}$$

Partielle Integration liefert

$$\begin{aligned} w_b(s) &= \frac{b^2}{4\pi^2}(u(bs)-\varepsilon v(bs)), \\ u(s) &= \begin{cases} \frac{\cos s-1}{s^2}+\frac{\sin s}{s}, & s\neq 0 \\ \frac{1}{2}, & s=0, \end{cases} \\ v(s) &= \begin{cases} \frac{2\cos s}{s^2}+\left(1-\frac{2}{s^2}\right)\frac{\sin s}{s}, & s\neq 0, \\ \frac{1}{3}, & s=0. \end{cases} \end{aligned}$$

Verknüpft man b mit h über

$$b = \frac{\pi}{h}$$

und wertet w_b nur an den Stellen $s = s_i$ aus, so vereinfacht sich das beträchtlich und man erhält:

$$w_b(s_i) = \frac{b^2}{2\pi^2}\cdot\begin{cases} 1/4-\varepsilon/6, & i=0, \\ -\varepsilon/(\pi^2 i^2), & i\neq 0,\ \text{gerade}, \\ -(1-\varepsilon)/(\pi^2 i^2), & i\ \text{ungerade}. \end{cases}$$

Es wurden viele andere Filter vorgeschlagen, u.a. der Filter mit Transferfunktion

$$\hat{\Phi}(\sigma) = \begin{cases} \operatorname{sinc}(\sigma\pi/2), & \sigma\geq 0, \\ 0, & \sigma > 1. \end{cases}$$

Beispiel 8.5 Die Fächerstrahlgeometrie wird am bequemsten auf die parallele Geometrie zurückgeführt. Vgl. NATTERER, V.1.2, S 111.

8.4 Weitere Rekonstruktionsalgorithmen

Es gibt praktisch zu jeder bewiesenen Umkehrformel mehrere Vorschläge für Rekonstruktionsalgorithmen.

Beispiel 8.6 (Die Fourierrekonstruktion) Dabei wird die Projektionsformel

$$\hat{f}(\sigma\xi) = (2\pi)^{(1-d)/2}(\mathcal{R}f)^\wedge(\sigma,\xi)$$

direkt ausgenutzt, indem man die rechte Seite (Fourier) invertiert. Die technischen Probleme liegen hier in der Diskretisierung (vgl. NATTERE, S 119ff).

Beispiel 8.7 (ρ-filtered layergram) Diese Methode beruht auf der Identität von Satz 7.3:

$$f(x) = \frac{1}{2}(2\pi)^{1-d}\mathcal{I}_x^{\alpha-d+1}\mathcal{R}^\natural\mathcal{R}f(x).$$

Mit $\alpha = d-1$ vereinfacht sie sich zu

$$f(x) = \frac{1}{2\cdot(2\pi)^{d-1}}\mathcal{I}_x^{1-d}\mathcal{R}^\natural\mathcal{R}f(x).$$

Da $\mathcal{I}_x$ über die Transferfunktion definiert war:

$$(\mathcal{I}_x^\alpha f)^\wedge(\lambda) = \frac{1}{|\lambda|^\alpha}\hat{f}(\lambda),$$

ergibt sich

$$f = \tilde{\hat{f}} = \frac{1}{2}(2\pi)^{1-d}\left(|\eta|^{d-1}(\mathcal{R}^\natural\mathcal{R}f)^\wedge\right)^\sim.$$

Somit wird eine Rücktransformation ausgeführt, gefolgt von einer d-dimensionalen Fouriertransformation, einer Multiplikation und einer Fourierrücktransformation. Dieser Ansatz ist nicht wesentlich verfolgt worden, da er erhebliche numerische Schwierigkeiten bereitet.

Wir schließen mit der vom Ansatz her einfachsten Methode, skizziert für die Ebene $\mathbb{R}^2$.

Beispiel 8.8 (Algebraische Rekonstruktion) Hierbei geht man direkt von einer Diskretisierung aus, nämlich von endlich vielen Linienintegralen

$$\int_{L_j} f(x)\,dx = g_j,\quad j = 1,\ldots,N,$$

wobei also $g_j = \mathcal{R}f(s_j, \xi_j)$ für endlich viele $(s_j, \xi_j) \sim L_j$. Nun ersetzt man f durch eine diskrete Version F: Wir nehmen an, f sei auf einem beschränkten Quadrat Q konzentriert ist. Dieses unterteilen wir gitterförmig in kleine Quadrate Q_k, z.B. $k = 1, \ldots, M = n^2$. Dann können wir F durch einen Vektor

$$F = (F_1, \ldots, F_M)^T$$

darstellen, wobei F_k der Mittelwert von f über Q_k ist. Das Linienintegral über L_j wird ersetzt durch eine Summe

$$g_j = \int_{L_j} f(x)\, dx \approx \sum_{k=1}^{M} a_{jk} F_k,$$

wobei

$$a_{jk} = \text{Länge von } (L_j \cap Q_k).$$

Mit der bekannten Matrix $A = (a_{jk})$ und dem gemessenen Vektor $g = (g_j)$ bekommen wir also ein lineares Gleichungssystem

$$AF = g$$

für die gesuchte Verteilung F.
Einerseits bereitet die Dimension des Problems Schwierigkeiten, denn $M = n^2$ kann ohne weiteres von der Größenordnung $10000^2 = 10^8$ sein. Andererseits trifft eine Gerade nur wenige Zellen, deren Anzahl ist von der Größenordnung $n \approx \sqrt{M}$, so daß die restlichen a_{jk} verschwinden.
Ein zweites Problem ist, daß mangels genügend vieler Projektionen das Gleichungssystem unbestimmt sein kann.
Das dritte Problem kommt von der Unsicherheit der Messungen; dies führt i.a. zur Inkonsistenz des realen Gleichungssystems.
Insgesamt tritt eine Fülle numerischer Probleme auf, deren Lösung unabhängig von der Computertomographie ist (vgl. Pseudoinverse, aczmarz-Methode).

Literatur

[1] H. BÄSSMANN & PH.W. BESSLICH (1991): *Bildverarbeitung Ad Oculos.* Theory and Applications. Springer Verlag

[2] R. CHELLAPPA & A. JAIN (1993): *Markov Random Fields. Theory and Applications.* Academic Press, Inc.: Boston etc.

[3] S.R. DEANS (1983): *The Radon Transform and Some of its Applications.* John Wiley & Sons: New York etc.

[4] FORSTER (1981): *Analysis 3. Integralrechnung im* $\mathbb{R}^n$. Vieweg Verlag: Braunschweig, Wiesbaden

[5] B. FORSTER & R. LASSER (2004): *Biomedical Imaging: An Overview.* this volume

[6] H. FÜHR, O.TREIBER & F. Wanninger (2003): Cluster-oriented Detection of Microcalcifications in Simulated Low-Dose Mammography. *Proceedings of 'Bildverarbeitung für die Medizin'*, Springer Verlag, 96-100

[7] R.C. GONZALEZ & P. WINTZ (1987): *Digital Image Processing. Second Edition.* Addison-Wesley

[8] G.T. HERMAN (1980): *Image Reconstruction from Projections.* Academic Press: New York etc.

[9] B. JÄHNE (2002): *Digital image processing. Concepts, algorithms, and scientific applications.* 5. überarbeitete Auflage, Springer Verlag: Berlin etc.

[10] D.A. LISLE (1996): *Imaging for Students.* Arnold: London etc. und Oxford University Press: New York

[11] A. MACOVSKI (1983): *Medical Imaging Systems.* Prentice-Hall Inc.: Englewood Cliffs, New Jersey 07632

[12] MORNEBURG H., SIEMENS AG (Hrsg.) (1995): *Bildgebende Systeme für die medizinische Diagnostik.* 3. Auflage, Publicis-MCD-Verlag

[13] F. NATTERER (1986): *The Mathematics of Computerized Tomography.* John Wiley & Sons: Chichester etc.

[14] J. RADON (1917): Über die Bestimmung von Funktionen durch ihre Ingetralwerte langs gewisser Mannigfaltigkeiten. *Berichte Sächsische Akademie der Wissenschaften*, 69:262–279,

[15] O. TREIBER, F. WANNINGER, H. FÜHR, W. PANZER, D. REGULLA & G. WINKLER (1998): An adaptive algorithm for the detection of microcalcifications in simulated low-dose mammography. *Physics in Medicine and Biology*, 48(3):449–466, .

[16] G. WINKLER (2003): *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods.* 2nd completely rewritten and revised edition. Applications of Mathematics 27: Springer Verlag: Berlin etc.

[17] G. WINKLER(2004): *Bildgebende Verfahren der medizinischen Diagnostik.* Dieser Band

[18] TZAY Y. YOUNG & KING-SUN FU (Eds.) (1986): *Handbook of Pattern Recognition and Image Processing*, Academic Press: Inc:San Diego etc.

[19] (1996) *Mathematics and Physics of Emerging Biomedical Imaging.* National Academic Press: Washington, D.C.

Teil III

PET, NMR, FTICRMS

Positron Emission Tomography and Molecular Imaging

Klaus Hahn *

Abstract

Due to improved biomolecular probes, PET technology comes again into the focus of interest. PET basics of physical, biological and signal processing aspects are discussed, and an application to cancer therapy in the framework of molecular imaging is presented.

*Institute of Biomathematics and Biometry
GSF - National Research Center for Environment and Health,
Postfach 1129, D-85758 Oberschleißheim, Germany,
hahn@gsf.de, http://ibb.gsf.de

1 Biochemical and Physical Prerequisites

PET (Positron Emission Thomography) is a biological imaging tool enabling quantitative regional tissue analysis. It utilizes radioactively labelled biological probes. After injection, atoms from these positron emitting isotopes tag to molecules of interest. The tracers dissolve in the tissue according to their absorbtion and change concentration with time. Some of the isotopes decay, each one emitting a positron and a neutrino. The neutrino passes through the tissue without interaction, the positron scatters with tissue electrons and annihilates with one electron in two .5 MeV photons, mainly back to back; Fig. 1 illustrates this process. Isotopes are produced by clinical cyclotrons.

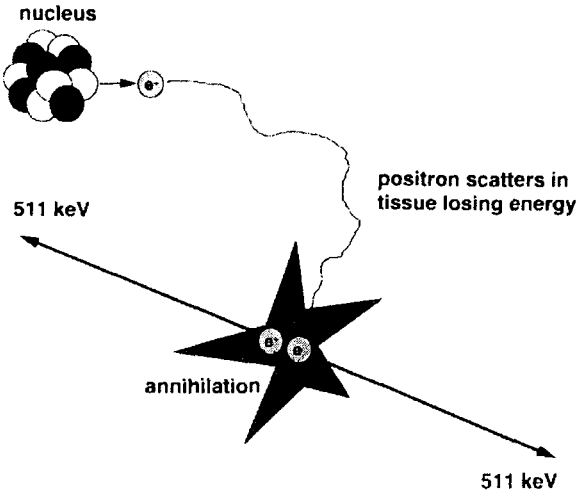


Figure 1: Positron emission and annihilation. A positron is emitted from a labelled probe molecule, losing energy by scattering from atomic electrons before annihilating with an electron to produce two 511-keV gamma rays which are emitted with an angular separation of 180^0. The distance the positron travels before annihilation typically is less than 2 mm. From [1].

Most important are those of carbon, nitrogen, oxygen and fluorine which are major constituents for example of the human body.
In Table 1 some important properties of such isotopes are collected. In Table 2 few tracers out of a very long list of known compounds are presented, their medical application field is indicated in the right column. The PET

Isotope	Half-life (min)	Energy (max) (keV)	Range (mm) (rms)
Carbon-11	20.4	960	0.69
Nitrogen-13	9.96	1190	0.91
Oxygen-15	2.07	1720	1.44
Fluorine-18	109.8	640	0.38

Table 1: Physical properties of commonly used positron-emitting isotopes. Energy is maximum possible positron energy (endpoint energy). The root mean square (rms) positron range is measured in water, see [3].

scanner detects the two .5 MeV gammas from decay. A plane scanner is symbolically displayed in Fig. 2. It consists of a circumferential array of scintillation detectors which register a signal when a gamma ray interacts with them. The detectors are connected to fast timing circuits which detect two simultaneous or "coincident" events on opposite sides of the object. The two opposite locations of detection define a straight line which meets the position of the positron labelled molecule.

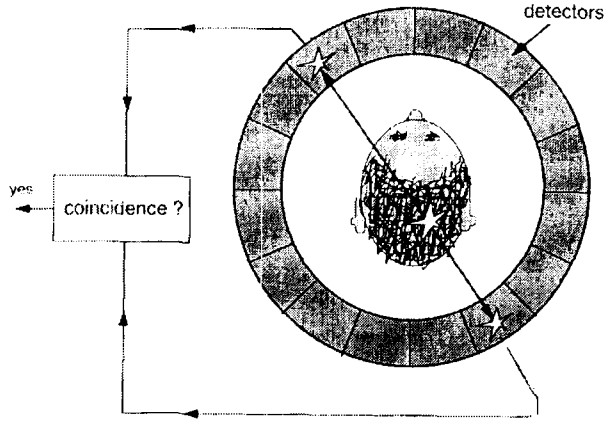


Figure 2: A plane PET scanner. From [1].

Labelled compound	Application
Physiological parameters	
[^{15}O]Water, [^{15}O]butanol, C[$^{15}O_2$]	Blood flow
^{11}CO, C^{15}O, ^{68}Ga-EDTA	Blood volume
^{68}Ga-EDTA, ^{82}Rb	Blood-brain barrier permeability
Metabolism and biosynthesis	
2-[^{18}F]Fluoro-2-deoxy-D-glucose, 2-[^{11}C]deoxy-D-glucose	Glucose metabolism
[^{11}C]Palmitic acid, [^{11}C]β-methylheptadecanoic acid	Free fatty acid metabolism
L-[^{13}N]Glutamate, L-[^{13}N]alanine, L-[^{13}N]aspartate	Transaminase activity
[1-^{11}C]Acetate	Krebs cycle function

Table 2: Partial list of biological probes labelled with positron emitters and their applications. From [1].

2 Signal Processing

In Fig. 3, left panel, a set of coincidence lines for a plane device, scanning a brain, is shown. For each single angle, all coincidences with distance r are collected in a sinogram in the right panel. The sinogram is the data basis for the reconstruction of the (in general 3 dimensional) image of radioactivity

distribution in the brain. Fortunately, distributions are characterized by the projections. The analytical key is the following result. Let f be a rapidly decreasing function[1] and denote by $\mathcal{R}_\xi f$ its projection along the direction of a unit vector ξ. Denote further by $\hat{h}$ the Fourier transform of a function h. Then

Theorem 1 (Projection-, Fourier-Slice Theorem) *For every rapidly decreasing function f on $\mathbb{R}^2$ one has*

$$(\mathcal{R}_\xi f)\hat{}(s) = \sqrt{2\pi}\hat{f}(s\xi) \quad \text{for every } s \in \mathbb{R}. \tag{1}$$

This assertion is illustrated in Fig. 4. The result gives immediately the *direct Fourier method* for the reconstruction of f:

Theorem 2 *A rapidly decreasing function f can uniquely be recovered from its Radon transform by Fourier inversion of both sides in (1). In particular, f is uniquely determined by all projections on the coincidence lines for all directions.*

The Radon inversion is explicitly given by the next result:

Theorem 3 *Let f be a rapidly decreasing function on $\mathbb{R}^2$. Then*

$$f(x) = -\frac{1}{4\pi^2}\int \frac{1}{q}\int_{S^1} (\mathcal{R}f)'(\langle x,\xi\rangle + q, \xi)\, d\xi\, dq.$$

Integration w.r.t. ξ ranges over the unit sphere S^1 in the Euclidean plane. A version more common in the literature uses

$$F_x(q) = \frac{1}{2\pi}\int_{S^1} (\mathcal{R}f)(\langle x,\xi\rangle + q, \xi)\, d\xi,$$

and f can be written as

$$f(x) = -\frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{F_x'(q)}{q}\, dq.$$

[1]For definitions, derivations, and background material the reader may consult the monograph S.R. Deans (1983), [2], or the contribution G. Winkler (2004), [6], to this volume.

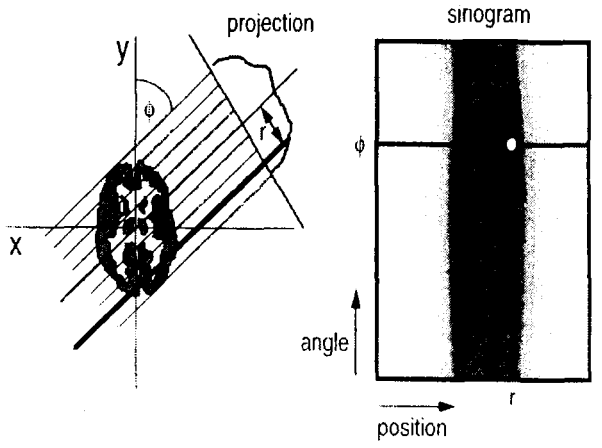


Figure 3: Raw PET data are stored in a 2D matrix known as a sinogram prior to image reconstruction. Each element in the sinogram represents the number of counts detected by a particular detector pair. The convention is such that the vertical axis represents the angle of the line of response and the horizontal axis represents the displacement from the center of the field of view. The events detected from the line of response shown in bold would be stored at the location represented by the white dot. Adapted from [1].

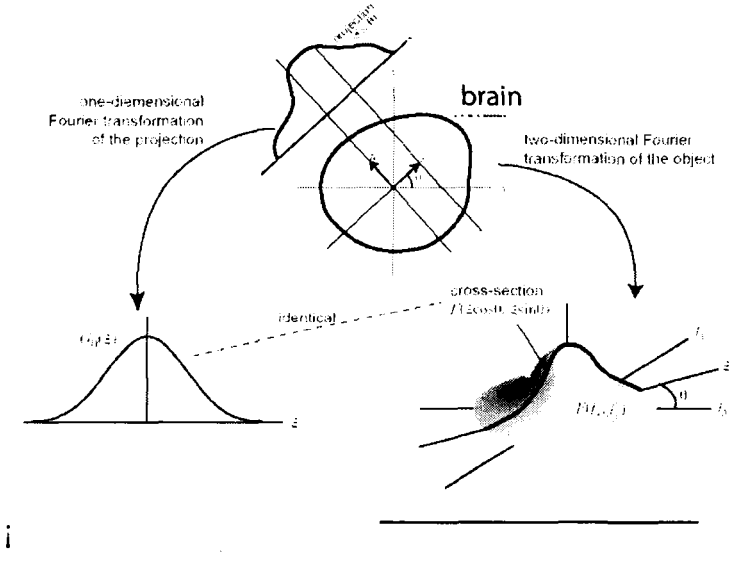


Figure 4: Application of the projection theorem. From [4].

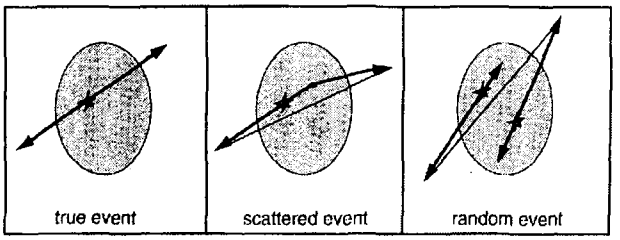


Figure 5: Three classes of coincidence events are detected by PET scanners. In addition to the true coincidence events, scattered and accidental (also known as random) events may be detected. Accidental coincidences randomly add events to the data set, whereas scatter results in the misspositioning of coincidence events. These types of events must be corrected for in order to obtain fully quantitative images. From Cherry and Phelps 1995.

3 Physical Degradations and Distortions

In a realistic data analysis, the influence of noise and of several other distortions of the signal has to be corrected. In addition to the good events detected by the PET scanner, two other classes of events contaminate data. There are the scattered events in which one or both of the .5 MeV gammas scatter on their way through the tissue, but are still detected. Such events are incorrectly positioned and produce reduction of contrast. The second class contains unrelated accidental coincidences producing also erroneous source positions, see Fig. 5.

In Fig. 6, the count rate versus in for a 3 dimensional brain study is presented. After correction for the effects of scattered and random coincidences, the lower curve is achieved. After this correction is performed, structures below twice the spatial resolution of the PET scanner will still suffer from partial volume effects. With decreasing resolution, there is underestimation of concentration, especially for small objects, see Fig. 7. The dimension of data acquisition is important as well. In Fig. 8, a two and a three dimensional PET acquisition are compared. The three dimensional method has better signal to noise ratios since the coincidences leaving the plane are also collected.

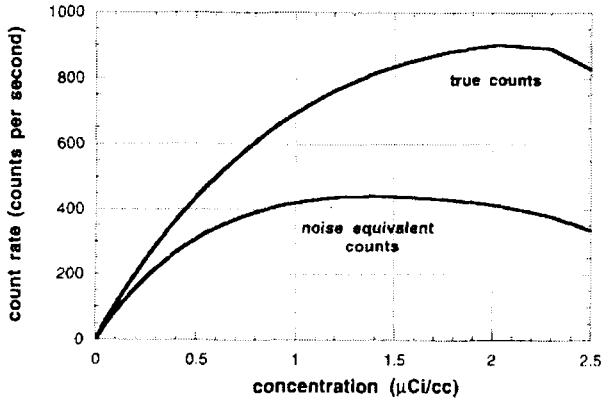


Figure 6: Count rate performance of the GE Advance scanner for 3D brain studies. The coincidence count rate peaks at 900,000cps at an activity concentration of 2μ Ci/ml. The lower curve is after correcting for the effects of scattered and random coincidence events and is known as the "noise effective count rate". This peaks at just over 4000,000 cps at a concentration of 1.25μ Ci/cc. In the absence of dead time, the relationship between concentration and counts would be linear. The plot shows that for this particular scanner, does which lead to concentrations greater than 1μ Ci/cc give no improvement in the number of events recorded by the scanner and only serve to increase the dose to the subject. The injected dose for a PET study should therefore be chosen to maximize the noise equivalent count rate while minimizing the dose to the subject. Data courtesy of Charles Stearns, GE Medical Systems, and Tom Lewellen and Steve Kohlmyer, University of Washington.

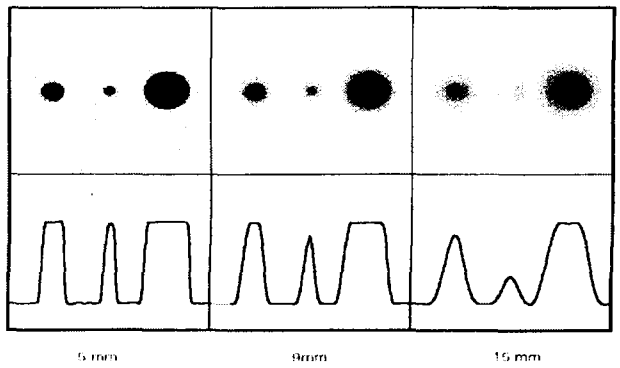


Figure 7: Illustration of the partial volume effect. The three cylinders are 1, 2 and 4*cm* in diameter and have identical concentrations as shown in the high resolution image at left. As the resolution degrades, the smaller cylinders progressively show an underestimation of the concentration. Brain structures are generally small and irregularly shaped and will suffer this same effect to different degrees. From [1].

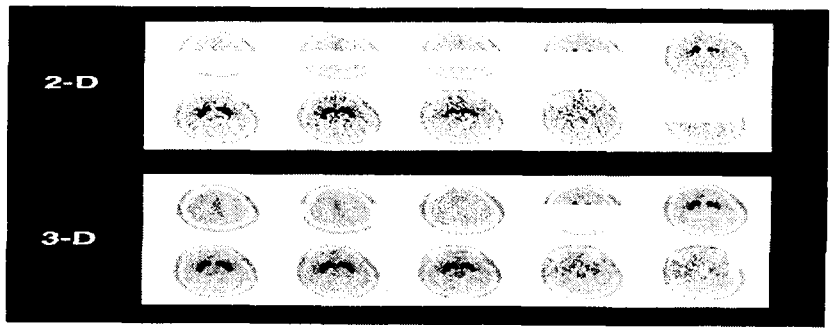


Figure 8: Comparison of 2D and 3D data acquisition and reconstruction in a L-6-[^{18}F]luoroDOPA study. A 20-min scan was acquired 1 hr postinjection using 2D acquisition, immediately followed by another 20-min scan using 3D acquisition. There is a substantial increase in signal to noise in the 3D data set for the same scan length and injected dose relative to the 2D data set. From [1].

4 What is 'Molecular Imaging'?

The following shortened citation from the Editorial 2003 of the official journal of the Academy of Molecular Imaging "Molecular Imaging and Biology" outlines today's issues of Molecular Imaging:

> The primary objective is to provide a forum to the discovery of molecular mechanisms of health and diseases through the use of imaging techniques. Some areas that are covered are: Molecular imaging investigations (i.e. PET, SPECT, MRI, etc.) of macromolecular targets (i.e. genes, receptors, enzymes, etc.) involved in significant biological processes. Technology involved in the design, synthesis, and evaluation of molecular probes used to recognize, image, and investigate macromolecular targets and the functions they perform. The overall objective is to translate basic science discoveries into molecular imaging of disease in patients, both to investigate the biological nature of disease and to establish new molecular imaging diagnostic procedures.

This is exemplified in the paper Ghambir (1999): The positron emitting compound FGCV (8-[18F]fluoroganciclovir) was used as a PET tracer to image the expression of the herpes simplex virus 1 thymidine kinase enzym (HSV1-TK). The tracer qualities were demonstrated in experiments which compared C6-stb-tk+ cells containing large amount of HSV1-TK with control cells. In addition, PET studies on living mice were performed, see Fig. 9, and compared with the sectioned autoradiography. FGCV proved high mobility in the biological target and strong affinity to HSV1-TK. HSV1-TK has been used as a key product-converting enzyme for a number of anticancer therapy approaches. The enzyme can convert less toxic ganciclovir into toxic compounds that result in cell death. Imaging the HSV1-TK activity can serve as fast control and can induce an improvement of toxic cancer therapy.

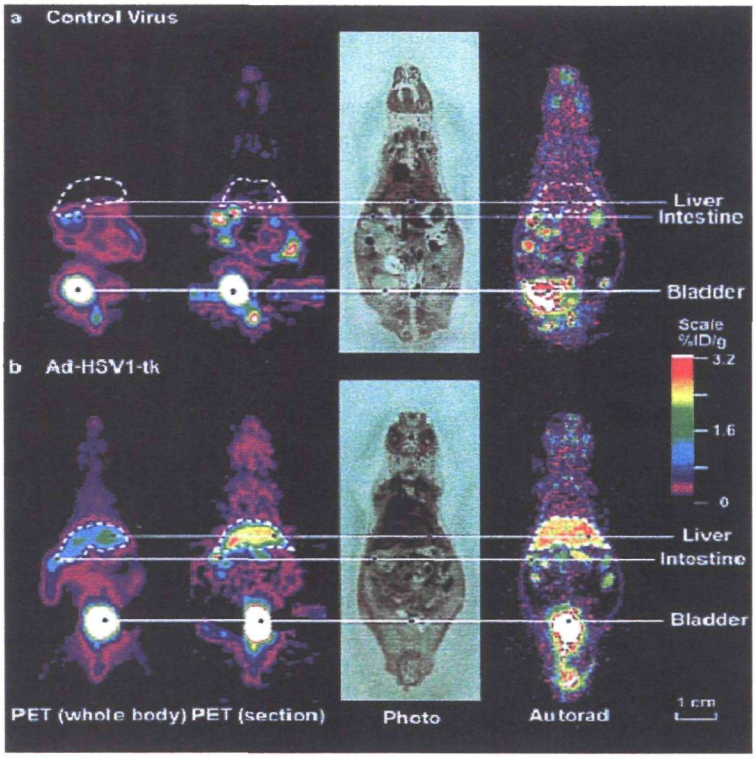


Figure 9: MicroPET and DWBA images of mice after AdCMV-HSV1-tk and control virus administration. Swiss-Webster mice were injected via the tail vein with 1.5×10^9 pfu of control virus (a) or 1.5×10^9 pfu of AdCMV-HSV1-tk virus (b). For each mouse, a whole-body mean coral projection image of the fluorine-18 activity distribution is displayed on the left. The liver outline, in white, was determined for both the FGCV signal and cryostat slice. The second images from the left are coronal sections, approximately 2-mm thick, from the microPET. After their PET scans, the mice were killed, frozen, and sectioned. The next images are photographs of the tissue sections (45-μm thickness) corresponding to approximately the midthickness of the microPET coronal section. The images on the right are DWBA (DWBA means 'digital whole-body autoradiography'). of these tissue sections. The color scale represents the FGCV % ID/g. Images are displayed on the same quantitative color scale to allow signal intensity comparisons among them. From [5].

References

[1] S. R. CHERRY & M. E. PHELPS (1995): Imaging Brain Function with PET. *Brain Mapping - the Methods*, Eds A. Toga & J. Mazziotta, Academic Press, 191–221.

[2] S.R. DEANS (1983): *The Radon Transform and Some of its Applications.* John Wiley & Sons: New York etc.

[3] S. E. DERENZO ET AL. (1982): Dynamic positron emission tomography in man using small bismuth germanate crystals. *Positron Annihilation*, Eds. G. Coleman, S. Sharman & L. Diana, North-Holland-Publ., 935–941.

[4] A. K. JAIN (2002): Radon transformation and projection theorem. *Pattern Information Processing*, Session 13, 1–4.

[5] S.S. GHAMBIR ET AL. (1999): Imaging advenoviral-directed reporter gene expression in living animals with positron emission tomography. *Proc. Natl. Acad. Sci. USA*, 96: 2333–2338.

[6] G. WINKLER(2004): *Mathematische Grundlagen der Radontransformation.* This volume.

Kompartmentmodelle in der Positronen-Emissions-Tomographie

Johannes Müller*

Zusammenfassung

Die wichtigsten Ideen der Modellbildung für Tracer-Dynamiken, gemessen von Positronen-Emissions-Tomographen, werden beschrieben. Ebenso werden die am häufigsten eingesetzten Verfahren zur Parameterschätzung dargestellt. Abschließend werden einige offene Probleme zusammengetragen.

*TU München, Zentrum für Mathematik,
Boltzmannstraße 3,
D-85748 Garching.
Johannes.Mueller@gsf.de, http://www-m12.ma.tum.de/ mueller/

Inhaltsverzeichnis

Kapitel 1

Einleitung

Mit der Positronen-Emissions-Tomographie (PET) stellt die Nuklearmedizin ein wichtiges diagnostisches Instrument zur Verfügung. Das Prinzip solcher Geräte beruht auf dem Zerfall gewisser radioaktiver Stoffe, bei dem Positronen freigesetzt wird. Rekombiniert das Positron mit einem Elektron, so entstehen zwei Gamma-Quanten, die in entgegengesetzte Richtungen abgestrahlt werden. Diese Gamma-Quanten werden von einem Ring aus Detektoren aufgefangen. Da eine Positron-Elektron-Rekombination zwei Quanten gleichzeitig erzeugt, können relevante Ereignisse von Hintergrundstrahlung recht zuverlässig unterschieden werden: zwei Detektoren müssen nahezu gleichzeitig ansprechen. Da weiter die beiden Quanten praktisch diametral entgegengesetzte Geschwindigkeiten haben, muss die Rekombination auf der direkten Verbindungsgeraden zwischen den beiden aktivierten Detektoren liegen. Grundsätzlich wird die Auflösung eines PET-Scanners allerdings durch die mittlere freie Fluglänge eines Positrons von 1-2 mm beschränkt [23].

Mit Hilfe adaptierter Varianten der Radon-Transformation kann (zumindest im Prinzip) die räumliche Dichte der radioaktiven Zerfälle bestimmt werden. Dabei entstehen durch die schlechte Gestelltheit des Problems die gleichen Schwierigkeiten, die sich auch bei Computer-Tomographie ergeben [17]. Auf diese Probleme soll hier nicht weiter eingegangen werden.

Die radioaktiven Substanzen sind an gewisse Tracer gebunden. Diese Tracer werden in die Blutbahn gebracht. Anhand der Dichte des Tracers können dann letztlich Bereiche des Gewebes (z.B. ein Tumor-Gewebe) im Körper lokalisiert werden. Häufig verwendet man einen Tracer, der zwar in den meisten Gewebetypen nachweisbar sein wird, dessen Dynamik sich aber je nach Gewebetyp unterscheidet. In diesem Fall muss die Dynamik modelliert und deren Parameter geschätzt werden.

Das Ziel der vorliegenden Arbeit ist ein Überblick über einige der momen-

tan wichtigsten Prinzipien der Modellierung bzw. Parameterschätzung. Eine vollständige Behandlung kann kaum angestrebt werden, da es sehr viele (zum großen Teil auch Tracer-spezifische) Varianten existieren; viele weichen auch nur in Details voneinander ab. Wir orientieren uns u.a. an den Übersichtsartikeln [19, 28]. In Paragraph 2 beschreiben wir die Messung und den Typ von Daten, den wir erhalten. In Paragraph 3 werden Beispiele für einige häufig genutzte Modelle gegeben, bzw. die Grundstruktur der Modelle erläutert. Methoden zur Parameterschätzung diskutieren wir in Paragraph 4. Im letzte Paragraph 5 werden schließlich einige spezielle Probleme angerissen.

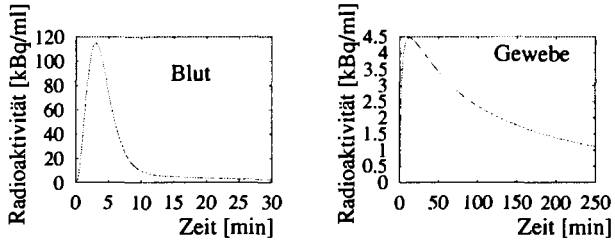


Abbildung 1.1: Skizze der Tracer-Dynamik im Blut (links) und im Gewebe (rechts). Daten dieses Types finden sich z.B. in der Publikation [3].

Kapitel 2

Datenerhebung durch PET-Scanner

Radioaktiv markierter Tracer wird in die Blutbahn gebracht. Durch das Blut verteilt sich die Substanz schnell im Körper. Dort diffundiert der Tracer in das Gewebe, und kann dort z.B. an spezifische Rezeptoren binden. Gleichzeitig filtert die Niere den Tracer wieder aus dem Blut. Ein Beispiel für die Zeitskalen und Größenordung ist in Abb. 1.1 skizziert. Man erkennt im Blut ein scharfes Maximum nach etwa 5 Minuten, das schnell wieder abklingt. Im Gewebe stellt man ebenfalls einen raschen Anstieg fest, der wesentlich langsamer wieder abklingt und dessen Maximum wesentlich niedriger liegt.

Die Zerfälle werden kontinuierlich gezählt, und alle 5-10 Minuten zu einem Datenpunkt zusammengefasst. Um die Radioaktivität im Blut zu bestimmen, wird Blut abgenommen (invasive Methoden). Da diese Prozedur für Probanden und Patienten sehr unangenehm ist, versucht man diese Blutabnahme zu vermeiden, und durch die Messung von Referenzregionen zu ersetzen (nicht-invasive Methoden). Wir werden darauf im Kapitel 4 noch näher eingehen.

Die beiden wichtigsten Merkmale der Daten sind also:

(1) Im Blut gibt es eine hohe, schnell wieder abklingenden initialen Spitze der Konzentration des Tracers. Die Zeitskala hier beträgt etwa 5-10 Minuten.

(2) Im Gewebe ist die Zeitskala wesentlich langsamer und liegt etwa bei ein bis zwei Stunden.

Kapitel 3

Modelle für die Tracer-Dynamik

Hier betrachten wir verschiedene Modelle für die Dynamik des Tracers in einem Voxel, i.e. an einer spezifischen Stelle im Körper. Der Tracer diffundiert vom Blut in das Gewebe und wieder zurück. Im Gewebe selbst kann der Tracer - je nach Typ - z.B. an spezifische Rezeptoren binden. Man muss die Dynamik eines jeden Tracers getrennt modellieren (weitere Beispiele, wahllos herausgegriffen, findet man etwa in [5, 20, 10]). Tatsächlich ist die prinzipielle mathematische Struktur der Modelle aber ähnlich.

3.1 Zwei häufige Modelle

Zwei Situationen sind besonders häufig:

(1) Der Tracer diffundiert in das Gewebe und wieder hinaus. Innerhalb des Gewebes findet keine spezifischen Reaktionen mehr statt. Verschiedene Gewebetypen unterscheiden sich nur durch die Raten, mit der der Tracer hinein- bzw. wieder hinaus diffundiert.

(2) Nachdem der Tracer in das Gewebe hineindiffundiert ist, kann er an einen spezifischen Rezeptor binden. Damit bleibt dieser spezifisch gebundene Tracer im Allgemeinen wesentlich längere Zeit im Gewebe als unspezifisch gebundener Tracer.

3.1.1 Keine Spezifische Bindung

Die wohl einfachste Situation ist gegeben, falls Tracer in das Gewebe hinein- und wieder hinaus diffundieren kann, aber keine weiteren Reaktionen stattfinden. Wir werden dieses Modell im Folgenden als Modell I bezeichnen.

Man unterscheidet zwei verschiedene Klassen (Kompartmente) des Tracers: Die Menge des Tracers im Blut, und die Menge des Tracers im Gewebe. Wir nehmen an, daß der Tracer zum Zeitpunkt $t = 0$ in die Blutbahn gebracht wurde. Sei (siehe Abb. 3.1)

$C_a(t)$	Menge des Tracers im Blut
$C_1(t)$	Menge des Tracers im lokalen Gewebe
k_1	Übergangsrate des Tracers vom Blut in das Gewebe
k_2	Übergangsrate des Tracers vom Gewebe in das Blut.

Dann wird das Modell beschrieben durch eine Differentialgleichung für $C_a(t)$,

$$\dot{C}_1(t) = -k_2 C_1(t) + k_1 C_a(t), \qquad C_1(0) = 0. \tag{3.1}$$

Der Verlust von Tracer durch radioaktiven Zerfall kann vernachlässigt werden. Falls invasive Methoden (Blutabnahme) verwendet wurden, so ist $C_a(t)$ eine bis auf Messfehler bekannte Funktion; bei nicht-invasivem Vorgehen ist $C_a(t)$ unbekannt.

$C_1(t)$ beschriebt die Radioaktivität in einem Voxel (kleinste aufgelöste Volumeneinheit), falls das Gewebe im Voxel homogen ist. Tatsächlich finden wir aber noch mindestens eine weitere Komponente in dem Voxel, nämlich das Blut. Das gemessene Signal ist also (bestenfalls) eine gewichtete Summe von $C_a(t)$ und $C_1(t)$,

$$\text{Signal}(t) \propto C_1(t) + \epsilon\, C_a(t).$$

Dabei ist der Koeffizient ϵ im Allgemeinen klein - es finden sich Aussagen in der Literatur derart, daß etwa 5% des Gewebes aus Kapillaren bestehen [19]. Oft nimmt man bei der Analyse $\varepsilon \approx 0$ an (siehe Kapitel 4).

3.1.2 Spezifische Bindung möglich

Der Tracer diffundiert in das Gewebe, genauso wie es das Modell I voraussetzt. Im Gewebe liegt er zunächst in einer nicht spezifisch gebundenen Form vor. Der Tracer vermag sich nun an spezielle Rezeptoren zu binden. Diese Bindung kann nur schwer aufgelöst werden, spezifisch gebundener Tracer verweilt im Gewebe länger als der unspezifisch gebundener Tracer. Wir führen innerhalb des Gewebes zwei Kompartmente ein, und erhalten (mit dem Kompartment, das das Blut darstellt) insgesamt drei Kompartmente (siehe Abb. 3.2):

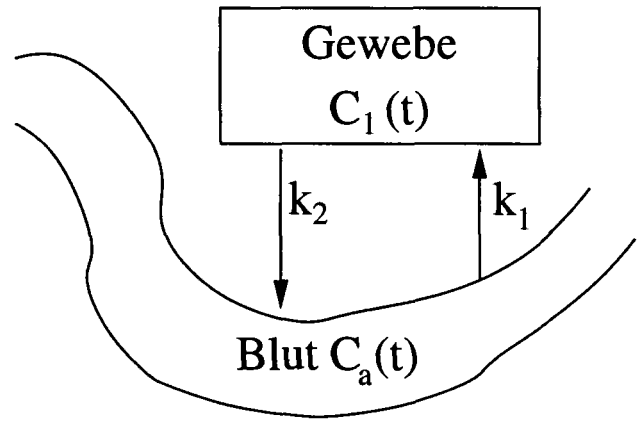


Abbildung 3.1: Struktur des Kompartmentmodells, falls keine spezifische Bindung o.ä. stattfindet (einfachstes Modell).

$C_a(t)$	Menge des Tracers im Blut
$C_1(t)$	Menge des Tracers im lokalen Gewebe, nicht spezifisch gebunden
$C_2(t)$	Menge des Tracers im lokalen Gewebe, spezifisch gebunden
k_1	Übergangsrate des Tracers vom Blut in das Gewebe
k_2	Übergangsrate des Tracers vom Gewebe in das Blut
k_3	Übergangsrate des Tracers unspezifischer Form zur spezifischen Bindung
k_4	Übergangsrate des Tracers von spezifischer Bindung in unspezifische Form.

Dieses Modell, das Standard-Modell der Tracer-Dynamik, bezeichnen wir später als Modell II. Für die Kompartmente C_1 und C_2 erhalten wir Differentialgleichungen

$$\begin{aligned} \dot{C}_1(t) &= -(k_2 + k_3)C_1(t) + k_4 C_2(t) + k_1 C_a(t) \\ \dot{C}_2(t) &= -k_4 C_2(t) + k_3 C_1(t) \end{aligned}$$

Wenn wir wieder annehmen, daß der Versuch zum Zeitpunkt $t = 0$ begann, so erhalten wir die Anfangsbedingung

$$C_1(0) = 0, \qquad C_2(0) = 0.$$

Das gemessene Signal ist nun eine Überlagerung der Beiträge der Kompartmente C_1 und C_2, zusammen mit einem (in der Regel kleinem) Beitrag des Signal aus dem Blut,

$$\text{Signal}(t) \propto C_1(t) + C_2(t) + \epsilon\, C_a(t).$$

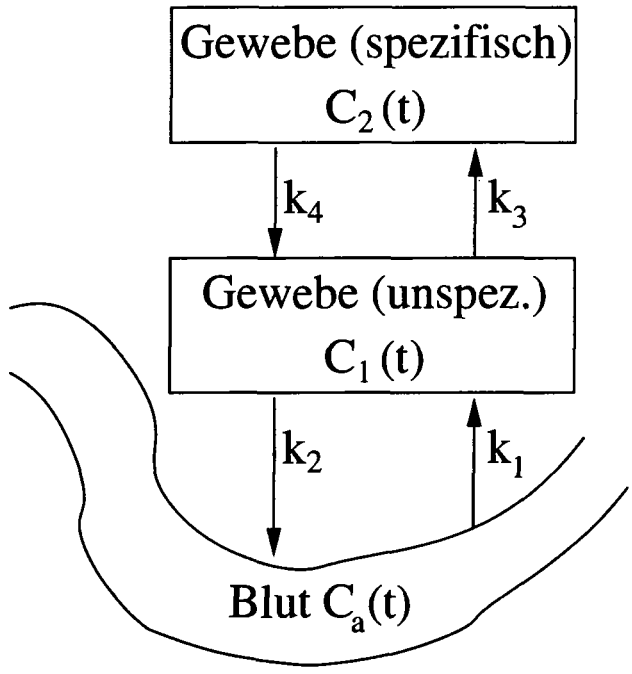


Abbildung 3.2: Struktur des Kompartmentmodells, falls auch spezifische Bindung möglich ist.

3.2 Allgemeine Struktur

Die Modelle nehmen die Form eines linearen, inhomogenen Differentialgleichungssystems an. Die Inhomogenität ist durch die Dichte des Tracers im Blut gegeben; die gemessenen Daten werden durch eine Projektion der Lösung der gewöhnlichen Differentialgleichung auf einen eindimensionalen Unterraum beschrieben. Das System fällt in die Klasse der Systemtheorie (siehe auch Gunn et al. [8, 9]),

$$\dot{x} = A\,x + y\,C_a(t), \qquad x(0) = 0.$$

Wir nehmen an, dass sich die Dynamik des Tracers im Gewebe durch n Kompartmente gut beschreiben lässt. Der Vektor $x(t) = (x_1(t), .., x_n(t))^T$ stellt die Dichte des Tracers zur Zeit t in den n Kompartmente dar. Der zeitlich konstante Vektor y beschreibt, in welche der n Kompartmente der Tracer direkt vom Blut aus gelangen kann. Das gemessene Signal, schließlich, ist die Summe der Signal aus allen Komponenten des Vektors, mit einer zusätzlichen Komponente des Signals aus dem Blut. Sei $\mathbf{e} = (1, 1, .., 1)^T$ der Vektor mit allen Einträgen gleich eins, dann können wir das gemessene Signal

darstellen als

$$\text{Signal}(t) \propto \mathbf{e}^T x(t) + \epsilon\, C_a(t).$$

Die Fragen der Systemtheorie (Identifizierbarkeit der Parameter, Rekonstruktion der Dynamik; weniger die Kontrollierbarkeit des Systems) sind auch im vorliegenden Fall von Interesse [8, 12]. Wir werden diese Fragen hier nur am Rande streifen und nicht systematisch untersuchen.

Bezeichnungen: Im Folgenden werden wir

$$S(t) = \mathbf{e}^T x(t) + \epsilon\, C_a(t), \qquad \overline{C}(t) = \mathbf{e}^T x(t)$$

benutzen. D.h., $S(t)$ steht für das gemessene Signal und $\overline{C}(t)$ für die radioaktive Substanz summiert über alle Kompartmente. In einigen Fällen wird parallel eine interessierende Region (Region Of Interest, ROI) und eine Referenz-Region betrachtet; erst Kombinationen aus Schätzungen beider Regionen liefert ein Ergebnis. In diesem Fall werden wir die Größen der Referenz-Region mit einem Strich kennzeichnen, also z.B. k_1' statt k_1 und $C_1'(t)$ statt $C_1(t)$ etc. schreiben.

3.2.1 Heuristische Folgerung für die Dynamik

Wir haben gesehen, daß der Tracer im Blut ein Maximum kurz nach Beginn des Versuchs besitzt, um schnell abzufallen. Danach wird die Dynamik, die die Evolution des Zustandes im Gewebe bestimmt, approximativ autonom werden. In Anlehnung an wie [18] unterschieden wir - heuristisch - drei Phasen des Tracer-Dynamik (hier nur für Modell II skizziert)

(I) Der Tracer des Blutes "läd" das erste Kompartment C_1 mit Tracer "auf". In dieser Phase wird die Dynamik von der Inhomogenität dominiert. Nur das Kompartment, das direkten Kontakt mit dem Tracer im Blut besitzt (C_1) wird stark verändert, das restliche Modell (C_2) bleibt einigermassen konstant.

(II) Nun ist der Tracer im Blut weitgehend verschwunden, das System ist annähernd autonom. Die Konzentration des Tracers im Blut – falls er überhaupt noch von Bedeutung ist – ändert sich auf einer sehr langsamen Zeitskala und kann in etwa als konstant angenommen werden. Der Zustand läuft entlang der schnellen Mannigfaltigkeit auf die langsame Mannigfaltigkeit zu.

(III) Jetzt ist der Zustand des Systems in der Nähe der langsamen Mannigfaltigkeit, und läuft entlang dieser wieder in den Ruhezustand zurück.

Für den Fall von zwei Kompartmenten im Gewebe / Modell II des letzten Kapitels ($n = 2$) sind diese Phasen in Abb. 3.3 angedeutet.

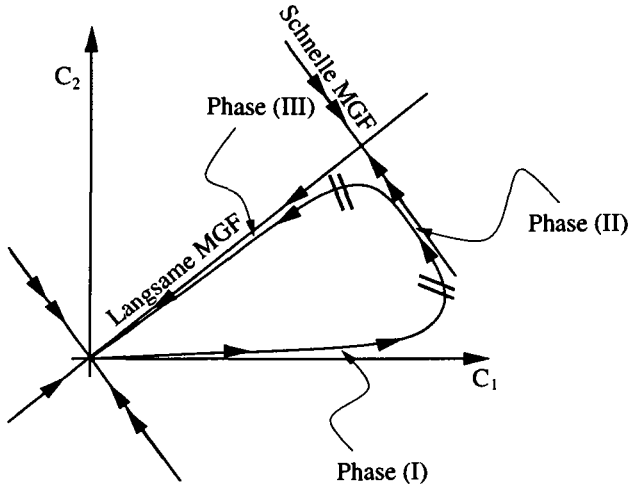


Abbildung 3.3: Phasen der Tracer-Dynamik (Modell I). Die schnelle / langsame Mannigfaltigkeiten sind durch Linien mit zwei Pfeilköpfen / einem Pfeilkopf gekennzeichnet. Der Zustand des Systems folgt der glatten Kurve, die im Nullpunkt beginnt, und zu diesem wieder zurückkehrt. Der Übergangzwischen den drei Phasen sind durch doppelte Querlinien angedeutet.

Kapitel 4

Schätzen von Parametern

4.1 Zu schätzende Größen:

Das primäre Ziel der Parameterschätzung ist nicht die Identifikation aller Parameter des Systems. Sollte dies gelingen, so mag man dadurch zusätzliche, interessante Informationen erhalten. Das eigentliche Ziel ist die Unterscheidung von kranken und gesunden Gewebe, bzw. die Identifikation von Gewebe mit speziellen Eigenschaften. Dafür genügt es in der Regel eine Kombination der Parameter zu schätzen, das heißt letztlich nur einen numerischen Wert (für ein gegebenes Voxel) zu bestimmen. Daher können die Informationen aus den Messungen einigermassen intensiv genutzt werden.

Modell I

Im Prinzip sind also alle Masszahlen sinnvoll, die den Typ des Gewebes gut unterschieden. Es hat sich herausgestellt, dass die Verhältnisse der Konzentrationen im Gleichgewicht (welche natürlich kaum vorkommen) in der Regel gute Kandidaten für diese Aufgabe sind. Für Modell I (Paragraph 3.1.1) finden wir, wenn wir in einem Gedankenexperiment annehmen, dass $C_a(t) \equiv C_a$ fest gegeben ist, asymptotisch

$$0 = \dot{C}_1 = -k_2 C_1 + k_1 C_a$$

und also

$$V_1 := \lim_{t\to\infty} \frac{C_1}{C_a} = \frac{k_1}{k_2}.$$

Daher werden wir für Modell I die Größe $V_1 = k_1/k_2$ als charakteristische Größe definieren, und versuchen, diese aus den Daten zu schätzen. Man muss sich darüber im Klaren sein, dass wir bei den Herleitungen dieser charakteristischen Größe keinen direkten Bezug zu den Messungen hatten, sondern

ein Gedankenexperiment betrachteten, und bei gegebenen, konstanten Tracerwerten im Blut das Verhältnis zwischen Tracer im Blut und Tracer im Gewebe bestimmten. Später werden wir sehen, dass diese Eigenschaft der Zielgrößen tatsächlich von sogenannten Gleichgewichtsmethoden zu deren Schätzung auch direkt ausgenutzt wird (Paragraphen 4.5.3 und 4.5.4).

Es ist plausibel, das Verhältnis der Konzentrationen zu betrachten, da man ansonsten nicht zwischen zwei verschiedenen Experimenten vergleichen kann - es ist unmöglich, C_a so genau zu normieren, dass direkte Vergleichbarkeit der Konzentration C_1 gewährleistet ist. Man muss nun eine Methode finden, V_1 aus den zur Verfügung stehenden dynamischen Daten zu schätzen.

Modell II

Je nach Methode der Messung (invasiv oder nicht-invasiv) bzw. Methode des Auswertens werden verschiedene Zielgrößen geschätzt. Allen ist gemein, daß sie insbesondere Informationen über das Verhältnis von k_3 und k_4 geben; die Unterschiede in diesen Parametern reflektieren die verschiedene Gewebetypen.

Gleichgewichtskonzentrationen:
Wir erhalten, wenn wir uns $C_a(t) \equiv C_a$ als konstant denken, asymptotisch

$$\begin{aligned} 0 = \dot{C}_1(t) &= -(k_2 + k_3)C_1(t) + k_4 C_2(t) + k_1 C_a(t) \\ 0 = \dot{C}_2(t) &= -k_4 C_2(t) + k_3 C_1(t) \end{aligned}$$

Definiere nun die Verhältnisse der Konzentrationen im Gleichgewicht

$$V_1 = C_1/C_a|_{\text{Gleichgewicht}}, \qquad V_2 = C_2/C_a|_{\text{Gleichgewicht}}.$$

(Bemerkung: Wir benutzten hier wie für Modell I den Buchstaben V_1 für das Verhältnis von C_a und C_1. Das ist gerechtfertigt, da Modell I für $k_3 = 0$ ein Grenzfall von Modell II ist). Dividieren wir beide Gleichungen durch C_a, so erhalten wir

$$0 = -(k_2 + k_3)V_1 + k_4 V_2 + k_1, \qquad 0 = -k_4 V_2 + k_3 V_1$$

und daher

$$\frac{V_1}{V_2} = \frac{k_4}{k_3}, \qquad V_1 = \frac{k_1}{k_2}, \quad \Rightarrow \quad V_2 = \frac{k_1 k_4}{k_2 k_3}.$$

Im Fall von Modell II wird V_2 als charakteristische Größe definiert, und man muss einen Weg finden, aus den gegebenen Messwerten diese Größe zu schätzen.

Bindungspotential:

Das Bindungspotential BP ist definiert als der Quotient zwischen k_3 und k_4, d.h. als Quotient der Konzentrationen von C_1 und C_2 im Gleichgewicht,

$$BP = \frac{k_3}{k_4} = \frac{V_1}{V_2}.$$

"Distributed Volumen Ratio":
Die "Distributed Volumen Ratio" (DVR) ist definiert als

$$\frac{V_1 + V_2}{V_1} = 1 + BP^{-1}$$

Auch diese Größe wird zur Beurteilung des Gewebetyps in dem betreffenden Voxel eingesetzt.

"Gjedde-Patlak Analyse":
Einige Tracer gehen eine irreversiblen Bindung ein ($k_4 = 0$). In diesem Fall werden die obigen Masszahlen trivial, da die rate k_4 multiplikativ auftaucht. Gjedde [6] und Patlak [24, 25, 29] schlagen (fußend auf Arbeiten von Sokoloff) den Ausdruck

$$\frac{k_1 k_3}{k_2 + k_3}$$

als charakterisierende Größe vor.

4.2 Experimentelle Voraussetzungen und Techniken der Schätzung

Experimentelle Voraussetzung der Schätzung:
Man kann zwei von den Voraussetzungen her unterschiedliche Situationen unterschieden: den invasiven Ansatz, bei dem durch Blutabnahme die Menge des Tracers im Blut $C_a(t)$ bekannt ist, und den nicht-invasiven Ansatz, bei dem auf die Blutabnahme verzichtet wird. In der Regel ergibt der invasive Ansatz ergibt genauere Daten. Er mathematisch einfacher zu handhaben und liefert meistens präzisere Resultate. Allerdings wird die Blutabnahme von vielen Patienten als unangenehm empfunden. Man versucht Methoden zu entwickeln, bei denen die Blutabnahme nicht mehr nötig ist, und greift statt auf die direkten Daten über den Tracer im Blut auf Messungen von Referenzregionen zurück. I.e., man ersetzt die Funktion $C_a(t)$ durch das Signal (oder dessen Transformierte) $S'(t)$ der Referenzregion. Insofern ist es meistens relativ einfach, aus einer invasiven Methode die entsprechende nicht-invasive Methode zu entwickeln. Man erkennt in der Tabelle 4.1, dass nicht

Invasive Methode			
Linearer Ansatz	Logan Analyse	§4.3.1	[15, 19]
	Gjedde-Patlak Analyse	§4.3.2	[6, 24, 25, 29]
	Lineares Modell	§4.3.3	[27, 16]
Nicht-linearer Ansatz	Expotentialpolynome	§4.4.1	[18]
	Quadrat-Fehler-Methode	§4.4.2	[18]
	Spektralanalyse	§4.4.3	[3, 21, 9]
Nicht-invasive Methode			
Linearer Ansatz	Logan Analyse	§4.5.1	[15, 19, 9, 13]
	Gjedde-Patlak Analyse	§4.5.2	[25, 29]
	Gleichgewichtsanalyse I	§4.5.3	[19, 28, 4]
	Gleichgewichtsanalyse II	§4.5.4	[19, 28, 4]
Nicht-linearer Ansatz	Vereinfachtes Modell	§4.6.1	[14]
	Basisfunktionen	§4.6.2	[7]

Tabelle 4.1: Die hier besprochenen Herangehensweisen der Datenanalyse, strukturiert nach invasiven/nicht-invasiven Methoden bzw. linearen und nicht-linearen Ansätzen (zusammen mit Literaturangaben resp. dem Paragraphen, in dem die Methode beschrieben wird).

alle invasiven Methoden als nicht-invasive Methode wieder auftauchen (z.B. der Ansatz "lineares Modell"). Dies ist aber kein prinzipielles Problem sondern spiegelt einfach die fehlende Präsenz dieser speziellen nicht-invasiven Methodik in der Literatur wieder.

Techniken für die Schätzung:
Von der Technik des Schätzens her, existieren wiederum zwei unterschiedliche Ansatzweisen: Entweder man versucht die Auswertung auf ein lineares Modell zurückzuführen, wobei (fast immer) der initiale Teil (I) der Dynamik vernachlässigt wird, oft genug auch der zweite Teil (II). Oder man benutzt die gesamte Zeitreihe, und ist dann auf nicht-lineare Methoden wie die Kleinste-Quadrate-Methode angewiesen.

Strukturierung der Methoden:
Um die Vielzahl verschiedener Methoden zu strukturieren, kann man die

beiden eben erläuterten Prinzipien heranziehen (diese Gliederung der Methoden ist natürlich willkürlich, und wird den Ansätzen sicherlich auch nicht völlig gerecht). Da es eine große Vielfalt von Ansätzen gibt, ist eine solche Strukturierung hilfreich. In Tabelle 4.1 sind die hier besprochenen Analysemethoden gelistet. Die größte Verbreitung in der praktischen Anwendung findet die Logan und Gjedde-Patlak Analyse. Hier werden im Wesentlichen das Problem der Parameterschätzung auf die Bestimmung einer Regressionsgeraden zurückgespielt. Der Ansatz "lineares Modell" entwickelt eine skalare, lineare Differentialgleichung höherer Ordnung für das Signal (d.h. die Information über die einzelnen Kompartmente wird nicht mehr benötigt) und integriert diese Gleichung so oft über die Zeit, bis eine lineare Beziehung zwischen dem Signal und dem (mehrfach) zeitintegrierten Signal entsteht. Die Parameter dieser linearen Beziehung sind die zu schätzenden Größen. Die Gleichgewichtsmethoden nutzen lokale Extrema des Signals dazu aus, um das instantane Gleichgewicht zur Schätzung der Verhältnisse der Kompartmente im Gleichgewicht heranzuziehen. Die Methoden "Expotentialpolynome", "Quadrat-Fehler", "Spektralanalyse", "vereinfachtes Modell" und "Basisfunktionen" sind nahe verwandt. Sie alle verwenden die Struktur der Lösung einer linearen nicht-autonomen Differentialgleichung (Faltung von Exponentialsummen mit dem Eingangssignal) und zielen auf die Schätzung der Exponenten dieser Faltung bzw. der Koeffizienten dieser Exponenten. Allein die Methodik der Schätzung unterscheiden sich.

4.3 Invasive Lineare/Graphische Methoden

Lineare Methoden werden häufig "graphisch" genannt, weil im einfachsten Fall die Analyse auf die Bestimmung einer Regressionsgeraden zurückgeführt werden kann. Und eine Regressionsgerade kann man graphisch darstellen.

Das interessanteste Beispiel für eine lineare, nicht-invasive Methode ist die Logan unf Gjedde-Patlak Analyse. Dieser Ansatz ist vor allem deshalb so interessant, da er am häufigsten von allen Analysemethoden eingesetzt wird.

Eine weitere lineare Methode ist bemerkenswert: Hier wird eine Differentialgleichung für das Signal hergeleitet. Da das Signal aus der Summe der Komponenten der Kompartmente besteht, erhält man eine Gleichung höherer Ordnung. Nun werden beide Seiten der Gleichung so lange über die Zeit integriert, bis alle Ableitungen verschwunden sind. Man erhält ein lineares Gleichungssystem, dessen Koeffizienten die zu schätzenden Parameter sind.

4.3.1 Logan Analyse

Hier stellen wir die einfachste Form der Logan Analyse zunächst für Modell I und danach für Modell II dar (siehe [15, 19]).

Modell I:
Wir wollen V_1 bestimmen. Wir kennen aus den Blutproben $C_a(t)$ recht genau, und wir kennen das Signal $C_1(t) + \epsilon C_a(t)$. Gemäß des Modells gilt

$$\frac{d}{dt}C_1 = k_1 C_a - k_2 C_1, \qquad C_1(0) = 0.$$

Integrieren wir beide Seiten nach t, so folgt

$$C_1(t) = k_1 \int_0^t C_a(\tau)\, d\tau - k_2 \int_0^t C_1(\tau)\, d\tau. \tag{4.1}$$

Division durch $k_2\, C_1(t)$ ergibt

$$\frac{\int_0^t C_1(\tau)\, d\tau}{C_1(t)} = \frac{k_1}{k_2}\frac{\int_0^t C_a(\tau)\, d\tau}{C_1(t)} - \frac{1}{k_2} = V_1 \frac{\int_0^t C_a(\tau)\, d\tau}{C_1(t)} - \frac{1}{k_2}. \tag{4.2}$$

V_1 kann nun als Steigung einer Geradengleichung bestimmt werden, wobei

$$\frac{\int_0^t C_1(\tau)\, d\tau}{C_1(t)}, \qquad \frac{\int_0^t C_a(\tau)\, d\tau}{C_1(t)}$$

die Rolle der abhängigen (unabhängigen) Variablen spielen.

Bias: Kennt man $C_a(t)$ und $C_1(t)$, so sind diese Größen natürlich leicht zu berechnen. Wir bestimmen wohl $C_a(t)$ direkt, nicht aber $C_1(t)$. Das Signal des PET-Scanners gibt nur Auskunft über $C_1(t) + \epsilon\, C_a(t)$. Dabei wird ϵ wesentlich bestimmt durch den relativen Anteil von Blut in dem Voxel. Dieser Anteil sollte klein sein. Wenn wir also die initiale Phase I der Dynamik abschneiden, und nur die Phasen II und III verwenden, so wird $\epsilon\, C_a(t)$ keinen großen Beitrag zum Signal liefern, d.h. wir erwarten, daß die Werte nach einer kurzen initialen Phase tatsächlich (approximativ) auf einer Geraden liegen. Ersetzen wir allerdings $C_1(t)$ durch das Signal $S(t) = C_1(t) + \epsilon\, C_a(t)$, so erhalten wir

$$\frac{\int_0^t S(\tau)\, d\tau}{S(t)} = \frac{\int_0^t C_1(\tau) + \epsilon C_a(\tau)\, d\tau}{C_1(t) + \epsilon\, C_a(t)}$$

Nach der initialen Phase, d.h. für $t \geq t_0$, ist $\epsilon\, C_a(t) \approx 0$, nicht aber $\int_0^t \epsilon\, C_a(\tau)\, d\tau$. Also,

$$\begin{aligned}\frac{\int_0^t S(\tau)\, d\tau}{S(t)} &\approx \frac{\int_0^t C_1(\tau) + \epsilon C_a(\tau)\, d\tau}{C_1(t)} = (V_1 + \epsilon)\frac{\int_0^t C_a(\tau)\, d\tau}{C_1(t)} - \frac{1}{k_2} \\ &\approx (V_1 + \epsilon)\frac{\int_0^t C_a(\tau)\, d\tau}{S(t)} - \frac{1}{k_2} \qquad \text{für } t \geq t_0.\end{aligned}$$

Das Blut läßt uns V_1 tendenziell überschätzen.

Modell II:
Für Modell II genügen die Informationen aus dem Blut und dem zu untersuchenden Voxel nicht allein. Wir benötigen zusätzlich noch die Messung einer Referenzregion, bei der wir annehmen können, daß $k_3 = 0$ ist. Es gehen also drei verschiedene Messungen in die Auswertung eines Voxels aus der interessierenden Region ("Region Of Interest", ROI) ein:

(1) Die Messung des Blutes $C_a(t)$.

(2) Die Messung einer Referenzregion. Wir werden die Größen aus der Referenzregion mit einem Strich kennzeichnen, also C_1', k_1' etc. verwenden.

(3) Die Messung eines Voxels aus der ROI. Für die Größen aus dem ROI benutzen wir die ursprünglichen Bezeichnungen, d.h. $C_1(t)$, k_1 etc.

Annahme: Wir nehmen an, daß $k_3' = 0$ und

$$V_1' = \frac{k_1'}{k_2'} = \frac{k_1}{k_2} = V_1.$$

Ohne diese Annahme können wir die Informationen aus dem Referenzbereich nicht auf die im ROI beziehen. In gewisser Weise ist diese Annahme willkürlich, und nicht streng zu begründen. Auf der anderen Seite scheint dieser pragmatische Ansatz gute Resultate zu erzielen, und daraus seine Rechtfertigung zu beziehen.

(A) Referenz Region: (Schätzung von V_1)
In der Referenzregion haben wir $k_3' = 0$, d.h. hier liegt Modell I vor. Wir haben im letzten Abschnitt mit Hilfe der Logan Analyse für Modell I eine Methode zum Schätzen von V_1 hergeleitet.

(B) ROI: (Schätzung von $V_1 + V_2$)
Im ROI können wir nicht mehr $k_3 = 0$ annehmen. Der Zustand im ROI ist somit bestimmt durch die drei Kompartmente $C_a(t)$, $C_1(t)$ und $C_2(t)$, das Signal durch

$$S(t) = \epsilon\, C_a(t) + C_q(t) + C_2(t) \approx C_1(t) + C_2(t)$$

Mit

$$A = \begin{pmatrix} -(k_2 + k_3) & k_4 \\ k_3 & -k_4 \end{pmatrix}, \qquad \mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

finden wir

$$\begin{aligned} \frac{d}{dt}\begin{pmatrix} C_1 \\ C_2 \end{pmatrix} &= A\begin{pmatrix} C_1 \\ C_2 \end{pmatrix} + C_a(t)k_1\mathbf{e}_1 \\ \Rightarrow \quad \frac{d}{dt}A^{-1}\begin{pmatrix} C_1 \\ C_2 \end{pmatrix} &= \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} + C_a(t)k_1 A^{-1}e_1 \end{aligned}$$

Sei $\mathbf{e} = (1,1)^T$ und $\overline{C}(t) = C_a(t) + C_1(t)$, so folgt

$$\mathbf{e}^T A^{-1}\begin{pmatrix} C_1(t) \\ C_2(t) \end{pmatrix} = \int_0^t \overline{C}(\tau)\,d\tau + \int_0^t C_a(\tau)\,d\tau \;\; k_1\,\mathbf{e}^T A^{-1}\mathbf{e}_1 \quad (4.3)$$

Dividiert man durch $\overline{C}(t)$ so erhält man

$$\frac{\int_0^t \overline{C}(\tau)\,d\tau}{\overline{C}(t)} = \frac{\mathbf{e}^T A^{-1}\begin{pmatrix} C_1(t) \\ C_2(t) \end{pmatrix}}{\overline{C}(t)} - k_1\,\mathbf{e}^T A^{-1}\mathbf{e}_1\,\frac{\int_0^t C_a(\tau)\,d\tau}{\overline{C}(t)}. \quad (4.4)$$

Da wir ein (in Phase II und III) asymptotisch autonomes gewöhnliches, lineares Differentialgleichungssystem betrachten, wissen wir, dass die Lösung gegen eine exponentiell wachsende Funktion strebt,

$$\begin{pmatrix} C_1(t) \\ C_2(t) \end{pmatrix} \propto e^{\hat{\lambda}t}\begin{pmatrix} \hat{C}_1 \\ \hat{C}_2 \end{pmatrix} \qquad \text{für } t \to \infty,$$

wobei $\hat{\lambda}$ der größte Eigenwert, und $(\hat{C}_1, \hat{C}_2)^T$ der entsprechende positive Eigenvektor darstellt. Also folgt, dass

$$\frac{\mathbf{e}^T A^{-1}\begin{pmatrix} C_1(t) \\ C_2(t) \end{pmatrix}}{\overline{C}(t)} \approx \text{const}$$

und wir finden asymptotisch die lineare Beziehung

$$\frac{\int_0^t \overline{C}(\tau)\,d\tau}{\overline{C}(t)} = \text{const} - k_1\,\mathbf{e}^T A^{-1}\mathbf{e}_1\,\frac{\int_0^t C_a(\tau)\,d\tau}{\overline{C}(t)}.$$

Die Steigung dieser Geraden ist

$$\begin{aligned} -k_1\,\mathbf{e}^T A^{-1}\mathbf{e}_1 &= -k_1\frac{1}{\det(A)}\mathbf{e}^T\begin{pmatrix} -k_4 & -k_4 \\ -k_3 & -(k_2+k_3) \end{pmatrix}\mathbf{e}_1 \\ &= k_1\frac{k_3+k_4}{k_4\,k_2} = \frac{k_1}{k_2} + \frac{k_1\,k_3}{k_4\,k_2} = V_1 + V_2 \end{aligned}$$

Bias: Wir ersetzen wieder $\overline{C}(t) = C_1(t) + C_2(t)$ durch das Signal

$$S(t) = \epsilon\, C_a(t) + C_1(t) + C_2(t) = \epsilon\, C_a(t) + \overline{C}(t).$$

Wieder werden wir einen Bias in die Schätzung tragen. Wenn t groß genug ist, dürfen wir $S(t) \approx \overline{C}(t)$ annehmen, aber $\int_0^t S(\tau)\,d\tau \not\approx \int_0^t \overline{C}(\tau)\,d\tau$. Daher finden wir

$$\begin{aligned}
\frac{\int_0^t S(\tau)\,d\tau}{S(t)} &\approx \frac{\int_0^t \overline{C}(\tau)\,d\tau}{\overline{C}(t)} + \epsilon\,\frac{\int_0^t C_a(\tau)\,d\tau}{\overline{C}(t)} \\
&= \text{const} + (V_1 + V_2)\,\frac{\int_0^t C_a(\tau)\,d\tau}{\overline{C}(t)} + \epsilon\,\frac{\int_0^t C_a(\tau)\,d\tau}{\overline{C}(t)} \\
&\approx \text{const} + (V_1 + V_2 + \epsilon)\,\frac{\int_0^t C_a(\tau)\,d\tau}{S(t)}
\end{aligned}$$

(C) Kombination der Schätzungen aus Referenzregion und ROI: (Schätzung von V_2)
Wir haben nun Methoden, sowohl V_1 als auch $V_1 + V_2$ zu schätzen. Die Differenz dieser beiden Schätzer ergibt einen Schätzer für V_2. In [28] finden sich Daten zu diesem Verfahren. Man erkennt schön, dass die Dynamik sich zunächst in die Asymptotik bewegen muss, bis ein linearer Zusammenhang erkennbar wird.

Bias: Wir sehen, daß die Schätzer für V_1 und $V_1 + V_2$ den gleichen, systematischen Bias beinhalten. Falls ϵ (i.e. der Anteil von Kapillaren im Gewebe des ROI bzw. der Referenzregion) vergleichbar sind, so heben sich diese Fehler gerade auf. Allerdings erhalten wir dadurch, dass wir durch $S(t)$ teilen, weitere Probleme, die zu einer verzerrten Schätzung führen können [28].

4.3.2 Gjedde-Patlak Analyse

Falls der Tracer irreversible bindet, d.h. im Kompartment C_2 verbleibt ($k_4 = 0$), kann die Logan Analyse offensichtlich nicht direkt angewandt werden ($V_2 = 0$). In diesem Fall beschreibt die von Gjedde [6] und Patlak [24, 25, 29] entwickelte Methode ein mögliches Vorgehen, das im Kern dem der Logan Analyse entspricht. Man versucht - unter gewissen Annahmen - aus dem Differentialgleichungssystem eine lineare Beziehung zwischen dem Signal und der Tracerdynamik im Blut (resp. der über die Zeit integrierten Tracerdynamik) zu finden. Steigung dieser linearen Gleichung wird der gewünschte, zu schätzende Parameter sein.

Der von Patlak entwickelte Ansatz erlaubt komplexere Modelle als unser Modell II - es dürfen beliebig viele reversible Kompartmente vorkommen,

solange sich genau ein irreversibles Kompartment im System befindet und das Spektrum der Übergangsmatrix gewisse Eigenschaften erfüllt. Da sich an dem grundlegenden Konzept nichts ändert, werden wir uns das Verfahren für das Modell II ansehen. Ausgehend von den Differentialgleichungen erhalten wir wegen $k_4 = 0$

$$\begin{aligned} C_1(t) &= \frac{k_1}{k_2+k_3}\int_0^t (k_2+k_3)e^{-(k_2+k_3)(t-\tau)}\,C_a(\tau)\,d\tau, \\ C_2(t) &= k_3\int_0^t C_1(\tau)\,d\tau = \frac{k_1k_3}{k_2+k_3}\int_0^t\int_0^\tau (k_2+k_3)e^{-(k_2+k_3)(t-\tau')}\,C_a(\tau')\,d\tau'\,d\tau. \end{aligned}$$

Annahme: Wir nehmen an, daß $k_2 + k_3 \gg 1$ (i.e., dass C_1 sich wesentlich schneller als C_a ändert).

Da für eine Funktion $f(t) \in C^0$ gilt

$$\lim_{k\to\infty} k\int_0^t e^{-k(t-\tau)}\,f(\tau)\,d\tau = f(t)$$

solange $t > 0$, finden wir approximativ für $k_2 + k_3$ groß, dass

$$C_1(t) = \frac{k_1}{k_2+k_3}C_a(t), \qquad C_2(t) = \frac{k_1k_3}{k_2+k_3}\int_0^t C_a(\tau)\,d\tau.$$

Damit folgt

$$S(t) = C_1(t) + C_2(t) + \epsilon C_a(t) = \frac{k_1k_3}{k_2+k_3}\int_0^t C_a(\tau)\,d\tau + \left(\epsilon + \frac{k_1}{k_2+k_3}\right)C_a(t)$$

und daher finden wir eine lineare Beziehung zwischen

$$S(t)/C_a(t) \text{ und } \int_0^t C_a(\tau)\,d\tau/C_a(t)$$

mit dem gesuchten Ausdruck $k_1k_2/(k_2 + k_3)$ als Steigung

$$\frac{S(t)}{C_a(t)} = \left(\frac{k_1k_3}{k_2+k_3}\right)\frac{\int_0^t C_a(\tau)\,d\tau}{C_a(t)} + \left(\epsilon + \frac{k_1}{k_2+k_3}\right).$$

4.3.3 Lineares Modell

Wir stellen hier eine vereinfachte Version dieser Methode, wie sie in [27] beschrieben wird, vor. Wir konzentrieren und auf Modell I. In [27] wird vor allem das Problem des inhomogenen Gewebes adressiert. Dies wollen wir an der Stelle nicht betrachten, kommen aber in Kapitel 5 darauf zurück.

Die Idee dieser Methode ist es, durch Differenzieren eine Differentialgleichung für das Signal zu erhalten. In diese Gleichung soll nur noch das Signal, die Messungen im Blut $C_a(t)$ und deren Ableitungen eingehen.

Annahme: Wir nehmen an, dass $\epsilon = 0$, d.h. $\overline{C}(t) \equiv S(t)$.

Wir suchen eine Differentialgleichung für $\overline{C}(t)$.

Modell I:
Da hier $\overline{C}(t) = C_1(t)$, haben wir direkt die gewünschte Differentialgleichung vorliegen,

$$\dot{\overline{C}}(t) = -k_2\overline{C}(t) + k_1 C_a(t).$$

Wir wählen eine Zerlegung der Zeitachse, d.h. Zeitpunkte

$$0 \le t_0 < t_1 < t_2 < \cdots < t_l \le T$$

wobei T der Zeithorizont, i.e. das Ende der Messung, bezeichnet. Dann integrieren wir die Differentialgleichung von t_i bis t_{i+1}, und erhalten mit den Definitionen

$$\tilde{C}^i = \int_{t_i}^{t_{i+1}} \overline{C}(\tau)\,d\tau, \qquad \Delta C^i = \overline{C}(t_{i+1}) - \overline{C}(t_i), \qquad \tilde{C}_a^i = \int_{t_i}^{t_{i+1}} \overline{C}_a(\tau)\,d\tau,$$

für $i = 0, .., l-1$

die Beziehung

$$\Delta C^i = -k_2\tilde{C}^i + k_1\tilde{C}_a^i.$$

Da man aus dem Signal ΔC^i, $\tilde{C}^i$ und $\tilde{C}_a^i$ bestimmen kann, führt dies auf ein lineares Modell

$$Y = X\,p$$

mit

$$Y = (\Delta C^0, .., \Delta C^{l-1})^T, \qquad X = \begin{pmatrix} \tilde{C}_a^0 & -\tilde{C}^0 \\ \vdots & \vdots \\ \tilde{C}_a^{l-1} & -\tilde{C}^{l-1} \end{pmatrix}, \qquad p = (k_1, k_2)^T.$$

Nimmt man nun an, dass der Fehler bei der Bestimmung der Designmatrix X wesentlich kleiner ist als der bei der Bestimmung der Daten Y (zur Berechnung von X integrieren wir, was zu einer Herausmittelung des Fehlers führt), und nimmt man weiter das übliche Fehlermodell an, so kann man p approximativ durch den Gauss-Schätzer bestimmen,

$$\hat{p} = (XX^T)^{-1}X^TY.$$

Damit haben wir Schätzer $\hat{k}_1$, $\hat{k}_2$ für k_1 und k_2, und können einen naiven Schätzer für V_1 definieren,

$$\hat{V}_1 = \frac{\hat{k}_1}{\hat{k}_2}.$$

Modell II

In diesem Fall müssen wir die Differentialgleichung für $\overline{C}(t) = C_1(t) + C_2(t)$ erst herleiten. Wiederholtes Differenzieren von $\overline{C}(t)$ führt auf

$$\begin{aligned}
\overline{C}(t) &= C_1(t) + C_2(t) \\
\dot{\overline{C}}(t) &= \dot{C}_1(t) + \dot{C}_2(t) = -k_2 C_1(t) + k_1 C_a(t) \\
\ddot{\overline{C}}(t) &= k_2\,(k_2 + k_3) C_1(t) - k_2\,k_4 C_2(t) - k_1\,k_2\,C_a(t) + k_1\,\dot{C}_a(t)
\end{aligned}$$

Aus diesen drei Gleichungen kann man nun $C_1(t)$, $C_2(t)$ eliminieren, und erhält

$$\ddot{\overline{C}}(t) + \frac{k_2(k_2+k_3)+k_1k_4}{k_2}\dot{\overline{C}}(t) + k_1k_4\overline{C} = \frac{k_1}{k_2}(k_2(k_2+k_3)+k_1k_4)C_a(t) + k_1\dot{C}_a(t)$$

bzw. mit

$$\theta_1 = \frac{k_2(k_2+k_3)+k_1k_4}{k_2}, \qquad \theta_2 = k_1k_4, \qquad \theta_3 = k_1$$

haben wir

$$\ddot{\overline{C}}(t) + \theta_1\dot{\overline{C}}(t) + \theta_2\overline{C} = \theta_1\theta_3 C_a(t) + \theta_3\dot{C}_a(t)$$

Im Prinzip können wir nun wie oben fortfahren, d.h. die Differentialgleichung zweimal über die Zeit integrieren, die Zeitachse in Intervalle aufteilen, und ein überbestimmtes Gleichungssystem für θ_1,..,θ_3 herleiten. Im Gegensatz zu oben wird dieses Gleichungssystem nicht mehr linear in den Parametern θ_1, θ_2 und θ_3 sein. Interessanterweise finden wir schon hier den Fall, dass nicht mehr alle Parameter aus den Daten geschätzt werden können: die Schätzung ist nur möglich für drei Größen, das ursprüngliche Modell beinhaltet aber vier Raten. Daher muss man entweder eine Rate aus einer Referenz-Region schätzen, oder aber eine Funktion der Werte θ_1,..,θ_3 als charakterisierende Größe nutzen.

Die Situation ist leicht unterschiedlich, wenn wir schon wissen, dass $k_4 = 0$ ist. Dann haben wir genügend Informationen, um $k_1k_3/(k_2+k_3)$ zu schätzen. Dies wurde u.a. von Logan vorgeschlagen [16].

4.4 Invasive Nicht-lineare Methoden

Die nicht-linearen Methoden benutzen in der Regel die explizite Lösung der Differentialgleichung. Diese Lösung beschreibt die Dynamik des Tracers. Hier nutzt man, dass diese Gleichungen linear und nieder-dimensional sind, sodass sie explizit gelöst werden können.

Allgemeine Struktur der Lösung:
Die Lösung eines Systems gewöhnlichen, linearer Differentialgleichung erster Ordnung mit Inhomogenität $C_a(t)$,

$$\dot{x} = Ax + yC_a(t) \qquad x(0) = 0$$

deren Fundamentalsystem nur reelle Eigenwerte besitzt, kann geschrieben werden als

$$x(t) = \sum_{i=1}^{n} \alpha_i e^{\lambda_i t} \otimes C_a(t),$$

wobei $\otimes$ als Konvolution definiert ist,

$$e^{\lambda_i t} \otimes C_a(t) = \int_{-\infty}^{t} e^{-\lambda_i (t-\tau)} C_a(\tau)\, d\tau.$$

und α_i Projektionen von y auf die entsprechenden Eigenräume bedeutet. Im Folgenden wird wieder $\epsilon = 0$ angenommen, i.e.

$$S(t) = \overline{C}(t) = \sum_{i=1}^{n} (\mathbf{e}^T \alpha_i)\ (e^{\lambda_i t} \otimes C_a)(t).$$

4.4.1 Vernachlässigen des ersten Teils der Zeitreihe

Wenn wir den ersten Teil der Zeitreihe vernachlässigen (i.e. Phase I), so können wir die Konzentration des Tracers im Blut häufig approximativ zu Null setzten [18]. Dann erhalten wir als Lösung der gewöhnlichen Differentialgleichung ein Exponentialpolynom,

$$S(t) = \sum_{i=1}^{n} (\mathbf{e}^T \alpha_i)\ (e^{\lambda_i t} \otimes C_a)(t) \approx \sum_{i=1}^{n} (\mathbf{e}^T \alpha_i) \int_0^{\infty} e^{-\lambda_i \tau} C_a(\tau)\, d\tau e^{\lambda_i t} = \sum_{i=1}^{n} \hat{\alpha}_i e^{\lambda_i t}.$$

Nun können Methoden der Parameterschätzung bei Exponentialpolynomen angewandt werden (siehe etwa [11, 22]). Leider sind diese Methoden nicht weit verbreitet. Bei Summen von Exponentialtermen ist der dominante Term (mit dem kleinsten λ) relativ leicht zu schätzen. Allerdings ist es schwierig, die anderen Terme verlässlich zu bestimmen. Man findet in der Literatur bzgl. PET vor allem die kleinste Fehler-Quadrat Methode [18] und kaum an dieses Problem angepasste Methoden.

4.4.2 Nutzung der gesamten Zeitreihe: Quadrat-Fehler-Minimierung

Mit Hilfe der Minimierung des quadratischen Fehlers kann man die Parameter von $S(t) = \sum_{i=1}^{n} (\mathbf{e}^T \alpha_i)\ (e^{\lambda_i t} \otimes C_a)(t)$ an die empirischen Daten anpassen [18]. Dabei ist zu erwarten, dass man in ähnliche Schwierigkeiten gerät wie beim anpassen von Exponentialsummen. Leider existieren kaum spezialisierte Methoden der Parameterschätzung, die die Struktur von $y(t)$ effizient ausnutzen. Selbst ein adaptiertes Fehlermodell, die das detaillierte Wissen über $C_a(t)$ nutzt findet man kaum (siehe [1]).

4.4.3 Nutzung der gesamten Zeitreihe: Spektralanalyse

Cunningham und Jones untersuchten die Verwendung des Simplexalgorithmus, um – bei gebebener Eingangsfunktion $C_a(t)$ – die Exponentialterme zu schätzen [3]. Dabei verwendeten die Autoren die Tatsache, dass die Summe $\sum_{i=1}^{n} (\mathbf{e}^T \alpha_i)\ e^{\lambda_i t} \otimes C_a(t)$ linear in $e^{\lambda_i t}$ ist. Wählt man sich künstlich (ohne den Hintergrund eines Modells, das *a priori* Wissen über Zahl und Größenordnung der λ_i impliziert) eine diskrete Menge möglicher Werte (letztlich eine Diskretisierung eines Intervalls in $\mathbb{R}$, in dem man Komponenten von λ finden möchte) so erhält man sofort ein endlichdimensionales, lineares Problem

$$S(t) = \sum_{i=1}^{n} (\mathbf{e}^T \alpha_i)\ (e^{\lambda_i t} \otimes C_a)(t)$$

in dem nur noch die Koeffizienten $\mathbf{e}^T \alpha_i$ geschätzt werden müssen. Dieses lineare Problem wird mit dem Simplexalgorithmus gelöst. Auf diese Weite erhält man eine Art Spektrum (λ_i, α_i). Dieses Spektrum beinhaltet alle Informationen, die die Situation in einem Voxel charakterisieren. Mehrere Untersuchungen scheinen die Leistungsfähigkeit der Methode zu belegen [21, 9]. Die Idee der Diskretisierung der Werte für λ ähnelt dem Basisfuntionsansatz [7].
Dies ist aber vielleicht momentan der interessanteste Ansatz. Lange Zeit hatte man wegen fehlender Rechnerleistung kaum Möglichkeiten, diese Ideen auf Voxelbasis zu verfolgen. Das ändert sich langsam.

4.5 Nicht-Invasive lineare Methoden

Bei den nicht-invasiven Methoden fehlt die Information über $C_a(t)$ völlig. Diese muss durch die Messung einer Referenz-Region ersetzt werden. Auch

hier gibt es eine Version der Logan und Gjedde-Patlak Analyse (siehe etwa [9, 25]). Wie oben werden wir die Größen, die sich auf die Referenz-Region beziehen mit einem Strich kennzeichnen und die Größen aus der interessierenden Region (ROI) unverändert belassen, d.h. k_1 ist die Rate k_1 der ROI während k_1' die entsprechende Rate der Referenz-Region bezeichnet.

4.5.1 Nicht-invasive Logan Analyse

Modell I/Referenz Region:
Aus Gleichung (4.1) folgt

$$\int_0^t C_a(\tau)\,d\tau = \frac{1}{k_1'}C_1'(t) + \frac{k_2'}{k_1'}\int_0^t C_1'(\tau)\,d\tau = \frac{1}{V_1'}\left(\frac{1}{k_2'}C_1'(t) + \int_0^t C_1'(\tau)\,d\tau\right) \quad (4.5)$$

Damit können wir aus dem Signal der Referenz-Region Informationen über den Tracer im Blut $C_a(t)$ bzw. dessen Integral erhalten. Die Schätzung der Parameter von Modell I steht hier nicht im Vordergrund; Modell I ist primär das Modell für die Referenz-Region, während die ROI durch Modell II beschrieben wird.

Modell II: (Variante 1)
Setzen wir (4.5) in (4.4) ein, so erhalten wir approximativ für groß Zeiten

$$\frac{\int_0^t \overline{C}(\tau)\,d\tau}{\overline{C}(t)} = \text{Const} + \frac{V_1+V_2}{V_1'}\,\frac{\frac{1}{k_2'}C_1'(t) + \int_0^t C_1'(\tau)\,d\tau}{\overline{C}(t)}$$
$$\text{für große Zeiten.} \quad (4.6)$$

Annahme: Wie vorher nehmen wir an, daß die Raten k_1, k_2 in der ROI und der Referenzregion annähernd gleich sind,

$$V_1 = V_1', \qquad k_2' = k_2.$$

Problem: Wir kennen k_2 nicht. Diese Konstante muss aus einer Population von Probanden bestimmt werden, in der Hoffnung, dass die Schwankungen der Werte dieser Konstante innerhalb der Population relativ klein ist.
Ergebnis: Wir können nicht V_2, sondern nur

$$\frac{V_2+V_3}{V_2}$$

schätzen. Diese Größe wird "Distribution Volumen Ration" (DVR) genannt, und analog wie V_2 zur Generierung von Bildern eingesetzt.

Modell II: (Variante 2)
Nun nehmen wir an, daß für große Zeiten

$$\frac{C_1'(t)}{\overline{C}(t)} \approx \text{Const.}$$

Diese Annahme muss bei jeder Messung geprüft werden. Da C_1 und C_1' aus verschiedenen Gewebebereichen stammen, wird man unter generischen Bedingungen erwarten, dass diese Funktionen asymptotisch exponentielle Funktionen mit unterschiedlichen Exponenten approximieren, sodass der Quotient $C_1'(t)/C_1(T)$ exponentiell schnell wächst oder fällt. Falls aber die Konstanz dieses Quotienten eintritt, benötigt man k_2 nicht sondern kann in Gleichung (4.6) den Term $C_1'(t)/(k_2\,\overline{C}(t))$ dem Achsenabschnitt der Geradengleichung zuschlagen,

$$\frac{\int_0^t \overline{C}(\tau)\,d\tau}{\overline{C}(t)} \approx \text{Const} + \frac{V_1+V_2}{V_1}\,\frac{\int_0^t C_1'(\tau)\,d\tau}{\overline{C}(t)} \qquad \text{für große Zeiten.}$$

Modell II: (Variante 3 / Ichise Analyse)
In [13, 19] wird eine multivariate Variante der Logan Analyse vorgeschlagen; damit ist es möglich, die Bestimmung der Konstante k_2 aus der Population zu umgehen. Das Vorgehen ist das Gleiche wie beim Logan Analyse, nur fasst man die Terme der rechten Seite von (4.6) als unabhängige Variablen auf,

$$\frac{\int_0^t \overline{C}(\tau)\,d\tau}{\overline{C}(t)} = \text{Const} + \frac{V_1+V_2}{k_2V_1}\left(\frac{C_1'(t)}{\overline{C}(t)}\right) + \frac{V_1+V_2}{V_1}\left(\frac{\int_0^t C_1'(\tau)\,d\tau}{\overline{C}(t)}\right)$$
$$\text{für große Zeiten.} \qquad (4.7)$$

Mittels bivariater, linearer Regression kann nun $(V_1 + V_2)/V_1$ und $(V_1 + V_2)/(k_2V_1)$ bestimmt werden.

4.5.2 Nicht-invasive Gjedde-Patlak Analyse

Es werden Zeitskalen-Argumente benutzt, um die Dynamik des Tracers im Kompartment C_1 eng an die im Blut zu binden. Wir bezeichnen wiederum mit einem Strich die Größen der Referenz-Region (in der $k_3 = 0$ gilt).

Dann folgt, falls $k_2 \gg 1$, dass

$$C_1'(t) = \frac{k_1'}{k_2'}C_a(t).$$

Damit können wir C_a durch C_1' ersetzten, und erhalten

$$\begin{aligned} S(t) &= C_1(t) + C_2(t) + \epsilon C_a(t) \\ &= \frac{k_1 k_3}{k_2 + k_3} \int_0^t C_a(\tau)\, d\tau + \left(\epsilon + \frac{k_1}{k_2 + k_3}\right) C_a(t) \\ &= \frac{k_1 k_3 k_2'}{(k_2 + k_3) k_1'} \int_0^t C_1(\tau)\, d\tau + \left(\epsilon + \frac{k_1}{k_2 + k_3}\right) \frac{k_1'}{k_2'} C_1'(t). \end{aligned}$$

Das ergibt wieder eine lineare Beziehung, diesmal zwischen $S(t)/C_1'(t)$ und $\int_0^t C_1'(\tau)\, d\tau / C_1'(t)$. Die Steigung beträgt $k_1 k_2' k_3/((k_2 + k_3)k_1')$. Dies ist zwar nicht der gewünschte Wert, ist aber dennoch geeignet, Gewebe verschiedener Eigenschaft bzgl. des Wertes k_3 zu unterscheiden.

4.5.3 Gleichgewichtsmethode I: Lokale Maxima der Daten

Diese wie die folgende Methode [19, 28] ist ausschließlich für Modell II interessant.

Annahmen: Bei dieser Methode gehen zwei Voraussetzungen ein.
(1) Es ist $C_1(t) = C_1'(t)$, das Kompartment C_1 ist das Gleiche für Referenz-Region und ROI.
(2) $C_1(t)$ und $C_2(t)$ besitzen zur gleichen Zeit ein lokales Maximum.

Beide Annahmen sind sehr stark; exakt können sie nicht erfüllt werden. Für gewisse Tracer scheinen sie allerdings approximativ gegeben zu sein [4]. Wenn wir die Annahmen akzeptieren, so können wir $C_2(t)$ als Differenz von $\overline{C}(t)$ und $C_1'(t)$ bestimmen,

$$C_2(t) = \overline{C}(t) - C_1(t).$$

Durch anpassen einer glatten Kurve an die Datenpunkte von $\overline{C}(t) - C_1(t)$ erhalten wir Informationen über Zeitpunkt und Größe der lokalen Maxima. In diesen Maxima verschwinden die zeitlichen Ableitungen, und wir haben

$$BP = \frac{k_3}{k_4} = \left.\frac{C_2}{C_1}\right|_{\text{Gleichgewicht}}.$$

In diesem Fall nehmen wir $BP = k_3/k_4$ als charakteristische Größe.

4.5.4 Gleichgewichtsmethode II: Transientes Plateau der Daten

Diese Methode ist ausschließlich für Modell II interessant. Sie ist eng verwandt mit dem zuletzt betrachteten Ansatz (lokales Maximum). Sie beruht auf den Annahmen:

Annahmen:
(1) $V_1 = V_1'$
(2) $(C_1 + C_2)/C_1'$ erreicht ein Plateau und ist längere Zeit konstant.
(3) $C_1 \approx C_1'$ in dem Plateau.

Obwohl wiederum diese Annahmen schwerlich exakt zu rechtfertigen sind, zeigt es sich, dass diese Methode bei einigen Tracern Ergebnisse liefert, die mit den Ergebnissen anderer Ansätze stark korrelieren.

Falls im transienten Gleichgewicht nicht nur $(C_1 + C_2)/C_1'$, sondern auch C_1 und C_2 konstant werden, so findet sich

$$BP = \left.\frac{C_2/C_a}{C_1/C_a}\right|_{\text{Gleichgewicht}} = \left. -1 + \frac{C_1 + C_2}{C_1}\right|_{\text{Gleichgewicht}} = \left. -1 + \frac{C_1 + C_2}{C_1'}\right|_{\text{Gleichgewicht}}$$

Tatsächlich lässt sich aus dem Verschwinden der zeitlichen Ableitung von $(C_1 + C_2)/C_1'$ die Konstanz von C_1 und C_2 mitnichten folgern; die wirkliche Rechtfertigung dieser Methode ist letztlich rein phänomenologisch [2].

4.6 Nicht-Invasive Nicht-lineare Methoden

Die alleine Struktur der Lösung ist natürlich für den invasiven- und den nicht-invasiven Fall die Gleiche. Da wir aber die Inhomogenität, d.h. die Dichte des Tracers im Blut $C_a(t)$, nicht explizit gegeben haben, müssen wir diese Information erst aus den Modellgleichungen der Referenzregion herausschälen.

4.6.1 Vereinfachung des Standard-Modells

Um die Informationen aus der Referenz-Region zu nutzen und auch, um nicht-lineare Methoden schnell anwenden zu können, wurde von Lammertsma und Hume eine Vereinfachung des Standard-Modells entwickelt [14]. Die Dynamik der ROI folgt wieder dem Standard-Modell (Modell II), die Messung einer Referenzregion genügt Modell I.

Vereinfachung / Schritt 1: Reduktion auf vier Parameter
In Modell I und Modell II haben wir zunächst insgesamt sechs Parameter: k_1, k_2, k_3, k_4, k_1' und k_2'. Wenn wir annehmen, dass $V_1 = V_1'$, so können wir einen Parameter eliminieren.

Annahme 1: $V_1 = V_1'$.

Definiere

$$R := k_1/k_1' \qquad \Rightarrow \qquad k_2' = k_2/R.$$

Löst man die Differentialgleichung für $C_1'(t)$ algebraisch nach $C_a(t)$ auf, und setzt das Ergebnis in die Gleichungen des Modells II ein, so erhält man

$$\begin{aligned}\dot{C}_1(t) &= -(k_2+k_3)C_1(t) + k_4 C_2(t) + R\left(\dot{C}_1'(t) - k_2 C_1'(t)/R\right)\\ \dot{C}_2(t) &= -k_4 C_2(t) + k_3 C_1(t)\end{aligned}$$

Hier hat man nur noch vier Parameter.

Vereinfachung / Schritt 2: Reduktion auf drei Parameter
Nun verwenden Lammertsma und Hume ein Zeitskalen-Argument. Die Autoren setzen voraus, dass die Zeitskala der Transitionen von C_1 nach C_2 und wieder zurück ist wesentlich schneller als die von C_a nach C_a und zurück. Dann kann man in der ROI den zeitlichen Verlauf von $\overline{C}(t) = C_1(t) + C_2(t)$ gut durch eine einzige Differentialgleichung beschreiben.

Annahme 2: $k_3, k_4 \gg k_1, k_2$.

Diese Annahme ist für einige Tracer akzeptabel. In dem Fall können wir das System reskalieren und schreiben

$$\begin{aligned}\dot{C}_1(t) &= -(k_2+k_3)C_1(t) + k_4 C_2(t) + R\left(\dot{C}_1'(t) - k_2 C_1'(t)/R\right)\\ \delta\dot{C}_2(t) &= -k_4 C_2(t) + k_3 C_1(t)\end{aligned}$$

wobei $0 < \delta \ll 1$. Also, $C_2 \approx (k_3/k_4)C_1$ und $\overline{C} \approx (1 + k_3/k_4)C_1$. Daher folgt für $\delta \to 0$, dass

$$\dot{\overline{C}}(t) = -k_2 C_1(t) + k_1 C_a(t) = -\tilde{k}_2 \overline{C}(t) + k_1 C_a(t).$$

mit $\tilde{k}_2 = k_2 k_4/(k_3 + k_4)$, d.h.

$$\tilde{k}_2 = k_2 \left(1 + \frac{k_3}{k_4}\right)^{-1} = k_2\,(1 + BP)^{-1}.$$

Drücken wir $C_a(t)$ wie oben durch $C_1'(t)$, $\dot{C}_1'(t)$ aus, so erhalten wir

$$\begin{aligned}\dot{\overline{C}}(t) &= -\tilde{k}_2\overline{C}(t) + R\left(\dot{C}_1'(t) - k_2 C_1'(t)/R\right)\\ \Rightarrow \quad \frac{d}{dt}\left[\overline{C}(t) - R\,C_1'(t)\right] &= -\tilde{k}_2\left[\overline{C}(t) - R\,C_1'(t)\right] - (k_2 - R\tilde{k}_2)C_1'(t)\end{aligned}$$

mit der Lösung

$$\begin{aligned}\overline{C}(t) &= R\,C_1'(t) + (k_2 - R\tilde{k}_2)\int_0^t e^{\tilde{k}_2(t-\tau)} C_1'(\tau)\,d\tau\\ &= R\,C_1'(t) + \left(k_2 - \frac{R\,k_2}{1-BP}\right) e^{\frac{k_2}{1-BP}t} \otimes C_1'(t) \qquad (4.8)\end{aligned}$$

Durch das Ausnutzen der Zeitskalen konnten wir einen weiteren Parameter wegskalieren (es kommt nur noch auf k_3/k_4 an, nicht mehr auf k_3 und k_4), sodass am Ende drei Parameter zu bestimmen bleiben.

4.6.2 Schätzung der Parameter für das Vereinfachte Standard-Modell

In [7] wird ein Hybrid-Algorithmus aus linearem Fit und nicht-linearer kleinste-Fehler-Quadrat Methode vorgestellt. Tatsächlich liegt dieser Ansatz sehr in der Nähe der Spektralanalyse (siehe Paragraph 4.4.3); im Wesentlichen wird hier die Zahl der Exponentialterme festgelegt, was bei der Spektralanalyse nicht der Fall ist. Gleichung (4.8) können wir schreiben als

$$\overline{C}(t) = \theta_1 C_1'(t) + \theta_2 \int_0^t e^{\theta_3 t} \otimes C_1'(t)\, d\tau$$

mit

$$\theta_1 = R, \quad \theta_2 = k_2 - \frac{R k_2}{1 - BP}, \qquad \theta_3 = \frac{k_2}{1 - BP}.$$

Wir finden

$$R = \theta_1, \quad k_2 = \theta_2 + R\theta_3, \qquad BP = \frac{k_2}{\theta_3} - 1,$$

d.h. können aus θ_1, θ_2 und θ_3 die Größe BP schätzen.

Wenn wir θ_3 kennen, so ist die Schätzung von θ_1 und θ_2 aus den Daten ein lineares Problem, welches leicht und effizient gelöst werden kann (Gauss-Schätzer mit Berücksichtigung der angemessenen Varianzstruktur). Um einen effizienten Algorithmus für die Schätzung von θ_3 zu entwickeln, wählen wir die Werte von θ_3 nicht aus $\mathbb{R}$, sondern erlauben nur diskrete Werte $\theta_{3,i}$, $i = 1, .., m$. Diese diskreten Werte müssen natürlich den relevanten Parameterbereich relativ dicht überdecken - hier kommt *a-priori*-Wissen zum Tragen. Im Beispiel wählen die Autoren $m = 100$ diskrete Werte zwischen 0.001/sec und 0.01/sec.
Um den optimalen Wert zu bestimmen, wird die lineare Regression für alle $\theta_3 = \theta_{3,i}$, $i = 1, .., m$, durchgeführt. Danach wird das Optimum, i.e. derjenige Parametersatz, der das kleinste Residuum besitzt, gewählt. Dieses Verfahren wird nur dadurch nicht-linear, dass wir wissen, dass es einen eindeutigen Wert für θ_3 gibt, der mit geschätzt wird. Es wird hier nicht akzeptiert, dass die Daten eine lineare Kombination mehrerer Basisfuntionen für verschiedene θ_3 sein könnte. Das lineare Verfahren von Cunningham und Jones [3] ist die direkte Verallgemeinerung, bei der simultan eine Linearkombination von

Lösungen mit verschiedenen Werten für θ_3 zugelassen werden, sodass eine Art exponentiellem Spektrums entsteht.

Bemerkung: Interessanterweise wurde dieses Verfahren vor der Logan Analyse entwickelt, erlangte aber längst nicht die Bedeutung des Logan Analyse. Der Hauptgrund dafür ist mangelnde Rechnerleistung zu dem Zeitpunkt, zu dem sich diese Verfahren entwickelten: Die oben beschriebende Prozedur muss für jedes Voxel durchgeführt werden; das ist ein sehr rechenintensives Vorgehen. Man kann allerdings erwarten, dass mit zunehmender Leistungsfähigkeit der Rechner diese Verfahren an Bedeutung gewinnen.

Kapitel 5

Probleme

Bisher haben wir die Idealsituation betrachtet. Es flossen allerdings Annahmen ein, die nicht gerechtfertigt sind. Einige diese Annahmen bzw. deren Abschwächung soll in diesem Kapitel diskutiert werden. Speziell streifen wir kurz die Bereiche: Effekt von Transportphänomenen im Blut, Transportphänomene der Positronen und Inhomogenität des Gewebes.

5.1 Transportphänomene - Zeitliche Verzögerung

Die Dichte des Tracers im entnommenen Blut ist im Allgemeinen nicht die Gleiche wie die in der Referenz-Region oder ROI. Zwischen der Ader, an der das Blut entnommen wird, und den entsprechenden Regionen liegen Transport- und Diffusionsprozesse. Man muss daher mit einer Verzögerung zwischen der gemessenen Dichte des Tracers im entnommenen Blut $C_m(t)$ und der Dichte in der betrachteten Region $C_a(t)$ rechnen. Dies kann man explizit durch

$$C_a(t) = C_m(t + \Delta t) + \tau \frac{d}{dt} C_m(t + \Delta t).$$

modellieren [27]. Dabei steht die Verzögerung um Δt für den Transport, und der Tiefpaßfilter für die Diffusion. Prinzipiell ist es mit diesem Ansatz möglich, die Effekte der Transport- und Diffusionsprozesse auf die Analysemethoden zu untersuchen.

5.2 Transportphänomene - Positronen

Das Signal entsteht dort, wo Positronen zerfallen. Dies kann aber einige Millimeter neben der Stelle sein, wo diese Positronen entstanden. Wenn zu einer festen Zeit die Tracerkonzentration durch $h(x)$ gegeben ist, so kann das gemessene Signal $g(x)$ zu dieser Zeit durch einen Integraloperator beschrieben werden,

$$g(x) = \kappa \int \sigma(x,y)h(y)\,dy$$

Dabei gibt κ die Normierung des Signals wieder, und $\sigma(x,y) = \sigma(x-y)$ ist durch den Transport der Positronen gegeben. D.h. $\sigma(z)$ ist die Wahrscheinlichkeit, ein Positron, welches am Ort $x = 0$ erzeugt wird, am Ort $x = z$ zu detektieren. Der Kern $T(z)$ glättet also die ursprüngliche Tracer-Dichte. Man nennt diesen Effekt "Partial Volume Effect". Rousset, Yiong und Ma [26] beschreiben einen Weg, diesem Phänomen entgegenzuwirken. Da der PET-Scanner die räumliche Struktur diskretisiert, betrachten die Autoren eine diskrete Variante dieser Glättung,

$$G = \Sigma H,$$

wobei H der Vektor der tatsächlichen Tracerdichte innerhalb der Voxel bedeutet, G der Vektor des gemessenen Signals in den Voxeln, und Σ eine Matrix, die die Glättung der Tracerdichte beschreibt. Letztlich muss man Σ invertieren. Da diese Matrix glättet, ist dies (obwohl die Inverse existiert) prinzipiell ein schlecht gestelltes Problem. In der Praxis aber ist dieser Effekt nicht sehr stark. Daher sind die Singulärwerte der Matrix Σ noch so weit von der Null entfernt, dass eine naive Invertierung die Fehler im nicht sehr verstärkt. Numerische und reale Experimente zeigen, dass dieser naive Ansatz zu einer Verbesserung des PET-Bildes führt.

5.3 Inhomogenes Gewebe - Probleme der Identifikation

Bisher nahmen wir an, dass im betrachteten Voxel nur ein Gewebetyp und eventuell ein kleiner, prinzipiell vernachlässigbarer Anteil Kapillargefäße vorhanden sind. Da ein Voxel aber relativ groß sein muss (erzwungen durch die mittlere freie Weglänge eines Positrons), wird diese Annahme häufig nicht zu halten sein. Der Artikel [27] beschreibt ein Modell für inhomogenen Gewebe, und zwei Methoden der Paramterschätzung. Dabei wird für einen Gewebetyp das Modell I zugrundegelegt.

5.3.1 Modell

Wir nehmen an, dass in einem Voxel zwei Gewebetypen existieren; jedes dieser Gewebe wird durch Modell I beschrieben. Das Modell muss also den Tracer im Blut $C_a(t)$ beschreiben, und den Tracer im Gewebetyp α bzw. Gewebetyp β. Seien nun

$C^\alpha(t)$, $C^\beta(t)$	Dichte des Tracers im Gewebe α bzw. β.
$C_a(t)$	Dichte des Tracers im Blut
k_1^i	Übergangsrate vom Blut nach Gewebe, $i \in \{\alpha,\ \beta\}$
k_1^i	Übergangsrate vom Gewebe in das Blut, $i \in \{\alpha,\ \beta\}$
w^i	relativer Anteil der Gewebetypen $i \in \{\alpha,\ \beta\}$, wobei $w^\alpha + w^\beta = 1$ und $w^i \geq 0$.

Wir erhalten die Differentialgleichungen

$$\begin{aligned} \dot{C}^i(t) &= -k_2^i C^i(t) + k_1^i C_a(t) \\ C^i(0) &= 0 \end{aligned} \tag{5.1}$$

Das gemessene Signal nun ist

$$S(t) = w^\alpha C w^\alpha(t) + w^\beta C w^\beta(t) + \epsilon\, C_a(t).$$

Wenn wir (wie üblich) $\epsilon\, C_a(t)$ vernachlässigen, so haben wir

$$\overline{C}(t) = w^\alpha C^\alpha(t) + w^\beta C^\beta(t).$$

5.3.2 Ansatz 1: Lineares Modell

Wir folgen dem Paragraph 4.3.3, und leiten eine Differentialgleichung höherer Ordnung für $\overline{C}(t)$ her. Durch wiederholtes Differenzieren finden wir

$$\begin{aligned} \overline{C}(t) &= w^\alpha C^\alpha(t) + w^\beta C^\beta(t) \\ \dot{\overline{C}}(t) &= -w^\alpha k_1^\alpha C^\alpha - w^\beta k_1^\beta C^\beta + (w^\alpha C^\alpha + w^\beta C^\beta) C_a \\ \ddot{\overline{C}}(t) &= -w^\alpha (k_1^\alpha)^2 C^\alpha - w^\beta (k_1^\beta)^2 C^\beta + (w^\alpha C^\alpha + w^\beta C^\beta)\dot{C}_a + (w^\alpha C^\alpha + w^\beta C^\beta) C_a \end{aligned}$$

Aus diesen drei Gleichungen kann man C^α, C^β eliminieren. Man findet

$$\ddot{\overline{C}}(t) + (k_1^\alpha + k_1^\beta)\dot{\overline{C}}(t) + k_1^\alpha k_1^\beta \overline{C}(t) = (w^\alpha k_2^\alpha + w^\beta k_2^\beta)\dot{C}_a(t) + (k_1^\alpha k_1^\beta)\left\{ w_\alpha \frac{k_2^\alpha}{k_1^\alpha} + w_\beta \frac{k_2^\beta}{k_1^\beta} \right\} C_a$$

mit den Anfangsbedingungen

$$\overline{C}(t)|_{t=0} = \dot{\overline{C}}(t)|_{t=0}$$

Aus biologischen Gründen darf man

$$C_a(t)|_{t=0} = \dot{C}_a(t)|_{t=0} = 0$$

ebenfalls voraussetzen. Setzten wir

$$\begin{aligned} \theta_a &= k_1^\alpha + k_1^\beta, \qquad \theta_b = k_1^\alpha k_1^\beta, \\ \theta_c &= w^\alpha k_2^\alpha + w^\beta k_2^\beta, \qquad \theta_d = (k_1^\alpha k_1^\beta)\left\{ w_\alpha \frac{k_2^\alpha}{k_1^\alpha} + w_\beta \frac{k_2^\beta}{k_1^\beta} \right\} \end{aligned}$$

so lautet die Gleichung

$$\begin{aligned} \ddot{\overline{C}}(t) + \theta_a \dot{\overline{C}} + \theta_b \overline{C} &= \theta_c \dot{C}_a + \theta_d C_a \\ \overline{C}(0) &= 0, \quad \dot{\overline{C}}(0) = 0, \quad C_a(0) = 0, \quad \dot{C}_a(0) = 0 \end{aligned}$$

Wir sehen dass wir also nur vier Parameter bestimmen können $(\theta_a, \theta_b, \theta_c, \theta_d)$, während das Modell fünf Parameter $(w^\alpha = 1 - w^\beta, k_1^\alpha, k_1^\beta, k_2^\alpha, k_2^\beta)$ besitzt. Die Parameter des Systems lassen sich nicht identifizieren. Heuristisch ist es leicht zu sehen, dass man vier Parameter bestimmen kann: asymptotisch wird sich das Signal aus zwei Exponential-Termen aufbauen. Diese sind aber durch Gewichte und Exponenten, d.h. durch vier Größen völlig festgelegt.

Die fünfte Information muss von außen kommen. In [27] wird angenommen, dass $V_1^\alpha = k_1^\alpha / k_2^\alpha$ bzw. $V_1^\beta = k_1^\beta / k_2^\beta$ bekannt sind. Mit diesem Vorwissen können die Parameter identifiziert werden.

Zweimaliges Integrieren liefert

$$\begin{aligned} \overline{C}(t) &= -\theta_a \int_0^t \overline{C}(\tau)\, d\tau - \theta_b \int_0^t \int_0^\tau \overline{C}(\tau')\, d\tau' d\tau \\ &+ \theta_c \int_0^t C_a(\tau)\, d\tau + \theta_d \int_0^t \int_0^\tau C_a(\tau')\, d\tau'\, d\tau \end{aligned} \tag{5.2}$$

Diese Gleichung kann man als lineares Modell auffassen. Das (über die Zeit integrierte) Signal bildet die Designmatrix X, die linke Seite der Gleichung den Datenvektor Y, und der Vektor $\theta = (\theta_a, \theta_b, \theta_c, \theta_d)^T$ beinhaltet die zu schätzenden Parameter,

$$Y = X\theta.$$

Wie üblich, kann man den naiven Gauss-Schätzer nutzen,

$$\hat{\theta} = (X^T X)^{-1} X^T Y.$$

Dabei sollte man eigentlich noch die Kovarianzstruktur des Datenvektors mit einbringen. Weiter geht die Messung nicht nur in den Datenvektor Y, sonder auch in die Designmatrix ein. Es ist nicht klar, inwieweit der Gauss-Schätzer in dieser Situation angemessene Schätzungen liefert.

5.3.3 Ansatz 2: Direkte Lösung der dynamischen Gleichungen

Die Differentialgleichungen können natürlich auch direkt gelöst werden. Man erhält

$$C^i(t) = k_1^i \int_0^t e^{-k_2^i(t-\tau)} C_a(\tau)\, d\tau \qquad \text{für } i \in \{\alpha,\ \beta\}.$$

Das gemessene Signal (bei Vernachlässigung des Signal aus dem Blut) ist nun

$$\overline{C}(t) = w^\alpha k_1^\alpha \int_0^t e^{-k_2^\alpha(t-\tau)} C_a(\tau)\, d\tau + w^\beta k_1^\beta \int_0^t e^{-k_2^\beta(t-\tau)} C_a(\tau)\, d\tau.$$

Wir sehen wieder, dass wir nur vier Parameterkombinationen identifizieren können, nämlich k_2^α, k_2^β, $w^\alpha k_1^\alpha$ und $w^\beta k_1^\beta$. Diese Parameter kann man z.B. mit der Kleinsten-Quadrate-Methode schätzen.

Nehmen wir (wie im letzten Abschnitt) an, dass V_1^α bzw. V_1^β bekannt sind, so erhalten wir genug Information um alle Parameter bestimmen zu können.

5.4 Fehlermodell

Ein regelrechtes statistisches Fehlermodell wird nur selten aufgestellt. So arbeiten Maeght et al. [18] mit Fehlern, die vor allem von dem Zählprozess stammen (Poisson-Verteilung). Gunn et al. [7] approximiert die Poisson-Verteilung durch eine Normalverteilung mit entsprechender Varianz. Auch Aston et al. [1] betrachten ein normalverteiltes Fehlermodell, wobei die Autoren vor allem auf die räumlichen Korrelationen zielen. Bei invasiven Verfahren kommt eine zusätzliche Schwierigkeit von den unterschiedlichen Messmethoden im Gewebe (PET) und in den Blutproben. So wird i.a. das Signal $S(t)$ und das Signal $C_a(t)$ eine unterschiedliche Varianzstruktur aufweisen.

Kapitel 6

Diskussion

Wir beschrieben die beiden am häufigsten genutzten Modelle der Modellierung von Tracer-Dynamik, und einige der verbreiteten Methoden die durch PET gewonnene Daten auszuwerten. Wir schlossen viele Problemkreise wie die Rekonstruktion der Tracerintensität, Reduktion von Rauschen der Daten etc. völlig aus. Es ging hier primär um die dynamischen Daten eines Voxels, und die Methoden diese Daten zu analysieren.

Es ist zu beobachten, dass fast alle Methoden eng an Modelle gebunden sind. Nur die Spektralanalyse ist weitgehend frei von Modellannahmen. Die am weitest verbreiteten Methoden sind sicher die sogenannten "graphischen Methoden", i.e. die Logan- und Gjedde-Patlak Analyse. Der Trend aber geht deutlich zu den nichtlinearen Schätzmethoden, die in irgendeiner Weise die Lösung der Differentialgleichung an die Daten anpasst.

Nicht allzugroße Beachtung findet die Spektralanalyse. Sie ist mathematisch am wenigsten gut untersucht. Die Leistungsfähigkeit der existierenden Algorithmen lassen erwarten, dass dort Verbesserungen möglich sind, insbesondere wenn angemessene Fehlermodelle mit einbezogen werden. Auch von der Anwendung her ist die Robustheit der Spektralanalyse gegen die Modellannahmen eine sehr interessante Eigenschaft. Z.B. das Problem der fehlenden Identifizierbarkeit der Dynamik, das schon in sehr einfachen Situationen zum Tragen kommt (ein-Kompartment Modelle für inhomogenes Gewebe) spielt auf der Ebene keine Rolle mehr.

Literaturverzeichnis

[1] Aston J.A.D, Cunningham V.J., Asselin, M.-C. et al. Positron Emission Tomography Partial Volume Correction: Estimation and Algorithms. *J. Cereb. Blood Flow Metab.* 22(2002) 1019–1034.

[2] Carson R.E., Channing M.A., Blasberg R.G. et al. Comparison of bolus and infusion methods for receptor quantitation: application to [18F]cylofoxy and positron emission tomography. *J. Cereb. Blood Flow Metab.* 13(1993) 24–42.

[3] Cunningham, V.J. and Jones, T. Spectral Analysis of Dynamic PET Studies. *J. Cereb. Blood Flow Metab.* 13(1993) 15–23.

[4] Farde L., Eriksson L., Blomquist G. and Halldin C. Kinetic analysis of central [11C]raclopride binding D2-dopamine receptors studied by PET - a comparison to the equilibrium analysis. *J. Cereb. Blood Flow Metab.* 9(1989) 696–708.

[5] Frost J.J., Douglass K.H., Mayberg H.S. et al. Multikompartmental Analysis of [11C]-Carfentanil Binding to Opiate Receptors in Humans Measured by Positron Emission Tomography. *J. Cereb. Blood Flow Metab.* 9(1989) 398–409.

[6] Gjedde A., Wienhard, K. et al. Comparative Regional Analysis of 2-Flourodeoxyglucose and Methylglucose Uptake in Brain of Four Stroke Patients. With Special Reference to the Regional Estimation of the Lumped Constant. *J. Cereb. Blood Flow Metab.* 5(1985) 163–178.

[7] Gunn R.N., Lammertsma A.A., Hume S.P. et al. Parametric Imaging of Ligand-Receptor Binding in PET Using a Simplified Reference Region Model. *Neuroimage* 6(1997) 279–287.

[8] Gunn R.N., Gunn S.R., Cunningham, V.J. Positron Emission Tomography Compartmental Models. *J. Cereb. Blood Flow Metab.* 21(2001) 635–652.

[9] Gunn R.N., Gunn S.R., Turkheimer F.E. et al. Positron Emission Tomography Compartmental Models: A Basis Pursuit Strategy for Kinetic Models. *J. Cereb. Blood Flow Metab.* 22(2002) 1425–1439.

[10] Hawkins R., Huang S.-C., Barrio J.R. et al. Estimation of Local Cerebral Protein Synthesis Rates with L-[1-11C]Leukine and PET: Methods, Model and Results in Animals and Humans. *J. Cereb. Blood Flow Metab.* 9(1989) 446–460.

[11] Holmström K. and Petersson J. A review of the parameter estimation problem of fitting positive exponential sums to empirical data. *Appl. Math. Comp.* 126(2002) 31–61.

[12] Howard B.E., Ginsberg M.D., Hassel W.R. at al. On the Uniqueness of Cerebral Blood Flow Measured by the *In Vivo* Autoradiographic Strategy and Positron Emission Tomography. *J. Cereb. Blood Flow Metab.* 3(1983) 432–441.

[13] Ichise M., Ballinger J.R., Golan H. et al. Noninvasive quantification of dopamine D2 receptors with iodine-123-IBF SPECT. *J. Nucl. Med.* 37(1996) 513–520.

[14] Lammertsma A.A and Hume S.P. Simplified Tissue Model for PET Receptor Studies. *Neuroimage* 4(1996) 153–158.

[15] Logan J. Graphical Analysis of PET Data Applied to Reversible and Irreversible Tracers. *Nucl. Med. Biol.* 27(2000) 661–670.

[16] Logan J., Fowler J.S. et al. Strategy of the Formation of Parametric Images Under Conditions of Low Injected Radioactivity Applied to PET Studies With Irreversible Monoamine Oxidase A Tracers [^{11}C]Clorgyline and Deuterium-Substituted [^{11}C]Clorgyline. *J. Cereb. Blood Flow Metab.* 22(2002) 1367–1376.

[17] Louis, A.K. *Inverse und schlecht gestellte Probleme.* Teubner, 1989.

[18] Maeght, J., Noll, D., Celer, A. and Farncombe, T. Methods for Dynamic SPECT Tomography. *Publications du laboratoire MIP, Université Paul Sabatier, Toulouse,* Preprint 26(1999).

[19] Meyer J.H. and Ichise M. Modeling of Receptor Ligand Data in PET and SPECT Imaging: A Review of Major Approaches. *J. Neuroimag.* 11(2001) 30–39.

[20] Murase K., Tanada Sh., Fujita H. et al. Kinetic Behavior of Technetium-99m-HMPAO in the Human Brain and Quantification of Cerebral Blood Flow Using Dynamic SPECT. *J. Nucl. Med.* 33(1992) 135–143.

[21] Murase K., Inuoue T., Fujiokra H. et al. An alternative approach to estimation of the brain perfusion index for measurement of cerebral blood flow using technetium-99m compounds. *Eur. J. Nucl. Med.* 16(1999) 1333-1339.

[22] Osborne M.R. and Smyth G.K. A Modified Prony Algorithm for Exponential Function Fitting. *SIAM J. Sci. Comp.* 16(1995) 119–138.

[23] Palmer, M.R. and Brownell, G. Annihilation Density Distribution Calculations for Medically Important Positron Emitters. *IEEE Trans. Med. Imag.* 11(1992) 373–378.

[24] Patlack C.S., Blasberg, R.G. et al. Graphical Evaluation of Blood-to-Brain Transfer Constants from Multiple-Time Uptake Data. *J. Cereb. Blood Flow Metab.* 3(1983) 1–7.

[25] Patlack C.S. and Blasberg, R.G. Graphical Evaluation of Blood-to-Brain Transfer Constants from Multiple-Time Uptake Data. Generalizations. *J. Cereb. Blood Flow Metab.* 5(1985) 584–590.

[26] Rousset, O.G., Ma Y and Evans A.C. Correction for Partial Volume Effect in PET: Principle and Validation. *J. Nucl. Med.* 39(1998) 904–911.

[27] Schmidt K. and Sokoloff L. A Computational Efficient Algorithm for Determining Regional Cerebral Blood Flow in Heterogeneous Tissues by Positron Emission Tomography. *IEEE Trans. Med. Imag.* 20(2001) 618–632.

[28] Slifstein M.and Laruelle M. Models and methods for derivation of *in vivo* neuroreceptor parameter with PET and SPECT reversible radiotracers. *Nucler Med. Biol.* 28(2001) 595–608.

[29] Wienhard, K. Measurement of glucose consumption using [^{18}F]lourodeoxyglucose. Methods 27(2002), 218–255.

Physikalische und mathematische Grundlagen der NMR-Spektroskopie

Frank Filbir*

Zusammenfassung

Wir stellen in knapper Form Grundlagen der NMR-Spektroskopie bereit. Zunächst wird die Spindynamik aus den Gesetzen der Quantenmechanik und der Elektrodynamik abgeleitet und die Gesamtmagnetisierung berechnet. Anschließend wird die Radio-Frequenz Impulsanregung erklärt. Im nächsten Abschnitt werden die Bloch-Gleichungen diskutiert. Schließlich wird im letzten Abschnitt auf die Signalzerlegung eingegangen.

*Institute of Biomathematics and Biometry
GSF - National Research Center for Environment and Health,
Postfach 1129, D-85758 Oberschleißheim, Germany,
filbir@gsf.de, http://ibb.gsf.de

1 Spindynamik

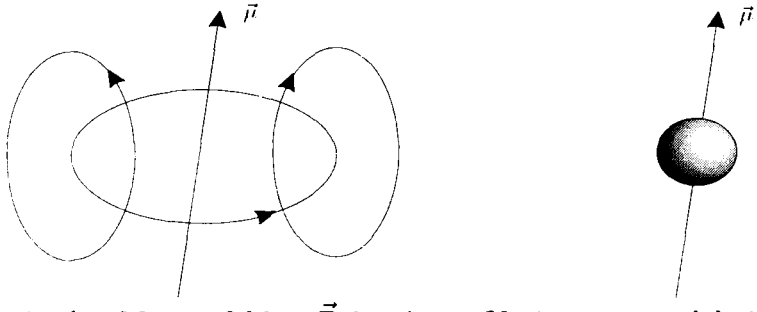
Elementarteilchen wie Protonen und Elektronen besitzen ein Drehmoment (Spin) $\vec{J}$. Der Spin eines Teilchens ist eine quantenmechanische Größe, d.h. $|\vec{J}|$ kann nicht beliebige Werte annehmen. Es gilt vielmehr

$$|\vec{J}| = \hbar \sqrt{I(I+1)}$$

mit $I = \frac{k}{2}$, $k \in \mathbb{N}_0$. Die Größe I heißt Spinquantenzahl, $\hbar$ ist die Plancksche Konstante. Es gilt

$$\hbar = \frac{h}{2\pi} \ , \ h = 6.6{\cdot}10^{-34} Js$$

Als bewegte elektrische Ladung erzeugen Protonen und Elektronen nach den Gesetzen der Elektrodynamik ein magnetisches Feld. Dies kann in Analogie zu einem Kreisstrom gesehen werden.



Analogie des Magnetfeldes $\vec{B}$ in einem Kreisstrom und bei einer bewegten elektrischen Ladung. $\vec{\mu}$ ist das magnetische Moment.

Der Zusammenhang zwischen dem Spin und dem magnetischen Moment ist durch

$$\vec{\mu} = \gamma \vec{J} \tag{1.1}$$

gegeben, wobei $\gamma \in \mathbb{R}$ das gyromagnetische Verhältnis ist. Das gyromagnetische Verhältnis ist abhängig vom Nukleid. Beispielsweise ist für ^1H, also Protonen, $\gamma = 42.48 \frac{MHz}{T}$.
Aus (1.1) ergibt sich

$$|\vec{\mu}| = \gamma \hbar \sqrt{I(I+1)}$$

Die Quantenmechanik liefert folgende Auswahlregeln für die Spinquantenzahl: Es ist

- $I = \frac{2n+1}{2}$, $n \in \mathbb{N}_0$, für Nukleide mit ungerader Atommasse (z.B. ^{13}C,^1H,^{31}P),
- $I = 0$, für Nukleide mit gerader Atommasse und gerader Ladungszahl,

- $I = n$, $n \in \mathbb{N}$, für Nukleide mit gerader Atommasse und ungerader Ladungszahl.

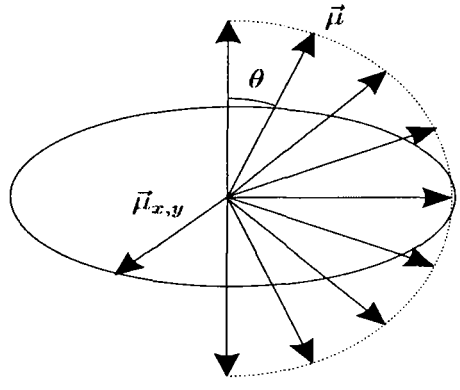
Der NMR-Spektroskopie sind nur solche Nukleide zugänglich, für die $I \neq 0$ gilt.
Ohne äußere Einflüsse und im thermischen Gleichgewicht sind alle magnetischen Momente $\vec{\mu}$ gleichwahrscheinlich, d.h. die Gesamtmagnetisierung eines Spinsystems (d.h. eine große Anzahl von Nukleiden mit gleicher Spinquantenzahl) ist Null.
Wir betrachten nun ein Spinsystem in einem homogenen Magnetfeld. Zur Vereinfachung nehmen wir an:

$$\vec{B}_0 = B_0 \, \vec{e}_z \tag{1.2}$$

Während ein makroskopischer magnetischer Dipol (z.B. eine Kompassnadel) sich in Richtung von $\vec{B}_0$ ausrichtet, kann im atomaren Bereich ein magnetischer Moment nach den Gesetzen der Quantenmechanik nur folgende Richtungen annehmen: $\vec{\mu} = \mu_x \vec{e}_x + \mu_y \vec{e}_y + \mu_z \vec{e}_z$ mit

$$\mu_z = \gamma \, \hbar \, m_I$$

mit $m_I \in \{-I, -I+1, ..., I-1, I\}$. Die Komponenten μ_x und μ_y sind zufällig (gleichverteilt). Wir erhalten anschaulich folgende Situation:



Es gilt

$$\vec{\mu}_{x,y} = \mu_x \, \vec{e}_x + \mu_y \, \vec{e}_y = |\vec{\mu}_{x,y}| \, (\cos(\xi) \, \vec{e}_x + \sin(\xi) \, \vec{e}_y)$$

wobei ξ gleichverteilt in $[0, 2\pi)$ und

$$|\vec{\mu}_{x,y}| = \sqrt{\mu^2 - \mu_z^2} = \gamma \, \hbar \, \sqrt{I(I+1) - m_I}$$

ist. Ferner gilt

$$\cos(\theta) = \frac{\mu_z}{\mu} = \frac{m_I}{\sqrt{I(I+1)}} \, .$$

Nach den Gesetzen der Elektrodynamik wirkt auf $\vec{J}$ im magnetischen Feld ein Drehmoment $\vec{\mu} \times \vec{B}_0$. Damit folgt

$$\frac{d\vec{J}}{dt} = \vec{\mu} \times \vec{B}_0 \quad \text{bzw.} \quad \frac{d\vec{\mu}}{dt} = \gamma\, \vec{\mu} \times \vec{B}_0 \tag{1.3}$$

Mit (1.2) folgt aus (1.3)

$$\frac{d\vec{\mu}}{dt} = \gamma\, B_0\, (\mu_y\, \vec{e}_x - \mu_x\, \vec{e}_y)$$

oder in Koordinatenform

$$\begin{aligned} \frac{d\mu_x}{dt} &= \gamma\, B_0\, \mu_y\,, \\ \frac{d\mu_y}{dt} &= \gamma\, B_0\, \mu_x\,, \\ \frac{d\mu_z}{dt} &= 0\,. \end{aligned}$$

Als Lösung erhalten wir

$$\mu_z(t) = \mu_z(0), \quad \begin{pmatrix} \mu_x(t) \\ \mu_y(t) \end{pmatrix} = e^{A\,t} \begin{pmatrix} \mu_x(0) \\ \mu_y(0) \end{pmatrix},$$

wobei $A = \begin{pmatrix} 0 & \gamma\, B_0 \\ -\gamma\, B_0 & 0 \end{pmatrix}$ ist.

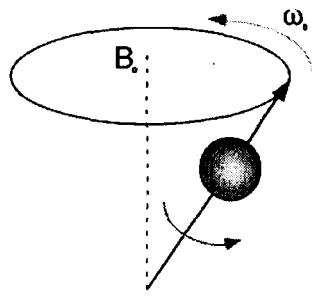
Explizit ergibt sich

$$\begin{aligned} \mu_x(t) &= +\mu_x(0)\cos(\omega_0 t) + \mu_y(0)\sin(\omega_0 t)\,, \\ \mu_y(t) &= -\mu_x(0)\sin(\omega_0 t) + \mu_y(0)\cos(\omega_0 t)\,, \\ \mu_z(t) &= \mu_z(0)\,, \end{aligned} \tag{1.4}$$

wobei

$$\omega_0 = \gamma\, B_0 \tag{1.5}$$

ist. Die Frequenz ω_0 heißt Larmor-Frequenz.
Das magnetische Moment präzediert mit der Larmor-Frequenz ω_0 um die z-Achse.



2 Gesamtmagnetisierung

Wir betrachten nun die Magnetisierung einer großen Anzahl von magnetischen Dipolen. Es sei N_0 die Anzahl der magnetischen Dipole im betrachteten System. Der Magnetisierungsfaktor $\vec{M}$ ist gegeben durch

$$\vec{M} = \sum_{n=1}^{N_s} \vec{\mu}_n \,.$$

Im Gleichgewicht und ohne äßere Einflüsse ist $\vec{M} = 0$.
Die potentielle Energie eines magnetischen Dipols im magnetischen Feld $\vec{B}_0$ ist gegeben durch

$$E = - < \vec{\mu}, \vec{B}_0 >= -\mu_z B_0 = -\gamma \hbar m_I B_0$$

Es ergeben sich somit die Energiezustände:

$$\hbar\omega_0 I\,, \quad \hbar\omega_0(I-1), \quad \cdots, \quad -\hbar\omega_0(I-1), \quad -\hbar\omega_0 I.$$

Die Energiedifferenz zwischen zwei aufeinanderfolgenden Energieniveaus ist

$$\triangle E = \hbar\omega_0.$$

Es ist

$$\begin{aligned} \vec{M} &= M_x\,\vec{e}_x + M_y\,\vec{e}_y + M_z\,\vec{e}_z \\ &= (\sum_{n=1}^{N_s} \vec{\mu}_{n,x})\,\vec{e}_x + (\sum_{n=1}^{N_s} \vec{\mu}_{n,y})\,\vec{e}_y + (\sum_{n=1}^{N_s} \vec{\mu}_{n,z})\,\vec{e}_z \end{aligned}$$

Wegen der Gleichverteilung der Komponenten $\mu_{n,x}$ und $\mu_{n,y}$ ist

$\sum_{n=1}^{N_s} \vec{\mu}_{n,x} = 0, \sum_{n=1}^{N_s} \vec{\mu}_{n,y} = 0.$ Es gilt daher

$$\begin{aligned} \vec{M} &= \sum_{n=1}^{N_s} \vec{\mu}_{n,z}\, \vec{e}_z \\ &= \gamma\, \hbar\, (N_I I + N_{I-1}(I-1) + \cdots + N_{-I+1}(-I+1) + N_{-I}(-I))\, \vec{e}_z\,, \end{aligned}$$

wobei N_I die Anzahl der Nukleide mit Spin I ist.
Die Magnetisierung ist nur dann von Null verschieden, falls

$$N_I I + N_{I-1}(I-1) + \cdots + N_{-I}(-I) \neq 0$$

ist.

Spezialfall: $I = 1/2$
In diesem Fall ist

$$\vec{M} = \frac{1}{2}\, \gamma\, \hbar\, (N_{\frac{1}{2}} - N_{-\frac{1}{2}})\, \vec{e}_z$$

Nach der Boltzmann-Statistik ist

$$\frac{N_{\frac{1}{2}}}{N_{-\frac{1}{2}}} = \exp(\frac{\triangle E}{KT_s})\,,$$

wobei $K = 1,28 \cdot 10^{-23} J/K$ die Boltzmann-Konstante und T_s die absolute Temperatur des Systems ist.
Wegen $\dfrac{\triangle E}{KT_s} > 0$ folgt

$$\frac{N_{\frac{1}{2}}}{N_{-\frac{1}{2}}} > 1 \Longleftrightarrow N_{\frac{1}{2}} - N_{-\frac{1}{2}} > 0$$

In der Praxis gilt $\triangle E \ll KT_s$, so dass

$$\frac{N_{\frac{1}{2}}}{N_{-\frac{1}{2}}} \approx 1 + \frac{\gamma \hbar B_0}{2KT_s}$$

bzw.

$$(1 + \frac{\gamma \hbar B_0}{2KT_s}) N_{\frac{1}{2}} - N_{-\frac{1}{2}} \approx N_s \frac{\gamma \hbar B_0}{2KT_s}$$

oder

$$N_{\frac{1}{2}} - N_{-\frac{1}{2}} \approx N_s \frac{\gamma \hbar B_0}{2KT_s}\,.$$

Damit ist

$$|\vec{M}| \approx \frac{\gamma^2 \hbar^2 B_0}{4KT_s} N_s .$$

Um die Magnetisierung des Spinsystems zu erhöhen haben wir folgende Möglichkeiten:

(a) Vergrößerung von B_0,

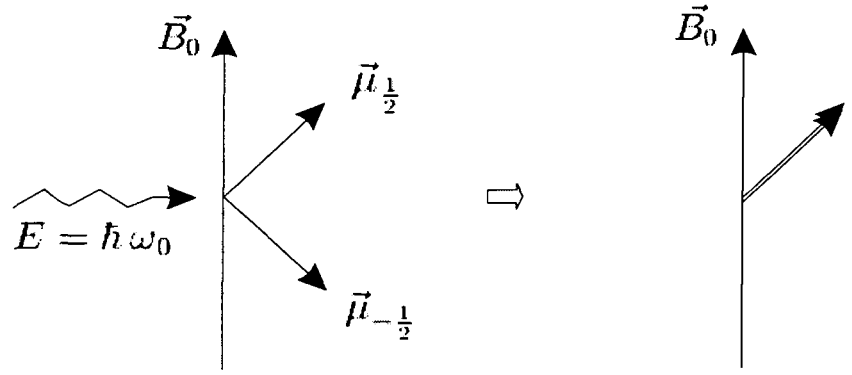
(b) Verkleinerung von T_s, d.h. Kühlung des Systems.

3 Radio-Frequenz Impulsanregung des Spinsystems

Strahlt man eine elektromagnetische Welle mit der Energie

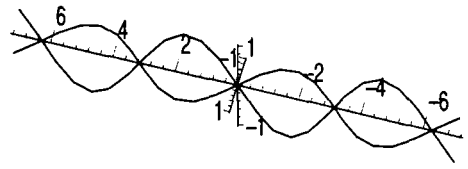
$$E = \hbar\omega$$

auf das Spinsystem, so kommt es bei der Larmor-Frequenz $\omega = \omega_0$ zu Absorption.



Die elektromagnetische Welle besitzt die Komponenten und $\vec{E}(\vec{r},t)$ und $\vec{B}(\vec{r},t)$ und es gilt $\vec{E}(\vec{r},t) \perp \vec{B}(\vec{r},t)$.
Eine linear polarisierte in y-Richtung laufende Welle besitzt die Komponenten

$$\begin{aligned} \vec{E}(\vec{r},t) &= E\cos(\omega(t-\frac{1}{c}y))\,\vec{e}_z\,, \\ \vec{B}(\vec{r},t) &= B\cos(\omega(t-\frac{1}{c}y))\,\vec{e}_x\,. \end{aligned}$$



Für die weiteren Überlegungen gehen wir davon aus, dass das eingestrahlte Magnetfeld ein homogenes Wechselfeld der Form

$$\vec{B}_1(t) = 2B_1(t)\cos(\omega t)\,\vec{e}_x$$

ist. Es gilt

$$\begin{aligned}\vec{B}_1(t) &= B_1(t)\,[\cos(\omega t)\vec{e}_x - \sin(\omega t)\vec{e}_y] + B_1(t)\,[\cos(\omega t)\vec{e}_x + \sin(\omega t)\vec{e}_y] \\ &=: \vec{B}_L(t) + \vec{B}_R(t)\end{aligned}$$

Diese Zerlegung in ein links- und ein rechtsdrehendes Magnetfeld hat folgende Begründung. Die Komponente $\vec{B}_L$ dreht sich in Richtung der Spinrotation, während $\vec{B}_R$ entgegengesetzt dreht.
Mit Hilfe der Quantenmechanik kann man zeigen, dass für Frequenzen ω, die in der Nähe der Larmor-Frequenz liegen, die gegen den Drehsinn des Spin rotierende Feldkomponente $\vec{B}_R$ keinen Einfluss auf $\vec{\mu}$ hat.
Wir können daher ohne Einschränkung ein Magnetfeld der Form

$$\vec{B}_1(t) = B_1(t)\cos(\omega t)\,\vec{e}_x - B_1(t)\sin(\omega t)\,\vec{e}_y$$

annehmen.
Nach den Gesetzen der Elektrodynamik erhalten wir somit

$$\frac{d\vec{\mu}}{dt} = \gamma\,\vec{\mu} \times (\vec{B}_0 + \vec{B}_1(t)) \quad \text{bzw.} \quad \frac{d\vec{M}}{dt} = \gamma\,\vec{M} \times (\vec{B}_0 + \vec{B}_1(t))$$

für die freie Präzesion.
Für die weitere Untersuchung dieser Bewegung führen wir einen Basiswchsel durch

$$\begin{aligned}\vec{u}_x &= \cos(\omega t)\,\vec{e}_x - \sin(\omega t)\,\vec{e}_y \\ \vec{u}_y &= \cos(\omega t)\,\vec{e}_x + \sin(\omega t)\,\vec{e}_y \\ \vec{u}_z &= \vec{e}_z\end{aligned}$$

Es ist

$$\begin{aligned}\frac{d\vec{u}_x}{dt} &= -\omega\sin(\omega t)\,\vec{e}_x - \omega\cos(\omega t)\,\vec{e}_y\,, \\ \frac{d\vec{u}_y}{dt} &= +\omega\cos(\omega t)\,\vec{e}_x - \omega\sin(\omega t)\,\vec{e}_y\,, \\ \frac{d\vec{u}_z}{dt} &= 0\,.\end{aligned}$$

Mit $\vec{\omega} = \omega \vec{u}_z$ folgt aus $\omega \vec{u}_y = \vec{\omega} \times \vec{u}_x$, $\omega \vec{u}_x = \vec{\omega} \times \vec{u}_y$

$$\frac{d\vec{u}_x}{dt} = \vec{\omega} \times \vec{u}_x\,, \quad \frac{d\vec{u}_y}{dt} = \vec{\omega} \times \vec{u}_y\,, \quad \frac{d\vec{u}_z}{dt} = \vec{\omega} \times \vec{u}_z\,.$$

Es gilt:

$$\vec{M} = M_x\vec{e}_x + M_y\vec{e}_y + M_z\vec{e}_z = M_x^u\vec{u}_x + M_y^u\vec{u}_y + M_z^u\vec{u}_z$$

und

$$\begin{aligned}\frac{d\vec{M}}{dt} &= \frac{dM_x^u}{dt}\vec{u}_x + \frac{dM_y^u}{dt}\vec{u}_y + \frac{dM_z^u}{dt}\vec{u}_z + M_x^u\frac{d\vec{u}_x}{dt} + M_y^u\frac{d\vec{u}_y}{dt} + M_z^u\frac{d\vec{u}_z}{dt} \\ &= \frac{\partial\vec{M}^u}{\partial u} + \vec{\omega} \times (M_x^u\vec{u}_x + M_y^u\vec{u}_y + M_z^u\vec{u}_z)\,,\end{aligned}$$

wobei $\vec{M}^u = M_x^u\vec{u}_x + M_y^u\vec{u}_y + M_z^u\vec{u}_z$ die Darstellung von $\vec{M}$ im rotierenden Koordinatensystem ist.
Insgesamt ist also

$$\frac{d\vec{M}}{dt} = \frac{\partial\vec{M}^u}{\partial t} + \vec{\omega} \times \vec{M}^u \tag{3.1}$$

$\frac{\partial\vec{M}^u}{\partial t}$ gibt die Relativbewegung im rotierenden Koordinatensystem wieder.

4 Die Bloch-Gleichungen

Die Bloch-Gleichungen lauten

$$\frac{d\vec{M}}{dt} = \gamma\,\vec{M} \times \vec{B} - \frac{1}{T_2}\,(M_x\vec{e}_x + M_y\vec{e}_y) - \frac{1}{T_1}\,(M_z - M_z^0)\,\vec{e}_z \tag{4.1}$$

Hierbei bedeuten:

M_z^0 der Betrag der Magnetisierung im thermischen Gleichgewicht bei einem homogenen äußeren Magnetfeld $\vec{B}_0 = B_0\,\vec{e}_z$.

T_1, T_2 sind Zeitkonstanten, welche den thermischen Relaxationsproszess beschreiben. D.h.: Nach den Gesetzen der Thermodynamik kehrt das Spinsystem nachdem es durch einen RF-Impuls ausgelenkt wurde, in das thermische Gleichgewicht zurück.

Man kann zwei Phasen bei der Spindynamik unterscheiden:

(a) RF-Anregung: Gewöhnlich ist die Dauer des RF-Impulses sehr kurz, d.h. $B_1(t) \neq 0$ für $t \in [0, T_{RF}]$, wobei $T_{RF} \ll T_i$, $i = 1, 2$. Während dieser Phase können die $\frac{1}{T_2}$ bzw. $\frac{1}{T_1}$ Terme in (4.1) vernachlässigt werden. Es ist also

$$\frac{d\vec{M}}{dt} = \gamma \vec{M} \times \vec{B} = \gamma \vec{M} \times (\vec{B}_0 + \vec{B}_1)$$

für $t \in [0, T_{RF}]$.

(b) Relaxation: Nach Abschaltung des RF-Impulses kehrt das System in das thermische Gleichgewicht zurück.

Die Trennung der Dynamik gelingt am Besten im rotierenden System

Es ist

$$\frac{d\vec{M}}{dt} = \gamma \vec{M} \times \vec{B}$$

Mit (3.1) ergibt sich:

$$\begin{aligned} \frac{\partial \vec{M}^u}{\partial t} &= \gamma \vec{M}^u \times \vec{B}^u - \vec{\omega} \times \vec{M}^u \\ &= \gamma \vec{M}^u \times (\vec{B}^u + \frac{1}{\gamma} \vec{\omega}), \end{aligned}$$

wobei $\vec{M}^u$ bzw. $\vec{B}^u$ die Vektoren im rotierenden Koordinatensystem sind. Das effektive Magnetfeld ist

$$\vec{B}_{eff} = \vec{B}^u + \frac{1}{\gamma} \vec{\omega} \tag{4.2}$$

Die Bloch-Gleichungen sind dann gegeben durch

$$\frac{\partial \vec{M}^u}{\partial t} = \gamma \vec{M}^u \times \vec{B}_{eff} - \frac{1}{T_2} (M_x^u \vec{u}_x + M_y^u \vec{u}_y) - \frac{1}{T_1} (M_z^u - M_z^0) \vec{u}_z \tag{4.3}$$

(a) RF-Anregung:
Wegen $T_{RF} \ll T_i$, $i = 1, 2$ ist

$$\frac{d\vec{M}}{dt} = \gamma \vec{M} \times \vec{B}\,.$$

bzw. im rotierenden System

$$\frac{\partial \vec{M}^u}{\partial t} = \gamma \vec{M} \times \vec{B}_{eff} \tag{4.4}$$

Es gilt $\vec{B}_1(t) = B_1 \cos(\omega t)\,\vec{e}_x - B_1 \sin(\omega t)\,\vec{e}_y$
Mit

$$\begin{pmatrix} B^u_{1,x} \\ B^u_{1,y} \end{pmatrix} = \begin{pmatrix} \cos(\omega t) & -\sin(\omega t) \\ \sin(\omega t) & \cos(\omega t) \end{pmatrix} \begin{pmatrix} B_1 \cos(\omega t) \\ -B_1 \cos(\omega t) \end{pmatrix} = \begin{pmatrix} B_1 \\ 0 \end{pmatrix}$$

gilt

$$\vec{B}^u_1(t) = B_1\,\vec{u}_x$$

Daher ist

$$\vec{B}_{eff}(t) = B_0\vec{u}_z + B_1\vec{u}_x + \frac{1}{\gamma}\vec{\omega} = (B_0 - \frac{\omega}{\gamma})\,\vec{u}_z + B_1\vec{u}_x$$

wobei $\vec{\omega} = -\omega\,\vec{u}_z$ verwendet wurde. (s.o.)
Für die Resonanzfrequenz $\omega = \gamma B_0$ ist somit

$$\vec{B}_{eff}(t) = B_1(t)\,\vec{u}_x$$

Somit liefert (4.4) im Fall der Resonanz $\omega = \gamma B_0$

$$\frac{\partial \vec{M}^u}{\partial t} = \gamma \vec{M}^u \times B_1(t)\,\vec{u}_x$$

In Koordinatenform erhält man das Dgl-System:

$$\frac{dM_x}{dt} = 0\,, \quad \frac{dM_y}{dt} = \gamma B_1(t)\,M_z\,, \quad \frac{dM_z}{dt} = -\gamma B_1(t)\,M_y$$

mit den Anfangsbedingungen $M_x(0) = 0\,, \quad M_y(0) = 0\,, \quad M_z(0) = M_z^0$
Folglich ist für $t < T_{RF}$

$$\begin{aligned} M^u_x(t) &= 0\,, \\ M^u_y(t) &= M^0_z \cos(\int_0^t \gamma B_1(\tilde{t})\,d\tilde{t})\,, \\ M^u_z(t) &= M^0_z \cos(\int_0^t \gamma B_1(\tilde{t})\,d\tilde{t})\,. \end{aligned}$$

Für die spezielle Funktion $B_1(t) = B_1\,\chi_{[0,T_{RF}]}(t)$ ist

$$\begin{aligned} M_x^u(t) &= 0\,, \\ M_y^u(t) &= M_z^0\cos(\gamma B_1 t)\,, \\ M_z^u(t) &= M_z^0\cos(\gamma B_1 t)\,. \end{aligned}$$

für $t \in [0, T_{RF}]$. Im stationären Koordinatensystem erhalten wir

$$\begin{pmatrix} M_x \\ M_y \\ M_z \end{pmatrix} = \begin{pmatrix} \cos(\omega t) & \sin(\omega t) & 0 \\ -\sin(\omega t) & \cos(\omega t) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} M_x^u \\ M_y^u \\ M_z^u \end{pmatrix}$$

$$\begin{aligned} M_x(t) &= M_z^0\sin(\omega t)\cos(\omega_1 t)\,, \\ M_y(t) &= M_z^0\cos(\omega t)\sin(\omega_1 t)\,, \\ M_z(t) &= M_z^0\sin(\omega_1 t)\,. \end{aligned}$$

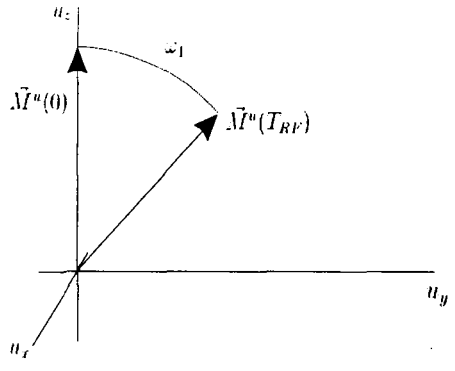
wobei $\omega_1 = \gamma B_1$.

(b) Relaxation:
Für $t > T_{RF}$ ist $B_1(t) \equiv 0$, d.h.

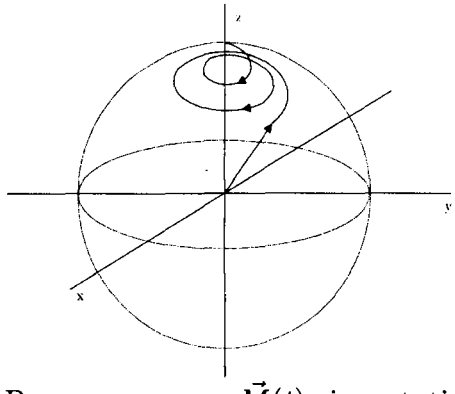
$$\vec{B}_{eff}(t) = (B_0 - \frac{\omega}{\gamma})\,\vec{u}_z + B_1\vec{u}_x = (B_0 - \frac{\omega}{\gamma})\,\vec{u}_z$$

und für $\omega = \gamma B_0$ gilt also $\vec{B}_{eff} = 0$.
Damit ergibt sich

$$\frac{\partial \vec{M}^u}{\partial t} = -\frac{1}{T_2}\,(M_x^u\vec{u}_x + M_y^u\vec{u}_y) - \frac{1}{T_1}\,(M_z^u - M_z^0)\,\vec{u}_z$$



Bewegung von $\vec{M}(t)$ im rotierenden System



Bewegung von $\vec{M}(t)$ im stationären Koordinatensystem

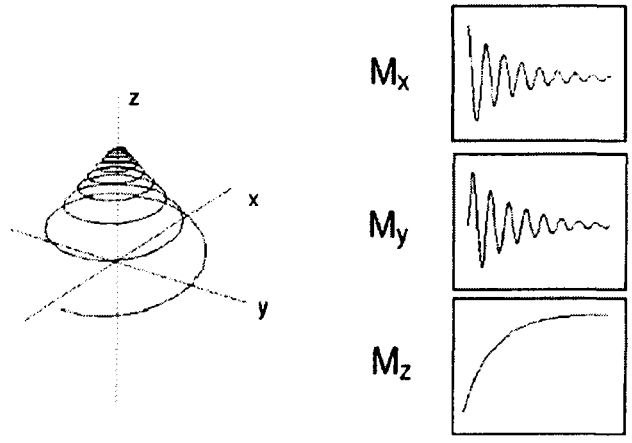
d.h.

$$\frac{d}{dt}\begin{pmatrix} M_x^u \\ M_y^u \\ M_z^u \end{pmatrix} = \begin{pmatrix} -\frac{1}{T_2} & 0 & 0 \\ 0 & -\frac{1}{T_2} & 0 \\ 0 & 0 & -\frac{1}{T_1} \end{pmatrix} \begin{pmatrix} M_x^u \\ M_y^u \\ M_z^u \end{pmatrix} + \frac{1}{T_1}\begin{pmatrix} 0 \\ 0 \\ M_z^0 \end{pmatrix}$$

Also ist

$$\begin{aligned} M_x^u(t) &= M_x(0)\, e^{-\frac{t}{T_2}}, \\ M_y^u(t) &= M_y(0)\, e^{-\frac{t}{T_2}}, \\ M_z^u(t) &= M_z(0)\, e^{-\frac{t}{T_1}} + M_z^0\,(1 - e^{-\frac{t}{T_1}}). \end{aligned}$$

$$\begin{aligned} M_x(t) &= (\cos(\omega t + \varphi_1)\, M_x(0) + \sin(\omega t + \varphi_1)\, M_y(0))\, e^{-\frac{t}{T_2}} \\ M_y(t) &= (-\sin(\omega t + \varphi_1)\, M_x(0) + \cos(\omega t + \varphi_1)\, M_y(0))\, e^{-\frac{t}{T_2}} \\ M_z(t) &= M_z(0)\, e^{-\frac{t}{T_1}} + M_z^0\,(1 - e^{-\frac{t}{T_1}}) \end{aligned}$$



Bewegung des $\vec{M}$-Vektors im festen Koordinatensystem. (Schraubenlinie)

Hierbei ist φ_1 die Phasenverschiebung, die durch die RF-Anregung hervorgerufen wurde.

5 Signalzerlegung

Als NMR-Signal wird eine elektrische Spannung gemessen, die durch die zeitliche Veränderung der Gesamtmagnetisierung und der damit verbundenen zeitlichen Veränderung des magnetischen Flusses in einer Spule induziert wird, die die Messprobe umgibt.

Der zeitabhängige Fluss ist gegeben durch

$$\Phi(t) = \int_{\Omega} < \vec{B}_r(\bar{r}), \vec{M}(\bar{r}, t) > d\bar{r} \tag{5.1}$$

wobei $\Omega \subset \mathbb{R}^3$ das Gebiet ist, das durch die Probe eingenommen wurd und $\vec{B}_r(\bar{r})$ ein Referenzfeld ist, was orthogonal zur Leiterebene liegt. (siehe Skizze) Nach dem Faradayschen Gesetz folgt dann für die induzierte Spannung

$$\begin{aligned} V(t) &= -\frac{d}{dt} \int_{\Omega} < \vec{B}_r(\bar{r}), \vec{M}(\bar{r}, t) > d\bar{r} \\ &= -\frac{d}{dt} \int_{\Omega} [B_{r,x}(\bar{r})\, M_x(\bar{r}, t) + B_{r,y}(\bar{r})\, M_y(\bar{r}, t) + B_{r,z}(\bar{r})\, M_z(\bar{r}, t)]\, d\bar{r} \end{aligned}$$

Physikalische Näherung: Die zeitliche Veränderung der M_z-Komponente ist sehr viel langsamer als die zeitlichen Veränderungen von M_x bzw. M_y. $\frac{d}{dt} M_z$ wird daher vernachlässigt.

Betrachtet man nun $V(t)$ für $t > T_{RF}$, so gilt

$$\begin{aligned} M_x(t) &= [M_x(\bar{r}, 0)\, \cos(\omega(\bar{r})t + \varphi_1(\bar{r})) + M_y(\bar{r}, 0)\, \sin(\omega(\bar{r})t + \varphi_1(\bar{r}))]\, e^{-\frac{t}{T_2(\bar{r})}} \\ M_y(t) &= [-M_x(\bar{r}, 0)\, \sin(\omega(\bar{r})t + \varphi_1(\bar{r})) + M_y(\bar{r})\, \cos(\omega(\bar{r})t + \varphi_1(\bar{r}))]\, e^{-\frac{t}{T_2(\bar{r})}} \end{aligned}$$

Daher folgt

$$\begin{aligned} \frac{\partial M_x(\bar{r}, t)}{\partial t} &= \omega(\bar{r})[-M_x(\bar{r}, 0)\, \sin(\omega(\bar{r})t + \varphi_1(\bar{r})) \\ &\quad + M_y(\bar{r}, 0)\, \cos(\omega(\bar{r})t + \varphi_1(\bar{r}))]\, e^{-\frac{t}{T_2(\bar{r})}} - \frac{1}{T_2(\bar{r})} e^{-\frac{t}{T_2(\bar{r})}} [\cdots] \\ \frac{\partial M_y(\bar{r}, t)}{\partial t} &= \omega(\bar{r})[-M_x(\bar{r}, 0)\, \cos(\omega(\bar{r})t + \varphi_1(\bar{r})) \\ &\quad + M_y(\bar{r}, 0)\, \sin(\omega(\bar{r})t + \varphi_1(\bar{r}))]\, e^{-\frac{t}{T_2(\bar{r})}} - \frac{1}{T_2(\bar{r})} e^{-\frac{t}{T_2(\bar{r})}} [\cdots] \end{aligned}$$

Da $\omega(\bar{r}) \gg \frac{1}{T_2(\bar{r})}$ ist, kann der zweite Term vernachlässigt werden. Ferner sei

$$\begin{aligned} B_{r,x}(\bar{r}) &= |B_{r,x,y}(\bar{r})| \cos(\varphi_r(\bar{r})), \\ B_{r,y}(\bar{r}) &= |B_{r,x,y}(\bar{r})| \sin(\varphi_r(\bar{r})). \end{aligned}$$

Daher ist

$$\begin{aligned} &< \vec{B}_r(\bar{r}), \vec{M}(\bar{r},t) >= |B_{r,x,y}(\bar{r})|\,\omega(\bar{r})\,[-M_x \sin(\omega t+\varphi_1)\cos(\varphi_r) \\ &\qquad +M_y \cos(\omega t+\varphi_1)\cos(\varphi_r) \\ &\qquad -M_x \cos(\omega t+\varphi_1)\sin(\varphi_r) - M_y \sin(\omega t+\varphi_1)\sin(\varphi_r)]\, e^{-\frac{t}{T_2(\bar{r})}} \\ = \;& |B_{r,x,y}(\bar{r})|\,\omega(\bar{r})\,[-M_x \sin(\omega t+\varphi_1+\varphi_r(\bar{r})) \\ &\qquad +M_y \cos(\omega t+\varphi_1+\varphi_r(\bar{r}))]\, e^{-\frac{t}{T_2(\bar{r})}} \end{aligned}$$

also

$$\begin{aligned} V(t) = \int_\Omega & \omega(\bar{r})\,|B_{r,x,y}(\bar{r})|\,[M_y(\bar{r},0)\cos(\omega(\bar{r})t+\varphi_1(\bar{r})+\varphi_r(\bar{r})) \\ & -M_x(\bar{r},0)\sin(\omega(\bar{r})t+\varphi_1(\bar{r})+\varphi_r(\bar{r}))]\, e^{-\frac{t}{T_2(\bar{r})}}\, d\bar{r} \end{aligned} \tag{5.2}$$

Wie (5.2) zeigt hängt das Signal von folgenden Größen wesentlich ab:

1. $M_x(\bar{r},0)$ bzw. $M_y(\bar{r},0)$, d.h. von der transversalen Magnetisierung zum Zeitpunkt $t=0$. (d.h. nach der Anregung)

2. $\omega(\bar{r})$, d.h. der Präzesionsfrequenz.

Für die Signaldetektoren wird $V(t)$ auf eine harmonische Schwingung $\cos(\omega_0 t)$ bzw. $\sin(\omega_0 t)$ moduliert:

$$\begin{aligned}
2\,\cos(\omega_0 t)\,V(t) &= \int_\Omega \omega(\bar{r})\,|B_{r,x,y}(\bar{r})|\,[2\,M_y(\bar{r},0)\,\cos(\omega(\bar{r})t \\
&\quad +\varphi_1(\bar{r})+\varphi_r(\bar{r}))\,\cos(\omega_0 t) - 2\,M_x(\bar{r},0)\,\sin(\omega(\bar{r})t+\varphi_1(\bar{r}) \\
&\quad +\varphi_r(\bar{r}))\,\cos(\omega_0 t)]\,e^{-\frac{t}{T_2(\bar{r})}}\,d\bar{r} \\
= \int_\Omega \omega(\bar{r})\,|B_{r,x,y}(\bar{r})|\,[&M_y(\bar{r},0)\,\cos((\omega(\bar{r})+\omega_0)t+\varphi_1(\bar{r})+\varphi_r(\bar{r})) \\
&+ M_y(\bar{r},0)\,\cos((\omega(\bar{r})-\omega_0)t+\varphi_1(\bar{r})+\varphi_r(\bar{r})) \\
&- M_x(\bar{r},0)\,\sin((\omega(\bar{r})-\omega_0)t+\varphi_1(\bar{r})+\varphi_r(\bar{r})) \\
&- M_x(\bar{r},0)\,\sin((\omega(\bar{r})+\omega_0)t+\varphi_1(\bar{r})+\varphi_r(\bar{r}))]\,e^{-\frac{t}{T_2(\bar{r})}}\,d\bar{r} \\
2\,\sin(\omega_0 t)\,V(t) &= \int_\Omega \omega(\bar{r})\,|B_{r,x,y}(\bar{r})|\,[-M_y(\bar{r},0)\,\sin((\omega(\bar{r})-\omega_0)t \\
&\quad +\varphi_1(\bar{r})+\varphi_r(\bar{r})) \\
&\quad + M_y(\bar{r},0)\,\sin((\omega(\bar{r})+\omega_0)t+\varphi_1(\bar{r})+\varphi_r(\bar{r})) \\
&\quad - M_x(\bar{r},0)\,\cos((\omega(\bar{r})-\omega_0)t+\varphi_1(\bar{r})+\varphi_r(\bar{r})) \\
&\quad + M_x(\bar{r},0)\,\cos((\omega(\bar{r})+\omega_0)t+\varphi_1(\bar{r})+\varphi_r(\bar{r}))]\,e^{-\frac{t}{T_2(\bar{r})}}\,d\bar{r}
\end{aligned}$$

Die modulierten Signale werden mit einem Tiefpassfilter bearbeitet. Dadurch werden die Terme mit dem $(\omega(\bar{r})+\omega_0)$-Argument entfernt.

Wenn wir

$$S(t) = 2\,V(t)\,(\cos(\omega_0 t) + i\sin(\omega_0 t))$$

setzen, erhalten wir:

$$\begin{aligned}
S(t) &= \int_\Omega \omega(\bar{r})\,|B_{r,x,y}(\bar{r})|\,[M_y(\bar{r},0)\,(\cos(\theta(\bar{r},t)) - i\sin(\theta(\bar{r},t))) \\
&\quad - M_x(\bar{r},0)\,(\sin(\theta(\bar{r},t)) + i\cos(\theta(\bar{r},t)))]\,e^{-\frac{t}{T_2(\bar{r})}}\,d\bar{r}
\end{aligned}$$

Fourier Transformation Ionen Zyklotron Resonanz Massenspektrometrie

Josef Obermaier *

Zusammenfassung

Es werden die physikalischen Grundlagen der (Fourier Transform Ion Cyclotron Resonance Mass Spectrometry) erläutert. Insbesondere wird die Zyklotrongleichung hergeleitet, sowie der Vorgang der Teilchenanregung und Teilchendetektion näher betrachtet. Schließlich werden die Begriffe Trappingpotential und reduzierte Zyklotronfrequenz erklärt.

*Institute of Biomathematics and Biometry
GSF - National Research Center for Environment and Health,
Postfach 1129, D-85758 Oberschleißheim, Germany,
josef.obermaier@gsf.de, http://ibb.gsf.de

In diesem Artikel wir das Konzept der unter FT ICR MS bekannten Methode der Massenspektrometrie kurz zusammengefasst. Für Detailfragen sei auf die umfangreiche Literatur zu diesem Thema verwiesen.

1 Das Physikalische Grundprinzip

Auf ein sich in einem Magnetfeld $\vec{\mathbf{B}} = (0, 0, B)$ mit Geschwindigkeit $\vec{\mathbf{v}} = (\dot{x}, \dot{y}, \dot{z})$ bewegendes Teilchen mit Masse m und Ladung q wirkt eine Kraft $\vec{\mathbf{F}}$, die sogenannte Lorentz-Kraft. Die Kraft wirkt senkrecht zu der durch $\vec{\mathbf{v}}$ und $\vec{\mathbf{B}}$ aufgespannten Ebene. Die Richtung läßt sich durch die sogenannte Rechte-Hand-Regel bestimmen. Die Kraft $\vec{\mathbf{F}}$ ist gegeben durch das Vektorprodukt

$$\vec{\mathbf{F}} = q\vec{\mathbf{v}} \times \vec{\mathbf{B}}. \tag{1}$$

Nach dem 2. Newtonschen Gesetz wirkt eine Kraft auf eine Masse beschleunigend und man erhält die Differentialgleichung

$$m\frac{d\vec{\mathbf{v}}}{dt} = q\vec{\mathbf{v}} \times \vec{\mathbf{B}}. \tag{2}$$

Die Lösung des dazugehörigen Systems

$$\begin{aligned} m\ddot{x} &= qB\dot{y}, \\ m\ddot{y} &= -qB\dot{x}, \\ m\ddot{z} &= 0 \end{aligned} \tag{3}$$

ergibt bei gegebenen Anfangsbedingungen die Bahnkurve eines Kreises in der xy-Ebene und eine gleichförmige Bewegung in z-Richtung mit Geschwindigkeit $\dot{z}(0)$, insgesamt also eine Schraubenlinie.
Durch Gleichsetzen von Zentripetalkraft und Lorentz-Kraft erhält man

$$m\frac{v_{xy}^2}{r} = qv_{xy}B\,, \tag{4}$$

dabei ist $v_{xy} = \sqrt{\dot{x}^2 + \dot{y}^2}$ und r der Radius des Kreises. Man berechnet hieraus den Radius

$$r = \frac{mv_{xy}}{qB} \tag{5}$$

und die Umlaufzeit

$$T = \frac{2\pi m}{qB}\,. \tag{6}$$

Für die Winkelgeschwindigkeit des Teilchens folgt

$$\omega = \frac{qB}{m}. \tag{7}$$

ω nennt man auch die Zyklotronfrequenz des Teilchens. Sie hängt allein von der Masse des Teilchens ab und nicht von dessen kinetischer Energie.
Ziel ist es ein Signal zu erzeugen, aus dem sich ω bestimmen lässt. Aus der Zyklotrongleichung (7) lässt sich die Masse m des Teilchens mit Hilfe der Zyklotronfrequenz bestimmen. Bei unseren modellhaften Überlegungen setzen wir voraus, dass die Ladung q für jedes Teilchen identisch ist. Die Teilchen werden durch zwei Kondensatorplatten in der xy-Ebene mit $z = -\frac{a}{2}$ und $z = \frac{a}{2}$ begrenzt, die im Falle $q > 0$ positiv und im Falle $q < 0$ negativ geladen sind. Sie verhindern das Entweichen der Teilchen in z-Richtung. Siehe hierzu Abschnitt 4: Trappingpotential und reduzierte Zyklotronfrequenz.

2 Teilchenanregung

Im Folgenden betrachten wir nur Ionen einer Teilchenart und gehen davon aus, dass sie sich in unangeregtem Zustand als Teilchenpaket in der Nähe des Ursprungs $(0, 0, 0)$ befinden. Die durch Wärmeenergie erzeugten Radien der Teilchen ergegeben sich zu

$$r = \frac{\sqrt{2mkTe}}{qB}, \tag{8}$$

dabei bezeichnet k die Boltzmann-Konstante und Te die Temperatur. Für ein Teilchen mit Masse $m = 100u$ und Einheitsladung ergibt sich bei einem Magnetfeld von 3 Tesla etwa ein Radius $r = 0.08$ mm. Die Radien sind im nicht angeregten Zustand zu klein, um ein Signal zu erzeugen. Außerdem sind die Phasen der Teilchen zufällig und heben sich in ihrer induktiven Wirkung gegenseitig auf.
Den Teilchen wird Energie zugeführt, um die Bahnradien zu vergrößern. Zudem sollen die Teilchen einer Art dabei in Phase gebracht werden.
Hierzu wird ein Paar Kondensatorplatten in der xz-Ebene mit $y = -\frac{a}{2}$ und $y = \frac{a}{2}$ angebracht und an ihnen eine Wechselspannung

$$V_{ex}(t) = V_0 \sin(\omega_{ex} t) \tag{9}$$

angelegt.

Dadurch entsteht zwischen den Platten ein elektrisches Feld $\vec{\mathbf{E}}_{ex} = (0, E_y, 0)$. Es gilt

$$E_y(t) = E_0 \sin(\omega_{ex} t) \tag{10}$$

und $E_0 = \frac{2\beta V_0}{a}$, wobei β eine von der Architektur der Ionenfalle abhängige Konstante ist. Die maximale Leistungsaufnahme des Teilchens geschieht im Resonanzfall $\omega_{ex} = \omega$.
Nimmt man an, dass das Teilchen innerhalb eines Umlaufs weiterhin annähernd eine Kreisbahn beschreibt (absorbierte Energie während der Zeit $T \ll$ kinetische Energie des Teilchens) und dass das elektrische Feld mit der Geschwindigkeit $\dot{y}$ des Teilchens in Phase ist, so erhält man als mittlere Leistungsaufnahme

$$\langle P \rangle = \frac{1}{T} \int_0^T q E_y(t) \dot{y}(t) dt = \frac{q E_0 r \omega}{2} .$$

Die kinetische Energie des Teilchens ist gegeben durch $E_{kin} = \frac{m}{2}\omega^2 r^2$. Aus $P = \dot{E}_{kin} = m\omega^2 r \dot{r}$ folgt $\dot{r} = \frac{qE_0}{2m\omega}$. Mit der Annahme $r(0) = 0$ erhält man eine zeitabhängige Darstellung des Radius

$$r(t) = \frac{E_0}{2B} t \,, \tag{11}$$

die massenunabhängig ist.
Die bei Anregung $\omega_{ex} = \omega$ geltende Bewegungsgleichung lautet

$$m \frac{d\vec{\mathbf{v}}}{dt} = q \vec{\mathbf{v}} \times \vec{\mathbf{B}} + q \vec{\mathbf{E}}_{ex} \,. \tag{12}$$

Sie besitzt eine Lösung, deren wesentlicher Anteil eine von den Anfangsbedingungen unabhängige Archimedische Spirale

$$(x(t), y(t)) = \Big(-\frac{E_0}{2B} t \sin(\omega t),\ -\frac{E_0}{2B} t \cos(\omega t) \Big) \tag{13}$$

ist. Der wachsende Radius der Spirale entspricht (11). Da die Teilchen als Paket auf einen größeren Radius angehoben werden, entsteht Phasenkohärenz. In einem realen FT ICR MS Experiment wird nicht nur eine einzige Teilchenart angeregt, sondern ein oder mehrere Frequenzbänder. Eine Vorgehensweise dabei ist, die für das gewünschte Frequenzspektrum notwendige Anregungsspannung V_{ex} durch inverse Fourier-Transformation zu ermitteln. Dieses Verfahren nennt man SWIFT.

3 Teilchendetektion

Wir gehen nun davon aus, dass nach Anregung N Teilchen einer Art koherent eine Kreisbahn mit Radius r durchlaufen, deren Mittelpunkt der Ursprung

ist. Nun werden zwei weitere Kondensatorplatten in der yz-Ebene mit $x = -\frac{a}{2}$ und $x = \frac{a}{2}$ angebracht. Die Teilchen lassen sich als rotierender Dipol modellieren, die einen Signalstrom

$$I_s(t) = \frac{Nq^2rB}{md} \sin(\omega t) \tag{14}$$

in den verbundenen Platten induzieren. Man misst nun die an einem Widerstand R anfallende Signalspannung $V_s(t)$ und verstärkt diese. Der Mess- und Verstärkungsvorgang lässt sich als eine zu R parallel geschaltete Kapazität C betrachten. Aus der Wechselstromkreistheorie folgt

$$V_s(t) = \frac{Nq^2rB}{ma} \cdot \frac{1}{\sqrt{1/R^2 + \omega^2C^2}} \cdot \sin(\omega t + \varphi), \tag{15}$$

mit $\varphi = \arctan(-\omega CR)$. Geht man etwa von der Annahme aus, dass $1/R \ll \omega C$ gilt, d.h. dass der Wechselstromkreis primär kapazitiv ist, so ergibt sich als Signalspannung

$$V_s(t) = \frac{Nqr}{aC} \sin(\omega t - \frac{\pi}{2}) . \tag{16}$$

Die Amplitude hängt nur von der Teilchenzahl und nicht von deren Massen ab. Damit spiegeln die Höhenverhältnisse der Peaks im Frequenzspektrum das Verhältnis der Teilchenhäufigkeiten wieder. Im realen Experiment werden je N_i Teilchen verschiedener Massen m_i durch Breitbandanregung auf einen gemeinsamen Radius r angehoben. Durch Überlagerung entsteht somit ein Gesamtsignal

$$V_s(t) = \frac{qr}{aC} \sum_i N_i \sin(\omega_i t + \varphi_i) . \tag{17}$$

Bei Sensitivitätsüberlegungen, die hier nicht angestellt werden, bezieht man eine Rauschspannung V_n, verursacht durch einen durch Wärmeentwicklung am Widerstand R entstandenen Rauschstrom I_n, mit ein. Durch Teilchenkollision kann das Gesamtsignal gedämpft werden. Wir verfolgen hier nur den ungedämpften Fall in der Darstellung (17).

Das Signal wird nun innerhalb einer Zeitspanne T_{acq} abgetastet und diskretisiert. Die diskreten, zeitabhängigen Daten werden durch FFT in diskrete, frequenzabhängige Daten transformiert. Das sich hieraus ergebende Frequenzspektrum weist an der Stelle ω_i einer angeregten Teilchenart mit Masse $m_i = qB/\omega_i$ einen Peak auf. Die Abtastfrequenz des zeitabhängigen Signals muss sich nach der höchsten Frequenz ω_i richten, die man im Spektrum erkennen will. Dies folgt aus dem sogenannten Nyquist-Kriterium. Das Frequenzspektrum wird als Betrag der Fourier-transformierten Daten aufgetragen.

Von großer Bedeutung ist die Massentrennung. Als Auflösung $\Delta\omega_{50\%}$ bezeichnet man die Breite eines Peaks bei mittlerer Höhe. Im umgerechneten Massenspektrum entspricht ihr eine Massenauflösung $\Delta m_{50\%}$.
Die Auflösungsstärke wird definiert als $\omega/\Delta\omega_{50\%}$ bzw. $m/\Delta m_{50\%}$. Aus (7) folgt

$$\frac{d\omega}{dm} = -\frac{qB}{m^2} = -\frac{\omega}{m} \tag{18}$$

und somit

$$\frac{\omega}{d\omega} = -\frac{m}{dm}\,. \tag{19}$$

Daraus leitet sich

$$\frac{m}{\Delta m_{50\%}} = -\frac{qB}{m\Delta\omega_{50\%}} \tag{20}$$

ab. Mit Hilfe weiterer Überlegungen bezüglich des Spektrums findet man für das ungedämpfte Signal

$$\Delta m_{50\%} = \frac{7.583m^2}{qBT_{acq}} \tag{21}$$

bzw.

$$\frac{m}{\Delta m_{50\%}} = \frac{0.132qBT_{acq}}{m}\,. \tag{22}$$

4 Trappingpotential und reduzierte Zyklotronfrequenz

Dem in den Abschnitten 1 bis 3 vorgestellten Konzept folgend, gelangen wir zu einer kubischen Architektur einer Ionenfalle, bestehend aus einem Würfel sich jeweils gegenüberliegender Kondensatorplatten der Länge a. In der Praxis tritt eine Vielzahl weiterer Architekturen auf, etwa ein Zylinder. Gemeinsam haben sie, dass durch Anlegen einer Trappingspannung an den Trappingplatten ein Potential ϕ entsteht, welches die konzeptionellen Überlegungen der Abschnitte 1 bis 3 verkompliziert. Das quadrupolare elektrostatische Trappingpotential ist gegeben durch

$$\phi(x,y,z) = V_{trap}\Big(\gamma + \frac{\alpha}{2a^2}(2z^2 - x^2 - y^2)\Big) \tag{23}$$

bzw. in radialer Darstellung

$$\phi(r,z) = V_{trap}\Big(\gamma + \frac{\alpha}{2a^2}(2z^2 - r^2)\Big), \tag{24}$$

dabei bezeichnet a den Plattenabstand, γ und α sind architekturabhängige Konstanten und V_{trap} bezeichnet die an einer Trappingplatte angelegte Spannung.

Das hierdurch erzeugte elektrische Feld ist gegeben durch

$$\vec{\mathbf{E}}_{trap} = -\left(\frac{d\phi}{dx}, \frac{d\phi}{dy}, \frac{d\phi}{dz}\right) \tag{25}$$

oder in radialer Darstellung

$$\vec{\mathbf{E}}_{trap} = -\left(\frac{d\phi}{dr}, \frac{d\phi}{dz}\right). \tag{26}$$

Die modifizierte Bewegungsgleichung (7) für das Teilchen lautet

$$m\frac{d\vec{\mathbf{v}}}{dt} = q\vec{\mathbf{v}} \times \vec{\mathbf{B}} + q\vec{\mathbf{E}}_{trap}\,. \tag{27}$$

In z-Richtung führt das Teilchen mit der Anfangsbedingung $\dot{z}(0) = 0$ eine Schwingung

$$z(t) = z(0)\cos(\omega_{trap}t) \tag{28}$$

aus. Dabei ist die Trappingfrequenz gegeben durch

$$\omega_{trap} = \sqrt{\frac{2qV_{trap}\alpha}{ma^2}}\,. \tag{29}$$

In radialer Darstellung ergibt sich nach dem 2. Newtonschen Gesetz folgende auf das Teilchen wirkende Kraft

$$m\omega_v^2 r = qB\omega_v r - \frac{qV_{trap}\alpha}{a^2}r\,, \tag{30}$$

dabei steht ω_v für die Winkelgeschwindigkeit des Teilchens. Die von r unabhängige quadratische Gleichung ergibt die beiden Lösungen

$$\omega_+ = \frac{\omega}{2} + \sqrt{\left(\frac{\omega}{2}\right)^2 - \frac{\omega_{trap}^2}{2}} \tag{31}$$

und

$$\omega_- = \frac{\omega}{2} - \sqrt{\left(\frac{\omega}{2}\right)^2 - \frac{\omega_{trap}^2}{2}}\,. \tag{32}$$

Man bezeichnet ω_+ als reduzierte Zyklotronfrequenz und ω_- als Magnetronfrequenz. Das Teilchen bewegt sich in der xy-Ebene mit einer Winkelgeschwindigkeit ω_+ auf einem Kreis, deren Mittelpunkte mit der Winkelgeschwindigkeit ω_- wiederum auf einer Potentiallinie kreisen. Die reduzierte Zyklotronfrequenz ω_+ ist die eigentlich gemessene Frequenz. Die Magnetronfrequenz ω_- ist im Normalfall vergleichsweise sehr klein. Durch elementare Berechnung erhält man

$$\frac{m}{q} = \frac{B}{\omega_+} - \frac{\alpha V_{trap}}{a^2\omega_+^2}\,. \tag{33}$$

Daraus leitet sich folgende Gleichung zur Kalibrierung ab

$$\frac{m}{q} = \frac{C_1}{\omega_+} + \frac{C_2}{\omega_+^2} \,. \tag{34}$$

C_1 und C_2 sind Konstanten, die sich mit Hilfe zweier bekannter Massen bestimmen lassen. Die kritische Masse $m_{critical}$ resultiert aus der Gleichheit $\omega_+ = \omega_- = \frac{\omega}{2}$. Mit ihr beginnt die Bewegung des Teilchens instabil zu werden. Es gilt

$$m_{critical} = \frac{qB^2a^2}{4V_{trap}\alpha} \,. \tag{35}$$

5 Bezeichnungen und Konstanten

x, y, z bezeichnet die Raumkoordinaten, $\dot{}$ bzw. $\ddot{}$ die erste bzw. zweite Ableitung nach der Zeit t. Die Pfeile $\vec{}$ kennzeichnen vektorielle Größen.

a	Plattenabstand	(in m)
α	architekturabhängige Konstante ($\alpha = 2.77373$ für kubische Architektur)	
B	Magnetfeld	(in Tesla)
β	architekturabhängige Konstante ($\beta = 0.72167$ für kubische Architektur)	
C	Kapazität	(in Farad)
$\Delta m_{50\%}$	Massenauflösung	(in kg)
$\Delta\omega_{50\%}$	Frequenzauflösung	(in s^{-1})
E_0	Amplitude des elektrischen Feldes bei Anregung	
$\vec{\mathbf{E}}_{ex}$	Elektrisches Feld bei Anregung	(in $N \cdot C^{-1}$)
E_{kin}	kinetische Energie	(in Joule)
$\vec{\mathbf{E}}_{trap}$	Elektrisches Feld durch Trappingspannung	
E_y	y-Komponente des elektrischen Feldes bei Anregung	
$\vec{\mathbf{F}}$	Kraft	(in Newton)
φ, φ_i	Teilchenphase	
γ	architekturabhängige Konstante ($\gamma = 0.33333$ für kubische Architektur)	
I_n	Rauschstrom	
I_s	Signalstrom	(in Ampere)
k	Boltzmann-Konstante $(1.38066 \cdot 10^{-23} JK^{-1})$	
m, m_i	Masse	(in kg)
$m_{critical}$	kritische Masse	
N, N_i	Teilchenanzahl	
ω, ω_i	Zyklotronfrequenz	(s^{-1})
ω_+	Reduzierte Zyklotronfrequenz	
ω_-	Magnetronfrequenz	
ω_{ex}	Anregungsfrequenz	
ω_{trap}	Trappingfrequenz	
ω_v	Winkelgeschwindigkeit allgemein	
P	Leistung	(in Watt)
$\langle P \rangle$	mittlere Leistung	

ϕ	Trappingpotential	(in Volt)
q	Teilchenladung	(in Coulombs)
r	Radius	(in m)
R	Widerstand	(in Ohm)
T	Umlaufzeit	(in s)
T_{acq}	Datenerhebungszeit	
Te	Temperatur	(in Kelvin)
u	Atommasseneinheit	$(1.66054 \cdot 10^{-27}\text{kg})$
$\vec{\mathbf{v}}$	Geschwindigkeit	
v_{xy}	xy-Komponente der Geschwindigkeit	
V_0	Amplitude der Spannung bei Anregung	(in Volt)
V_{ex}	Spannung bei Anregung	
V_n	Rauschspannung	
V_s	Signalspannung	
V_{trapp}	Trappingspannung	

Literatur

[1] A.G. MARSHALL ET AL. (1998): Fourier Transform Ion Cyclotron Resonance Mass Spectrometry: A Primer. *Mass Spectrometry Reviews*, 17: 1–35.

Der Artikel enthält eine umfangreiche Literaturliste zum Thema.