Monte Carlo Simulations in Positron Emission Tomography Reconstruction

Full matrix, dual matrix, and system matrix compression

Dissertation

zur Erlangung des Grades eines Doktors
der Naturwissenschaften
der Fakultät für Mathematik und Physik
der Eberhard-Karls-Universität zu Tübingen

vorgelegt von
Niklas Sebastian Rehfeld
aus Berlin

2007

Tag der mündlichen Prüfung: 9. Mai 2007

Dekan:	Prof. Dr. Schopohl
1. Berichterstatter:	Prof. Dr. Dr. Schick
2. Berichterstatter:	PD Dr. Sibylle Ziegler

Summary

In positron emission tomography (PET) emission density images are formed by photon coincidence measurements. This process is complicated, particularly with regard to the photons that can be scattered in the inhomogeneous patient. A method to incorporate Monte Carlo simulations into the image formation process to model the scattering is presented. This is achieved by simulating the system matrix that describes the map from emission density to detected coincidences. The problem of the very large size of the matrix is met by fitting and B-spline compression of Monte Carlo results. A dedicated Monte Carlo code for system matrix calculation using variance reduction techniques is presented to reduce simulation time. Other desirable properties like reduced sensitivity to Monte Carlo noise and the possibility for sequential compression are met by the presented compression method. In proof-of-principle simulations of single ring scanners it is shown that the matrices compressed by this scheme are good approximations to the uncompressed matrices and that scatter artifacts in the images are strongly suppressed. In the last part, noise in the images introduced by the noise of the Monte Carlo simulated system matrices is investigated and quantified.

Zusammenfassung

In der Photonenemissionstomographie (PET) werden Bilder der Aktivitätsverteilung aus Photonen-Koinzidenzmessungen errechnet. Das ist insbesondere wegen der Streuung der Photonen im inhomogenen Patienten kompliziert. In der vorgestellten Methode werden Monte Carlo Simulationen in der Bildberechnung benutzt, um die Photonenstreuung zu bestimmen. Dabei wird die Systemmatrix, die die Abbildung der Aktivitätsverteilung auf meßbare Koinzidenzen beschreibt, mittels Monte Carlo Simulationen berechnet. Durch Parametrisierung und B-Spline Komprimierung wird die Größe der Matrix soweit reduziert, dass die Speicherung im Hauptspeicher möglich ist. Es wird ein Monte Carlo Programm vorgestellt, dass auf Systemmatrix Berechnungen spezialisiert ist und Varianzreduktionsmethoden verwendet. Andere wünschenswerte Eigenschaften wie geringe Anfälligkeit gegenüber Monte Carlo Rauschen und die Möglichkeit einer schrittweisen Kompression werden durch das vorgestellte Kompressionschema erfüllt. In Simulationen von Ein-Ring-Scannern wird beispielhaft gezeigt, dass die komprimierten Matrizen gute Näherungen der unkomprimierten Matrizen sind und dass Streuartefakte in den Bildern stark reduziert sind. In einem letzten Teil wird das Bildrauschen, das durch das Monte Carlo Rauschen der Matrizen verursacht wird, untersucht und quantifiziert.

Contents

1. Introduction **1**

1.1. Motivation and thesis outline 5

2. Physics of PET **7**

2.1. Positron emission and annihilation 7

2.2. The detection system 8

2.3. Geometry and interactions 10

2.3.1. Scattered coincidences 12

2.3.2. Attenuation 13

2.3.3. Random correction 14

2.4. Acquisition 15

2.4.1. PET 15

2.4.2. PET/CT 16

3. Image Reconstruction **17**

3.1. The Radon transform 17

3.2. Iterative algorithms 20

3.2.1. The system matrix 21

3.2.2. The objective function 22

3.2.3. Maximum likelihood expectation maximization 23

3.3. 3D scanners 25

3.4. Scatter correction 25

4. Monte Carlo Code **27**

4.1. Particle tracing 27

4.1.1. Linear attenuation coefficients 27

4.1.2. Tracing of particles in voxelized phantoms 29

4.2. Simulated particle detection 34

4.3. Variance reduction 34

4.3.1. Forced detection 35

4.3.2. Stratification 38
4.4. Implementation and parallelization 42

5. System Matrix Compression **43**
5.1. Goals and requirements 44
5.2. Properties of the system matrix 44
5.3. Compression scheme 45
5.3.1. Principle 47
5.3.2. Increasing robustness of compression scheme 53
5.3.3. Increasing compression speed 54
5.3.4. Read-out 55
5.3.5. Memory saving for 2D scanners and outlook for 3D-scanner matrix compression 56

6. Implemented Reconstruction Algorithms **59**
6.1. Monte Carlo maximum likelihood expectation maximization 59
6.1.1. Full matrix 59
6.1.2. No scatter modeling 60
6.1.3. Compressed matrix 60
6.2. Dual matrix maximum likelihood expectation maximization 61
6.3. A hybrid approach 62

7. Evaluation **65**
7.1. Simulated phantoms and scanner geometries 65
7.2. Measures used for quantification 67
7.3. Verification of the Monte Carlo code 68
7.4. Compressed matrix 69
7.4.1. Comparison of full matrix and compressed matrix 70
7.4.2. Comparison of reconstructed images 72
7.4.3. B-spline order and grid dimensions 77
7.5. The influence of Monte Carlo noise on the reconstructed images 79
7.5.1. Propagation of noise in iterative reconstructions 79
7.5.2. Convergence and noise propagation of the full matrix and the dual matrix algorithm 81
7.5.3. Discussion 89
7.6. Performance 91

8. Conclusions and Outlook **93**

A. Calculations **111**
A.1. Gaussian sampling 111
A.2. Variance of detected weighted counts 111

B. Paper: The influence of noise in full Monte Carlo ML-EM and dual ... **113**

C. Conference proceedings **125**
C.1. Monte Carlo noise in full Monte Carlo ML-EM and dual matrix recsontructions in 127
C.2. Reconstruction of PET images with a compressed Monte Carlo based system matrix 133
C.3. Compression of a Monte Carlo based system matrix for iterative reconstruction of PET 141

1. Introduction

Imaging is an important branch of medical diagnostics. Medical imaging methods provide the physician with information about the location of normal or pathological processes and structures in the human body. This information can be obtained indirectly by measuring physical properties of tissue. These properties include electron density, the density of hydrogen nuclei and nuclear spin relaxation times, or elastic properties. Medical imaging methods that can perform this task are planar X-ray imaging, (X-ray-)computed tomography (CT), (nuclear) magnetic resonance imaging (MRI)[1], and ultrasound imaging. The advantage of these methods is the usually rather high resolution (especially of X-ray based images) and the often low noise in the images. However, these methods are usually not well suited to measure and visualize biological or biochemical processes, because the measured physical quantities do often not provide the required information. Contrast agents (for all methods) and scan parameters (especially for MRI) can be used to add further physiological information, but still the possibilities are rather limited.

Other methods, based on emission measurements, are better suited to provide information about biochemical processes. In emission measurements a radioactive substance is brought into the patient (usually by injection) and the photons that leave the body are detected. The photons are either a direct product of the decay process or originate from annihilation of the positron in the vicinity of the decaying radionuclide. The latter effect is used in positron emission tomography (PET) and the first in scintigraphy and single photon emission computed tomography (SPECT). X-ray imaging, computed tomography, scintigraphy, single photon computed tomography and positron emission tomography are all based on ionizing radiation. The two first methods, however, are based on *transmission measurement*s. Those methods use an external source of photons on one side and detectors on the other side of the patient to measure the attenuation.

In emission tomography[2], radioactive tracers or biomarkers are used to visualize

[1]The word "nuclear" is usually not used in medical context, because it sounds dangerous.

[2]The word tomography stands for a method to obtain images that represent slices of the scanned patient/object, nowadays also often used as a synonym for volume imaging.

biochemical processes under investigation. These tracers are molecules which are designed to accumulate in certain regions of the body under specific circumstances. Usually these molecules are similar or even identical to molecules that are part of the human metabolism. In order to be localizable, at least one atom of the tracer is radioactive. In the case of PET this atom is a β^+-emitter, in the case of SPECT it is a photon emitter. Common β^+-emitters in nuclear medicine are F-18 ($\tau_{1/2} = 110\,\mathrm{min}$), O-15 (2.0 min), C-11 (20.5 min) and N-13 (10.0 min). Common photon emitters are $^{99\mathrm{m}}$Tc(6 h) or I-123 (13 h). In contrast to SPECT radionuclides, PET radionuclides are rather short lived and therefore special infrastructure is needed. The radioactive atoms that can be used in PET are very common in biochemical molecules. Tracer molecules range from the very common ^{18}F-FDG (fluorodeoxyglucose, PET) which resembles glucose and therefore gets accumulated in regions of high energy consumption like tumors, over oxygen-sensitive tracers like F-MISO (PET) that can be used to locate hypoxic areas, I-123 (SPECT) that can be used in thyroid diagnostic, and $^{99\mathrm{m}}$Tc (SPECT) that is used in bone scans, to very specific ligands for selected metabolic processes (PET, SPECT).

The tracers in SPECT are located by detecting the decay photons that leave the patient. Before the photons reach the detector they must pass a collimator. A collimator is made of a highly absorbing material (such as tungsten or lead) that is placed in front of the detectors and which restricts the detected photons to those that reach the detector within a certain incident angle interval. In this way the origin of the emission can be localized to a part of the patient. The information of many of such photons that are detected after passing the collimators can be used to determine the most likely emission density (or activity distribution). This process, the calculation of images using the given measured data, is called reconstruction. On one side, the collimators provide more detailed information about the origins of the decays, but on the other side the number of detected photons is strongly reduced. In SPECT usually 99.9% or even more of the emitted photons are blocked by the collimators. This leads to a reduced sensitivity and noise in the reconstructed images. In constrast to transmission tomography where the position of the source of photons is known (and which therefore defines together with the position of detection the path of the photons), the collimation therefore increases strongly the noise in the measured data. An additional problem of emission tomography is the limited amount of tracer that can be injected due to the dose that is deposited in the patient. In contrast to transmission tomography this is more limiting, because dose is not only deposited during the examination, but also later due to the remaining tracer in the patient. Therefore, there is a great benefit in increasing the fraction of detected photons.

In positron emission tomography an alternative method is used to localize the position of decay. The positron is annihilated in the vicinity of the decay position (submillimeter to several millimeter according to tissue) and two photons are created that travel approximately in opposite directions. In PET, a coincidence measurement with two detectors replaces the collimation that is used in SPECT. When two photons are measured in coincidence they are supposed to originate from the same decay. In ring PET the detectors are organized in rings around the patient. In analogy to transmission tomography where two points define the origin of the decay (the X-ray source and the detector), in PET the two detectors reduce the possible decay position to those points from where the photons can reach both detectors. Under the assumption that none of the two photons were scattered, this possible decay area is reduced to a tube that is defined by the surfaces of the two detectors that are in coincidence. This region between the two detectors is usually called line of response(LOR) or tube of response. Due to the reduction of possible origin positions, this process is sometimes also called electronic collimation.

Older PET scanner are two dimensional PET scanners. Two dimensional scanners are scanners where coincidences between detectors of different rings are blocked. The blocking is achieved by lead or tungsten rings between the detector rings that reach further into the inside of the scanner. These high density rings are called septa. Two dimensional scanners therefore use electronic as well as conventional collimation. The septa reduce the problem of localization of the decay position to a two-dimensional problem.[3] Since a PET scanner has many detectors ($\approx \mathcal{O}(10^4)$) organized in usually $\approx 15-30$ rings, this leads to a reduction of the very large reconstruction problem to many but much smaller sub problems. This simplification that makes data processing and especially reconstruction much easier, is bought with the reduced number of detected photons that is caused by the collimating septa.

Collimation in 2D PET is therefore achieved by two effects: collimation by the septa and by electronic collimation. Collimation by septa is less effective and in addition reduces strongly the number of coincidences. In most modern scanners septa are therefore removed and the scanners are working in 3D mode. This causes a very strong increase in the complexity: Much more detector-detector combinations generate coincidences and the idealizing concept of lines of responses becomes less correct in view of the strong increase of scatter. In modern scanners therefore the measured data is re-organized and simplified before reconstruction. This simplification reduces the accuracy of the reconstructed images. The main problem, however, is caused by the

[3] In real 2D PET scanners usually coincidence between neighboring rings are allowed, but larger ring differences are blocked. The problem is therefore quasi two dimensional.

increased scatter. Due to the complex and inhomogeneous patient a correct treatment of the scatter is very difficult.

The measured data (the detected coincidences) depends in very good approximation linearly on the (unknown) emission density. When the emission density is discretized, it is therefore possible to describe the map from emission density to measured data by a matrix. This matrix is called system matrix and it is in general ill-conditioned, not quadratic, and very large (around $10^6 \times 10^8$ matrix elements for a full 3D system). When scatter is neglected this matrix is quite sparse. The storage of the matrix including scatter is impossible due to its size. The storage of the scatter-free matrix is difficult [Johnson, 1997, Kehren, 2001] and only possible when symmetries are used that are not present in the matrix including scatter. The elements of the scatter-free matrix are therefore often calculated on-the-fly when they are needed. The calculation of the elements of the full matrix (including scatter) is however more time consuming and many non-zero elements exist. The images are reconstructed iteratively by first starting with an image guess (like a uniform image). The system matrix can then be used to calculate hypothetical measured data. This operation is called projection. This hypothetical data is compared to the real measured data and (using the information of the system matrix) a better guess of the emission density is calculated (and so on). This process is called back-projection.

Due to the size of the matrix a direct incorporation of scatter in the reconstruction is not possible. Therefore the scatter-free matrix is used in the iterative reconstruction process. Scatter can be estimated by analytical methods [Bailey and Meikle, 1994] often using further information like the energy of the detected photons [Grootoonk et al., 1996]. Analytical scatter calculation is very difficult in inhomogeneous media (like the patient) and is therefore only very approximate. Monte Carlo(MC) simulations are better suited to calculate the scatter contribution, because they can also be used in inhomogeneous environments and because of the underlying inherently probabilistic quantum mechanical physics. Unfortunately, these simulations are slow and it is therefore important how to include these simulations in the reconstruction process.

There are several possibilities to include Monte Carlo simulations into the reconstruction algorithms. The least difficult but also least accurate way is to reconstruct an approximate estimate of the emission density without considering scatter and to simulate the data that would be obtained by this emission density. This data can then be used to correct the measured data [Levin et al., 1995]. A more advanced method simulates scatter in the projector of the iterative algorithm (which is straight forward but time consuming), but use simply the scatter-free matrix elements in the back-projector [Ollinger, 1996, Watson, 2000, Beekman et al., 2002, Werling et al., 2002].

This approach will be called dual matrix approach. The incorporation of scatter into the back-projector is problematic, because it requires the storage of the matrix.

The reconstruction of images using an algorithm with the full matrix including scatter in the projector and back-projector for a human sized PET scanner is not possible due to the size of the matrix. For similar (less storage demanding problems) this is however possible. The usage of this full matrix approach for problems like small animal PET imaging (ignoring scatter in the animal, but correctly simulating the geometry of the scanner and detector scatter)[Rafecas et al., 2003, 2004b,a], small animal PET imaging with simplified animal and low number of detectors [Shoukouhi et al., 2004, Shoukouhi, 2005] or human SPECT imaging [Lazaro et al., 2004b, Buvat et al., 2003, Lazaro et al., 2004a, 2005] was shown recently.

1.1. Motivation and thesis outline

The usage of a Monte Carlo based system matrix including patient scatter in reconstruction (full matrix approach) would solve many problems that are present in PET image reconstruction, because many effects can be included into the simulations that are otherwise difficult or impossible to handle. The system matrix describes the map from the emission density to the measured data. In modern scanners, the emission density is usually described by roughly $\mathcal{O}(10^6)$ voxels, and the scanner consists of $\mathcal{O}(10^4)$ detectors, allowing roughly $\mathcal{O}(10^7)$ detector-detector combinations (LORs). The system matrix of such a scanner therefore comprises roughly $\mathcal{O}(10^{13})$ elements. This large number of elements is very problematic to store and to calculate.

Current reconstruction and scatter correction methods for human PET systems therefore do not make use of a stored full matrix. Due to increasing computer speed and storage capacities it is however worthwhile to think about storage methods and the behavior and performance of algorithms that use such a matrix. This thesis deals with these two topics.

After the introduction of PET physics, the process of image reconstruction is explained. Thereupon, the implemented Monte Carlo code is presented. This code is needed for the simulations and is designed for fast system matrix calculations. A compression scheme is introduced that allows a sufficiently efficient compression of the matrix. The choice of the compression method is motivated by matrix properties and by computational constraints.

In the first part of the evaluation section the simulated scanners and phantoms are presented and measures for the quantification of the quality of the compressed matrices

are introduced. It is shown that the code simulates correctly by comparing simulation results with Geant4 simulations. Then the compression scheme is evaluated. Due to the Monte Carlo based evaluation it was possible to verify the patient scatter aspect of the reconstruction problem and to use a single ring scanner simulation as a proof-of-concept. The reduced number of detectors allowed the comparison of compressed matrices to full matrices directly (otherwise not possible due to memory limitations). Apart from the direct comparison of the matrices, reconstructed images using compressed matrices are compared to reconstructed images using the full matrix or dual matrix approach. Finally, noise propagation during iterative reconstruction algorithms using the dual matrix and the full matrix approach is simulated and compared. This completes the proof of principle to use a full Monte Carlo generated system matrix for reconstruction and shows that the usage of such a matrix for 3D scanners is promising to increase the signal to noise ratio and to allow the reconstruction of more quantitative images.

2. Physics of PET

A PET examination starts with the injection of a radioactive tracer into the patient. The tracer is then metabolized and the density of radionuclides evolves with the elapsed time. The patient is placed in the scanner and during one time interval (static image acquisition) or several time intervals (dynamic image acquisition[1]) the escaping photons (and especially photons in coincidence) are detected. The collected information is then used to form images of the estimated emission density at one or several time points. The physical processes that are responsible for the image formation therefore include the decay of a radionuclide, the annihilation of the positron and the creation of photons, possible interactions in the patient, and finally the detection of the photons.

2.1. Positron emission and annihilation

During the decay of the radionuclide that is incorporated in the tracer molecule a positron and a neutrino are created. The neutrino leaves the patient, but the positron usually travels some distance (the positron range), losing most of its kinetic energy by causing ionization and excitation, before it annihilates with an electron of the surrounding tissue. Two photons ($E \geq 0.511\,\mathrm{keV}$) are created that travel approximately in opposite directions. The deviation from 180° is caused by the residual momentum of the positron and electron and is called non-collinearity. The mean expected non-collinearity (FWHM[2] around 0.6°, [Jan et al., 2004]) as well as the positron range (see Table 2.1, [Haber et al., 1990, Levin and Hoffman, 1999, Harrison et al., 1999, Sanchez-Crespo et al., 2004]) depend on the emitting nucleus and on the surrounding tissue, but are both rather small. In Table 2.1 FWHM and FWTM are stated to emphasize the non Gaussian character of the positron range. To a small extent (around 0.5 %, depending on the energy of the positron and the material) it is possible that

[1] In modern scanners often the coincidences are recorded as single events with time tags (list mode). In this way it is later possible to assign the coincidences to different time intervals (frames). This process is called framing. It is therefore possible to choose the time intervals after the measurement.

[2] FWHM=full width half maximum, FWTM=full width third maximum

nucleus	maximal energy	FWHM	FWTM
F-18	634 keV	0.19 mm	0.91 mm
C-11	960 keV	0.28 mm	1.70 mm
N-13	1198 keV	0.33 mm	2.12 mm
O-15	1732 keV	0.41 mm	3.10 mm

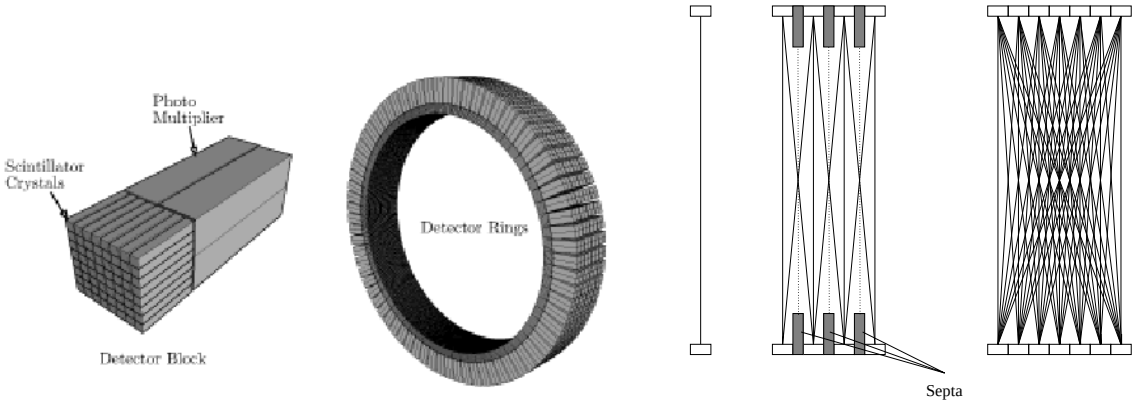
Table 2.1.: Maximal positron energy [Bushberg et al., 2002] and positron range in soft tissue for different radio-nuclei [Sanchez-Crespo et al., 2004].

more than two photons of lower energy are created [Harpen, 2004]. In this case the electron and the positron form a molecule-like system (positronium) before they annihilate. Since the probability for this effect is very small and the energy of the photons usually is below the energy threshold of the detectors, this effect is to be neglected. In order to obtain information about the location of the emission, detectors are located outside the object (the patient).

2.2. The detection system

A ring PET scanner usually consists of several rings of scintillator crystals (for example 24 rings with 384 detectors). Common crystals are BGO ($Bi_4Ge_3O_{12}$), LSO (Lu_2SiO_5: Ce), and GSO (Gd_2SiO_5: Ce) [Knoll, 2000, Humm et al., 2003]. In general several crystals are organized in detector blocks (for example 8×8 crystals per block, in this case the scanner would consist of 3 block detector rings) which are connected to the same number or often a smaller number (like 2×2) of photo multiplier tubes (PMT) that amplify the signal. When a photon leaves the patient and hits a crystals it is converted to lower energy photons which in turn are amplified by the photo multiplier tubes. When a smaller number of PMT is used for the readout, logic circuits like Anger logic circuits calculate the most likely position where the original high energy photon entered the crystals. Often neighboring scintillator blocks are combined to larger systems (called buckets), to simplify electronics. Intra-bucket coincidences are then impossible to detect.

In recent small animal PET scanners avalanche photo diodes(APD) are used [Nuyts, 2000], which, in contrast to PMTs, are much less affected by magnetic fields and therefore good candidates for multi modal PET/MRI scanners. Even more recent research has been pursued in using Geiger mode avalanche photo diodes which are basically grids of very small APDs run in Geiger mode (each small cell APD being



(a) Example for a scanner [Wikipedia, PET] (here ECAT EXACT HR+).

(b) Single ring, 2D scanner, and 3D scanner. Possible detector-detector combinations (LORs).

Figure 2.1.: Ring PET scanner.

therefore binary) and directly detect the high energy photons. While the energy in conventional crystal/PMT and crystal/APD is proportional to the current (or at least derivable from the current) in Geiger mode APD the energy is derived from the number of activated small APDs in a detector unit.

Independent of the kind of detector, in PET the detected events are either stored in histograms (histogram mode) or lists (list mode). While in the first approach the count number of a LOR is incremented when the appropriate event is detected, the latter stores each event separately. It is always possible to generate histograms from list mode data, but not vice versa, because when the events are binned, information like the energy and time of the events gets lost.

All detection systems have limits to the rate at which events may be processed. The electronics or also intrinsic detector characteristics like crystal afterglow [Humm et al., 2003, Knoll, 2000] might be the limiting factor by having a finite maximum rate. In PET, usually, pulse pile-up is the main reason for this so called dead time. This means that several coincidences occur, but they cannot be detected as single events, because they occur so close together. Therefore, less events are detected than truly happen. The corresponding loss is called dead time loss. At high count-rates such losses can become very significant.

As mentioned before, there exist basically two kinds of ring PET scanners, 2D and 3D scanners. Ideal 2D scanners consist of one detection ring observing only one slice of the patient. The set of all LORs of this ideal 2D system is called sinogram. Real 2D scanners consist of several detection rings, each ring shielded by so called septa from

the other rings (Fig. 2.1(b)). These septa are made of high-density and well absorbing material such as tungsten. Usually scanners in 2D mode also allow coincidences between opposing detectors in adjacent rings. Two non-transversal sinograms (oblique sinograms), are usually averaged to form an artificial transversal sinogram. Nowadays, many scanners are 2D/3D or dedicated 3D scanners. These scanner can either retract the septa or do not possess septa at all. The advantage of the scanner running in 3D mode is the increased sensitivity. While many photons get absorbed in the septa of the 2D scanner, these photons can reach the detectors in 3D scanners. The disadvantage is the increase in randoms, and more severely the increase in scattered photons.

2.3. Geometry and interactions in the patient

In ring PET (that is considered solely in this thesis) the patient is surrounded by detectors that work in coincidence mode. This means when two photons are detected

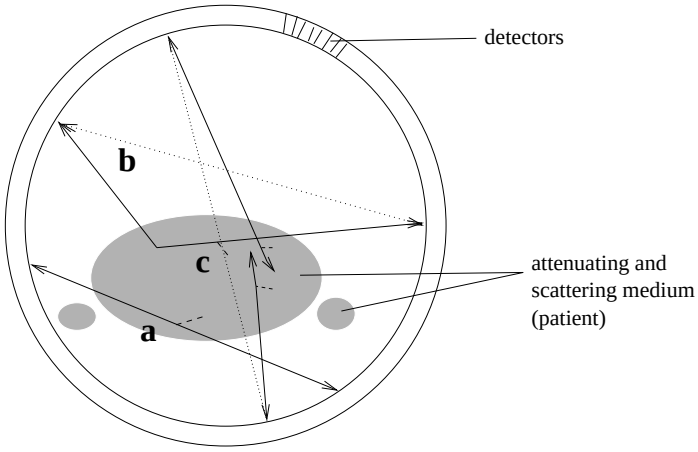


Figure 2.2.: The principle of PET: a tracer accumulates in certain regions of the body, the radionuclide emits a positron which annihilates and two photons are created. A true direct event (**a**), a scattered event (**b**), and a random event (**c**) are shown.

by different detectors within a coincidence time window, they are considered to originate from the same positron. The smaller this time window the more valid is this proposition. However, due to travel time differences, even two photons originating from the same annihilation process usually reach the detectors at different times. The coincidence time window should be chosen in such a way that all true coincidences can be detected. For human PET scanners usually coincidence windows 2τ of at least

interaction	mass attenuation coefficient $\mu/\rho(\mathrm{cm}^2/\mathrm{g})$
Compton effect	9.58×10^{-2}
Rayleigh scattering	2.15×10^{-4}
Photo effect	1.78×10^{-5}

Table 2.2.: Mass attenuation coefficients for photons in H_2O at 0.511 keV [NIST]. Not only in water, but in almost all human tissue the Compton effect is by far the most likely interaction at photon energies $200 - 1000$ keV that are relevant for PET [Kinahan et al., 2003].

a few nanoseconds are required. When both photons have not been scattered, the position where the positron was annihilated must lay on the line (or better "tube") that is defined by the position of the two detectors (see Fig. 2.2). This line is called line of response (LOR). When the photons originate from the same decay process, the coincidence is called a true coincidence. True coincidences comprise scattered coincidences (at least one of the photons is scattered) and not scattered coincidences or direct coincidences[3]. Usually the number of measured scattered coincidences is labeled S while the number of coincidences of not scattered photons is labeled T and called the trues. The origin of the two photons of a scattered coincidence is not located on the LOR. Scattered events are therefore unwanted. The interactions that are responsible for the scattering of the photons (see Table 2.2 and section 4.1.2) are mostly inelastic (Compton effect) and to a small extent elastic (Rayleigh scattering).

In addition it is also possible that a random coincidence is detected. A random coincidence is caused by two positrons and is not a true coincidence. This kind of event can occur when the time difference between two positron annihilations is so small that two of the four photons that originate from different positrons can be detected within the coincidence time window. The two other photons from this annihilations are either absorbed (Photo effect), their energy drops below the energy threshold of the detectors(Compton effect) or the photons are simply not detected (either due to geometrical reasons or due to non-ideal detector efficiency). Three (or even four) simultaneously detected events are usually discarded. The number of detected random coincidences is called randoms R. *Single events* (one detector detects a photon, the other not) can also be recorded and can give insight into the frequency of random events (see 2.3.3).

[3]In some publications true coincidences are defined to be a coincidence from not scattered photons only. Direct coincidences are sometimes (but not in this thesis) defined to be coincidences between detectors of the same ring (in 3D PET).

Time information is not used in conventional PET. PET using time information to further localize the annihilation position is called time of flight PET (TOF-PET). The detector physics and electronics limits the time resolution to around $\Delta t = 0.2 - 1.2\,\mathrm{ns}$ [Moses and Derenzo, 1999, Defrise et al., 2005, Conti et al., 2005, Vandenberghe et al., 2006] and therefore the possible spatial resolution along the line of response to roughly $\Delta x = c\Delta t/(2n) \approx 3 - 15\,\mathrm{cm}$ (c/n=effective speed of light between the two detectors, $n(511\,\mathrm{keV}) \approx 1.0$). The sole usage of time information to determine the annihilation position is therefore not practicable. Recently, however, Conti et al. [2005] showed that the usage of time of flight information can be used to improve considerably the signal to noise ratio of the reconstructed images for a modern PET scanner for humans. Other recent results show that TOF PET can also be used to reduce the angular sampling while compromising only little the resolution [Vandenberghe et al., 2006]. Time of flight does not eliminate scattered coincidences.

2.3.1. Scattered coincidences

While direct coincidences are best suited for image reconstruction, detected and uncorrected scattered coincidences degrade the image. The degradation has two reasons. Firstly, scattered photon pairs give less accurate information about the origin of the annihilation position. While in the case of true unscattered coincidences the positron must have been annihilated somewhere on the LOR, in the case of the scattered events the origin can also lie in the vicinity of the LOR or even further away depending on the scatter angle (limited by the energy threshold of the detectors) and the size of the scanner. This leads to a reduced effective resolution of the scanner. Secondly, all reconstruction algorithms so far are based on some approximate and often simplified scatter treatment. This wrong scatter modeling might not only lead to resolution reduction but also to wrong activity distributions. Since the scatter is patient dependent, correct scatter treatment is very difficult.

The ratio

$$f_\mathrm{S} = \frac{\#\text{ of scattered coincidences}}{\#\text{ of all true coincidences (including scatter)}} \tag{2.1}$$

is called scatter fraction. A small scatter fraction is desirable. The scatter fraction mainly depends on the scanner (around $10 - 20\%$ for 2D scanners and around $40+\,\%$ for 3D scanners according to Adam et al. [1999] and Lodge et al. [2006]) and also on the patient, because for a larger patient the ratio of scattered photons to not scattered photons increases. In human PET scanners most photons are scattered in the patient,

but some photons are scattered in the gantry (the surrounding of the patient opening i.e. the detection system + supporting structure) [Adam et al., 1999].

The scatter fraction can be reduced by choosing a high lower energy threshold for the detectors. The lower energy threshold for the detectors is usually set between 250 and 450 keV. Due to finite energy resolution of the detector, the energy threshold cannot be set arbitrarily close to 511 keV. This would reduce considerably the number of detected true direct coincidences.

2.3.2. Attenuation

All the aforementioned interactions, predominantly Compton effect, but also Photo effect and Rayleigh scattering lead to a reduced number of direct counts. This effect is called attenuation. Attenuation has to be considered in the process of reconstructing the images, because otherwise the emission density inside the patient is underestimated. While there is an extra correction in conventional reconstruction (see also section 2.4), in the presented Monte Carlo based reconstruction this is considered implicitly in the Monte Carlo system matrix.

In contrast to SPECT where it is rather difficult to estimate the attenuation caused by the patient, attenuation correction is rather straightforward in PET. Each photon of the two photons that are created at the annihilation point, often named *pink* and *blue* photon for convenience, can interact during their way to the detectors.

$$\begin{aligned} p_{\text{pink}} &= e^{-\int_{x_0}^{x_1} \mathrm{d}x\ \mu(x)} \\ p_{\text{blue}} &= e^{-\int_{x_0}^{x_2} \mathrm{d}x'\ \mu(x')} \qquad \text{with} \quad x' = -x \\ p_{\text{pink}\wedge\text{blue}} &= p_{\text{pink}} \cdot p_{\text{blue}} = e^{-\int_{x_1}^{x_2} \mathrm{d}x\ \mu(x)} \end{aligned} \tag{2.2}$$

The probabilities that the photons will not interact with the patient on their way (from x_0 to x_1 or to x_2) to the detector are p_{pink} and p_{blue} respectively ($\mu(x)$ being the total linear attenuation coefficient at location x). The probability that both photons reach the detectors is $p_{\text{pink}\wedge\text{blue}}$ which equals the probability of a photon starting at one of the two involved detectors and reaching the other detector without interaction. The latter probability can therefore be obtained by a transmission scan and the probability does not depend on the position of the annihilation point on the LOR. It is then either possible to correct the measured counts for LOR i by multiplying them with attenuation correction factor

$$e^{+\int_{x_1(i)}^{x_2(i)} \mathrm{d}x\ \mu(x)}$$

or to consider

$$e^{-\int_{x_1(i)}^{x_2(i)} \mathrm{d}x\ \mu(x)}$$

during iterative reconstruction and incorporate this factor into the model of the scanning system (described by the system matrix).

2.3.3. Random correction

Random coincidences are even less desirable than scattered photons, because they carry no information about the origin of the photon pair. Fortunately, their frequency can be estimated rather straightforwardly.

Random correction can be performed applying two basic approaches [Brasse et al., 2005, Knoll, 2000], either by delayed coincidence correction and by estimating the randoms from single counts. In coincidence measurements, when a signal at time t in

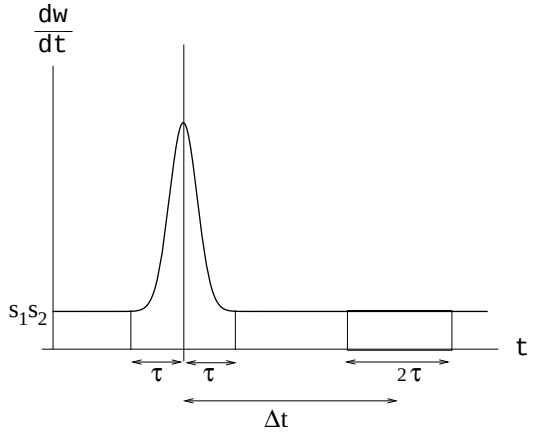


Figure 2.3.: Differential time spectrum. The abscissa t is the time interval length or time delay, the ordinate is the count rate w for an infinitesimal coincidence window dt at time delay t.

one detector is measured and the other detector detects a photon within a coincidence time window $[t-\tau, t+\tau]$, it can be assumed that the detected photons are correlated and come from the same event. Unfortunately, due to the finite time window τ, also photons *not* originating from the same positron, hence random events, are counted. Fig. 2.3 shows the typical differential count rates that can be expected for different time delays t and an infinitesimal coincidence window dt. For large time delays (finite coincidence time window $[t+\Delta t-\tau, t+\Delta t+\tau]$) the detected counts are not correlated

and in this case the random count rate is [Knoll, 2000]

$$\bar{r} = 2\tau s_1 s_2 \,. \tag{2.3}$$

For small time delays t this is still true, but in addition true coincidences are detected. It is therefore possible to subtract $\bar{r}$ from the total count rate for the coincidence time window $[t-\tau, t+\tau]$. This approach is therefore based on the estimate of $\bar{r}$ by the single count rates s_1 and s_2 of the two involved detectors. This approach essentially introduces no additional noise [Brasse et al., 2005], but can introduce bias, if some of the factors in (2.3) are estimated wrongly. The second approach on the other hand introduces additional noise, but does not introduce a bias [Brasse et al., 2005]. Instead of measuring the single rates s_1 and s_2 the mean random rate $\bar{r}$ is directly measured using this strongly delayed window (see Fig. 2.3) about Δt. For both methods different improvement schemes, like "smoothed delays method" or other singles based approaches exist [Brasse et al., 2005].

2.4. Acquisition

After injection of the tracer and after some chosen time delay for perfusion, the photons emitting from the patient can be measured. This process is called emission scan.

2.4.1. PET

In conventional PET in addition a blank and a transmission scan are used to correct for the attenuation in the patient (see section 2.3.2). Both scans are usually obtained by a rod source that rotates around the patient (transmission scan) or in the empty scanner (blank scan). Ideally the transmitting photons should also be in the energy range around 511 keV. For this reason usually either ^{137}Cs (γ-emitter, 662 keV) or ^{68}Ge($\tau_{1/2} = 288$ d) is used. The latter source decays to ^{68}Ga which primarily decays by positron emission [Bushberg et al., 2002]. Using the information obtained by both scans it is possible to calculate the attenuation caused by the patient. The blank scan can be used to correct for differently responding detectors (like different amplifier gains) [Defrise et al., 1991]. This is called normalization. The transmission scan can be performed before or after injection, but the latter adds noise from the emitting photons.

Apart from the mentioned needed corrections, the data has to be corrected for the exponential decay of activity. Since the scan time and the half life of the radio-tracer is

of similar order, the reduction of the activity during the scan cannot be neglected. This is especially important for dynamic studies. Dead time loss should be also corrected for [Mazoyer et al., 1985].

2.4.2. PET/CT

A PET/CT scanner is a combined system of a PET and a CT scanner [Beyer et al., 2001, 2002]. The obtained images combine the advantage of showing exact and very fine morphological information (CT-image) as well as information about tracer uptake (PET-image) in the same coordinate system. Especially in the case when highly specific tracers are used, the additional information from the CT image can be crucial to associate the sometimes very localized higher uptake to some morphological structure. While it would be ideal to have both scanners at the same position, nowadays both machines are placed next to each other due to hardware limitations. The patient first passes the CT scanner and is then moved into the PET scanner. While a PET scan can take from a couple of minutes to up to an hour per bed position, a similar CT scan can be easily performed in less than 15 s. The shorter the PET scan the noisier are the obtained images due to bad statistics of the measured sinograms.

In PET/CT the transmission scans in general are replaced by the CT scan [Beyer et al., 2004]. The advantage of the CT scan is the great increase in accuracy of the attenuation map and the fast acquisition. One disadvantage is the different energy of the photons (around $40 - 140$ keV [Kinahan et al., 1998, 2003] where Photo effect and Compton effect are both important). The attenuation has to be calculated for photons of 511 keV where the Compton effect is dominant. The calculation is therefore not trivial and cannot be exact, since the exact atomic numbers Z and mass number A of the involved atoms and their concentration ratio are unknown. Usually linear attenuation coefficients at CT energy are mapped to linear attenuation coefficients at PET energy by using simple piece-wise linear functions based on some assumptions on the tissue [Nuyts and Stroobants, 2005, Kinahan et al., 2003]. This problem occurs also in MC simulation. In section 4.1.1 it is shown how this problem is approached. A perhaps more severe disadvantage (being an advantage at the same time) is the fast transmission scan used for attenuation correction. Any shift of the patient during the PET acquisition with respect to the CT acquisition can then lead to artifacts induced by wrongly attenuation corrected LORs. This especially applies for lung motion, which is averaged in PET, but the obtained CT image and CT-based attenuation map might only show one snapshot of the motion.

3. Image Reconstruction

The process of calculating the emission density given the detected coincidences is called image reconstruction. Numerous algorithms exist to perform this task [Qi and Leahy, 2006]. The algorithms can be classified into two groups of algorithms: iterative reconstruction algorithms and algorithms that lead to the emission density by direct inversion (often called analytic reconstruction algorithms). The latter are usually based on the Radon transform.

3.1. The Radon transform and its inverse

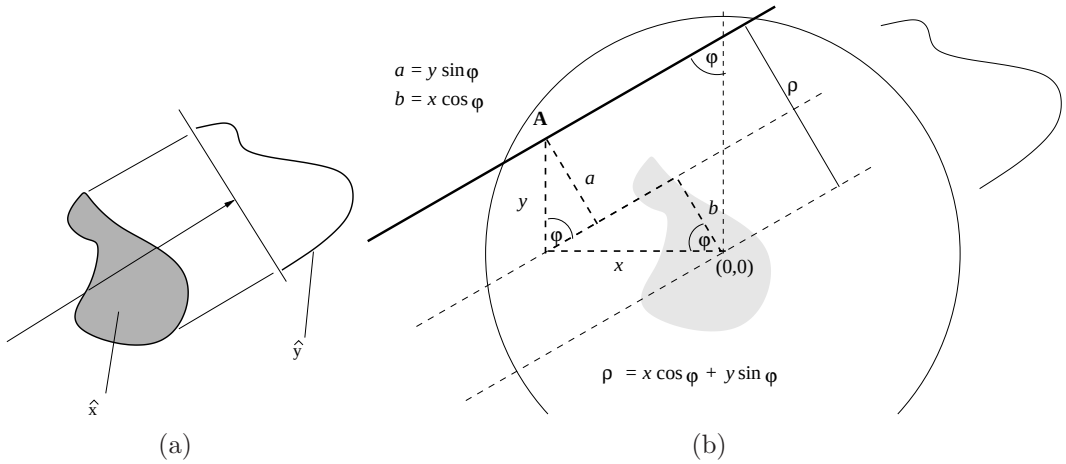


Figure 3.1.: The Radon transform

In all ray based tomographic imaging like computed tomography, single photon emission computed tomography and also PET, the images are obtained indirectly by measuring projections. A projection $\hat{y}_\varphi$ is obtained by integrating the density of an object $\hat{x}$ (emission density in PET, electron density in CT) along lines with angle φ, reducing the two-dimensional image to one dimension (see Fig. 3.1(a) and Fig. 3.2). The set of all projections for angles $\varphi \in [0, \pi[$ can be described by a two-dimensional

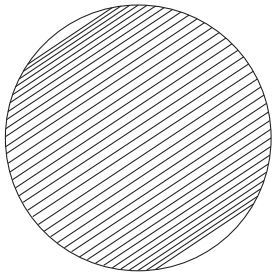


Figure 3.2.: Example: LORs that belong to a projection in ring PET.

function $\hat{y}(\varphi, \rho) \equiv \hat{y}_\varphi(\rho)$. This function is obtained by applying the Radon transform $\mathcal{R}$ [Toft, 1996, Kak and Slaney, 1999] on the object density function[1]

$$\hat{y}(\varphi, \rho) \equiv \mathcal{R}(\hat{x}(x, y)) \equiv \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathrm{d}x\mathrm{d}y \, \delta(\rho - x\cos\varphi - y\sin\varphi)\hat{x}(x, y)\,. \tag{3.1}$$

In PET (and also CT) $\hat{y}(\varphi, \rho)$ is called sinogram. The sinogram of a point like emission density at (x_0, y_0) is a sinusoidal function

$$\rho - x_0\cos\varphi - y_0\sin\varphi \stackrel{!}{=} 0 \Longrightarrow \rho = x_0\cos\varphi + y_0\sin\varphi. \tag{3.2}$$

This is the reason for the name "sinogram".

Using the Fourier transform[2] ($\mathcal{FT}$) and the inverse Fourier transform($\mathcal{FT}^{-1}$), it is possible to relate the sinogram and the emission density in an elegant manner. The two-dimensional Fourier transform of the emission density is

$$\hat{X}(k_\mathrm{x}, k_\mathrm{y}) \equiv \mathcal{FT}_{(x,y)\to(k_\mathrm{x},k_\mathrm{y})}\hat{x}(x, y) \equiv \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \hat{x}(x, y)e^{-i2\pi(k_\mathrm{x}x+k_\mathrm{y}y)}\mathrm{d}x\mathrm{d}y\,. \tag{3.3}$$

With the introduction of polar frequency parameters $k_\mathrm{x} = \nu\cos\varphi$, $k_\mathrm{y} = \nu\sin\varphi$ equa-

[1] Here $\delta(*)$ is the delta distribution.

[2] In the following derivation the Fourier transform based on ordinary frequency ν instead on the circular frequency $\omega = 2\pi\nu$ is used. In this way the transformation is unitary without a factor $1/\sqrt{2\pi}$. This form is often used in the field of signal processing.

tion (3.3) becomes

$$\begin{aligned}
\hat{X}(\nu\cos\varphi,\nu\sin\varphi) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\hat{x}(x,y)e^{-i2\pi(x\nu\cos\varphi+y\nu\sin\varphi)}\mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\hat{x}(x,y)\delta(\rho-x\cos\varphi-y\sin\varphi)e^{-i2\pi\nu\rho}\mathrm{d}x\mathrm{d}y\right)\mathrm{d}\rho \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\hat{x}(x,y)\delta(\rho-x\cos\varphi-y\sin\varphi)\mathrm{d}x\mathrm{d}y\right)e^{-i2\pi\nu\rho}\mathrm{d}\rho \\
&= \int_{-\infty}^{\infty}\hat{y}(\varphi,\rho)e^{-i2\pi\nu\rho}\mathrm{d}\rho \qquad (3.4)\\
&\equiv \mathcal{FT}_{\rho\to\nu}\hat{y}(\varphi,\rho)\equiv\hat{Y}(\varphi,\nu)\,.
\end{aligned}$$

The two-dimensional Fourier transform of the emission density equals[3] the one dimensional Fourier transform (with respect to ρ) of the sinogram. This result is called Fourier slice theorem. It is therefore possible to calculate $\hat{x}(x,y)$ from $\hat{y}(\varphi,\rho)$ and vice versa.

The most used inversion scheme, however, is filtered back-projection(FBP). This scheme is derived by using the inverse Fourier transform of the emission density

$$\hat{x}(x,y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\hat{X}(x,y)e^{i2\pi(k_{\mathrm{x}}x+k_{\mathrm{y}}y)}\mathrm{d}k_{\mathrm{x}}\mathrm{d}k_{\mathrm{y}} \qquad (3.5)$$

and introducing polar coordinates

$$\begin{aligned}
\hat{x}(x,y) &= \int_0^{2\pi}\int_0^{\infty}\nu\hat{X}(\nu\cos\varphi,\nu\sin\varphi)e^{i2\pi\nu(x\cos\varphi+y\sin\varphi)}\mathrm{d}\nu\mathrm{d}\varphi \\
&= \int_0^{\pi}\int_{-\infty}^{\infty}|\nu|\hat{X}(\nu\cos\varphi,\nu\sin\varphi)e^{i2\pi\nu(x\cos\varphi+y\sin\varphi)}\mathrm{d}\nu\mathrm{d}\varphi \\
&\overset{(3.4)}{=} \int_0^{\pi}\underbrace{\int_{-\infty}^{\infty}|\nu|\left(\int_{-\infty}^{\infty}\hat{y}(\varphi,\tilde{\rho})e^{-i2\pi\nu\tilde{\rho}}\mathrm{d}\tilde{\rho}\right)e^{i2\pi\nu\overbrace{(x\cos\varphi+y\sin\varphi)}^{\rho}}\mathrm{d}\nu}_{z(\varphi,\rho)\equiv\mathcal{FT}^{-1}_{\nu\to\rho}\left(|\nu|\mathcal{FT}_{\tilde{\rho}\to\nu}(\hat{y}(\varphi,\tilde{\rho}))\right)}\mathrm{d}\varphi\,. \qquad (3.6)
\end{aligned}$$

The emission density is therefore calculated by filtering the projections according to

[3]Exactly when the signal processing version of the Fourier transform is used, but up to a normalization factor when the angular frequency version is used.

ρ with a ramp filter and then this filtered sinogram $z(\varphi, \rho)$ is back-projected.

$$\begin{aligned}\hat{x}(x,y) &= \int_0^{\pi} z(\varphi, x\cos\varphi + y\sin\varphi)\mathrm{d}\varphi \\ &= \int_0^{\pi}\int_{-\infty}^{\infty} z(\varphi,\rho)\delta(\rho - x\cos\varphi - y\sin\varphi)\mathrm{d}\rho\mathrm{d}\varphi \end{aligned} \tag{3.7}$$

The last equation (3.7) has a similar structure like (3.1) and is therefore named back-projection. The ramp filter is a high pass filter. Although being theoretically correct, this strong high pass filter is not ideal, because the measured sinograms are usually quite noisy. In real application deviations of the ramp filter like Shepp-Logan, Hann, Hamming, or Butterworth filters [Kehren, 2001] are used that mimic a ramp filter for low frequencies but reduce high frequency parts of the projections.

Because of the finite number of detectors the sinogram of a PET scanner is always discrete. There exist discrete versions of the Fourier slice theorem and FBP [Toft, 1996, Kak and Slaney, 1999] which will not be introduced in detail, because both algorithms are not used in this dissertation. A discrete projection(see Fig. 3.2) is the set of all LORs with the same angle φ. One value of such a discrete projection, the number of detected counts of a LOR is called bin [Alenius, 1999]. It can be seen that the LORs are not equally spaced, especially close to the gantry [Fahey, 2002]. The projections used in FBP must therefore be corrected for this effect. This correction is called arc correction.

3.2. Iterative algorithms

The above introduced approaches based on the inversion of the Radon transform are based on geometrical considerations and assume that only true direct coincidences are detected. They therefore lack the possibility to correct for scatter during reconstruction. It is possible to formulate the problem to be the solution of a set of linear equations that include all relevant physics. For this purpose the phantom/patient is discretized. The patient is decomposed into volume elements/basis functions $\beta(\mathbf{r} - \mathbf{r}_i)$ located at grid node positions $\mathbf{r}_i$ and the activity is represented by a linear combination of such basis functions. The decomposition is usually a regular grid with N_V voxels (in this case the basis functions are B_0-spline basis functions), but also more general basis functions are possible [Lewitt, 1992, Fessler, 2004]. The activity is approximated

by

$$\text{activity}(\mathbf{r}) = \sum_{i=1}^{N_\text{V}} \hat{x}_i \beta(\mathbf{r} - \mathbf{r}_i) \tag{3.8}$$

The activity can therefore be represented by a N_V-dimensional vector $\hat{\mathbf{x}}$.

3.2.1. The system matrix

Because the emitted photons are not influencing each other, the system response of two different voxels filled with activity is independent. The system response of several voxels (coincidences, the sinogram) is therefore a linear combination of the system response of the single voxels (coincidences, unit sinogram). In other words, there is a linear relation between the emission density $\hat{\mathbf{x}}$ and the sinogram $\hat{\mathbf{y}}$

$$\hat{\mathbf{y}} = M\hat{\mathbf{x}}\,. \tag{3.9}$$

This relation is very general and the system matrix M describing the map from activity to sinogram can also include scatter and attenuation.

The goal is the calculation of $\hat{\mathbf{x}}$ for a measured sinogram $\hat{\mathbf{y}}$ and a given model of the scanner described by the system matrix M. Usually M is not a quadratic matrix and, more severely, equation (3.9) is inherently ill posed like all inversion problems in tomographic imaging. In addition, especially for 3D scanners, the matrix M is very large (around $10^6 \times 10^8$ elements or even more). While ideally the system matrix should include all physics like attenuation and scatter in the patient, randoms, scanner geometry, detectors (and normalization if not done before) and electronics, due to its immense size it is not possible to include all this. Usually normalization and attenuation, the scanner geometry and some kind of detector modeling are included into the matrix. Randoms are not included (see section 2.3.3). Most difficult is the inclusion of patient scatter into the matrix, because this requires a recalculation of the matrix for each patient, and in contrast to scanner dependent contributions it is not possible to reduce the problem due to scanner symmetry, because the patient is asymmetric. In contrast to small animal PET the scatter contribution of the patient cannot be neglected in human 3D PET. Instead of addressing the problem directly by including scatter into the matrix, scatter is treated conventionally during the reconstruction or even guessed before the reconstruction which always involves some approximations or assumptions (see section 3.4).

3.2.2. The objective function

Since in real applications the matrix M is not invertible, an approximate solution of (3.9) must be calculated differently. While minimizing

$$|\hat{\mathbf{y}} - M\hat{\mathbf{x}}|^2 \tag{3.10}$$

gives an approximate solution, especially for low statistics measurements the minimization of

$$F(\bar{\mathbf{x}}) \equiv P(\hat{\mathbf{y}}|\bar{\mathbf{x}}) = \prod_{j=1}^{N_\mathrm{L}} e^{-\bar{y}_j} \frac{\bar{y}_j^{\hat{y}_j}}{\hat{y}_j!} \qquad , \quad \bar{\mathbf{y}} = M\bar{\mathbf{x}} \tag{3.11}$$

results in better images. $P(\hat{\mathbf{y}}|\bar{\mathbf{x}})$ is the likelihood that the a sinogram $\hat{\mathbf{y}}$ is measured given the mean emission density $\bar{\mathbf{x}}$ and assuming that each LOR $\hat{y}_j$ varies according to Poissonian statistics. $\bar{\mathbf{y}}$ is the expected mean value for $\hat{\mathbf{y}}$. The goal is to find the unknown mean $\bar{\mathbf{x}}$ of the activity, given the measured sinogram $\hat{\mathbf{y}}$. The mean emission density $\bar{\mathbf{x}}$ that maximizes (3.11) represents the most likely emission density. Since logarithmizing is a monotone transformation and therefore preserving the maximum, it is possible to maximize the log likelihood

$$\begin{aligned} \tilde{f}(\bar{\mathbf{x}}) \equiv \log(F(\bar{\mathbf{x}})) &= \sum_{j=1}^{N_\mathrm{L}} \left(-\bar{y}_j + \hat{y}_j \log \bar{y}_j - \log\left(\hat{y}_j!\right)\right) \\ &= \sum_{j=1}^{N_\mathrm{L}} \left(-\sum_{i=1}^{N_\mathrm{V}} m_{ji}\bar{x}_i + \hat{y}_j \log\left(\sum_{i=1}^{N_\mathrm{V}} m_{ji}\bar{x}_i\right) - \underbrace{\left[\log\left(\hat{y}_j!\right)\right]}_{\text{constant}} \right) \end{aligned} \tag{3.12}$$

instead of (3.11). This is advantageous, because every product is reduced to a sum. The last addend can be ignored, because it is constant. The Hessian of the function $f = \tilde{f} + \sum_j^{N_\mathrm{L}} \log\left(\hat{y}_j!\right)$ is

$$\frac{\partial^2 f}{\partial \bar{x}_k \partial \bar{x}_l} = -\sum_j^{N_\mathrm{L}} \hat{y}_j \frac{m_{jk} m_{jl}}{\left(\sum_{i=1}^{N_\mathrm{V}} m_{ji}\bar{x}_i\right)^2} \tag{3.13}$$

and therefore

$$\sum_{k,l=1}^{N_\mathrm{V}} \bar{x}_k \frac{\partial^2 f}{\partial \bar{x}_k \partial \bar{x}_l} \bar{x}_l = -\sum_{j=1}^{N_\mathrm{L}} \hat{y}_j \left(\sum_{k=1}^{N_\mathrm{V}} \frac{m_{jk}\bar{x}_k}{\sum_{i=1}^{N_\mathrm{V}} m_{ji}\bar{x}_i} \right)^2 \leq 0 \qquad \forall\, x \tag{3.14}$$

negative semi-definite. For all $\hat{y}_j \geq 0$, $\forall m_{jk} > 0$, and $\bar{x}_i > 0$ the inequality (3.14) is even strictly fulfilled[4]. For positive emission density there exists therefore unique maximizer.

3.2.3. Maximum likelihood expectation maximization

Since the emission density is always positive, equation (3.12) has to be maximized subject to the constraint

$$\bar{x}_i \geq 0\,. \tag{3.15}$$

This constraint can be incorporated into (3.12) by the introduction of the Lagrange function L with the Lagrange parameter λ

$$L(\bar{\mathbf{x}}, \boldsymbol{\lambda}) \equiv f(\bar{\mathbf{x}}) - \sum_{i=1}^{N_\mathrm{V}} \lambda_i \bar{x}_i \tag{3.16}$$

and requiring (Nocedal and Wright [1999]) that the Karush-Kuhn-Tucker conditions are satisfied:

$$\nabla_{\bar{x}_k} L(\bar{\mathbf{x}}, \boldsymbol{\lambda}) = \sum_{j=1}^{N_\mathrm{L}} \left(-m_{jk} + \hat{y}_j m_{jk} \frac{1}{\sum_{i=1}^{N_\mathrm{V}} m_{ji} \bar{x}_i} \right) - \lambda_k = 0 \qquad \forall k \tag{3.17}$$

$$\bar{x}_i \geq 0 \quad \forall i \tag{3.18}$$

$$\lambda_i \bar{x}_i = 0 \quad \forall i\,. \tag{3.19}$$

The last constraint (3.19) is also known as the complementary slackness. When multiplying (3.17) with $\bar{x}_k$ and inserting (3.19), it is possible to get rid of the Lagrange parameter λ due to the complementary slackness. This leads to

$$\bar{x}_k \nabla_{x_k} L(\bar{\mathbf{x}}, \boldsymbol{\lambda}) = \sum_{j=1}^{N_\mathrm{L}} \left(-m_{jk} \bar{x}_k + \frac{\hat{y}_j m_{jk} \bar{x}_k}{\sum_{i=1}^{N_\mathrm{V}} m_{ji} \bar{x}_i} \right) \stackrel{!}{=} 0 \tag{3.20}$$

or

$$\bar{x}_k = \frac{\bar{x}_k}{\sum_{j=1}^{N_\mathrm{L}} m_{jk}} \sum_{j=1}^{N_\mathrm{L}} \left(\frac{\hat{y}_j m_{jk}}{\sum_{i=1}^{N_\mathrm{V}} m_{ji} \bar{x}_i} \right)\,. \tag{3.21}$$

[4] When online random subtraction is performed, negative values for $\hat{y}_j$ are possible. In this case usually the sinogram values are set to zero or a different objective function that considers random subtraction is used [Fessler, 2004].

It is always possible [Vardi et al., 1985] to simplify (3.21) by rescaling the matrix and the activity.

$$x_i^{'} = \bar{x}_i \sum_{j=1}^{N_\mathrm{L}} m_{ji} \tag{3.22}$$

$$m_{ij}^{'} = \frac{m_{ij}}{\sum_{j'=1}^{N_\mathrm{L}} m_{j'i}} \tag{3.23}$$

In this way (3.21) simplifies to

$$x_k^{'} = x_k^{'} \sum_{j=1}^{N_\mathrm{L}} \left(\frac{\hat{y}_j m_{jk}^{'}}{\sum_{i=1}^{N_\mathrm{V}} m_{ji}^{'} x_i^{'}} \right) = \left[\mathbf{C}\left(\mathbf{x}^{'}\right) \right]_k \tag{3.24}$$

Equation (3.24) is a fixed point equation. Applying a fixed point algorithm [Johnson and Sofer, 2000] to (3.24) yields

$$\mathbf{x}^{'(\alpha+1)} = \mathbf{C}\left(\mathbf{x}^{'(\alpha)}\right) \tag{3.25}$$

This is the update equation of the ML-EM algorithm applied to emission tomography. Vardi et al. [1985] showed (following the idea of Dempster et al. [1977]) that (3.25) converges to the maximum of (3.11) in the presence of more detectors than voxel. Equation (3.25) is often splitted into two coupled equations

$$\begin{aligned} \mathcal{P}: &\quad \bar{\mathbf{y}}^{(\alpha+1)} = M\mathbf{x}^{('\alpha)} \\ \mathcal{B}: &\quad x_i^{'(\alpha+1)} = x_i^{'(\alpha)} \sum_{j=1}^{N_\mathrm{L}} \left(\frac{\hat{y}_j m_{ji}}{\bar{y}_j^{(\alpha+1)}} \right) \end{aligned} \tag{3.26}$$

with a projector $\mathcal{P}$ and a back-projector $\mathcal{B}$.

There exist a vast number of variants of the ML-EM algorithm and also several other iterative algorithms [Fessler, 2004]. Variants of the ML-EM algorithm are usually introduced to speed up the rather slow performance of the ML-EM algorithm. A very common algorithm is OSEM, ordered subset expectation maximization [Hudson and Larkin, 1994], which is a block iterative not convergent, but much faster, descendant of the ML-EM. This algorithm can often be found in commercial scanners. In block iterative algorithms all pixels are updated using a subset of the measured data at one time. In contrast, in the ML-EM algorithm all voxels are updated at once using all available data. In other algorithms like the row action maximum likelihood

algorithm (RAMLA) [Browne and De Pierro, 1996] or algebraic reconstruction technique (ART) [Herman and Meyer, 1993] all voxels are updated once using one row (with respect to the system matrix) of the data once at a time. These algorithms are called row action algorithms. The opposite are sequential algorithms that update only one pixel using all data at each iteration like SAGE [Fessler, 2004] or coordinate descent.

Usually objective functions should include a regularization term. This can lead to algorithms that are less ill-posed, and might converge faster. There exists a large number of possible regularization terms [Ollinger and Fessler, 1997, Alenius, 1999].

3.3. Reconstruction with 3D scanners

The aforementioned reconstruction methods work well with 2D scanners. Reconstruction using 3D scanners can be performed by producing artificial 2D sinograms (either by single-slice, by multi-slice, or by the more advanced Fourier-slice rebinning which is often used in commercial scanners together with OSEM) [Kehren, 2001] or by directly using the 3D data. The approximate 2D sinograms can be used in the same way as described in the previous sections.

For 3D data, there exist generalizations of Fourier slice reconstruction and of FBP [Toft, 1996], but due to the increased amount of scatter, analytical methods are not preferable. Iterative reconstruction algorithms can be used for 3D PET data without modification, but due to the large amount of data the reconstruction is very time consuming. In general, therefore the 3D data is simplified. The 3D data can be divided into LORs from the same ring (transversal sinograms) or in LORs from different rings (oblique sinograms). Usually, the transversal sinograms are used unmodified, but neighboring oblique sinograms are combined. This reduces the size of the data, but also the resolution [Kehren, 2001] and results in an increase of the detected counts of such combined sinograms. In the same way the neighboring bins of similar angle (of the same oblique or transversal sinogram) can be grouped together. This process is called mashing and also reduces the resolution and the data size [Fahey, 2002].

3.4. Scatter correction

There are different ways to treat the scatter in positron emission tomography. Due to the problem of accurate modeling of the system matrix (see chapter 1), many different approximation schemes [Bergström et al., 1983, Ollinger, 1996, Zaidi, 2000,

Beekman et al., 2002, Zaidi and Koral, 2004] were proposed that can roughly be grouped into four methods [Markiewicz et al., 2004]:

1. pre-correction of the data
2. post-correction of the image
3. incorporation of scatter in the projector (dual matrix)
4. incorporation of scatter in the matrix (full matrix)

In the first method the the sinogram is corrected for scatter and the corrected data is used in the reconstruction algorithm. Different approaches exist to estimate the scatter. It can be estimated by using energy information of the detected photons (energy window based scatter correction) [Grootoonk et al., 1996], by simulating the scatter by means of Monte Carlo methods [Levin et al., 1995] or by (iterative) convolution of the estimated true image with a scatter kernel to obtain scatter data that can be subtracted [Bailey and Meikle, 1994]. The two latter pre-correction methods need therefore a first guess of the activity. All pre-corrected data can be used in principle also in analytical reconstruction like FBP. The post correction methods improve the image that is reconstructed with uncorrected data [Zaidi and Koral, 2004]. This can for example be done by reconstructing an image using the scatter estimate and subtract this obtained scatter image from the image that was obtained for uncorrected data [Lercher and Wienhard, 1994]. The last two scatter correction groups either model the scatter in the projector only (dual matrix) [Ollinger, 1996, Beekman et al., 2002] or incorporate the scatter in the matrix [Rafecas et al., 2004b, Shoukouhi, 2005, Lazaro et al., 2005]. Both methods usually use Monte Carlo simulation and only work together with iterative reconstruction. Monte Carlo based dual and especially full matrix reconstructions are theoretically superior to the other correction methods since they model the scatter based on physical processes in each iteration and do not require a first estimate of the activity distribution. Both methods are further explained in chapter 6.1 and chapter 6.2.

4. A Fast Monte Carlo Code for System Matrix Calculations

The purpose of the Monte Carlo code YaPRA was the simulation of the system matrix. Several Monte Carlo codes for PET like GEANT4 (generic), SimSET (dedicated) [Harrison] exist. While generic codes usually carry an overhead that increases computation time, dedicated codes are trimmed down to become fast. Generic codes are not a good choice for the very time consuming task of system matrix calculation. Existing dedicated codes are used to simulate sinograms for a given extended emission density. System matrix calculation is based on the simulation of sinograms of single voxels. The decision was therefore to develop a new Monte Carlo code that could trace particles in the patient and that was optimized to perform this task.

This codes is inspired by two publications [Haynor et al., 1990, 1991] that explain the variance reduction techniques used in SimSET. Parts of the main photon tracking algorithm (without variance reduction techniques) as well as the pseudo random number generator RANMAR were taken from the Monte Carlo code XVMC for dose calculation used in radiotherapy [Fippel, 1999, 2000].

4.1. Particle tracing

4.1.1. Linear attenuation coefficients

The simulation requires knowledge about the probability of particle interaction at a given location in the patient. This probability depends on the tissue. In medical situations, it is not possible to obtain an exact map of the tissue of the patient. It is therefore necessary to estimate at which position which mean cross section should be used. This approximation usually is based on tomographic images. Since the cross sections for the Compton effect and the Photo effect mostly depend on the electron density of the material, the assumed cross sections are based on density images that can be obtained by CT scanners or by attenuation maps obtained by rotating rod sources (see section 2.4.1). This assignment of density values to cross sections or

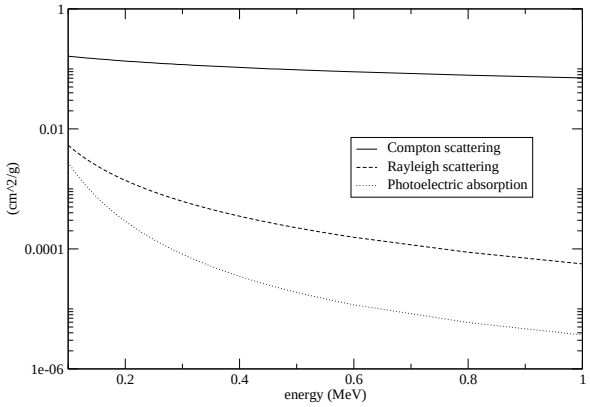


Figure 4.1.: Mass attenuation coefficients for H_2O. Taken from the photon cross sections data base XCOM [NIST]

linear attenuation coefficients is called "segmentation". Usually this segmentation is discrete (e.g. GEANT, SimSET), meaning that for a certain density interval a specific tissue with the corresponding cross sections is assumed. In the presented code the segmentation is continuous. This approach, which is adopted from , is based on the observation that in the case of body tissue (and $E \gtrsim 0.2\,\mathrm{MeV}$) the cross section for Compton scattering (and Photo effect) varies weakly for materials of comparable density. Thus, it is reasonable to assume a mapping

$$\rho \longrightarrow \mu_\rho, \tag{4.1}$$

especially, if the density information is the only information given. Here ρ is the mass density, and μ_ρ is the linear attenuation coefficient that can be expected when the material is body tissue of the density ρ. A similar map as in XVMC was used[1]:

$$\text{Compton effect:}\quad \mu_\rho^{\mathrm{C}}(E) \approx \begin{cases} \mu_{\mathrm{H_2O}}^{\mathrm{C}}(E)\,\rho/\rho_{\mathrm{H_2O}} & \rho \le 1\,\mathrm{g/cm^3} \\ \mu_{\mathrm{H_2O}}^{\mathrm{C}}(E)\,(0.85\,\rho/\rho_{\mathrm{H_2O}} + 0.15) & \rho > 1\,\mathrm{g/cm^3} \end{cases} \tag{4.2}$$

$$\text{Photo effect:}\quad \mu_\rho^{\mathrm{P}}(E) \approx \begin{cases} \mu_{\mathrm{H_2O}}^{\mathrm{P}}(E)\,\rho/\rho_{\mathrm{H_2O}} & \rho \le 1.1\,\mathrm{g/cm^3} \\ \mu_{\mathrm{H_2O}}^{\mathrm{P}}(E)\,\rho/\rho_{\mathrm{H_2O}}\,(1 + 8\sqrt{\rho/\rho_{\mathrm{H_2O}} - 1.1}) & \rho > 1.1\,\mathrm{g/cm^3} \end{cases}$$

[1]Some small simplifications for low density linear attenuation coefficients of the Photo effect compared to XVMC are used. Because the Photo effect is unlikely ($\approx 1\%$), the corresponding deviations are neglegible.

This mapping was obtained by fit to ICRU data [Fippel, 2000, ICRU, 1992]. The linear attenuation coefficients for the Compton effect $\mu_\rho^{\mathrm{C}}(E)$ and for the Photo effect $\mu_\rho^{\mathrm{P}}(E)$ are functions of the respective linear attenuation coefficients $\mu_{\mathrm{H_2O}}^{\mathrm{C}}(E)$ and $\mu_{\mathrm{H_2O}}^{\mathrm{P}}(E)$ for water, the density ρ, and the energy E of the photon.

4.1.2. Tracing of particles in voxelized phantoms

In the simulation the positron range as well as the rest energy before annihilation is assumed to be zero. Therefore, the simulation starts with two photons at the position of the emission that travel in opposite directions. Since the rest energy of the positron and the electron is zero, the two photons have the energy 511 keV and the non-collinearity is zero.

The random generator used in the simulation was the pseudo random number generator RANMAR which was taken from XVMC [Fippel, 1999, 2000]. This random number generator provides a uniform random number RND $\in [0, 1[$.

The direction of the two photons is determined by

$$\begin{aligned} \varphi &= 2\pi \mathrm{RND} \\ \vartheta &= \arccos\left(\mathrm{RND}\right) . \end{aligned} \tag{4.3}$$

Equation (4.3) guarantees that this direction is uniformly sampled.

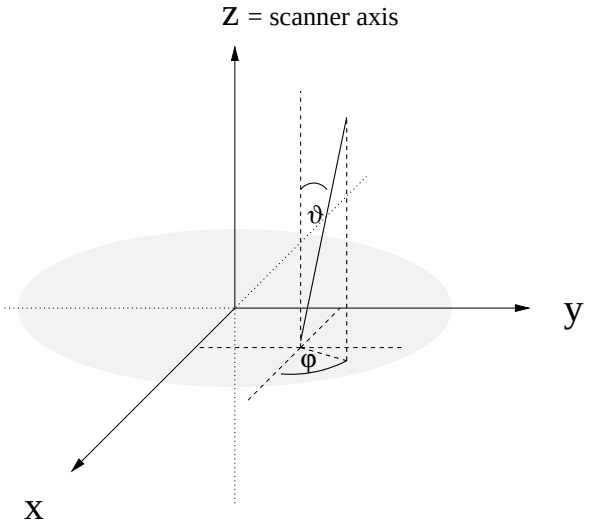


Figure 4.2.: Definition of the angles and the coordinate system with respect to the tomograph.

The probability that a particle does interact before it covered the distance x in a material with the (total) linear attenuation coefficient μ is

$$P(x) = 1 - \exp(-\mu x) \tag{4.4}$$

and the probability that the particle does interact at position x is

$$p(x) = \frac{\partial P}{\partial x} = \mu \exp(-\mu x)\,. \tag{4.5}$$

In the Monte Carlo code this probability distribution is used to obtain the location of the position where the next interaction will occur. The problem of sampling from this non uniform probability function $p(x)$ was overcome by using the *inverse transform method*. When $p(x)$ should be sampled in the interval $[0, x_{\text{leave}}[$, then the normalized cumulative distribution function $F(x; x_{\text{leave}})$ can be interpreted itself as a uniformly distributed random number RND in the interval $[0, 1[$

$$F(x; x_{\text{leave}}) = \frac{P(x)}{\int_0^{x_{\text{leave}}} p(x')\,\mathrm{d}x'} = \frac{1 - \exp\left(-\mu x\right)}{1 - \exp\left(-\mu\, x_{\text{leave}}\right)} \stackrel{!}{=} \text{RND}\,, \tag{4.6}$$

and by inversion

$$\mu x = -\ln\left(1 - \text{RND} \cdot \left(1 - \exp\left(-\mu\, x_{\text{leave}}\right)\right)\right) \tag{4.7}$$

it is possible to sample the distance x according to the non uniform probability distribution $p(x)$ by using the uniformly sampled RND. Equation (4.7) will be used later in section 4.3.1 when the variance reduction techniques are explained. In the Monte Carlo simulation without variance reduction it is possible that the particle never interacts with the phantom This implies $x_{\text{leave}} \to \infty$ and (4.7) simplifies to

$$l \equiv \mu x = -\ln\left(1 - \text{RND}\right). \tag{4.8}$$

A nice property of (4.8) is the fact that the right part does only depend on the random number RND. The introduced dimensionless variable l is called attenuation path length henceforth. When the medium is homogeneous, the next interaction therefore occurs after the distance l/μ. In a medium with varying linear attenuation

coefficient, l becomes

$$l = \int_C \mathrm{d}x' \, \mu(x') \approx \sum_i^C \Delta_i^C \mu_i \,. \tag{4.9}$$

The right part of (4.9) is the simplification in the case of a voxelized phantom. Unfortunately, the set of paths $\{\Delta_i^C\}$ is in general not equally spaced and has to be calculated successively.

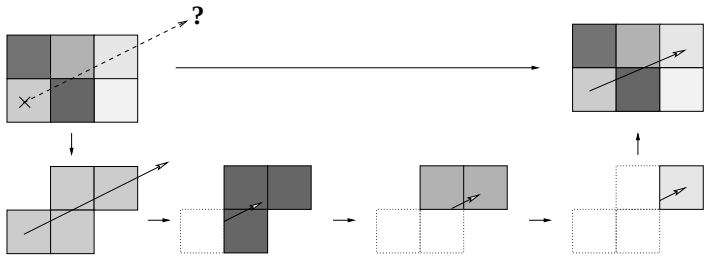


Figure 4.3.: Determination of the location of the next interaction. The steps that are needed to determine the interaction position in a voxelized phantom are shown. At each step it is shown how far the photon would travel if the rest of the voxels had the same linear attenuation coefficient as the present voxel. When this distance is smaller than the distance to the next voxel, the successive reduction of l ends (lower right image). Dark voxels represent high density, bright voxels low density.

Fig. 4.3 shows how the distance to the position of interaction is determined. Let us assume that voxel i is the voxel in which the particle starts. In this voxel i with the linear attenuation coefficient μ_i the path length would be $x = l/\mu_i$ If x is smaller than the distance to the next voxel along the particle's path, the interaction occurs at the position specified by x within this voxel. In the other case l is reduced by $\Delta_i^C \mu_i$ (with Δ_i^C = traveled distance within voxel i). Then the voxel index i is replaced by the index of the neighboring voxel in which the particle travels in a straight line and the the same process is started again. Eventually, the particle interacts within the phantom or leaves it without any interaction. In the latter case it can be checked if the particle hits the detector surface.

Photo effect

This interaction was determined by calculating the linear attenuation coefficients of the Compton effect and the Photo effect according to section 4.1.1. Rayleigh scattering was

neglected, because the relevance is rather small and there exist no reasonable map from density (or electron density) to the corresponding linear attenuation coefficient. The Photo effect was included, because it is very easy to incorporate the effect. Whenever the interaction was a Photo effect the photon history simply ended. This happened whenever

$$\mathrm{RND} < \frac{\mu_{\mathrm{Photo}}}{\mu_{\mathrm{Compton}} + \mu_{\mathrm{Photo}}} \,. \tag{4.10}$$

Compton effect

As mentioned before (see section 2.3), the most common interaction at photon energies between 0.2 MeV and 1 MeV in body tissue is by far the Compton effect. The Compton effect describes the inelastic collision of a photon with an electron (of mass m_{e}). Since the considered energy range $E > 0.2\,\mathrm{MeV}$ is beyond atomic or molecular binding energy in tissue, the electron can be considered as being at rest. Hence, in good approximation the energy of the scattered photon (4.11) can be calculated assuming conservation of relativistic momentum and energy.

$$E_{\mathrm{new}} = E R(E,\vartheta) \quad \text{with} \quad R(E,\vartheta) = \frac{1}{1+\alpha(1-\cos\vartheta)} \quad \text{and} \quad \alpha = \frac{E}{m_{\mathrm{e}}c^2} \tag{4.11}$$

The Klein-Nishina [Knoll, 2000] formula describes the differential cross section of an incoming photon of energy E with a free electron at rest

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}(E,\vartheta) = \frac{1}{2} r_{\mathrm{el}} \left(R(E,\vartheta) - R(E,\vartheta)^2 \sin^2\vartheta + R(E,\vartheta)^3 \right) \tag{4.12}$$

with $r_{\mathrm{el}} = \frac{1}{4\pi\epsilon_0}\frac{e^2}{m_{\mathrm{e}}}$ being the classical electron radius. Sampling from this probability distribution has to be divided into two parts, because two variables $\vartheta_{\mathrm{scatter}}$ and $\varphi_{\mathrm{scatter}}$ need to be determined. The latter variable, the azimuthal angle $\varphi_{\mathrm{scatter}}$ is obtained by

$$\varphi_{\mathrm{scatter}} = 2\pi \mathrm{RND} \,. \tag{4.13}$$

The scatter angle $\vartheta_{\text{scatter}}$ is calculated using the probability distribution

$$\begin{aligned} p_E(\vartheta) &= \frac{\mathrm{d}\sigma}{\mathrm{d}\vartheta}(E,\vartheta) = \int_0^{2\pi} \frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}(E,\vartheta)\sin\vartheta\,\mathrm{d}\varphi \\ &= \pi r_{\text{el}} \frac{\left(\frac{1}{1+\alpha(1-\cos\vartheta)} + \alpha(1-\cos\vartheta) + \cos^2\vartheta\right)\sin\vartheta}{(1+\alpha(1-\cos\vartheta))^2}\,. \end{aligned} \tag{4.14}$$

The inverse transform method (see section 4.1.2) is used to sample from this non uniform probability function $p_E(\vartheta)$. Again, the normalized cumulative distribution function $P_E(\vartheta;\vartheta_0,\vartheta_{\max})$ can be interpreted itself as a uniformly distributed random number RND in the interval $[0,1[$:

$$\text{RND} \equiv P_E(\vartheta;\vartheta_0,\vartheta_{\max}) = \int_{\vartheta_0}^{\vartheta} \bar{p}_E(\vartheta')\,\mathrm{d}\vartheta' = \int_{\vartheta_0}^{\vartheta} \frac{p_E(\vartheta')\,\mathrm{d}\vartheta'}{\int_{\vartheta_0}^{\vartheta_{\max}} p_E(\vartheta'')\,\mathrm{d}\vartheta''} \tag{4.15}$$

Its inverse $P_E^{\text{inv}}(\text{RND})$ provides the random number $X = \vartheta$ distributed according to $p_E(\vartheta)$.

$$X = P_E^{\text{inv}}(\text{RND}) \tag{4.16}$$

In case it is impossible to form P^{inv} like in the case of $p_E(\vartheta)$, it is reasonable to use a discrete approximation $\tilde{P}^{\text{inv}}$ of P^{inv}.

$$\text{RND} \xrightarrow{\tilde{P}^{\text{inv}}} \begin{cases} X_0 & \text{for RND} \in [0, \frac{1}{n}[\\ X_1 & \text{for RND} \in [\frac{1}{n}, \frac{2}{n}[\\ \dots & \\ X_n & \text{for RND} \in [\frac{n-1}{n}, 1[\end{cases} \tag{4.17}$$

The inverse function can be represented by a vector with $n+1$ elements. This vector is calculated by numerical integration of $\bar{p}_E(\vartheta)$. The integration is approximated by a Riemann sum with respect to the mth regular subdivision on $[\vartheta_0,\vartheta_{\max}]$ with $m \gg n$. Whenever the Riemann sum exceeds $1/n$, $2/n$, $3/n$, $\dots (n-1)/n$ at positions ϑ_1, ϑ_2,

$\vartheta_3, \ldots, \vartheta_{n-1}$ an approximate value for $X_{i=0\cdots n}$ is calculated by forming

$$X_i = \begin{cases} 1/2\,(\vartheta_{i+1} + \vartheta_i) & \text{for } 0 \leq i < n \\ 1/2\,(\vartheta_{\max} + \vartheta_n) & \text{for } i = n. \end{cases}$$

In this way by sampling uniformly the interval $[0, 1[$ is possible to get an approximate scatter angle $\vartheta_{\text{scatter}} = \tilde{P}^{\text{inv}}(\text{RND})$. A discretization with $\vartheta_0 = 0$, $\vartheta_{\max} = \pi$, $n = 5000$ and $m = 10\,n$ yields reasonable results. The formation of this vector X_i is repeated for different energies. Since the energy dependence of $\bar{p}_E(\vartheta)$ is not very strong, the used energy separation of approximately 125 eV is reasonably small enough. Altogether, this requires the storage of less than 10^6 values and provides a very fast way to calculate the scatter angle $\vartheta_{\text{scatter}}$. The energy of the scattered photon is calculated easily by applying equation (4.11).

4.2. Simulated particle detection

The focus of the simulations is the incorporation of patient scatter into the system matrix. The simulated model of the scanner uses simplified detectors. Photons that hit the detector surface and exceed a certain energy threshold are counted as being detected. Together with dead time (which is not simulated), this leads to an overestimation of the absolute number of counts when compared to more realistic scanners, but should give good insight into relative distributions which was important for the evaluation of the proposed compression scheme. In some simulations Gaussian energy blurring

$$E_{\text{blurred}} = E + \delta E \tag{4.18}$$

was applied to the energy of the photons before they hit the detector surface and were thresholded. δE was Gaussian distributed around mean 0 with standard deviation $\sigma = \frac{R\sqrt{m_e c^2}\sqrt{E}}{2\sqrt{2\ln 2}}$ and with the energy resolution R at $E = m_e c^2 = 511\,\text{keV}$.

4.3. Variance reduction techniques

Although the previously discussed way to simulate photon transport in the phantom is quite fast, the number of useful LORs per emitted photon pair is rather small. Only a small fraction of the started photons hit a detector, most leave the phantom without ever passing the detector ring or are scattered and not detected because the energy

drops below the threshold of the detectors . The number of detected events is limited even more because two photons are needed to form a coincidence. The goal is therefore the improvement of the fraction of detectable coincidences. In order to achieve this, two different variance reduction techniques were used. These techniques favor certain outcomes (here the detection of a coincidence) over outcomes that are not wanted or that do not carry useful information (here photons that leave the simulated volume without being detected or which are absorbed). These variance reduction techniques are inspired by [Haynor et al., 1990, 1991].

4.3.1. Forced detection

The first technique mostly improves the number of detected scattered events and is based on the method of importance sampling. In importance sampling, a given probability distribution is replaced by an alternative probability distribution that favors wanted outcomes. The resulting bias is removed by the normalization of the outcome with the correct weight. Translated into the problem of reducing the variance of the detected coincidences, favorable outcomes are detected coincidences. Forced detection improves therefore the detection of a coincidence by improving the detection of single photons.

For a simulated photon in the phantom (either just created or already scattered) four possibilities exist: (a) either it escapes and is detected, (b) the particle escapes without detection, (c) the particle does not leave the phantom, but its energy drops below the lower energy threshold of the detectors[2] or (d) the particle stays in the phantom with energy larger than the energy threshold. While the three first possibilities represent final states of a particle, the last represents an intermediate state. The first outcome is wanted, the second and third are not wanted and the last is undefined, but has to be pursued due to this reason.

Since case (a) and partially case (d) are favorable, it is reasonable to ensure that photons only leave the phantom in the direction of the scanner ring(s) and that photons are forced to stay in the phantom. While the first is always favorable, the second is only favorable to a certain degree. The number of interactions inside the phantom (scatter order) of the modified simulations should not increase considerably the average scatter order in the original simulation without variance reduction techniques, because otherwise too much time is spent on the simulation of rarely occurring events. This suggests to link the modified simulation somehow to the original simulation. This is done as follows:

[2]Photo electric absorption is treated like a photon of zero energy.

In an original simulation a photon is traced through the phantom. During its way copies of the photon are created with correct weights that are forced to interact within the phantom and scattered in directions that lead to a hit on the detector ring(s). These forced photons are counted. When the original photon leaves the phantom, the tracing of this photon ends and the photon is discarded to avoid double counting.

The interaction forcing inside the phantom can be achieved by using (4.7):

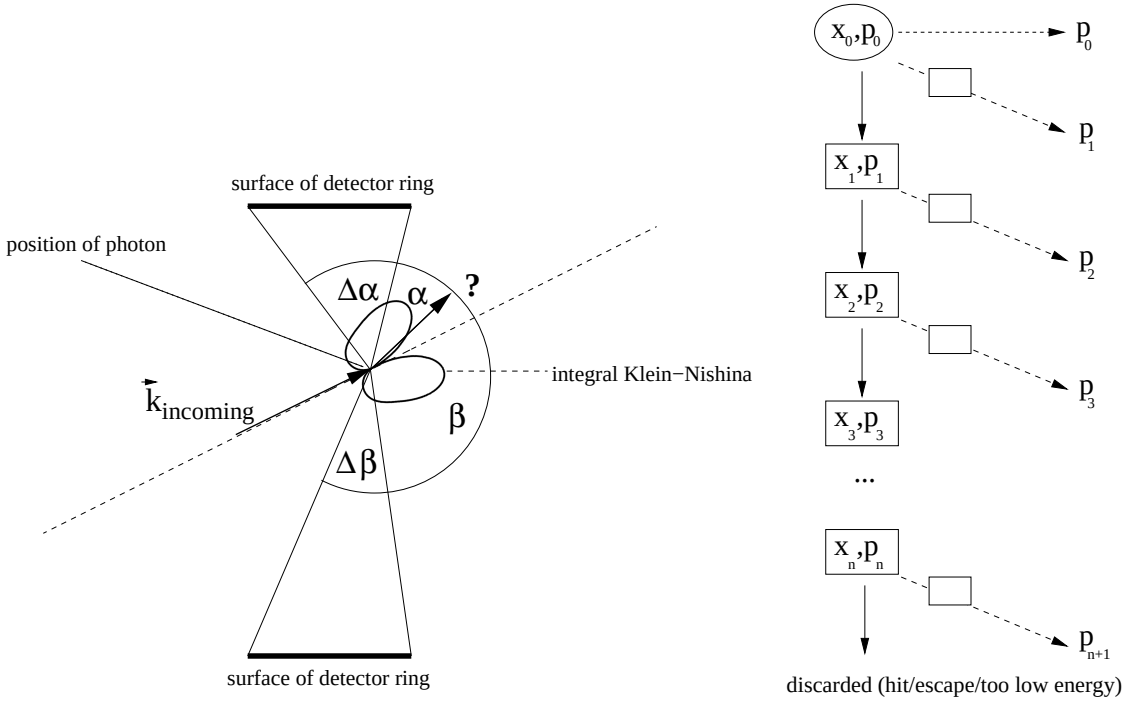
$$\hat{l} = -\ln\left(1 - \mathrm{RND} \cdot \left(1 - \exp\left(-\hat{x}_{\mathrm{leave}}\right)\right)\right) \quad \text{with} \quad \hat{x}_{\mathrm{leave}} = \sum_{i}^{C} \Delta_i^C \mu_i \tag{4.19}$$

Here $\hat{x}_{\mathrm{leave}}$ is the attenuation path length of the photon that would leave the patient on a straight line C. In contrast to (4.8), the right part of equation (4.19) also depends on the linear attenuation coefficients of the voxels (in $\hat{x}_{\mathrm{leave}}$). It is therefore necessary to calculate $\hat{x}_{\mathrm{leave}}$ *before* this interaction forcing is done. A random number $\mathrm{RND} \in [0, 1[$ always enforces an interaction in the phantom somewhere on the line C. The position of the next interaction is then determined in the same way as in section 4.1.2 by reducing successively the attenuation path $\hat{l}$ by $\Delta_i^C \mu_i$.

The forcing of the scattered (copied) photon in the direction of the detector ring(s) is done in the following way. First the azimuthal angle φ is chosen randomly from the interval $[0, \pi[$. This specifies a plane of interaction. Usually this plane of interaction intersects with the detector ring(s). Depending on the location and direction of the plane relative to the scanner, it is possible that the plane intersects along two segments (usually most likely for photons that are scattered inside the scanner), along one segment (when the plane is approximately transversal) or not at all (only possible when the photon is scattered outside the scanner and the plane is approximately transversal). One example of the most likely case of two intersecting segments is shown in Fig. 4.4(a).

For simplicity the plane in Fig. 4.4(a) is aligned along the symmetry axis of the scanner, but in general this is not the case. When the scattering angle ϑ is chosen in such a way, that the photon travels towards these segments, and no further scattering occurs, the photon hits the scanner surface. The segments are computed by calculating the intersection points of the two rings that define the inner opening at the front side and the back side of the scanner. This results in up to four points on the plane of interaction. The correct angle intervals are determined by a case differentiation. The scattering angle was chosen according to the integral Klein-Nishina formula[3] in these

[3]Here, integral Klein-Nishina formula means that the differential Klein-Nishina is integrated over φ



(a) **Sideview of the scanner** Instead of sampling the integral Klein-Nishina distribution in the interval $[-\pi, \pi[$ it is only sampled in the intervals $\Delta\alpha$ and $\Delta\beta$.

(b) **Forced detection** Instead of photons of the original simulation (arrows), copied photons are counted (dashed arrows).

Figure 4.4.: (a) direction forcing and (b) the calculation of the weight of the forced photons. In (b) each rectangle represents an interaction. Dashed lines represent forced photons, and the solid line represents the original simulation. The corresponding probabilities can be found in equation (4.24).

interval(s).

A photon inside the scanner can always be forced in the direction of the scanner and also for a photon outside the scanner this is often the case. The code should therefore also be suited well to simulate out of field of view (FOV) scatter. The two proposed changes must be inserted in the simulation process without introducing any bias. In Fig. 4.4(b) it can be seen how this is achieved. First, when the photon is started, it is checked whether it would hit the surface of a detector if not scattered. If this is the case, a hit is stored using the weight $p_0 = p_{\text{leave}}$ accounting for the probability of this event. Then a copy of the original photon is forced to interact during its passage in the phantom and after sampling randomly φ, the scatter angle ϑ is sampled as described ensuring a direction towards the detector ring surface. The weight of this photon has

(see section 4.3.1).

to be adapted for the probability that (b) the photon interacts within the phantom (p_{int}), (c) the process is a Compton scatter and not a Photo effect (or Rayleigh, which at present is not handled) (p_{ratio}), (d) the probability that the photon is scattered in the direction towards the scanner ring (p_{dir}) and (a) no further interaction occurs (p_{leave}).

$$\text{a)} \qquad p_{\text{leave}}(\mathbf{x}, \mathbf{p}) = \exp\left(-\hat{x}_{\text{leave}}(\mathbf{x}, \mathbf{p})\right) \tag{4.20}$$

$$\text{b)} \qquad p_{\text{int}}(\mathbf{x}, \mathbf{p}) = 1 - p_{\text{leave}}(\mathbf{x}, \mathbf{p}) \tag{4.21}$$

$$\text{c)} \qquad p_{\text{ratio}}(\mathbf{x}, E) = \frac{\mu_{\text{Compton}}(\mathbf{x}, E)}{\mu_{\text{total}}(\mathbf{x}, E)} \tag{4.22}$$

$$\text{d)} \qquad p_{\text{dir}}(\mathbf{x}, \mathbf{p}; \varphi_{\text{leave}}) = \text{ (see text below)} \tag{4.23}$$

Here $\hat{x}_{\text{leave}}(\mathbf{x}, \mathbf{p})$ is again the attenuation path along the straight way of the unscattered photon (starting at position $\mathbf{x}$ with momentum $\mathbf{p}$ until it leaves the phantom). The change of the weight due to the change of the scattering angle (p_{dir}) is basically the ratio of the gray area in Fig. 4.4(a) and the "white+gray" area being two times the integral of the integral Klein Nishina distribution. In Fig. 4.4(b) it can be seen how photons of higher scatter order are forced to be detected. While the original photon is tracked in a normal fashion, directly after each interaction a copied photon is forced to interact on its way through the phantom. The scatter angle is chosen from those that guarantee a hit on the scanner surface. The weights in Fig. 4.4(b) are:

$$p_n = \begin{cases} p_{\text{leave}}(\mathbf{x}_0, \mathbf{p}_0) & \text{for} \quad n = 0 \\ p_{\text{int}}(\mathbf{x}_0, \mathbf{p}_0)\, p_{\text{ratio}}(\mathbf{x}_{\text{f}}, E_{\text{f}})\, p_{\text{dir}}(\mathbf{x}_{\text{f}}, \mathbf{p}_{\text{f}}; \varphi_{\text{leave}})\, p_{\text{leave}}(\mathbf{x}_{\text{f}}, \mathbf{p}_{\text{leave}}) & \text{for} \quad n = 1 \\ \left(\prod_{i=1}^{n-1} p_{\text{ratio}}(\mathbf{x}_i, E_i)\right) p_{\text{int}}(\mathbf{x}_{n-1}, \mathbf{p}_{n-1}) & \\ \qquad \times p_{\text{ratio}}(\mathbf{x}_{\text{f}}, E_{\text{f}})\, p_{\text{dir}}(\mathbf{x}_{\text{f}}, \mathbf{p}_{\text{f}}; \varphi_{\text{leave}})\, p_{\text{leave}}(\mathbf{x}_{\text{f}}, \mathbf{p}_{\text{leave}}) & \text{for} \quad n \geq 2\,. \end{cases} \tag{4.24}$$

One emission can therefore lead to several "detected" coincidences of photons of usually very small weight. This forced detection scheme therefore introduces a correlation between the counts of different LORs.

4.3.2. Stratification

So far, the quality of the estimate for the scattered detected events is increased considerably. The corresponding estimate for the unscattered photons, however, is almost

not improved at all. The sole usage of forced detection leads to a disproportionate emphasis on scatter coincidences.

The reason why rather few coincidences that are not scattered are detected is due to the fact that the uniformly sampled starting direction (equation (4.3)) of most photon pairs is not directed towards the detector ring. This uniform sampling is therefore highly inefficient. In order to improve the sampling, the set of all starting angles was divided into N_Ω disjoint subsets Ω_i, called stratification cells. A starting probability $n_i \in]0,1]$ with $\sum_i n_i \equiv 1$ and a starting weight w_i are assigned to each subset with index i. The starting probability n_i is the probability that the starting angle lies in the corresponding cell i. This allows us to increase the probability that a photon pair is started in certain directions while the probability for other directions (like in z-direction, the direction of the scanner axis) can be decreased. The predefinition of the starting weights

$$w_i = \frac{F_i}{n_i} \tag{4.25}$$

with

$$F_i = \frac{|\Omega_i|}{4\pi} \qquad \text{with} \quad |\Omega_i| \stackrel{\text{def}}{=} \int_{\Omega_i} \mathrm{d}\Omega \;\; \text{and} \;\; \sum_i |\Omega_i| = 4\pi \tag{4.26}$$

avoids any bias. While heuristically chosen n_i (large n for directions approximately towards scanner ring and small n for directions close to z-axis) can already result in a strong acceleration of the code, it is possible to maximize the acceleration for a given set of stratification cells $\{\Omega_i\}$ by minimizing the variance $\sigma^2 \equiv \sigma^2(\Sigma_y)$ of the sum of all detected coincidences

$$\Sigma_y \equiv \sum_{j=1}^{N_\mathrm{L}} y_j = \sum_{k=1}^{\breve{N}} \breve{w}_k \,. \tag{4.27}$$

Since usually stratification is used together with forced detection, it can happen that multiple small weighted coincidences are detected for a single emission only. Also it is important to notice, that $\breve{w}_k$ is the weight of a detected coincidence, therefore $\breve{w}_k = \breve{w}_k^{\text{photon1}} \breve{w}_k^{\text{photon2}}$ and not the weight of single photons in contrast to section 4.3.1.

It is shown in appendix A.2 that for reasonable large $\breve{N}$ ($\breve{N}$=total number of detected counts) the variance of the detected counts can be approximated by the sum of the

squared detected weights $\breve{w}$

$$\sigma^2 \approx \sum_{k=1}^{\breve{N}} \breve{w}_k^2 \,. \tag{4.28}$$

Furthermore, it is possible to estimate the variance σ^2 for arbitrary n_i, w_i based on previously[4] recorded weights $\breve{w}_{i,k}^{\mathrm{P}}$ (kth weight from stratification cell i) and the number of the corresponding emissions N_i^{P} from cell i.

$$\sigma^2 \propto \sum_i \frac{n_i}{N_i^{\mathrm{P}}} \sum_k (w_i \breve{w}_{i,k}^{\mathrm{P}})^2 \overset{(4.25)}{=} \sum_i \frac{1}{n_i} \underbrace{\frac{F_i^2}{N_i^{\mathrm{P}}} \sum_k (\breve{w}_{i,k}^{\mathrm{P}})^2}_{\equiv \pi_i^2} = \sum_i \frac{\pi_i^2}{n_i} \tag{4.29}$$

Equation (4.29) assumes that the starting weights of the detected weights $\breve{w}_{i,k}^{\mathrm{P}}$ are equal to one (for this reason N_i^{P} can be used). This is not a restriction. For non-uniform starting weights, it is possible to simulate the particles with uniform starting weights, then store $\breve{w}_{i,k}^{\mathrm{P}}$ (for stratification purpose) and later multiply the weights with the non uniform starting weight in order to restore the initial simulation.

It is possible to find n_i that minimize σ^2 by minimizing

$$\sum_i \frac{\pi_i^2}{n_i} \qquad \text{subject to} \quad \sum_i n_i = 1 \,. \tag{4.30}$$

Equation (4.30) can be solved with the help of Lagrange parameters. The solution is

$$n_i = \frac{\pi_i}{\sum_i \pi_i} \,. \tag{4.31}$$

The new starting weights w_i can be calculated with the help of equation (4.25). The optimal number of emission per cells n_i can be therefore calculated after at least a first simulation was performed that resulted in $\breve{w}_{i,k}^{\mathrm{P}}$ and N_i^{P}. In a second run the starting angles and starting weights are then chosen according (4.31) and (4.25). From then on it is in principle possible to update $\breve{w}_{i,k}^{\mathrm{P}}$ and N_i^{P} after each started photon pair and calculate optimal n_i and w_i based on more accurate simulations. Since this update also requires some calculations and therefore computation time, a reasonable update interval was 100-500 emissions. This interval should be also chosen depending on the

[4]The superscript $^{\mathrm{P}}$ stands for *previously*, and $\breve{}$ for *recorded*.

number of stratification cells and the number of initially simulated emissions in the first run.

The first run should result in a sufficient number of recorded particles in all stratification cells in order to provide a reliable estimate. Although the subsequent simulations are used to improve the accuracy of $\breve{w}_{i,k}^{\mathrm{P}}$ and N_i^{P}, this first estimate needs to be good enough to avoid a totally wrong assignment of n_i.

The Monte Carlo code should be optimized for system matrix calculations. This implies that the activity is usually not distributed over the whole phantom, but located in some well defined small volume representing a voxel or other basis function. This means that it is easy to predict starting directions that can result in direct counts and starting directions that will never lead to direct counts. It is reasonable to divide the starting directions into cells that do not result in direct coincidences and into cells that result primarily in direct coincidences. It is not difficult in the case of voxels to calculate the maximal angle $\vartheta_{\max}$ below which only scattered coincidences can occur (Fig. 4.5).

The number of cells is chosen to be rather small, since also very high noise matrices (with less than 10^4 emissions/voxel in the case of 2D scanners) should be possible to simulate. More stratification cells require a larger number of simulated emissions in the first run. In order to have fewer cells, and for simplicity, only ϑ (and not φ) was used to define the border between neighboring cells.

$$\Omega_i = [\vartheta_{i-1}, \vartheta_i] \times [0, 2\pi] \qquad \vartheta_i \in \ ]0, \pi/2] \tag{4.32}$$

Due to the fact that two photons are started in opposite directions, it is enough to sample $\vartheta_i \in [0, \pi/2[$ and $\varphi \in [0, 2\pi[$. Larger intervals would result in redundant information. In the simulations usually only five stratification cells were used together with 1000-2000 emitted photon pairs in the first run. This resulted in good enough estimates of $\breve{w}_{i,k}^{\mathrm{P}}$ and N_i^{P} even in the case of ideal one-ring scanners with very low scatter fractions below 5%. Due to the requirement that sometimes matrices of very bad statistics are to be simulated, it is necessary to bias the number of starting photon pairs already in the first run. This can be done without violating (4.25) by choosing

$$n_i^{\mathrm{initial}} = \frac{1}{2}\left(F_i + \frac{1}{N_\Omega}\right) . \tag{4.33}$$

Since the cells that can lead to direct coincidences are smaller than the other cells, this choice improves the number of detected direct coincidences.

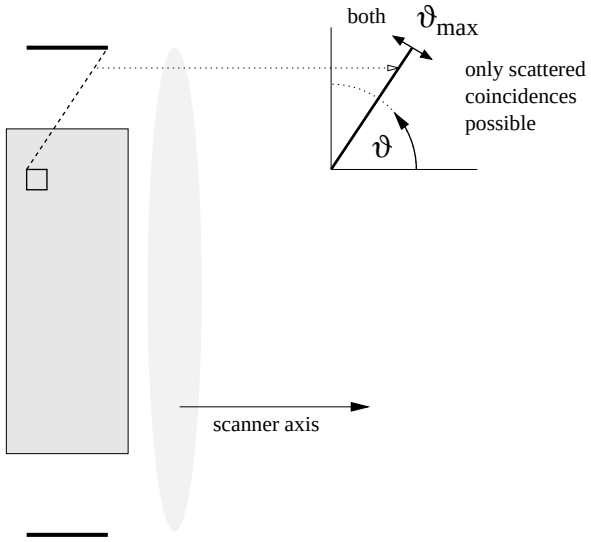


Figure 4.5.: Randomly positioned photon pairs within the boundary of the marked voxel started into directions $\vartheta < \vartheta_{\max}$ can only be detected when they were scattered.

A method to avoid inefficiently small weights or very large weights of the simulated photons, weight control [Haynor et al., 1990, 1991] was not implemented. This method can be used to further increase the efficiency.

4.4. Implementation and parallelization

The Monte Carlo code was implemented in C++ [Schildt, 1998, 2000, Stroustrup, 2000]. System matrix calculation and sinogram calculation is parallelized using the parallel virtual machine library, PVM 3.4.4 [PVM 3.4.4] using an extension called PVM++ of Wilhelmi [2001-06-27], a C++ - library providing convenient wrapper classes for PVM. The simulations were parallelized by dividing the set of voxels into disjoint sets. If this was not possible (like for sinograms of single voxels, or if the number of voxels was not a multiple of the number of processors), the emission of some voxels were distributed amongst the processors. The simulations usually ran on a 16 processor cluster with identical processors. The lack of load balancing therefore did not influence the performance.

5. System Matrix Compression

The key to high quality and quantitative imaging is a correct internal model of the scanner and the relevant physical processes in the patient. This includes scatter in inhomogeneous phantoms/patients, the simulation of different radionuclides, and temporally varying patients (for example breathing, heart beat). Monte Carlo simulations are very well suited for this task. Proposals how to include Monte Carlo simulations into the reconstruction algorithms of emission tomography have recently been made. The proposals are considering SPECT [Beekman et al., 2002, Buvat et al., 2003, Lazaro et al., 2004a,b, 2005], small animal PET [Rafecas et al., 2003, 2004a,b, Shoukouhi et al., 2004, Shoukouhi, 2005] or human PET systems [Levin et al., 1995, Ollinger, 1996, Watson, 2000, Beekman et al., 2002, Werling et al., 2002]. Ideally, the Monte Carlo simulations are used directly to calculate the matrix elements.

In the case of SPECT [Buvat et al., 2003, Lazaro et al., 2004a,b, 2005] and small animal PET [Rafecas et al., 2003, 2004a,b, Shoukouhi et al., 2004, Shoukouhi, 2005] the incorporation of scatter directly into the system matrix was achieved, because the matrix was small enough to be stored. The storage is necessary, because on-the-fly Monte Carlo simulations of the matrix elements during the reconstruction are too time consuming. The reconstruction of images for human PET scanners was improved by using Monte Carlo simulations in the projector of the reconstruction algorithm but not in the back-projector. The storage of the system matrix of a human scanner including scatter caused by the patient is not possible due to the high scatter fraction, the missing symmetry, and the large amount of detectors and voxels. In this chapter a novel way how to incorporate Monte Carlo based scatter into the reconstruction algorithm is presented. This is achieved by using a compressed matrix that can be stored in memory. A compression scheme for the system matrix is introduced that can achieve a sufficient compression and on-the-fly read-out during the reconstruction. This system matrix including scatter can be used in any iterative reconstruction algorithm.

5.1. Goals and requirements

The aim of the matrix compression is the storage of the matrix in the memory. In addition to this important goal there are several other requirements that have to be considered when thinking about a compression scheme. The uncompressed matrix is orders of magnitude larger than the available memory. A simultaneous compression of the matrix is therefore not possible. The compression scheme should allow a sequential calculation. This means that it should only be necessary to keep a small part of the original (uncompressed) matrix in the memory at the same time. Another requirement is the speed of the compression and extraction (read-out). Here, the read-out speed is more limiting than the compression speed, because the matrix must be read out at each iteration step. The read-out speed should be at least faster than the simulation of the matrix. Preferably, the read-out is much faster. Depending on the number of iterations, it might be still reasonable that the compression is slower than the simulation, but of course also here the time needed for compression should in general not exceed the simulation time. Lastly, the compression scheme should not (or only very weakly) affect the resolution of the scanner and in good approximation correctly describe the matrix.

5.2. Properties of the system matrix

Not only do the aforementioned requirements influence the choice of the compression method, but also the properties of the system matrix. The system matrix consists of many matrix elements that are non-zero due to scatter and a small percentage of non-zero elements that also comprise direct (unscattered) counts. Due to the large values of the latter elements, they dominate the sinogram. The values of the scatter-only elements are much smaller. However, due to the much larger number of such elements they also influence the reconstruction.

In Fig. 5.1 sinograms of single voxels are shown. These unit sinograms represent the columns of the system matrix. The sinograms are depicted as gray value images. A sinogram is usually shown in a (ρ, φ) coordinate system: the projection angle φ varies vertically and the distance ρ (or bin number) varies horizontally. A horizontal line in Fig. 5.1 is therefore a projection at a fixed angle. The sinograms were obtained by simulations with 10^8 emissions in an inhomogeneous phantom using variance reduction techniques. Typical projections of neighboring voxels at the same angle φ are shown in Fig. 5.2. The sinusoidal form of the sinograms in Fig. 5.1 can be obtained by a Radon transform (and therefore independently of the patient), but the value depends

on the attenuation in the patient.

The scatter-free unit sinograms (only photons of scatter order zero are used to form coincidences) are strongly peaked (see Fig. 5.2(a)). The position and the value (logarithmic scale!) of the maximal bin is clearly changing for neighboring voxels. The part of the unit projections that are caused solely by scattered photons (the scatter tails) in the projections are much broader (Fig. 5.2(b)). The maximum of this unit scatter projections is usually located at the position of the maximum of the scatter-free projection. For unit sinograms of voxels at or close to the border of the phantom this might not always be true (Fig. 5.1(d)), but for the vast majority of the voxels inside the patient boundary this applies. The shape of neighboring voxels only changes slowly (see Fig. 5.2(b)). The shape is not shift-invariant (compare Fig. 5.1(c) and Fig. 5.1(d)) and also inhomogeneity inside the phantom can influence the shape of the scatter sinograms (see Fig. 5.1(b) or Fig. 5.1(c)).

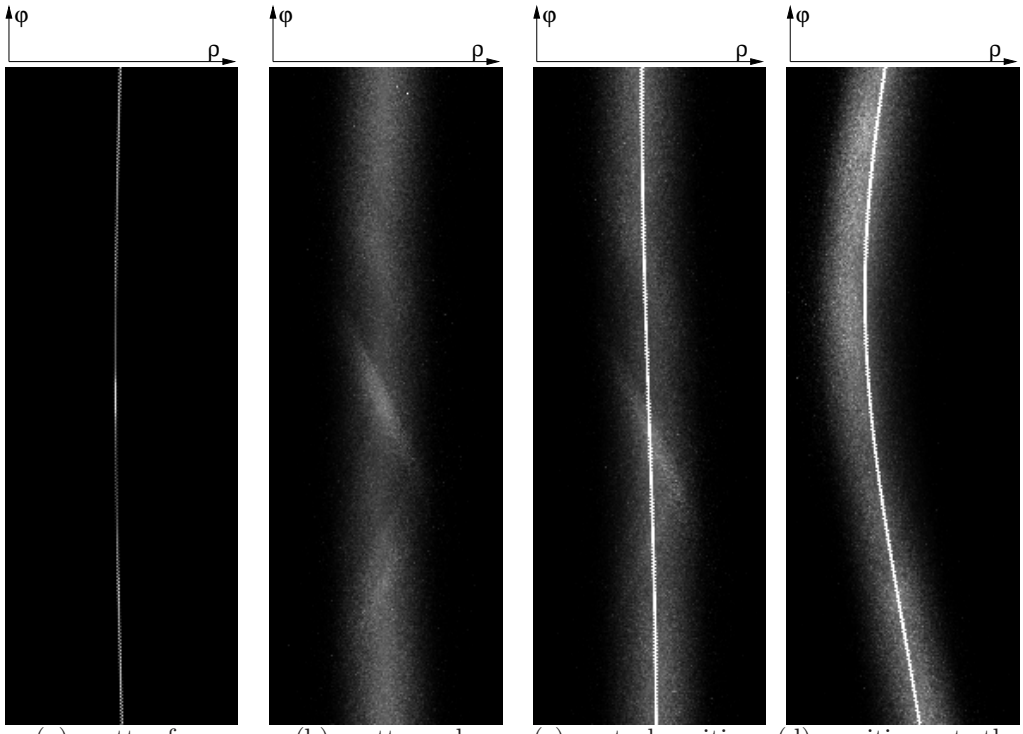
All these figure are obtained by simulations using a very high number of photons and applying variance reduction techniques. System matrices that are calculated by Monte Carlo simulations within a more realistic time (and often also measured sinograms) are much noisier. In low photon number simulations only very few single scattered photons are simulated. This leads to very noisy scatter projections (Fig. 5.3).

5.3. Compression scheme

Fig. 5.3 suggests to save the matrix in a sparse manner. Sparse storage means that only non-zero elements of the matrix are stored. For the following reason the obtained reduction in size is not large enough. A 3D scanner has usually at least around $N_\mathrm{D} \approx 10^4$ detectors and should provide volume imaging with around $N_\mathrm{V} \approx 10^6$ voxels. The number of matrix elements that are non-zero due to unscattered coincidences can be estimated by $N_\mathrm{direct} \approx \mathcal{O}(N_\mathrm{V} N_\mathrm{D}) = 10^6 \cdot 10^4 = 10^{10}$. Therefore, it is already difficult to store the direct coincidences of the matrix in a sparse manner. The sparse storage of the whole matrix of 3D scanners with present computers is not feasible.

The total number of matrix elements N is given by $N \sim \mathcal{O}(N_\mathrm{V} N_\mathrm{L})$ where N_V is the number of voxels (or number of columns of the matrix) and N_L the number of lines of response (LORs) (or number of rows). A substantial reduction of the size of the matrix requires that both the number of columns and the number of rows of the matrix are effectively reduced.

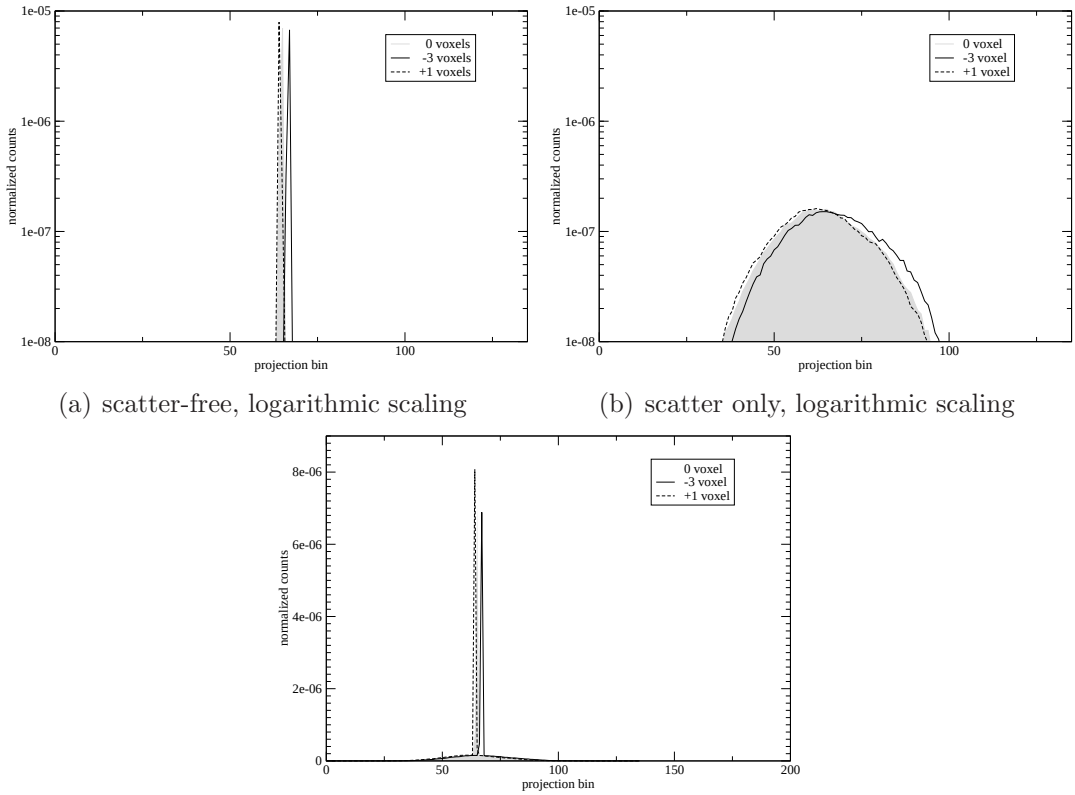
The requirements and properties of the matrix (section 5.2) rule out some approaches that appear reasonable at first glance. All approaches based on (rotational) symmetries



(a) scatter-free (b) scatter only (c) central position (d) position at the border

Figure 5.1.: Typical sinograms of single voxels. Fig. (a) and (b) show scatter-free and scatter-only sinograms of the same single voxel, respectively. Fig. (c) and (d) show both scatter and direct counts, (c) is a sinogram of a rather central voxel and the last image (d) shows the sinogram of a single voxel placed at the border of the phantom (but not at the border of the field of view). Gray value scaling was applied in order to visualize scatter in (b), (c), and (d): the whole dynamic range of gray values is used for a subset of sinogram values (windowing). Small values are visualized by black and large values by white.

will fail due to to the asymmetric patient. Possible approaches to compress the matrix can be basically divided into three schemes. In one scheme single rows or groups of few rows of the matrix are compressed simultaneously. This is not reasonable, because it is difficult to change the Monte Carlo simulations in such a way that the rows can be simulated successively. In a second scheme the whole matrix is compressed at once. This is not possible due the memory constraints during the compression. The last scheme comprises the compression of single columns or groups of a small number of columns. Single columns (or a small number of columns) can be simulated directly.



(a) scatter-free, logarithmic scaling

(b) scatter only, logarithmic scaling

(c) both, linear scaling

Figure 5.2.: Projections of sinograms of single voxels that are in the vicinity of each other. Two voxels are placed next to each other and a third voxel is placed 3 (or 4) voxels from the others.

Therefore symmetry considerations (like in Johnson [1997] or Kehren [2001]), compression schemes that make use of the transformation of the whole matrix (like straight forward Fourier transform or wavelet transforms) or statistical methods that need the whole matrix (like principal component analysis of the whole matrix) are not good choices.

The presented method compresses successively small groups of columns of the system matrix (without using symmetries). In the next sections this compression method is described. To motivate the final method, first the basic idea is explained.

5.3.1. Principle

As discussed, the compression scheme should provide a way to effectively compress the matrix in the voxel and in the LOR domain. A compression in voxel space (i.e. a

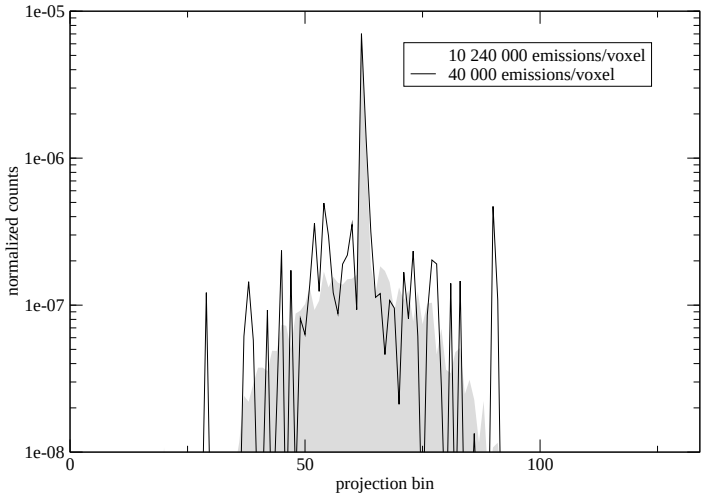


Figure 5.3.: A projection of a typical high and low statistic simulation using variance reduction techniques. In the low count simulation it is difficult to recognize the shape of the scatter tail.

reduction of the number of columns of the matrix) is problematic to achieve without degrading the resolution. Therefore the part of the matrix mainly responsible for the resolution (or small structures respectively high space frequencies), namely the scatter-free part of the matrix, is separated from the rest of the matrix. This can be done and does not infringe the aforementioned requirements, because the scatter-free or attenuation matrix A can be calculated on-the-fly or be stored due to its very sparse nature [Kehren, 2001].

This suggests to decompose the matrix M like

$$M = A + S, \tag{5.1}$$

where S is the scatter only part of the system matrix M. The matrix A should factor in geometrical effects, attenuation, and detector efficiency and normalization (if the data is not corrected for it). The matrix A can be calculated approximately by conventional methods [Kehren, 2001]. The remaining goal is therefore the reduction of the storage size of matrix S.

The proposed compression of matrix S consists of two steps: A compression of the lines of response (LORs) for a given voxel and a compression in the voxel domain. In the following the compression for a single ring scanner (one transversal sinogram) is explained. The compression in the LOR domain is based on the assumption that the

left and right parts of the projections of the unit scatter sinogram of a single voxel can be approximated well by a function of few parameters. The sum of a Gaussian and an exponential function (see Fig. 5.4) was chosen. For voxels inside the patient, Monte Carlo simulations (see Fig. 5.2) as well as measurements of rod sources at different positions inside a phantom [Bergström et al., 1983] show that this is a good choice.

The left (L) and right (R) part of the scatter projections are described by functions $g^{\mathrm{L/R}}_{\varphi,\mathbf{x}}(\rho) = h^{\mathrm{L/R}}_{\varphi,\mathbf{x}}(\rho - \rho^0)$ defined by

$$h^{\mathrm{L/R}}_{\varphi,\mathbf{x}}(\rho) = \exp\left(a^{\mathrm{L/R}}_{\varphi,\mathbf{x}} + b^{\mathrm{L/R}}_{\varphi,\mathbf{x}}\rho\right) + \exp\left(c^{\mathrm{L/R}}_{\varphi,\mathbf{x}} + d^{\mathrm{L/R}}_{\varphi,\mathbf{x}}\rho^2\right), \tag{5.2}$$

where $\rho^0 = \rho^0(\varphi, \mathbf{x})$ is the *geometrically expected maximum* of the unit scatter projection at the angle φ for a voxel at position $\mathbf{x} \equiv (x_1, x_2, x_3)^{\mathrm{T}}$ (see Fig. 5.4).

The geometrically expected maximum ρ^0 is determined by assuming that the maximum of the projections of the direct and scattered counts coincide. Since the position of the maximum of the unscattered (direct) counts can be determined by a simple Radon transform, ρ^0 can be calculated by geometrical considerations.

This assumption fails in the case of voxels that are located outside the phantom/patient, because there the real maximum of the scatter projections does usually not coincide with the geometrically expected maximum and a stronger deviation of the Gaussian/exponential shape can be expected. In the following it is assumed that the patient boundary can be determined accurately as it is the case for PET/CT scanners and ρ^0 will be calculated by the Radon transform. Voxels that are outside the patient boundary need not to be reconstructed, because by definition no activity is to be found there. The wrong position of the geometrically expected maximum has therefore no influence and in addition this leads to a reduction of the number of columns of the system matrix. Scatter from outside the FOV can be included directly in the system matrix by adding columns to the system matrix that represent unit sinograms of large areas (for example sectors) outside the FOV. These columns consist only of scattered coincidences and need not to be compressed because of the small number of these voxels. The incorporation of scatter into the system matrix therefore offers a consistent way of out of FOV scatter treatment.

The compression is achieved by only storing the parameters $a_{\varphi,\mathbf{x}}$, $b_{\varphi,\mathbf{x}}$, $c_{\varphi,\mathbf{x}}$, and $d_{\varphi,\mathbf{x}}$ that describe these Gaussians and exponentials (see equation (5.2)). The number of parameters needed to describe the matrix S is therefore so far (with n_φ being the number of projections):

$$n^{\mathrm{2D}}_{\mathrm{L}} = 8 n_\varphi N_{\mathrm{V}}. \tag{5.3}$$

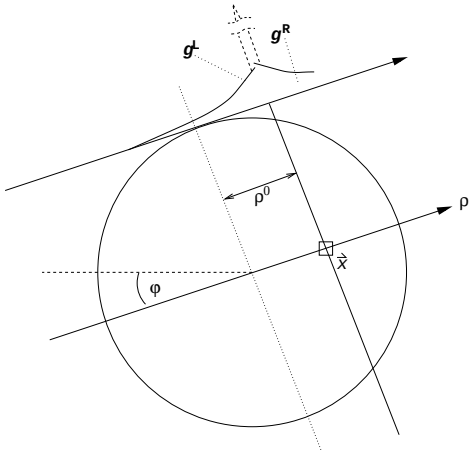


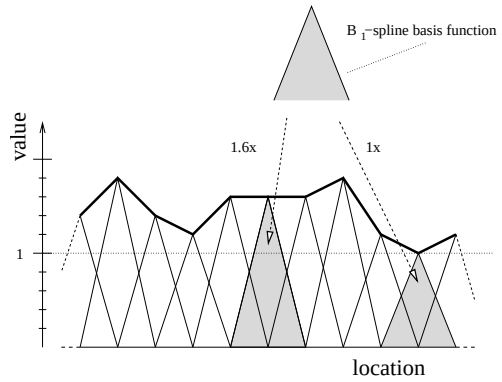
Figure 5.4.: Radon transform of the center of a voxel yields the geometrical expected maximum of a projection of the scatter. The scatter tails are approximated by the sum of an exponential and a Gaussian.

A characteristic of the system matrix described in section 5.2 provides the idea for the additional compression in voxel space. It can be assumed that the shape of the unit scatter sinograms (described by the parameters) varies more slowly from voxel to voxel than the counts in the unscattered sinograms of matrix A or the position ρ^0 of the geometrically expected maximum of the scatter. Since the latter quantities can be calculated on-the-fly, a reduction in the number of parameters leads directly to a reduction in memory consumption. This reduction is achieved by B-spline approximation of the parameters in the voxel domain. In the present implementation B_1-spline and B_2-spline curves are chosen, because these continuous functions are based on B-spline basis functions [Unser, 1999] that have the smallest support while being still continuous. The small finite support is important for a reasonable read-out speed during reconstruction. In the following, the term *kernel* is used for a multidimensional (here 2D or 3D) B-spline-basis function (see Fig.5.5). B-spline curves that are composed of fewer kernels than voxels yield the claimed reduction in memory consumption.

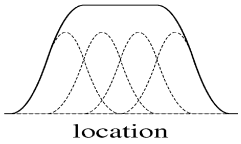
The functions $g^{\mathrm{L/R}}_{\varphi,\mathbf{x}}(\rho)$ are approximated by

$$\begin{aligned} g^{\mathrm{L/R}}_{\varphi,\mathbf{x}}(\rho) &= h^{\mathrm{L/R}}_{\varphi,\mathbf{x}}(\rho-\rho^0) \approx \tilde{h}^{\mathrm{L/R}}_{\varphi,\mathbf{x}}(\rho-\rho^0) \\ &= B_n(\mathbf{x}) \otimes \sum_{\boldsymbol{\kappa}} \bar{h}^{\mathrm{L/R}}_{\varphi,\boldsymbol{\kappa}}(\rho-\rho^0)\,\delta(\mathbf{x}-\mathbf{x}_{\boldsymbol{\kappa}})\,. \end{aligned} \tag{5.4}$$

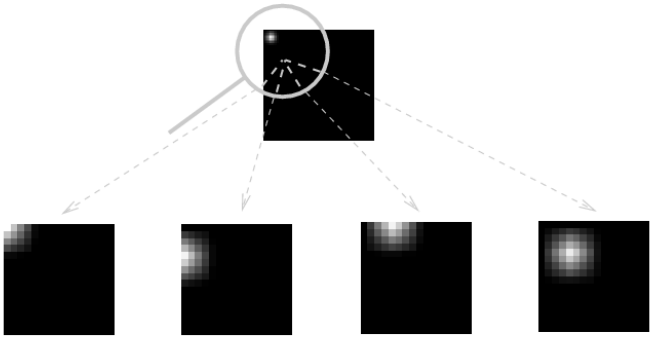
Here δ is the delta function, $\otimes$ the convolution, and $\boldsymbol{\kappa}$ is a (vector-) index specifying



(a) B_1-spline function constructed by using B_1-spline basis functions



(b) B_2-spline



(c) Discrete B_1-2D-spline kernels in the voxel domain

Figure 5.5.: B-spline as a superposition of B-spline basis functions. Low order B-spline functions (like B_1 or B_2) are evaluated by adding only few basis functions at a given location.

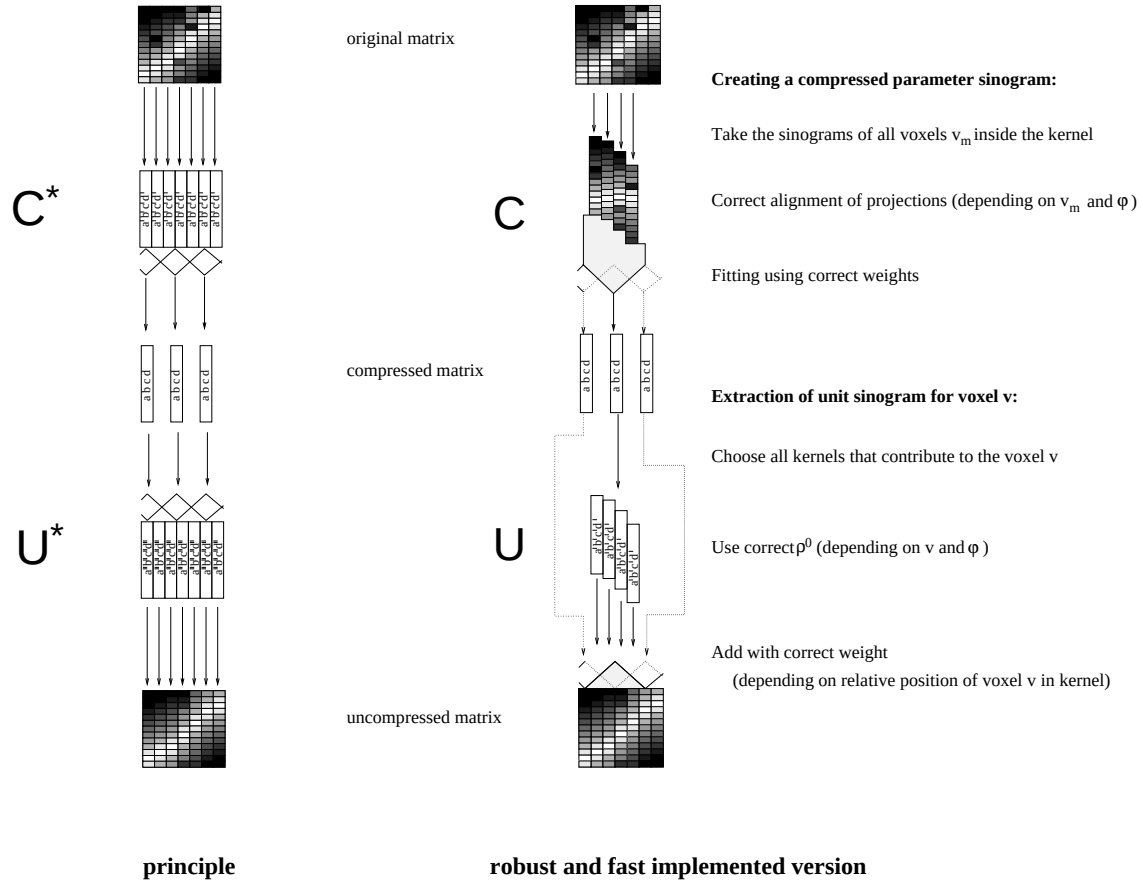


Figure 5.6.: Schematic diagram showing the compression and the decompression of the system matrix (left: principle, right: used improved scheme) and explaining the formation of a compressed parameter sinogram (upper text) and the extraction of one unit sinogram from the compressed matrix (lower text).

a B-spline node. Equation (5.4) is the convolution of a support limited kernel with a weighted comb function. For a given voxel v at position $\mathbf{x}$ this summation is reduced to few terms only (see figure 5.5). Instead of $a_{\varphi,\mathbf{x}}$, $b_{\varphi,\mathbf{x}}$, $c_{\varphi,\mathbf{x}}$, and $d_{\varphi,\mathbf{x}}$ only the parameters $a_{\varphi,\boldsymbol{\kappa}}$, $b_{\varphi,\boldsymbol{\kappa}}$, $c_{\varphi,\boldsymbol{\kappa}}$, and $d_{\varphi,\boldsymbol{\kappa}}$ that describe $\bar{h}^{\mathrm{L/R}}_{\varphi,\boldsymbol{\kappa}}(\rho-\rho^0)$ need to be stored. Since the number of B-spline nodes $n_{\boldsymbol{\kappa}}$ is smaller than the number of voxels $n_{\mathbf{v}}$, this leads to a reduction of the parameters.

These parameters must guarantee good approximate functions $\tilde{h}^{\mathrm{L/R}}_{\varphi,\mathbf{x}}(\rho-\rho^0)$. The left pictogram in Fig. 5.6 shows how this could be achieved. Firstly, the projections are parametrized. Since the parameters vary slowly, they are then approximated by a B-spline in the voxel domain. The compression of a matrix with this basic method

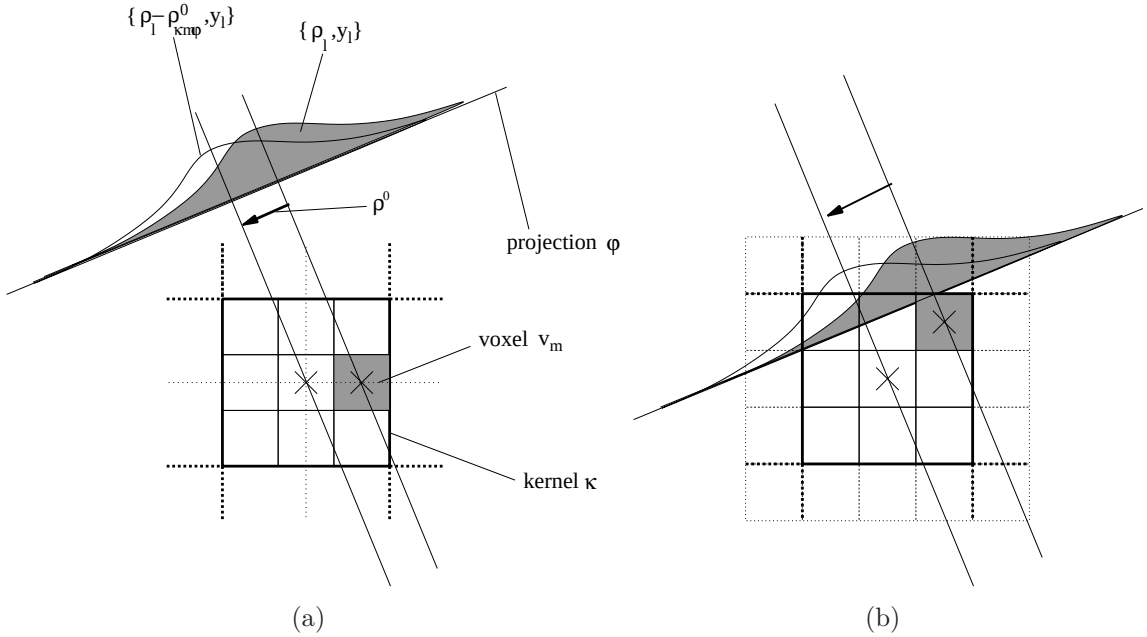


Figure 5.7.: Unit projections of single voxels inside a kernel are shifted in such a way that the geometrical expected maxima coincide.

would already result in a compressed scatter matrix of sufficiently small size. The method must however be modified in order to meet other requirements of section 5.1.

5.3.2. Increasing robustness of compression scheme

The compression consists in finding good approximate values for the as, bs, cs, and ds that define the functions $\bar{h}^{\mathrm{L/R}}_{\varphi,\boldsymbol{\kappa}}(\rho-\rho^0)$. This necessarily involves a fitting algorithm and the reduction of the voxel grid to the coarser spline node grid. The fitting algorithm can lead to unreasonable results if not enough data points or too few trustworthy data points are provided. This problem is strongly reduced by collecting and aligning (Fig. 5.7) the unit scatter sinograms of the voxels for a given kernel and using these collected data for fitting. Let the set

$$\omega_{v,\varphi} = \{(\rho_l; y_{v,\varphi,l})\}_l \tag{5.5}$$

represent a discrete (arc corrected) unit scatter projection at angle φ of voxel v. The pair $(\rho_l; y_{v,\varphi,l})$ represents the position ρ_l and value $y_{v,\varphi,l}$ of bin l of such a projection. Instead of using only single projections as input for the fitting algorithm, projections at angle φ of all voxels $v_m = v_1, \ldots, v_{n_{\boldsymbol{\kappa}}}$ that are located inside kernel $\boldsymbol{\kappa}$ are aligned (shifted by the geometrically expected maximum $\rho^0_{v_m,\varphi}$ of voxel m) and collected.

This collection $\Omega_{\boldsymbol{\kappa},\varphi}$ of data (a "collected projection") is then processed by the fitting algorithm.

$$\Omega_{\boldsymbol{\kappa},\varphi} = \{(r_{\boldsymbol{\kappa},\varphi,\alpha}; y_{\boldsymbol{\kappa},\varphi,\alpha})\}_\alpha \equiv \bigcup_{m=1}^{n_{\boldsymbol{\kappa}}} \{(\rho_l - \rho^0_{v_m,\varphi}; y_{v_m,\varphi,l})\}_l \tag{5.6}$$

Here, a new index $\alpha \equiv l n_{\boldsymbol{\kappa}} + m$ is introduced. Equation (5.6) shows such a "collected projection". The available information for fitting is strongly increased. This allows stable fitting even if the unit scatter sinogram of single voxels are extremely noisy. A drawback is however the large amount of points which results in very long fitting procedures if the data is not preprocessed.

5.3.3. Increasing compression speed

Even though this large increase of available data points facilitates stable fitting for matrices of very bad statistics, the high number of available points leads to a very long fitting procedure. A strong acceleration was achieved by grouping the collected data points of a projection into intervals $I^{(t)} = [\rho^{(t-1)}, \rho^{(t)}[$ before fitting (see Fig. 5.8). Each interval was represented by a point at weighted mean position $\rho^{(t)}$ with the weighted mean number of counts $y^{(t)}$ and a weighted mean "standard deviation" $\sigma^{(t)}$.

$$\begin{aligned} \Sigma^{(t)}_{\boldsymbol{\kappa},\varphi} &= \sum_{\alpha\in G^{(t)}} B_n(\mathbf{x}_m - \mathbf{x}_{\boldsymbol{\kappa}}); \quad \rho^{(t)}_{\boldsymbol{\kappa},\varphi} = \frac{1}{\Sigma^{(t)}_{\boldsymbol{\kappa},\varphi}} \sum_{\alpha\in G^{(t)}} B_n(\mathbf{x}_m - \mathbf{x}_{\boldsymbol{\kappa}})\rho_{\alpha,\varphi} \\ y^{(t)}_{\boldsymbol{\kappa},\varphi} &= \frac{1}{\Sigma^{(t)}_{\boldsymbol{\kappa},\varphi}} \sum_{\alpha\in G^{(t)}} B_n(\mathbf{x}_m - \mathbf{x}_{\boldsymbol{\kappa}}) y_{\alpha,\varphi} \quad ; \quad \sigma^{(t)}_{\boldsymbol{\kappa},\varphi} = \frac{1}{\sqrt{y^{(t)}_{\boldsymbol{\kappa},\varphi}\Sigma^{(t)}_{\boldsymbol{\kappa},\varphi}}} \end{aligned} \tag{5.7}$$

Here $G^{(t)}$ is the set of all points that are located in the interval $I^{(t)}$. It is important to notice that $\sigma^{(t)}_{\boldsymbol{\kappa},\varphi}$ is not a standard deviation but a heuristic way to account for the importance of a LOR by using information about the location relative to the kernel $\boldsymbol{\kappa}$ and the number of the detected counts $y^{(t)}_{\boldsymbol{\kappa},\varphi}$. The intervals $I^{(t)}$ close to $\rho = 0$ were chosen to be smaller than the intervals at large $|\rho|$. In this way it was possible to obtain a good fit close to the steeper and more important maximum at $\rho = 0$ while the number of points is reduced considerably. The borders of the intervals $I^{(t)}$ were chosen to be

$$\rho^{(t)} \propto -\ln\left(1 - \frac{t}{t_{\max}}\right) \qquad \text{with} \quad t \in \mathbb{N}_0 \wedge t < t_{\max}\,. \tag{5.8}$$

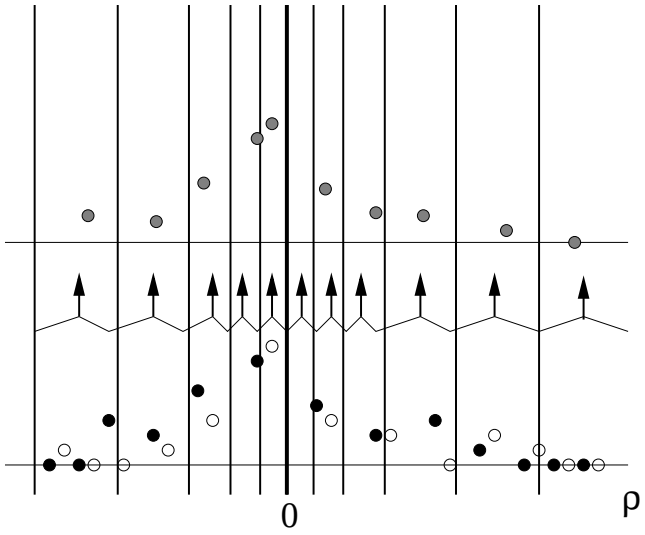


Figure 5.8.: Pictogram showing the reduction of points to be fitted by grouping them into intervals. In the lower part an example of two projections (one black, the other white) is shown. The projections are aligned in such a way that the geometrical expected maxima coincide (at $\rho = 0$). For each interval a mean point is calculated (upper part, gray circles).

The proportionality factor was chosen so that around $t_{\max} = 15$ to 25 points were used to be fitted. The fitting was performed by the Levenberg-Marquardt algorithm using the GNU scientific library [Galassi et al., 2006]. The fewer points, the faster the Levenberg-Marquardt algorithm [Press et al., 2002]. A minimum of points should be used in order to correctly describe the scatter.

5.3.4. Read-out

The extraction of the unit-sinograms from the compressed matrix (see Fig. 5.6U) was achieved by evaluating the parametrized and shifted unit projections of all kernels $\boldsymbol{\kappa}$ that cover the voxel $\mathbf{x}$ in question. These evaluated projections then are weighted according to the position of the voxel inside the respective kernel and added to form the extracted unit-projection (see equation (5.4)). All projections of the voxel then form the unit-sinogram or column of the matrix.

5.3.5. Memory saving for 2D scanners and outlook for 3D-scanner matrix compression

Using the described compression scheme, in total only

$$n_{\mathrm{LV}}^{\mathrm{2D}} = 8 n_{\varphi} n_{\mathrm{x}} n_{\mathrm{y}} \tag{5.9}$$

parameters have to be stored. Here $n_{\mathrm{x}} n_{\mathrm{y}}$ is the number of B-spline kernels needed to describe the 2D B-spline function. For a n_{R}-ring 3D scanner with span=1 and maximum ring difference RD=$n_{\mathrm{R}} - 1$ the number of parameters would be

$$n_{\mathrm{LV}}^{\mathrm{3D}} = 4 n_{\varphi} n_{\mathrm{x}} n_{\mathrm{y}} n_{\mathrm{z}} (n_{\mathrm{R}} + 1) n_{\mathrm{R}} \tag{5.10}$$

under the reasonable assumption that the same or a very similar compression approach can be used for oblique sinograms (for the definition of *span* or *maximum ring difference* see for example [Kehren, 2001, Fahey, 2002]). In contrast a discrete non-spare storage of the matrix S would require the storage of

$$N = \frac{1}{2} n_{\varphi} n_{\mathrm{bins}} N_{\mathrm{V}} (n_{\mathrm{R}} + 1) n_{\mathrm{R}}, \tag{5.11}$$

values. This corresponds to a compression ratio of

$$\mathrm{ratio} = \frac{8 n_x n_y n_z}{n_{\mathrm{bins}} N_{\mathrm{V}}} \,. \tag{5.12}$$

A significant reduction is therefore possible ($N_{\mathrm{V}} \gg n_{\mathrm{x}} n_{\mathrm{y}} n_{\mathrm{z}}$, $n_{\mathrm{bins}} \gg 8$).

In 3D scanners there can in addition be oblique unit-sinograms and it is possible that unit-sinograms have *no* direct coincidences. This might require a change of the functions that describe such scatter projections, especially the geometrically expected maximum has to be calculated differently for oblique sinograms. But the principle of fitting of scatter projections and B-spline compression in the voxel domain should be still feasible. In addition, some of these unit-sinograms might have so few counts that they can be set to zero (see Fig. 5.9).

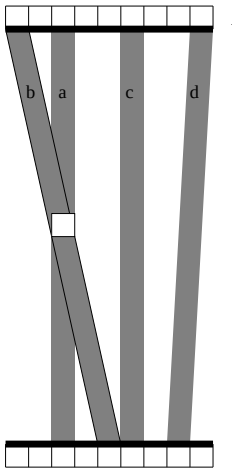


Figure 5.9.: Possible unit-sinograms for a voxel of a 3D scanner. (a) transversal sinogram, (b) oblique sinogram with direct coincidences and (c) and (d) transversal and oblique sinograms without direct coincidences. Depending on the distance of the latter to the voxel, they might be neglected.

6. Implemented Reconstruction Algorithms

All implemented reconstruction algorithms are based on the maximum likelihood expectation maximization. This algorithm was chosen because it is well understood and widely used. Accelerated (but not necessarily convergent) versions of it like ordered subset expectation maximization are used in clinical scanners. Accelerated version of ML-EM were not used in this work, however, because often they are not guaranteed to converge and because the correct physical modeling of the scatter and not the convergence speed should be investigated. All reconstruction algorithms used a uniform starting image with 1's in all voxels if not otherwise mentioned.

6.1. Monte Carlo maximum likelihood expectation maximization

6.1.1. Full matrix

The full matrix approach is the most straightforward way to include scatter into the reconstruction process. This is achieved by simulating uniform activity within voxels and storing the obtained coincidences in the columns of the full matrix M and using this matrix directly in the ML-EM reconstruction algorithm. The matrix M was not stored in a sparse manner, because matrices with a higher number of simulated emissions anyhow had a substantial non-zero fraction. The equations

$$\begin{aligned} \mathcal{P}_{\text{full}}: \quad & \mathbf{y}^{(k+1)} = M\mathbf{x}^{(k)} \\ \mathcal{B}_{\text{full}}: \quad & x_i^{(k+1)} = x_i^{(k)} \sum_{j=1}^{N_{\text{L}}} \left(\frac{y_j^{\star} m_{ji}}{y_j^{(k+1)}} \right) . \end{aligned} \tag{6.1}$$

describe this algorithm. Here k stands for the iteration number, $\mathbf{y}$ is the measured

(or in this case the simulated) sinogram, and M and m_{ji} the matrix and the matrix elements respectively. The algorithm is divided into a projector $\mathcal{P}_{\text{full}}$ and a back-projector $\mathcal{B}_{\text{full}}$ as described in section 3.2. The full matrix approach can only be used in problems of reduced size like in the considered proof-of-principle single ring scanner simulations. Otherwise the storage of the matrix would be impossible.

6.1.2. No scatter modeling

The same algorithm like in section 6.1.1 was used to investigate the performance of ML-EM reconstruction without consideration of scatter. In that case the matrix M in (6.1) was simply replaced by the scatter-free matrix A. The images that were reconstructed using this algorithm were mostly used to identify and compare regions where artifacts due to the incorrect scatter treatment occur.

6.1.3. Compressed matrix

Other images were reconstructed by using the compressed scatter matrix $\hat{S}$. The principle of this algorithm is shown in (6.2). Here U is the decompression operator.

$$\begin{aligned} \mathcal{P}_{\mathsf{UC}}: \quad & \mathbf{y}^{(k+1)} = (A + (\mathsf{U}\hat{S}))(\mathbf{x}^{(k)}) \\ \mathcal{B}_{\mathsf{UC}}: \quad & x_i^{(k+1)} = x_i^{(k)} \sum_{j=1}^{N_{\text{L}}} \left(\frac{y_j^{\star}(A + (\mathsf{U}\hat{S}))_{ji}}{y_j^{(k+1)}} \right) . \end{aligned} \tag{6.2}$$

The reconstruction was *not* performed by uncompressing the whole matrix at once which (6.2) could suggest (and which would be not possible for larger problems). The matrix was uncompressed in a column-wise manner. Equation (6.3) describes the implementation.

$$\begin{aligned} \mathbf{y}^{(n)} &= \text{MULT}(A, \hat{S}; \mathbf{x}^{(n)}) \\ y_j'^{(n)} &= \frac{y_j^{\star}}{y_j^{(n)}} \\ \mathbf{f}^{(n)} &= \text{TMULT}(\mathbf{y}'^{(n)}; A, \hat{S}) \\ x_i^{(n+1)} &= x_i^{(n)} f_i^{(n)} \end{aligned} \tag{6.3}$$

The first line of (6.3) describes the projector and the three other lines the back-

projector. The operators MULT and TMULT that represent matrix and transverse matrix multiplication with the compressed matrix $(A, \hat{S})$ are defined as follows:

$$\mathbf{Y} = \mathrm{MULT}(A, \hat{S}; \mathbf{x}): \quad \mathbf{Y} = \sum_i x_i \left(\mathbf{col}_i(A) + \mathbf{COL}_i(\hat{S}) \right) \tag{6.4}$$

$$\mathbf{X} = \mathrm{TMULT}(\mathbf{y}; A, \hat{S}): \quad [\mathbf{X}]_i = \left\langle \left(\mathbf{col}_i(A) + \mathbf{COL}_i(\hat{S}) \right), \mathbf{y} \right\rangle \tag{6.5}$$

Here $\mathbf{COL}_i(\hat{S})$ is the extracted (and uncompressed) column of compressed matrix $\hat{S}$ for voxel i, $\mathbf{col}_i(A)$ the ith column of A, and $\langle \cdot, \cdot \rangle$ the inner product. The expected geometrical maximum can be calculated by using the voxel index i (and therefore the position of the voxel). In this way the projection can be aligned correctly before added (see also section 5.3.4). In the compressed matrix algorithm only voxels of non-zero density were used.

6.2. Dual matrix maximum likelihood expectation maximization

Dual matrix expectation maximization was also implemented. This algorithm uses different physics in the projector and in the back-projector. The back-projector is based solely on the scatter-free matrix A alone and does therefore not include scatter. The projector does model scatter. Scatter is incorporated in the projector by running a Monte Carlo simulation at each iteration.

$$\begin{aligned} \mathcal{P}': \quad & \mathbf{y}^{(k+1)} = \mathbf{MC}'(\mathbf{x}^{(k)}) \\ \mathcal{B}_{\mathrm{DM}}: \quad & x_i^{(k+1)} = x_i^{(k)} \sum_{j=1}^{N_{\mathrm{L}}} \left(\frac{y_j^\star a_{ji}}{y_j^{(k+1)}} \right) \end{aligned} \tag{6.6}$$

There is no need to store a matrix including scatter, but the algorithm is not guaranteed anymore to converge. The algorithm (6.6) is also not ideal from a numerical point of view. Especially for low count simulations, it is likely that the denominator $y_j^{(k+1)}$ (which is obtained by Monte Carlo simulation $[\mathbf{MC}'(\mathbf{x}^{(k)})]_j$) becomes zero for given a index j, but the nominator $y_j^\star a_{ji}$ is greater than zero. This is very likely especially for LORs tangential to the phantom boundary. Simulations confirmed this effect and therefore this algorithm was not further used. Instead, a slightly different algorithm was used. The problem of a zero denominator was avoided by introducing a projector

that uses the scatter-free matrix A.

$$\begin{aligned} \mathcal{P}_{\mathrm{DM}}: \quad & \mathbf{y}^{(k+1)} = A\mathbf{x}^{(k)} + \mathbf{s}^{(k)} \\ \mathcal{B}_{\mathrm{DM}}: \quad & x_i^{(k+1)} = x_i^{(k)} \sum_{j=1}^{N_{\mathrm{L}}} \left(\frac{y_j^{\star} a_{ji}}{y_j^{(k+1)}} \right) \end{aligned} \tag{6.7}$$

$$\text{with } [A]_{ji} \equiv a_{ji}$$

Here the sinogram $\mathbf{s}^{(k)} = \mathbf{MC}_{\mathrm{scatter}}(\mathbf{x}^{(k)})$ is calculated by a scatter-only simulation. Like (6.6), (6.7) is still not guaranteed to converge, but the numerical problem of the denominator becoming zero is avoided. The problem of convergence is based on the relatively unpredictable behavior of the algorithm. Since the projector is changed at every iteration step due to the Monte Carlo simulation, the progress made in one iteration can be canceled partially in the next iteration. A similar problem arises from the fact that projector and back-projector differ in terms of scatter, therefore not leading to optimal search directions.

The first iteration was performed with $s^{(0)} \equiv 0$ and a first guess of the activity and the total number of emissions was obtained. Then a fixed fraction p of the guessed emissions $x_i^{(k)}$ were simulated in the following iterations. The obtained simulated scatter sinograms $s^{(k)}$ were scaled by a factor $1/p$ in order to correct for the lower number of simulated emissions.

6.3. A hybrid approach

In order to show the potential a stored compressed matrix, an algorithm that is a mixture of the dual matrix and compressed matrix algorithm is introduced:

$$\begin{aligned} \mathbf{y}^{(k+1)} &= c_1 (A + S^{\mathrm{comp}})\, \mathbf{x}^{(k)} + c_2 (A\mathbf{x}^{(k)} + \mathbf{s}^{(k)}) \\ x_i^{(k+1)} &= x_i^{(k)} \sum_{j=1}^{N_{\mathrm{L}}} \left(\frac{y_j^{\star} (a_{ji} + s_{ji}^{\mathrm{comp}})}{y_j^{(k+1)}} \right) \end{aligned} \tag{6.8}$$

Here c_1, c_2 are constant parameters that must fulfill $c_1 + c_2 = 1$. Since this algorithm was only introduced to show the additional flexibility which is obtained by the compressed matrix, values $c_1 = c_2 = 0.5$ were chosen ad-hoc. Better values might be found. Like in the dual matrix algorithm, in this algorithm the voxel index is only

running over non-zero density voxels and the scatter sinogram $\mathbf{s}^{(k)}$ was simulated using a fraction p of the emission density $\mathbf{x}^{(k)}$.

7. Evaluation

7.1. Simulated phantoms and scanner geometries

In the simulations, different phantoms and scanner geometries were used. As a phantom either an inhomogeneous cylindrical (phantom A and A', Fig. 7.1, Table 7.1) or a inhomogeneous generalized cylinder of elliptic cross section (phantom B, Fig. 7.2, Table 7.1) was used in the simulations. The cylinder was placed in the center of the scanner, but the elliptic cylinder was placed off-centrally. Directly at the border of the phantoms intermediate density/activity voxel values were used in order to better approximate the cylindrical shape (like the partial volume effect).

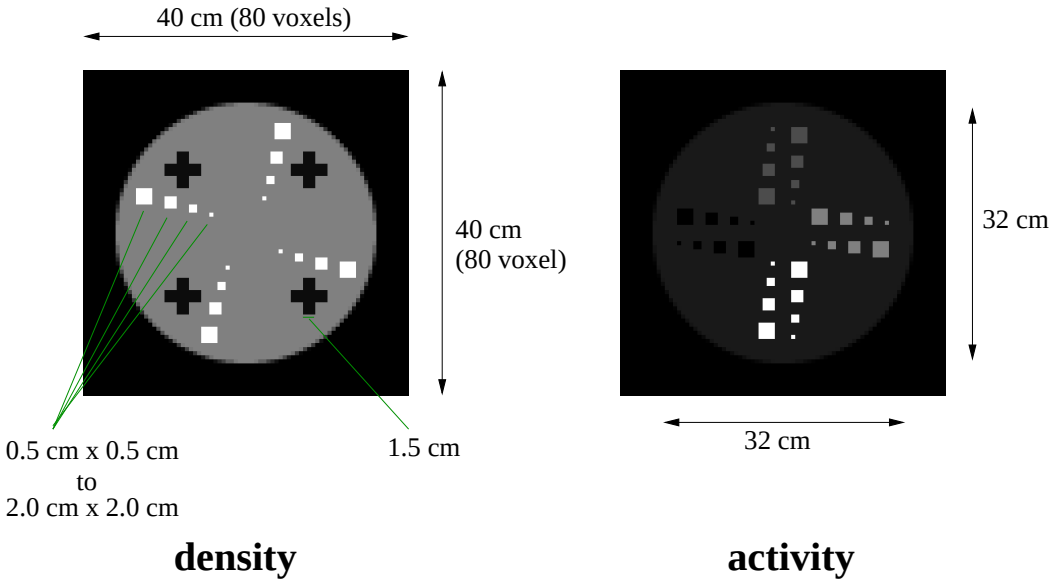


Figure 7.1.: Phantom A[A']. Density: 0(outside) : 0.1[0.26](crosses) : 1(cylinder) : 2[1.46](spots) g/cm^3. Activity ratio: 0(outside,spots) : 1(cylinder) : 3(spots) : 5(spots) : 10(spots).

The density and the activity grid were always arranged in such a way that the center of the scanner, of the density grid, and of the activity grid coincided. Phantom A'

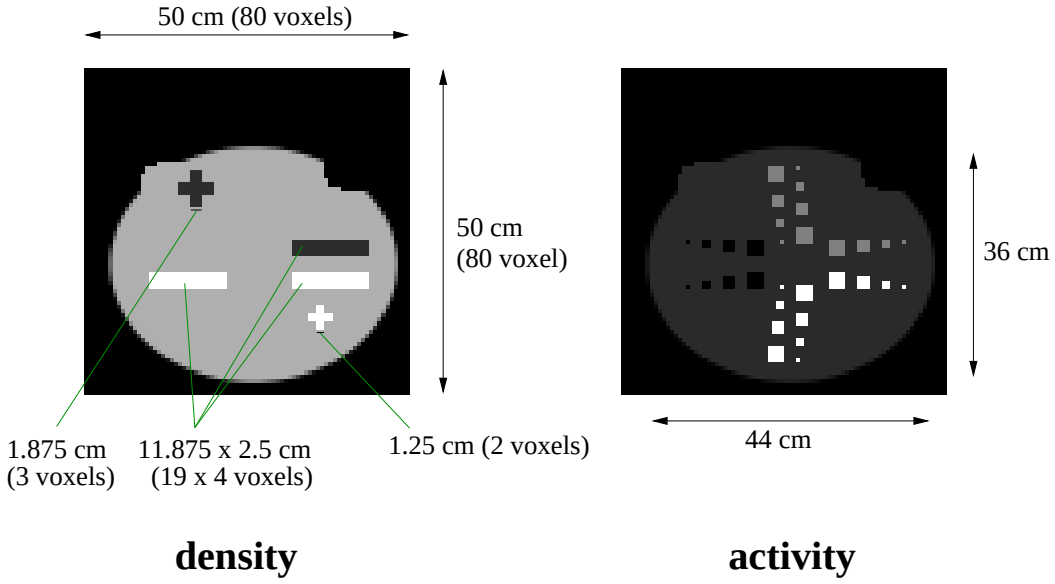


Figure 7.2.: Phantom B. Density: 0(outside) : 0.26(crosses) : 1(cylinder) : 1.46(spots) g/cm^3, activity ratio: 0(outside,spots) : 1(cylinder) : 3(spots) : 6(spots)

phantom	# of voxels	voxel size (density)	voxel size (activity)	⊘
A	$80 \times 80 \times 1$	$5 \times 5 \times 100\,\mathrm{mm}^3$	$5 \times 5 \times 6.45\,\mathrm{mm}^3$	32 cm
A'	$80 \times 80 \times 1$	$5 \times 5 \times 100\,\mathrm{mm}^3$	$5 \times 5 \times 6.45\,\mathrm{mm}^3$	32 cm
B	$80 \times 80 \times 1$	$6.25 \times 6.25 \times 100\,\mathrm{mm}^3$	$6.25 \times 6.25 \times 6.45\,\mathrm{mm}^3$	44/36 cm

Table 7.1.: Phantom dimensions.

scanner	# of rings	# of detectors	⊘	detector ring depth
a	1	384	82.4 cm	0.645 cm
b	1	384	82.4 cm	10 cm

Table 7.2.: Scanner dimensions

setup	phantom	scanner	# of projection bins	# of projection angles
Aa	A	a	95/96	384
A'b	A'	b	95/96	384
Ba	B	a	135/136	384
Bb	B	b	135/136	384

Table 7.3.: Simulated geometries.

differed from phantom A only in terms of density. The density in A' was chosen in such a way that it matched densities of bone and lung as defined in GEANT4. Two single ring scanners were simulated. Scanner 'a' was an idealized two-dimensional scanner with a detector ring depth of 0.645 cm, scanner 'b' was a scanner with a large detector ring depth of 10 cm (see Table 7.2). The latter scanner was simulated in order to investigate scanners with large scatter fraction or large span. Multiple ring scanners were not simulated in order to allow the direct storage of the full matrix. The number of detectors per ring was always 384 and the number of voxels was $80 \times 80 \times 1$. The dimension of the grid describing the density and the grid describing the activity distribution differed for phantom B (see Table 7.1). The number of projection bins was larger for phantom B, because more bins were needed to cover the larger field of view.

7.2. Measures used for quantification

The outcome of the simulations and reconstructions is multidimensional (sinograms and images). These outcomes should be compared with reference sinograms or images. This can be accomplished by using a metric. The outcome and the reference can both be represented by a vector ($\mathbf{v}$ and $\mathbf{v}^{\text{ref}}$ respectively). With this convention the following metric can be defined:

$$(\mathcal{R}^K, \mathcal{R}^K) \longrightarrow \mathcal{R} \tag{7.1}$$

$$(\mathbf{v}, \mathbf{v}^{\text{ref}}) \longrightarrow \text{NRMSE}(\mathbf{v}, \mathbf{v}^{\text{ref}}) \equiv \frac{1}{N_{\text{E}}(\mathbf{v}^{\text{ref}})} \sqrt{\frac{1}{K} \sum_{i=1}^{K} (v_i - v_i^{\text{ref}})^2} \tag{7.2}$$

$$\text{with} \quad N_{\text{E}}(\mathbf{v}^{\text{ref}}) = \frac{1}{K} \sum_{i=1}^{K} v_i^{\text{ref}} \tag{7.3}$$

This metric NRMSE is called normalized root mean squared error and is applied to reconstructed images and simulated sinograms. Whenever images are compared, this metric is called xNRMSE and the expression sNRMSE is used to make clear that sinograms are evaluated. A similar measure was used when there was the need to

quantify the variance of a set of images $\{\mathbf{x}_\alpha\}$ (represented as a set of vectors).

$$(\underbrace{\mathcal{R}^K, \cdots, \mathcal{R}^K}_{N_\alpha \text{ times}}; \mathcal{R}^K) \longrightarrow \mathcal{R} \tag{7.4}$$

$$(\{\mathbf{x}_\alpha\}; \mathbf{x}^{\text{true}}) \longrightarrow \text{NRMV}\left(\{\mathbf{x}_\alpha\}; \mathbf{x}^{\text{true}}\right) \equiv \frac{1}{N_{\text{E}}(\mathbf{x}^{\text{true}})} \sqrt{\frac{1}{K} \sum_{i=1}^{K} \sigma_i^2} \tag{7.5}$$

$$\text{with} \quad \sigma_i^2 = \frac{1}{N_\alpha - 1} \sum_{\alpha=1}^{N_\alpha} (x_{i,\alpha} - \bar{x}_i)^2 \tag{7.6}$$

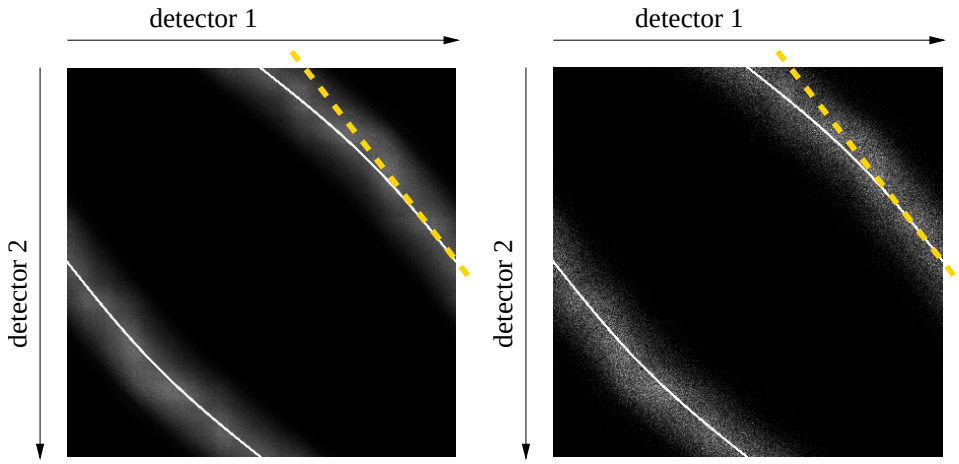
$$\bar{x}_i = \frac{1}{N_\alpha} \sum_{\alpha=1}^{N_\alpha} x_{i,\alpha} \tag{7.7}$$

The true image $\mathbf{x}^{\text{true}}$ is used to scale the NRMV (normalized root mean variance). Apart from these metrics gray value images and profiles are used to give further insight into the local quality of the outcomes.

7.3. Verification of the Monte Carlo code

The Monte Carlo code (chapter 4) was compared against GEANT4 simulations using setup A'b. The large detector ring depth was used because of the slow performance of GEANT4. In both codes idealized detectors (no Gaussian filtering) were used and in both cases Rayleigh scattering was neglected. In this way it was possible to check the simulation of attenuation and scatter in the phantom/patient using the same physics.

In the GEANT4 simulations 10^8 positrons of zero kinetic energy were created. In the YaPRA simulation 10^6 photon pairs were simulated using variance reduction techniques. Both sinograms were then normalized and compared. The total number of detected coincidences as well as unscattered coincidences agreed both within 1%. In Fig. 7.3 it can be seen that the Monte Carlo code YaPRA describes correctly the geometric effects. Fig. 7.4 shows the quantitative agreement between the two codes. Due to the variance reduction techniques, Monte Carlo noise is suppressed stronger in the YaPRA simulations (left part of the LOR profile). In the right part of Fig. 7.4 (LOR profile bin numbers of 160 or higher) there is a good agreement between the results of the YaPRA and the GEANT4 Monte Carlo code. That region includes coincidences of unscattered photons. The oscillations occur, because sometimes only parts of the detectors can be reached by photons without being scattered and sometimes the whole detector can be reached.



(a) YaPRA (b) GEANT4

Figure 7.3.: Simulated detected coincidences using YaPRA and GEANT4 of an off-central single voxel with activity. White represents large values and black small values. The same gray value scaling was applied to both figures in order to show scattered coincidences. The dashed lines mark the profiles shown in Fig. 7.4.

7.4. Compressed matrix

In this section the reconstruction of images using the proposed compressed matrix is evaluated. The evaluation can be divided into three parts. The first part comprises the comparison of the compressed matrix with an uncompressed reference matrix. This reference matrix cannot be calculated exactly. Therefore, a matrix simulated with a very high number of photon pairs is used instead. In the second part of the evaluation, images reconstructed using the compressed and using the high statistic matrix are compared. Further, images reconstructed with other methods like the dual matrix approach or with the introduced hybrid method were evaluated for comparison. In the last part results with modified compression parameters are presented.

All system matrices were calculated using Monte Carlo simulation with the variance reduction techniques stratification and forced detection. Scatter matrices S^{ref} with $1,280,000$ simulated emissions per voxel were used as the reference scatter matrix. The low count scatter matrices S^{orig} with $40,000$ simulated emissions per voxel were compressed to the compressed matrices S^{comp}. The compression was achieved by using $10 \times 10 \times 1$ B_1-spline-kernels. In this way a total reduction of the storage size of the scatter matrices by a factor of 1076 could be achieved (384 projections with 134 or 135

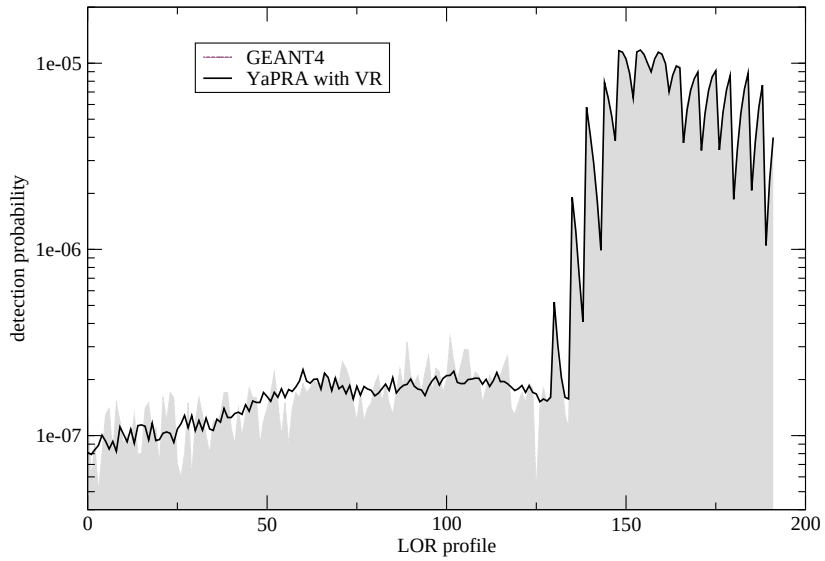


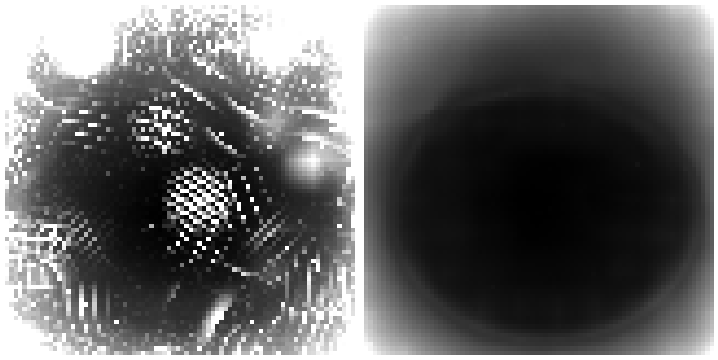
Figure 7.4.: Profiles through the histograms of Fig. 7.3. In regions where only scatter coincidences occur the Monte Carlo code YaPRA suppresses Monte Carlo noise (left part of the graph). In the region where direct (unscattered) coincidence occur, the probability of detection follows the GEANT4 simulation in good agreement.

bins, no interleaving). For the reconstruction also a high statistic scatter-free matrix A with $10,240,000$ simulated emission per voxel was calculated. This matrix A was used in the back-projector of the dual matrix approach and the matrix $M = A + S^{\text{ref}}$ was used in the full matrix approach. Two setups were simulated. In one setup the phantom was placed in a single ring scanner with a small ring detector depth (setup Ba) and in the other in a single ring scanner with a large detector ring depth (setup Bb). Both setups are described in section 7.1.

7.4.1. Comparison of full matrix and compressed matrix

The sNRMSE was used to compare directly the compressed matrix S^{comp} and the uncompressed reference matrix S^{ref}. For each column i of the matrix the $\text{sNRMSE}_i = \text{sNRMSE}(\mathbf{s}_i^{\text{comp}}, \mathbf{s}_i^{\text{ref}})$ was calculated. Here $\mathbf{s}_i^{\text{comp}}$, $\mathbf{s}_i^{\text{ref}}$ are the column i of S^{comp} and S^{ref}, respectively. Because each column of the matrices represents the sinogram of a single voxel, gray value images can be created that are meaningful. The gray value of a pixel represents the sNRMSE of the corresponding column of the matrix. In this way it was possible to get a good overview over the local quality of the compressed

matrices. These sNRMSE images can be seen in Fig. 7.5. Small values are represented

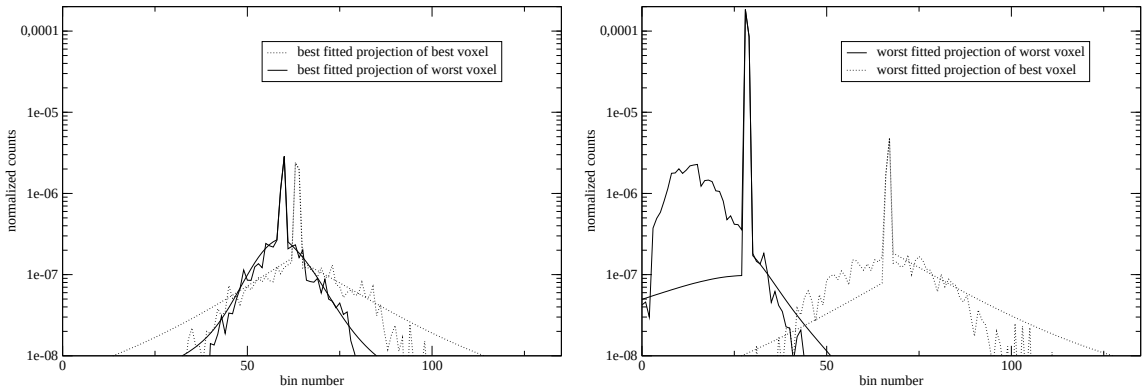


(a) setup Ba (b) setup Bb

Figure 7.5.: sNRMSE images for setup Ba and Bb. Black corresponds to a small and white to a large error. In Fig. 7.5(a) the size of the B-spline kernel can be seen. When the values for a kernel are not well fitted, the whole area appears white.

by black color and large values by white color. Due to the small amount of scatter in setup Ba, there the fitting did sometimes not perform ideally (circular white regions in Fig. 7.5(a) or white noise-like areas). The circular white regions can be explained by bad parameters of a kernel. This mainly occurred for zero-density voxels (outside the phantom) that were not used for reconstruction. The white noise-like regions are probably caused by projections that are fitted by too steep Gaussians. This can lead to aliasing effects which, however, did not influence the reconstructed images due to the small scatter fraction. The larger scatter fraction of setup Bb is the reason that the simulated matrix S^{orig} for this setup is less noisy. This is the explanation for the relatively homogeneous sNRMSE image (see Fig. 7.5(b)). Especially for voxels inside the phantom boundary this is the case. Almost no spline grid effect can be seen.

Fig. 7.6 provides insight into the quality of the fits. The column with the smallest sNRMSE and the column with the largest sNRMSE are chosen and the best and worst fitted projection of each column are shown. It should be stressed that the fits to the low statistics matrix S^{orig} (40,000 emissions per voxel) are compared to the high statistics reference matrix S^{ref} (1,280,000 emissions per voxel). Although there can be stronger deviations at the border of the phantom (Fig. 7.6(b)) the fits to this low statistics matrix are quite stable. The deviation in Fig. 7.6(b) (worst voxel) can be explained by the used fitting algorithm. Rising scatter tails were suppressed during the Levenberg-Marquardt iteration by setting the parameters in the exponent



(a) Best fitted projection

(b) Worst fitted projection

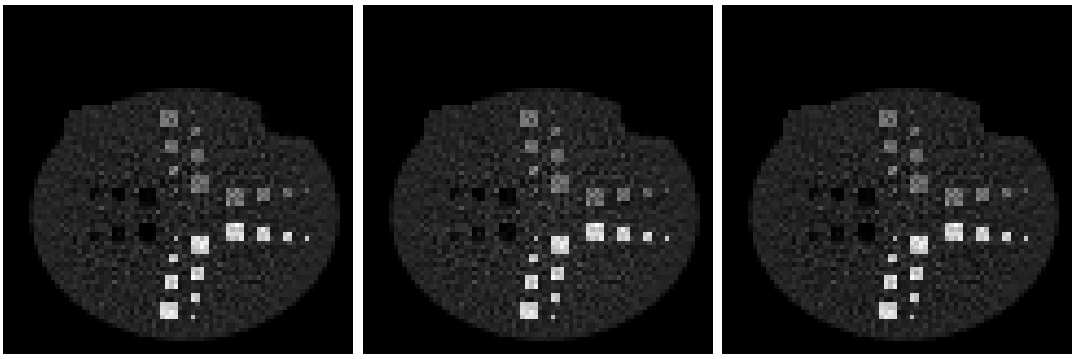
Figure 7.6.: Comparison of projections of $A + S^{\text{comp}}$ with projections of M for setup Bb. Projections of voxels with smallest sNRMSE (in the middle of the phantom) and largest sNRMSE (at the border of the phantom) are shown. The quality of the fit was quantified by the sum of the squared errors for each bin. This measure was used to find the best and worst projection.

of the Gaussian and the exponential to large negative values in order to avoid a scatter maximum that does not coincide with the expected maximum. This means that if the fitting fails totally, the compressed matrix is approximated by matrix A (no scatter). At the border of the phantom this might happen (Fig. 7.6(b)), but it has no visible influence on the reconstructed image.

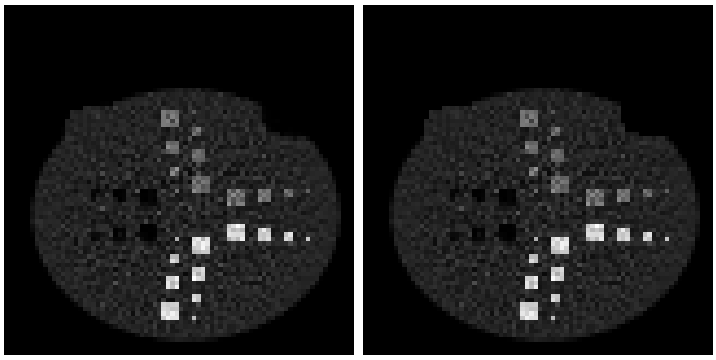
7.4.2. Comparison of reconstructed images

The images were reconstructed using the reconstruction algorithms that are introduced in chapter 6. Each algorithm was iterated 500 times. In the dual-matrix approach roughly 3.2×10^6 emissions were simulated in the projector (in the hybrid approach 1.6×10^6 emissions). The emissions for the dual matrix approach was chosen in such a way that the reconstruction time matched approximately the compressed matrix method. The sinograms $\mathbf{y}_j^\star$ were simulated using 6.4×10^{10} and 3.2×10^{10} emissions for setup Ba and setup Bb respectively. Due to this large number of simulated particles the dependency of the quality of the reconstructed images on the noise in the sinogram could be strongly reduced. The simulations of the sinograms did not make use of variance reduction techniques.

Fig. 7.7 shows the reconstructed images that were obtained after 500 iterations of the respective reconstruction methods for setup Ba. Due to the small scatter fraction of this setup the images cannot be distinguished by visual impression. Even without



(a) M (b) A (c) compressed

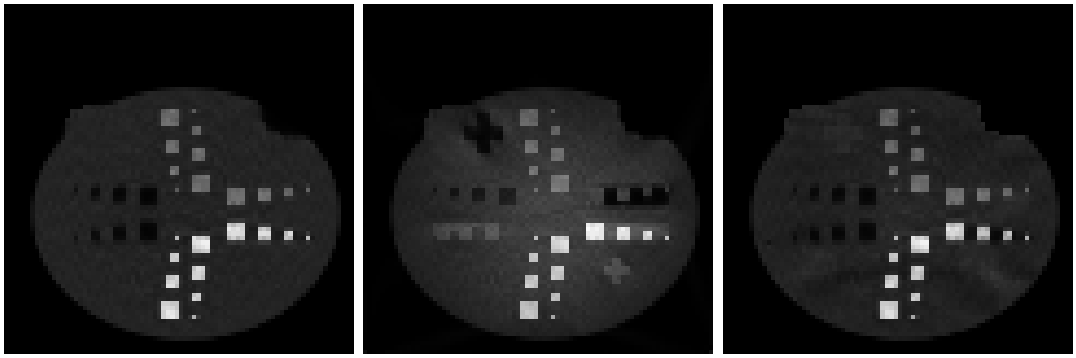


(d) dual matrix (e) hybrid

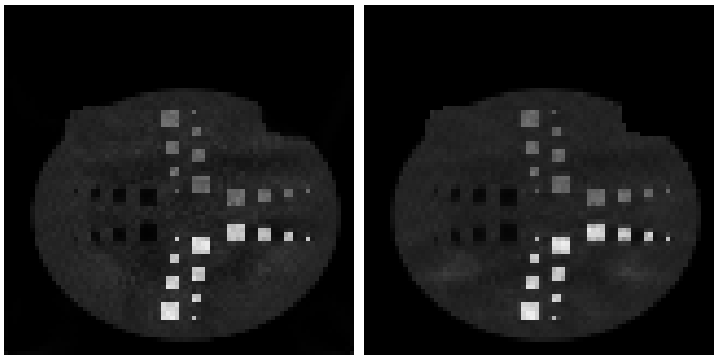
Figure 7.7.: Reconstructed images for setup Ba (small detector ring depth) using different reconstruction algorithms at iteration step 500.

scatter treatment (Fig. 7.7(b)) there are no visible artifacts. This is not the case for setup Bb (Fig. 7.8). The images that are reconstructed solely using matrix A (Fig. 7.8(b)), show severe artifacts in regions where the density differs from the background density of the cylinder. Those artifacts are largely suppressed when using the compressed matrix (Fig. 7.8(c)), dual matrix (Fig. 7.8(d)), or hybrid (Fig. 7.8(e)) approach. The full matrix approach (Fig. 7.8(a)) was considered to be the ideal case.

In Fig. 7.9(a), Fig. 7.9(b), and Fig. 7.9(c) diagonal profiles through the reconstructed images are shown. While for setup Ba all reconstruction methods results in practically the same images, there are some small deviations in the images of setup Bb. The profile of Fig. 7.8(b) is also shown to mark the position of high density and low density voxels.

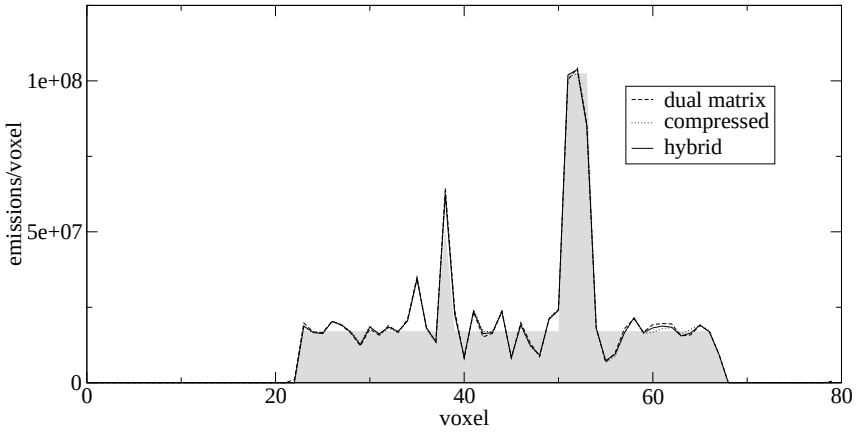


(a) M (b) A (c) compressed

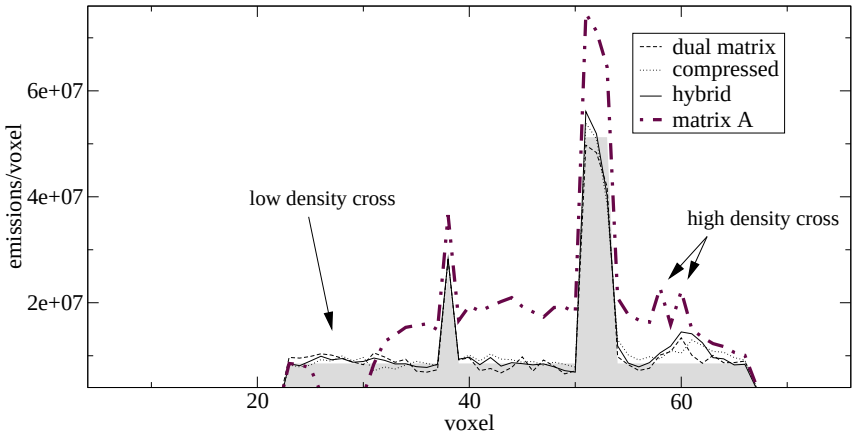


(d) dual matrix (e) hybrid

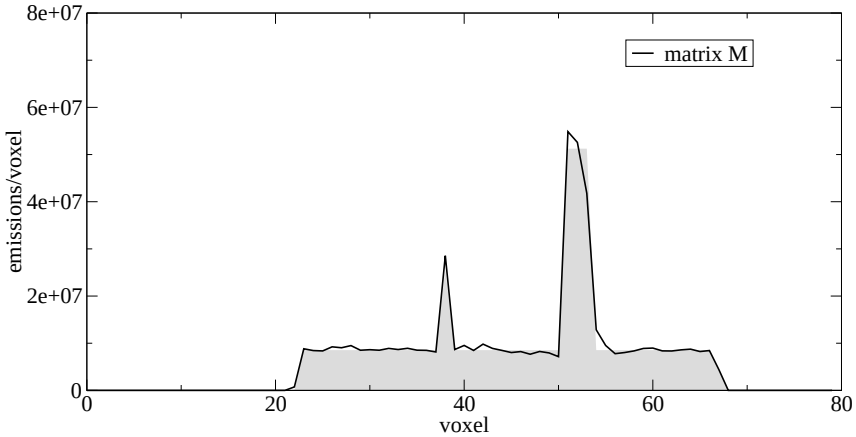
Figure 7.8.: Reconstructed images for setup Bb using different reconstruction algorithms at iteration step 500.



(a) Setup Ba



(b) Setup Bb using scatter free matrix A, compressed matrix M, dual matrix approach or hybrid approach



(c) Setup Bb using matrix M

Figure 7.9.: Diagonal profiles (upper left corner = voxel 0, lower right corner = voxel 79; see Fig. 7.7 and Fig. 7.8) through the reconstructed images at iteration number 500.

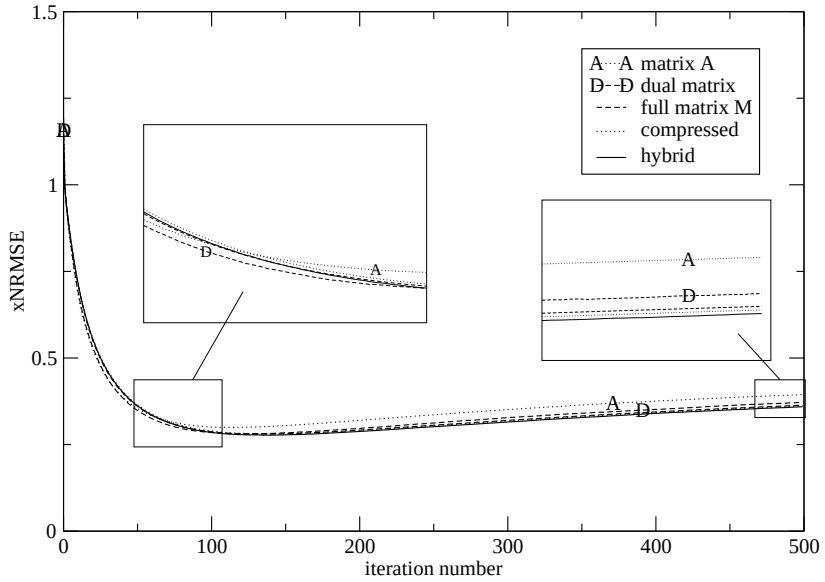


Figure 7.10.: xNRMSE of the reconstructed images for setup Ba using different reconstruction methods.

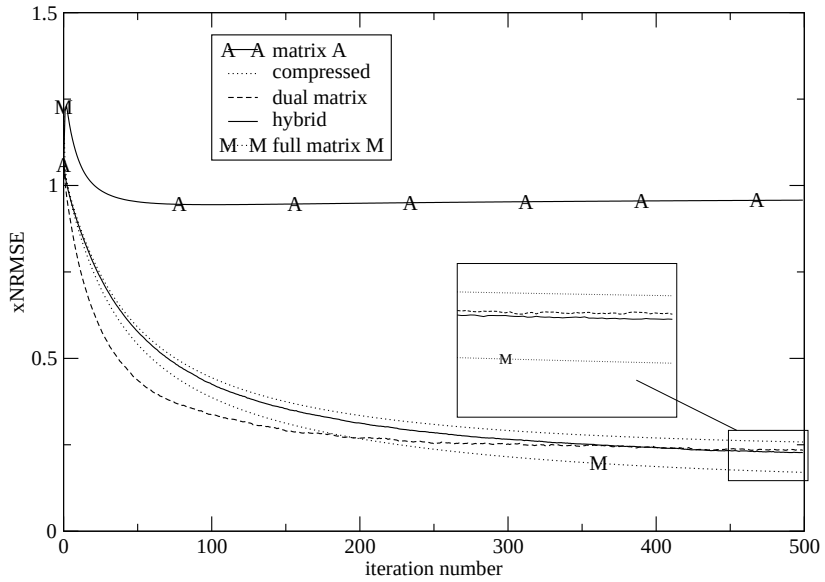


Figure 7.11.: xNRMSE of the reconstructed images for setup Bb using different reconstruction methods.

In Fig. 7.10 and Fig. 7.11 a measure for the closeness of the reconstructed activity to the true activity (the xNRMSE) is plotted as a function of iteration number. It can be seen that images reconstructed with the compressed matrix (compressed matrix ML-EM and hybrid approach) show a similar xNRMSE-shape like images reconstructed with matrix M for both setup Ba and setup Bb, whereas the dual matrix approach results in a different shape (stronger early convergence and then staying rather constant). The hybrid approach performs best for both scanners.

7.4.3. B-spline order and grid dimensions

It is possible to influence the compression quality by changing the B-spline order or by changing the dimensions of the spline node grid.

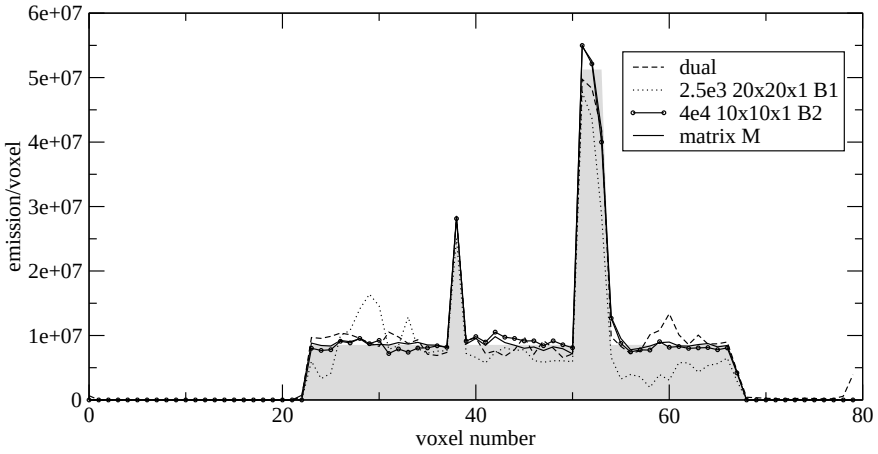
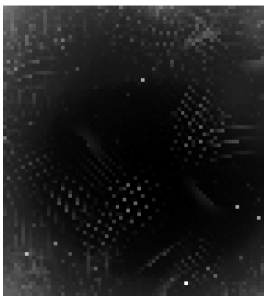
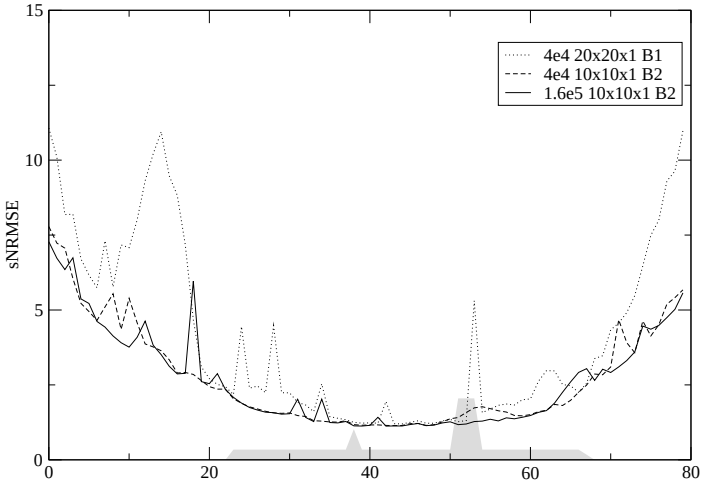


Figure 7.12.: Improvement by adjusting the number of nodes, using B_2-spline, and odd spline grid size/voxel size ratio: Diagonal profiles (upper left corner = voxel 0) through the reconstructed images at iteration number 500 (setup Bb).

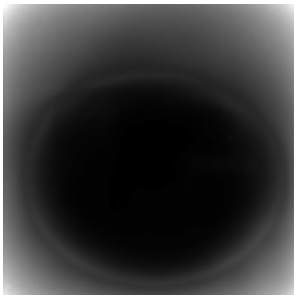
It was possible to improve the reconstructed images using a compressed matrix that is compressed using B_2-spline kernels instead of B_1-spline kernels and by slightly increasing the size of the spline node grid from a $10 \times 10 \times 1$ grid that covered exactly the volume to be reconstructed (node to node distance exactly 1/9 of 50 cm, used in the previous sections) to a $10 \times 10 \times 1$ grid with node-to-node distance of 1/8.3831987 of 50 cm. The first measure reduced the susceptibility to Monte Carlo noise due to the larger support of the kernels. The latter reduced aliasing like effects (see Fig. 7.13). Both measures improved convergence (Fig. 7.14).



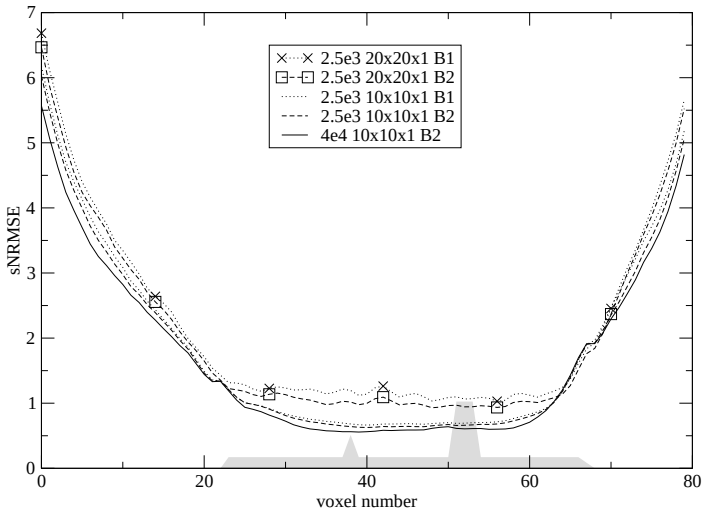
(a) setup Ba



(b) diagonal profile through 7.13(a)



(c) setup Bb



(d) diagonal profile through 7.13(c)

Figure 7.13.: Gray value images of sNRMSE_i for a compressed matrix with 4×10^4 emissions/voxel and $10 \times 10 \times 1 - \mathrm{B}_2$ compression for setup Ba (7.13(a)) and setup Bb (7.13(c)) and diagonal (upper left - lower right corner) profiles through these sNRMSE images and sNRMSE images obtained by compressing a 2.5×10^3 matrix.

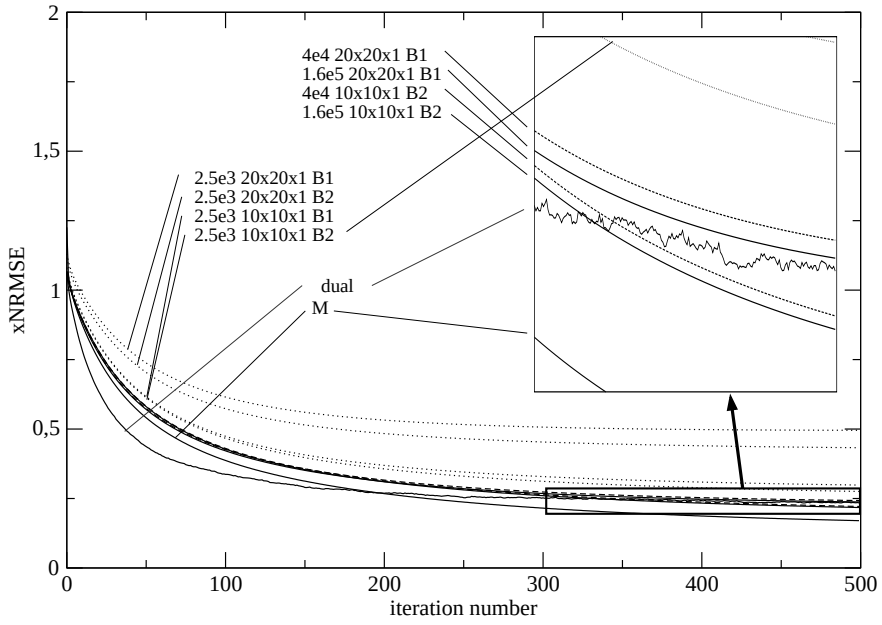


Figure 7.14.: xNRMSE of the reconstructed images for setup Bb using different reconstruction methods and statistics.

7.5. The influence of Monte Carlo noise on the reconstructed images

7.5.1. Propagation of noise in iterative reconstructions

An interesting and important aspect of iterative reconstruction algorithms is the propagation of error. The sources of error are the noise in the sinogram, the incorrect system matrix, and (in the case of the dual matrix approach) noise introduced by the Monte Carlo simulation in the projector. The problem is therefore to find the relation:

$$\begin{matrix}\text{error in sinogram or}\\ \text{error in matrix or}\\ \text{error in forward projector}\end{matrix} \longrightarrow \text{error in reconstructed image at iteration } k \quad (7.8)$$

The propagation of sinogram noise in reconstruction with the ML-EM algorithm was investigated theoretically [Barrett et al., 1994] and by simulations [Wilson et al., 1994]. The problem of the investigation of the influence of noise in the system matrix is the long calculation time needed for the simulation of the matrices. The error introduced by the system matrix can be divided into two parts. A systematical error

introduced by wrong modeling and (in the case of a Monte Carlo based system matrix) a statistical error introduced by Monte Carlo noise. While it is difficult to find a relation between wrong modeling and the resulting error in the reconstructed images due to problem of precisely defining the modeling error, the investigation of the statistical error and its influence on the reconstructed images is straight forward. If the system matrix is calculated by Monte Carlo simulations and the modeling is considered to be correct, there should be a monotone relation between any error measure and the number of simulated emissions. In other words, the word "error" in (7.8) can be replaced by "noise", and this "noise" in the sinogram, matrix, or forward projection can be related to the number of simulated emissions.

The variance of a voxels i at an iteration step k can the be calculated by taking N reconstructed images $\mathbf{x}_\alpha^{(k)}$ $(\alpha = 1, \cdots, N)$ using N matrices that were calculated with different seeds.

$$\sigma^2_{i\ \mathrm{matrix}}(k) = \frac{1}{N-1} \sum_{\alpha=1}^{N} \left(x_{i,\alpha}^{(k)} - \bar{x}_i^{(k)} \right)^2 \tag{7.9}$$

$$\bar{x}_i^{(k)} \equiv \text{ mean value of voxel } i \text{ at iteration } k$$

The same approach is applicable to the sinogram error by keeping the matrix seed fixed. For consistency the same number of sinograms was used. The sinograms were simulated without using variance reduction techniques.

In the case of the DM reconstructions three sources of error exist: the sinogram, the matrix A, and the Monte Carlo scatter projection (leading to the sinogram $\mathbf{s}^{(k)}$). The influence of each source can be measured again by varying the seed of the corresponding Monte Carlo simulation and keeping the two other seeds constant.

The normalized root mean variance NRMV introduced in (7.5) on page 68 is used as a global metric (in contrast to the voxelwise metric of (7.9)) to define the noise in the reconstructed images. The NRMV is calculated at each iteration step. It is possible to relate the NRMV to the other global metric NRMSE. At iteration k the image $\mathbf{x}^{(k)}$ is obtained by successive application of the projector and back-projector operators on the starting image $\mathbf{x}^{(0)}$. The back-projector is containing explicitly the sinogram $\mathbf{y}^\star$ while both depend on the system matrix M.

$$\mathbf{x}^{(k)} = \left(\prod_{n=1}^{k} \mathcal{BP} \right) \mathbf{x}^{(0)} \equiv \mathbf{F}^{(k)}_{\mathbf{y}^\star, M}(\mathbf{x}^{(0)}) \tag{7.10}$$

With the help of a Taylor expansion it is possible to estimate the total error $\epsilon_i(k) = x_i(k) - x_i^{\text{true}}$.

$$\begin{aligned}{\epsilon_i}^2(k) \approx & \underbrace{\sum_j \left(\frac{\partial F_i^{(k)}}{\partial y_j^\star}\right)^2 \Delta y_j^{\star 2}}_{\approx \sigma_{i\ \text{sinogram}}^2(k)} \\ & + \underbrace{\sum_{j,l,r,s} \frac{\partial F_i^{(k)}}{\partial m_{jl}} \frac{\partial F_i^{(k)}}{\partial m_{rs}} \operatorname{Cov}(m_{jl}, m_{rs})}_{\approx \sigma_{i\ \text{matrix}}^2(k)} \\ & + (\text{convergence error})^2 \end{aligned} \tag{7.11}$$

Here the noise in the sinogram bins ($\Delta y_j^\star$) and the matrix elements (Δm_{ji}) is not correlated, because different simulations (with different Monte Carlo seeds) are started. A correlation between two matrix elements, however, is in principle possible due to the variance reduction techniques that are used for the simulation of the elements. The relation (7.11) is Gauss' law of error propagation modified due to the potential correlation of the matrix elements $\operatorname{Cov}(m_{jl}, m_{rs})$ and with an additional term to account for the fact that the algorithm is not converged. This latter addend decreases to zero for large k in case of convergence. Therefore, for large k the convergence error can be neglected and the following inequality can be derived using the triangle inequality.

$$\text{NRMSE}(k) \lesssim \text{NRMV}_{\text{sinogram}}(k) + \text{NRMV}_{\text{matrix}}(k) \tag{7.12}$$

This inequality above is also valid when using higher order Taylor expansions which is necessary in the case of large errors.

7.5.2. Convergence and noise propagation of the full matrix and the dual matrix algorithm

The noise propagation was investigated for two reconstruction algorithms: the full matrix and the dual matrix approach. The simulation of the matrices is very time consuming. For this reason only are rather small number of $N = 9$ scatter-free matrices A_α and also nine full matrices M_α (including scatter and direct coincidences) were simulated using different Monte Carlo seeds. The simulated scanner was scanner 'a' (detector ring depth of 0.645 cm) and the used phantom was phantom A (see Fig. 7.1 and Table 7.1). For a fixed sinogram then images were reconstructed using these nine

emissions per voxel	non-zero elements
160,000	40.5 %
40,000	23.0 %
10,000	9.8 %
160,000 (no scatter)	1.9 %

Table 7.4.: The influence of the number of simulated particles (with variance reduction) on the fraction of non-zero elements in the matrix (setup Aa).

matrices.

In order to investigate the influence of the number of simulated emissions on the error in the reconstructed emission, sinograms and matrices of different statistics were simulated. The sinograms were either simulated with 5×10^9 or 1×10^9 emissions (no variance reduction). This corresponds roughly to 30 and 5 min scans of (average) 6 Becquerel/ml initially. The detectors were idealized (and no Gaussian energy filtering was applied)[1].

The system matrices A_α and M_α were calculated simulating a fixed number of photon pairs per voxel. Due to the high number of voxels a simulation without variance reduction techniques was not possible. Both stratification and forced detection were used in the simulation. The number of simulated emissions per voxel were 1×10^4, 4×10^4, and 1.6×10^5. In Table 7.4 the fraction of non-zero elements for matrices of different statistics can be seen. The total number of matrix elements was 234,700,800. The total number of simulated emissions (in the case of the 160,000-matrix) was 1.024×10^9 emissions.

In Fig.7.15 the NRMSE of the reconstructed images with matrices of different statistics can be seen. The figure shows the typical property of the iterative solution of an (unregularized) ill-posed problem: after a relatively fast convergence the algorithm starts to "focus" on the noise and to drift away from the true solution.

Clearly, better statistics resulted in a smaller NRMSE, but the difference between the 1.6×10^5-matrix and the 4×10^4-matrix was already much smaller than the difference between the latter and the 1×10^4-matrix. Better the matrix statistics lead to a shift of the minimum (best agreement between true and reconstructed image) towards higher iterations.

Fig. 7.16 shows the NRMSE using the same matrices as in Fig. 7.15, but applying the algorithm to the low count sinogram with 1×10^9 emissions. The error introduced

[1]Therefore, the statistics of the sinograms was better than in real experiments (with the same emission density).

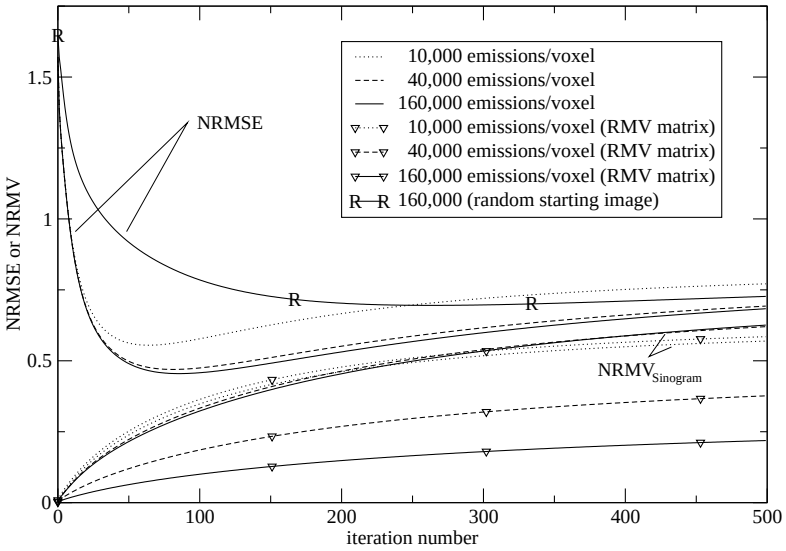


Figure 7.15.: Error vs. iteration number for full matrix reconstructions with different statistics as indicated. The R-R graph shows the NRMSE when using a starting image with random voxels $x_i \in [0, 2[$ instead of a uniform image with voxels $x_i \equiv 1$. The sinogram was simulated with 5×10^9 emissions in total. Iteration number zero corresponds to errors of images after the first iteration.

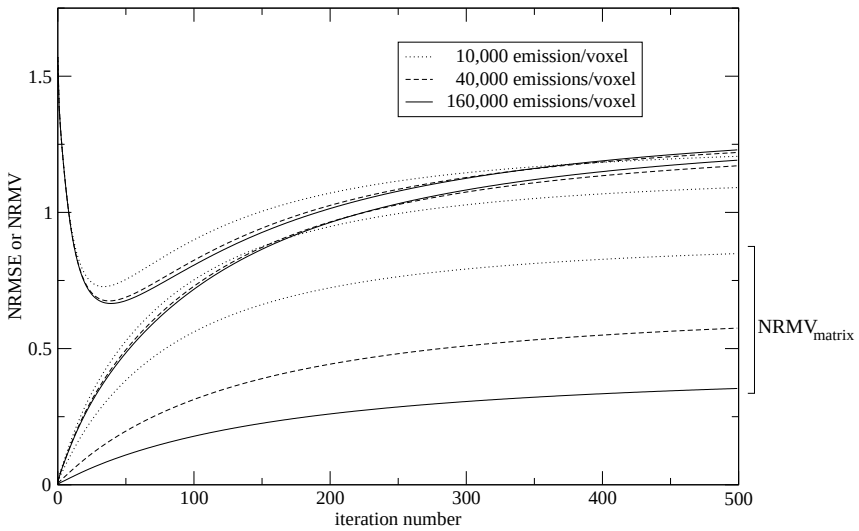


Figure 7.16.: Error vs. iteration number for full matrix reconstructions of a sinogram simulated with 1×10^9 emissions in total and matrices of different statistics.

by the sinogram is bigger. The shape of the NRMSE at higher iterations is mostly determined by this error. The bigger total noise induced error results in a shift of the minimum of the NRMSE towards early iterations. At the first iterations, there is only a small deviation between the corresponding NRMSE curves in Fig. 7.16 and Fig. 7.15. This is caused by the strong influence of the starting image which is in both cases the same. In both figures the validity of inequality (7.12) for large iteration numbers can be verified. Fig. 7.17 shows the relative importance of the matrix error

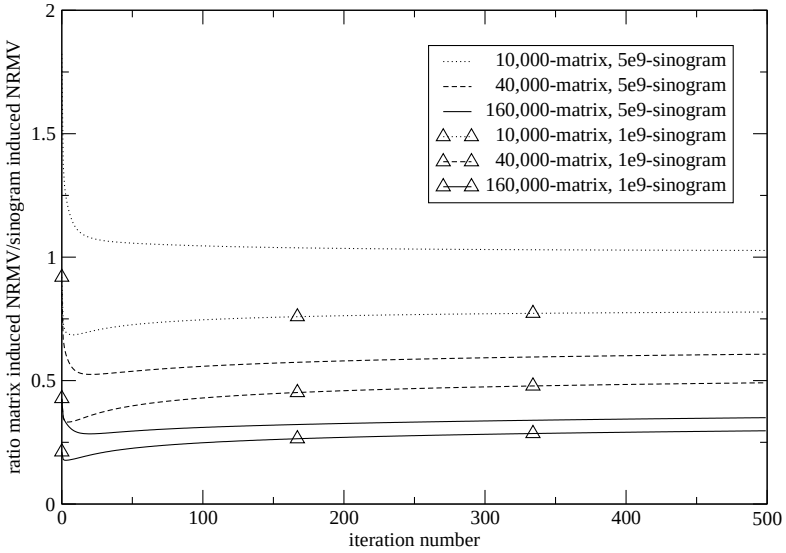


Figure 7.17.: The ratio of the $\mathrm{NRMV}_{\mathrm{matrix}}/\mathrm{NRMV}_{\mathrm{sinogram}}$ for a sinogram with 1×10^9 emissions and sinogram with 5×10^9 emissions is shown for different matrix and sinogram statistics.

vs. the sinogram error in the case of full matrix reconstruction. The matrix error became comparable to the error caused by the sinogram for the 5×10^9-sinogram and the 10,000-matrix (i.e. the ratio is approximately one). The ratio can be used to determine the required number of simulated emissions per voxel: the error caused by the sinogram should be larger than the error caused by the matrix. At higher iteration numbers the ratio of the influences of both error sources seems to be independent of the iteration number.

Fig. 7.18 shows the NRMSE for full matrix and dual matrix reconstructions with different statistics. For the dual matrix reconstructions a fraction p of 0.001, 0.0001 or 0.00001 of the guessed activity was simulated in the forward Monte Carlo simulation, corresponding to approximately 5×10^6, 5×10^5 or 5×10^4 simulated emissions. More simulated particles led to smaller NRMSE as expected.

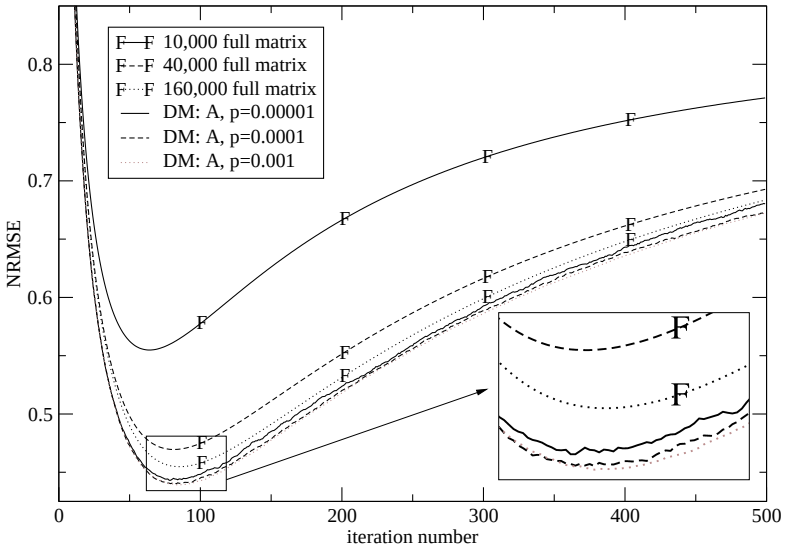


Figure 7.18.: NRMSE for full matrix reconstructions and dual matrix reconstructions. Sinogram 5×10^9 emissions.

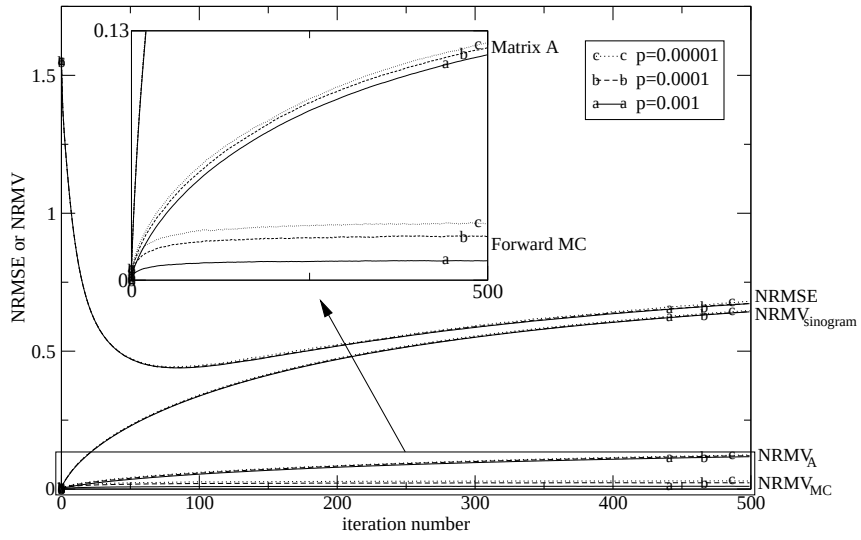


Figure 7.19.: NRMSE and NRMV of dual matrix reconstruction with matrix A for the sinogram with 5×10^9 emissions and different fractions of simulated forward scatter. The small figure shows more enlarged NRMV curves caused by the matrix A and by the forward Monte Carlo simulation.

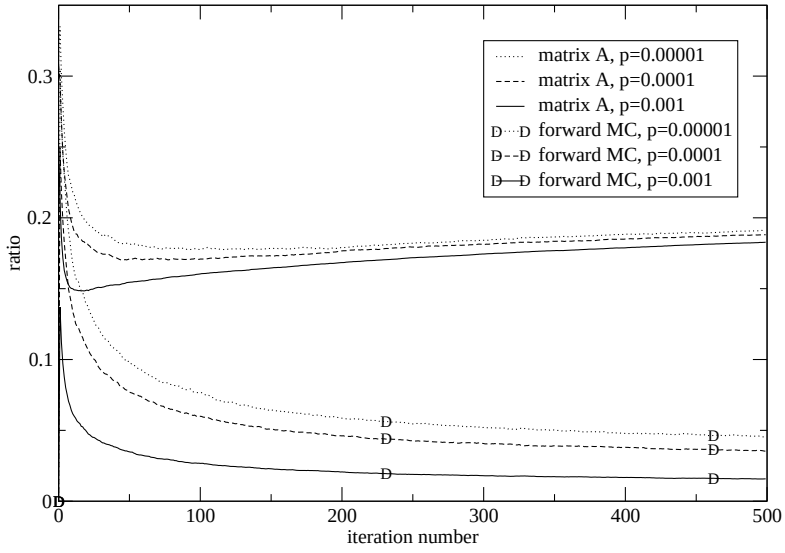


Figure 7.20.: Ratios $\mathrm{NRMV_A}/\mathrm{NRMV_{sinogram}}$ and $\mathrm{NRMV_{Monte\ Carlo\ scatter}}/\mathrm{NRMV_{sinogram}}$ in dual matrix reconstruction

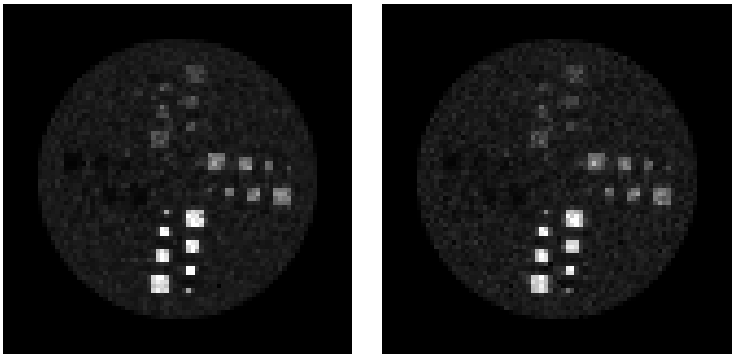
A characteristic feature of the dual matrix reconstruction is the faster initial convergence. This feature might, however, be related to the uniform starting image. Very typical are also the noise like fluctuations that can be seen in the $p = 0.00001$-curve. This feature is also very weakly present in the $p = 0.0001$ curve (and extremely weekly in the $p = 0.001$ curve). These fluctuations decreased the more particles were simulated in the forward Monte Carlo simulation.

Fig. 7.19 shows the contribution of the different sources of error for the dual matrix reconstruction. Clearly, the error caused by the sinogram dominated. The error introduced by the system matrix A had similar features like the matrix induced error in the full matrix reconstruction.

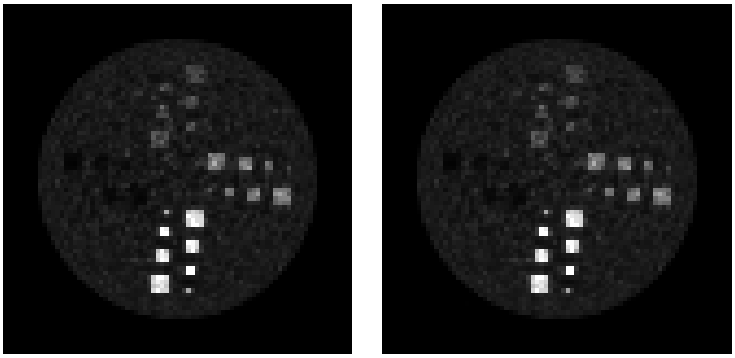
The error introduced by the forward Monte Carlo scatter increased initially and soon stayed rather constant. This suggests that the algorithm converges to some mean solution and oscillates randomly around this solution with rather constant mean amplitude.

In analogy to Fig. 7.17, Fig. 7.20 shows the importance of the error introduced by matrix A and the forward Monte Carlo scatter simulation relative to the sinogram induced error. Again the error caused by the sinogram dominates. The small error due to the forward Monte Carlo scatter simulation can be explained by the fact that the system was a 2D system with low scatter contribution. Therefore the scatter free matrix A which occurs both in the projector and back-projector mostly determines

the convergence properties. The error ratio introduced by matrix A shows similar properties like the error ratio in full matrix reconstruction (an almost constant non-zero ratio at high iteration numbers). This is not the case for the error introduced by the Monte Carlo forward simulation.



(a) $160,000$-matrix, iteration 85 (b) $10,000$-matrix, iteration 64



(c) DM, $p = 0.001$, iteration 83 (d) DM, $p = 0.00001$, iteration 78

Figure 7.21.: Reconstructed images for the 5×10^9-emissions-sinogram at minimal NRMSE

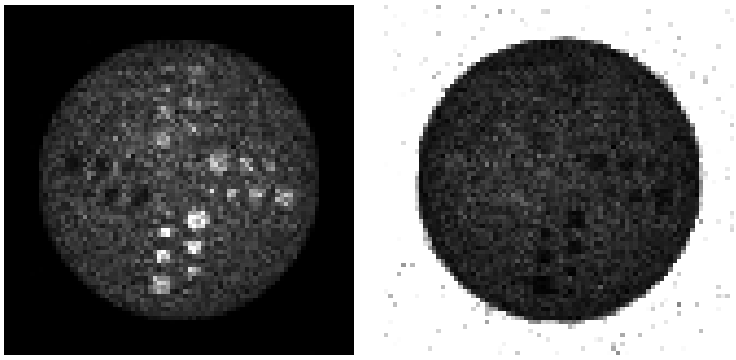
Fig. 7.21 shows the reconstructed images. Clearly the difference is very small which can be accounted to the small scatter fraction.

In the following figures, the voxel dependency of the error is shown. Images of minimal NRMSE were chosen. The likelihood of the EM-algorithm was not used to define the iteration number of the images, because the likelihood itself depends on the quality of the matrix but not on the difference to the true activity. The choice of the minimal NRMSE as a criteria for selection in general leads to different iteration numbers. This should be kept in mind when comparing images. The absolute

error (standard deviation) or the relative error (standard deviation divided by the mean value) for each voxel is visualized as the gray value of the corresponding voxel (white ≡ big error, black ≡ small error) .

Fig. 7.22 shows the absolute and relative error caused by the sinogram. The absolute error did not depend much on the reconstruction method. More emissions in a voxel resulted in a larger absolute error, which is in agreement with the study on noise properties of the EM algorithm by Wilson *et al* Wilson et al. [1994].

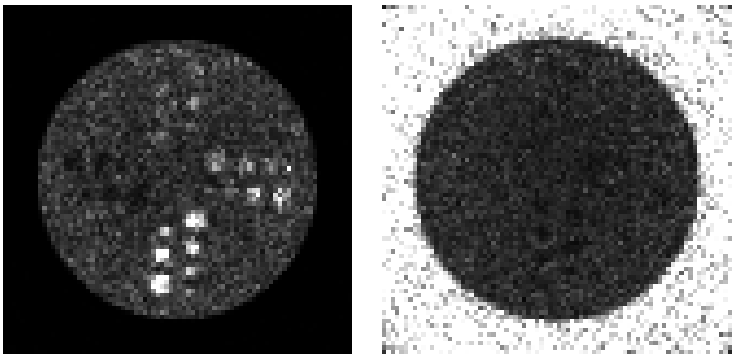
On the other hand, more emissions lead to a reduced relative error. In both full matrix and dual matrix reconstruction the error seemed not to depend on the phantom density of the voxels. This might be different for scanners with large scatter fraction. The density change at the border of the phantom, however, clearly influenced the relative error introduced by the sinogram. This is probably caused by the fact, that both algorithms converged much faster outside the phantom to small (zero) activities.



(a) abs. error, iteration 85 (b) rel. error, iteration 85

Figure 7.22.: Sinogram induced absolute and relative error for the 5×10^9-emissions-sinogram and the 160,000-matrix at minimal NRMSE (white ≡ big error, black ≡ small error). The very large relative error of voxels outside the cylinder often exceeded the gray value scaling. The error of these voxels is therefore represented by white color.

In dual matrix reconstruction (Fig. 7.24), the error caused by the scatter-free matrix A has the same features. In the case of the error introduced by the forward scatter simulation a similar behavior can be seen, but in addition the noise outside the phantom seems to be structured (see Fig. 7.24(d)). This can be seen only in the relative error images, because the reconstructed activity outside the phantom is small. Therefore, this effect is negligible for single ring scanners with low scatter fraction.



(a) abs. error, iteration 64 (b) rel. error, iteration 64

Figure 7.23.: Matrix induced absolute and relative error for the 5×10^9-sinogram and the 10,000-matrix at minimal NRMSE (white $\equiv$ big error, black $\equiv$ small error)

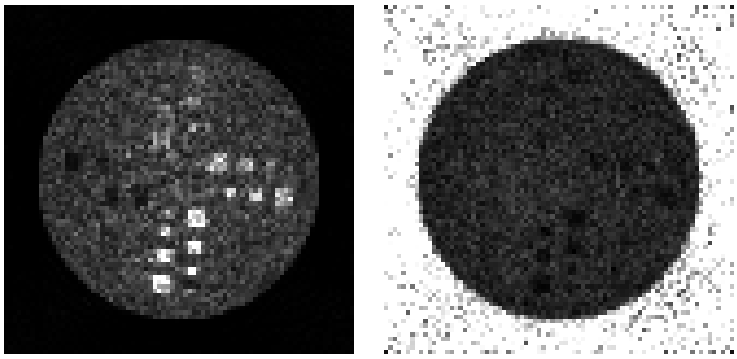
7.5.3. Discussion

The log-likelihood that is to be maximized by the algorithm is based on the approximated noisy matrix and the approximated noisy sinogram. The algorithm does therefore not converge to the maximum of the ideal problem but to a shifted maximum (image x^∞). Before converging to this shifted maximum, the NRMSE can even become smaller than at higher iterations, because the uniform starting image encourages smooth reconstructed images which often agree better with the original image x^{true}. The fading influence of this starting image at higher iteration numbers leads to a positive slope of the NRMSE. The NRMSE eventually approaches asymptotically $\text{NRMSE}(x^\infty) > \text{NRMSE}(x^{\text{true}}) \equiv 0$.

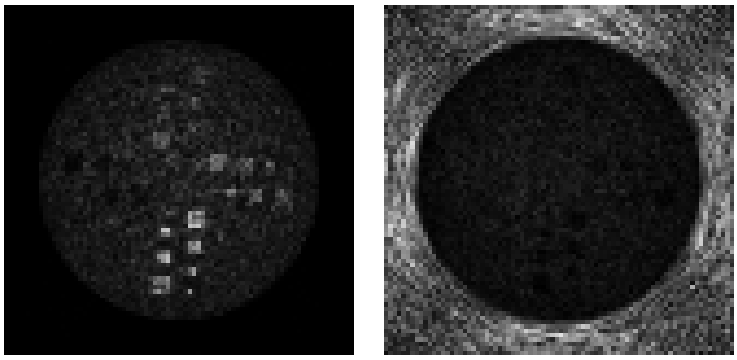
This explanation can be verified by looking at graph R in Fig. 7.15 where the starting image is an image with random voxel values between zero and two. The agreement between the true image and this random starting image is much smaller. This leads to an almost horizontal slope at higher iterations and a shift of the minimum of the NRMSE towards higher iterations.

The fading influence of the starting image during the reconstruction process is associated with a growing influence of the given information (the matrix and the sinogram) on the reconstructed images. Since the matrix and the sinogram are noisy, the corresponding NRMVs should grow as well. This can be verified in Fig. 7.15 which shows monotonously increasing NRMVs.

In both Fig. 7.15 and in Fig. 7.16 the NRMV curve of the sinogram induced



(a) abs. error, matrix A induced (b) rel. error, matrix A induced



(c) abs. error, forward scatter induced (d) rel. error, forward scatter induced

Figure 7.24.: Error in dual matrix reconstruction with $p = 0.00001$ at iteration 78=minimal NRMSE (white $\equiv$ big error, black $\equiv$ small error)

error with the 10,000 matrix is below the two other curves using the higher statistics matrices for iteration numbers $\gtrsim 230$ and $\gtrsim 170$ respectively. This means that the NRMV of the sinogram somehow depends on the number of simulated emission of the matrix. This can have three reasons. The derivative of $F^{(k)}_{\mathbf{y}^\star,M}$ in (7.11) depends implicitly on the matrix. Therefore, the NRMV of the sinogram should be positioned randomly around some mean NRMV curve. In the case of bad matrix statistics this deviation should be larger and the considered curve could be positioned below the two other curves. However, this is not very likely, because the NRMV is the average over many voxels. Secondly, in the case of few simulated emissions, the non-zero fraction of the matrix is small. This might lead to a reduced rank of the matrix which should also influence the mentioned derivative. Thirdly, it is possible that due to a large matrix error higher order terms in the Taylor expansion can not be neglected anymore. This

would lead also to a dependency of the considered NRMV on the matrix statistics.

Simulations of a similar one-ring scanner with larger detector ring width (and phantom), and therefore scatter fraction, could give more insight into the noise propagation and performance of the two algorithms in 3D scanners. Preliminary simulations of such a scanner showed that relative to the full matrix algorithm the position of the minimum of the NRMSE curve of the dual matrix algorithm is positioned at smaller iteration numbers, while the minimal NRMSE is increasing. It can be therefore expected that in the case of 3D scanners the dual matrix algorithm is initially converging faster than the full matrix algorithm.

While the position of the minimum of the dual matrix approach relative to the full matrix approach seems to be rather sensitive to the scatter fraction, it can be expected that the qualitative shape of the NRMV curves (monotonously increasing influence of matrix or sinogram noise) should stay the same. For large iteration numbers, there should be an upper bound NRMSE (x^{∞}) for the NRMSE as discussed at the beginning of this section. Together with (7.11) this suggests that there exists an upper bound for the NRMVs as well, independently of the scatter fraction. The property of the $\mathrm{NRMV}_{\mathrm{MC}}$ to stay at a constant level after only few iterations is probably also present in the case of 3D scanners. It can be expected that the influence of this forward scatter simulation will increase when more scatter is present. While in the case of 3D scanners the variance of the voxels will most likely also be correlated to the activity of these voxels, there might be an additional correlation to the density which was not present in the 2D case.

7.6. Performance

Although the main purpose of the simulations in this chapter is the proof of concept of the compression scheme as well as the investigation of the noise propagation, the performance of the used algorithms and methods is important.

Due to the very large number of simulated photons, the calculation of the system matrices were parallelized. A small computer cluster of eight two-processor boards with AMD Athlon MP 2800+ was used for this purpose. The calculation time for a matrix with 10,000, 40,000, and 160,000 emissions/voxel was around 3 1/2, 14, and 50 minutes.

The compression of a matrix (section 7.4) on a single AMD Athlon MP 2800+ computer was performed in around 4 1/2 minutes. In Table 7.5 the time needed per iteration step for the different reconstruction algorithms (section 7.4) can be seen. A

resident matrix was used in the case of the full matrix approach. This is the reason for the short iteration time. If the matrix elements had been re-calculated at each iterations step, the algorithm would perform impractically slowly. The compression

Method	Time per iteration
resident full matrix	ca. 5 sec
dual matrix	ca. 2 min
compressed matrix	ca. 2 min
hybrid	ca. 2 1/2 min

Table 7.5.: Time needed per iteration using a single AMD MP 2800+ processor.

of the matrix as well as the reconstruction algorithms can be parallelized. Groups of spline-kernels or single spline-kernels can be compressed in parallel. This was not implemented but should allow the compression of the matrix on the 16 processor cluster in around 20 seconds. The presented reconstruction algorithms are fully parallelizable [Fessler, 2004]. Therefore a reduction of the time per iteration from 2 minutes to roughly 10 seconds can be expected, when the this cluster is used.

The simulations of 10^8 photon pairs took more than 8 h with GEANT4 (no variance reduction), around 90 minutes with YaPRA and variance reduction, and around 7 1/2 minutes without variance reduction.

8. Conclusions and Outlook

The reconstruction of the emission density in PET examinations relies on the correct modeling of the system. In contrast to small animal imaging where the scatter and perhaps even the attenuation in the animal might be neglected, a good approximate model of a PET scan of humans must comprise the scanner *and* the patient. Especially 3D scanners, which become more and more common due to their increased sensitivity, are very difficult to model because of their complicated system response. The efficient and correct modeling of the attenuation and the scanner geometry and hardware is demanding. However, the complexity is introduced by the scatter in the patient that varies from scan to scan and that is completely asymmetrical.

Monte Carlo simulations are well suited to approximate the attenuation and especially the scatter in the patient. Different approaches to incorporate these simulations in the reconstruction process were investigated. For this purpose a fast Monte Carlo code that is capable of tracking photons in the patient was implemented. Stratification and forced detection were used in the Monte Carlo simulations. These variance reduction techniques were optimized for system matrix calculation. Sinograms could be simulated without variance reduction techniques in a reasonable time. This was advantageous, because the simulations then showed similar statistical properties like measurements. Reconstructions were performed based on maximum likelihood expectation maximization. The ideal way of scatter treatment, the simulation of the whole matrix including patient scatter (full matrix approach) is impracticable for 3D scanners, because of the simulation time, the reconstruction time (due to the large number of non-zero elements), and especially because the storage of the matrix in memory is not feasible with present hardware ($\mathcal{O}(10^{13+})$ matrix elements).

Three different ways to include Monte Carlo calculated scatter into the reconstruction were investigated. The full matrix approach was compared to the incorporation of Monte Carlo scatter in the projector only (dual matrix approach, storage of the matrix not necessary), and to a new method that uses a compressed Monte Carlo matrix. Single ring scanners were simulated as a proof of principle and because in this way the uncompressed matrix could be kept in memory for comparative reconstructions.

The compression method reduced the size of the scatter part of the matrix so that

the storage of the matrix for a 3D scanner should become feasible. The compression scheme separated scatter from direct coincidences. The compression of this scatter matrix is based on a parametrization of the columns of the matrix (i.e. sinograms of single voxels) and an approximate description of the change between columns. The implementation allowed a compression of very noisy system matrices. Together with a reasonable compression speed, this allowed the calculation of system matrices of single ring scanners within several minutes on a small computer cluster.

In a second part, the convergence and especially the noise propagation of dual matrix ML-EM and full matrix ML-EM was investigated. The approximate dual matrix approach showed a faster early convergence in the case of a uniform starting image, but the compressed approach yielded less variance of voxel values. Due to the low scatter fraction of the simulated single ring scanner, the noise propagation was dominated by the uncertainty of unscattered coincidences. The propagation of noise of the sinogram, of the noise in the matrix elements and in the forward projection was investigated. As the influence of the starting images decreases the influence of the matrix and the sinogram increases. The noise in the sinogram will be given by the measurement, but the system matrix will always be simulated and can be improved by longer simulations or faster computers. Theoretical considerations and extrapolation of the results showed that there is an upper bound for the introduced error in the reconstructed images. The influence of the noise in the forward projector Monte Carlo simulation (dual matrix approach) on the reconstructed images is more difficult. The simulations showed that this noise introduces an error that oscillates randomly.

In summary, three different possibilities to incorporate Monte Carlo simulations into the reconstruction process were investigated. In a proof of concept a novel way to circumvent the matrix storage problem by compressing the matrix was introduced and compared to the incorporation of scatter only in the forward projector (dual matrix or relative approaches). The proposed scatter matrix compression scheme allows the storage of the matrix for a 3D human scanner with patient scatter which otherwise would be impossible. The presented results form the basis of further projects which may include more realistic simulations, the investigation of a greater variety of algorithms, and the improvement of the implementation of the compressed matrix approach.

It would therefore be worthwhile to simulate realistic 3D scanners. This implies the implementation of a compression scheme for oblique sinograms similar to the presented scheme for transversal sinograms and a more realistic modeling of detectors, which were simplified in the presented simulations. A fast but precise incorporation of the detectors in the simulations should be possible when the detector response (depend-

ing on the incoming photon location, energy, and incidence angle) is parametrized. These parameters could be obtained either by measurements or by detailed simulations (for example with generic Monte Carlo codes like GEANT4). Monte Carlo based reconstruction is very well suited for studies on noise propagation and therefore alternative reconstruction algorithms like ordered subset expectation maximization [Hudson and Larkin, 1994], RAMLA [Browne and De Pierro, 1996] or primal dual approaches [Johnson, 1997, Johnson et al., 2000, Johnson and Sofer, 2000] should be investigated. Regularization which should insure less noisy images and sometimes an improved convergence speed should be investigated in terms of noise propagation and together with a compressed matrix. Finally, reconstructions based on Monte Carlo simulations offer great advantage where conventional reconstruction techniques are limited or erroneous. This includes temporally varying patients which can be simulated easily by Monte Carlo simulations. Triggered and listmode PET/CT or future simultaneous PET/MR scans can be used as the basis for such simulations. "Dirty" isotopes like ^{86}Y or ^{124}I that can be used to monitor cancer treatment are very problematic to model correctly with conventional methods due to their prompt γ emissions but can be simulated well by Monte Carlo simulations.

Bibliography

L.-E. Adam, J. S. Karp, and G. Brix. Investigation of scattered radiation in 3D whole-body positron emission tomography using Monte Carlo simulations. *Phys. Med. Biol.*, 44:2879, 1999.

Sakari Alenius. *On noise reduction in iterative image reconstruction algorithms for emission tomography: Median root prior.* PhD dissertation, Tampere University of Technology, Tampere, 1999. URL `http://research.nokia.com/people/sakari_alenius/thesis.pdf`.

Dale L. Bailey and Steven R. Meikle. A convolution-substraction scatter correction method for 3D PET. *Phys. Med. Biol.*, 39:411, 1994.

Harrison H. Barrett, Donald W. Wilson, and Benjamin M. W. Tsui. Noise properties of the EM algorithm: I. theory. *Phys. Med. Biol.*, 39:833, 1994.

Freek J. Beekman, Hugo W. A. M. de Jong, and Sander van Geloven. Efficient fully 3-D iterative SPECT reconstruction with Monte Carlo-based scatter compensation. *IEEE Trans. Med. Imag.*, 21(8):867, 2002.

M. Bergström, L. Eriksson, C. Bohm, G. Blomqvist, and J. Litton. Correction for scattered radiation in a ring detector positron camera by integral transformation of the projections. *J. Comput. Assist. Tomogr.*, 7(1):42, 1983.

T. Beyer, C. C. Watson, C. C. Meltzer, D. W. Townsend, and R. Nutt. The biograph: A premium dual-modality PET/CT tomograph for clinical oncology. *electromedica*, 69:120, 2001.

T. Beyer, Y. Y., and S. Kaepplinger. PET/CT tomography using a new PET detector material for ultra-fast imaging in clinical oncology. *electromedica*, 70:151, 2002.

Thomas Beyer, Gerald Antoch, Stefan Müller, Thomas Egelhof, Lutz S. Freudenberg, Jörg Debatin, and Adreas Bocktisch. Acquisition protocol considerations for combined PET/CT imaging. *J. Nucl Med.*, 6945(1/Suppl):25S–35S, 2004.

David Brasse, Paul E. Kinahan, Carole Lartizien, Claude Comtat, Mike Casey, and Christian Michel. Correction methods for random coincidences in fully 3D whole body PET: Impact on data and image quality. *J. Nucl. Med.*, 46(5):859–867, 2005.

Jolyon Browne and Alvaro De Pierro. A row-action alternative to the EM algorithm for maximizing likelihoods in emission tomography. *IEEE Trans. Med. Imag.*, 15(5): 687–699, 1996.

Jerrold T. Bushberg, J. Anthony Seibert, Edwin M. Leidholdt Jr., and John M. Boone. *The essential of physics of medical imaging*. Lippincott Williams & Wilkins, Philadelphia, 2002.

Irène Buvat, Delphine Lazaro, and Vincent Breton. Fully 3D Monte Carlo reconstruction in SPECT: Proof of concept and is it worthwile? In *The VIIth International Conference on Fully 3D Reconstruction in Radiology and Nuclear Medicine*, 2003.

Maurizio Conti, Bernard Bendriem, Mike Casey, Mu Chen, Frank Kehren, Christian Michel, and Vladimir Panin. First experimental results of a time-of-flight reconstruction on an LSO PET scanner. *Phys. Med. Biol.*, 50:4507–4526, 2005.

Daniel J. de Vries, Stepehen C. Moore, Robert E. Zimmermann, Stefan P. Mueller, Bernard Friedland, and Richard C. Lanza. Development and validation of Monte Carlo simulation of photon transport in an anger camera. *IEEE Trans. Med. Imag.*, 9(4):430–438, 1990.

M. Defrise, D. W. Townsend, D. Bailey, A. Geissbuhler, C. Michel, and T. Jones. A normalization technique for 3D PET data. *Phys. Med. Biol.*, 36:939–952, 1991.

M. Defrise, Micheal E. Casey, Christian Michel, and Maurizio Conti. Fourier rebinning of time-of-flight PET data. *Phys. Med. Biol.*, 50:2749–2763, 2005.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Ser. B*, 39:1, 1977.

Frederic H. Fahey. Data aquisition in PET imaging. *Journal of Nuclear Medicine Technology*, 30(2):39–49, 2002.

Jeffrey A. Fessler. Statistical methods for image reconstruction, 2004. Short course at NSS/MIC conference in Rome.

Matthias Fippel. Fast Monte Carlo dose calculation for photon beams based on the VMC electron algorithm. *Med. Phys.*, 26:1466, 1999.

Matthias Horst Fippel. Entwicklung eines schnellen Monte-Carlo-Verfahrens zur Dosisberechnung in der Strahlentherapie, 2000. Habilitation, Eberhard-Karls-Universität zu Tübingen.

M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, M. Booth, and F. Rossi. *GNU Scientific library reference manual*. Network Theory Limited, 5 Royal Park, Bristol BS8 3AL, United Kingdom, revised second ediction (v1.8) edition, August 2006. URL `http://www.gnu.org/software/gsl/`.

S. Grootoonk, T. J. Spinks, D. Sashin, N. M. Spyrou, and T. Jones. Correction for scatter in 3D brain PET using a dual energy window method. *Phys. Med. Biol.*, 41: 2757–2774, 1996.

S. F. Haber, S. E. Derenzo, and D. Uber. Application of mathematical removal of positron range blurring in positron emission tomography. *IEEE Trans. Med. Imag.*, 37(3):1293–1299, 1990.

Michael D. Harpen. Positronium: Review of symmetry, conserved quantities and decay for the radiological physicist. *Med. Phys.*, 31(1):57–61, 2004.

R. L. Harrison, M. S. Kaplan, S. D. Vannoy, and T. K. Lewellen. Positron range and coincidence non-collinearity in SimSET. In *Nuclear Science Symposium, Conference Record*, volume 3, pages 1265–1268, 1999.

Robert Harrison. Simulation system for emission tomography (SimSET). URL: `http://depts.washington.edu/~simset/html/simset_main.html`.

D. R. Haynor, R. L. Harrison, T. K. Lewellen, A. N. Bice, C. P. Anson, S. B. Gillispie, R. S. Miyaoka, K. R. Pollard, and J. B. Zhu. Improving the efficiency of emission tomography simulations using variance reduction techniques. *IEEE Trans. Nucl. Sci.*, 37(2):749, 1990.

David R. Haynor, Robert L. Harrison, and Thomas K. Lewellen. The use of importance sampling techniques to improve the efficiency of photon tracking in emission tomography simulations. *Med. Phys.*, 18(5):990, 1991.

Gabor T. Herman and Lorraine B. Meyer. Algebraic reconstruction techniques can be made computationally efficient. *IEEE Trans. Med. Imag.*, 12(3):600–609, 1993.

H. Malcolm Hudson and Richard S. Larkin. Accelerated image reconstruction using ordered subset of projection data. *IEEE Trans. Med. Imag.*, 13(4):594, 1994.

John L. Humm, Anatoly Rosenfeld, and Alberto Del Guerra. From PET detectors to PET scanners. *Eurpean Journal of Nuclear Medicine and Molecular Imaging*, 30(11):1574–1597, 2003.

International Commission on Radiation Units and Measurements ICRU. Photon, electron, proton and neutron interaction data for body tissues. *ICRU-Report 46*, 1992.

S. Jan, G. Santin, D. Strul, S. Staelens, K. Assie, D. Autret, S. Avner, R. Barbier, M. Bardies, P. M. Bloomfield, D. Brasse, V. Breton, P. Bruyndockx, I. Buvat, A. F. Chatziioannou, Y. Choi, Y. H. Chung, C. Comtat, D. Donnarieix, L. Ferrer, S. J. Glick, C. J. Groiselle, D. Guez, P-F. Honore, S. Kerhoas-Cavata, A. S. Kirov, V. Kohli, M. Koole, M. Krieguer, D. J. van der Laan, F. Lamare, G. Largeron, C. Lartizien, D. Lazaro, M. C. Maas, L. Maigne, F. Mayet, F. Melot, C. Merheb, E. Pennacchio, J. Perez, U. Pietrzyk, F. R. Rannou, M. Rey, D. R. Schaart, C. R.Schmidtlein, L. Simon, T. Y. Song, J-M. Vieira, D. Visvikis, R. Van der Walle, E. Wieers, and C. Morel. GATE: a simulation toolkit for PET and SPECT. *Phys. Med. Biol.*, 49:4543–4561, 2004.

Calvin A. Johnson. *Nonlinear Optimization for Volume PET Reconstructions*. PhD dissertation, George Mason University, Fairfax, Virginia, 1997. UMI Microform 9810569.

Calvin A. Johnson and Ariela Sofer. A primal-dual method for large-scale image reconstruction in emission tomography. *SIAM J. Optim.*, 11(3):691, 2000.

Calvin A. Johnson, Jürgen Seidel, and Ariela Sofer. Interior-point methodology for 3-D pet reconstruction. *IEEE Transactions on medical imaging*, 19:271, 2000.

Aviniash C. Kak and Malcom Slaney. *Principles of Computerized Tomographic Imaging*. Society of Industrial and Applied Mathematics, 1999. URL `http://rvl4.ecn.purdue.edu/~malcolm/pct/pct-toc.html`.

Frank Kehren. *Vollständige iterative Rekonstruktion von dreidimensionalen Positron-Emissions-Tomogrammen unter Einsatz einer speicherresidenten Systemmatrix auf Single- und Multiprozessor-Systemen*. PhD dissertation, Forschungszentrum Jülich, Jülich, Germany, 2001.

P. E. Kinahan, D. W. Townsend, T. Beyer, and D. Sashin. Attenuation correction for a combined 3D PET/CT scanner. *Med. Phys.*, 25(10):2046, 1998.

Paul E. Kinahan, Bruce H. Hasegawa, and Thomas Beyer. X-ray-based attenuation correction for positron emission tomography/computed tomography scanners. *Seminars in Nuclear Medicine*, XXXIII(3):166–179, July 2003.

Glenn F. Knoll. *Radiation Detection and Measurement, 3rd ed.* John Wiley & Sons, Inc., New York, 2000.

D. Lazaro, V. Breton, Z. El Bitar, and I. Buvat. Effect of noise and modeling error on the reliability of fully 3D Monte Carlo reconstruction in SPECT. In *Conference Record NSS/MIC*. IEEE, 2004a.

Delphine Lazaro, Vincent Breton, and Irene Buvat. Feasibility and value of fully 3D Monte Carlo reconstruction in single-photon emission computed tomography. *Nuclear Instruments and Methods in Physics Research A*, 527:195–200, 2004b.

Delphine Lazaro, Z. El. Bitar, Vincent Breton, D. Hill, and Irene Buvat. Fully 3D Monte Carlo reconstruction in SPECT: a feasibility study. *Phys. Med. Biol.*, 50: 3739–3754, 2005.

Martin J. Lercher and Klaus Wienhard. Scatter correction in 3-D PET. *IEEE Trans. Med. Imag.*, 13(4):649–657, 1994.

Craig S. Levin and Edward J. Hoffman. Calculation of positron range and its effet on the fundamental limit of positron emission tomography system spatial resolution. *Phys. Med. Biol.*, 44:781, 1999.

Craig S. Levin, M. Dahlbohm, and Edward J. Hoffman. A Monte Carlo correction for the effect of compton scattering in 3-D PET brain imaging. *IEEE Trans. Nucl. Sci.*, 42:1181, 1995.

Robert M. Lewitt. Alternatives to voxels for image representation in iterative reconstruction algorithms. *Phys. Med. Biol.*, 37(3):705–716, 1992.

Martin A. Lodge, Ramsey D. Badawi, Richard Gilbert, Pablo E. Dibos, and Bruce R. Line. Comparison of 2-dimensional and 3-dimensional acquisition for ^{18}F PET oncology studies performed on an LSO-based scanner. *J. Nucl. Med.*, 47(1):23–31, 2006.

Markiewicz, A. J. Reader, M. Tamal, P. J. Julyan, and D. L. Hastings. Towards an accurate voxel-based analytical unified scatter and attenuation system model for 3D PET. *IEEE Nuclear Science Symposium Conference Record*, 4:2199 – 2203, Oct 2004.

B. M. Mazoyer, M. S. Roos, and R. H. Huesman. Dead time correction and counting statistics for positron tomography. *Phys. Med. Biol.*, 30(5):385–399, 1985.

William W. Moses and S. E. Derenzo. Prospects for time-of-flight PET using LSO scintillator. *IEEE Trans. Nucl. Sci.*, 46:474–478, 1999.

National Institute of Standard and Technology NIST. XCOM: Photon cross sections data base. URL `http://physics.nist.gov/PhysRefData/Xcom/Text/XCOM.html`.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer-Verlag, New York, 1999.

Johan Nuyts. Nuclear medicine technology and techniques, 2000. URL `http://perswww.kuleuven.ac.be/~u0015224/`.

Johan Nuyts and Sigrid Stroobants. Reduction of attenuation correction artifacts in PET-CT. In *Conference Record NSS/MIC*. IEEE, 2005.

John M. Ollinger. Model-based scatter correction for fully 3D PET. *Phys. Med. Biol.*, 41(1):153–176, 1996.

John M. Ollinger and Jeffrey A. Fessler. Positron-emission tomography. *IEEE Signal Processing Magazine*, January:43–55, 1997.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge University Press, February 2002.

PVM 3.4.4. parallel virtual machine. URL `http://www.csm.ornl.gov/pvm/`.

Jinyi Qi and Richard M. Leahy. Iterative reconstruction techniques in emission computed tomography. *Phys. Med. Biol.*, 51:R541–R578, 2006.

M. Rafecas, G. Böning, B. J. Pichler, E. Lorenz, M. Schwaiger, and S. I. Ziegler. Inter-crystal scatter in a dual layer, high resolution LSO-APD positron emission tomograph. *Phys. Med. Biol.*, 48(7):821, 2003.

M. Rafecas, B. Mosler, M. Dietz, M. Pogl, A. Stamatakis, D. P. McElroy, and S. I. Ziegler. Use of a Monte Carlo-based probability matrix for 3-D iterative reconstruction of MADPET-II data. *Trans. Nucl. Sci.*, 51(5(2)):2597– 2605, 2004a.

Magdalena Rafecas, Guido Böning, Bernd J. Pichler, Eckhart Lorenz, Markus Schwaiger, and Sibylle I. Ziegler. Effect of noise in the probability matrix used for statistical reconstruction of PET data. *IEEE Trans. Nucl. Sci.*, 51(1):149–156, 2004b.

Alejandro Sanchez-Crespo, Pedro Andreo, and Stig A. Larsson. Positron flight in human tissue and its influence on PET image spatial resolution. *European Journal of Nuclear Medicine and Molecular Imaging*, 31(1):44–51, 2004.

Herbert Schildt. *C/C++ programmers reference*. Osborne/McGraw-Hill, 2000.

Herbert Schildt. *Teach yourself C++*. Osborne/McGraw-Hill, 1998.

S. Shoukouhi, P. Vaska, S. Southekal, D. Schlyer, M. Purschke, V. Dzordzhadze, C. Woody, S. Stoll, D. L. Alexoff, D. Rubins, A. Villanueava, and S. Krishnamoorthy. Statistical 3D image reconstruction for the RatCAP PET tomograph using a physically accurate, Monte Carlo based system matrix. In *Conference Record NSS/MIC*. IEEE, 2004.

Sepideh Shoukouhi. *Image Reconstruction and Image Performance Simulation of RatCAP (Rat Conscious Animal PET)*. PhD thesis, Stony Brook University, 2005.

Bjarne Stroustrup. *C++ Programmiersprache*. Addison-Wesley, 2000.

Peter Aundal Toft. *The Radon Transform - Theory and Implementation*. PhD dissertation, Department of Mathematical Modelling – Section for Digital Signal Processing – Technical University of Denmark, 2800 Lyngby, Denmark, 1996. URL `http://pto.linux.dk/PhD/`.

Michael Unser. Splines: A perfect fit for signal and image processing. *IEEE Signal Processing Magazine*, 16(2):22–38, 1999.

Stefaan Vandenberghe, Margaret E. Daube-Witherspoon, Robert M. Lewitt, and Joel S. Karp. Fast reconstruction of 3D time-of-flight PET data by axial rebinning and transverse mashing. *Phys. Med. Biol.*, 51:1603–1621, 2006.

Y. Vardi, L. A. Shepp, and L. Kaufman. A statistical model for positron emission tomography. *Journal of the American Statistical Association*, 80:7, 1985.

C. C. Watson. New, faster, image-based scatter correction for 3D PET. *IEEE Trans. Nucl. Sci.*, 47(4):1587–1594, 2000.

Alexander Werling, Olaf Bublitz, Josef Doll, Lars-Eric Adam, and Gunnar Brix. Fast implementation of the single scatter simulation algorithm and its use in iterative image reconstruction of PET data. *Phys. Med. Biol.*, 47:2947, 2002.

Wikipedia(PET). Positronen-emissions-tomography. URL `http://de.wikipedia.org/wiki/Positronen-Emissions-Tomographie`.

Sebastian Wilhelmi. *PVM++*, 2001-06-27. URL `http://pvm-plus-plus.sourceforge.net/`.

Donald W. Wilson, Benjamin M W Tsui, and Harrison H Barrett. Noise properties of the EM algorithm: Ii. Monte Carlo simulations. *Phys. Med. Biol.*, 39:847–871, 1994.

Habib Zaidi. Comparative evaluation of scatter correction techniques in 3d positron emission tomography. *European Journal of Nuclear Medicine*, 27:1813, 2000.

Habib Zaidi and Kenneth F. Koral. Scatter modelling and compensation in emission tomography. *European Journal of Nuclear Medicine and Molecular Imaging*, 31(5): 761–782, 2004.

Index

p_{dir}, 38
p_{leave}, 38
p_{int}, 38
p_{ratio}, 38
^{18}F-FDG, 2
$\mathcal{FT}$, 18
$\mathcal{FT}^{-1}$, 18
$\hat{x}_{\text{leave}}$, 36
2D PET scanner, 3
3D PET scanner, 8

activity distribution, 2
algebraic reconstruction technique, 25
analytic reconstruction algorithm, 17
APD, 8
arc correction, 20
ART, 25
attenuation, 13, 15
attenuation correction factor, 13
attenuation path length, 30
avalanche photo diodes, 8

B-spline basis function, 50
back-projection, 4, 20
back-projector, 24, 60
BGO, 8
bin, 20
biomarkers, 1
blank scan, 15

C-11, 2
coincidence, 35
coincidence time window, 10
collimation, 2
complementary slackness, 23
Compton effect, 11, 31, 32
computed tomography, 1, 17
crystal afterglow, 9
CT, 1, 16

dead time, 9
dead time loss, 9
delayed coincidence, 14
detector blocks, 8
direct coincidence, 11
discrete projection, 20
dual matrix approach, 5, 93
dual matrix expectation maximization, 61
dynamic image acquisition, 7

electronic collimation, 3
emission density, 2, 7, 17
emission measurement, 1
emission scan, 15
emission tomography, 1
energy window based scatter correction, 26
expected geometrical maximum, 61

F-18, 2
FBP, 19, 20
field of view, 37, 67
filtered back-projection, 19
fixed point equation, 24
Fourier slice theorem, 19, 20
Fourier transform, 18
FOV, 37
frame, 7
framing, 7
full matrix, 26, 59
full matrix approach, 5
FWHM, 7
FWTM, 8

gantry, 13, 20
GEANT4, 27, 68
Geiger mode avalanche photo diodes, 8
geometrically expected maximum, 56
GSO, 8

histogram mode, 9

I-123, 2
image reconstruction, 17
importance sampling, 35
interaction forcing, 36
inverse Fourier transform, 18
inverse transform method, 33
iterative reconstruction algorithm, 17

Karush-Kuhn-Tucker conditions, 23

Lagrange function, 23
Lagrange parameter, 23
likelihood, 22
line of response, 3, 11
linear attenuation coefficient, 28, 31
list mode, 7, 9
log likelihood, 22
LOR, 11
lower energy threshold, 13
LSO, 8

magnetic resonance imaging, 1
mashing, 25
mass attenuation coefficient, 11
maximum likelihood expectation maximization, 59
MC, 4
ML-EM algorithm, 24
Monte Carlo, 4
MRI, 1

N-13, 2
non-collinearity, 7, 29
normalization, 15
normalized root mean squared error, 67
normalized root mean variance, 68, 80
NRMSE, 67
NRMV, 68

O-15, 2
oblique sinogram, 10, 25
ordered subset expectation maximization, 24, 59, 95
OSEM, 24

parallel virtual machine, 42
PET, 1, 15, 16
PET/CT, 16
Photo effect, 31
photo multiplier tubes, 8
PMT, 8
positron emission tomography, 1
positron range, 7
positronium, 8
projection, 4, 17

projector, 24, 60
pseudo random number generator, 27, 29
PVM, 42
PVM++, 42

Radon transform, 17, 18, 44
RAMLA, 25, 95
ramp filter, 20
random coincidence, 11
random events, 14
RANMAR, 27, 29
Rayleigh scattering, 31
reconstruction, 2
ring PET scanner, 8
RND, 29
row action algorithm, 25

scatter fraction, 12
scatter order, 35
scatter projection, 49
scatter-free unit sinogram, 45
scattered coincidence, 11
scintillator crystals, 8
segmentation, 28
sensitivity, 2
septa, 3, 9
SimSET, 27
single event, 11
single photon emission computed tomography, 1, 17
sinogram, 9, 18, 21, 44
small animal PET, 21
sNRMSE, 67
sparse, 45
SPECT, 1
static image acquisition, 7
stratification cells, 39
system matrix, 4, 14, 21

time of flight PET, 12
TOF-PET, 12
tomography, 1
tracer, 1
transmission scan, 13, 15
transversal sinogram, 25
true coincidence, 11
trues, 11
tube of response, 3

ultrasound imaging, 1
unit scatter projection, 45
unit sinogram, 21, 44

variance reduction techniques, 35
volume imaging, 1

weight control, 42

xNRMSE, 67
XVMC, 28, 29

Acknowledgments

To Markus Alber, whom I owe the topic of the thesis, for financial support, for fruitful discussions, for the help with publications, and for (sometimes cynical) explanations of the world of (life) science.
To Prof. Schick for his support and kindness.

To my collegues:
To Matthias Söhn for numerous discussions, help with Mathematica, C++, Qt...
To Matthias Fippel for Monte Carlo support and his calm manner.
To Martin Soukup for Monte Carlo related discussions, Geant4 simulations, and his reliable presence in the office to order pizza late at night.
To Marcin Sikora and Urszula Jeleń for computer related help and climbing fun.
To Filippo Ammazzalorso for taking over my HTML duties.
To Tatiana Kleshneva and to Jan Muzik for being nice office mates.
For their assistance in teaching of medical technicians, I would like to thank Daniela Thorwarth with whom I could teach several semesters together, and also Christoph Baum and Oliver Dohm who introduced me to the teaching and Zanzem Atem Tung who later took over my job.
To Matthias Birkner with his calm attitude.
To Annemarie Bakai and Gustav Meedt for integration at the beginning and swimming.
To Freddy Haryanto for his good mood.

In addition I would like to express my special thanks
To Magdalena Rafecas for discussions about reconstruction and her very helpful manner. To Sibylle Ziegler for literature help and integration. To Maria-José Martinez for help concerning PET/CT. To Irène Buvat for discussions and reading my papers before submitting.

To my partner Margit Kiechle for her indulgence and patience.
And to my parents for their constant support.

A. Calculations

A.1. Gaussian sampling

Random numbers δE (section 4.2) that vary according to Gaussian statistics with standard deviation σ were obtained by applying the Box-Muller method to two uniformly distributed random variables $\mathrm{RND}_1 \in [0, 1[$ and $\mathrm{RND}_2 \in [0, 1[$.

$$\rho = \sqrt{(2 \max(0\,, -\ln(1 - \mathrm{RND}_1))} \tag{A.1}$$

$$\phi = 2\pi \mathrm{RND}_2 \tag{A.2}$$

$$\delta E = \sigma \rho \cos \phi \tag{A.3}$$

A.2. Variance of detected weighted counts

Let y be the counts of a LOR or the sum of the counts of LORs. Then y is a random variable calculated by

$$y = \sum_{k=1}^{\breve{N}} \breve{w}_k \,, \tag{A.4}$$

where the detected weights $\breve{w}$, as well as the number of detected weights $\breve{N}$ are random numbers as well. The variance of y is given by

$$\sigma^2 = \langle \breve{N} \rangle \sigma^2_{\breve{w}} + \sigma^2_{\breve{N}} \langle \breve{w} \rangle^2 \,. \tag{A.5}$$

Since the number of emissions follows Poissonian statistics, N is also distributed according to this statistics. The mean and the variance of N are therefore

$$\langle \breve{N} \rangle = \breve{N} \quad \text{and} \quad \sigma^2_{\breve{N}} = \breve{N} \,. \tag{A.6}$$

The variance of $\breve{w}$ is given by

$$\sigma^2_{\breve{w}} = \frac{1}{\breve{N}-1}\left(\sum_{k=1}^{\breve{N}} \breve{w}_k^2 - \frac{y^2}{\breve{N}}\right) \quad \text{with} \quad \frac{y}{\breve{N}} = \langle \breve{w} \rangle \,. \tag{A.7}$$

The preceding equations lead to

$$\sigma^2 = \frac{\breve{N}}{\breve{N}-1}\left(\sum_{k=1}^{\breve{N}} \breve{w}_k^2 - \frac{y^2}{\breve{N}}\right) + \frac{y^2}{\breve{N}} \quad \xrightarrow{\breve{N}\gg 1} \quad \sigma^2 = \sum_{k=1}^{\breve{N}} \breve{w}_k^2 \tag{A.8}$$

For reasonably large $\breve{N}$ the variance of the counts can be approximated as the sum of the squared detected weights. This derivation follows de Vries et al. [1990].

B. Paper: The influence of noise in full Monte Carlo ML-EM and dual matrix reconstructions in positron emission tomography

published in *Medical Physics*, 33(9):3498-3507, September 2006, doi: 10.1118/1.2239165

The influence of noise in full Monte Carlo ML-EM and dual matrix reconstructions in positron emission tomography

Niklas Rehfeld and Markus Alber
Sektion für Biomedizinische Physik, Universitätsklinikum Tübingen, Hoppe-Seyler-Str. 3, 72076 Tübingen, Germany

(Received 30 January 2006; revised 29 June 2006; accepted for publication 29 June 2006; published 31 August 2006)

Monte Carlo (MC) simulations in positron emission tomography (PET) play an important role in detector modeling and algorithm testing. Whereas the simulations are widely used in a forward projection manner to accomplish this task, ideally they should be included into the reconstruction process itself. It is therefore desirable to investigate the convergence properties and the propagation of MC noise of these kinds of reconstruction algorithms. MC simulations were integrated into the maximum likelihood expectation maximization (ML-EM) algorithm in two different ways. In the full matrix approach the system matrix was calculated by running MC simulations, including scatter. This matrix was used in both the projector and the backprojector. In the dual matrix (DM) approach, MC simulations were used to incorporate scatter in the projector, whereas the backprojector only comprised attenuation. Repeated reconstructions with different MC seeds allowed a statistical analysis of the error at each iteration step and made it possible to investigate separately the propagation of the MC noise that was introduced by the sinogram, by the projector, and by the matrix. Both approaches resulted in similar images, but the DM approach with unmatched projector and backprojector yielded a faster initial convergence when compared to the ideal full matrix approach. The analysis of the noise sources for the modeled single ring scanner in full matrix reconstruction showed that the noise introduced by the matrix became comparable to the noise introduced by the sinogram when using a matrix that was simulated with 10 000 emissions/voxel.
[DOI: 10.1118/1.2239165]

Key words: positron emission tomography, reconstruction, Monte Carlo, noise

I. INTRODUCTION

Monte Carlo (MC) simulations in positron emission tomography (PET) play an important role in detector modeling and algorithm testing.[1,2] Whereas the simulations are widely used in a forward projection manner to accomplish this task, ideally they should be included into the reconstruction process itself as proposed by Floyd *et al.*[3] (henceforth labeled the *full matrix* approach). The usage of a system matrix including all relevant parts in reconstruction (modeling of isotope, patient, scanner geometry, and detectors) should solve problems introduced to reconstruction by using simplified matrices. For human scanners the correct treatment of patient scatter plays a dominant role. The incorporation of this kind of scatter into the matrix is therefore crucial for full matrix reconstruction.

In the field of single photon emission computed tomography (SPECT), Lazaro *et al.*[4,5] reconstructed images with a matrix calculated by MC simulations including patient scatter. Another approach also using MC simulations to accurately calculate the (re-)projector was used by Beekman *et al.*[6] However, their algorithm (*dual matrix ordered subset*, DM-OS) used a simplified backprojector rendering of a storage of the matrix unnecessary. Both approaches can also be used in the case of PET reconstructions. While the first approach is theoretically superior and at the same time computationally very demanding, the second approach reduces the computational burden at the expense of a possible inaccuracy.

Therefore it appears desirable to compare these two methods for PET and investigate the reconstruction accuracy in light of the MC noise. For this purpose we developed a fast parallelized ring-PET MC code YaPRA to fulfill the task of scatter and attenuation simulation in the patient and the formation of a system matrix.

Simulations were run and matrices were calculated to accomplish three different tasks: First, the convergence properties of the full matrix reconstruction for different matrix statistics were investigated. Second, the noise propagation in the reconstruction process was examined. In order to be able to distinguish between the error introduced by the noisy sinogram or the noisy matrix, a method was developed to quantify the respective errors by running multiple MC simulations with different seeds. Lastly, the convergence properties and the error propagation of the full matrix approach and the *dual matrix maximum likelihood expectation maximization* (DM) approach were compared.

The simulated scanner was a single ring scanner. A three dimensional (3D) scanner could not be simulated, because the storage requirement of the full system matrix for such a scanner exceeded the available memory by far. Although the results of a single ring scanner should differ from more interesting results of 3D scanners, some features of the results

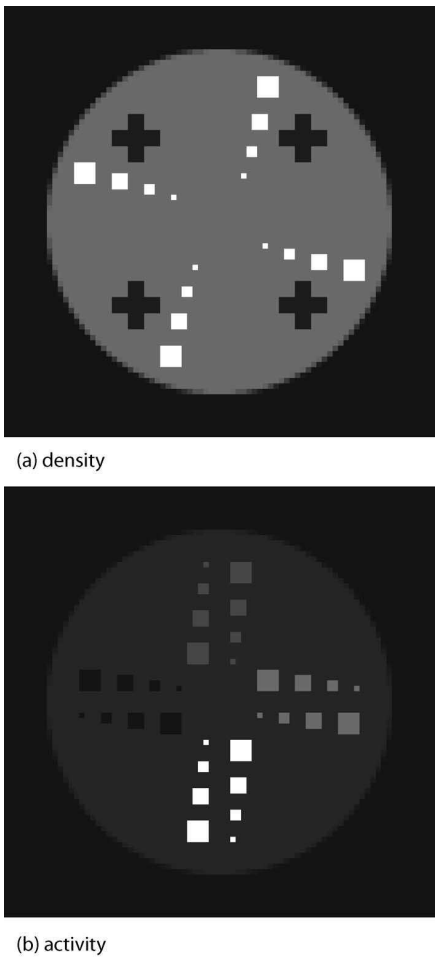


FIG. 1. Cylindrical phantom with density 0 (outside), 0.1 (cross), 1 (cylinder), and 2 (spots) g/cm^3 and activity ratio 0 (black):1:3:5:10 (white).

should be qualitatively similar or be weakly present, making the investigation of a single ring scanner worthwhile.

II. METHODS

A. Geometry and phantom

The ring of the scanner was modeled to be the surface of a cylinder divided into 384 detector units. The detector ring width of the cylinder was 6.45 mm and the radius was 41.21 cm. As a phantom a voxelized water cylinder of radius 16 cm was used. The depth of the water cylinder was 10 cm. Its center was located at the center of the scanner. The density of the phantom was approximated by a $80\times80\times1$ voxel grid of voxel dimensions $5\times5\times100$ mm^3 covering a total volume of $40\times40\times10$ cm^3 [see Fig. 1(a)]. The activity was described by the same grid, but reducing the voxels in the z direction to the depth of the scanner (6.45 mm) [Fig. 1(b)]. The scatter fraction of this ideal single ring scanner and the phantom shown in Fig. 1 was 4.2%.

All lines of responses (LORs) with a minimum of 96 detectors between the registrating pair of detectors were used for reconstruction. The LORs that were not taken into account lay outside the volume to be reconstructed.

B. Monte Carlo code

For the simulations, our MC code YaPRA for ring PET scanners was used. Similar to SimSET,[7,8] it uses the variance reduction techniques *stratified sampling and forced detection* (a specialization of importance sampling) in the phantom/patient. Both techniques, however, differed slightly from the ones used in the SimSET code. The detectors were idealized. Photons that hit the detector surface and exceeded a certain energy (here 350 keV) were counted. Dead time and single events were not simulated.

In order to translate density information into linear attenuation coefficients, the method of Fippel[9,10] was used. In contrast to the usual (discrete) segmentation approach it used functions fitted to ICRU data to map from density to linear attenuation coefficients. The advantage of this approach is the possibility to map directly from CT data to linear attenuation coefficients for the full bandwidth of human tissue.

Stratified sampling similar to the stratification described by Haynor *et al.*[7,8] was used to lessen inefficient starting directions for the photon pairs. Since the MC code is mainly used for system matrix calculation, stratified sampling for each voxel was performed.

The forced detection also followed the scheme described by Haynor *et al.*, but differed in the way of choosing proper scatter angles. Like in their approach, the distribution determining the scatter angles was a modified Klein-Nishina distribution, compensating the difference to the correct distribution by adjusting the weights of the photons properly. However, a different replacement for the Klein-Nishinia distribution was chosen, using an even smaller support. First the azimuthal scattering angle, φ, was sampled using a uniform distribution in the interval $[0,\pi]$. The choice of φ fixed a plane of interaction in which the incoming momentum vector and the outgoing (scattered) momentum vector of the photon must lie. Instead of sampling ϑ by using the (integral) Klein-Nishina distribution like in the unforced case, the possible angles that are used for sampling were reduced to those that guarantee a hit on the scanner surface (see Fig. 2). The contrast to the method of Haynor *et al.* lies in the fact that these proper intervals were calculated on the fly, assuring always the right scattering angles for every direction and position of the incoming photon. The advantage of our method is based upon the fact that for scanners without septa and/or photons positioned out of the field of view, good forced detection can still be achieved.

In order to speed up the calculation, a cluster of eight two-processor computers (AMD MP 2800+) together with the PVM library (PVM=*parallel virtual machine*) were used, which allowed a reduction of the calculation time by a factor of roughly 16.

The correctness of the code was tested by running simulations with YaPRA and with GEANT4.70pl with the same geometry and type of interactions (ideal detectors, no Rayleigh scattering). For this purpose, a scanner like in Sec. II A

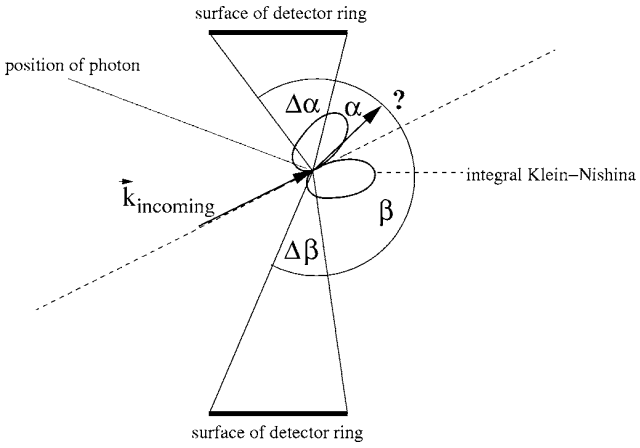


FIG. 2. Side view of the scanner: Instead of sampling the integral Klein-Nishina distribution in the interval $\vartheta \in [-\pi, \pi]$, it was only sampled in the intervals $[\alpha, \alpha+\Delta\alpha]$ and $[-\beta-\Delta\beta, -\beta]$, which ensured a hit on the detector if no further interactions occurred. The borders of these intervals were calculated using the intersection points of the two rings defining the scanner cylinder and the plane of interaction. Depending on the position and direction of the photon before interaction and the uniformly sampled angle φ, 0–2 valid intervals for ϑ did exist. Shown is the most likely case of two intervals.

was simulated that had a larger detector ring width of 10 cm in order to cope with the slower performance of GEANT4. The density grid was the same as described in Sec. II A, but the grid of the activity was adjusted to match the density grid (and the detector ring width). In order to facilitate the segmentation in GEANT4 slightly different densities [as compared to Fig. 1(a)] like lung and bone were used instead of 0.1 and 2 g/cm^3 in these comparative simulations. The sinograms of voxels (almost central and off-central) were compared (10^8 emissions in GEANT4 and 10^6 emissions in YaPRA with variance reduction). The detected counts (direct+scatter) as well as the primaries alone agreed to within 1%. In addition, no systematic difference in the sinograms could be found.

C. Sinogram and system matrix

In order to stay as realistic as possible, sinograms were calculated without applying variance reduction techniques. In this way the number of simulated particles could be directly translated into Becquerel/ml with the expected statistic uncertainties corresponding to real measurements. The number of simulated emissions were 1×10^9 and 5×10^9, which corresponds roughly to 5 and 30 min scans with (average) 6 Becquerel/ml initially.

The system matrix was calculated simulating a fixed number of photon pairs per voxel. Due to the high number of voxels, a simulation without variance reduction techniques was not possible. Both stratification and forced detection were used in the simulation. The number of simulated emissions per voxel were 1×10^4, 4×10^4, and 1.6×10^5. The elements m_{ji} of the matrix M were normalized to be the probability of the detection of an event in LOR j given one photon pair was started in voxel i.

TABLE I. The influence of the number of simulated particles on the fraction of nonzero elements in the matrix.

Emissions per voxel	Nonzero elements
160 000	40.5%
40 000	23.0%
10 000	9.8%
160 000 (no scatter)	1.9%

In Table I the fraction of nonzero elements for matrices of different statistics can be seen. The total number of matrix elements was 234,700,800. The maximal number of simulations (in the case of the 160 000-matrix) was 1.024×10^9 simulations. The calculation time on the 16 processor cluster was approximately 48 min for this matrix. The mere calculation time for the 10 000 matrix was less than four minutes.

D. Reconstruction

Images were reconstructed using the well known maximum likelihood expectation maximization algorithm (ML-EM). In the first reconstruction method the system matrix M including scatter was used in the projector $\mathcal{P}_{\text{full}}$ and also in the backprojector $\mathcal{B}_{\text{full}}$ of the algorithm. Therefore the same physics was used in the projector and in the backprojector. This approach with a matrix including scatter was called full matrix approach. With $m_{ji} \equiv [M]_{ji}$, $x_i \equiv [\mathbf{x}]_i$, and $y_j^* \equiv [\mathbf{y}^*]_j$ being the matrix elements, the unknown activity and the measured or (in this case) simulated sinogram, respectively, this algorithm is defined by the two equations:

$$\mathcal{P}_{\text{full}}: \quad \mathbf{y}^{(k+1)} = M\mathbf{x}^{(k)} \tag{1}$$

$$\mathcal{B}_{\text{full}}: \quad x_i^{(k+1)} = x_i^{(k)} \sum_{j=1}^{N_L} \left(\frac{y_j^* m_{ji}}{y_j^{(k+1)}} \right). \tag{2}$$

Here k is the iteration number and N_L is the number of LORs.

The ML-EM algorithm was used because it is clearly defined and well understood, and it forms the basis for several algorithms. The starting image was a uniform image with 1's in all voxels. This influenced, of course, the reconstruction, especially at early iterations. In order to show this effect, one reconstruction was performed using a starting image with random values $x_i \in [0,2]$ in each voxel.

A second set of images was reconstructed using the dual matrix (DM) reconstruction. DM is a reconstruction technique with unmatched projector/backprojector pairs. The projector is modeled by MC simulation (and therefore includes scatter in the patient), yet the backprojector does not include scatter. The advantages of this approach are the fact that the MC system matrix need not be stored and that fewer particles have to be simulated in total (in case of a reasonable number of iterations). The disadvantage is the relatively unpredictable behavior of the algorithm. Since the projector is changed at every iteration step due to the MC simulation, the progress made in one iteration can be canceled partially in

the next iteration. A similar problem arises from the fact that projector and backprojector differ in terms of scatter, therefore leading to suboptimal search directions.

The DM reconstruction was performed in the following manner: First, a MC matrix A was calculated ignoring all scattered events (simulated with 160 000 emissions/voxel). This matrix was used in the backprojector $\mathcal{B}_{\mathrm{DM}}$. The projector $\mathcal{P}_{\mathrm{DM}}$. used this matrix as well, but added an additional scatter sinogram $s^{(k)}$. This sinogram was calculated by performing a scatter-only MC simulation of the calculated activity $\mathbf{x}^{(k)}$ at iteration k. The efficiency of these simulations was enhanced by the variance reduction techniques described in Sec. II B:

$$\mathcal{P}_{\mathrm{DM}}: \quad \mathbf{y}^{(k+1)} = A\mathbf{x}^{(k)} + \mathbf{s}^{(k)}, \tag{3}$$

$$\mathcal{B}_{\mathrm{DM}}: \quad x_i^{(k+1)} = x_i^{(k)} \sum_{j=1}^{N_L} \left(\frac{y_j^* a_{ji}}{y_j^{(k+1)}} \right), \tag{4}$$

with $[A]_{ji} \equiv a_{ji}$.

The first iteration was performed with $s_0 \equiv 0$ and a first guess of the activity and the total number of emissions was obtained. Then a fixed fraction $p = 0.001, 0.0001$, or $0.000\,01$ of the guessed emissions $x_i^{(k)}$ were simulated in the following iterations. The obtained simulated scatter sinograms $s^{(k)}$ were scaled by a factor $1/p$ in order to correct for the lower number of simulated emissions.

E. Evaluation

In order to quantify the closeness to the true solution $\mathbf{x}^{\mathrm{true}}$, the *normalized root mean squared error* (or NRMSE for short) of the reconstructed images was calculated for each iteration step:

$$\mathrm{NRMSE} = \frac{1}{N_E} \sqrt{\frac{1}{N_V} \sum_{i=1}^{N_V} (x_i - x_i^{\mathrm{true}})^2},$$

$$N_V = \text{number of voxels},$$

$$N_E = \text{average number of emission per voxel}, \tag{5}$$

i.e.,

$$= \frac{5 \times 10^9}{N_V} \quad \text{or} \quad \frac{1 \times 10^9}{N_V}.$$

The NRMSE can be assigned to different sources of error. At early iterations, the algorithm is far from being converged and the major source of error can be accounted to this fact. If this were the only source of error, clearly the NRMSE should be monotonically decreasing. However, this is not the case. The inexact matrix and the noisy sinogram must be responsible for the positive slope of the NRMSE at higher iterations.

It is possible to quantify the error induced by the sinogram by simulating N sinograms that are identical but do have different MC seeds. In this way, for each iteration step k and each voxel i the variance $\sigma^2_{i\ \mathrm{sinogram}}(k)$ caused by the noise in the sinogram can be calculated:

$$\sigma^2_{i\ \mathrm{sinogram}}(k) = \frac{1}{N-1} \sum_{\alpha=1}^{N} (x_{i,\alpha}^{(k)} - \bar{x}_i^{(k)})^2,$$

$$\bar{x}_i^{(k)} \equiv \text{mean value of voxel } i \text{ at iteration } k. \tag{6}$$

Although the values $x_{i,\alpha}^{(k)}$ are probably not exactly following Gaussian statistics, the second central moment is nevertheless a useful measure of the error.

The same approach is applicable to the system matrix. The calculation of several matrices with different seeds is, of course, more time consuming than just calculating the sinograms. For this reason only a small number of matrices were calculated (nine matrices). The same number of sinograms was simulated for consistency. Because of the small number of N simulations a further analysis by means of bootstrap methods was not pursued.

In the case of the DM reconstructions, three sources of error exist: the sinogram, the matrix A, and the MC scatter projection (leading to the sinogram $s^{(k)}$). The influence of each source can be measured again by varying the seed of the corresponding MC simulation and keeping the two other seeds constant.

Analogously to the NRMSE, a measure for the total induced error can be introduced, *normalized root mean variance*, or for short, NRMV henceforth:

$$\mathrm{NRMV}(k) = \frac{1}{N_E} \sqrt{\frac{1}{N_V} \sum_{i=1}^{N_V} \sigma_i^2(k)}. \tag{7}$$

It is possible to relate this quantity to the previously introduced NRMSE. At iteration k the image $\mathbf{x}^{(k)}$ is obtained by successive application of the projector and backprojector operators on the starting image $\mathbf{x}^{(0)}$. The backprojector is containing explicitly the sinogram $\mathbf{y}^*$ while both depend on the system matrix M.

$$\mathbf{x}^{(k)} = \left(\prod_{n=1}^{k} \mathcal{B}\mathcal{P} \right) \mathbf{x}^{(0)} \equiv \mathbf{F}^{(k)}_{\mathbf{y}^*, M}(\mathbf{x}^{(0)}). \tag{8}$$

With the help of a Taylor expansion, it is possible to estimate the total error $\epsilon_i(k) = x_i(k) - x_i^{\mathrm{true}}$,

$$\epsilon_i^2(k) \approx \underbrace{\sum_j \left(\frac{\partial F_i^{(k)}}{\partial y_j^*} \right)^2 \Delta y_j^{*2}}_{\approx \sigma^2_{i\ \mathrm{sinogram}}(k)} + \underbrace{\sum_{j,l,r,s} \frac{\partial F_i^{(k)}}{\partial m_{jl}} \frac{\partial F_i^{(k)}}{\partial m_{rs}} \mathrm{Cov}(m_{jl} m_{rs})}_{\approx \sigma^2_{i\ \mathrm{matrix}}(k)} + (\text{convergence error})^2. \tag{9}$$

Here the noise in the sinograms bins (Δy_j^*) and the matrix elements (Δm_{ji}) are not correlated, because different simulations (with different MC seeds) are started. A correlation between two matrix elements, however, is, in principle, possible due to the variance reduction techniques that are used

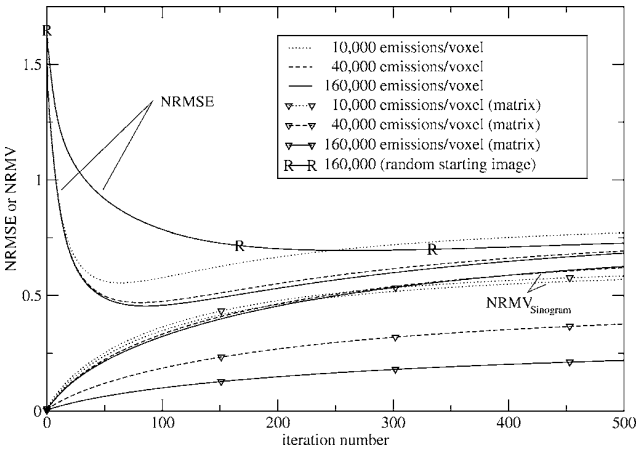


FIG. 3. Error versus iteration number for full matrix reconstructions with different statistics, as indicated. The R-R graph shows the NRMSE when using a starting image with random voxels $x_i \in [0,2]$ instead of a uniform image with voxels $x_i \equiv 1$. The sinogram was simulated with 5×10^9 emissions in total. Iteration number zero corresponds to errors of images after the first iteration.

for the simulation of the elements. The relation (9) is Gauss' law of error propagation modified due to the potential correlation of the matrix elements $\mathrm{Cov}(m_{jl},m_{rs})$ and with an additional term to account for the fact that the algorithm is not converged. This latter addend decreases to zero for large k in case of convergence. Therefore, for large k the convergence error can be neglected and the following inequality can be derived using the triangle inequality:

$$\mathrm{NRMSE}(k) \lesssim \mathrm{NRMV}_{\mathrm{sinogram}}(k) + \mathrm{NRMV}_{\mathrm{matrix}}(k). \qquad (10)$$

This inequality is also valid when using higher order Taylor expansions, which is necessary in the case of large errors.

III. RESULTS

In Fig. 3, the NRMSE of the reconstructed images with matrices of different statistics can be seen. The figure shows the typical property of the iterative solution of an (unregularized) ill-posed problem: after a relatively fast convergence the algorithm starts to " focus" on the noise and to drift away from the true solution.

Clearly better statistics resulted in a smaller NRMSE, but the difference between the 1.6×10^5 matrix and the 4×10^4 matrix was already much smaller than the difference between the latter and the 1×10^4 matrix. The better the matrix statistics the more the minimum (best agreement between true and reconstructed image) is shifted toward higher iterations.

Figure 4 shows the NRMSE using the same matrices as in Fig. 3, but applying the algorithm to the low count sinogram with 1×10^9 emissions. The error introduced by the sinogram is bigger. The shape of the NRMSE at higher iterations is mostly determined by this error. The bigger total noise induced error results in a shift of the minimum of the NRMSE toward early iterations. At the beginning, there is only a small deviation between the corresponding NRMSE curves in Figs. 3 and 4. This is caused by the strong influence of the starting image, which is in both cases the same. In both figures the validity of inequality (10) for large iteration numbers can be verified.

Figure 5 shows the relative importance of the matrix error vs. the sinogram error in the case of full matrix reconstruction. The matrix error became comparable to the error caused by the sinogram for the 5×10^9 sinogram and the 10 000 matrix (i.e., the ratio is approximately one). The ratio can be used to determine the required number of simulated emissions per voxel: the error caused by the sinogram should be larger than the error caused by the matrix. At higher iteration numbers the ratio of the influences of both error sources is independent of the iteration number.

Figure 6 shows the NRMSE for full matrix and DM reconstructions with different statistics. For the DM reconstructions, a fraction p of 0.001, 0.0001 or 0.000 01 of the guessed activity was simulated in the forward MC simulation, corresponding to approximately 5×10^6, 5×10^5, or 5×10^4 simulated emissions. More simulated particles led to smaller NRMSE, as expected.

A characteristic feature of the DM reconstruction is the faster initial convergence. Very typical are also the noise like

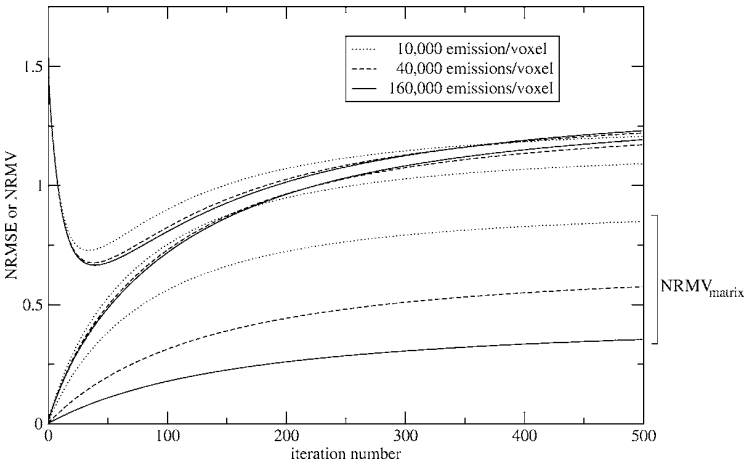


FIG. 4. Error versus iteration number for full matrix reconstructions of a sinogram simulated with 1×10^9 emissions in total and matrices of different statistics.

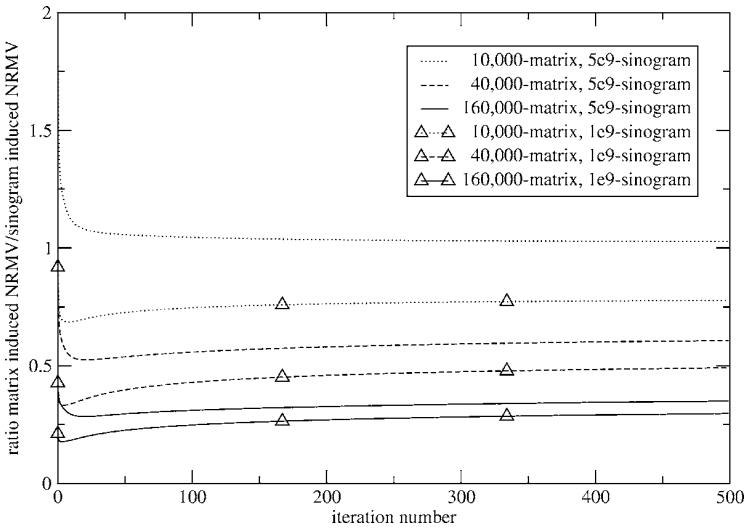


FIG. 5. The ratio of the $NRMV_{matrix}/NRMV_{sinogram}$ for a sinogram with 1×10^9 emissions and a sinogram with 5×10^9 emissions is shown for different matrix and sinogram statistics.

fluctuations that can be seen in the $p=0.000\,01$ curve. This feature is also very weakly present in the $p=0.0001$ curve (and extremely weak in the $p=0.001$ curve). These fluctuations decreased the more particles were simulated in the forward MC simulation.

Figure 7 shows the contribution of the different sources of error for the DM reconstruction. Clearly, the error caused by the sinogram dominated. The error introduced by the system matrix A had similar features like the matrix induced error in the full matrix reconstruction.

The error introduced by the forward MC scatter increased initially and soon stayed rather constant. This suggests that the algorithm converges to some mean solution and oscillates randomly around this solution with rather constant mean amplitude.

In analogy to Fig. 5, Fig. 8 shows the importance of the error introduced by matrix A and the forward MC scatter simulation relative to the sinogram induced error. Again, the error caused by the sinogram dominates. The small error due to the forward MC scatter simulation can be explained by the fact that the system was a 2D system with a low scatter contribution. Therefore the scatter free matrix A that occurs both in the projector and backprojector mostly determines the convergence properties. The error ratio introduced by matrix A shows similar properties, like the error ratio in full matrix reconstruction (a constant nonzero ratio at high iteration numbers). This is not the case for the error introduced by the MC forward simulation.

Figure 9 shows the reconstructed images. Clearly the difference is not very big, which can be accounted to the relatively small scatter fraction.

In the following figures, the voxel dependency of the error is shown. Images of minimal NRMSE were chosen. The likelihood of the EM algorithm was not used to define the iteration number of the images, because the likelihood itself depends on the quality of the matrix, but not on the difference to the true activity. The choice of the minimal NRMSE as a criteria for selection in general leads to different iteration numbers. This should be kept in mind when comparing images. The absolute error (standard deviation) or the relative error (standard deviation divided by the mean value) for each voxel is visualized as the gray value of the corresponding voxel (white ≡ big error, black ≡ small error).

Figure 10 shows the absolute and relative error caused by the sinogram. The absolute error did not depend much on the reconstruction method. More emissions in a voxel resulted in a larger absolute error, which is in agreement with a study on noise properties of the EM algorithm by Wilson *et al.*[11]

On the other hand, more emissions lead to a reduced relative error. In both full matrix and DM reconstruction, the error seemed not to depend on the phantom density of the voxels. The density change at the border of the phantom, however, clearly influenced the relative error introduced by the sinogram. This is probably caused by the fact that both algorithms converged much faster outside the phantom to small (zero) activities.

In Fig. 11 the error caused by the system matrix is visualized. A density dependency inside the phantom could not

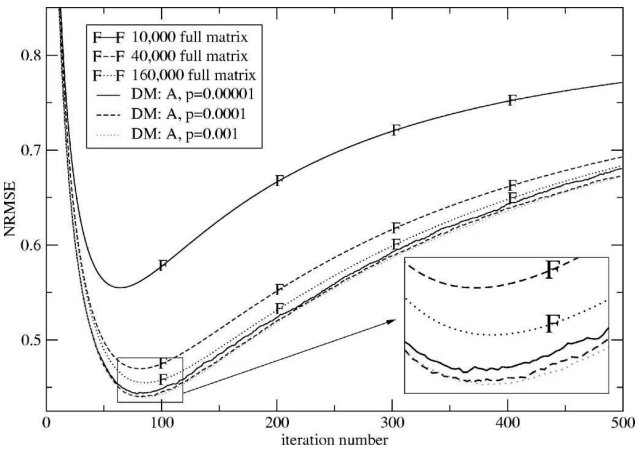


FIG. 6. NRMSE for full matrix reconstructions and DM reconstructions. Sinogram 5×10^9 emissions.

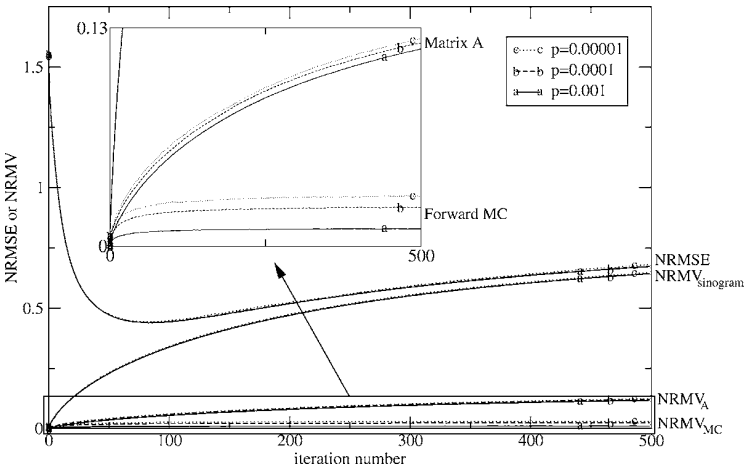


FIG. 7. NRMSE and NRMV of DM reconstruction with matrix A for the sinogram with 5×10^9 emissions and different fractions of simulated forward scatter. The small figure shows more enlarged NRMV curves caused by the matrix A and by the forward MC simulation.

be found. Again a higher number of emissions lead to an increased absolute error but a reduced relative error. The same behavior can be seen in DM reconstruction (Fig. 12) for the error caused by the scatter-free matrix A. In the case of the error introduced by the forward scatter simulation, a similar behavior can be seen, but, in addition, the noise outside the phantom seemed to be structured [see Fig. 12(d)]. However this could be seen only in the relative error images, because the reconstructed activity outside the phantom is small. Therefore at least in 2D this effect seems to be negligible.

IV. DISCUSSION

The log-likelihood that is to be maximized by the algorithm is based on the approximated noisy matrix and the approximated noisy sinogram. The algorithm does therefore not converge to the maximum of the ideal problem but to a shifted maximum (image $\mathbf{x}^{\infty}$). Before converging to this shifted maximum, the NRMSE can even become smaller than at higher iterations, because the uniform starting image encourages smooth reconstructed images that often agree better with the original image $\mathbf{x}^{\text{true}}$. The fading influence of this starting image at higher iteration numbers leads to a positive slope of the NRMSE. The NRMSE eventually approaches asymptotically $\text{NRMSE}(\mathbf{x}^{\infty}) > \text{NRMSE}(\mathbf{x}^{\text{true}}) \equiv 0$.

This explanation can be verified by looking at graph R in Fig. 3, where the starting image is an image with random voxel values between zero and two times the mean expected activity. The agreement between the true image and this random starting image is much smaller. This leads to an almost horizontal slope at higher iterations and a shift of the minimum of the NRMSE toward higher iterations.

The fading influence of the starting image during the reconstruction process is associated with a growing influence of the given information (the matrix and the sinogram) on the reconstructed images. Since the matrix and the sinogram are noisy, the corresponding NRMVs should grow as well. This can be verified in Fig. 3, which shows monotonously increasing NRMVs.

In both Fig. 3 and in Fig. 4 the NRMV curve of the sinogram induced error with the 10 000 matrix is below the two other curves using the higher statistics matrices for iteration numbers $\gtrsim$230 and $\gtrsim$170, respectively. This means that the NRMV of the sinogram somehow depends on the number of the simulated emission of the matrix. This can have three reasons. The derivative of $F^{(k)}_{\mathbf{y}^*,M}$ in (9) depends implicitly on the matrix. Therefore, the NRMV of the sinogram should be positioned randomly around some mean NRMV curve. In the case of bad matrix statistics this deviation should be larger and the considered curve could be positioned below the two other curves. However, this is not very likely, because the NRMV is the average over many voxels. Second, in the case of few simulated emissions, the nonzero fraction of the matrix is small. This might lead to a reduced rank of the matrix which should also influence the mentioned derivative. Third, it is possible that due to a large matrix error higher order terms in the Taylor expansion cannot be neglected anymore. This would lead also to a dependency of the considered NRMV on the matrix statistics.

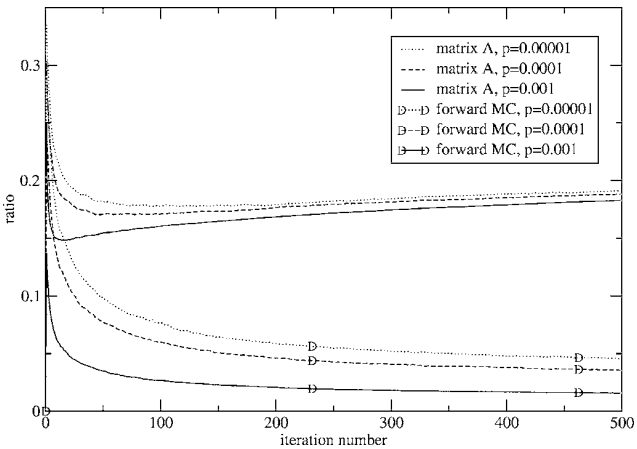
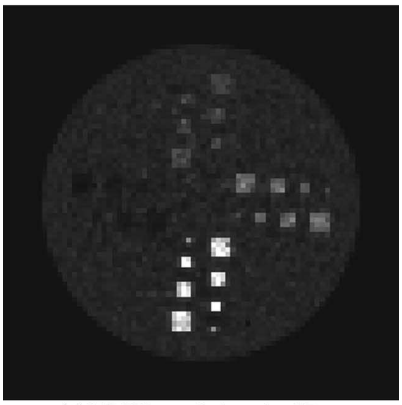
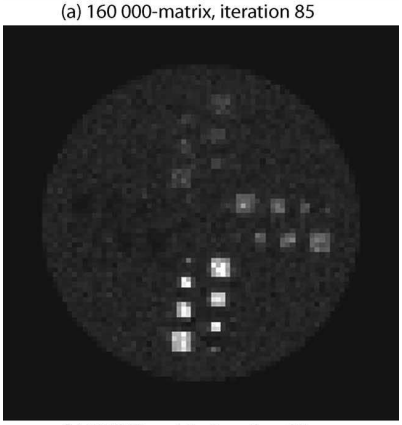
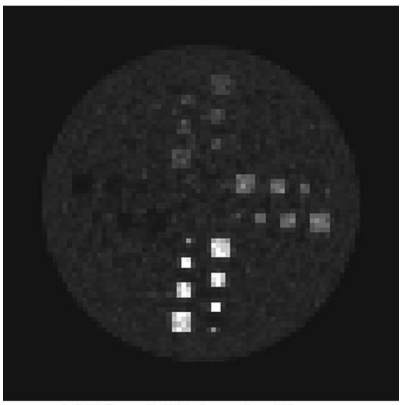


FIG. 8. Ratios $\text{NRMV}_A/\text{NRMV}_{\text{sinogram}}$ and $\text{NRMV}_{\text{MC scatter}}/\text{NRMV}_{\text{sinogram}}$ in DM reconstruction.


(a) 160 000-matrix, iteration 85


(b) 10 000-matrix, iteration 64


(c) DM, p=0.001, iteration 83

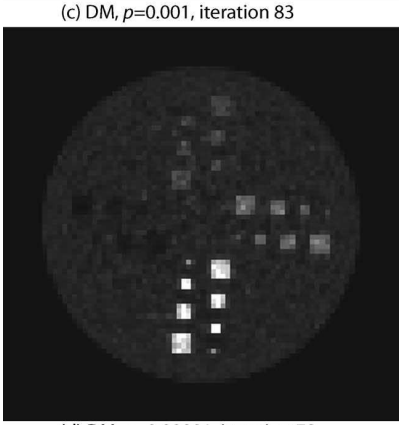
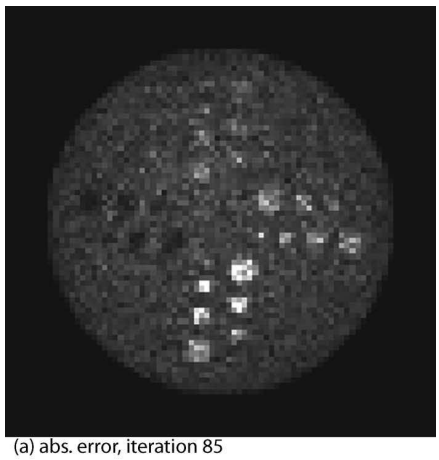

(d) DM, p=0.00001, iteration 78

FIG. 9. Reconstructed images for the 5×10^9 emissions sinogram at minimal NRMSE.


(a) abs. error, iteration 85

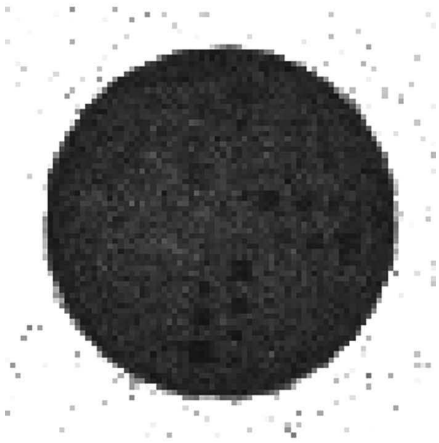

(b) rel. error, iteration 85

FIG. 10. Sinogram induced absolute and relative error for the 5×10^9 emissions sinogram and the 160 000 matrix at minimal NRMSE (white ≡ big error, black ≡ small error). The very large relative error of voxels outside the cylinder often exceeded the gray value scaling. The error of these voxels is therefore represented by white color.

An additional algorithm was also tested. The algorithm is closely related to the DM algorithm

$$\mathcal{P}': \quad \mathbf{y}^{(k+1)} = \mathbf{M}\mathbf{C}'(\mathbf{x}^{(k)}) \tag{11}$$

$$\mathcal{B}_{\mathrm{DM}}: \quad x_i^{(k+1)} = x_i^{(k)} \sum_{j=1}^{N_L} \left(\frac{y_j^* a_{ji}}{y_j^{(k+1)}} \right). \tag{12}$$

It uses a projector $\mathcal{P}'$ that only consisted of a MC simulation (with unscattered *and* scattered events) with the same number of simulated emissions like in $\mathcal{P}_{\mathrm{DM}}$. This tested algorithm that is not using the matrix A in the projector proved to be clearly inferior to the DM algorithm. Even a reconstruction with p=0.001 (or roughly 5×10^6 emissions per iteration) was unstable. Problems arose always when a LOR $y_i^{(k+1)}$ in the denominator of (12) was zero while the nominator $y_j^* a_{ji}$ was not zero. This problem occurred mainly for LORs tangential to the phantom boundary. This unstable behavior of the algorithm could be avoided by defining a projector $\mathcal{P}_{\mathrm{DM}}$, as proposed in Sec. II D.

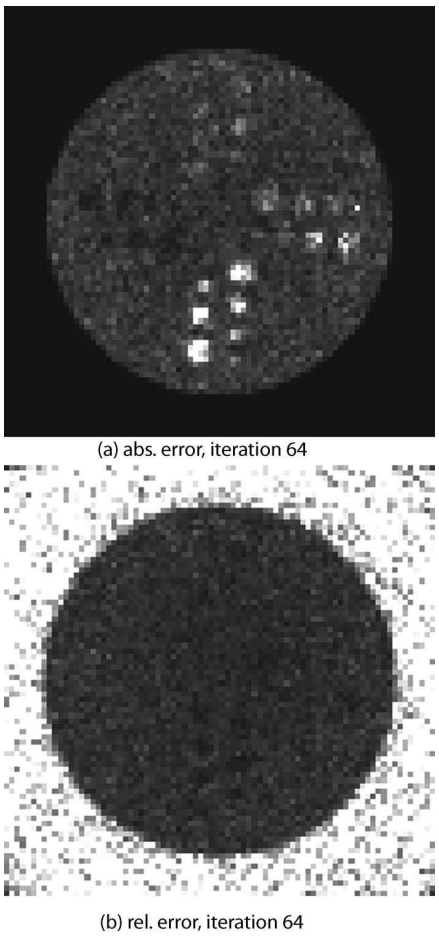


FIG. 11. Matrix induced absolute and relative error for the 5×10^9 sinogram and the 10 000 matrix at minimal NRMSE (white ≡ big error, black ≡ small error).

Simulations of a similar one-ring scanner with a larger detector ring width (and phantom), and therefore scatter fraction, could give more insight into the noise propagation and performance of the two algorithms in 3D scanners. Preliminary simulations of such a scanner showed that relative to the full matrix algorithm the minimum of the NRMSE curve of the DM algorithm is positioned at smaller iteration numbers, while the minimal NRMSE is increasing. It can be therefore expected that in the case of 3D scanners the DM algorithm is initially converging faster than the full matrix algorithm.

While the position of the minimum of the DM approach relative to the full matrix approach seems to be rather sensitive to the scatter fraction, it can be expected that the qualitative shape of the NRMV curves (monotonously increasing influence of matrix or sinogram noise) should stay the same. For large iteration numbers, there should be an upper bound $\mathrm{NRMSE}(\mathbf{x}^\infty)$ for the NRMSE, as discussed at the beginning of this section. Together with (9), this suggests that there exists an upper bound for the NRMVs as well, independently of the scatter fraction. The property of the $\mathrm{NRVM}_{\mathrm{MC}}$ of staying at a constant level after only few iterations is probably

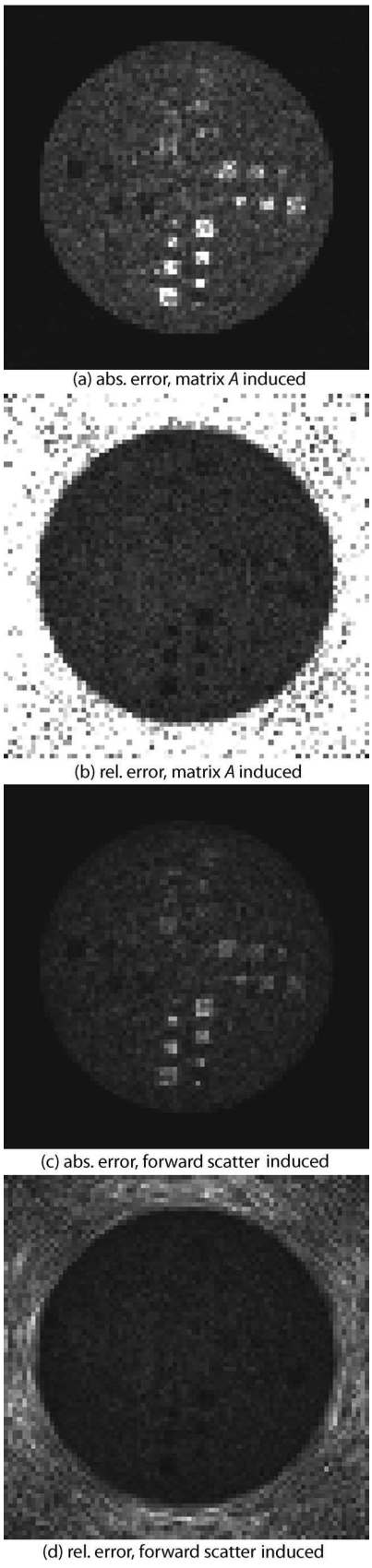


FIG. 12. Error in DM reconstruction with $p = 0.000\,01$ at iteration 78 = minimal NRMSE (white ≡ big error, black ≡ small error).

also present in the case of 3D scanners. It can be expected that the influence of this forward scatter simulation will increase. While in the case of 3D scanners the variance of the voxels will most likely also be correlated to the activity of these voxels, there might be an additional correlation to the density that was not present in the 2D case.

Apart from the crucial question of the performance and noise propagation in the case of 3D scanners, it would also be interesting to compare the convergence and noise properties of the full matrix approach and DM approach using different starting images (like images obtained by filtered backprojection, which might lead to superior convergence of the full matrix approach), and also to investigate different reconstruction algorithms (like, for example, primal-dual methods[12,13]) or the influence of regularization.

V. CONCLUSIONS

A method to quantify the error introduced by MC simulations in reconstruction algorithms was introduced. The noise propagation and performance of the ML-EM algorithm in the full matrix and DM version were investigated in a single ring scanner.

The full matrix simulations showed that the noise introduced by the matrix and by the sinogram became comparable in the case of reconstructions with a system matrix with a relatively small number of simulated emissions per voxel (10 000 with variance reduction). This matrix was simulated in less than four minutes calculation time on a small computer cluster, which shows that the much more demanding task of the simulation of a matrix of a 3D scanner might become feasible. The storage of the matrix, however, is still problematic, although approaches like compression of the matrix[14] were presented that might reduce the severity of this limiting factor.

A simpler way to avoid the storage problem is to use the DM algorithm for reconstruction. In a broader sense, not only the DM-OS algorithm from Beekman *et al.*, but also the well known single scatter simulation algorithm[15] falls into this class of algorithms. In the case of the investigated idealized one-ring scanner and a uniform starting image, the DM algorithm performed slightly better than the algorithm using the full scatter matrix. Although theoretically the DM approach should be inferior to the full matrix approach, because it is not guaranteed to converge and additional noise by the forward MC simulation is introduced into the algorithm, it performed very well.

ACKNOWLEDGMENTS

We would like to thank M. Fippel for his help, many useful discussions, and for permission to include and modify parts of his code XVMC in our MC code. In addition, we would like to thank M. Soukup for comparative simulations with GEANT4.

[1] I. Buvat and I. Castiglioni, "Monte Carlo simulations in SPET and PET," Q. J. Nucl. Med. **46**, 48 (2002).
[2] H. Zaidi, "Relevance of accurate Monte Carlo modeling in nuclear medical imaging," Med. Phys. **26**, 574 (1999).
[3] C. E. Floyd, Jr., R. J. Jaszczak, K. L. Greer, and R. E. Coleman, "Inverse Monte Carlo as a Unified Reconstruction Algorithm for ECT," J. Nucl. Med. **27**, 1577 (1986).
[4] D. Lazaro, Z. E. Bitar, V. Breton, D. Hill, and I. Buvat, "Fully 3D Monte Carlo reconstruction in SPECT: A feasibility study," Phys. Med. Biol. **50**, 3739 (2005).
[5] D. Lazaro, V. Breton, and I. Buvat, "Feasibility and value of fully 3D Monte Carlo reconstruction in single-photon emission computed tomography," Nucl. Instrum. Methods Phys. Res. A **527**, 195 (2004).
[6] F. J. Beekman, H. W. A. M. de Jong, and S. van Geloven, "Efficient fully 3-D iterative SPECT reconstruction with Monte Carlo-based scatter compensation," IEEE Trans. Med. Imaging **21**, 867 (2002).
[7] D. R. Haynor, R. L. Harrison, T. K. Lewellen, A. N. Bice, C. P. Anson, S. B. Gillispie, R. S. Miyaoka, K. R. Pollard, and J. B. Zhu, "Improving the efficiency of emission tomography simulations using variance reduction techniques," IEEE Trans. Nucl. Sci. **37**, 749 (1990).
[8] D. R. Haynor, R. L. Harrison, and T. K. Lewellen, "The use of importance sampling techniques to improve the efficiency of photon tracking in emission tomography simulations," Med. Phys. **18**, 990 (1991).
[9] M. Fippel, "Fast Monte Carlo dose calculation for photon beams based on the VMC electron algorithm," Med. Phys. **26**, 1466 (1999).
[10] M. H. Fippel, "Entwicklung eines schnellen Monte-Carlo-Verfahrens zur Dosisberechnung in der Strahlentherapie," Habilitation, Eberhard-Karls-Universität zu Tübingen, 2000.
[11] D. W. Wilson, B. M. W. Tsui, and H. H. Barrett, "Noise properties of the EM algorithm: II. Monte Carlo simulations," Phys. Med. Biol. **39**, 847 (1994).
[12] C. A. Johnson, J. Seidel, and A. Sofer, "Interior-point methodology for 3-D PET reconstruction," IEEE Trans. Med. Imaging **19**, 271 (2000).
[13] C. A. Johnson and A. Sofer, "A primal-dual method for large-scale image reconstruction in emission tomography," SIAM J. Optim. **11**, 691 (2000).
[14] N. Rehfeld, M. Fippel, and M. Alber, "Reconstruction of PET Images with a Compressed Monte Carlo Based System Matrix—A Comparison to Other Monte Carlo Based Algorithms," in *Conference Record NSS/MIC*, IEEE, 2005.
[15] A. Werling, O. Bublitz, J. Doll, L.-E. Adam, and G. Brix, "Fast implementation of the single scatter simulation algorithm and its use in iterative image reconstruction of PET data," Phys. Med. Biol. **47**, 2947 (2002).

C. Conference proceedings

C.1. Monte Carlo noise in full Monte Carlo ML-EM and dual matrix recsontructions in positron emission tomography

published in *Nuclear Instruments and Methods in Physics Research A*, 571:211–214, 2007, doi: 10.1016/j.nima.2006.10.065
presented at the 1st European Conference on Molecular Imaging Technology, Marseille, May 2006

Available online at www.sciencedirect.com
ScienceDirect
Nuclear Instruments and Methods in Physics Research A 571 (2007) 211–214
NUCLEAR INSTRUMENTS & METHODS IN PHYSICS RESEARCH Section A
www.elsevier.com/locate/nima

Monte Carlo noise in full Monte Carlo ML-EM and dual matrix reconstructions in positron emission tomography ☆

Niklas Rehfeld*, Markus Alber

Sektion für Biomedizinische Physik, Klinik für Radioonkologie Universitätsklinikum Tübingen, Hoppe-Seyler-Str. 3, 72076 Tübingen, Germany

Available online 7 November 2006

Abstract

Monte Carlo (MC) simulations in positron emission tomography (PET) play an important role in detector modeling and algorithm testing. Nowadays, these simulation are also increasingly used for scatter correction during reconstruction. This can be done ideally by using MC simulations to calculate the system matrix including scatter (*full matrix approach*). Another approach to incorporate MC simulations into the reconstruction is using a MC based projector and attenuation based back-projector, avoiding the storage of the matrix (*dual matrix* (DM) *approach*). It appears desirable to compare these two methods for PET and investigate the reconstruction accuracy in the light of MC noise. For this purpose a method to estimate the error introduced by the matrix, the sinogram or the projector based on repeated simulations with different MC seeds is introduced. Simulations of a single ring scanner (due to storage limitations) were performed.

PACS: 87.57.−s

Keywords: Positron emission tomography; Reconstruction; Monte Carlo; Noise

1. Introduction

In the field of single photon emission computed tomography [1] as well as in the case of small animal positron emission tomography (PET) [2,3] system matrices were simulated in the last years by means of Monte Carlo (MC) methods to accurately model the scanner [1–3] and the scatter in the patient or small animal [1,3]. Although the correct simulation of the system matrix with MC methods of sufficient statistics guarantees the correct treatment of difficult scanner geometries and scatter in the reconstruction, this approach is problematic due to very long simulation times and very large system matrices. The problem of storing the matrix can be avoided by using a MC based projector including scatter, but using a simpler back-projector without scatter [4]. Especially in the case of human PET the latter approach is very appealing, because the size of the system matrix. Therefore, it appears desirable to compare these two methods for PET and investigate the reconstruction accuracy in the light of the MC noise.

2. Methods

2.1. Geometry and MC simulations

The modeled scanner was an ideal one-ring scanner with 0.645 cm scanner width and a diameter of 82.4 cm. The phantom was described by $80 \times 80 \times 1$ voxels and had the total dimensions of $40 \times 40 \times 0.645\,\mathrm{cm}^3$ (see Fig. 1). The simulated phantom was cylindrical with density 0(outside), 0.1, 1(cylinder), and $2\,\mathrm{g/cm}^3$ and an activity distribution with the activity ratios 0(outside):1(cylinder):3:5:10. The used fast parallelized ring-PET MC code YaPRA concentrated on phantom scatter and used ideal detector physics: no dead time simulation, ideal energy resolution with detection of photons $\geqslant 350$ keV. Singles were not simulated.

☆Paper presented at the 1st European Conference on Molecular Imaging Technology, Marseille, May 2006.

*Corresponding author.
E-mail address: niklas.rehfeld@med.uni-tuebingen.de (N. Rehfeld).

0168-9002/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.nima.2006.10.065

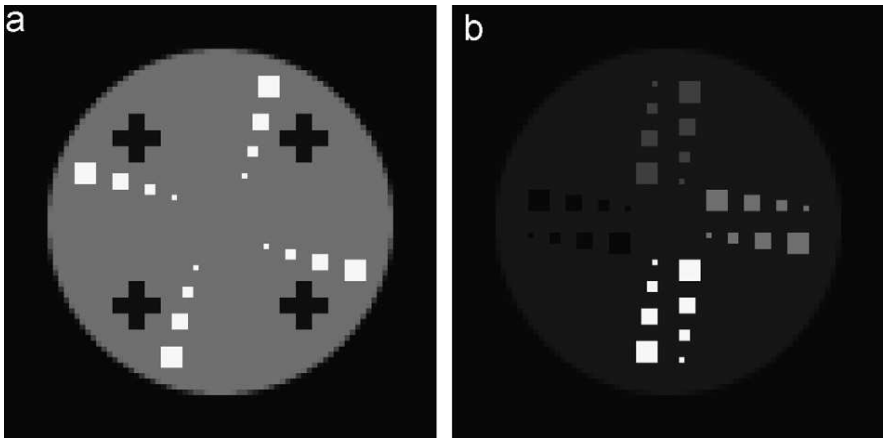


Fig. 1. Phantom: (a) density; (b) activity $\mathbf{x}^{\text{true}}$.

2.2. Sinograms and system matrices

The sinogram was simulated without using variance reduction techniques with 5×10^9 emissions (high statistics) and 1×10^9 emission (lower statistics) which corresponds roughly to 5 and 30 min-scans with (average) 6 Becquerel/ml initially. The system matrices including scatter (M) were calculated with 160 000, 40 000, and 10 000 emission per voxel using the variance reduction techniques stratification and forced detection. The scatter-free matrix (A) was simulated with 160 000 emission per voxel using the same variance reduction techniques.

2.3. Reconstruction

The images were either reconstructed using the ML-EM algorithm with a matrix M of different statistics (full matrix approach) or using the attenuation only matrix A in the back-projector and a forward projector $\mathscr{P}$ including MC scatter estimates based on the previously reconstructed activity $\mathbf{x}^{(k)}$ at iteration number k (dual matrix (DM)-approach). In both cases a uniform image ($\mathbf{x} \equiv 1$) was the starting image.

Full matrix:

$$x_i^{(k+1)} = x_i^{(k)} \sum_{j=1}^{N_L} \left(\frac{y_j^\star m_{ji}}{[M\mathbf{x}^{(k)}]_j} \right)$$

DM:

$$\mathscr{P} : \mathbf{y}^{(k+1)} = A\mathbf{x}^{(k)} + \mathbf{s}^{(k)}$$

$$\mathscr{B} : x_i^{(k+1)} = x_i^{(k)} \sum_{j=1}^{N_L} \left(\frac{y_j^\star a_{ji}}{y_j^{(k+1)}} \right)$$

with $[A]_{ji} \equiv a_{ji}$, $[M]_{ji} \equiv m_{ji}$, $[\mathbf{x}]_i \equiv x_i$, $[\mathbf{y}]_j \equiv y_j$, $\mathbf{y}^\star \equiv$ simulated sinogram, $N_L \equiv$ number of LORs, and $\mathbf{s}^{(k)} \equiv$ sinogram obtained by scatter only MC simulation using $\mathbf{x}^{(k)}$.

The forward scatter simulation was performed with a fraction $p = 0.001, 0.0001, 0.00001$ of the reconstructed emissions $\mathbf{x}^{(k)}$. A normalization of $\mathbf{s}^{(k)}$ with $1/p$ assured the correct scatter fraction in the projector.

2.4. Evaluation

In order to quantify the closeness to the true solution $\mathbf{x}^{\text{true}}$ the *normalized root mean squared error* (NRMSE) of the reconstructed images was calculated for each iteration step. The error introduced by the matrix was estimated by simulating matrices using different MC seeds but the same activity. The same approach was used to estimate the sinogram induced error. In both cases nine simulations were run. In this way, for each iteration step k and each voxel i the variance $\sigma_i^2(k)$ caused by the noise in the matrix (or sinogram) could be calculated:

$$\sigma_i^2(k) = \frac{1}{8} \sum_{\alpha=1}^{9} (x_{i,\alpha}^{(k)} - \bar{x}_i^{(k)})^2$$

where $\bar{x}_i^{(k)} \equiv$ mean value of voxel i at iteration k.

In the case of the DM reconstruction three sources of error exist: the sinogram, the matrix A, and the MC scatter projection (leading to the sinogram $\mathbf{s}^{(k)}$). The influence of each source can be measured again by varying the seed of the corresponding MC simulation and keeping the two other seeds constant.

Analogously to the NRMSE, a measure for the total induced error (*normalized root mean variance* (NRMV)) can be introduced:

$$\text{NRMV}(k) = \frac{1}{N_E} \sqrt{\frac{1}{N_V} \sum_{i=1}^{N_V} \sigma_i^2(k)}$$

where N_V is the number of voxels and N_E the average number of emissions per voxel.

3. Results

Fig. 2(a) (sinogram with 5×10^9 emission) and Fig. 2(b) (1×10^9 emissions) show the NRMSE curves for matrices of different statistics. In addition a measure for the introduced error (the NRMV) is shown. The sinogram

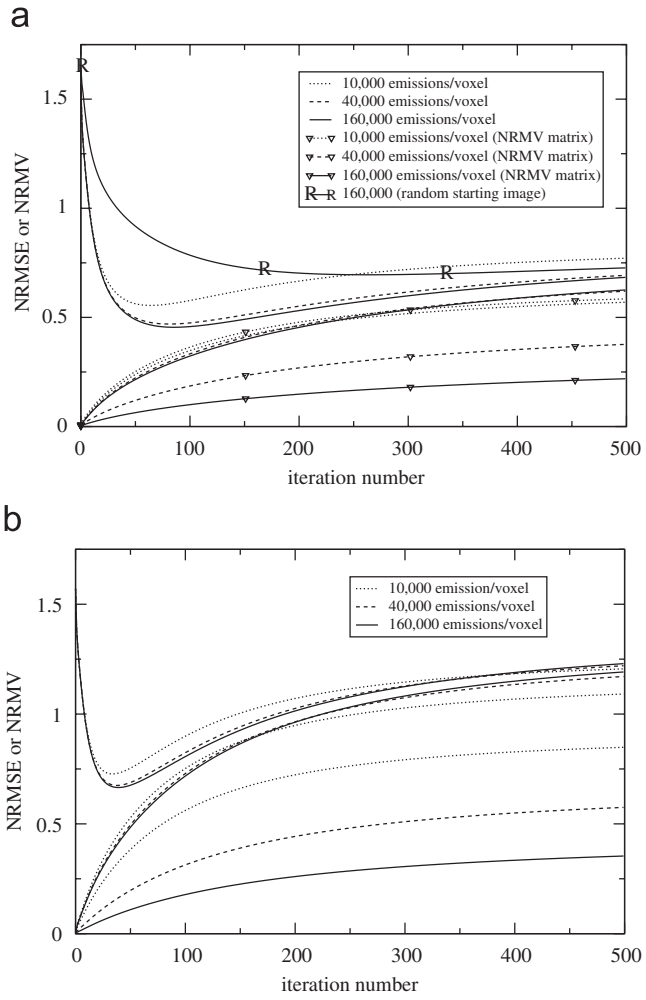


Fig. 2. Error vs. iteration number for full matrix reconstructions with different matrix statistics as indicated. (a) The R–R graph shows the NRMSE when using a starting image with random voxels $x_i \in [0,2[$ instead of a uniform image with voxels $x_i \equiv 1$. The sinogram was simulated with 5×10^9 emissions in total, (b) Sinogram simulated with 1×10^9 emissions in total. The three lowest curves show the NRMV of the matrix. The three upper curves represent the NRMSE.

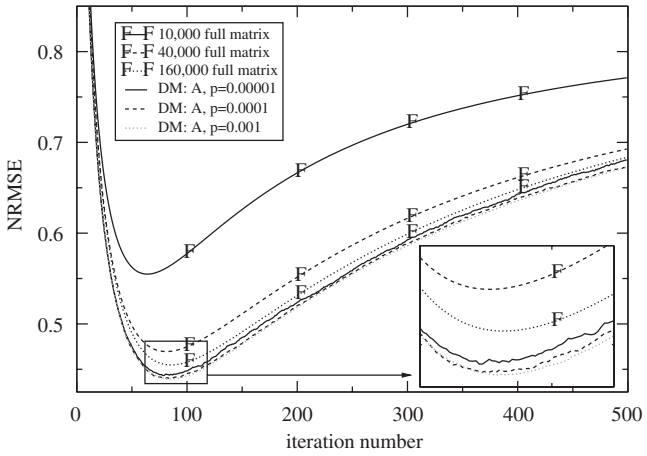


Fig. 3. NRMSE for full matrix reconstructions and DM reconstructions. Sinogram 5×10^9 emissions.

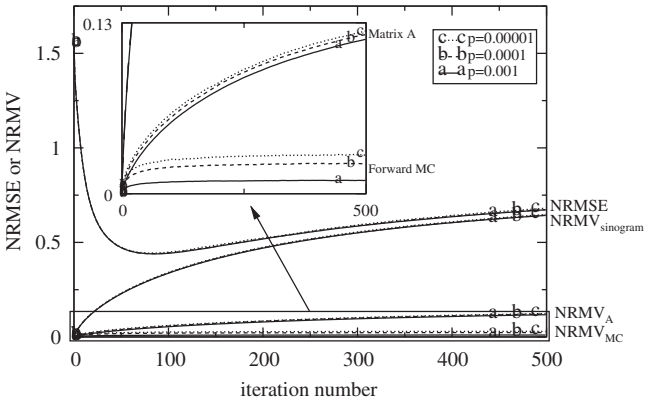


Fig. 4. NRMSE and NRMV of DM reconstruction with matrix A for the sinogram with 5×10^9 emissions and different fractions p of simulated forward scatter.

induced as well as the matrix induced error is monotonously increasing as the influence of the starting image is decreasing. The matrix induced error became comparable to the sinogram induced error in the case of the 10 000-emissions/voxel matrix (160 000 emissions in total). This matrix could be calculated in less than 4 min on a 16 processor cluster.

In Fig. 3 the NRMSE of the full matrix and DM approach can be seen. The difference in the 10 000, 40 000 and 160 000 curves of the full matrix approach are mostly due to different statistics of unscattered counts. Differences of the DM NRMSE is due to different scatter statistics.

In Fig. 4 the NRMSE and NRMV (sinogram, attenuation matrix A and forward MC scatter) is shown for reconstructions with different p. The forward MC scatter induced error is rather small due to the very small scatter fraction of the ideal one ring scanner (4.2%) and shows characteristic noise like fluctuations due to the MC projector. In scanners with higher scatter fractions an increase in the NRMV and an increase of the amplitude of the fluctuations can be expected.

In Fig. 5(a) and (b) the voxel dependency of the error ($\sigma_i = \sqrt{\sigma_i^2}$) of the sinogram induced and matrix M induced error can be seen. The relative error for voxel i is $\sigma_i/\bar{x}_i$. While the absolute error increases with higher activity, the relative error decreases. This behavior can be seen for all kinds of induced error: sinogram, matrix M as well as matrix A (Fig. 5(c)) and forward MC (Fig. 5(d)). The relative error of the forward MC induced error is inhomogeneously structured outside the phantom in contrast to the other induced errors. This effect is very small and cannot be seen in the absolute error images. A density dependent effect inside the cylinder could not be found.

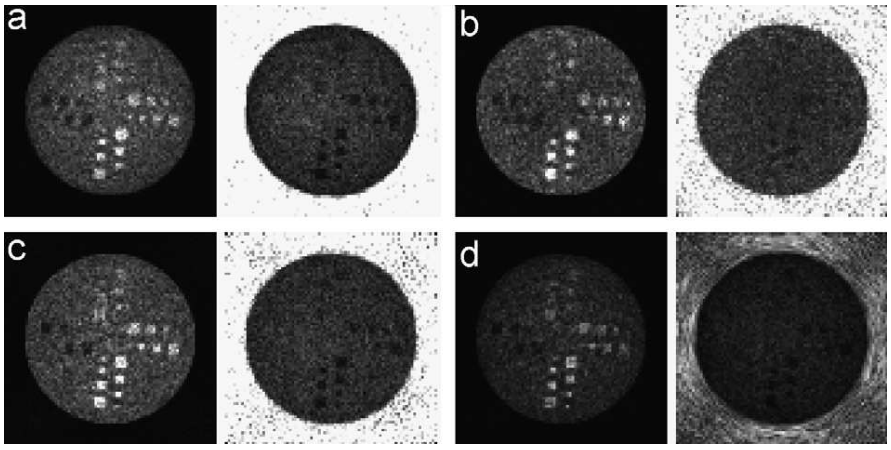


Fig. 5. (a) and (b) σ-images of full matrix reconstructions and (c) and (d) σ-images of DM reconstruction with $p = 0.00001$, both at minimal NRMSE with 5×10^9 simulated emission for the sinogram (white $\equiv$ big error, black $\equiv$ small error): (a) sinogram-σ: absolute (left), relative (right); 160 000-matrix; (b) matrix-M-σ: absolute (left), relative (right); 10 000-matrix; (c) matrix-A-σ: absolute (left), relative (right); 160 000-matrix; and (d) forward MC-σ: absolute (left), relative (right).

4. Conclusions

A method to quantify and investigate the propagation of error during a reconstruction that uses either MC simulated system matrices or MC simulated projectors was introduced. The error propagation of an one-ring scanner was investigated by simulations. Simulations of this ideal system show that the DM approach is converging faster initially in the considered case of a uniform starting image.

References

[1] D. Lazaro, et al., Phys. Med. Biol. 50 (2005) 3739.
[2] M. Rafecas, et al., IEEE Trans. Nucl. Sci. NS-51 (1) (2004) 149.
[3] S. Shoukouhi, Image reconstruction and image performance simulation of RatCAP (rat conscious animal PET), Ph.D. Dissertation, Stony Brook University, 2005.
[4] F.J. Beekman, et al., IEEE Trans. Med. Imag. 21 (8) (2002) 867.

C.2. Reconstruction of PET images with a compressed Monte Carlo based system matrix – a comparison to other Monte Carlo based algorithms

published in *Nuclear Science Symposium Conference Record, 2005 IEEE* Volume 4, 23-29 Oct. 2005 Page(s):2286 - 2290, doi: 10.1109/NSSMIC.2005.1596791.

Reconstruction of PET Images with a Compressed Monte Carlo Based System Matrix – a Comparison to Other Monte Carlo Based Algorithms

Niklas Rehfeld, *Student Member, IEEE*, Matthias Fippel, and Markus Alber

***Abstract*— A new method to compress the system matrix of a PET scanner calculated by Monte Carlo (MC) simulations is introduced. The proposed method reduces the size of the matrix drastically and allows a considerable reduction in the number of simulated particles. The images reconstructed with such a compressed matrix are compared to images reconstructed with other MC based algorithms, namely the dual-matrix algorithm [1] and the full-matrix algorithm (reconstruction using the uncompressed MC matrix).**

***Index Terms*— positron emission tomography, reconstruction, Monte Carlo, system matrix, compression.**

I. Introduction

Monte Carlo (MC) simulations are widely used in the field of positron emission tomography, mainly for the purpose of scanner development and reconstruction algorithm testing. The simulations also can be used to accurately simulate the system matrix (including the patient), however (a) the very long simulation time and (b) especially the immense storage demands make this direct approach not practicable.

The common way to avoid the latter shortcoming and to reduce the simulation time (but still to use MC simulations in reconstruction) is to accurately simulate the (re-)projector but to approximate the back-projector ([2], [1]). The approximated back-projector usually only comprises attenuation information, but lacks any scatter information.

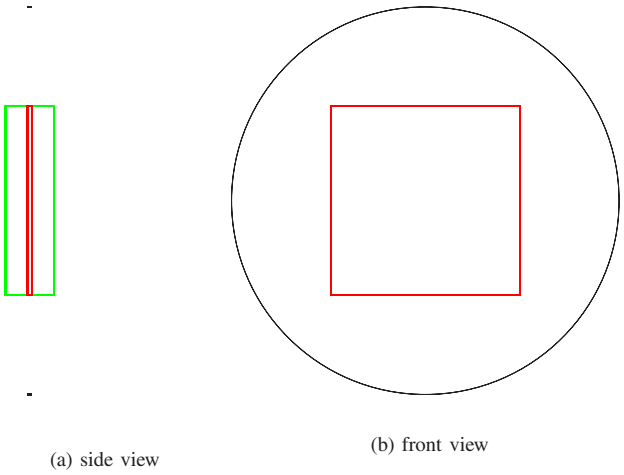
A new approach solving the problem of drawback (b) is the compression of the MC-matrix. A method to compress the matrix of a 2D PET scanner including a voxelized density phantom is introduced in this work. A generalization for the 3D case with oblique sinograms should be possible. The achieved compression factor is of the order of 10^3. In addition, a drastic reduction of the number of simulated particles can be obtained while achieving similar reconstructed images.

A matrix which is compressed by this method is used to reconstruct images and the images are compared to images reconstructed with the full MC-matrix (i.e. a MC matrix used for the projector and back-projector) and to images reconstructed with a dual-matrix approach as proposed by Beekman *et al*[1].

II. Methods

A. Geometry and Monte Carlo simulations

The modeled scanner consisted of one ring of depth 0.645 cm, a diameter of 82.4 cm, and 384 detectors. The density was described by 80x80x1 voxels with dimension 5x5x100 mm^3 and placed in the center of the scanner. A phantom, a heterogeneous density cylinder of diameter 32 cm, was approximated using this voxel grid. The activity was modeled by 80x80x1 voxels of dimension 5x5x6.45 mm^3.

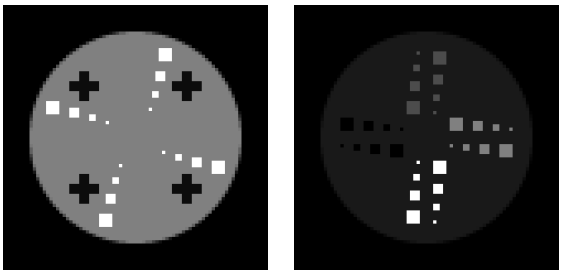


(a) side view (b) front view

Fig. 1. Side and front view: black=scanner, green=density, red=activity

The simulations were performed using a self written MC-code YaPRA, which is tracing photons in voxelized density phantoms, but implements only simple detector modeling by accepting all photons in an energy window (here 350-650 keV). The sinogram was simulated using 5×10^9 photon pairs without applying variance reduction techniques. In this way realistic noise properties could be obtained. The number of particles roughly corresponded to a 30 min scan with 6 kBq/ml initially. In the case of the system matrix this approach was too time consuming. Therefore the matrix was simulated with stratified sampling and forced detection, similar to the SimSET techniques [3]. Density to attenuation coefficients conversion according to M. Fippel[4] was used in order to obtain attenuation coefficients from density information. For the simulations a cluster of eight two-processor-computers with AMD XP2800+ were used. The MC simulation of a matrix with 4×10^4 photon pairs per voxel (2.56×10^8 photon pairs in total) needed approximately 12 min.

0-7803-9221-3/05/$20.00 ©2005 IEEE



(a) Density: 0, 0.1, 1, 2 g/cm^3

(b) Activity: ratio 0:1:3:5:10

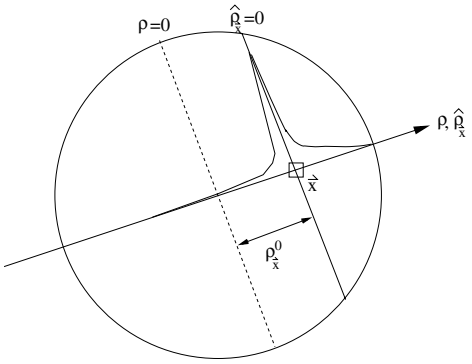


Fig. 2. Radon transform of the center of a voxel yields the geometrical expected maximum of a projection of the scatter

B. Compression

Since the attenuation-matrix $\mathbf{A}$ (no scatter, only attenuation) alone can be approximated well and fast by numerous algorithms and a substantial compression would most likely lead to image artifacts, the MC matrix $\mathbf{M}$ is divided into two matrices $\mathbf{M} = \mathbf{A} + \mathbf{S}$ ($\mathbf{S}$ = scatter only matrix) assuming that $\mathbf{A}$ is calculated on-the-fly or stored in a sparse manner.

The compression of matrix $\mathbf{S}$ is twofold. Firstly, the scatter tails of the projections of the sinograms of the matrix were fitted by functions $g(\rho) = \exp(a + b(\rho - \rho^0)) + \exp(c + d(\rho - \rho^0)^2)$ (see Fig. 4 and 5), shifted by the geometrically expected maximum ρ^0 of the scatter (see Fig. 2). Each scatter sinogram of matrix $\mathbf{S}$ is therefore described by 8×384 parameters, achieving a compression factor of 96/8=12 (96 being the number of bins per projection). MC simulations as well as measurements [5] showed that exponential scatter tails are a reasonable approximation.

A further compression is achieved by B_1-spline interpolation of the parameters in the voxel domain. These interpolating functions have the nice property of being continuous and that the corresponding kernels are small and of finite support, which is rather crucial for a reasonable calculation time in reconstruction.

In principle there are two different ways to perform this additional compression. One approach is to obtain the parameters for each voxel by fitting the projections and then to approximate these parameters by B-splines. The disadvantage of this approach is the large number of detected events needed, because only the events of one voxel contribute to a projection. Too few events result in unstable fits.

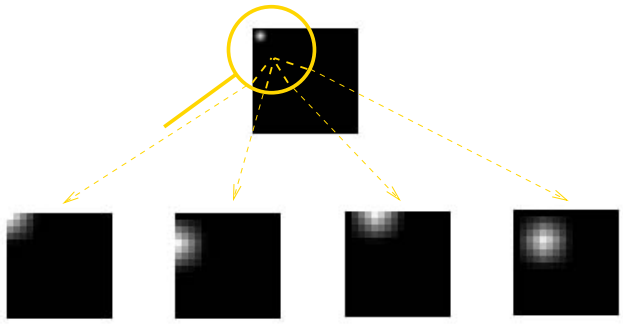


Fig. 3. Discrete B_1-spline kernels in the voxel domain

The other approach is to use the sinograms of the voxels belonging to a B-spline kernel (see Fig. 3) and to collect all detected events (with the appropriate weights). The advantage is the increased number of events used to determine the parameters: if n voxels belong to a B-spline-kernel, on average n times more events contribute to the parameter-fits. Unfortunately, the sinograms of the kernel cannot simply be added before fitting because this would result in a flattened and smeared scatter peak. When the expected geometrical scatter maximum of the projections for all voxels in the kernel are aligned, however, only the shape of the tails is averaged but not the position of the peak.

The usage of $20 \times 20 \times 1$ (or $5 \times 5 \times 1$) B-splines-kernels (instead of $80 \times 80 \times 1$ voxels) resulted therefore in a total reduction of the matrix size by a factor of $12 \times 4 \times 4 = 192$ (or $12 \times 16 \times 16 = 3072$). Only voxels inside the cylinder (i.e. with non-zero density) were considered resulting in an additional reduction of the size of the matrix by almost a factor of 2.

The Levenberg-Marquardt algorithm was used for fitting. Since especially for large kernels many points have to be fitted, the data was re-organized before fitting, grouping several points at the far ends of the scatter tails into one point, but grouping less and less points the more ρ approaches the expected peak of the scatter. This assured a correct fit close to the scatter peak and proved to result in rather stable fits and fast fitting.

C. Reconstruction and evaluation methods

Maximum likelihood expectation maximization (ML-EM) without regularization was used for reconstruction. The normalized mean squared error (NMSE) of the reconstructed activity (compared to the true activity) was plotted vs. iteration number. The minimal value of this plot as well as well as the slope for higher iteration numbers was used as a measure for the correctness of the matrix. Matrix $\mathbf{A}$ was calculated by a high statistics MC simulation with 1.024×10^7 photon pairs/voxel. Dual matrix (DM) reconstruction was performed with $\mathbf{A}$ as the

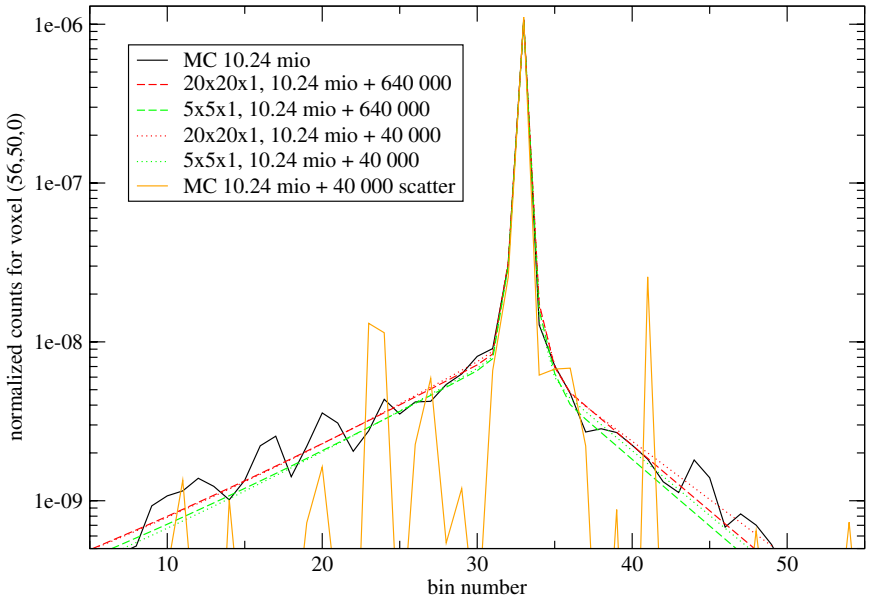


Fig. 4. A projection of the matrix (voxel (56,50,0)) and compressed versions of the projections.

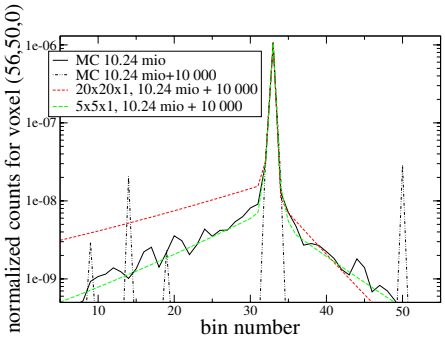
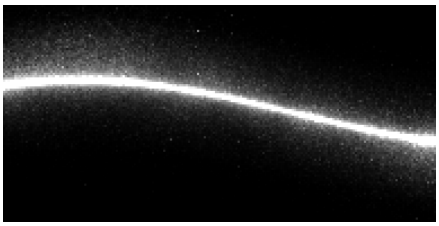
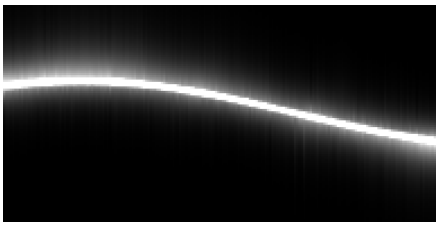


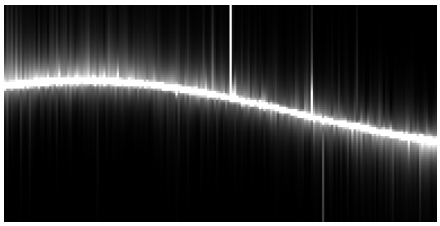
Fig. 5. A projection of the matrix (voxel (56,50,0)). Only 10 000 photon pairs are started in this voxel to estimate the scatter. It can be seen that the larger kernel in the case of the $5\times5\times1$-compression assures a more stable fit.



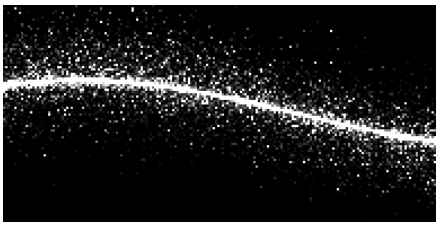
(a) 1.024×10^7 photon pairs/voxel



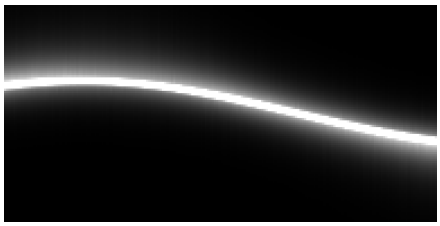
(b) $20 \times 20 \times 1$, direct: 1.024×10^7 + scatter: 6.4×10^5 photon pairs/voxel



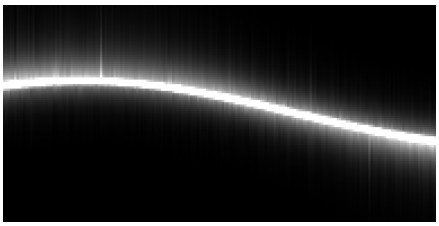
(c) $20 \times 20 \times 1$, direct: 1.024×10^7 + scatter: 4×10^4 photon pairs/voxel



(d) direct: 1.024×10^7 + scatter: 4×10^4 photon pairs/voxel



(e) $5 \times 5 \times 1$, direct: 1.024×10^7 + scatter: 6.4×10^5 photon pairs/voxel



(f) $5 \times 5 \times 1$, direct: 1.024×10^7 + scatter: 4×10^4 photon pairs/voxel

Fig. 6. Sinograms of the voxel (56,50,0). Gray-values are windowed to $[0,10^{-8}]$ normalized counts in order to show scatter. It can be seen that the $5 \times 5 \times 1$-compression resulted in more stable fits in the case of low statistics scatter (compare Fig. 6(c) to Fig. 6(f)), but also in a less correct description of the scatter tails in the case of higher scatter statistics (compare Fig. 6(b) to Fig. 6(e)).

back-projector and $P(x_i) = \mathbf{A}x_i + s_i$ as the projector, where s_i is the (pure) scatter sinogram of the activity x_i at iteration i.

III. Results

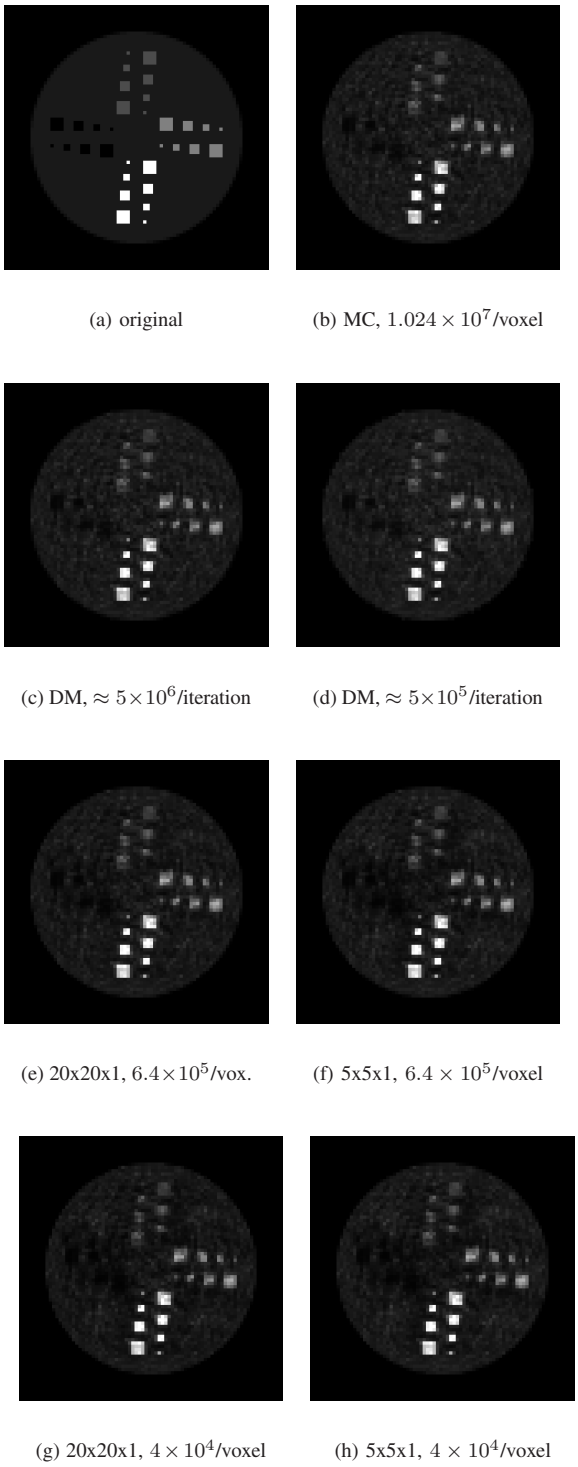
Images were reconstructed in seven different ways: Using a MC matrix including scatter (1.024×10^7 photon pairs/voxel simulated), using DM with roughly 5×10^6 or 5×10^5 photon pairs/iteration for the scatter sinogram s, and using compressed matrices (matrix $\mathbf{A}$+ matrix $\mathbf{S}$ with 6.4×10^5 or 4×10^4 photon pairs/voxel with either 20x20x1 or 5x5x1 spline compression).

It can be seen that the visual results did not differ too much (Fig. 7). This was due to the reason that the simulated scanner was an ideal 2D scanner. The reconstruction methods only differed in the way of treating scatter, and since the scatter fraction was small, the difference was not very big.

The NMSE plots (Fig. 8) were more meaningful. A positive slope is a sign for an improper sinogram (noisy) or a not correct matrix (noisy or wrong). It can be seen in Fig. 8 that the full-MC-matrix reconstruction as well as the compressed matrix reconstruction result in similar slopes at higher iteration numbers. The fact that a compressed matrix ($20 \times 20 \times 1$ with 6.4×10^5 photon pairs simulated per voxel for scatter) resulted in better images than the full matrix showed that the the description of the scatter tail as the sum of a Gaussian and an exponential function is good enough. Larger spline kernels proved to be superior when used for low statistics simulations (compare 20x20x1 compression and 5x5x1 compression when simulating 4×10^4 photon pairs in Fig. 8). This is in agreement with the fact that they provide more data points and therefore should result in more stable fits. For higher statistics simulation, however, smaller kernels were superior, because the higher number of splines allowed a more subtle description of the matrix.

The DM approach is initially faster converging. This, however, cannot be accounted for the fact that the "matrix" is in better agreement with the reality. Strictly speaking this algorithm is not an EM - algorithm, because two different matrices are used.

The fact that the MC-matrix is not performing best may rely on the fact that this matrix was calculated simulating 1.024×10^7 photon pairs per voxel including scatter, whereas all others use this number of pairs for direct counts plus additional simulated particles for scatter. More likely is however simply a better approximation of the "reality" by the fitted smooth scatter tails of the 20x20x1 compressed matrix than by the discrete high statistics MC simulation (see Fig. 4).



(a) original (b) MC, 1.024×10^7/voxel

(c) DM, $\approx 5 \times 10^6$/iteration (d) DM, $\approx 5 \times 10^5$/iteration

(e) 20x20x1, 6.4×10^5/vox. (f) 5x5x1, 6.4×10^5/voxel

(g) 20x20x1, 4×10^4/voxel (h) 5x5x1, 4×10^4/voxel

Fig. 7. Reconstructed images. The images with minimal NMSE were chosen.

IV. Conclusion

A new way to incorporate MC scatter in the reconstruction process was introduced. The method allowed a significant reduction of the matrix size and simulation time. Images obtained were comparable with images obtained using the dual matrix algorithm or the full MC matrix approach. A larger spline kernel could be used to counterbalance a decrease in the statistics, providing a handle to reduce the simulation time without degrading the images too much. When using the same spline compression, the quality of the images could

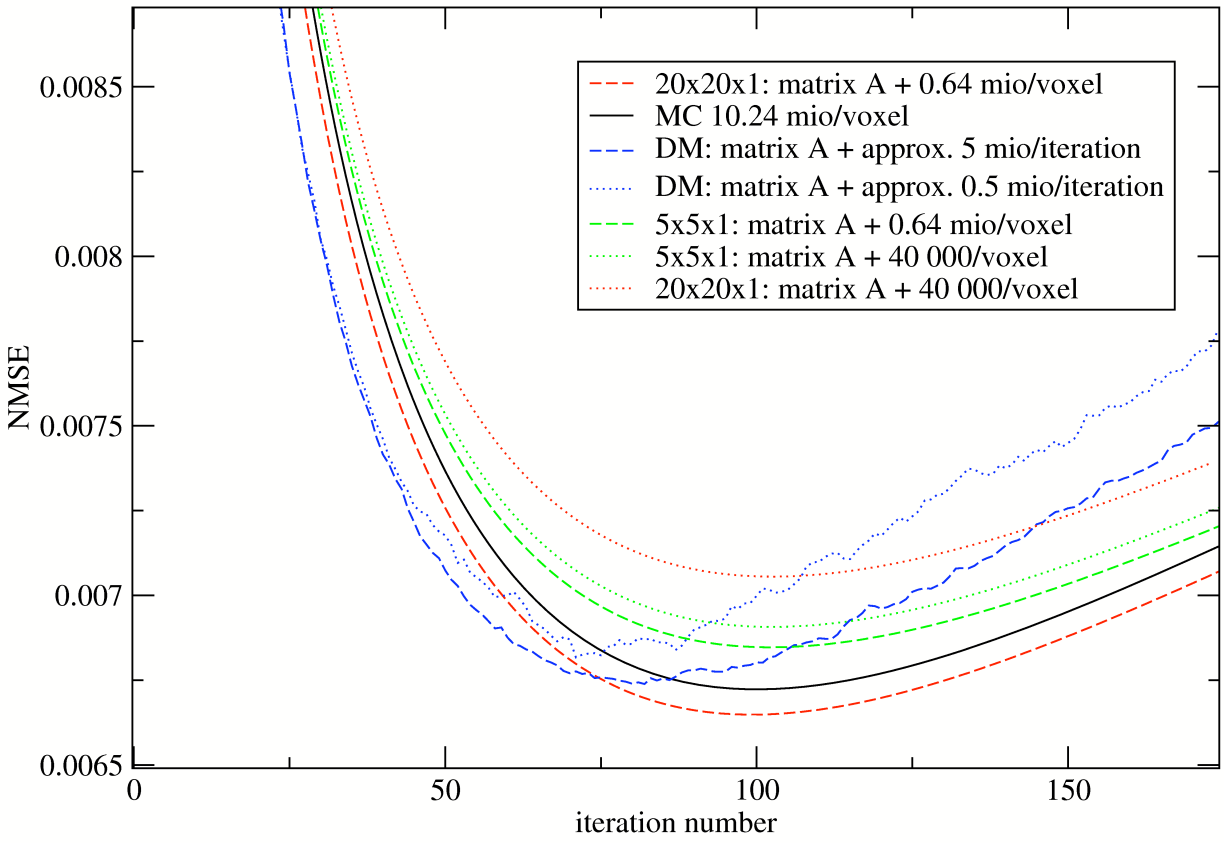


Fig. 8. NMSE vs. iteration number

be increased directly by increasing the number of simulated particles. Together with the possibility to model other physical processes (like detector physics), the proposed reconstruction method – which is not restricted to the EM algorithm – provides a straight forward and direct way to improve the correctness of the reconstructed images.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft DFG.

References

[1] F. J. Beekman *et al*, "Efficient fully 3-D iterative SPECT reconstruction with Monte Carlo-based scatter compensation," *IEEE Trans. Med. Imag.*, vol. 21(8), p. 867, 2002.

[2] A. Werling *et al*, 'Fast implementation of the single scatter simulation algorithm and its use in iterative image reconstruction of PET data," *Phys. Med. Biol.*, vol. 47, p. 2947, 2002.

[3] D. R. Haynor *et al*, "The use of importance sampling techniques to improve the efficiency of photon tracking in emission tomography simulations," *Med. Phys.*, vol. 18(5), p. 990, 1991.

[4] M. Fippel, "Fast Monte Carlo dose calculation for photon beams based on the VMC electron algorithm," *Med. Phys.*, vol. 26, p. 1466, 1999.

[5] M. Bergström *et al*, "Correction for scattered radiation in a ring detector positron camera by integral transformation of the projections," *J. Comput. Assist. Tomogr.*, vol. 7(1), p. 42, 1983.

C.3. Compression of a Monte Carlo based system matrix for iterative reconstruction of PET images

published in *Nuclear Science Symposium Conference Record, 2004 IEEE* Volume 6, 16-22 Oct. 2004 Page(s):3945 - 3947 Vol. 6, doi: 10.1109/NSSMIC.2004.1466741.

Compression of a Monte-Carlo Based System Matrix for Iterative Reconstruction of PET Images

Niklas Rehfeld, *Student Member, IEEE*, Markus Alber, Matthias Fippel, and Fridtjof Nüsslin

***Abstract*— Modern human PET-scanners often have high scatter fractions due to the lack of septa. This work treats the scatter in a straight forward manner by calculating the system matrix elements directly with Monte-Carlo (MC) methods. A parametric compression method was used to scale down memory consumption. The resulting images (reconstructed with either the compressed or with the uncompressed MC-matrix) were compared.**

***Index Terms*— positron emission tomography, Monte Carlo, system matrix, compression.**

INTRODUCTION

Modern PET scanners require a correct treatment of scattered events due to their high scatter fraction. Monte Carlo (MC) simulations are the methods of choice for correct scatter calculation. Unfortunately, they are very time consuming.

One approach to overcome this shortcoming is to use accelerated MC calculations in combination with fast reconstruction algorithms. Beekman *et al* [1] showed that this is possible for single photon emission computed tomography (SPECT). Single scatter simulation [2], [3], a comparable technique for PET, is using random scatter points to estimate the scatter.

The other way to handle this problem is to calculate the matrix once at the beginning and store it, retrieving matrix elements from memory whenever necessary. Recently, Rafecas *et al* [4] used a MC-based system matrix for reconstruction in a small animal PET scanner, storing the matrix with the help of a data base management system. The advantage of a stored matrix over the first approach relies in the much faster retrieval of matrix elements which facilitates not only the usage of scatter in the forward projection but also in other parts of the reconstruction algorithm.

For human 3D-PET-scanners direct storage of the system matrix is however prohibitive, because of the larger number of detectors and scattering in the patient which can be neglected in small animal PET systems. A compression of the matrix might resolve this storage problem. Not only the resulting matrix is smaller, but also the compression can lessen effects which occur due to low MC statistics.

This work presents first investigations in the two dimensional case. In two dimensions the matrix can be stored without problems and images reconstructed by either the uncompressed matrix or the compressed matrix can be compared.

METHODS

The simulations were performed by a dedicated ring-PET MC-code which is in parts derived from the MC-code XVMC used in radiotherapy [5]. Detection forcing, implemented in a similar manner originally in SimSET [6], has been used. Stratification which should result in additional speed-up of the simulations has not yet been implemented. Absorption due to the photo effect has been neglected but could be easily implemented if needed. The simulations were performed in voxelized density phantoms using linear attenuation coefficients calculated from the density information as described in [5].

The system matrix was filled with data obtained by successive MC sub-simulations. Each sub-simulation started a defined number of photon pairs randomly positioned in a voxel. The obtained sinograms for each voxel j formed columns of the system matrix P_{ij} where i represented a line of response (LOR).

The rows (constant ϕ, variable ρ) in a sinogram of a single voxel j were modeled by functions $f_{\phi j}(\rho)$. These functions were composed out of three parts: a central discrete part and two lateral mono-exponentials (see Fig. 1).

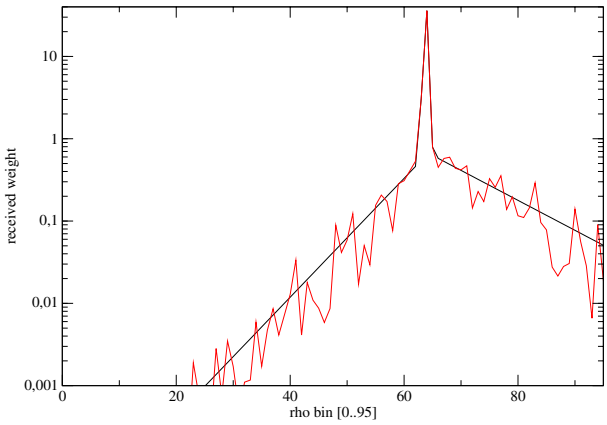


Fig. 1. Compression of a row of an off-central voxel.

The left and right slopes were fitted in the log-plot by a least squares method to straight lines, assuming Poisson statistics . A similar method has been proposed by Bergström for experimental data [7].

The simulated scanner had a diameter of 96 cm and consisted of one ring with 4 cm depth and 384 detectors. The larger than usual depth was used to obtain better statistics. The detectors

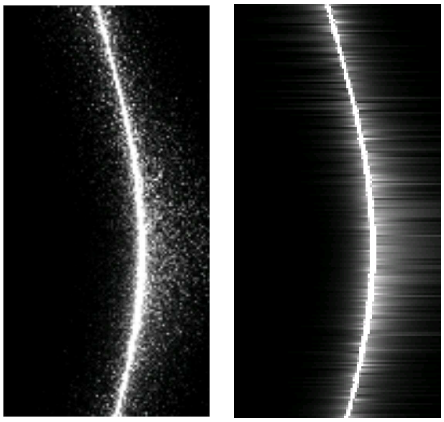


Fig. 2. Uncompressed (left) and compressed (right) sinogram of an off-central voxel (windowed gray-values to show scatter).

of the scanner were considered to be ideal: 100% efficiency, curved along the nappe of the cylinder, with no depth and without inter-detector or inter-crystal spacing. An energy cut-off for energies below 300 keV was assumed for the detectors.

As a phantom a water filled cylinder ($\varnothing = 30\,\text{cm}$, $4\,\text{cm}$ long) with an L-shaped hole was used. The cylinder was filled with uniform activity, whereas two areas were left without activity and other two areas were filled with $4\times$ and $2\times$ the background activity (see Fig. 3 right).

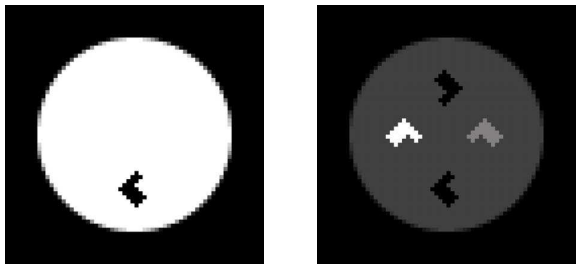


Fig. 3. Left: density, right: activity (ratio 0:1:2:4)

The whole volume to be reconstructed was $40 \times 40 \times 4\,\text{cm}^3$ ($64 \times 64 \times 1$ voxel). The images were reconstructed using maximum likelihood expectation maximization (ML-EM) without any regularization. The reconstruction started with uniform activity in all voxels.

Results

Using 5×10^7 started photon pairs for the simulation of the sinogram of the phantom and 5×10^6 started photon pairs for each voxel of the system matrix (150,208,512 elements), the reconstructed images which can be seen in Fig. 4 were obtained.

The image reconstructed with the compressed matrix was more grainy. The reason for this could be either some badly

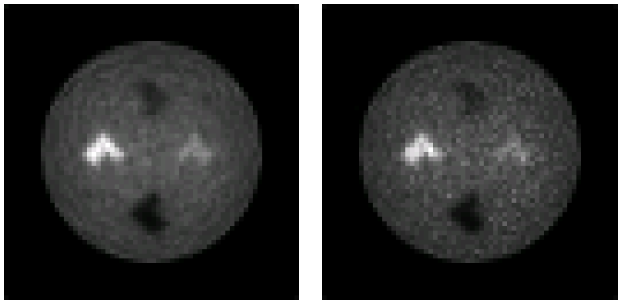


Fig. 4. Reconstructed with uncompressed matrix (left) and compressed matrix (right), both after 20 iterations of ML-EM.

fitted rows or a faster convergence of the algorithm when using the compressed matrix. Since different matrices were used, a direct comparison of the log likelihood was unfortunately not the best way of evaluating the convergence properties. Looking at a compressed sinogram of the matrix (Fig. 2 right) revealed that a fit/compression in ϕ direction would most likely lessen the number of badly fitted rows considerably.

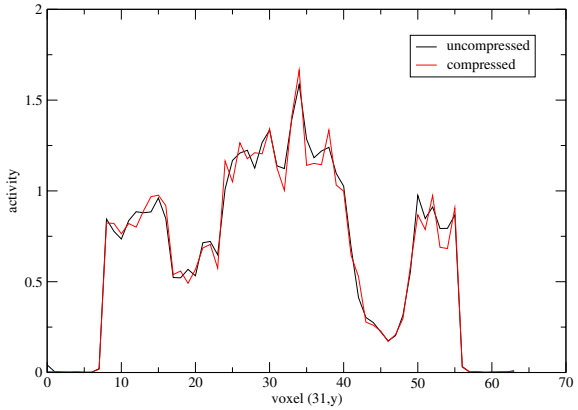


Fig. 5. Vertical profile at $x = 31$.

A vertical (Fig. 5) and horizontal (Fig. 6) profile through the middle of the reconstructed activity distributions were taken. The less smooth profiles through the activity reconstructed with the compressed matrix (Fig. 4 right) confirm the grainy impression of this image. For both matrices, the higher activity concentrations ($4\times$ and $2\times$ background activity) matched very well, whereas the zero activity regions reached only approximately 25% of the background activity in air (lower L-shape) and even only 50% in water (upper L-shape). Outside the cylinder there was virtually no activity reconstructed.

The proposed fitting-compression scheme did reduce the matrix size about a factor of 26. The system matrix was calculated using the the PVM (**p**arallel **v**irtual **m**achine)-library on 14 2.8-GHz-processors. The calculation time was about 22 h.

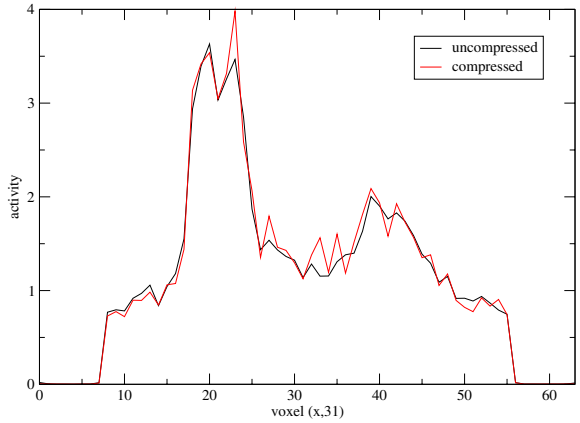


Fig. 6. Horizontal profile at $y = 31$.

CONCLUSION

Images reconstructed with MC-based system matrices and ML-EM are rather smooth and activity artefacts outside the phantom are not present at all. Compression of the matrix did not improve the image quality, but resulted in more grainy images. Since regularization was not used, it is difficult to judge if the graininess can be accounted for a faster convergence and therefore less dependence on the starting image or for the fact that the fitting procedure does not work well for some data. Older simulations with worse statistics, showed that compression improved image quality, especially smoothness.

Altogether it can be concluded that the system matrix can be compressed by parameterization and using this compressed matrix still yields reasonable reconstructed images. However, further refinement of the compression scheme is needed. Further research to improve the compression scheme as well as the MC- code is therefore worthwhile and the effect of regularization should be examined.

ACKNOWLEDGMENTS

The authors would like to thank S. Ziegler and M. Rafecas for their very helpful suggestions. This work was supported by the Deutsche Krebshilfe e.V.

REFERENCES

[1] F. J. Beekman, H. W. A. M. de Jong, and S. van Geloven, "Efficient fully 3-D iterative SPECT reconstruction with Monte Carlo-based scatter compensation," *IEEE Trans. Med. Imag.*, vol. 21(8), p. 867, 2002.

[2] A. Werling, O. Bublitz, J. Doll, L.-E. Adam, and G. Brix, "Fast implementation of the single scatter simulation algorithm and its use in iterative image reconstruction of PET data," *Phys. Med. Biol.*, vol. 47, p. 2947, 2002.

[3] J. M. Ollinger, "Model-based scatter correction for fully 3D PET," *Phys. Med. Biol.*, vol. 41(1), pp. 153–176, 1996.

[4] M. Rafecas, G. Böning, B. J. Pichler, E. Lorenz, M. Schwaiger, and S. I. Ziegler, "Effect of noise in the probability matrix used for statistical reconstruction of PET data," *IEEE Trans. Nucl. Sci.*, vol. 51(1), pp. 149–156, 2004.

[5] M. Fippel, "Fast Monte Carlo dose calculation for photon beams based on the VMC electron algorithm," *Med. Phys.*, vol. 26, p. 1466, 1999.

[6] D. R. Haynor, R. L. Harrison, and T. K. Lewellen, "The use of importance sampling techniques to improve the efficiency of photon tracking in emission tomography simulations," *Med. Phys.*, vol. 18(5), p. 990, 1991.

[7] M. Bergström, L. Eriksson, C. Bohm, G. Blomqvist, and J. Litton, "Correction for scattered radiation in a ring detector positron camera by integral transformation of the projections," *J. Comput. Assist. Tomogr.*, vol. 7(1), p. 42, 1983.