Image Reconstruction Methods in Positron Tomography

David W. Townsend and Michel Defrise
Division de Médecine Nucléaire
Hôpital Cantonal Universitaire de Genève
Geneva, Switzerland

Lectures given in the Academic Training Programme of CERN 1992–1993
Physics and mathematics for medical imaging

GENEVA
1993
Propriété littéraire et scientifique réservée pour tous les pays du monde. Ce document ne peut être reproduit ou traduit en tout ou en parti sans l'autorisation écrite du Directeur général du CERN, titulaire du droit d'auteur. Dans les cas appropriés, et s'il s'agit d'utiliser le document à des fins non commerciales, cette autorisation sera volontiers accordée.

Le CERN ne revendique pas la propriété des inventions brevetables et dessins ou modèles susceptibles de dépôt qui pourraient être décrits dans le présent document ; ceux-ci peuvent être librement utilisés par les instituts de recherche, les industriels et autres intéressés. Cependant, le CERN se réserve le droit de s'opposer à toute revendication qu'un usager pourrait faire de la propriété scientifique ou industrielle de toute invention et tout dessin ou modèle décrits dans le présent document.

Literary and scientific copyrights reserved in all countries of the world. This report, or any part of it, may not be reprinted or translated without written permission of the copyright holder, the Director-General of CERN. However, permission will be freely granted for appropriate non-commercial use.

If any patentable invention or registrable design is described in the report, CERN makes no claim to property rights in it but offers it for the free use of research institutions, manufacturers and others. CERN, however, may oppose any attempt by a user to claim any proprietary or patent rights in such inventions or designs as may be described in the present document.

ISSN 0007-8328
ISBN 92-9083-051-4
Image Reconstruction Methods
in Positron Tomography

David W. Townsend and Michel Defrise
Division de Médecine Nucléaire
Hôpital Cantonal Universitaire de Genève
Geneva, Switzerland

Lectures given in the
Academic Training Programme of CERN
1992–1993
Physics and mathematics for medical imaging

GENEVA
1993
Abstract

In the two decades since the introduction of the X-ray scanner into radiology, medical imaging techniques have become widely established as essential tools in the diagnosis of disease. As a consequence of recent technological and mathematical advances, the non-invasive, three-dimensional imaging of internal organs such as the brain and the heart is now possible, not only for anatomical investigations using X-rays but also for studies which explore the functional status of the body using positron-emitting radioisotopes. This report reviews the historical and physical basis of medical imaging techniques using positron-emitting radioisotopes. Mathematical methods which enable three-dimensional distributions of radioisotopes to be reconstructed from projection data (sinograms) acquired by detectors suitably positioned around the patient are discussed. The extension of conventional two-dimensional tomographic reconstruction algorithms to fully three-dimensional reconstruction is described in detail.
## CONTENTS

1. INTRODUCTION.................................................................1

2. DATA ACQUISITION...........................................................7
   2.1 Lines of response.......................................................7
   2.2 Sinograms for a single detector ring..............................9
   2.3 Sampling the sinogram..............................................12
   2.4 Sinograms for multiple rings of detectors......................14
   2.5 Sinograms for full 3-D acquisition...............................15
   2.6 List mode data acquisition.......................................16

3. IMAGE RECONSTRUCTION..................................................16
   3.1 Filtered backprojection in 2-D: the continuous case...........16
   3.2 Filtered backprojection in 2-D: the discrete implementation...21
   3.3 Filtered backprojection in 3-D: the continuous case...........23
   3.4 Reconstruction in 3-D: the discrete implementation............29
   3.5 Iterative reconstruction methods................................32
1. INTRODUCTION

Positron emission tomography* (PET) is a diagnostic technique which allows the in vivo measurement of the distribution of a tracer labelled with a positron emitting nuclide. The importance of this technique in medicine originates from the existence of neutron-deficient isotopes of carbon (\(^{11}\)C), oxygen (\(^{15}\)O) and nitrogen (\(^{13}\)N), three of the major elements occurring in biological tissues and physiological processes. Organic molecules such as water, carbon dioxide and ammonia can therefore be labelled with a radioactive marker without using artificial, inorganic elements which could modify metabolism. This approach is in contrast to conventional nuclear medicine (single photon imaging), where the use of more "exotic" nuclides such as \(^{99m}\)Tc or \(^{201}\)Tl complicates a physiological interpretation of the data. The short half life of \(^{11}\)C (20.4 min), \(^{15}\)O (2.07 min) and \(^{13}\)N (9.96 min) is a further advantage from a dosimetry point of view. However, since these short-lived isotopes are cyclotron-produced, the cyclotron must be located in close proximity to the hospital. There are, in addition, a number of longer-lived isotopes (such as the widely used \(^{18}\)F, half life 109.8 min) which can be obtained by delivery from a remote cyclotron, and in-house generator-produced isotopes\(^1\) such as \(^{68}\)Ga (68 min) and \(^{82}\)Rb (75 s), which are obtained from longer-lived parent nuclei. Together with the complexity and cost of the imaging equipment (scanner), the requirement for an in-house cyclotron explains the limited number of PET centres found worldwide (around 100) compared with other medical imaging installations such as X-ray tomography and magnetic resonance.

A key property of positron tomography is the possibility to record positron annihilations by coincidence detection. By detecting the two 511 keV photons emitted in opposite directions from the annihilation of a positron with an atomic electron, it is possible to localize the decay of a positron-emitting nuclide along the line joining the two detectors. This coincidence technique, referred to as "electronic collimation", eliminates the need for any physical collimation. The accuracy of electronic collimation is limited by the acollinearity of the two annihilation photons\(^2\) (typically a 0.5° FWHM deviation from 180°), and by the distance between the points of emission and annihilation of the positron\(^3,4\) (positron range, see table 1.1). In practice,

<table>
<thead>
<tr>
<th>Maximum energy</th>
<th>(^{18})F</th>
<th>(^{11})C</th>
<th>(^{13})N</th>
<th>(^{15})O</th>
<th>(^{68})Ga</th>
</tr>
</thead>
<tbody>
<tr>
<td>MeV</td>
<td>0.633</td>
<td>0.959</td>
<td>1.197</td>
<td>1.738</td>
<td>1.898</td>
</tr>
<tr>
<td>Most probable energy</td>
<td>MeV</td>
<td>0.202</td>
<td>0.326</td>
<td>0.432</td>
<td>0.696</td>
</tr>
<tr>
<td>Half-life</td>
<td>min</td>
<td>109.8</td>
<td>20.4</td>
<td>9.96</td>
<td>2.07</td>
</tr>
<tr>
<td>max. path length in water</td>
<td>mm</td>
<td>2.4</td>
<td>5.0</td>
<td>5.4</td>
<td>8.2</td>
</tr>
<tr>
<td>radial range in water (FWHM)</td>
<td>mm</td>
<td>1.0</td>
<td>1.1</td>
<td>1.4</td>
<td>1.5</td>
</tr>
<tr>
<td>radial range in water (FWTM)</td>
<td>mm</td>
<td>1.8</td>
<td>2.2</td>
<td>2.8</td>
<td>3.6</td>
</tr>
</tbody>
</table>

(FWHM: Full Width at Half Maximum; FWTM: Full Width at Tenth Maximum)

\* tomography, from the greek "temnein", meaning to cut.
however, the spatial resolution of modern positron tomographs is still limited to typically 6 mm by photon statistics (i.e. by the limited number of recorded positron decays) and by the intrinsic detector resolution, rather than by the physical limitations due to photon-photon acollinearity and positron range. When justified by the instrument resolution and signal-to-noise, the resolution loss caused by these effects can be recovered to some extent using deconvolution techniques because the form of the blurring function can be measured independently. The effective resolution improvement, however, is restricted by the ill-posed nature of deconvolution.

Together with the existence of naturally-occurring positron-emitting nuclides, “electronic collimation” is undoubtedly the most important factor responsible for the success of positron tomography. This is better understood by considering the imaging of nuclides decaying by single photon emission. In this case, no useful information can be extracted from the position where a photon has been detected. It is necessary to collimate the detector in such a way that not only the position, but also the momentum i.e. direction, of the photon can be inferred. Collimation in single photon imaging is achieved using thick lead (or other high Z materials such as tungsten) shielding drilled with holes and placed in front of a large sodium iodide crystal (gamma camera). The size and number of the holes determine both the resolution and the sensitivity of the detector, and collimator design then amounts to an often frustrating compromise between resolution and sensitivity. By eliminating the need for collimators, coincidence detection in PET leads to much higher sensitivities than single photon imaging (factors of 20 or more). As will be seen, however, some degree of collimation may nevertheless be needed in PET to suppress random coincidences between photons originating from independent positron decays, and scattered coincidences involving one or both photons which have undergone Compton scattering (fig.1).

Figure 1: The three types of coincident events detected by a PET scanner

A further advantage of coincidence detection is the simplification of the attenuation correction. Only 14% of the annihilation pairs produced at the centre of a 20 cm diameter water sphere escape from the sphere without undergoing any interaction, and an accurate attenuation correction is essential to quantify the local tracer concentration. With coincidence detection, the correction factor applied to the positron decays recorded along a “line of response” joining two detectors, does not depend on the localization of each decay along that line. This correction factor can be determined by means of an independent transmission scan performed with an external source, prior to injecting the tracer in the patient. In single photon imaging, on the other hand, it is impossible to assign a constant attenuation correction factor to each line of response, since the attenuation depends on the unknown position of each decay along the line of response. This is the reason that attenuation correction is more complex in single photon imaging than in PET, and approximate correction methods must be used.
Finally, coincidence detection offers the possibility to exploit differential time-of-flight (TOF) measurements to localize each positron decay along the line of response joining the two detectors. The time-of-flight technique, which could in principle eliminate the need for image reconstruction, was proposed by Anger in 1974\textsuperscript{11}, and implemented successfully in a number of scanners [for recent references see Lewellen et al\textsuperscript{12} and Mazoyer et al\textsuperscript{13}]. The time-of-flight approach remains difficult to exploit however, since it requires the use of fast crystals such as CsF or BaF\textsubscript{2} which have lower stopping power than bismuth germanate (Bi\textsubscript{4}Ge\textsubscript{3}O\textsubscript{12}, or BGO), the scintillator most-widely used in conventional PET scanners. The resulting reduction in sensitivity is not compensated by the limited accuracy (at best 5 cm) of TOF localization. The time-of-flight technique, however, is well adapted to studies at very high count rates\textsuperscript{13,14}.

The imaging of molecules labelled with positron emitting radionuclides was first proposed by Wrenn et al\textsuperscript{15}, only a few years after the birth of nuclear medicine. The first rectilinear scanner was designed by Brownell and Sweet\textsuperscript{16} from a single pair of sodium iodide scintillation detectors with a grid (raster) scanning motion producing a rough planar map of the tracer distribution in the head. Improved scanners were rapidly developed, based either on area detectors (such as a single-crystal Anger gamma camera) or on multi-crystal detectors. Recent prototype scanners\textsuperscript{17,18} based on multi-crystal arrangements with wire-chamber read-out cannot easily be classified in either family. The classification of scanners as area detectors or as multi-crystal detectors is therefore somewhat arbitrary, but will nevertheless be used here to follow the development of positron tomography.

The first family of scanners was based on area detectors, either single crystal detectors (Anger gamma camera)\textsuperscript{19}, or, later, modified multi-wire proportional chambers\textsuperscript{20}. An early design of Anger\textsuperscript{21}, for example, consists of a non-collimated gamma camera in coincidence with a single gamma counter. As discussed by Webb\textsuperscript{22}, this arrangement is geometrically equivalent to simple x-ray radiography, and has no tomographic capability. The sensitivity was later improved by placing in coincidence two opposing stationary gamma cameras\textsuperscript{23}. A section through the tracer distribution was reconstructed by backprojecting the data in a plane parallel to the detectors (see section 3.1), the value at each point of this plane being reconstructed by summing the number of coincidences detected in all lines of response passing through the point. A limited degree of tomographic resolution was thus achieved, in the sense that a point source located outside the reconstructed plane was defocused (fig. 2). This, however, is still far from an "exact" tomographic reconstruction, which, as was realized later\textsuperscript{23,24}, requires the

![Diagram](https://via.placeholder.com/150)

**Figure 2:** A pair of area detectors imaging a point source located in plane A. Lines of response through S are shown. Backprojection yields a focused image of S in plane A and a defocused image in plane B.
rotation of the two area detectors during acquisition, and the development of appropriate reconstruction algorithms\cite{80}. It is important to note that positron scanners based on a pair of area detectors do not require a collimator, each point on one detector being in coincidence with any point on the other detector. These fully three-dimensional scanners therefore optimize the geometric sensitivity (i.e. the acceptance angle), taking full advantage of the electronic collimation capability of coincidence detection. A rotating area detector scanner was developed based on a pair of wire chambers equipped with converters of high Z material to convert the 511 keV gamma rays into photoelectrons. Ionisation created by the photoelectrons within the chamber is used to localise the position of the incident photons. Such a scanner was developed by A. Jeavons at CERN in the early eighties\cite{27}, using 5 mm thick lead stacks drilled with a large number of holes as converters; this scanner was operational at the University Hospital in Geneva from 1983 until 1988.

However, despite extensive and considerable effort, tomographs based on gamma cameras\cite{25} and wire-chambers\cite{26,31} have failed to match the remarkable performance of the second family of scanners, based on a large number of individual small scintillation detectors. The main reasons for this are the low intrinsic sensitivity and poor count rate capabilities of the area detectors. To date, all commercial scanners are based on multiple crystal detectors, with the exception of the PennPET scanner which consists of an hexagonal arrangement of rectangular sodium iodide area detectors\cite{32}.

One of the first multiple crystal scanners was built at Brookhaven National Laboratory\cite{33} in 1962 from 32 sodium iodide crystals coupled individually to photomultipliers. This first device could be described as fully three-dimensional: the 32 detectors were arranged on the surface of a sphere surrounding the patient's head. Unfortunately, the small number of detectors did not provide adequate 3D sampling and sensitivity, and suitable algorithms to handle such data were unknown at that time. Thus, the 32 detectors were eventually re-arranged into a circular configuration. The circular configuration (ring design) is still the basis of most modern tomographs, although considerable improvements have been introduced since the early days in Brookhaven. Among the most significant improvements, are the increased number of detectors, which has attained 600 in a single ring positron scanner at the Donner laboratory\cite{34,35}, and the replacement of sodium iodide by bismuth germanate\cite{36,37}, resulting in a much higher sensitivity for a given crystal depth. Furthermore, the development of computerized tomography in the seventies eventually opened the way to an exact two-dimensional reconstruction of the tracer distribution in the transaxial section defined by the ring of detectors. This led, in 1976, to the first commercial PET scanner, the ECAT I designed and built by E.G. & G. Ortec Inc in the USA.

The next step in the development of multi-crystal scanners was the construction of tomographs to image a number of transaxial sections simultaneously, thus yielding a 3D reconstruction of the tracer distribution. The simplest way to achieve this is to stack a number of rings of detectors, each ring yielding the reconstruction of a single transaxial section through the body. In this multi-ring design, coincidences were initially recorded only between detectors belonging to the same ring, in such a way that each transaxial slice (one per ring) was reconstructed independently from the others. Thus, three-dimensional image reconstruction was factorized into a set of 2D problems, enabling the well-established two-dimensional algorithms of computerized tomography to be used. This was an essential advantage over scanners based on area detectors, which require fully three-dimensional reconstruction algorithms. Such 3D algorithms were only developed after 1979, and at a rather slow pace since a limited number of groups were interested in the 3D approach. It was obvious, however, that the available detectors in a multi-ring tomograph could be utilized more efficiently by recording "multi-plane"
coincidences between detectors in any two rings. The now classical multi-ring design goes one step in this direction by recording coincidences between detectors in two adjacent rings; these data, however, are assigned to a so-called “cross” slice located between the planes of the two adjacent rings, and this slice is again reconstructed with the usual two-dimensional technique. Hence, a tomograph with 16 rings, such as the ECAT 953B (CTI/Siemens, Knoxville, Tenn) can reconstruct 16 “direct” slices and 15 “cross” slices. Factorization of the 3D reconstruction is maintained, and this approach is not able to incorporate coincidences between non contiguous rings, since the lines of response corresponding to such events cross several transaxial slices, and cannot be assigned to one single transaxial slice.

As multi-ring scanners are designed to image a number of parallel transaxial slices independently, it is natural to shield the different rings of detectors from each other, in order to suppress the background due to scattered and random coincidences. This is achieved by annular shields (fig. 3), or septa (a term from the plural of septum, a wall separating cavities, as in plants and animals). The optimal design of these septa has been studied in the literature\textsuperscript{38,39}, but the necessity to have septa at all has rarely been questioned, even though the presence of collimators in a PET scanner (other than shielding for activity outside the field of view) contradicts the basic principle of electronic collimation.

![Figure 3: Schematic representation of a 5-ring cylindrical scanner (thick line) with annular interplane septa (grey) and the central patient entry port. The distance between the rings is increased for clarity.](image)

The potential benefit of operating a multi-ring scanner without septa was realised as early as 1982. The neuro-ECAT scanner\textsuperscript{40} consisting of three octagonal arrays (i.e. three rings) of 88 BGO detectors per ring, was equipped with septa which could be inserted in, or retracted from, the scanning unit, under push-button control. Hence, two modes of operation were available: a high contrast mode with septa extended, allowing images to be obtained only within the detector planes; and a high efficiency mode with septa removed offering the possibility to image cross-planes. An increase in sensitivity by a factor of about 1.5 was observed when retracting the septa, but this was accompanied by an increase in the scatter fraction (for the particular set up used by the authors) from 9.2% to 22.7% of the total counts\textsuperscript{41}. The reconstruction, however, was still done on a slice by slice basis, using only coincidences between detectors belonging either to the same ring, or to two adjacent rings.

At about the same time, Mullani et al\textsuperscript{42} proposed to improve the sensitivity of a 4-ring time-of-flight scanner by utilizing all multi-plane coincidences. The authors noted that including these data in the reconstruction process had not been possible with non time-of-flight scanners.
since the multi-plane coincidences "do not behave in a manner suitable for the reconstruction algorithms; therefore some of these coincidences have been classified as useless and discarded". Using the time-of-flight information, on the other hand, 3D reconstruction from multi-plane data was relatively straightforward, although a satisfactory theory of 3D time-of-flight image reconstruction was in fact only presented recently by Grangeat and Mallon\textsuperscript{43}.

Thus, the incorporation of septa in multi-ring tomographs was motivated not only by the need to reject scattered and random coincidences, but also by the lack of appropriate fully three-dimensional reconstruction algorithms. Furthermore, with the 3 or 4 ring scanners available in the early eighties, the gain to be expected from septa removal was rather modest. The number of pairs of detectors in coincidence, and therefore the sensitivity, increases by a factor of \( N^2/(3N-2) \) when the septa are removed from an \( N \) ring scanner, with a further gain of a factor of about 2 due to the additional shadowing effect of the septa.

In parallel with the rapid progress of the multi-ring camera, investigations of tomographs based on rotating planar area detectors were continued by a few groups\textsuperscript{26,30,44}. A number of designs of volume PET scanners have also been proposed recently\textsuperscript{32,45-47,17,18}. In particular, an interesting approach has been taken by G. Charpak at CERN\textsuperscript{17} and E. Bateman at the Rutherford Laboratory in developing a hybrid configuration of BaF\textsubscript{2} crystals inside a wire chamber, using TMAE to convert the UV light emitted from the crystals into visible light. The aim is to combine the high stopping power of scintillation crystals with the high spatial resolution of wire chambers. All PET scanners based on such volume detectors require fully three-dimensional reconstruction, i.e. the recovery of a function defined in \( \mathbb{R}^3 \) from its two-dimensional parallel projections (see section 3). The first results in this field were obtained by Vainstein and Orlov\textsuperscript{48} and Orlov\textsuperscript{49} for the recovery of the structure of biomacromolecules from electron micrographs. Further developments led to a clear mathematical understanding of the problem, and to the definition of 3D inversion formulae suitable not only for rotating area detector scanners, but also for the multi-ring scanners which, when operated without septa, have a similar geometry.

The concept of 3D acquisition using multi-ring scanners without inter-plane septa reappeared in 1988\textsuperscript{50-52}. The important sensitivity improvement, reflected in the quality of the reconstructed images, eventually led to the design by CTI/Siemens of a 16 ring scanner, the ECAT 953B, equipped with retractable septa and with appropriate hardware to collect multi-plane data. As a result of the increased axial length of these modern scanners compared to the neuro-ECAT, the sensitivity improvement achieved by retracting the septa from the 953B was a factor of about six. The possibility of successfully operating multi-ring scanners without septa can also be exploited in the design of a lower cost scanner based on two partial-ring arrays of BGO detectors. The arrays are mounted on a rotating gantry to ensure complete angular sampling. Such a scanner, designed in collaboration with CTI/Siemens, and assembled at CERN, has been operational since 1991 at the University Hospital in Geneva.

A major difficulty with the 3D mode is that the acquisition system must handle much higher data rates, typically up to 0.5 MHz. In addition, there is an increase in the numerical complexity of the image reconstruction, since typically 500,000 voxels must be reconstructed from up to 15 million discrete data samples. As part of the EEC-funded Harmony project, and in collaboration with Geneva University Hospital and the University of Lausanne, a CERN group is currently developing a fast, versatile and scalable acquisition and reconstruction system based on a network of transputers.

Measurement of the regional tissue concentration of a positron-emitting tracer is achieved by acquiring sufficient annihilation coincidence events to enable the tracer distribution within the tissue to be reconstructed. This report will describe some of the computational aspects of the
measurement procedure: the acquisition of coincidence events in a form suitable for reconstruction, the techniques used to reconstruct an internal tracer distribution $f(x,y,z)$ from externally-detected coincidence events, and the computer implementation of the appropriate algorithms. For clarity, the reconstruction methods will be discussed first with continuous data sampling, even though, in practice, data is available only at discrete sample points. For the discrete algorithms, it is necessary to decompose $f(x,y,z)$ into small volume elements, or voxels, and to reconstruct the tracer distribution as a three-dimensional matrix of voxels; the tracer concentration is thus assumed to be constant within each voxel. The voxel dimensions are related to the precision with which the coincident photons can be spatially localised, i.e. the spatial resolution of the tomograph.

2. DATA ACQUISITION

2.1 Lines of response

Coincidences acquired during a PET scan occur in a random order and are usually sorted into distinct angular directions before reconstruction. To illustrate this sorting procedure, consider a pair of detectors in coincidence. Annihilation photons that arise within a voxel intersected by the line joining the centres of the two detectors will be detected provided that their direction of emission is along, or close to, the same line. Since the number of positrons emitted (and hence annihilations occurring) in a given time is proportional to the tracer concentration within the voxel, the number of photons detected is also proportional to tracer concentration, even though a single pair of detectors may capture only a small fraction of the total number of emissions from the voxel. A line joining two detectors in coincidence is called a line of response (LOR).

During a PET scan, coincidences are detected and assigned to their appropriate LORs, the current contents of which are continually updated. In practice, since the detectors are finite in size (with a rectangular cross-section of typical dimensions 5 to 10 mm), the LOR is replaced by a 'tube' (fig.4a). The number of annihilation photon pairs (events) associated with each line, or tube, of response is proportional to the sum of the tracer concentration within that entire tube. This sum approximates the line integral of the tracer concentration along the LOR. The closeness of the sum to a true line integral is evidently related to the geometry of the tubes, i.e. the finite size of the detectors. The image reconstruction algorithms developed in section 3 attempt to recover the tracer distribution function $f(x,y,z)$ from line integrals of the function. Thus, the concept of a line integral is fundamental to the reconstruction procedure.

![Figure 4a: The tube of response for two detectors in coincidence](image-url)
Figure 4b: A 1-D parallel projection of a 2-D section

Figure 4c: A 2-D parallel projection of a 3-D volume
The set of all line integrals intersecting the tracer distribution at the same given angle, but at different spatial positions, is termed a parallel projection. Parallel projections can be one dimensional as in the case of a single 2-D slice projected onto a line (fig. 4b), or two dimensional as in the case of the projection of a 3-D distribution onto the plane orthogonal to the projection direction (fig. 4c).

In reality, for PET, a number of important corrections must be applied to the acquired event count for each LOR in order to obtain true integrals of tracer concentration. These corrections are:

- detector sensitivity normalisation: this correction accounts for the fact that different pairs of detectors may have systematically differing responses to the same activity concentration.
- deadtime correction: at very high tracer concentrations, detectors may cease to respond linearly to changes in concentration.
- randoms correction: 'accidental' coincidences between uncorrelated photons increase the apparent tracer concentration. The importance of this effect depends upon the detection rate of single photons and the time window used to define a coincident event.
- attenuation and scatter correction: photons may be absorbed or scattered by body tissue before reaching the detectors. Attenuation results in an underestimate of the tracer concentration, while the misidentification of scattered photons as true coincidences leads to an overestimate. Attenuation may be corrected fairly easily using a transmission scan, although the procedure will introduce additional noise into the reconstruction. The correction of scattered photons presents considerably more of a problem.

For the purposes of this discussion, it will be assumed that all the necessary corrections have been applied and that the value obtained for each LOR is, within statistical error, a good approximation to the integral of tracer concentration along that LOR.

2.2 Sinograms for a single detector ring

Incoming coincidences are assigned in real time to their respective LORs and stored within the acquisition memory of the tomograph, which is organised into sinograms. A sinogram is simply an ordered way of storing LORs by grouping them into sets of parallel projections. Consider a ring of detectors where the LORs are restricted to a single transverse slice through the tracer distribution, denoted by f(x,y). At a given angle, say \( \phi_0 \), a one dimensional parallel projection of the tracer distribution is obtained by grouping all the LORs with angle \( \phi = \phi_0 \). Each LOR in the set represents the line integral of tracer concentration in the direction \( \phi = \phi_0 \), but at different distances (s) from the centre of the ring, where \( s \) and \( \phi \) are defined as in fig. 5a. A sinogram is simply an (s,\( \phi \)) plot, with each LOR represented by a single point on the plot (fig. 5b). The set of LORs corresponding to a parallel projection is therefore one complete row (fig. 5c), and subsequent sets for different values of \( \phi \) are stored as consecutive rows. Such a plot, or 2-D matrix, is termed a sinogram, with the vertical (column) axis representing projection angle and the horizontal (row) axis representing spatial position within the projection.

Each point in a sinogram is thus associated with a corresponding LOR in the field of view of the detector ring. The set of LORs through a fixed point within the field of view describes a sinusoidal curve on the sinogram, as shown in fig. 5d, hence the origin of the term sinogram. For a point at the centre of the ring, the locus on the sinogram is a straight line, whereas for points away from the centre, the amplitude of the curve increases with increasing distance from the centre.
Figure 5a: The coordinate system for projections of a 2-D function

Figure 5b: A line of response corresponding to one point in a sinogram
Figure 5c: A parallel projection corresponding to one row in a sinogram

Figure 5d: The sinusoidal curve in the sinogram for a point off centre
2.3 Sampling the sinogram

Although the sinogram has been introduced as a matrix of discrete values, it may be considered as a continuous function of angle $\phi$ and projection variable $s$, denoted by $p(s, \phi)$. Indeed, the image reconstruction formulae presented in section 3.1 assume that the function $p(s, \phi)$ is available for any $s$ and $\phi$. In practice, of course, this is not the case, since $p(s, \phi)$ is only measured at specific points defined by the geometry of the ring of discrete detectors. A discretized reconstruction algorithm (section 3.2) must then be used.

The sinogram samples obtained from $N_D$ discrete detectors arranged in a ring of radius $R_d$ are given by:

$$\phi_j = \pi j/N_D, \quad j = 0..N_D-1,$$

$$s_{k,j} = R_d \sin(\pi k/N_D), \quad \text{for } k+j \text{ even}.$$  

Sinogram sampling is shown in fig.6 for a simplified ring consisting of 20 detectors. It can easily be seen that the sinogram sampling characteristics of a ring PET scanner differ from a rectangular grid in two respects:

- the sample points for two adjacent parallel projections (sinogram rows) are shifted by half a detector spacing with respect to one another. This sampling scheme actually results in slightly higher image resolution than with the samples aligned\(^{55}\).

- the sample spacing across a projection (fixed $j$) is not a constant. However, provided the radius of the field of view (i.e. the range of $s_{k,j}$) is small compared with $R_d$, this effect will be small.

Reconstruction algorithms based on this sampling scheme are called fan-beam algorithms, where a fan formed by one detector in coincidence with an arc of detectors opposite corresponds to the set of LORs along a diagonal line in the sinogram, as shown in fig.6. While fan-beam algorithms\(^{56-58}\) have been used extensively in X-ray tomography, parallel-beam algorithms\(^ {59,60}\) have been preferred for PET. Thus, in order to satisfy the parallel-beam requirement of a rectangular sampling grid, the samples are interpolated as shown in fig.6. The sinogram sample points are then given by:

$$\phi_j = \pi j/M, \quad j = 0..M-1 \text{ where } M = N_D/2,$$

$$s_{k,j} = R_d \sin(\pi k/M) = \pi k R_d/M, \quad k = -N/2+1,...,N/2,$$

where $N$ is the number of projection samples. Sinograms sampled according to this rectangular grid are reconstructed using the discrete parallel beam algorithms described in section 3.2. Typical values for the sinogram dimensions of a whole body tomograph are 192 projection samples $\times$ 256 angular samples, with $\Delta s = \pi R_d/M = 3$ mm and $\Delta \phi = \pi/M = 0.7^\circ$.

**Wobble:** a number of schemes have been proposed to improve the sampling for a given tomograph geometry. The approach that has been incorporated into most commercial scanners is that of wobble\(^{61-64}\). In this procedure, the entire tomograph is displaced during scanning by moving the centre of the detector ring around a small circle, typically 1 cm in radius. In practice, the centre is moved to, for example, four distinct points on the circumference of this circle. Data is collected at each of the four points. Such motion is termed wobble, and takes place continuously throughout the scan.

To illustrate the effect of wobble, suppose the centre of the tomograph is moved to the point $(\omega \cos \alpha, \omega \sin \alpha)$ as shown in fig.7, where $\omega$ is the radius of the wobble circle. An LOR which samples the sinogram at $(s, \phi)$ is shifted radially to a point at $(s + \omega \sin(\phi-\alpha), \phi)$.  

\[12\]
Wobbling the tomograph can therefore be used to generate an arbitrary number of intermediate projection samples, depending only on the number of wobble points. The additional samples are then again interpolated into a rectangular sinogram sampling grid; the dimensions of the sinogram will be considerably expanded by this procedure, requiring increased storage capacity. For example, a typical whole body tomograph operated with a four-point wobble generates a sinogram of dimensions $384 \times 512$ with an angular sampling interval of $\Delta\phi/2$ and a projection sampling interval of $\Delta s/2$ compared with the non-wobbled sampling of $\Delta\phi$ and $\Delta s$.

It is important to note that while techniques such as wobble may improve spatial resolution by decreasing the projection sampling interval, they do not provide a corresponding increase in tomograph sensitivity. For a given scan time, the same total number of counts are simply shared between more LORs, and hence the improvement in spatial resolution is at the expense of an increase in image noise. There will obviously be a limit to the improvement that can be achieved using wobble, beyond which an increase in the number of wobble points has no further effect on image quality. From a statistical viewpoint, an improvement in spatial
resolution of a factor of two should be accompanied by a corresponding increase of at least a factor of eight in photon statistics\textsuperscript{6,65}, and therefore, in practice, wobble will be most effective in high count rate studies.

![Diagram of wobble position 1 and wobble position 2](image)

**Figure 7:** Sinogram sampling for two wobble positions

### 2.4 Sinograms for multiple rings of detectors

In the previous discussion, the tomograph geometry was limited to a single ring of detectors. Modern tomographs, however, consist of multiple rings of detectors, with the latest designs containing up to 24 rings. Between each ring, lead or tungsten annuli, or septa, shield the detectors from obliquely-incident radiation, as shown in fig.8. The length of the septa are chosen such that coincidences are accepted not only between detectors in the same ring, but also between detectors in adjacent rings. During a scan, sinograms are therefore acquired not only for the LORs from each detector ring, but also for the LORs between adjacent rings. Sinograms for LORs within the same ring are called direct plane sinograms, whereas sinograms for LORs between adjacent rings are termed cross plane sinograms.

In practice, the angle between a cross plane LOR and a direct plane (denoted by $\theta$ in fig.8) is small ($<1^\circ$) and is usually ignored. Instead, the cross plane LORs are taken to be parallel to the direct plane LORs and to lie in a transverse plane which, axially, is positioned mid-way between the two adjacent direct planes. Furthermore, the LOR between detector a in ring 1 and detector b in ring 2 is summed with the LOR between detector a in ring 2 and detector b in ring 1, i.e. $a_1b_2 + a_2b_1$, fig.8. This results in the photon statistics acquired for the cross planes being up to a factor 2 greater than for direct planes, although the actual increase depends upon the septa design. A tomograph with $n$ rings of detectors acquires data for $n$ direct plane sinograms and $(n-1)$ cross plane sinograms, i.e. a total of $2n-1$ sinograms. Note that in
this case the three-dimensional tracer distribution \( f(x, y, z) \) is recovered by performing a set of 
\( 2n-1 \) independent 2-D reconstructions, one for each sinogram, i.e. \( f(x, y; z_i) \), for \( i = 1..2n-1 \), 
where the \( z \)-axis is parallel to the axis of the tomograph. The tracer distribution \( f(x, y, z) \) is then 
obtained by stacking the reconstructed 2-D slices \( f(x, y; z_i) \); the thickness of each slice is 
denoted by \( \Delta z \).

![Diagram of Direct and Cross Planes for a Four Ring Camera with Septa](image)

**Figure 8:** Direct and cross planes for a four ring camera with septa

### 2.5 Sinograms for full 3-D acquisition

Recent developments in tomograph design\(^{32,45,66}\) include a reduction in the length of 
interplane septa, and the possibility to acquire coincidence data with the septa removed 
entirely\(^{50,52,67}\). The implication of these developments is that coincidences can now be 
acquired between detectors separated axially by more than one ring. For such data, the cross 
plane approximation is no longer satisfactory since the oblique LORs cannot be assigned to 
single transaxial planes. Instead, it is necessary to take the angle of the LOR into account by 
performing a full, three-dimensional reconstruction. The techniques used will be described in 
section 3.

Data acquisition without septa is a straightforward extension of acquisition with septa. 
The LORs are again stored in sinogram form, although the sinograms with polar angle \( \theta \neq 0 \) 
can no longer be assigned to individual transaxial planes. As explained above, while such an 
approximation may be adequate for LORs between detectors in adjacent rings, it is inadequate 
for LORs between detectors separated axially by several centimeters. All LORs between 
detectors in a given pair of rings \( i, j \) are stored in two sinograms \( A_{ij}(s, \phi) \) and \( A_{ji}(s, \phi) \), with 
the angle \( \phi \) in the range \([0, \pi]\) to avoid redundancy. Thus, a 16-ring tomograph acquires 256 
(i.e. \( 16 \times 16 \)) sinograms during a scan with septa retracted. Note that one row (\( \phi \) constant) in a 
sinogram with \( |i-j| > 1 \) can no longer be interpreted as a one dimensional parallel projection. 
This is because the angle \( \theta \) varies along the sinogram row, although the effect will be small 
provided the field of view is small compared with the diameter of the detector ring. For a 
whole-body tomograph, the variation of \( \theta \) within a projection will be less than 5\% for a 40 cm 
diameter field of view.
2.6 List mode data acquisition

Some tomographs incorporate the possibility to acquire data in list mode, so called because, instead of sinograms, a list of the detector numbers of individual coincidences is collected, interleaved with tag words at fixed time intervals. The main reason for this approach, which greatly increases the requirement for on-line data storage, is the possibility to interleave the coincidence data with the status of an external signal. Generally, such a signal will be physiological, relating to some state of the patient at the instant the particular coincidence was acquired. The most common example is the electro-cardiogram (ECG) signal used in gated cardiac studies. An example of a non-physiological signal would be the availability of energy information for each coincidence event that could then be used to correct for background due to scattered photons.

At the end of the scan, the list mode data must be sorted 'off-line' into sinograms, but in this case different sets of sinograms can be assembled by sorting the coincidences according to the particular values of the external signal, e.g. separate intervals in the cardiac cycle. Thus, this sorting procedure results in a complete set of normalized and corrected sinograms, \( p_\sigma(s,\phi) \), as input to the image reconstruction process, for each distinct value \( \sigma \) of the external signal.

3. IMAGE RECONSTRUCTION

In the previous section, it has been shown that the coincidence data acquired by a positron tomograph are, after normalisation and correction for various distortions and background effects, essentially parallel projections of the tracer distribution. These projections are stored in the form of sinograms. Techniques based on reconstruction from projections are therefore used to recover the three-dimensional tracer distribution.\(^6\text{-}\text{54,55,68}\) The most common technique in PET as in X-ray tomography (CT) is filtered backprojection, which will be described in section 3.1 for a continuous sampling of the sinograms. Discrete sinogram sampling will be introduced in section 3.2, and the reconstruction algorithm modified accordingly. The extension of these algorithms to incorporate the oblique angle data acquired without septa (section 2.5 above) will be discussed in section 3.3, although the procedure is not straightforward owing to the non-uniform distribution of the oblique LORs within the field of view. The discretized version of the 3-D algorithm will be presented in section 3.4. Finally, in section 3.5, some alternative approaches based on iterative reconstruction methods will be discussed.

3.1 Filtered backprojection in 2-D: the continuous case

For simplicity, consider the reconstruction, from the corresponding sinogram, of a single transverse slice for a fixed value of the axial coordinate \( z \). For this slice, the tracer distribution is represented by a 2-D function \( f(x,y) \) (since \( z \) is fixed), and the concentration is assumed to be constant in the axial direction over a distance \( \Delta z \), the 'thickness' of the slice. As was explained above, each sinogram value \( p(s,\phi) \) approximates a line integral through \( f(x,y) \) at a particular angle \( \phi \) and at a particular distance \( s \) from the origin. The problem, therefore, is to recover \( f(x,y) \) from the measurements \( p(s,\phi) \), assuming that \( f(x,y) \) is defined for a finite, circular region of radius \( R \) called the field of view, i.e. for values of \( x,y \) such that \( \sqrt{x^2+y^2} < R \). The line integrals \( p(s,\phi) \) may be written explicitly as:

\[
p(s,\phi) = \int dt \left( s \cos \phi - t \sin \phi, s \sin \phi + t \cos \phi \right)
\]

(3.1)

where, as shown in fig.5a, the integration variable \( t \) is along a line orthogonal to \( s \). In this section, it is assumed that \( p(s,\phi) \) is measured for any \( |s| < R \) and for any angle \( 0 < \phi < \pi \), i.e.
s and φ are continuous variables. The function \( p(s, \phi) \) is termed the Radon transform of \( f(x,y) \), and the solution of equation 3.1 for \( f(x,y) \) is unique. This solution has been given in a number of mathematically equivalent forms, the first of which was due to Radon. In two dimensions, the Radon transform is identical to the X-ray transform.

Insight into the relationship between a parallel projection \( p(s, \phi) \) and the original tracer distribution \( f(x,y) \) can be obtained by taking the 1-D Fourier transform of \( p(s, \phi) \) with respect to \( s \).

\[
P(v, \phi) = \int ds \ p(s, \phi) \ e^{-2\pi i s v}
\]

Substituting for the projection \( p(s, \phi) \) from equation (3.1) gives:

\[
P(v, \phi) = \int \int ds \ dt \ f(s \cos \phi - t \sin \phi, s \sin \phi + t \cos \phi) \ e^{-2\pi i s v}
\]

\[
= \int \int dx \ dy \ f(x, y) \ e^{-2\pi i (x \cos \phi + y \sin \phi) v}
\]

\[
= F(v \cos \phi, v \sin \phi)
\]

Therefore, the 1-D Fourier transform of the projection data is the 2-D Fourier transform of tracer concentration along a line in Fourier space at the same angle as the projection angle.

In most cases, the reconstruction of PET data is based on the filtered backprojection approach for parallel projections. This algorithm originated from work in a number of fields, including radio astronomy and electron microscopy. A similar approach, modified to incorporate divergent as opposed to parallel projections, is currently used in X-ray computer-assisted tomography. The implementation for PET requires two steps:

**Step 1: Convolution or filtering:**

Each projection \( p(s, \phi) \) at a given \( \phi \) is convolved with a function (kernel) \( h(s) \) to yield a modified projection \( p_F(s, \phi) \), where:

\[
p_F(s, \phi) = \int ds' \ p(s', \phi) \ h(s-s')
\]  

(3.2)

The kernel is the inverse Fourier transform of the ramp function, \( lv1 \), related to the Jacobian of the transformation to polar coordinates (\( v, \phi \)). In practice, for reasons to be discussed below, it is necessary to limit the high-frequency behaviour of the ramp by multiplying by a low-pass window function \( W(v) \), such that \( h(s) \) is given by

\[
h(s) = \int dv \ lv1 \ W(v) \ e^{-2\pi i s v}
\]  

(3.3)

The convolution may be equivalently performed in Fourier space where the convolution operation reduces to a simple multiplication. Thus:

\[
p_F(s, \phi) = \int dv \ P_F(v, \phi) \ e^{+2\pi i s v}
\]  

(3.4)

\[
P_F(v, \phi) = lv1 \ W(v) \int ds \ p(s, \phi) \ e^{-2\pi i s v}
\]

The Fourier transform of the convolving function \( h(s) \), i.e. \( lv1 \ W(v) \), is called a filter, and hence equation (3.4) is referred to as the filtering step and \( p_F(s, \phi) \) as filtered projections.
Step 2: Backprojection:

The reconstruction \( f_R(x,y) \) of the tracer function \( f(x,y) \) is formed from the filtered projections by redistributing the filtered projection value \( p_F(s,\phi) \) uniformly along the straight line \((s,\phi)\):

\[
f_R(x,y) = \int_0^\pi d\phi \ p_F(x \cos \phi + y \sin \phi, \ \phi)
\]

(3.5)

This operation is called backprojection since formally it is the reverse of the projection process, the value of \( f_R(x,y) \) at the point \((x,y)\) being obtained by summing the contributions from all lines \((s,\phi)\) that pass through the point \((x,y)\), i.e. those with coordinate \( s = x \cos \phi + y \sin \phi \), (fig.5d).

In practice, PET images are reconstructed from sinograms using discretized versions of equations (3.4) and (3.5), (section 3.2 below). However, to see how this procedure works in the continuous case, it is instructive to consider a simple example, such as a tracer distribution \( f(x,y) \) consisting of two disks with the same concentration but different radii (fig.9). In the first step, the parallel projection of \( f(x,y) \), for \( \phi = 0 \), is shown. The height of the projection of the large disk is greater than that of the small disk, so that backprojection without filtering would

Figure 9: Reconstruction of a simulated 2-D image consisting of two disks with different radii and equal tracer concentrations
result in the large disk appearing to have a greater concentration. The filtering operation (step 2) enhances small structures corresponding to high frequencies, by multiplying the frequency spectrum of the projection by the ramp function \( I_{ml} \). As a consequence of this operation, the filtered projection contains both positive and negative values. The role of these negative values is to cancel the excess of positive contributions that would otherwise occur at each point in the image during backprojection. An obvious case is a voxel lying outside both disks where the tracer concentration is initially zero. Without the negative contributions, the reconstructed value in such a voxel would be positive rather than zero. The final step (3) is to backproject the filtered projections so as to recover the original function.

Fig. 10 shows a single slice of a cerebral blood flow study using \(^{15}\text{O}_2\). In fig. 10a the unfiltered sinogram is shown and in fig. 10b the filtered sinogram. The result of backprojecting the unfiltered sinogram (fig. 10c) and the filtered sinogram (fig. 10d) clearly demonstrates the improved detail and contrast in the filtered image.

Figure 10: A cerebral blood flow study obtained using \(^{15}\text{O}_2\) showing (a) the unfiltered sinogram, (b) the filtered sinogram, (c) the image reconstructed from the unfiltered sinogram, and (d) the image reconstructed from the filtered sinogram
An important feature of the kernel $h(s)$ that appears in equation (3.2) is its dependence on the distance $s-s'$ rather than on the separate coordinates $s$ and $s'$. This property relates to the fact that the kernel is shift-invariant, meaning that all projection values are filtered using the same kernel. Any translation of the tracer distribution $f(x,y)$ results in a corresponding translation of the projection data $p(s,\phi)$, without modifying the form of equation (3.1). The simple form of equation (3.4) is then a direct consequence of the shift-invariance of $h(s)$. In practice, however, shift-invariance is only approximately satisfied due to the sensitivity variations within an LOR and to differing LOR cross sections. Nevertheless, these variations are, in general, small, and shift-invariance is a reasonable approximation for the reconstruction of direct and cross plane sinograms. Unfortunately, shift-invariance is not satisfied in the case of three-dimensional acquisition, as will be seen later.

The filtered backprojection algorithm yields an exact inversion of the X-ray transform (equation (3.1)) if $W(v) = 1$ for all $v$. Note that in this case the integral (3.3) diverges and the kernel $h(s)$ is then defined as a distribution. Under such a condition, the filtering operation with the kernel $h(s)$ must correctly recover arbitrarily large frequencies, a procedure which, in practice, is clearly not possible. Hence, the concept of an exact solution has no meaning because:

- the discrete sampling of the sinograms (see section 2.3) sets an upper bound on the maximum frequency that can be recovered. According to the sampling theorem this frequency will be of order $1/2\Delta s$, assuming a sufficiently fine angular sampling interval $\Delta \phi$. The inclusion of higher frequencies in the reconstruction will result in aliasing artifacts.

- the measured sinograms will include both statistical and systematic errors. In particular, the power spectrum (i.e. the squared modulus of the Fourier transform) of the exact, noise-free projection $p_{e}(s,\phi)$ typically decreases more rapidly than the corresponding power spectrum of measurement noise, $n(s,\phi)$, where the measured projection $p(s,\phi) = p_{e}(s,\phi) + n(s,\phi)$. Thus, above a certain frequency (which depends on the signal to noise ratio) the Fourier transform of $p(s,\phi)$ is dominated by noise, and the application of the ramp filter to higher frequencies results in the amplification of noise, degrading the quality of the reconstructed image.

These are the main reasons for the use of a low-pass window $W(v)$ which eliminates any contribution from frequencies above a certain limit $v_c$. The appropriate choice of $v_c$ is determined by both the sampling distance (which limits the maximum recoverable frequency), and the photon statistics (which limit the signal to noise ratio). The spatial resolution of the final image is determined by the maximum frequency recovered by the reconstruction, and hence depends not only on the detector resolution ($\Delta s$) but also on the photon statistics acquired during the scan.

A typical low-pass window used in PET reconstructions is the generalized Hamming window:

\[
W(v) = \begin{cases} 
\alpha + (1-\alpha) \cos(\pi v/v_c), & v \leq v_c, \\
0, & v > v_c.
\end{cases}
\]  

(3.6)

The parameter $\alpha$, for a given cut-off frequency $v_c$, controls the smoothness of the transition from $v = 0$ up to $v = v_c$. For $\alpha = 1$, $W(v) = 1$ for $v \leq v_c$ and $W(v) = 0$ for $v > v_c$, and the filter $W(v)$ is a simple ramp function up to the cut-off frequency, after which it is zero (fig.11). Unfortunately, photon statistics are not usually adequate for such an enhancement of the higher frequency components, while the abrupt decrease to zero at $v = v_c$ may result in
oscillatory reconstruction artifacts in the image. Instead, a more smoothly varying window, such as that obtained with $\alpha = 0.5$, is usually preferred (fig.11). A considerable number of low-pass window functions have appeared in the literature\textsuperscript{73}, although the only properties of $W(v)$ that have significant influence on image quality are the frequency cut-off $v_c$, and the way in which the function approaches $v = v_c$.

Before addressing the discretization problem, it is worth mentioning the modification to the filtered backprojection algorithm introduced by the measurement of time-of-flight information, although it is beyond the scope of these lectures to discuss time-of-flight (TOF) tomographs in detail\textsuperscript{74,75}. Time-of-flight tomographs measure the time difference between the arrival of coincident photons with a current precision of around 330 picoseconds\textsuperscript{76}, corresponding to an uncertainty of 5 cm in the point of emission. Thus, while such precision is insufficient to eliminate the need for an image reconstruction algorithm, it does give some additional information on the location of the emission point. In practice, the backprojection operation can be modified to incorporate this information so that, rather than distributing the projection data uniformly along the LOR as for a conventional tomograph, more weight can be assigned to the most probable point of emission, as determined by the time-of-flight measurement. For a given total number of coincidences, restricting the backprojection operation to a localised region around the emission point results in better signal to noise in the reconstructed image than that obtained with a non-TOF tomograph.

3.2 Filtered backprojection in 2-D: the discrete implementation

In practice, as described in section 2.3, the sinograms are sampled at discrete points. Parallel projections are obtained at $M$ angles $\phi_j = \pi j/M$, $j = 0..M-1$, and each parallel projection is sampled at $N$ equally-spaced sample points $s_k = k\Delta s$, $k = -N/2+1,..,N/2$. The filtered backprojection algorithm must be discretized appropriately to reconstruct an image from the available samples\textsuperscript{54,60,68}.

\textit{Step 1: Filtering}:

Equation (3.2) can be discretized using the trapezoidal quadrature rule to give:

$$p_{F}(k\Delta s, \phi_j) = \Delta s \sum_{n=-N/2+1}^{N/2} p(n\Delta s, \phi_j) h(k - n\Delta s)$$

$$k = -N/2+1,..,N/2$$

(3.7)
The sum approximates the integral of equation (3.2) in the range \([-N/2+1 \Delta s, N/2 \Delta s]\). If the function \(f(x,y) \neq 0\) outside a circle of radius \(N\Delta s/2\), reconstruction artifacts result from the truncation of \(p(s,\phi)\). The filtered projection will in general have non-zero values outside the measured range \([-N/2+1 \Delta s, N\Delta s/2]\) but these need not be computed since they do not contribute to the reconstruction \(f_R(x,y)\) within the circular field of view.

**Step 2: Backprojection**:

In the same way, equation (3.5) can be discretized for the \(M\) projections \(\phi_j\) to give:

\[
f_R(x,y) = \frac{\pi}{M} \sum_{j=0}^{M-1} p_F(x \cos \frac{\pi j}{M} + y \sin \frac{\pi j}{M}, \phi_j) \tag{3.8}
\]

The sum is computed at each point \((x,y)\) required for a square grid of voxels, i.e. the discretized image matrix, where \((x,y)\) is a point at the centre of a voxel. For a given \((x,y)\), the expression \(x \cos (\pi j/M) + y \sin (\pi j/M)\) will not in general coincide with one of the \(N\) measured samples of \(p_F(k\Delta s, \phi)\). Interpolation between adjacent measured samples is therefore necessary to evaluate the sample at \(x \cos (\pi j/M) + y \sin (\pi j/M)\); linear interpolation offers a reasonable compromise between accuracy and calculation speed:

\[
p_F(s, \phi_j) = \{ (k+1) - s/\Delta s \} p_F(k\Delta s, \phi_j) + \{ s/\Delta s - k \} p_F((k+1)\Delta s, \phi_j) \tag{3.9}
\]

for any \(s\) where \(k \leq s/\Delta s < (k+1)\) and \(k = -N/2+1, \ldots, N/2\).

Voxel-driven (or pixel-driven for a single, 2-D slice) backprojection is an efficient implementation of equation (3.8). In this approach, all image voxels are processed sequentially for each projection in turn. Thus, for a given projection, each voxel address \((x,y)\) is used to compute the address of the projection sample that contributes to that voxel, according to equation (3.8); if necessary, the interpolation scheme of equation (3.9) is used. The current voxel contents are then summed with the interpolated projection sample. This procedure is then repeated for all voxels, and for all projections. In practice, it is more efficient to pre-interpolate or 'stretch' the projection prior to backprojection by interpolating a fixed number of, say eight, intermediate samples between each pair of measured samples\(^{35}\). Then, during backprojection, samples are obtained by a fast nearest-neighbour interpolation algorithm rather than by linear interpolation.

Apart from the numerical problems that inevitably arise with the discrete implementation of a continuous solution, it is interesting to estimate to what extent the discrete result is a reliable representation of the continuous solution. It is expected that the discrete result will be at least a reasonable approximation to the continuous solution provided that the sampling of the sinogram fulfills certain conditions set down by Shannon's sampling theorem\(^{35}\).

For example, consider a tracer distribution \(f(x,y)\) that is zero outside a disk of radius \(R\). The Fourier transform of such a distribution will have infinite extent, but assume that it is essentially confined to a disk of radius \(\Omega\) in frequency space. That is:

\[
F(v_x, v_y) = \int \int dx \, dy \, f(x,y) \, e^{-2\pi i (v_x x + v_y y)} = 0 \tag{3.10}
\]

for \(\sqrt{v_x^2 + v_y^2} > \Omega\)

\[\text{for } \sqrt{v_x^2 + v_y^2} > \Omega\]
Shannon's sampling theorem then states that $f(x,y)$ can be recovered accurately provided that the sampling of the data satisfies the conditions:

$$\Delta s \leq 1/2\Omega \quad \text{and} \quad \Delta \phi = \pi/M \leq 1/2\Omega R \quad (3.11)$$

Then, using the limiting values for $\Delta s$ and $\Delta \phi$ from equation (3.11), and since $\Delta s = 2R/N$, the relationship $M = N\pi/2$ is obtained. In a typical whole-body tomograph, $N\pi/2 = 300$, while $M = 256$, which is good agreement for a consistent sinogram sampling scheme. Since $\Delta s$ and $\Delta \phi$ are fixed by the scanner geometry, when equations (3.11) are approximately satisfied, the maximum spatial frequency that can be recovered from the data is $1/2\Delta s$, which is called the Nyquist frequency. Thus, in the absence of statistical noise, an appropriate choice of cut-off for the low-pass window, equation (3.6), would be $\nu_c = 1/2\Delta s$, particularly as the data do not contain any information about higher frequencies. However, in the presence of noise, it may be necessary to set $\nu_c$ to a lower value to limit noise amplification.

An illustration of the artifacts arising from angular or projection (radial) undersampling is shown in fig.12. In fig.12a, the cerebral flow study of fig.10 is reconstructed from 192 radial samples and 256 projections, which satisfies the conditions in equation (3.11). In fig.12b, the same scan is reconstructed using only 96 radial samples per projection, and in fig.12c the scan is reconstructed from 64 projections, the voxel size being kept the same for all reconstructions. While the quality of both undersampled images is clearly degraded, radial undersampling results in more severe artifacts.

### 3.3 Filtered backprojection in 3-D: the continuous case

It has been seen in section 2.4 that, for a tomograph comprising multiple rings of detectors, a three-dimensional image of the tracer distribution $f(x,y,z)$ can be built up by stacking a set of independent 2-D slices $f(x,y; z_i)$ for $i = 1..2n-1$, where $n$ is the number of detector rings. Associated with each slice is either a direct or cross plane sinogram, which is used as input data to the 2-D reconstruction algorithm outlined in section 3.2 above. Each slice is reconstructed independently of the adjacent slices. However, as explained in section 2.5, this procedure is no longer possible when coincidences are acquired between all rings of detectors.
A full, 3-D acquisition necessitates a full 3-D reconstruction algorithm in order to correctly incorporate the sinograms for the oblique LORs.

The problem, therefore, is to recover the tracer concentration \( f(x,y,z) \) from a set of line integrals in three dimensions. Representing \( f(x,y,z) \) in vector notation as \( f(x) \), the line integrals are defined by:

\[
p(s,t,\phi,\theta) = \int d\xi \ f(\xi + \xi \hat{u})
\]  

(3.12)

where \( \hat{u} \) is a unit vector with components \((-\sin\phi \cos\theta, \cos\phi \cos\theta, \sin\theta)\) which specifies the direction of the line along which integration is performed. This equation is the 3-D analogue of equation (3.1) above. The values of the function \( f(x,y,z) \) at the points \((\xi + \xi \hat{u})\) are summed for all values of \( \xi \); these points lie on a line in the direction of \( \hat{u} \) which passes through the point \( \xi \).

The components of \( \xi \) are related to the projection coordinates \((s,t)\) by:

\[
\xi = (s \cos\phi + t \sin\phi \sin\theta, s \sin\phi \cos\sin\theta, t \cos\theta)
\]

As shown in fig.13, the vectors \( \xi \) and \( \hat{u} \) are orthogonal and, when \( \theta = 0 \), equation (3.12) reduces to the 2-D case, equation (3.1), for the transverse plane with \( z = t \).

Reconstruction of the tracer distribution \( f(x,y,z) \) from the line integrals \( p(s,t,\phi,\theta) \) is therefore a problem of inverting the X-ray transform\(^{70}\) in three dimensions. Unlike the 2-D case, in 3-D the Radon transform is not equivalent to the X-ray transform, equation (3.12). Whether or not inversion is possible depends upon the set of line integrals which have been measured by the tomograph\(^{49,78}\). Assume initially that line integrals are measured for a limited set of directions \( 0 \leq \phi < \pi \) and \(-\Psi \leq \theta < \Psi \), but for all coordinate pairs \((s,t)\) for which the straight line \((s,t,\phi,\theta)\) intersects the 3-D field of view. This means that, while not all possible 2-D parallel projections (section 2.1, fig.4c) are measured because of the limited range of \( \theta \), those projections which are measured are complete. The importance of completely measured projections, and the relationship to the measured sinograms will be discussed later.

![Figure 13: The definition of the 3-D coordinate system for 2-D parallel projections](image)

2-D Parallel Projection

\[ p(s, t, \phi, \theta) \]
Although not all possible line integrals through the field of view are available (since $|\theta| < \Psi < \pi$), the solution of equation (3.12) is nevertheless unique, and can be obtained by a straightforward generalisation of equations (3.2) to (3.5) above\textsuperscript{79-82}.

**Step 1: Convolution or filtering:**

Each 2-D parallel projection is convolved with a kernel $h(s,t,\theta)$, independent of $\phi$, to give the filtered projection $p_F(s,t,\phi,\theta)$:

$$p_F(s,t,\phi,\theta) = \int \int ds' \ dt' \ p(s',t',\phi,\theta) \ h(s-s',t-t',\theta) \quad (3.13)$$

As for the 2-D case, this filtering operation may be more efficiently performed using the Fourier convolution theorem than by implementing equation (3.13) directly.

**Step 2: Backprojection:**

The 2-D filtered projections are backprojected by redistributing the values $p_F(s,t,\phi,\theta)$ uniformly along the line $(s,t,\phi,\theta)$ so as to form the reconstructed image:

$$f_R(x,y,z) = \int \int \int_0^\pi \ d\phi \ d\theta \ \cos\theta \ p_F(s,t,\phi,\theta) \quad (3.14)$$

with $s = x \cos\phi + y \sin\phi$ and $t = x \sin\phi \sin\theta - y \cos\phi \sin\theta + z \cos\theta$.

The expressions in equation (3.14) for the projection coordinates $(s,t)$ ensure that the line $(s,t,\phi,\theta)$ passes through the point $(x,y,z)$. Thus, as in the 2-D case, the reconstructed value $f_R(x,y,z)$ at the point $(x,y,z)$ is obtained by summing the contributions from all the backprojection lines $(s,t,\phi,\theta)$ which pass through $(x,y,z)$.

In analogy with equation (3.3), the convolution kernel $h(s,t,\theta)$ may be expressed as the Fourier transform of a filter function $H(v_s,v_t,\theta)$ multiplied by a low-pass window function $W(v_s,v_t)$:

$$h(s,t,\theta) = \int \int dv_s \ dv_t \ H(v_s,v_t,\theta) \ W(v_s,v_t) \ e^{2\pi i (sv_s + tv_t)} \quad (3.15)$$

where $(v_s,v_t)$ are frequency space cartesian coordinates corresponding to the projection coordinates $(s,t)$. Equivalently, polar coordinates $(\rho,\alpha)$ can be defined by:

$$v_s = \rho \cos\alpha, \ v_t = \rho \sin\alpha, \text{ with } \rho^2 = v_s^2 + v_t^2$$

An appropriate filter function $H(v_s,v_t,\theta)$ has been published by Colsher\textsuperscript{80} expressed in frequency space polar coordinates on the projection plane as:

$$H(\rho \cos\alpha, \rho \sin\alpha, \theta) = \frac{\pi \rho}{\arcsin \frac{\sin\Psi}{Z}} \quad \text{if } Z \geq \sin\Psi$$

$$= 2\rho \quad \text{if } Z < \sin\Psi \quad (3.16)$$

where $Z = \sqrt{(\cos^2\alpha + \sin^2\alpha \sin^2\theta)}$ and $|\theta| \leq \Psi$.

As in 2-D (equation (3.3)), the filter function is proportional to the modulus of the frequency, $\rho$. The same considerations apply as in 2-D concerning high frequency noise amplification, hence the reason for the low-pass window function in equation (3.16).
addition, since equation (3.13) is a convolution equation, the measured data set \( p(s, t, \phi, \theta) \) is required to be shift-invariant.

Unfortunately, in practice, the three-dimensional data set acquired by a multi-ring PET tomograph does not satisfy the conditions for shift-invariance. This is easily seen by considering a single point concentration of tracer. When the point is located at the axial centre of the tomograph, more of the annihilation photons are emitted along LORs than when the same point is displaced axially to one of the outer slices (fig.14a). Thus the same tracer concentration appears to be greater when imaged at the centre of the field of view than when imaged at the edge. For any general tracer distribution, the particular cylindrical geometry of the multi-ring PET scanner results in the 2-D parallel projections becoming increasingly truncated as \( \theta \) increases (fig.14b). Such a situation is not even approximately shift-invariant, with up to seven or eight times as many LORs passing through the central slice as through the extreme edge slices. Under these conditions, the filtered backprojection algorithm cannot be applied to reconstruct the data.

If, however, some angle \( \psi \) can be defined such that, for \(|\theta| \leq \psi\), say, the corresponding parallel projections are completely measured, then the algorithm could successfully be applied to a subset of the three-dimensional data, rejecting those partially-measured projections with angles \(|\theta| > \psi\)^79,80. Indeed, this was the condition under which the algorithm was originally derived earlier in this section - that all projections for \(|\theta| \leq \Psi\) were completely measured. From fig.14c, it can be seen that, as the tracer distribution extends axially from the centre, the angle \( \psi \) for which all projections are completely measured becomes progressively smaller. Finally, when the tracer distribution covers the full axial extent of the tomograph, \( \psi \approx 0 \) and the situation reduces to that of the independent, two-dimensional slices which were discussed in section 3.1, i.e. sinograms with \(|\theta| > 0\) are not included in the reconstruction.

One of the simplest solutions to the problem of how to incorporate the oblique LORs from a multi-ring tomograph into a 3-D reconstruction using filtered backprojection is based on the following idea^83-85: since the projections which give rise to shift-variance are incomplete, and it is desired to include them in a filtered backprojection reconstruction, why not complete these projections before including them. Then the reconstruction will again involve only completely-measured projections and can be performed, as outlined earlier in this section, by 3-D filtered backprojection.

The problem of how to complete the partially-measured projections can be resolved by using the one set of complete projections that are measured, i.e. those for \( \theta = 0 \), to reconstruct an initial image using the 2-D algorithm described in section 3.1. With this initial image, it is then possible to simulate the projection process mathematically and create the unmeasured parts of a projection by 'forward projecting' along LORs that do not exist in the actual tomograph. The procedure is equivalent to increasing the effective axial extent of the tomograph such that all projections for \(|\theta| \leq \psi\) are complete, as shown in fig.14d, where \( \psi = \Psi \), the maximum acceptance angle of the tomograph. It is obvious that this approach does not replace additional rings of detectors which actually measure the required LORs.

The steps in the algorithm are therefore as follows:

1. Reconstruct a first estimate of the image using only the direct and cross plane sinograms (\( \theta = 0 \)) and a 2-D reconstruction technique such as equations (3.4) and (3.5) above.
2. Based on this image estimate, reproject the non-existent LORs in order to complete the partially-measured projections for \(|\theta| > 0\). If desired, \(|\theta|\) can extend out to \( \Psi \), the maximum acceptance angle of the tomograph. However, at these angles the number of measured LORs may be small compared with the number that have been obtained by reprojection.

26
3. With a set of complete projections for all $|\theta| \leq \psi$, reconstruct a final image incorporating all sinograms using equations (3.13) to (3.16) above.

In the ideal situation of continuously-sampled and noise-free data, the final image from step 3 will be identical to the initial estimate from step 1. It is only in the presence of statistical noise that the incorporation of the additional oblique LORs serves to improve the signal to noise of the final 3-D reconstruction.

Figure 14a: Axial shift variance of a point source

Figure 14b: The truncation of 2-D projections with increasing ring difference ($\theta$)
Figure 14c: The dependence of the angle $\psi$ on the axial extent of the tracer distribution ($\psi < \Psi$)

Extended detectors

Figure 14d: Extending the effective axial length of the scanner by forward projection ($\psi = \Psi$)
3.4 Reconstruction in 3-D: the discrete implementation

In marked contrast to the situation for 2-D reconstruction, the study of problems related to the numerical implementation of 3-D reconstruction algorithms for discretely-sampled data is still in its infancy\textsuperscript{52,67,83,84,86}. Understanding of the subject will doubtless improve in the coming years, and therefore the discussion in this section will be restricted to a description of the current trend.

There are two reasons why the discretization of equations (3.13) and (3.14) is not as straightforward as the two-dimensional counterpart, equation (3.3). Firstly, the multi-angle data are still acquired in the form of sinograms (section 2.5), and it is therefore necessary to relate the sinograms to parallel projections. Secondly, the filtering step explicitly requires two-dimensional parallel projections \( p(s,t,\phi,\theta) \) to be formed. A parallel projection is obtained by projecting the 3-D tracer distribution onto a plane perpendicular to the projection direction (fig.5b). A cylindrical detector such as a multi-ring tomograph does not provide such projections because, with the exception of \( \theta=0 \), a projection 'plane' is not perpendicular to the projection direction. Thus, unless certain approximations are satisfied, rebinning of the data from a cylindrical detector is necessary in order to obtain true parallel projections.

Each sinogram \( \delta_{ij}(s,\phi) \) contains LORs between a given pair of rings with indices \( i \) and \( j \). Consider the set of sinograms for a fixed ring index difference \( \delta = i-j \). For example, for an eight ring tomograph, the sinogram set with \( \delta=3 \) will contain sinograms with indices \( [i,j] \) equal to \([4,1], [5,2], [6,3], [7,4] \) and \([8,5]\), with each sinogram distinguished by the axial \((z)\) coordinates \( z_i \) and \( z_j \) of the centres of the two rings \( i \) and \( j \). Such a sinogram set is related to a parallel projection by:

\[
\delta_{ij}(s,\phi) = p(s,t,\phi,\theta). \tag{3.17}
\]

The sinogram variables \( s \) and \( \phi \) are directly those of the parallel projection, whereas the projection coordinate \( t \) is related to the mean axial coordinate \((z_i+z_j)/2\), and the projection angle \( \theta \) is related to the difference \((z_i-z_j)\). The exact relationships are:

\[
\sin \theta = \frac{(z_i - z_j)}{\sqrt{4(R_d^2 - s^2) + (z_i - z_j)^2}} \tag{3.18}
\]

\[
t = \cos \theta \frac{(z_i + z_j)}{2}
\]

where again \( R_d \) is the radius of the detector ring. When both the radius of the field of view (R) and the maximum axial extent of the tomograph are small compared with \( R_d \), \( \sin \theta \) may be approximated by:

\[
\sin \theta = \frac{(z_i - z_j)}{2R_d}. \tag{3.18a}
\]

Since, with this approximation, \( \theta \) no longer depends upon the projection coordinate \( s \), each sinogram row is a one dimensional parallel projection. Thus, a two dimensional parallel projection at azimuthal angle \( \phi \) is obtained by extracting, from each sinogram for a fixed ring difference \( \delta \), the row corresponding to \( \phi \). Clearly, if the approximations used above are not valid, the PET sinogram data will require rebinning in order to obtain true parallel projections. Iterative\textsuperscript{87} or non-iterative\textsuperscript{88,89} algorithms which do not explicitly require parallel projections will evidently not be affected by these approximations.

Therefore, a sampled set of two-dimensional parallel projections are obtained from the sinogram data at each of \( M \) equally-spaced angles \( \phi = \pi \ell/M, \ell = 0..M-1 \), and at polar angles \( \theta_\delta = \arcsin(\delta \Delta d/2 R_d), \delta = -n+1...,n-1 \). Each parallel projection is sampled at \( s_k = k\Delta s, \)
k = \text{-}N/2+1,...,N/2 and at \( r = r\Delta d \cos \theta \), for \( r = 1..n_r \), where \( n_r \) is the number of sinograms in the set and can be expressed as a matrix of dimensions \( N \times n_r \) given by:

\[
\mathcal{M}_{\delta,k}(k,r) = \mathcal{A}_{r+\delta,r}(k\Delta s, \phi_k),
\]

(3.19)

with \( k = \text{-}N/2+1,...,N/2 \) and \( r = 1..n_r \).

It has been implicitly assumed in this section that the 2-D projections are complete, i.e. that the data are shift invariant. In the previous section, it was seen that this is generally not the case and that unmeasured sinograms must be obtained by a forward projection procedure. This amounts to extending the range of the sinogram index \( r \) such that \( \mathcal{A}_{r+\delta,r}(k\Delta s, \phi_k) \) samples the entire 2-D parallel projection. The total number of unmeasured sinograms that have to be forward projected to complete all parallel projections depends on the radial and axial extent of the tracer distribution, and on \( \psi \). Typically, it is equal to the number of measured sinograms. In the following description of the discrete algorithm, it will be assumed that the matrix \( \mathcal{M}_{\delta,k}(k,r) \) completely samples the corresponding parallel projections for all \( \theta \), where \( |\theta| \leq \psi \).

**Step 1: Filtering:**

Using the sampling scheme for \( s \) and \( t \) described above, the sampled version of the convolution kernel \( h(s,t,\theta) \) which appears in equation (3.13) is \( h_s(k\Delta s,r\Delta d) \), where the \( \theta \)-dependence has been written explicitly as a dependence on the ring difference \( \delta \). Then, by analogy with equation (3.7), equation (3.13) can be discretized as:

\[
\mathcal{M}_{\delta,k}^F(k,r) = \Delta s \Delta t \sum_{n=-N/2+1}^{N/2} \sum_{u=1}^{n_r} \mathcal{M}_{\delta,k}(n,u) h_\delta((n-k\Delta s, u-r\Delta t))
\]

(3.20)

where \( \Delta t = \Delta d \cos \theta \). Again, as in 2-D reconstruction, it is more efficient to implement the convolution using the 2-D Fourier convolution theorem and the Fast Fourier Transform (FFT). The FFT is used to compute the 2-D discrete Fourier transform (DFT)\(^9\) of \( \mathcal{M}_{\delta,k} \), and then the transformed matrix is multiplied, element by element, by the product of the filter (equation (3.16)) and a low-pass window function, evaluated at the appropriate frequency. The angle \( \Psi \) in equation (3.16) must be replaced by \( \psi \), the maximum \( \theta \) for which all projections are complete. An inverse two-dimensional DFT then yields the filtered matrix \( \mathcal{M}_{\delta,k}^F(k,r) \), i.e. a sampled version of the continuous 2-D function \( p_F(s,t,\theta) \) which appears in equation (3.14). This filtering procedure is then repeated for all projections \( \kappa = 0..M-1 \) and for all ring differences \( \delta = -\delta_{\text{max}}..\delta_{\text{max}} \), where \( \delta_{\text{max}} \) is the ring index difference for \( |\theta| = \psi \).

**Step 2: Backprojection:**

The discrete version of three-dimensional backprojection, equation (3.14), is obtained by first making a change of variables. This is necessary because, in practice, the sampled variables are not \( t \) and \( \theta \) but the axial coordinates of the detector rings \( z \), which are related to \( t \) and \( \theta \) by equation (3.18). Suppose that the \( z \) are discrete samples of some continuous axial coordinate \( \zeta \), and that the values of this coordinate at the two intersection points of an LOR with the cylindrical detector array are \( \zeta_a \) and \( \zeta_b \). The filtered parallel projection data \( \mathcal{F}_z \zeta_a \zeta_b (s,\phi) \) can be written as a filtered, continuously-sampled set of sinograms \( \mathcal{A}_{\zeta_a \zeta_b}^F(s,\phi) \), where the discrete dependence on ring difference has been replaced by an explicit dependence on the continuous axial coordinate \( \zeta \). Replacing \( t \) and \( \theta \) by \( \zeta_a \) and \( \zeta_b \), equation (3.14) becomes:
\[ f_r(x,y,z) = \int_0^\pi \int_{-\eta}^\eta d\phi \, d(\zeta_a - \zeta_b) \, A_{\zeta_a \zeta_b}^F(s,\phi) \, J(s,\zeta_a,\zeta_b) \]  
(3.21)

where \( s = x \cos\phi + y \sin\phi \)

\[ \zeta_a + \zeta_b = 2z + (x \sin\phi - y \cos\phi) \frac{(\zeta_a - \zeta_b)}{\sqrt{(R_d^2 - s^2)}} \]  
(3.22)

and

\[ \eta = 2 \sqrt{R_d^2 - s^2} \tan\psi \]  
(3.23)

\[ J(s, \zeta_a, \zeta_b) = \frac{4(R_d^2 - s^2)}{4(R_d^2 - s^2) + (\zeta_a - \zeta_b)^2}^{3/2} \]

Finally, equation (3.21) can be discretized using standard quadrature and interpolation techniques with the variables \( s, \phi, \zeta_a, \zeta_b \) replaced by their sampled equivalents \( s_k, \phi_k, z_i, z_j \), respectively, and with \( A_{\zeta_a \zeta_b}^F(s,\phi) \) replaced by the sampled version \( A_{\zeta_a \zeta_b}^F(k,r) \).

In practice, the implementation of equation (3.21) is particularly simple when both the field of view and the axial length of the tomograph are small compared with \( R_d \). With this approximation, both \( \eta \) and the Jacobian of the transformation \( J(s,\zeta_a,\zeta_b) \) reduce to constants, and the integrals can be interchanged. In particular, \( \eta \) becomes:

\[ \eta = 2R_d \tan\psi = \delta_{\text{max}} \Delta d \]  
(3.24)

with \( \delta_{\text{max}} \leq n-1 \).

Equation (3.22) can be interpreted as follows: For each voxel \((x,y,z)\), and with \( \phi \) and \((\zeta_a - \zeta_b)\) given, equation (3.22) is used to compute the corresponding values of \( s \) and \( \zeta_a \) such that the straight line defined by \((s,\phi,\zeta_a,\zeta_b)\) passes through \((x,y,z)\). If this line does not physically intersect the tomograph, the value of \( \zeta_a \) does not lie within the allowed axial range \([0, n\Delta d]\) and hence the corresponding sinogram has not been measured. However, such a situation would not arise once the projections have been completed by the forward projection procedure and the axial range extended to include all possible values of \( \zeta_a \). The forward projection procedure is implemented by discretizing equation (3.12)\(^{91,92}\).

A simple illustration of three-dimensional reconstruction, and particularly the effects of shift variance, is shown in fig.15. The image is that of a cylinder containing a uniform distribution of tracer. The cylinder extends for the full axial length of the tomograph and should ideally appear as a uniform disk with the same intensity in each of the 15 slices imaged by the tomograph. Fig.15a shows the 15 slices reconstructed from 64 measured sinograms without correcting for shift variance, i.e. without completing the partially-measured projections. The effect is to cause an apparent axial decrease in tracer concentration from the central slice out towards the edge slices. In fig.15b, shift invariance has been re-established by forward projecting an additional 56 sinograms which are included in the reconstruction. The cylinder now appears comparable in all slices, and represents an accurate image of the original tracer distribution.
Figure 15: A uniform 20 cm diameter cylinder reconstructed using a 3-D algorithm with (a) no compensation for axial shift variance, and (b) compensation for shift variance by forward projecting the unmeasured LORs.

3.5 Iterative reconstruction methods

The reconstruction algorithms presented in the previous sections are transform methods based on direct inversion of the X-ray transform, derived assuming continuous sampling, and then discretized for application to sampled data. However, it is possible to develop an
alternative approach which incorporates, from the beginning, the discrete nature of the measured data\textsuperscript{54,93,94}. Thus, the discrete LORs which sample the continuous projection data \( p(s, \phi) \) are written as components of a one-dimensional vector of measurements \( n^*(d) \), \( d = 1 \ldots D \), where, for \( M \) projections and \( N \) samples per projection, \( D = N \times M \). Similarly, the true tracer concentration \( f(x, y) \), discretized as an image of \( L \times L \) pixels, is represented by a one-dimensional vector \( \lambda(b) \), \( b = 1 \ldots B \), \( B = L \times L \). These two vectors are related by:

\[
n^*(d) = \sum_{b=1}^{B} p_i(b, d) \lambda(b)
\]  

(3.25)

where \( p_i(b, d) \) is the probability that photons from a positron annihilation in voxel \( b \) are detected in LOR \( d \). Equation (3.25) represents a set of linear equations which can, in principle, be solved for \( \lambda(b) \), given the probabilities \( p_i(b, d) \) and the measurements \( n^*(d) \). An important advantage of this discrete approach is the potential to explicitly incorporate the physics of the detection process into \( p_i(b, d) \), rather than assume, as with transform methods, that the LORs are close approximations to line integrals which adequately sample the tracer distribution. In particular, effects such as scatter, attenuation, shift-variance and LOR profiles can be incorporated directly into \( p_i(b, d) \), although in practice the inclusion of these effects is not easy. Attenuation, for example, involves additional patient-dependent measurements (a transmission scan), requiring the recalculation of \( p_i(b, d) \) for each reconstruction. Thus, the \( p_i(b, d) \) are usually taken as simply the intersection length of the LOR \( d \) with the voxel \( b \), normalised such that the total probability of detection of a photon emitted from voxel \( b \) is unity:

\[
\sum_{d=1}^{D} p_i(b, d) = 1
\]  

(3.26)

The solution of the set of linear equations (3.25) is not straightforward, both because of the large number of equations involved (e.g. even for a single slice, some 50 000 equations in 16 000 unknowns must be solved), and because the equations are ill-conditioned. As it would be impractical to solve such sets of equations by direct inversion, effort has been devoted to the investigation of alternative approaches which, in this context, are generally referred to as Series Expansion Techniques\textsuperscript{94,95}.

A common feature of these alternative techniques is that they attempt to approach an acceptable solution by a series of successive estimates to \( \lambda(b) \). Starting with an initial estimate, or 'guess' \( \lambda_0(b) \), which may, for example, be a uniform image, the algorithms attempt progressively to improve upon this estimate, at the same time ensuring that the estimate satisfies known properties of tracer distributions. At each step, the current image estimate \( \lambda_k(b) \) is used to generate, by a forward projection procedure, a vector of projection data \( n_k(d) \) which is compared with the measurements \( n^*(d) \). The aim is then to make the \( n_k(d) \) consistent with \( n^*(d) \) by modifying the image estimate \( \lambda_k(b) \) to produce a new estimate \( \lambda_{k+1}(b) \). Precise details of the modification procedures differ from one algorithm to another, and a good review of the various possibilities has been given by Herman\textsuperscript{54}. The important point is that all such algorithms iterate progressively towards a solution rather than arriving at one by direct inversion of the image transform.

The possibility to incorporate prior information, or constraints, into iterative methods is a significant advantage\textsuperscript{96-99}. Such prior information is particularly useful in guiding the algorithm towards an acceptable solution when the usual constraints imposed by the data are weak, for example in the case of low signal to noise ratio or poor sampling. One such constraint is positivity, which ensures that, as they are supposed to represent tracer concentration,
successive $\lambda_k(b)$ are always positive. Note that the solution from the filtered backprojection algorithm is not guaranteed to be positive, negative values in the reconstructed image originating from negative values in the kernel $h(s)$.

It is evident that the necessity to compute a succession of image estimates must considerably increase the reconstruction time, particularly as each estimate, or iteration, takes approximately the same time as a single filtered backprojection, and typically 10 to 100 iterations may be required to reach an acceptable image. Transform methods have, therefore, been preferred for the reconstruction of PET data, even though in situations such as low signal to noise, the use of iterative techniques can indeed improve the quality of the image.

In addition to the large amounts of computer time required to reach a solution, a further difficulty of iterative methods is to decide when an acceptable solution has been reached. The choice of the appropriate number of iterations is analogous to selecting the appropriate cut-off frequency $v_c$ for the window function in equation (3.3): too few iterations and the resulting image is over-smooth, while too many iterations and the image shows excessive noise amplification. The establishment of objective and efficient stopping rules has proved to be difficult.\textsuperscript{100,101}

A valid criticism, and a further incentive to pursue an iterative approach, is that transform methods do not really take into account the statistical nature of the projection data. Data acquired by counting photons are essentially random variates of a process that obeys Poisson, or counting, statistics. Thus, in particular, the sampled projection data $p(k\Delta s, \pi j/M)$, $k = -N/2+1,...,N/2$, $j = 0..M-1$, are a set of Poisson random variables with rates related to the values of the corresponding line integrals. When a large number of annihilation photons are acquired for an LOR, the measurement should be reasonably close to the actual mean value for the line integral, a situation which is implicitly assumed by transform methods. However, when few photons are acquired, the measurement may be a very poor estimate of the line integral.

In order to incorporate explicitly the Poisson nature of the data, Shepp and Vardi\textsuperscript{102} proposed a model for PET based on equation (3.25). In their model, $\lambda(b)$ is the true tracer concentration and the projection of $\lambda(b)$ into LOR $d$ must be the contribution of that voxel to the expected mean value in LOR $d$. When summed over all voxels, the result, denoted by $\lambda^*(d)$, is the expected mean value for LOR $d$, i.e.

$$\lambda^*(d) = \sum_{b=1}^{B} \lambda(b) \ p_t(b,d) \quad d=1..D$$  \hspace{1cm} (3.27)

The corresponding measurements $n^*(d)$ are therefore Poisson random variables with means $\lambda^*(d)$. The aim of the reconstruction algorithm is then to maximise the likelihood that a set of Poisson processes with means $\lambda^*(d)$ gives rise to projection data $n^*(d)$. Since the $p_t(b,d)$ are given, this procedure will find the image $\lambda(b)$ that is the most likely to have generated the projection data $n^*(d)$, where the likelihood is defined as:

$$L(\lambda) = \prod_{d=1}^{D} \frac{\lambda^*(d)^{n^*(d)}}{n^*(d)!}$$  \hspace{1cm} (3.28)

In order to maximise this likelihood, Shepp and Vardi used the expectation–maximisation (EM) algorithm.\textsuperscript{103} They showed that an image maximising the expression in equation (3.28) can be computed from the E-M iteration:
\[
\lambda_{k+1}(b) = \lambda_k(b) \sum_{d=1}^{D} \frac{n^*(d)}{\sum_{b'} p_r(b',d) \lambda_k(b')} 
\] 

(3.29)

with the \( p_r(b,d) \) normalised as in equation (3.26).

With increasing iterations it was found that, as expected, initially the algorithm approached a reasonable reconstruction, especially in low-count situations where the true mean value for an LOR might vary considerably from the measurement. However, further investigation showed that if the iterations continued, the likelihood increased, whereas the image began to degenerate, becoming increasingly noisy. It is, therefore, important to terminate the reconstruction before this degeneration begins, even though the likelihood function may not be a maximum. The origin of this effect lies in the fact that the measurements \( n^*(d) \) are Poisson random variables with noise, and that insistence on an exact fit to the data will result in an image dominated by noise in a way similar to that resulting from the choice of \( \nu_c \) too large. The iteration procedure should then terminate when the fit to the data is as close as can be expected, within the noise level. However, as mentioned above, the choice of the stopping point is difficult, and research continues in order to define appropriate rules for use with real data.

The E-M algorithm is an attractive approach since both the positivity of successive image estimates and correct normalisation are guaranteed, and it has been shown to yield images superior to those of filtered backprojection when the signal to noise ratio is small. This, and other similar algorithms, are therefore the subject of continuing research in an attempt to limit the effects of noise amplification and to accelerate convergence. For PET, the major obstacle at the present time to the use of iterative methods is the inevitable computational burden. Recently, attempts have been made to design and construct special hardware for iterative reconstruction\textsuperscript{104}, but at the present time the use of iterative techniques for three-dimensional reconstruction is still rather impractical.
REFERENCES


4. Derenzo S.E., Precision measurement of annihilation point spread distributions for medically important positron emitters, in *Proc. 5th conference on positron annihilation*, Sendai, Japan, 819-824, 1979.


36


37


