4 Image Reconstruction Algorithms in PET*

Michel Defrise, Paul E Kinahan and Christian J Michel

Introduction

This chapter describes the 2D and 3D image reconstruction algorithms used in PET and the most important evolutions in the last ten years: the introduction of 3D acquisition and reconstruction and the increasing role of iterative algorithms. As will be seen, iterative algorithms improve image quality by allowing more accurate modeling of the data acquisition. This model includes the detection, the photon transport in the tissues, and the statistical distribution of the acquired data, i.e. the noise properties. The popularity of iterative methods dates back to the seminal paper of Shepp and Vardi on the maximum-likelihood (ML) estimation of the tracer distribution. Practical implementation of this algorithm has long been hindered by the size of the collected data, which has increased more rapidly than the speed of computers. Thanks to the introduction of fast iterative algorithms in the nineties, such as the popular Ordered Subset Expectation Maximization (OSEM) algorithm, iterative reconstruction has become practical. Reconstruction time with iterative methods nevertheless remains an issue for very large 3D data sets, especially when multiple data sets are acquired in whole-body or dynamic studies. Speed, however, is not the only reason why filtered-backprojection (FBP) remains important: analytic algorithms are linear and thereby allow an easier control of the spatial resolution and noise correlations in the reconstruction, a control which is mandatory for quantitative data analysis.

The chapter is organized as follows. First, the organization of the data acquired in 2D mode is described, and the reconstruction problem is defined. The third section reviews the classical analytic reconstruction of 2D tomographic data and describes the FBP method, which remains a workhorse of tomography. Iterative reconstruction is presented in the following section, where the accent is set on the key concepts and on their practical implications. Owing to the wide variety of iterative methods, only the popular ML-EM and OSEM methods are described in detail, though this does not entail any claim that these algorithms are optimal. The last sections concern the reconstruction of data acquired in 3D mode. Three-dimensional FBP is described, as well as fast *rebinning* algorithms, which reduce the redundant 3D data set to *synthetic* 2D data that can be processed by analytic or iterative 2D algorithms. *Hybrid algorithms* combining rebinning with a 2D iterative algorithm are introduced, and the chapter concludes with a discussion of the practical aspects of fully 3D iterative reconstruction.

Presented here as a separate chapter, image reconstruction cannot be understood independently of the other steps of the data-processing chain, including data acquisition, data corrections (described in chapters 2, 3, 5), as well as the quantitative or qualitative analysis of the reconstructed images. The variety of algorithms for PET reconstruction arises from the fact that there is no such thing as an *optimal* reconstruction algorithm. Different algorithms may be preferred depending on factors such as the signal-to-noise ratio (number of collected coincident events in the emission and transmission scans), the static or dynamic character of the tracer distribution, the practical constraints on the processing time, and, most importantly, the specific clinical *task* for which the image is reconstructed. It is

* Figures 4.1–4.11 are reproduced from Valk PE, Bailey DL, Townsend DW, Maisey MN. Positron Emission Tomography: Basic Science and Clinical Practice. Springer-Verlag London Ltd 2003, 91–114.

important to keep this observation in mind when discussing reconstruction as an isolated topic.

2D Data Organization

Line of Responses

A PET scanner counts coincident events between pairs of detectors. The straight line connecting the centers of two detectors is called a *line of response* (LOR). Unscattered photon pairs recorded for a specific LOR arise from annihilation events located within a thin volume centered around the LOR. This volume typically has the shape of an elongated parallelipiped and is referred to as a *tube of response*.

To each pair of detectors d_a, d_b is associated an LOR $\mathcal{L}_{d_a, d_b}$ and a sensitivity function $\psi_{d_a, d_b}(\vec{r} = (x, y, z))$ such that the number of coincident events detected is a Poisson variable with a mean value

$$< p_{d_a, d_b} > = \tau \int_{\text{FOV}} d\vec{r} f(\vec{r}) \psi_{d_a, d_b}(\vec{r}) \tag{1}$$

where τ is the acquisition time and $f(\vec{r})$ denotes the tracer concentration. We assume that the tracer concentration is stationary and that $f(\vec{r}) = 0$ when $\sqrt{}(x^2 + y^2) > R_F$, where R_F denotes the radius of the field-of-view (FOV). The reconstruction problem consists of recovering $f(\vec{r})$ from the acquired data p_{d_a, d_b}, $\{d_a, d_b\} = 1 \cdots, N_{\text{LOR}}$, where N_{LOR}, the number of detector pairs in coincidence, can exceed 10^9 with modern scanners.

The model defined by Eq. (1) is *linear* and hence implies that nonlinear effects due to random coincidences and dead time be pre-corrected. In the absence of photon scattering in the tissues, the sensitivity function vanishes outside the tube of response centered on the LOR. In such a case, the accuracy of the spatial localization of the annihilation events is determined by the size of the tube of response, which in turn depends on the geometrical size of the detectors and on other factors such as the photon scattering in the detectors, or the variable depth of interaction of the gamma rays within the crystal (*parallax error, figure 2.26*).

We have so far considered a scanner comprising multiple small detectors. Scanners based on large-area, position-sensitive detectors such as Anger cameras can be described similarly if viewed as consisting of a large number of very small virtual detectors.

Analytic reconstruction algorithms assume that the data have been pre-corrected for various effects such as randoms, scatter and attenuation. In addition, these algorithms model each tube of response as a mathematical line joining the center of the front face of the two crystals[1]. This means that the sensitivity function $\psi_{d_a, d_b}(\vec{r})$ is zero except when $\vec{r} \in \mathcal{L}_{d_a, d_b}$. With this approximation, the data are modeled as *line integrals* of the tracer distribution:

$$\langle p_{d_a, d_b} \rangle = \int_{\mathcal{L}_{d_a, d_b}} d\vec{r} f(\vec{r}) \tag{2}$$

Sinogram Data and Sampling

The natural parameterization of PET data uses the indices (d_a, d_b) of the two detectors in coincidence, as in Eq. (1). However, there are several reasons to modify this parameterization:

- The natural parameterization is often poorly adapted to analytic algorithms. This is why raw data are usually interpolated into an alternative *sinogram* parameterization described below.
- The number of recorded coincidences N_{events} in a given scan may be too small to take full advantage of the nominal spatial resolution of the scanner. In such a case, undersampling by grouping neighboring LORs reduces the data storage requirements and the reconstruction time without significantly affecting the reconstructed spatial resolution, which is primarily limited by the low count density.

Another approach to reduce data storage and processing time when N_{LOR} » N_{events} consists of recording the coordinates (d_a, d_b) of each coincident event in a sequential data stream called a *list-mode* data set. Additional information such as the time or the energy of each detected photon can also be stored. In contrast to undersampling, list-mode acquisition does not compromise the accuracy of the spatial localization of each event. But the fact remains that the number of measured coincidences may be too low to exploit the full resolution of the scanner.

Let us define the standard parameterization of 2D PET data into *sinograms*. Consider a transaxial section $z = z_0$ measured using a ring of detectors. Figure 4.1 defines the variables s and ϕ used to parameterize a straight line (an LOR) with respect to a Cartesian coordinate system (x, y) in the plane. The *radial* variable s is

[1] When the depth of interaction is accounted for, LORs are defined by connecting photon interaction points projected on the long axis of the crystals [1].

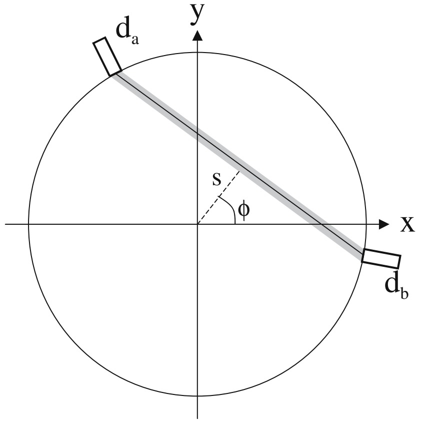

Figure 4.1. Schematic representation of a ring scanner. A tube of response between two detectors d_a and d_b is represented in grey with the corresponding LOR, which connects the center of the front face of the two detectors. The sinogram variables s and ϕ define the location and orientation of the LOR.

the signed distance between the LOR and the center of the coordinate system (usually the center of the detector ring). The *angular variable* ϕ specifies the orientation of the LOR. Line integrals of the tracer distribution are then defined as

$$p(s,\phi,z_0)=\int_{-\infty}^{\infty} dt\, f(x=s\cos\phi-t\sin\phi,\; y=s\sin\phi+t\cos\phi,\; z=z_0) \qquad (3)$$

where t, the integration variable, is the coordinate along the line. In the presentation of the 2D reconstruction problem below, we will omit the z arguments in the functions p and f.

The next section describes how a function $f(x, y)$ can be reconstructed from its line integrals measured for $|s| < R_F$ and $0 \leq \phi < \pi$. The mathematical operator mapping a function $f(x, y)$ onto its line integrals $p(s, \phi)$ is called the *x-ray transform*[2], and this operator will be denoted X, so that $p(s, \phi) = (Xf)(s, \phi)$. The function $p(s, \phi)$ is referred to as a *sinogram*, and the variables (s, ϕ) are called sinogram variables. This name was coined in 1975 by the Swedish scientist Paul Edholm because the set of LORs containing a fixed point (x_0, y_0) are located along a sinusoid $s = x_0 \cos \phi + y_0 \sin \phi$ in the (s, ϕ) plane, as can be seen from Eq. (3). For a fixed angle $\phi = \phi_0$, the set of parallel line integrals $p(s, \phi_0)$ is a *1D parallel projection* of f.

At the line integral approximation, and after data pre-correction, the PET data provide estimates of the x-ray transform for all LORs connecting two detectors, i.e., $p_{d_a,d_b} \simeq p(s, \phi)$, where the parameters (s, ϕ) correspond to the radial position and angle of $\mathcal{L}_{d_a,d_b}$. Thus, the geometrical arrangement of discrete detectors in a scanner determines a set of samples (s, ϕ) in sinogram space. The most common arrangement is a ring scanner: an even number N_d of detectors uniformly spaced along a circle of radius $R_d > R_F$[3]. Each detector, in coincidence with an arc of detectors on the opposite side of the ring, defines a *fan* of LORs (figure 3.6), and the corresponding sampling of the sinogram is:

$$\begin{aligned} s_{j,k} &= R_d\cos((2k-j)\pi/N_d) & k&=0,\ldots,N_d-1 \\ \phi_j &= j\pi/N_d & j&=0,\ldots,N_d-1 \end{aligned} \qquad (4)$$

where the pair of indices j, k corresponds to the coincidences between the two detectors with indices $d_a = j - k$ and $d_b = k$. Due to the curvature of the ring, each parallel projection j is sampled non-uniformly in the radial variable, with a sampling distance $\Delta s \simeq 2\pi R_d/N_d$ near the center of the FOV (i.e. for $s \simeq 0$). The radial samples of two adjacent parallel projections j and $j + 1$ are shifted by approximately $\Delta s/2$, as can be seen by shifting only one end of a LOR (Fig. 4.2).

For practical and historical reasons, it is customary in PET to reorganize the data on a rectangular sampling grid

$$\begin{aligned} s_k &= k\Delta s & k&=-N_s,\ldots,N_s \\ \phi_j &= j\Delta\phi & j&=0,\ldots N_\phi-1 \end{aligned} \qquad (5)$$

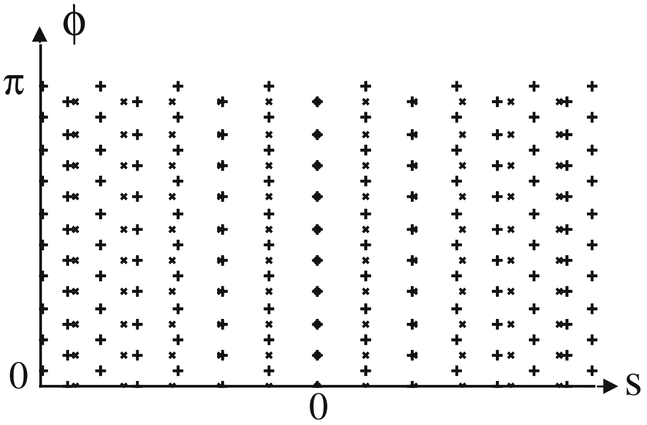

Figure 4.2. Representation of the sinogram sampling for a ring scanner with 20 detectors. The interleaved pattern provided by the LORs connecting detector pairs is shown by +'s. Note the decrease of the radial sampling distance at large values of s, which is exaggerated here because the plot extends to 90% of the ring radius. PET acquisition systems reorganize these data into the rectangular sampling pattern (see equation (5)) shown by ×'s.

[2] In 2D, the x-ray transform coincides with the Radon transform, see [2].

[3] If the depth of interaction is not measured, an effective value of R_d is used that accounts for the mean penetration of the 511 keV gamma rays into the crystal.

with $\Delta\phi = 2\pi/N_d$, $N_\phi = N_d/2$, and a uniform radial sampling interval $\Delta s = R_d\pi/N_d$ equal to half the spacing between adjacent detectors in the ring. The *parallel-beam sampling* defined by Eq. (5) will be used in the rest of the chapter. In this scheme the line defined by a sample (j, k) no longer coincides with a measured LOR connecting two detectors. The reorganization into parallel-beam data therefore requires an interpolation (usually linear interpolation) to redistribute the counts on the rectangular sampling grid (Eq. (5)). This interpolation entails a loss of resolution, which is usually negligible owing to the relatively low SNR in PET[4]. In addition, the geometry of some scanners is not circular, but hexagonal or octagonal. Resampling is then needed anyway if standard analytic algorithms are to be used.

When the average number of detected coincidences per sinogram sample is small, undersampling is often applied to reduce the storage and computing requirements. Angular undersampling (increasing $\Delta\phi$) is called *transaxial mashing* in the PET jargon. The mashing factor defined by $m = \Delta\phi N_d/(2\pi)$ is usually an integer so that undersampling simply amounts to summing groups of m consecutive rows (j's) in the sinogram. Angular undersampling results in a loss of resolution, which is smallest at the center of the FOV and maximum at its edge. Therefore, the maximum allowed mashing factor depends not only on the SNR but also on the radius R_F of the reconstructed FOV: for a fixed SNR, we can allow more mashing for a brain scan than for a whole-body study. Radial undersampling (increasing Δs) tends to generate more severe artifacts, and is rarely used. A rule of thumb to match the radial and angular sampling is the relation $\Delta\phi \simeq \Delta s/R_F$, which is derived using Shannon's sampling theory [2].

Multi-slice 2D Data

So far we have discussed data sampling for a single ring scanner located in the plane $z = z_0$. Multi-ring scanners are stacks of N_R rings of detectors spaced axially by Δz and indexed as $r = 0, \cdots, N_R - 1$ [3]. The coincidences between two detectors belonging to the same ring r are organized in a *direct* sinogram $p(s, \phi, z = r\Delta z)$ as described in the previous section. This is the sinogram of the function $f(x, y, z = r\Delta z)$ (Fig. 4.3). Multi-ring scanners also collect coincidences between detectors located in a few adjacent rings, i.e. between one detector in some ring r and another detector in one of the rings $r + d$, with $d = -d_{2D,\max}, \cdots, d_{2D,\max}$. The LORs connecting such detector pairs are not transaxial, but the maximum *ring difference* $d_{2D,\max}$ is chosen to be small enough (typically 5) that the angle between these oblique LORs and the transaxial planes ($\vartheta \simeq d_{2D,\max}\,\Delta z/(2R_d)$) can be neglected[5].

Consider first the LORs between detectors in adjacent rings r and $r + 1$. These data are assembled in a 2D sinogram $p(s, \phi, z = (r + 1/2)\Delta z)$ and used to reconstruct a transaxial slice that is approximated as lying midway, axially, between the two detector rings. Each sample in this *cross-plane* sinogram is the average of two LORs: on the one hand the LOR connecting a detector d_a in ring r to a detector d_b in ring $r + 1$, and on the other hand the LOR connecting detectors d_a in ring $r + 1$ and d_b in ring r. Indeed, these two LORs coincide if we neglect the small angles $\pm\vartheta$ they form with the transaxial plane. One effect of the introduction of the cross-plane sinograms is to increase the sampling rate in the axial direction so that instead of reconstructing N_r image planes of thickness Δz, we end up with $2N_r - 1$ image planes separated by $\Delta z/2$.

More generally, the LORs between rings $r - j$ and $r + j$, with $j = 0, 1, 2, .. \leq d_{2D,\max}/2$ are added to form the direct sinogram of slice $z = r\Delta z$, and the LORs between rings $r - j + 1$ and $r + j$, with $j = 0, 1, 2, .. \leq (d_{2D,\max} + 1)/2$ are added to form the cross-plane sinogram of slice $z = (r + 1/2)\Delta z$. There are an odd number of ring pairs contributing to the direct plane sinograms and an even number of ring pairs contributing to the cross-plane sinograms (Fig. 4.3). For small values of $d_{2D,\max}$,

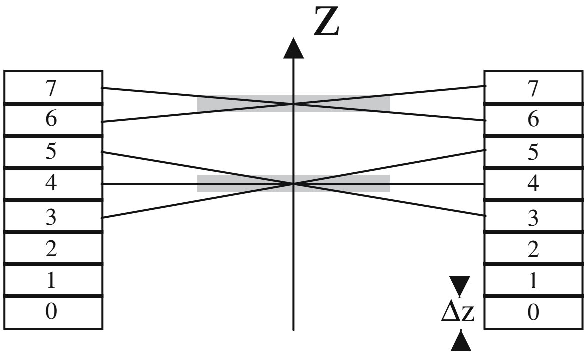

Figure 4.3. Longitudinal view of a multi-ring scanner with $N_r = 8$ rings, operated in 2D mode, illustrating the formation of sinograms for two transaxial slices (in grey), with $d_{2D,\max} = 2$. The sinogram for the cross slice at $z = 13\triangle z/2$ (top) is obtained by averaging the coincidences between two rings pairs $(r_a, r_b) = (6, 7)$ and $(7, 6)$. The sinogram for the direct slice at $z = 8\triangle z/2$ (bottom) is obtained by averaging the coincidences between three rings pairs $(r_a, r_b) = (3, 5), (5, 3)$ and $(4, 4)$.

[4] Parallel-beam resampling is used by some CT scanners despite more severe requirements in terms of spatial resolution.

[5] For $d = \pm 1$ this approximation is of the same order as when resampling the sinogram to parallel beam.

there is thus a significant difference in the number of LORs contributing to the different types of sinograms, and therefore also a difference in the corresponding SNRs (see also figure 3.9). As $d_{2\mathrm{D,max}}$ increases, the SNR of both types of sinograms increases and the differences diminish; however, there is a degradation in the image resolution as we will see later in the single-slice rebinning algorithm. In practice, the value of $d_{2\mathrm{D,max}}$ is chosen to balance these trade-offs, with typical values ranging from 3 to 11.

Analytic 2D Reconstruction

Properties of the X-ray Transform

In this section, we solve the inverse 2D x-ray transform. A closed-form solution of the integral equation, Eq. (3) is first derived assuming a continuous sampling of the sinogram variables over $(s, \phi) \in [-R_F, R_F] \times [0, \pi]$. An approximation to this exact solution will then be written in terms of the discrete data samples (defined by Eq. (5)), leading to the standard *filtered-backprojection* algorithm (FBP). We refer for this section to the comprehensive books by Natterer [2, 4], Kak and Slaney [5], Barrett and Swindell [6], and Barrett and Myers [7].

First, two properties of Eq. (3) should be stressed:

- The problem is *invariant for translations* in the sense that the x-ray transform of a translated image $f_t(x, y) = f(x - t_x, y - t_y)$ is $(Xf_t)(s, \phi) = (Xf)(s - t_x \cos\phi - t_y \sin\phi, \phi)$. Translating the image simply shifts each sinogram row.
- The problem is *invariant for rotations* in the sense that the x-ray transform of a rotated image $f_\theta(x, y) = f(x \cos\theta - y \sin\theta, x \sin\theta + y \cos\theta)$ is $(Xf_\theta)(s, \phi) = (Xf)(s, \phi + \theta)$.

These two invariances, and also the algorithms described in the next sections, are valid only when the scanner measures *all* line integrals crossing the support of the image (the disc of radius R_F), so that the sinogram is sampled over the complete range $(s, \phi) \in [-R_F, R_F] \times [0, \pi]$. When this condition is not satisfied, the problem is called an *incomplete data problem* (among many references, see [2] Ch. VI, [4, 8, 9]). This happens in particular with hexagonal or octagonal scanners such as the Siemens/CPS HHRT, where the gaps between adjacent flat panel detectors cause unmeasured diagonal bands in the sinogram [10]. Before applying the FBP algorithm presented below, the incompletely measured sinograms must first be completed by estimating the missing LOR data. When the gaps in the sinogram are not too wide, simple interpolation can be used, but more sophisticated techniques have been proposed [11, 12]. An alternative is to apply iterative reconstruction techniques, which are less sensitive to the specific geometry. We note, however, that the use of iterative methods does not provide a solution for the missing data problem. Rather it simplifies the introduction of prior knowledge which can partially compensate for the missing data.

The Cornerstone of Tomographic Reconstruction: The Central Section Theorem

Tomographic reconstruction relies on Fourier analysis. Recall that the Fourier transform of a function $f(x, y)$ is defined by

$$\begin{aligned}(\mathcal{F}f)(\nu_x, \nu_y) &= F(\nu_x, \nu_y)\\ &= \int_{\mathbb{R}^2} dx\, dy\, f(x, y) \exp(-2\pi i(x\nu_x + y\nu_y))\end{aligned} \qquad (6)$$

and is inverted by changing the sign of the argument of the complex exponential

$$\begin{aligned}(\mathcal{F}^{-1}F)(x, y) &= f(x, y)\\ &= \int_{\mathbb{R}^2} d\nu_x\, d\nu_y\; F(\nu_x, \nu_y) \exp(2\pi i(x\nu_x + y\nu_y))\end{aligned} \qquad (7)$$

We use ν_x and ν_y to denote the frequencies associated to x and y respectively, and denote the Fourier transform of a function, e.g., f, by the corresponding upper case character, e.g., F. These definitions are extended in the obvious way to N dimensions.

A key property of the Fourier transform is the *convolution theorem*, which states that the Fourier transform of the convolution of two functions f and h,

$$(f * h)(x, y) = \int_{\mathbb{R}^2} dx' dy' f(x', y') h(x - x', y - y') \qquad (8)$$

is the product of their Fourier transforms:

$$(\mathcal{F}(f * h))(\nu_x, \nu_y) = (\mathcal{F}f)(\nu_x, \nu_y) \cdot (\mathcal{F}h)(\nu_x, \nu_y) \qquad (9)$$

In signal- or image-processing terms, convolving f with h amounts to *filtering* f with a shift-invariant (i.e. invariant for translations) point spread function h. The convolution theorem simplifies convolution by reducing it to a product in frequency space. In general, the Fourier transform is useful for all problems that are invariant for translation, and therefore also for tomographic reconstruction as will now be shown.

The *central section* theorem, also called the projection slice theorem, states that the 1D Fourier transform of the x-ray transform Xf with respect to the radial variable s is related to the 2D Fourier transform of the image f by

$$P(\nu,\phi) = \mathrm{F}(\nu_x = \nu\cos\phi, \nu_y = \nu\sin\phi) \tag{10}$$

where

$$P(\nu,\phi) = (\mathcal{F}p)(\nu,\phi) = \int_{\mathbb{R}} ds\; p(s,\phi)\exp(-2\pi i s\nu) \tag{11}$$

and ν is the frequency associated to the radial variable s.

This theorem is easily proven by replacing the x-ray transform $p(s, \phi) = (Xf)(s, \phi)$ in the right hand side of Eq. (11) by its definition (Eq. (3)) as a line integral of f. Thus the 1D Fourier transform of a parallel projection of an image f at an angle ϕ determines the 2D Fourier transform of that image along the radial line in frequency plane (ν_x, ν_y) that forms an angle ϕ with the ν_x axis. The implication for reconstruction is the following: if we measure all projections $\phi \in [0, \pi]$, the radial line sweeps over the whole frequency plane and thereby allows the recovery of $F(\nu_x, \nu_y)$ for all frequencies $(\nu_x, \nu_y) \in \mathbb{R}^2$. The image f can then be reconstructed by inverse 2D Fourier transform (Eq. (7)).

The discrete implementation of the inversion formula combining Eqs. (11), (10) and (7) is referred to as the *direct Fourier reconstruction*. This algorithm is numerically efficient because the discretized 2D Fourier transform (Eq. (7)) can be calculated with the FFT algorithm. The 2D FFT requires as input the values of F on a square grid $(\nu_x = k\Delta, \nu_y = l\Delta)$, $(k, l) \in Z^2$, which does not coincide with the polar grid of samples provided by the data (see the right hand side of Eq. 10). Direct Fourier reconstruction therefore involves a 2D interpolation to map the polar grid onto the square grid. This interpolation is often based on gridding techniques similar to those used for magnetic resonance imaging [13, 14, 15].

The Filtered Backprojection Algorithm

The FBP algorithm is the standard algorithm of tomography. It is equivalent to the direct Fourier reconstruction in the limit of continuous sampling, but its discrete implementation differs.

The FBP inversion explicitly combines Eqs. (11), (10) and (7). Straight-forward manipulations involving changing from Cartesian (ν_x, ν_y) to polar (ν, ϕ) coordinates lead to a two-step inversion formula (Fig. 4.4):

$$f(x,y) = (X^* p^F)(x,y) = \int_0^\pi d\phi\, p^F(s = x\cos\phi + y\sin\phi, \phi) \tag{12}$$

where the *filtered projections* are

$$p^F(s,\phi) = \int_{-R_F}^{R_F} ds' p(s',\phi) h(s-s') \tag{13}$$

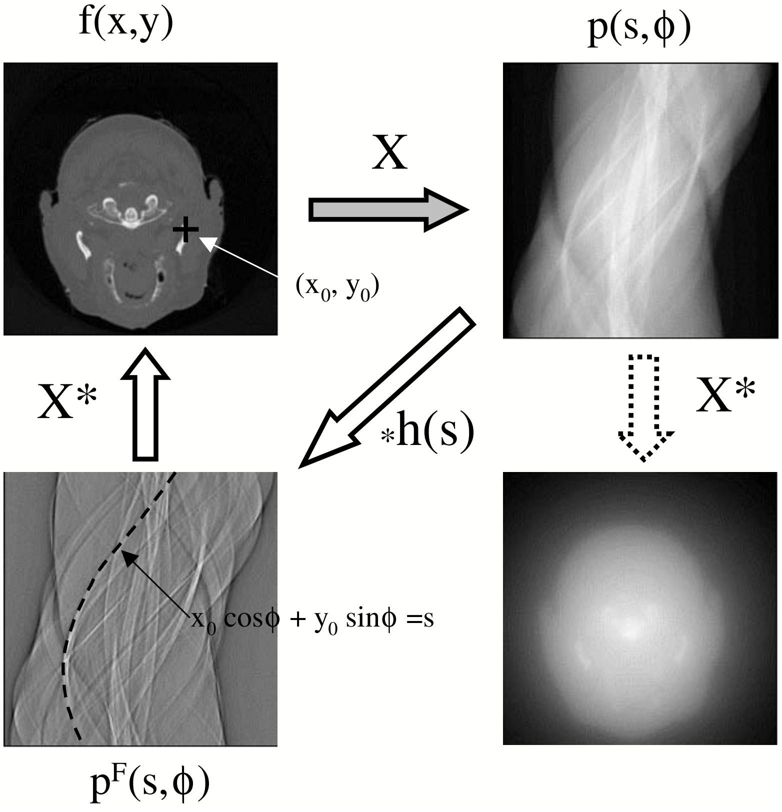

Figure 4.4. Illustration of 2D filtered backprojection. The top row shows a brain section and its sinogram $p = Xf$. The backprojection X^*p of the sinogram (bottom right) is the 2D convolution of f with the point spread function $1/\sqrt{(x^2 + y^2)}$ and illustrates the blurring effect of line integration. The filtered sinogram p^F obtained by 1D convolution with the ramp filter kernel has enhanced high frequencies, and when backprojected, yields the original image f, up to noise and discretization errors.

and the *ramp filter* kernel is defined as

$$h(s) = \int_{-\infty}^{\infty} d\nu\,|\nu|\,\exp(2\pi i s\nu) \qquad (14)$$

Three remarks are in order.

(i) The operator X^* mapping p^F onto f in Eq. (12) is called the *backprojection* and is the dual of the x-ray transform. Geometrically, $(X^*p^F)(x, y)$ is the sum of the filtered data p^F for all lines that contain the point (x, y).

(ii) The convolution (Eq. (13)) can be expressed using the convolution theorem as $P^F(\nu, \phi) = |\nu|\, P(\nu, \phi)$.

(iii) The integral (Eq. (14)) defines the kernel h as the inverse 1D Fourier transform of the ramp filter function $|\nu|$. This integral does not converge in the usual sense, and h is only defined as a generalized function (see chapter 2 in [7]).

Discrete Implementation of the FBP

The discrete implementation of Eqs. (12) and (13) using the measured samples of $p(s, \phi)$ described in the section on sinogram data and sampling, above (Eq. (5)), involves four approximations:

(i) The approximation of the kernel $h(s)$ by an *apodized* kernel

$$h_w(s) = \int_{-\infty}^{\infty} d\nu\,|\nu|\; w(\nu)\exp(2\pi i s\nu) \qquad (15)$$

where $w(\nu)$ is a low-pass filter which suppresses the high spatial frequencies, and will be discussed later in the section on the ill-posedness of the inverse X-ray transform.

(ii) The approximation of the convolution integral by a discrete quadrature. Usually standard trapezoidal quadrature is used:

$$p^F(k\Delta s,\phi_j) \simeq \Delta s \sum_{k'=-N_s}^{N_s} p(k'\Delta s,\phi_j)\,h_w((k-k')\Delta s) \qquad (16)$$
$$k = -N_s, \ldots, N_s$$

The calculation of this discrete convolution can be accelerated using the discrete Fourier transform (FFT) (see [16] section 13.1). In this case, some care is needed when defining the discrete filter: to avoid bias, this filter must be calculated as the FFT of the sampled convolution kernel $h_w(k\Delta s)$, $k = 0, \pm 1, \pm 2, \ldots$, and not by simply sampling the continuous filter function $|\nu| w(\nu)$.

(iii) The approximation of the backprojection by a discrete quadrature

$$f(x,y) \simeq \Delta\phi \sum_{j=0}^{N_\phi-1} p^F(s = x\cos\phi_j + y\sin\phi_j, \phi_j) \qquad (17)$$

for a set of image points (x, y) (usually a square pixel grid)[6].

(iv) The estimation of p^F $(s = x\cos\phi_j + y\sin\phi_j, \phi_j)$ in Eq. (17) from the available samples p^F $(k\Delta s, \phi_j)$. This is usually done using linear interpolation:

$$p^F(s,\phi_j) \simeq (k+1-\frac{s}{\Delta s})p^F(k\Delta s,\phi_j) + (\frac{s}{\Delta s}-k)p^F((k+1)\Delta s,\phi_j) \qquad (18)$$

where k is the integer index such that $k\Delta s \le s < (k+1)\Delta s$. Instead of linear interpolation some implementations apply a faster nearest-neighbor interpolation to filtered projections which have first been linearly interpolated on a finer grid (typically sampled at a rate $\Delta s/4$).

Remarkably, most FBP implementations only use simple tools of numerical analysis, such as linear interpolation and trapezoidal quadrature, despite many attempts to demonstrate the benefits of more sophisticated techniques.

The Ill-posedness of the Inverse X-ray Transform

Like many problems in applied physics, the inversion of the x-ray transform is an *ill-posed problem*: the solution f defined by Eqs. (11), (10) and (7) does not depend continuously on the data $p(s, \phi)$. Concretely, this means that an arbitrarily small perturbation of p due to measurement noise can cause an arbitrarily large error on the reconstructed image f. We refer to Bertero and Boccacci [20] and Barrett and Myers [7] for an introduction to the concept of ill-posedness and its implication in tomography. Intuitively, ill-posedness can be understood by noting that the ramp filter $|\nu|$ amplifies the high frequencies during the filtering step $P(\nu, \phi) \to P^F(\nu, \phi) = |\nu| P(\nu, \phi)$. The power spectrum of a typical image decreases rapidly with increasing frequencies, whereas the noise power spectrum decreases in general slowly[7]. Consider a hypothetical perturbation of the data $p(s, \phi) \to p(s, \phi) + \cos(2\pi\nu_0 s)/\surd\nu_0$ for some $\nu_o > 0$. This perturbation becomes arbitrarily small when ν_0 tends to ∞, but the corresponding

[6] Alternative and faster implementations of the backprojection have been proposed [17, 18, 19].

[7] In the so-called *white noise* limit, the noise power spectrum is constant.

perturbation of the filtered projection is easily seen to be $p^F(s,\phi) \to p^F(s,\phi) + \sqrt{\nu_0}\cos(2\pi\nu_0 s)$ and is arbitrarily large for large ν_0. This artificial example illustrates the fact that the ill-posedness of the inverse x-ray transform (and of most inverse problems) arises from high-frequency perturbations.

This discussion suggests that the reconstruction can be stabilized by filtering out the high frequencies. This is achieved by introducing a low-pass apodizing window $w(\nu)$ as in Eq. (15). A window frequently used in tomography is the Hamming window

$$\begin{aligned} w_{ham}(\nu) &= (1+\cos(\pi\nu/\nu_c))/2 && |\nu|<\nu_c \\ &= 0 && |\nu|\geq\nu_c \end{aligned} \qquad (19)$$

where ν_c is some cut-off frequency. The rectangular window

$$\begin{aligned} w_{rec}(\nu) &= 1 && |\nu|<\nu_c \\ &= 0 && |\nu|\geq\nu_c \end{aligned} \qquad (20)$$

results in a better spatial resolution, but introduces ringing artifacts near sharp boundaries. Figure 4.5 illustrates the apodized window and the convolution kernel $h_w(s)$. The choice of the cut-off frequency must take two factors into account:

- Given the radial sampling distance Δs in the sinogram, Shannon's sampling theory states that the maximum frequency that can be recovered without aliasing is $1/2\Delta s$. The cut-off frequency is therefore constrained by $\nu_c \leq 1/2\Delta s$.
- As we have seen, stabilization requires suppressing high frequencies. Therefore, lower values of ν_c are selected when the signal-to-noise ratio (i.e. the number of detected coincidences) is low.

The stability of the discrete FBP can be analyzed assuming a Poisson distribution for the measurement noise. Consider the reconstruction of a disc of radius R containing a uniform tracer distribution, from 2D PET data comprising N_{events} coincident events. Neglecting attenuation, scatter and random, the relative variance of the reconstructed image at the center of the disc can be shown [21] to be

$$\text{variance} f(x=0, y=0) \simeq \frac{\pi^3}{6}\frac{(R/\Delta s)^3}{N_{\text{events}}} \qquad (21)$$

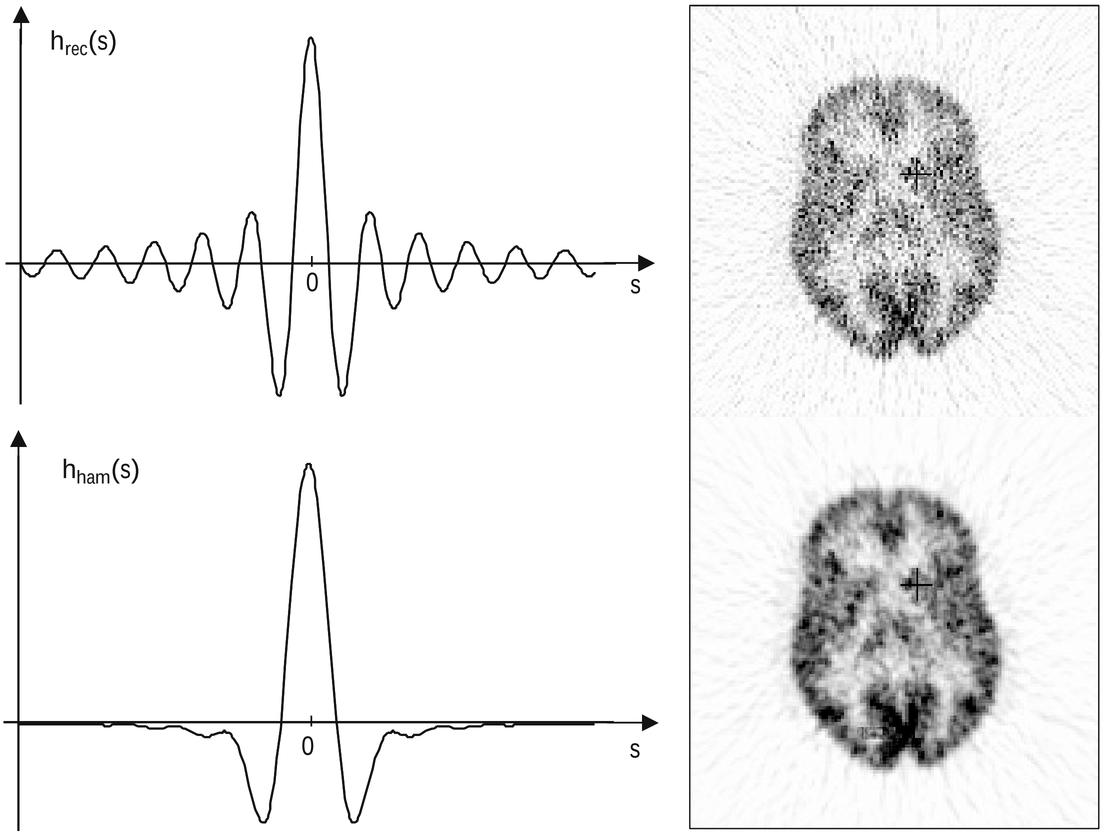

Figure 4.5. The convolution kernels corresponding to the rectangular window in equation (20) (top), and to the Hamming window (19) (bottom) are shown with arbitrary vertical scales. The smaller width of the central lobe of $h_{rec}(s)$ results in higher spatial resolution in the reconstruction, while the larger side lobes, compared to $h_{ham}(s)$ indicate a higher sensitivity to noise. A transaxial slice of an FDG brain scan reconstructed using FBP with these two windows is shown on the right.

where Δs is the radial sampling and a rectangular window with $\nu_c = 1/2\Delta s$ has been used. This result means that the number of detected events, hence the scanner sensitivity, should be multiplied by a factor of 8 when the spatial resolution Δs is halved. This is to be compared with the factor 4 increase that suffices in the absence of tomographic reconstruction, e.g. if perfect time-of-flight information is available, or if f is obtained from planar scintigraphy as in single photon imaging. The supplementary factor of 2 reflects the ill-posed character of the inverse x-ray transform. For a multi-slice 2D reconstruction, an additional factor of 2 must be included if the axial resolution is also halved, leading to a 16 fold increase of the number of counts when the isotropic resolution is halved. When an improvement in detector resolution is not matched by an increase in sensitivity, a cut-off frequency ν_c smaller than the Nyquist frequency $1/2\Delta s$ must be used to limit noise. In such a case, the improvement in detector resolution is not fully translated in the reconstructed image resolution. The improvement nevertheless remains beneficial because the modulation transfer function is enlarged at the lower frequencies $|\nu| \leq \nu_c$, allowing better recovery coefficients for small structures.

Iterative Reconstruction

This section introduces the major concepts of the iterative reconstruction algorithms, which play an increasingly important role in clinical PET. These algorithms rely on a discrete representation of both the data and the reconstructed image, in contrast with the analytic algorithms, which are derived assuming a continuous data sampling and introduce the discrete character of the data a posteriori. We begin this section with a general discussion of the ingredients of an iterative algorithm: the data model, the image model, the objective function, and the optimization algorithm. We refer to [22, 23] for more details. The various possible choices for each of these ingredients explains the wide variety of iterative algorithms in the literature. One specific algorithm will be described in detail in the section on ML-EM and OSEM (below).

One of the strengths of iterative algorithms is that they are largely independent of the acquisition geometry. Therefore, the concepts presented below apply equally to 2D and to 3D PET data.

The General Ingredients of Iterative Reconstruction Algorithms: Data Model

The data are represented using Eq. (1). To simplify notations, a single index j is used to denote the detector pair (d_a, d_b), and the mean number of events detected for one LOR is then rewritten as

$$\langle \vec{p} \rangle = \{\langle p_j \rangle = \tau \int_{FOV} d\vec{r} f(\vec{r}) \Psi_j(\vec{r}), j = 1, \ldots, N_{LOR}\} \quad (22)$$

Any linear physical effect can be modeled in the sensitivity function Ψ_j : attenuation and scatter (assuming a known density map), gaps in the detectors, non-uniform resolution of the detectors, etc. The accuracy of the physical model ultimately determines the accuracy of the reconstruction. Nevertheless, approximate models are often used to limit the computational burden, and these approximations are justified for low-count studies where image quality is primarily limited by noise. Many approaches can be found, ranging from a simple line integral model (as for FBP) up to a highly accurate model required for high SNR studies with small-animal scanners. A clever exploitation of the symmetries of the scanner and the use of lookup tables, as described in Qi et al. [24], allows the computational costs of such a complex modeling to be kept to a reasonable level.

Eq. (22) represents the *mean* value of the data. The *statistical distribution* of each LOR data p_j around its mean value $< p_j >$ must also be modeled. An inaccurate statistical model results not only in a sub-optimal variance, but also in a bias. Usually, the "raw data" p_j are counted numbers of detected photon pairs and are distributed as independent Poisson variables. The *likelihood function* then has the form

$$\Pr\{\vec{p} \mid f\} = \prod_{j=1}^{N_{LOR}} \exp(-\langle p_j \rangle) \langle p_j \rangle^{p_j} \,/\, p_j! \quad (23)$$

Due to the various forms of data pre-processing, the actual distribution of the data presented to the algorithm often deviates from the Poisson model. If the number of counts per bin is high enough, the distribution is approximately Gaussian

$$\Pr\{\vec{p} \mid f\} = \prod_{j=1}^{N_{LOR}} \frac{1}{\sqrt{2\pi}\,\sigma_j} \exp(-\frac{(p_j - \langle p_j \rangle)^2}{2\sigma_j^2}) \quad (24)$$

and the variance σ_j^2 of each LOR can be estimated knowing the data pre-processing steps. A more general Gaussian model with a non-diagonal covariance matrix may be needed if the pre-processing introduces correlations between LORs.

Another variant has been proposed to model data pre-corrected for random coincidences[8]. When this is done by subtracting delayed coincidences, the subtracted data (the "prompts" minus the "delayed") are no longer Poisson variables. An approximate model, the *shifted Poisson model*, for the distribution of such pre-corrected data has been proposed in [25, 26].

We conclude this section on data modeling with a few words on the reconstruction of transmission data acquired with monoenergetic photons of energy E. Typically, E = 511 keV if a positron source such as ^{68}Ge/^{68}Ga is used or E = 662 keV for a ^{137}Cs single photon source. These data are also distributed as Poisson variables but with mean values

$$\langle\vec{p}\rangle=\{\langle p_j\rangle=p_j^0\exp(-\int_{FOV}d\vec{r}\mu(\vec{r},E)\Psi_j^{tr}(\vec{r})) \quad j=1,\ldots,N_{LOR}\} \tag{25}$$

instead of Eq. (22). Here, p_j^0 is the mean number of coincident events in the reference (blank) scan, Ψ_j^{tr} is the sensitivity function for the transmission data, and $\mu(\vec{r}, E)$ is the attenuation coefficient to be reconstructed. The difference between this model and Eq. (22) shows that specific iterative algorithms are needed for transmission data [27, 28, 29, 30]. An alternative is to apply algorithms developed for emission data to the logarithm $\log(p_j^0 / p_j)$, but this approach introduces significant biases because the logarithm of the data is not a Poisson variable any more. Examples of these biases are given in [31].

The Image Model: Basis Functions and Prior Distribution

Iterative algorithms model the image as a linear combination of basis functions

$$f(x,y) \simeq \sum_{i=1}^{P} f_i b_i(x,y) \tag{26}$$

Most algorithms use contiguous and non-overlapping pixel basis functions, which partition the field of view:

$$\begin{aligned} b_i(x,y) &= 1 \quad |x-x_i| < \Delta x/2 \text{ and } \quad |y-y_i| < \Delta x/2 \\ &= 0 \quad |x-x_i| \geq \Delta x/2 \text{ or } \quad |y-y_i| \geq \Delta x/2 \end{aligned} \tag{27}$$

with $i = (i_x, i_y)$ and the center of the i^{th} pixel is $(x_i = i_x\Delta x, y_i = i_y\Delta x)$. The pixel size is $\Delta x = \Delta s/Z$, where Z is the *zoom factor*.

The pixel basis function is not band-limited: its Fourier transform $(\mathcal{F}b_i)(\nu_x, \nu_y)$ decreases slowly at large frequencies due to the discontinuity at the boundary of the pixel. This property is at odds with the fact that the frequencies larger than $\nu_c = 1/2\Delta s$ cannot be recovered from sampled data (see The Ill-posedness of the Inverse X-ray Transform, above). An alternative proposed by Lewitt [32] consists of using smooth basis functions which are essentially band-limited. Significant improvements in image quality have been demonstrated using truncated Kaiser–Bessel functions, dubbed *blobs* [33]. These radially symmetrical basis functions have a compact support, but they do overlap, which increases the processing time unless the spacing and size of the basis functions are carefully chosen. At the time of writing, most iterative algorithms are still based on discontinuous basis functions, but at least one clinical scanner implements blobs.

In principle, the choice of the basis functions determines the image model and reduces image reconstruction to the estimation of a vector $\{f_i, i = 1, \cdots, P\}$, usually with the constraint $f_i \geq 0$. The constraint implicit in this discrete representation[9] helps to stabilize the reconstruction, but may be insufficient. In such a case, a small perturbation of the data vector $\vec{p}$ still causes an unacceptably large perturbation of the reconstruction $\vec{f}$. The set of admissible images must then be further restricted. Several techniques can be used for this purpose, we focus here on the popular *Bayesian scheme* (see, for example, [34, 35, 7]).

In the Bayesian scheme, regularization is achieved by considering the image as a random vector with a prescribed probability distribution $\Pr(\vec{f})$. This distribution is called the *prior distribution* (or simply "the prior"). Typically, the prior enforces *smoothness* by assigning a low probability to images having large differences $|f_{i_x, i_y} - f_{i_x \pm 1, i_y \pm 1}|$ between neighboring pixels. One says that large differences between neighboring pixels are *penalized* by the prior. In practice, priors are defined empirically because the clinically relevant prior information is usually too complex to be expressed mathematically. We will see in the section on the cost function (below) how the prior is incorporated in the reconstruction.

Priors based on a Gaussian distribution with a uniform (i.e. shift-invariant) covariance are in essence equivalent to the linear smoothness constraint introduced in the FBP algorithm by low pass windows $w(\nu)$ discussed earlier. More sophisticated priors can

[8] In contrast with the randoms, the contribution of scattered coincidences is linearly related to true coincidences and hence can in principle be included in the model $\Psi_j(\vec{r})$.However, the scatter background is more often subtracted from the data prior to reconstruction for the sake of numerical efficiency. See chapter 5.

[9] Mathematically we constrain $f(x, y)$ to belong to the P-th dimensional space of functions spanned by the b_i.

improve image quality, especially sharpness, in specific situations and for specific tasks, but may introduce subtle nonlinear biases and noise correlations.

An attractive class of prior distributions exploits a registered anatomic, MR or CT, image of the patient [36, 37, 38, 39, 40]. This image defines likely boundaries between regions in which uniform tracer concentration is expected. These boundaries can be incorporated in a prior that enforces smoothness only between pixels belonging to the same anatomical region. Despite promising results, that approach still needs further validation and comparison with the alternative approach in which the MR or CT prior information is exploited visually using, for example, image fusion techniques.

Let us finally stress that the 2D or 3D nature of the image model is independent of the fact that the data are acquired in 2D or 3D mode. Indeed, true 3D image models based, for example, on 3D blobs and on 3D smoothness constraints are useful even when the data are collected independently for each slice (or rebinned, see section on 3D analytic reconstruction by rebinning (below)) [41]. For dynamic or gated PET studies, mixed basis functions depending on both the time and the spatial coordinates can be defined to model the expected behavior of the tracer kinetics [42, 43].

The System Matrix

We can now summarize the assumptions in the two previous sections. Putting the image model (Eq. (26)) into the data model (Eq. (22)) reduces the problem to a *set of linear equations*:

$$\langle p_j \rangle = \sum_{i=1}^{P} a_{j,i} f_i \qquad j = 1,\ldots,N_{LOR} \tag{28}$$

where the elements of the *system matrix* are

$$a_{j,i} = \tau \int_{FOV} d\vec{r}\, b_i(\vec{r})\Psi_j(\vec{r}) \qquad j = 1,\ldots,N_{LOR}; i = 1,\ldots,P \tag{29}$$

A line integral model including only attenuation correction and normalization generates a sparse system matrix a with elements simple enough to be calculated on the fly. More accurate models that include scatter lead to densely populated matrices, which are complex to calculate. A practical algorithm then requires a compromise between accuracy, required storage, and speed. A useful approach is to factor a as a product of matrices, each of which models a specific aspect of the data acquisition [24].

A direct inversion of the linear system (Eq. (28)) with the $< p_j >$ replaced by the measured data p_j is impractical for two reasons:

- The discrete system is *ill-conditioned*: the condition number of a is large[10]. Consequently, the solution[11] of Eq. (28) is unstable for small perturbations $p_j - < p_j >$ of the data. Ill-conditioning is the discrete equivalent of the ill-posedness of the inverse x-ray transform discussed in the section on the ill-posedness of the inverse X-ray transform (above).
- Numerically, the inversion of matrix a is hindered by its very large size (typically $P = 10^6$ unknowns and $N_{\text{LOR}} \simeq 10^6$ up to $N_{\text{LOR}} \simeq 10^9$ in 3D PET).

The first problem is solved by incorporating prior knowledge in a cost function. The second, numerical problem, is solved by optimizing the cost function by successive approximations.

The Cost Function

The key ingredient of an iterative algorithm is a cost function $Q(\vec{f} = (f_1, \cdots, f_p), \vec{p})$, which depends on the unknown image coefficients and on the measured data. $Q(\vec{f}, \vec{p})$ is also called the objective function. The reconstructed image estimate f^* is defined as one that maximizes Q:

$$\vec{f}^* = \arg\max_{\vec{f}} Q(\vec{f}, \vec{p}) \tag{30}$$

with usually the constraint $f_j \geq 0$. The role of the cost function is to enforce (i) a good fit with the data, i.e., Eq. (28) should be approximately satisfied, (ii) the prior conditions on the image model.

In the Bayesian framework, the cost function is the *posterior probability distribution*

$$\Pr\{\vec{f} \mid \vec{p}\} = \frac{\Pr\{\vec{p} \mid \vec{f}\}\Pr\{\vec{f}\}}{\Pr\{\vec{p}\}} \tag{31}$$

The first factor in the numerator of the right hand side is the data likelihood (given, for example, by the Poisson model (Eq. (23)), and the second factor is the prior probability discussed above in the section on the image model. The denominator is independent of $\vec{f}$ and can be dropped. Maximizing the posterior proba-

[10] The condition number of a matrix is the ratio between its largest and smallest singular values.

[11] Or the generalized Moore–Penrose solution if a is singular or $\vec{p} \notin$ range a, see [20].

bility is equivalent to maximizing its logarithm, and the cost function becomes

$$Q(\vec{f} \mid \vec{p}) = \log \Pr\{\vec{p} \mid \vec{f}\} + \log \Pr\{\vec{f}\} \qquad (32)$$

The first term penalizes images which do not well fit the data, whereas the second one stabilizes the inversion by penalizing images which are deemed a priori "unlikely". An image maximizing $Q(\vec{f}, \vec{p})$ is called a *maximum a posteriori* (MAP) estimator. When the log-likelihood is Gaussian (Eq. (24)), the first term in Eq. (32) is a quadratic function and the algorithm maximizing $Q(\vec{f}, \vec{p})$ is called a *penalized weighted least-square* method [44, 45].

Ideally, the maximum of Q should be unique. Uniqueness is guaranteed when the cost function is convex, i.e., when the Hessian matrix

$$H_{i,j} = \frac{\partial^2 Q(\vec{f}, \vec{p})}{\partial f_i \partial f_j} \qquad i, j = 1, \ldots, P \qquad (33)$$

is negative definite for all feasible $\vec{f}$. Non-convex cost functions may still have an unique global maximum, but they can also have local maxima, which complicate the optimization.

Optimization Algorithms

The cost function (assuming it has an unique global maximum) defines the looked-for estimate $\vec{f}^*$ of the tracer distribution. To actually calculate $\vec{f}^*$, an optimization algorithm is needed. Such an algorithm is a prescription to produce a sequence of image estimates $\vec{f}^n$, $n = 0, 1, 2, \cdots$, which should converge asymptotically to the solution:

$$\lim_{n-\infty} \vec{f}^n = \vec{f}^* \qquad (34)$$

Asymptotic convergence is not the only requirement: the optimization algorithm should be stable, efficient numerically, and ensure fast convergence independently of the choice of the starting image $\vec{f}^0$. A further property is that of monotonic convergence, which guarantees that $Q(\vec{f}^{n+1}, \vec{p}) \geq Q(\vec{f}^n, \vec{p})$ at each iteration. Though not strictly needed, monotonic convergence is useful in practice and is often the key property used to prove asymptotic convergence.

In principle, the choice of the optimization algorithm should not influence the solution, which is defined by Eq. (30). In practice, however, the image that will be used is produced by a necessarily finite number of iterations and thereby does depend on the algorithm.

When the cost function is differentiable and a non-negative solution is required, the solution $\vec{f}^*$ must satisfy the *Karush–Kuhn–Tucker conditions*:

$$\begin{aligned} (\nabla Q(\vec{f}^*, \vec{p}))_j &= 0 \qquad f_j^* > 0, j = 1, \ldots, P \\ &\leq 0 \qquad f_j^* = 0 \end{aligned} \qquad (35)$$

where the gradient of the cost function is the vector with components

$$\nabla Q(\vec{f}^*, \vec{p}))_j = \frac{\partial Q(\vec{f}, \vec{p})}{\partial f_j} |_{\vec{f} = \vec{f}^*} \qquad (36)$$

When positivity is not enforced, the Karush–Kuhn–Tucker condition reduces to the first line of Eq. (35). If in addition the cost function is quadratic (e.g., with a Gaussian log-likelihood), optimization reduces to a set of P linear equations in P unknowns. With a Gaussian likelihood without prior, these equations are the so-called *normal equations* corresponding to Eq. 28 [7, 20].

There is a considerable literature on optimization, and even within the field of tomography a wide variety of methods have been proposed. A detailed overview (see [7, 16, 22, 46]) is beyond the scope of this chapter, but it may be useful to briefly list a few basic tools that can be used to develop iterative methods. The major difficulty is that the system of equations (Eq. (35)) is large, strongly coupled, and often non-linear. Many algorithms are based on the replacement at each iteration of the original optimization problem (Eq. (30)) by an alternative problem which is easier to solve because

- it has a much smaller dimensionality, and/or
- the modified cost function is quadratic in its unknowns, or even better separable in the sense that its gradient is a sum of functions each depending on a single unknown parameter f_j.

Standard examples include:

(i) *Gradient-based methods*. The prototype is the steepest-ascent method, which reduces the problem to a one-dimensional optimization along the direction defined by the gradient. The n^{th} iteration is defined by:

$$\begin{aligned} \vec{f}^{n+1} &= \vec{f}^n + \alpha_n \nabla Q(\vec{f}^n, \vec{p}) \\ \alpha_n &= \arg \max_{\alpha} Q(\vec{f}^n + \alpha \nabla Q(\vec{f}^n, \vec{p}), \vec{p}) \end{aligned} \qquad (37)$$

The step length α_n maximizes the cost function along the gradient direction, taking into account possible constraints such as positivity.

(ii) *Methods using subsets of the image vector.* Only a subset of the components of the unknown image vector (i.e., a subset of voxels) is allowed to vary at each iteration, while the value of the other components is kept constant. A different subset of voxels is allowed to vary at each iteration. In the *coordinate ascent algorithm*, a single voxel is varied at each iteration, according to:

$$\begin{aligned} f_j^{n+1} &= f_j^n & j \neq J(n) \\ &= \arg\max_{f_j} Q((f_1^n, \ldots, f_{j-1}^n, f_j, f_{j+1}^n, \ldots, f_p^n), \vec{p}) & j = J(n) \end{aligned} \tag{38}$$

where $J(n)$ defines the order in which voxels are accessed in successive iterations, e.g., $J(n) = n \bmod P$.

(iii) *Methods based on surrogate cost functions.* The original cost function $Q(\vec{f}, \vec{p})$ is replaced at each step by a modified objective function $\tilde{Q}(\vec{f}, \vec{f}^n, \vec{p})$ that satisfies the following conditions [47]:

- $\tilde{Q}(\vec{f}, \vec{f}^n, \vec{p})$ can easily be maximized with respect to $\vec{f}$, e.g., it is quadratic or separable,
- $\tilde{Q}(\vec{f}, \vec{f}, \vec{p}) = Q(\vec{f}, \vec{p})$
- $\tilde{Q}(\vec{f}, \vec{f}^n, \vec{p}) \leq Q(\vec{f}, \vec{p})$

The two last conditions ensure that the next image estimate

$$\vec{f}^{n+1} = \arg\max_{\vec{f}} \tilde{Q}(\vec{f}, \vec{f}^n, \vec{p}) \tag{39}$$

monotonically increases the value of the cost function: $Q(\vec{f}^{n+1}, \vec{p}) \geq Q(\vec{f}^n, \vec{p})$. The ML-EM algorithm (next section), the least-square ISRA algorithm [48, 49], and Bayesian variants [50] can be derived using surrogate functions.

(iv) *Block-iterative methods* use at each iteration only a subset of the data. They are called *row-action* methods when a single datum is used at each iteration as in the ART algorithm. The OSEM method (see next section) and its variants are also block-iterative methods. While allowing significant acceleration of the optimization, these methods do not guarantee a monotonic increase of the cost function. In addition, the iterated image estimates tend asymptotically to cycle between S slightly different solutions, where S is the number of subsets. Appropriate under-relaxation can be used to alleviate the problem.

ML-EM and OSEM

The most widely used iterative algorithms in PET are the ML-EM (maximum-likelihood expectation maximization) algorithm and its accelerated version OSEM (Ordered Subset EM). The ML-EM method was introduced by Dempster et al in 1977 [51] and first applied to PET by Shepp and Vardi [52] and Lange and Carsson [27]. The algorithm is akin to the Richardson–Lucy algorithm developed for image restoration in astronomy (see, for example, [20]). The OSEM variation of the ML-EM algorithm, proposed in 1994 by Hudson and Larkin was the first iterative algorithm sufficiently fast for clinical applications.

The cost function in the ML-EM and OSEM algorithms is the Poisson likelihood (Eq. (23)). Putting Eq. (28) into Eq. (23), taking the logarithm, and dropping the terms that do not depend on the unknowns f_i, we get

$$Q(\vec{f}, \vec{p}) = \sum_{j=1}^{N_{LOR}} \{-\sum_{i=1}^{P} a_{j,i} f_i + p_j \log(\sum_{i=1}^{P} a_{j,i} f_i)\} \tag{40}$$

If the matrix a is non-singular, this cost function is convex and defines a unique image.

The EM iteration is a mapping of the current image estimate $\vec{f}^n$ onto the next estimate $\vec{f}^{n+1}$:

$$\vec{f}_i^{n+1} = f_i^n \frac{1}{\sum_{j'=1}^{N_{LOR}} a_{j',i}} \sum_{j=1}^{N_{LOR}} a_{j,i} \frac{p_j}{\sum_{i'=1}^{P} a_{j,i'} f_{i'}^n} \qquad i = 1, \ldots, P \tag{41}$$

Usually, the first estimate is a uniform distribution $f_i^1 = 1$, $i = 1, \ldots, P$. The sum over i' in the denominator of the second factor in the right hand side is a forward projection and corresponds to Eq. (28): therefore the denominator is the average value $<p_j^n>$ that would be measured if $\vec{f}^n$ was the true image. The sum over j in the numerator is a multiplication with the transposed system matrix and represents the backprojection of the ratio between the measured and estimated data. Finally, the denominator in the first factor is equal to the sensitivity of the scanner for pixel i.

The ML-EM iteration has several remarkable properties:

- The cost function increases monotonically at each iteration, $Q(\vec{f}^{n+1}, \vec{p}) \geq Q(\vec{f}^n, \vec{p})$,
- The iterates $\vec{f}^n$ converge for $n \rightarrow \infty$ to an image $\vec{f}^*$ that maximizes the loglikelihood,
- All image estimates are non-negative if the first one is,
- The algorithm can easily be implemented with list-mode data [53, 54, 55, 56] because the only LORs that contribute to the backprojection sum over j in Eq. 41 are those for which at least one event has been detected ($p_j \geq 1$).

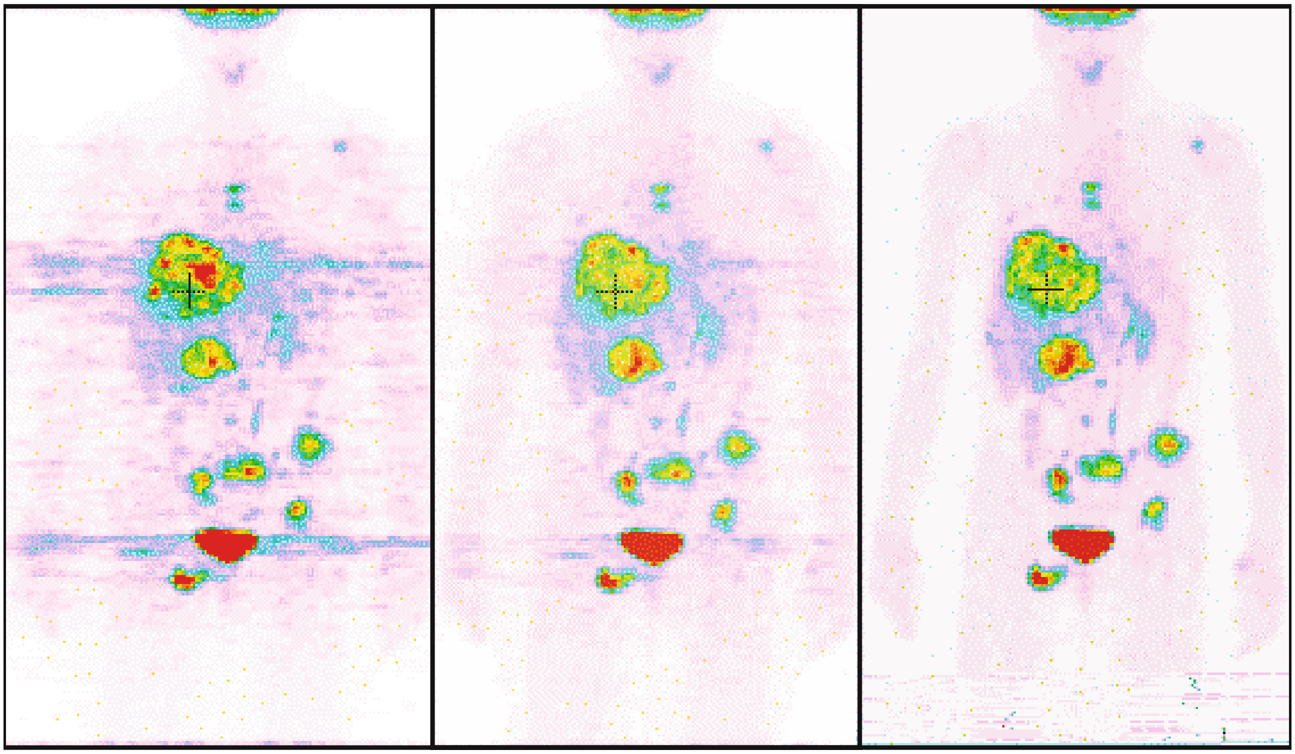

Figure 4.7. Comparison between the FBP and the OSEM reconstruction of a 2D FDG whole-body study, showing a frontal section. The algorithms used for the reconstruction of the transmission scan and of the emission scan are FBP-FBP (left), OSEM-FBP (center), OSEM-OSEM (right).

The ML-EM iteration (Eq. (41)) is then applied incorporating the data from one subset only. Each subset is processed in a well-defined order, usually in a periodic pattern where subset $J_{n \bmod S}$ is used at iteration n[12]:

$$f_i^{n+1} = f_i^n \frac{1}{\sum_{j' \in J_{n \bmod S}} a_{j',i}} \sum_{j \in J_{n \bmod S}} a_{j,i} \frac{p_j}{\sum_{i'=1}^{P} a_{j,i'} f_{i'}^n}$$
$$i = 1, \ldots, P \qquad (42)$$

Empirically, the convergence is accelerated by a factor $\simeq S$ with respect to ML-EM. But the asymptotic convergence to the maximum-likelihood estimator is no longer guaranteed. In fact, OSEM tends to cycle between S slightly different image estimates. To minimize the adverse effects of this behavior it is recommended to keep the number of 1D parallel projections in each subset equal to at least 4. In addition, several authors suggest progressively decreasing the number of subsets during iteration. Finally, we only mention here the *row action maximum likelihood* (RAMLA) algorithm [63] and the *rescaled block-iterative ML-EM* algorithm [64]. These two algorithms for maximum-likelihood estimation with a Poisson distribution are closely related to OSEM, but guarantee asymptotic convergence under certain conditions.

Compared to FBP reconstructions, some qualitative characteristics of images reconstructed from Poisson data using ML-EM or OSEM are:

- Reduced streak artifacts
- A better SNR in regions of low tracer uptake, resulting in particular in a better visibility of the contours of the body
- Some non-isotropy and non-uniformity of the spatial resolution, especially when the range of values of the attenuation correction factor is large, as e.g. in the chest
- A slower convergence for regions of low tracer uptake than for regions of high tracer uptake.

Figure 4.7 illustrates some of these properties.

Finally, some comments are in order about data corrections prior to reconstruction with ML-EM or OSEM. Physical effects such as detector efficiency variations, attenuation, scattered and random coincidences, etc., must be accounted for to obtain quantitatively correct

[12] In the OSEM jargon, such an iteration is called a *sub-iteration*, and an *iteration* denotes a set of S consecutive sub-iterations, corresponding to one pass through the whole data set.

images. With analytical algorithms such as FBP the data are corrected before reconstruction to comply with the line integral model. With the ML-EM algorithm, on the contrary, pre-correction must be avoided because it would destroy the Poisson character of the data and thereby could bias the reconstruction. This means that the ML-EM algorithm should be applied to the raw data, and that all physical effects should be included in the system matrix as described in the section on the system matrix (above). A full modeling, however, can be impractical when the system matrix is too large to be pre-computed. A faster, approximate, procedure consists of including only the most significant effect – attenuation – in the system matrix. Corrections for scattered and random coincidences can be of the order of 50%, but are often less. Attenuation correction, however, involves multiplication by factors ranging from 5 to more than 100 ! The attenuation correction is multiplicative and can easily be incorporated in the ML-EM iteration as shown by Hebert and Leahy [65],

$$f_i^{n+1} = f_i^n \frac{1}{\sum_{j'=1}^{N_{LOR}} a_{j',i} / \alpha_{j'}} \sum_{j=1}^{N_{LOR}} a_{j,i} \frac{p_j}{\sum_{i'=1}^{P} a_{j,i'} f_{i'}^n} \qquad i = 1, \ldots, P \tag{43}$$

where the p_j are the data corrected for all effects except attenuation, α_j is the pre-computed attenuation correction factor[13] for LOR j, and the system matrix a does not include the effect of attenuation. This attenuation-weighting (AW) of the ML-EM algorithm is easily extended to the attenuation-weighted OSEM algorithm (AW-OSEM). The AW-OSEM approach has been shown to perform almost as well as algorithms that model all physical effects, with only modest increases in computation time over OSEM applied to pre-corrected sinogram data [66].

The previous approach can also be applied to other multiplicative corrections such as the normalization for detector efficiency variations. For more complex, e.g., non-linear, relations between the raw data p_j and the corrected data p_j^c , an approximate statistical modeling can be achieved by applying the ML-EM algorithm to *scaled* data $p_j^s = \beta_j p_j^c$, where $\beta_j = \langle p_j^c \rangle / \mathrm{var}(p_j^c)$ is a low-variance (smoothed) estimate of the ratio between the mean and the variance of the corrected data. With this choice of β_j the scaled data satisfy the same relation $\langle p_j^s \rangle \simeq \mathrm{var}(p_j^s)$ as data obeying Poisson statistics, and it is therefore reasonable to reconstruct them using the ML-EM algorithm [30]. This yields the following iteration:

$$f_i^{n+1} = f_i^n \frac{1}{\sum_{j'=1}^{N_{LOR}} \beta_{j'} a_{j',i}} \sum_{j=1}^{N_{LOR}} a_{j,i} \frac{p_j^s}{\sum_{i'=1}^{P} a_{j,i'} f_{i'}^n} \qquad i = 1, \ldots, P \tag{44}$$

In the case of the attenuation correction, $p_j^c = \alpha_j p_j$, and one easily checks that $\beta_j = 1/\alpha_j,\ and\ p_j^s = p_j$, so that Eqs. (44) and (43) coincide.

When the data are acquired in true mode as the difference between the prompt and delayed coincidences, the shifted Poisson model (described earlier) leads to the following modified ML-EM algorithm [26],

$$f_i^{n+1} = f_i^n \frac{1}{\sum_{j'=1}^{N_{LOR}} a_{j',i} / \alpha_{j'}} \sum_{j=1}^{N_{LOR}} a_{j,i} \frac{p_j + 2\bar{r}_j + \bar{s}_j}{\sum_{i'=1}^{P} a_{j,i'} f_{i'}^n + \alpha_j (2\bar{r}_j + \bar{s}_j)} \qquad i = 1, \ldots, P \tag{45}$$

where the p_j are the data corrected for random and scatter (but not attenuation), α_j and $a_{j,i}$ are as in equation (43), and $\bar{r}_j$ and $\bar{s}_j$ are low-variance estimates of the random and scatter background in LOR j. The mean random $\bar{r}_j$ is generally estimated using variance reduction techniques or from the single photon data (see sections Randoms Variance Reduction and Estimation from Single Rates in next chapter). The mean scatter ¯ $\bar{s}_j$ is estimated using a model based scatter model (see section Simulation-based Scatter Correction in next chapter).

Variance and resolution with non-linear reconstruction algorithms

Predicting and controling the statistical properties and the resolution of reconstructed PET images is of paramount importance for quantitative applications of PET and for task oriented performance studies using numerical observers. For clinical PET, a good awareness of these properties helps minimizing the probability of erroneous image interpretations.

Denote the "true" image by $\vec{f}$, the measured data vector by $\vec{p}$, and the mean data by $< \vec{p} >= A\vec{f}$, where A is the system matrix (see Eq. (28)). Consider any

[13] The ratio between the blank and transmission scans.

specific algorithm denoted by $\mathcal{T}$ (for example 100 ML-EM iterations with a uniform initial image estimate). The reconstruction is then $\vec{f}^* = \mathcal{T}(\vec{p})$, and the reconstruction error is

$$\vec{f}^* - \vec{f} = (\mathcal{T}(\vec{p}) - <\mathcal{T}(\vec{p})>) + (<\mathcal{T}(\vec{p})> - \vec{f}) \qquad (46)$$

where $<\mathcal{T}(\vec{p})>$ denotes the mean value of the reconstructed image, which could be estimated by averaging a large number of images reconstructed by applying the algorithm $\mathcal{T}$ to statistically independent realizations of the random data vector $\vec{p}$. The first term in the RHS of Eq. (46) is the *statistical error* due to the fluctuations of data $\vec{p}$ around its mean value $<\vec{p}>$. The statistical error is characterized by the covariance matrix

$$V_{j,j'} = <(\mathcal{T}(\vec{p})_j - <\mathcal{T}(\vec{p})_j>)\,(\mathcal{T}(\vec{p})_{j'} - <\mathcal{T}(\vec{p})_{j'}>)> \\ j,j' = 1, \dots, P \qquad (47)$$

the diagonal elements of which give the variance of each reconstructed pixel value. The second term in the RHS of Eq. (46) is the *systematic error* or *bias*: even the mean value of the reconstructed image is not exact because of sampling, apodization, finite number of iterations, etc.

For a linear reconstruction algorithm the image covariance can easily be determined once we know the statistical (e.g. Poisson) properties of the data. In addition, with a linear algorithm, the systematic error is fully characterized by the *point response* defined as the reconstruction of the mean data of a point source located in a voxel $j_0 \in [1, \dots, P]$. While the point response depends in general on the position of the voxel j_0 relative to the scanner, it does not depend on the strength of the point source, or on whether that source is sitting or not over some background. This allows an unambiguous definition of the resolution, using parameters such as the FWHM of the point response. For the FBP algorithm, in particular, the statistical error and the bias are determined by the apodized ramp filter, and the trade-off between these two errors is well understood (see the section Ill-posedness of the Inverse X-ray Transform).

For non-linear algorithms such as ML-EM, the derivation of analytical expressions for the covariance matrix is complex. More importantly, the point response becomes object dependent. To understand this important point, consider any data set $\vec{p}$, measured e.g. as a "normal" whole-body tracer distribution. Consider also some additional point source $\Delta\vec{f}$ located in voxel j_0: $\Delta f_j = \delta_{j,j0}$, and denote the corresponding mean contribution to the data by $\Delta\vec{p} = A\Delta\vec{f}$. Then the non-linearity of the algorithm $\mathcal{T}$ means that, in general,

$$\mathcal{T}(\vec{p} + \Delta\vec{p}) \neq \mathcal{T}(\vec{p}) + \mathcal{T}(\Delta\vec{p}) \qquad (48)$$

A concrete consequence of this non-linearity can be observed when the ML-EM algorithm is used with the small number of iterations typical of clinical practice: the reconstruction of a unit "point source" sitting on top of a uniform background broadens when the strength of the background is increased. Similarly, the anisotropy of the attenuation correction factors for an elongated object such as the chest at the level of the shoulders is translated by ML-EM into an anisotropy of the point response: if the point source is located in an ellipsoidal attenuating medium with long axis along the x-axis, the point response takes an ellipsoidal shape with long axis along the y-axis. One should therefore interpret with care results on the "reconstructed resolution of ML-EM" obtained for isolated point or line sources. Similar observations hold for MAP or other non-linear algorithms.

A local infinitesimal point response function, depending both on the data $\vec{p}$ and on the position at voxel j_0, can be defined as the image

$$\vec{\delta}_{\vec{p},j_0} = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}(\mathcal{T}(\vec{p} + \varepsilon \triangle \vec{p}) - \mathcal{T}(\vec{p})) = \frac{\partial \mathcal{T}(A\vec{f})}{\partial f_{j_0}} \in \mathbb{R}^P \qquad (49)$$

Approximate expressions and efficient numerical techniques have been developed [67] to calculate this point response, as well as methods to design a penalty term log $Pr\{\vec{f}\}$ in Eq. (32)[14] that guarantee homogeneous resolution [68]. An alternative approach to improve the homogeneity of the resolution consists in pursuing the ML-EM iteration beyond the point where the image is deemed acceptable, and in post-filtering this image with an appropriate filter [69].

The image covariance (Eq. (47)) can be estimated numerically by reconstructing a large number of data sets simulated with statistically independent pseudorandom noise realizations. An alternative for maximum-likelihood algorithms is to calculate the Fisher information matrix, the inverse of which is related by the Cramer-Rao theorem to the covariance of the ML estimator (see e.g. [7]). An approximate expression of the covariance, for the more relevant case where ML-EM iteration is stopped well before convergence, was derived in [57], and validated numerically in [58].

[14] This penalty depends on the data and can no longer be interpreted as a real Bayesian prior.The algorithm is then better referred to as a penalized likelihood method.

The major conclusion is that the variance of the ML-EM reconstruction is roughly proportional to the image itself, i.e.

$$V_{j,j} \simeq C < f_j >^2 \qquad j = 1, \ldots, P \tag{50}$$

for some constant C depending on the object and on the number of iterations. Thus, the ML-EM reconstructions have lower variance in regions of low tracer uptake, thereby allowing good detectability in these regions. This is in contrast with FBP reconstructions, in which the noise arising from the high uptake regions spreads more uniformly over the whole FOV, resulting in particular in the well-known streak artefacts.

3D Data Organization

Two-dimensional Parallel Projections

We have seen in the previous section that 2D data acquired with a ring scanner can be stored in a sinogram $p(s, \phi)$. If the data are modeled as line integrals, as for analytic algorithms, the sinogram is a set of 1D parallel projections of $f(x, y)$ for a set of orientations $\phi \in [0, \pi]$. Similarly, the LORs measured by a volume PET scanner can be grouped into sets of lines parallel to a direction specified by a unit vector $\vec{n} = (n_x, n_y, n_z) = (-\cos\theta \sin\phi, \cos\theta \cos\phi, \sin\theta) \in S^2$ where S^2 denotes the unit sphere. The angle θ is the angle between the LOR and the transaxial plane, so that the data acquired in a 2D acquisition therefore correspond to $\theta = 0$. The set of line integrals parallel to $\vec{n}$ is a *2D parallel projection* of the tracer distribution:

$$p(\vec{s}, \vec{n}) = \int_{\mathbb{R}} dt\, f(\vec{s} + t\vec{n}) \tag{51}$$

where the position of the line is specified by the vector $\vec{s} \in \vec{n}^{\perp}$, which belongs to the *projection plane* $\vec{n}^{\perp}$ orthogonal to $\vec{n}$.

Consider a cylindrical scanner with N_r rings of radius R_d, extending axially over $0 \leq z \leq L$, where $L = N_r\Delta z$. Assuming continuous sampling, this scanner measures all LORs such that the line defined by $(\vec{s}, \vec{n})$ has two intersections with the lateral surface of the cylinder (these intersections are the positions of the two detectors in coincidence). The set of measured orientations is

$$\Omega(\theta_{\max}) = \{\vec{n} = (\phi, \theta) \mid \phi \in [0, \pi), \theta \in [-\theta_{\max}, +\theta_{\max}]\} \tag{52}$$

with $\tan\theta_{\max} = L / 2\sqrt{R_d^2 - R_F^2}$, where R_F is the radius of the transaxial FOV. However, for each $\theta \neq 0$, not all LORs parallel to $\vec{n}$ and crossing the FOV of the scanner are measured. That is, the parallel projection $p(\vec{s}, \vec{n})$ is measured only for some subset of LORs $\vec{s} \in M(\vec{n}) \subset \vec{n}^{\perp}$. One says that this projection is *truncated*.

Two important properties of the 3D data can already be stressed:

(i) 3D data are *redundant* since four variables are required to parameterize $p(\vec{s}, \vec{n})$ (two for the orientation $\vec{n}$ and two for the vector $\vec{s}$) whereas the image only depends on three variables (x, y, z).
(ii) 3D data are not invariant for translation as in the 2D case because the cylindrical detector has a finite length and the measured projections are truncated.

The vector $\vec{s}$ can be defined by its components (s, u) on two orthonormal basis vectors in $\vec{n}^{\perp}$.

$$\vec{s} = s(\cos\phi, \sin\phi, 0) + u(\sin\theta\sin\phi, -\sin\theta\cos\phi, \cos\theta) \tag{53}$$

The variable s coincides with the 2D radial sinogram variable of Eq. (3). We will thus write $p(\vec{s}, \vec{n}) = p(s, u, \phi, \theta)$. The subset $p(s, u, \phi, 0)$ is the 2D sinogram of the slice $z = u$.

The LORs measured by a PET scanner do not uniformly sample the variables (s, u, ϕ, θ), and therefore interpolation is needed to reorganize the raw data into parallel projections. This holds both for multi-ring scanners and for scanners based on flat panel detectors.

Oblique Sinograms

Some analytic algorithms use an alternative parameterization of the parallel projections, where the vari-

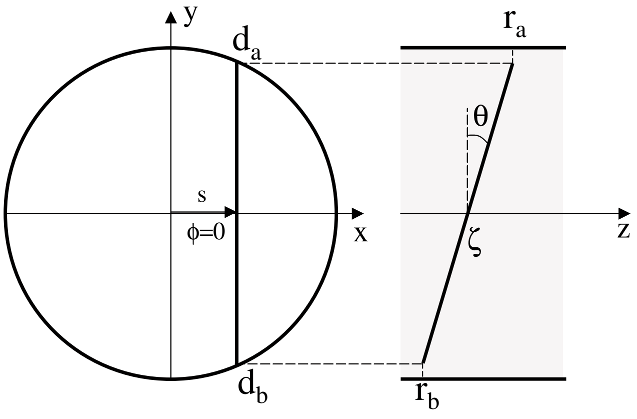

Figure 4.8. A transverse and a longitudinal view of a multi-ring scanner. An LOR connecting a detector d_a in ring r_a to a detector d_b in ring r_b is shown, with the four variables (s, ϕ, ζ, θ) used for the oblique sinogram parameterization. The particular LOR represented has $\phi = 0$.

able u in Eq. (53) is replaced by the axial coordinate $\zeta = u/\cos\theta$, the average of the axial coordinates of the two detectors in coincidence. One defines weighted parallel projections

$$\begin{aligned} p_s(s,\phi,\zeta,\theta) &= p(s,\zeta\cos\theta,\phi,\theta)\cos\theta \\ &= \int_{\mathbb{R}} dt' f(s\cos\phi - t'\sin\phi, s\sin\phi + \\ &\quad t'\cos\phi, \zeta + t'\tan\theta) \end{aligned} \tag{54}$$

The domain of the variables is $|s| \le R_F$, $\phi \in [0, \pi)$, $|\theta| \le \arctan\left(L/2\sqrt{R_d^2 - s^2}\right)$, and $\zeta \in \left\lfloor |\tan\theta|\sqrt{R_d^2-s^2}, L - |\tan\theta|\sqrt{R_d^2-s^2}\right\rfloor$ (Fig. 4.8). For each pair ζ, θ the function $p_s(.,.,\zeta,\theta)$ is called an *oblique sinogram* by analogy with Eq. (3). The similarity with the 2D format makes this oblique sinogram format suited to the analytic rebinning algorithms, which reduce the 3D data to 2D data.

Consider now the discrete sampling of the oblique sinograms. The measured LORs connecting detector d_a in ring r_a to detector d_b in ring r_b corresponds to parameters (s, ϕ, ζ, θ) in Eq. (48), where s and ϕ are determined as in the 2D case (Eq. (4)), and the axial variables are determined by

$$\begin{aligned} \tan\theta &= (r_b - r_a)\Delta z / \left(2\sqrt{R_d^2 - s^2}\right) \\ \zeta &= (r_a + r_b)\Delta z/2 \end{aligned} \tag{55}$$

If the radius of the FOV is small, θ in Eq. (55) is approximately independent of s. With this approximation, the coincidences between two rings r_a and r_b can be used to build an oblique sinogram $p_s(.,.,\zeta,\theta)$ with $\zeta = (r_a + r_b)\Delta z/2$ and $\tan\theta = (r_b - r_a)\Delta z/(2R_d)$.

To save storage and computation, some volume scanners use axial angular undersampling by averaging sets of sinograms with adjacent values of θ. The degree of undersampling is characterized by an odd integer parameter S, called the *span*. The resulting sampling is non-interleaved:

$$\begin{aligned} \tan\theta &= i_\theta S\Delta z/(2R_d) & i_\theta &= -i_{\max},\ldots,+i_{\max} \\ \zeta &= i_z\Delta z/2 & z_{\min}(i_\theta) &\le i_z \le 2N_r - 2 - z_{\min}(i_\theta) \end{aligned} \tag{56}$$

where $z_{min}(i_\theta) = \max(0, |i_\theta|S - S/2)$. Each sample (i_θ, i_z) is obtained by averaging data from all pairs of rings such that

$$\begin{aligned} i_\theta S - S/2 \le r_b - r_a &\le i_\theta S + S/2 \\ i_z &= r_b + r_a \end{aligned} \tag{57}$$

The sampling scheme is often illustrated on a 2D diagram, the "Michelogram", in which each grid point represents one ring pair and each sampled oblique sinogram (i_θ, i_z) is represented by a line segment connecting the contributing pairs r_a, r_b (Fig. 4.9). Just as for the azimuthal undersampling ("mashing", see end of sinogram data and sampling section, above), a good

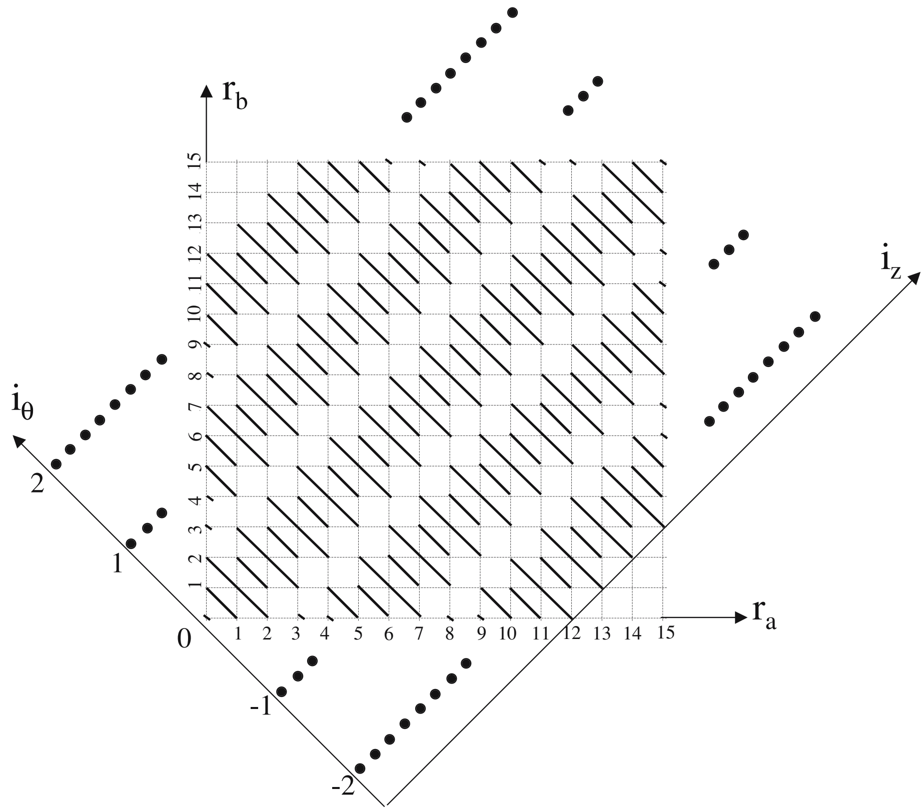

Figure 4.9. A Michelogram for a 16-ring scanner, illustrating the axial sampling with a span $S = 5$ and a maximum ring difference defined by $i_{max} = 2$. Each grid point corresponds to a ring pair, and each diagonal line segment links the ring pairs (2 or 3 except at the edge of the FOV) that are averaged to form one oblique sinogram. The samples located outside the square (dots) are the unmeasured oblique sinograms needed to obtain a shift-invariant response. In the 3DRP algorithm, these missing sinograms are estimated by forward-projecting an initial 2D reconstruction of the direct segment $i_\theta = 0$.

choice of the span S depends both on the SNR and on the radius of the FOV. Values between 3 (for high-statistics brain studies) and 9 (for low-count, whole-body studies) are standard.

When the radius of the FOV is large, more accurate interpolation is needed to reorganize the raw data into parallel projections according to Eq. (55), but the sampling pattern (Eq. (56)) can be kept.

3D Analytic Reconstruction by Filtered-backprojection

The Central Section Theorem

The central section theorem (Eq. 10) can be generalized to 3D, and states that

$$P(\vec{v},\vec{n}) = F(\vec{v}) \qquad \vec{v} \in \vec{n}^{\perp} \tag{58}$$

where

$$P(\vec{v},\vec{n}) = \int_{n^{\perp}} d\vec{s}\, p(\vec{s},\vec{n}) \exp(-2\pi i \vec{s} \quad \vec{v}) \tag{59}$$

is the 2D Fourier transform of a parallel projection and F is the 3D Fourier transform of the image. Note that as the integral in Eq. (59) is over the whole projection plane $\vec{n}^{\perp}$, the central section theorem is only valid for non-truncated parallel projections.

Geometrically, this theorem means that a projection of direction $\vec{n}$ allows the recovery of the Fourier transform of the image on the central plane orthogonal to $\vec{n}$ in 3D frequency space. A corollary is that the image can be reconstructed in a stable way from a set of non-truncated projections $\vec{n} \in \Omega \quad \subset S^2$ if and only if the set Ω has an intersection with any equatorial circle on the unit sphere S^2. This condition is due to Orlov [70]. The equatorial band $\Omega(\theta_{\max})$ in Eq. (52) satisfies Orlov's condition for any $\theta_{\max} > 0$.

The *direct 3D Fourier reconstruction* algorithm is a direct implementation of Eq. (58) [71]. This technique involves a complex interpolation in frequency space, and has not so far been used in practice. However, Matej [15] recently demonstrated a significant gain of reconstruction time compared to the standard FBP.

3D Filtered Backprojection

Following the same lines as for the 2D FBP inversion, Eq. (58) leads to a two-step inversion formula for a set of non-truncated 2D projections with orientations $\vec{n} \in \Omega$, where Ω is a subset of the unit sphere that satisfies Orlov's condition. The reconstructed image is a 3D backprojection

$$f(\vec{r}) = \int_{\Omega} d\vec{n}\, p^F(\vec{s} = \vec{r} - \cdot(\vec{r} \quad \vec{n})\vec{n},\vec{n}) \tag{60}$$

which, as in 2D, is the sum of the *filtered projections* p^F for all lines containing the point $\vec{r}$. The filtered projections are given by

$$p^F(\vec{s},\vec{n}) = \int_{\vec{n}^{\perp}} d\vec{s}'\, p(\vec{s}' ,\vec{n}) h_C(\vec{s}-\vec{s}',\vec{n}) \tag{61}$$

In this equation, the 2D convolution kernel $h_C(\vec{s})$ is the 2D inverse Fourier transform of the filter function due to Colsher [72]:

$$H_C(\vec{v},\vec{n}) = \{\int_{\Omega} d\vec{n}'\,\delta(\vec{v}\cdot\vec{n}')\}^{-1} = \frac{|\vec{v}|}{L_{\Omega}(\vec{v})} \qquad \vec{v} \in \vec{n}^{\perp} \tag{62}$$

where δ is the Dirac delta function, and $L_{\Omega}(\vec{v})$ is the arc length of the intersection between Ω and the great circle normal to $\vec{v}$ (Fig. 4.10). Orlov's condition ensures that $L_{\Omega}(\vec{v}) > 0$. An expression of this filter in terms of the variables v_s, v_u, ϕ, θ can be found in [72]. Like the ramp filter, Colsher's filter is proportional to the modulus of the frequency. In contrast to the 2D case, however, the filter depends on the angular part of the

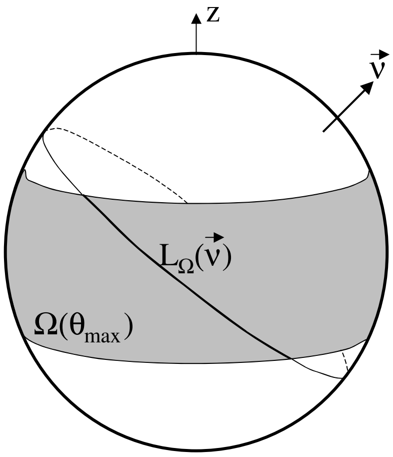

Figure 4.10. Each vector $\vec{n}$ on the unit sphere S^2 is the direction of one 2D parallel projection $p(\vec{s},\vec{n})$. The set of directions $\Omega(\theta_{max})$ measured by a cylindrical scanner (equation (52)) is shown as a grey subset. The Fourier transform $F(\vec{V})$ can be recovered from any projection along the measured (thick line) segment of the great circle orthogonal to $\vec{V}$. The reciprocal $1/L_{\Omega}$ of the length of this segment is the angular part of the reconstruction filter.

frequency. Another specificity of 3D reconstruction, due to the redundancy of the 3D data, is that the reconstruction filter is not unique [73]. Colsher's filter, however, yields the reconstructed image with the *minimal variance* under fairly general assumption on the data statistics [74].

The discretization of the 3D FBP algorithm is based as in 2D on replacing integrals by trapezoidal quadratures and on linear interpolation in $\vec{s}$ for the 3D backprojection. The 3D backprojection is the most time-consuming step in the algorithm and various techniques have been proposed to accelerate this procedure (see [17] and references therein). The 2D convolution is implemented in frequency space as:

$$p^F(\vec{s},\vec{n}) = \int_{\vec{n}^\perp} d\vec{\nu}\; h_C(\vec{\nu},\vec{n})\; w(\vec{\nu})\; P(\vec{\nu},\vec{n})\; \exp(2\pi i\vec{s}\cdot\vec{\nu}) \quad (63)$$

where $P(\vec{\nu}, \vec{n})$ is the 2D Fourier transform of the non-truncated projection and $w(\vec{\nu})$ is an apodizing window, which plays the same stabilizing role as in 2D (see the remark below Eq. (16) and reference [75] for details on the discrete implementation using the 2D FFT).

The Reprojection Algorithm

The 3D FBP algorithm is valid only for non-truncated parallel projections. In almost all PET studies, the tracer distribution extends axially over the whole FOV of the scanner, and the only non-truncated parallel projections are those with $\theta = 0$. For sampled data, the equality $\theta = 0$ is replaced by $\theta < \theta_0$ for some small maximum oblicity angle θ_0, which corresponds typically to the maximum ring difference $d_{2D,max}$ incorporated in a 2D acquisition.

The standard analytic reconstruction algorithm for volume PET scanners is the *3D reprojection algorithm* (3DRP) [76], which consists of four steps:

(i) Reconstruct a first image estimate $f_{2D}(\vec{r})$ by applying the 2D FBP algorithm to the non-truncated data subset $\theta < \theta_0$.
(ii) Forward project $f_{2D}(\vec{r})$ to estimate the unmeasured parts $p(\vec{s} \notin M(\vec{n}), \vec{n})$ of a set of 2D parallel projections $\vec{n} \in \Omega(\theta_{max})$,
(iii) Merge the measured and estimated data to form non-truncated projections,
(iv) Reconstruct these merged data with the 3D FBP algorithm described in the previous section.

In general, a value of θ_{max} smaller than the scanner maximum axial acceptance angle is used to limit the amount of missing data, which must be estimated and backprojected. With a 24-ring scanner, using $d_{max} = 19$ instead of the maximum value $N_r - 1 = 23$ still incorporates 95% of the data.

Images reconstructed with the 3DRP algorithm share many features with 2D FBP reconstructions, including linearity (the reconstructed FWHM in a given point is the same for a cold and for a hot spot) and the prevalence of streak artifacts in low-count studies. One difference with 2D reconstructions is the axial dependence of the spatial resolution, due to the increasing contribution of the estimated data near the edges of the axial FOV (see Fig. 4.9). This property of 3DRP reflects the non-uniform sensitivity of the volume PET scanner. Clearly, *any* analytic or iterative algorithm has to somehow reflect this property in the reconstruction. With the rebinning algorithms described below, the lower sensitivity in the edge slices is translated in an increased variance rather than in a degraded spatial resolution.

3D Analytic Reconstruction by Rebinning

The high sensitivity of a PET scanner operated in 3D mode is directly related to the large number of sampled LORs, which is much larger than the number of reconstructed pixels: $N_{LOR} >> P$ (by a factor proportional to N_r). We have already mentioned in the previous section that this data redundancy results in the non-uniqueness of the reconstruction filter. From the practical point of view, redundancy increases the data storage requirements and the computational load for reconstruction and data correction.

This observation has motivated the development of *rebinning* algorithms. A rebinning algorithm is an algorithm that estimates the ordinary sinogram (Eq. (3)) of each sampled transaxial section $z \in [0, L]$, i.e.

$$p_{reb}(s,\phi,z) = \int_{-\infty}^{\infty} dt\; f(x = s\cos\phi - t\sin\phi, y = s\sin\phi + t\cos\phi, z) \quad (64)$$

from the measured oblique sinograms $p_s(s, \phi, \zeta, \theta)$ defined by Eq. (54). Each rebinned sinogram is then reconstructed separately using a 2D reconstruction algorithm. This procedure is illustrated in Fig. 4.11.

Rebinning would be trivial for noise-free data because one easily checks by comparing Eqs. (54) and (3) that

$$p_{reb}(s,\phi,z) = p_s(s,\phi,\zeta = z,\theta = 0) \quad (65)$$

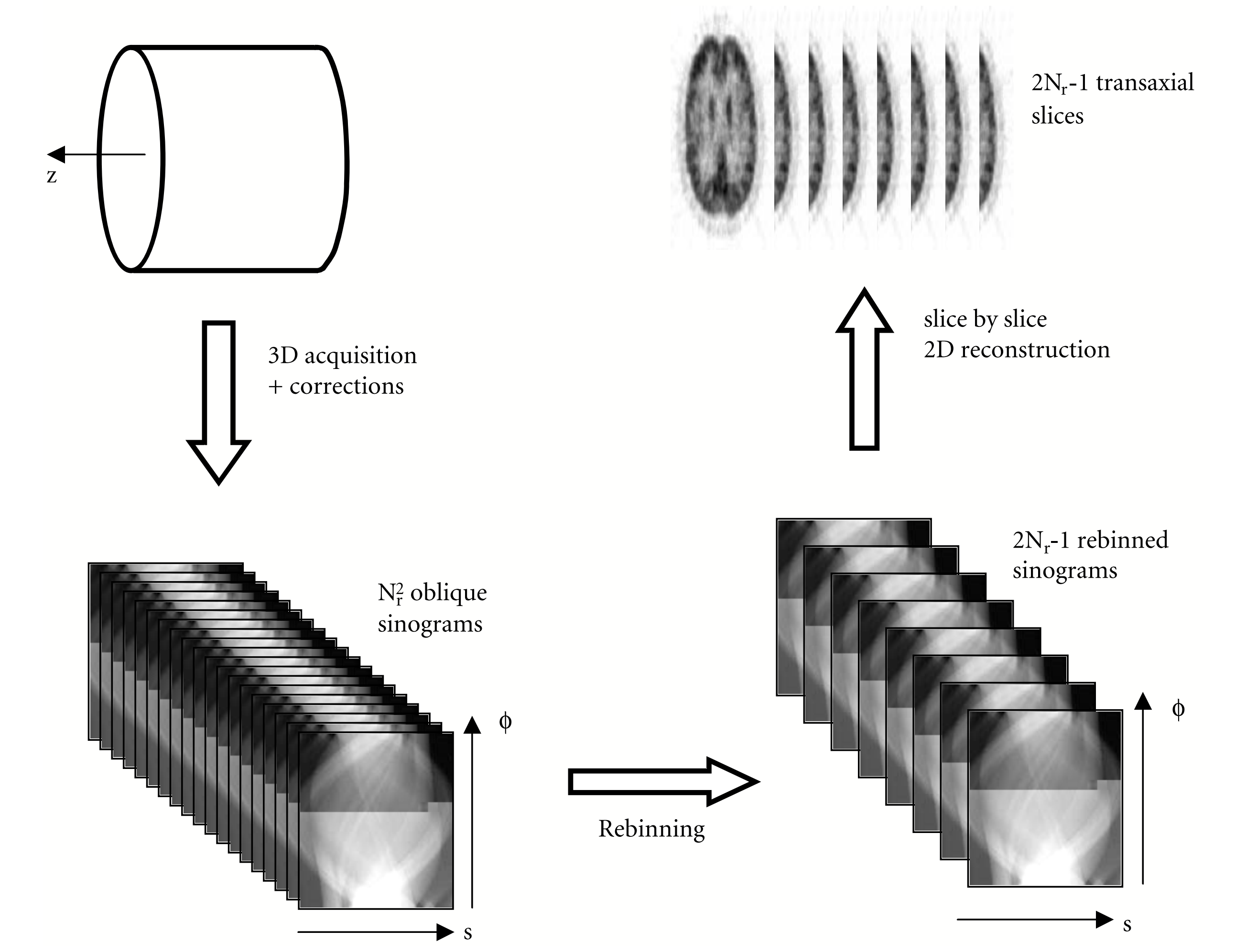

Figure 4.11. Schematic representation of the principle of a rebinning algorithm for 3D PET data.

In the presence of noise, however, an efficient rebinning method should optimize the SNR by exploiting the whole set of oblique sinograms to estimate p_{reb}.

Several approximate [77, 78, 79, 80, 81] and exact [82, 83] rebinning methods have been published. We only summarize the two algorithms that have been most used in practice.

The Single-slice Rebinning Algorithm (SSRB)

This approximate algorithm [77] is based on the assumption that each measured oblique LOR only traverses a single transaxial section within the support of the tracer distribution. Referring to the third argument of f in Eq. (54), this assumption amounts to neglecting the product R_F tan θ, where R_F , the radius of the FOV, is the maximum value of the variable t'. Using this approximation, Eq. (65) can be extended to

$$p_{reb}(s,\phi,z) \simeq p_s(s,\phi,\zeta = z,\theta = 0) \tag{66}$$

and by averaging all available estimates, SSRB defines the rebinned sinograms by

$$p_{ssrb}(\mathrm{s},\phi,\mathrm{z}) = \frac{1}{2\theta_{\max}(\mathrm{s},\mathrm{z})} \int_{-\theta_{\max}(s,z)}^{\theta_{\max}(s,z)} d\theta p_s(s,\phi,\zeta = z,\theta) \tag{67}$$

where $\theta_{\max}(s,z) = \arctan\left(\min\left[z, L-z\right] / \sqrt{R_d^2 - s^2}\right)$ is the maximum axial aperture for an LOR at a distance s from the axis in slice z. The algorithm is exact for tracer distributions which are linear in z, of the type $f(x, y, z) = a(x, y) + zb(x, y)$. For realistic distributions, the accuracy of the approximation will decrease with increasing R_F and $\theta_{\max}$. Axial blurring and transaxial distortions increasing with the distance from the axis of the scanner are the main symptoms of the SSRB approximation.

The discrete implementation of the SSRB algorithm is simply the extension of the technique described in the multi-slice 2D data section (above) to build 2D data with a multi-ring scanner operated in 2D mode, with d_{2Dmax} replaced by a larger value d_{max}. The choice of d_{max} entails a compromise between the systematic errors (which increase with d_{max}) and the reconstructed image variance (which increases with decreasing d_{max}).

The Fourier Rebinning Algorithm (FORE)

The approximate Fourier rebinning algorithm [81] is more accurate than the SSRB algorithm and extends the range of 3D PET studies that can be processed using rebinning algorithms. The main characteristics of FORE is that it proceeds via the 2D Fourier transform of each oblique sinogram, defined as

$$P_s(\nu,k,\zeta,\theta) = \int_0^{2\pi} d\phi \exp(-ik\phi) \int_{\mathbb{R}} ds \exp(-2\pi i s\nu) \times p_s(s,\phi,\zeta,\theta) \quad k \in Z, \nu \in \mathbb{R} \tag{68}$$

where k is the azimuthal Fourier index. Rebinning is based on the following relation between the Fourier transforms of oblique and direct sinograms:

$$P_s(\nu,k,z,0) \simeq P_s(\nu,k,\zeta = z + k \tan\theta / (2\pi\nu),\theta) \tag{69}$$

For each θ such that the oblique sinogram ζ, θ is measured (see Eq. (54)), the RHS yields an independent estimate of the direct data $\theta = 0$. FORE then averages all these estimates to optimize the SNR. The accuracy of the approximation (Eq. (69)) breaks down at low frequencies ν. Therefore, for all frequencies below some small threshold, the Fourier transform of the rebinned data is estimated using the SSRB approximation.

The main steps of the FORE algorithm are:

(i) Initialize a stack of Fourier transformed sinograms $P_{fore}(\nu, k, z)$,
(ii) For each oblique sinogram ζ, θ
  a. Calculate the 2D Fourier transform $P_s(\nu, k, \zeta, \theta)$,
  b. For each frequency component (ν, k), increment $P_{fore}(\nu, k, \zeta - k \tan\theta/(2\pi\nu))$ by $P_s(\nu, k, \zeta, \theta)$,
(iii) Normalize $P_{fore}(\nu, k, z)$ for the varying number of contributions it has received,
(iv) Take the 2D inverse Fourier transform to get the rebinned data $p_{fore}(s, \phi, z)$.

Like all analytic algorithms, FORE assumes that the data p_s (s, ϕ, ζ, θ) are line integrals of the tracer distribution and that each oblique sinogram is sampled over the whole range $(s, \phi) \in [-R_F, R_F\ ]\ x\ [0, \pi]$. Therefore, the raw data must be corrected for all effects including detector efficiency variations, attenuation, and scattered and random coincidences, before applying FORE. Also, when the data are incomplete due to gaps in the detector assembly, the sinograms must be filled as discussed in the section on properties of the inverse 2D radon transform (above). Refer to [81] for a detailed description and for the derivation of FORE.

In practice, FORE is sufficiently accurate when the axial aperture θ_{max} is smaller than about 20°, though the limit depends on the radius of the FOV and on the type of image. Beyond 20°, artifacts similar to those observed with SSRB (at lower apertures) appear [84]: degraded image quality at increasing distance from the axis. Two variations of FORE, the FOREJ and FOREX rebinning algorithms [82, 83], are exact in the limit of continuous sampling, and have been shown to overcome this loss of axial resolution when reconstructing high statistics data acquired with a large aperture scanner [85]. However, the current implementation of the FOREJ algorithm [82] is more sensitive to noise than FORE since the correction term involves a second derivative of the data with respect to the axial coordinate ζ, and the application to low statistics data remains questionable.

Hybrid Reconstruction Algorithms for 3D PET

The future evolution of image reconstruction in PET will most probably lead to the generalized utilization of iterative algorithms, both for 2D and for 3D data. As shown in the next section, it is straightforward to extend iterative methods, such as OSEM, to fully 3D scanning. These algorithms have the potential to model accurately the data acquisition, the measurement noise, and also the prior information on the tracer distribution. In contrast, analytic algorithms are bound to the line integral representation of the data. Even though some physical effects can be incorporated in pre- or post-processing steps, an accurate modeling of the Poisson statistics of the data is difficult with analytic methods. To date, however, the computational burden of fully 3D iterative algorithms remains a major issue for some applications involving multiple acquisitions, or for research scanners such as the HRRT which sample a very large number of LORs. The current practice of undersampling these data (see above) to accelerate reconstruction is contradictory with the aim of accurate modeling claimed by iterative methods.

This limitation has led to the application of *hybrid algorithms* for 3D PET data [41, 66, 91]. These algorithms first rebin the 3D data into a multi-slice set of ordinary sinogram data, using e.g. the SSRB method, or, more often, FORE. Each rebinned sinogram is then reconstructed using some 2D iterative algorithm. This hybrid approach provides a significant time gain with respect to fully 3D iterative reconstruction.

The two components of hybrid algorithms, rebinning and iterative methods, have been discussed in previous sections. In this section, we briefly discuss the interplay between these two elements, the main difficulty being to model the rebinned data that are presented to the 2D iterative algorithm. We focus on the application of FORE followed by a 2D OSEM reconstruction but the same problems would arise with other combinations, such as SSRB followed by an iterative minimization of a 2D penalized weighted least-square (PWLS) cost function [86].

One of the major benefits of iterative reconstruction arises from a correct modeling of the data statistics, which allows to weight each LOR according to its variance. This is the reason why improved image quality is obtained by reconstructing the raw, uncorrected data with a system matrix incorporating the effects of attenuation, normalisation and scatter, rather than by reconstructing pre-corrected data with a system matrix modeling only the detector's geometric response. Ideally, therefore, we would like to develop a hybrid algorithm in which un-corrected rebinned data are reconstructed by means of a 2D iterative algorithm including the effects of attenuation, etc. This approach is impossible because the FORE Eq. (69) must be applied to fully pre-corrected data as discussed at the end of the previous section. The rebinned data must then be reconstructed with a 2D iterative algorithm which does *not* model the pre-corrected physical effects.

One solution to improve the statistical model is to *de-correct* the data for the physical effects after the rebinning. This de-correction restores Poisson-like statistics to the rebinned data, and the physical effects can then be reintroduced in the system matrix. If we hypothesize that the most important effect is that of attenuation, we can decorrect for attenuation only and then reconstruct the de-corrected rebinned data with AW-OSEM (see Eq. (43)). This approach is referred to as the FORE+OSEM(AW) algorithm. Note that this algorithm is still approximate: even in the absence of attenuation and scatter, the rebinned sinograms are not independent Poisson variables because of the complex linear combination of the 3D data during FORE rebinning. Strictly speaking, it is inappropriate to reconstruct the rebinned data using the OSEM algorithm derived for independent Poisson data, and it is preferable to use a weighted least-square method [87] or the NEC scaling technique [30] (Eq. (44)). In each case, one needs to estimate the variance of the rebinned data [88] and also, ideally, the covariance [89].

Finally, modeling the shift-variant detector response of the data. Several implementations have been described for 3-D data, based on the Space Alternating Generalized EM [93], on ML-EM and OSEM [94], on Bayesian estimation [26, 95], and on the row-action maximum likelihood [90]. All algorithms of the ML-EM type described above, for instance, can be readily generalized to 3D PET by replacing the system matrix $a_{j,i}$ describing the acquisition geometry (equation (29)) by its 3D equivalent, which takes into account the axial coordinates of the LORs. For block-iterative methods such as OSEM (see Eq. (42)), the set of LORs parameterized by the two transaxial sinogram indices s_k, ϕ_j in Eq. (5) and by the two axial coordinates i_θ, i_z in Eq. (56) must be divided into subsets. Most implementations simply subdivide the azimuthal index ϕ_j, exactly as in the 2D case. Each subset then contains all axial samples i_θ, i_z.

The benefit expected from fully 3D iterative reconstruction is easily demonstrated for scanners with large polar aperture, particularly in the presence of gaps. Fig. 4.12, for example, shows a high resolution phantom measured with the HHRT brain scanner. The bottom image was reconstructed with FORE+OSEM (AW), while the top one was reconstructed with OSEM3D(ANW), where "ANW" indicates that both the normalization and attenuation corrections are incorporated in the system matrix. The horizontal streak artifacts in the coronal section of the FORE+OSEM(AW) image are attributed to the gap filling step prior to FORE. Blurring can also be observed on the 3 brightest rods at the edge of the cylinder. When the polar angle is smaller, as with many clinical scanners, the bias introduced by FORE is small, and the benefit of 3D reconstruction is harder to visualize, especially at

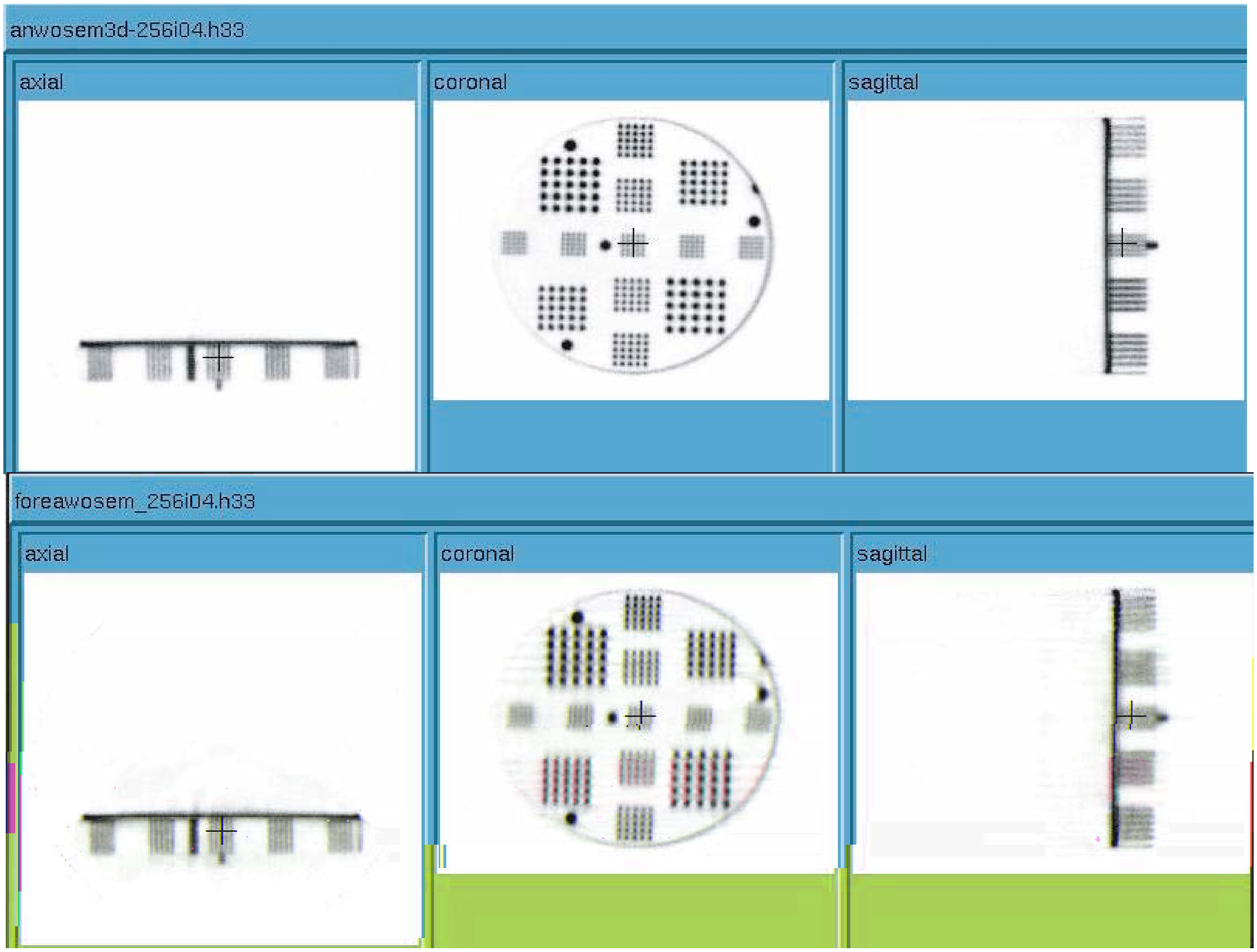

Figure 4.12. High resolution phantom data acquired on the HRRT: comparison of a fully 3D iterative reconstruction using OSEM3D(ANW) (top) and of a hybrid reconstruction with FORE+OSEM(AW) (bottom). Both images were reconstructed with 4 iterations and 16 subsets. The phantom is oriented vertically in the FOV of the scanner and the vertical axis on the coronal section is parallel to the axis of the scanner (courtesy K. Wienhard, Köln).

low count statistics and when regularization is achieved by post-reconstruction smoothing. This is illustrated by Fig. 4.13, which shows whole-body patient data processed with FORE+OSEM(AW) and OSEM3D(ANW). Note the similarity between the two reconstructions, even in regions with high attenuation (shoulder and neck for this patient with arms up).

In contrast with the algorithms illustrated above, the fully-3D image reconstruction developed by Leahy et al. [24, 26, 95, 96] is based on an extensive system model. The algorithm incorporates a shifted Poisson model that includes the statistics of true, scattered and random coincidences, as well as positron range, annihilation photon acolinearity, attenuation, sinogram sampling, detector dead-time and efficiency, block detector effects, and the spatially varying detector resolution due to parallax (depth of interaction) and Compton scatter in the scintillators (Chapter 2). Although the size of the system matrix is reduced using a factorized model and by taking advantage of symmetries, the computation time is necessarily longer than with simplified system models. This lead us to considerations of the potential for parallel-processing of image reconstructions on processor arrays.

Parallel Implementation of Iterative Reconstruction

The need for parallel implementation of the ML-EM algorithm was already recognized in the mid-eighties. Pioneering work proposed the use of a cluster of commodity PCs [97] or dedicated hardware [98]. But as soon as commercial parallel systems became available, dedicated algorithms were developed on high-end computers such as transputers [99, 100], hypercubes [101], meshes [102], rings [103], fine-grain message-

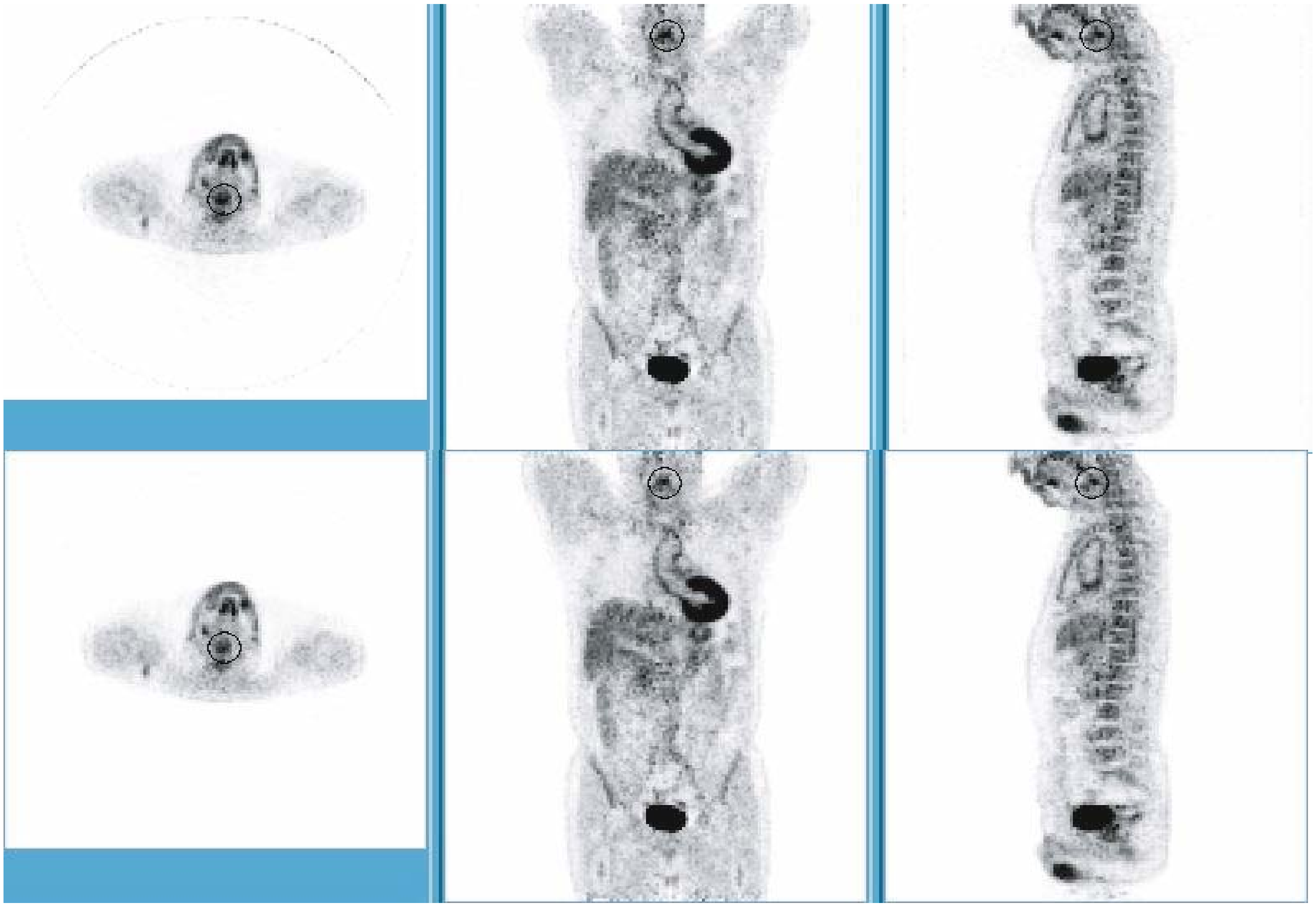

Figure 4.13. Whole-body FDG scan on an HR+ tomograph, reconstructed using FORE+OSEM(AW) (top) and OSEM3D(ANW) (bottom), in both cases with 4 iterations and 16 subsets. A 3D gaussian filter with FWHM 4 mm was applied after reconstruction. The orthogonal views are passing through the cursor (small circle in the neck area).

passing machines [104], linear arrays of DSPs [105] to cite a few examples. Recent efforts concentrated on using clusters of multi-processor PCs, sometimes called component off the shelf (COS), and combine both shared and distributed memory approaches. This choice is dictated by the cost/performance ratio of the hardware, by its flexibility and by the possibility to upgrade the system with faster and cheaper hardware in this very competitive market. One key problem in distributed computing is to optimize the balance between computation and communication amongst the nodes. The ultimate goal is to keep individual processors busy all the time by interleaving I/O and computation. A good measure of the performance of a parallel algorithm is how well the speed-up factor scales linearly with the number of nodes. In their work, Shattuck et al [106] describe a parallel implementation of the MAP-PCG reconstruction [26] using a master-slave model with 9 dual PC nodes. The work of Vollmar [107] describes a parallel extension of the OSEM3D reconstruction [92] and is also using a masterslave model with 7 quad PC nodes. By calculating the system matrix on the fly and neglecting the physics of the detection system these authors could handle very large reconstruction problems on the HRRT. The HRRT scanner acquires generally data in span 3(9) with a maximum ring difference of 67, which generates 3D data of 983 MB (326MB). The work of Jones et al [108] is another parallel extension of the OSEM3D reconstruction [92]. It uses a single program multiple-data (SPMD) rather than a master-slave model. These authors have shown that image space decomposition (ISD) and projection space decomposition (PSD) were roughly equivalent since the communication burden was large at forward projection when using ISD but was also large at backprojection when using PSD. However, by developing an efficient I/O subsystem and reorganizing the data, these authors finally favored the PSD model [109]. The performance of this parallel implementation of OSEM3D was shown to scale relatively well up to 16 nodes (32 processors). A commercial implementation of this computing cluster uses 8 nodes of dual Pentium 4 Xeon at 3.0 Ghz, and performs one iteration of OSEM3D in about 20 min for a 3D sinogram set of 983 MB and an image size of 256×256×207 (27 MB). Finally, the PARAPET initiative, currently known as the STIR project [110], has developed a generic, multi-platform and multi-scanner, implementation of OSEM3D using an object-oriented library [111, 112]. The parallel implementation uses a master-slave model and a PSD scheme. On a 12-node Parsytec CC system it provides a factor 7 speed-up compared to serial mode.

Acknowledgment

This work has been supported in part by NIH Grant CA-74135 and by the grant G.0174.03 of the FWO (Belgium).

References

1. Virador PRG, Moses WW, Huesman RH. Reconstruction in PET cameras with irregular sampling and depth of interaction capability. IEEE Trans Nucl Sc 1998;NS-45:1225–30.
2. Natterer F. The mathematics of computerized tomography. New York: Wiley, 1986.
3. Townsend DW, Isoardi RA, Bendriem B. Volume imaging tomographs. In: Bendriem B, Townsend DW (eds) The theory and practice of 3D PET, Dordrecht: Kluwer Academic, pp 111–32, 1998.
4. Natterer F,Wuebbeling F. Mathematical methods in image reconstruction. SIAM Monographs on Mathematical Modeling and Computation, 2001.
5. Kak AC, Slaney M. Principles of computerized tomographic imaging. New York: IEEE Press, 1988.
6. Barrett HH, Swindell W. Radiological Imaging. New York: Academic Press, 1981.
7. Barrett HH, Myers KJ, Foundations of Image Science. New York: Wiley, 2004.
8. Ramm AG, Katsevich AI. The radon transform and local tomography. New York: CRC Press, 1996.
9. Quinto ET. Exterior and limited-angle tomography in nondestructive evaluation. Inverse Problems 1998;14:339–53.
10. Wienhard K et al. The ECAT HRRT: Performance and first clinical application of a new high-resolution research tomograph. In: Records of the 2000 IEEE Medical Imaging Symposium, Lyon, France.
11. Karp JS, Muehllehner G, Lewitt RM. Constrained Fourier space method for compensation of missing data in emission-computed tomography. IEEE Trans Med Imag 1988;MI-7:21–5.
12. de Jong, H, Boellaard R, Knoess C, Lenox M, Michel C, Casey M, Lammertsma A. Correction methods for missing data in sinograms of the HRRT PET scanner. IEEE Trans Nucl Sc 2003;NS-50:1452–1456.
13. O'Sullivan JD. A fast sinc function gridding algorithm for Fourier inversion in computer tomography. IEEE Trans Med Imag 1985;MI-4:200–7.
14. Schomberg P, Timmer J. The gridding method for image reconstruction by Fourier transforms. IEEE Trans Med Imag 1995;MI-14:596–607.
15. Matej S, Lewitt RM. 3D FRP: Direct Fourier reconstruction with Fourier reprojection for fully 3D PET. IEEE Trans Nucl Sc 2001;NS-48:1378–85.
16. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes in C, Cambridge University Press 1992.
17. Egger ML, Joseph C, Morel VC. Incremental beamwise backprojection using geometrical symmetries for 3D PET reconstruction in a cylindrical scanner geometry. Phys Med Biol 1998; 43:3009–24.
18. Brady ML. A fast discrete approximation algorithm for the Radon transform. SIAM J Comp 1998;27:107–19.
19. Brandt A, Mann J, Brodski M, Galun M. A fast and accurate multilevel inversion of the Radon transform. SIAM J Appl Math 1999;60:437–62.
20. Bertero M, Boccacci P. Introduction to inverse problems in imaging. Bristol: IOP Publishing, 1998.
21. Huesman RH. The effects of a finite number of projection angles and finite lateral sampling of projections on the propagation of statistical errors in transverse section reconstruction. Phys Med Biol 1977;22:511–21.

22. J Fessler, 2001 NSS/MIC statistical image reconstruction short course notes, www.eecs.umich.edu/ fessler/papers/talk.html
23. Leahy R, Qi J. Statistical approaches in quantitative positron emission tomography. Statistics and Computing 2000;10:147–65.
24. Qi J, Leahy RM, Cherry SR, Chatziioannou, Farquhar TH. High-resolution 3D Bayesian image reconstruction using the micro-PET small animal scanner. Phys Med Biol 1998;43:1001–13.
25. Yavuz M, Fessler J. New statistical models for random precorrected PET scans. In: Duncan J, Gindi G (eds), Information Processing in Medical Imaging, 15th International Conference, Berlin: Springer, pp 190–203, 1998.
26. Qi J, Leahy R, Hsu C, Farquhar T, Cherry S. Fully 3D Bayesian image reconstruction for ECAT EXACT HR+. IEEE Trans Nucl Sci 1998;NS-45:1096–1103.
27. Lange K, Carson R. EM reconstruction algorithms for emission and transmission tomography. J Comp Ass Tomo 1984;8:306–16.
28. Mumcuoglu EU, Leahy R, Cherry SR, Zhou Z. Fast gradient-based methods for Bayesian reconstruction of transmission and emission PET images. IEEE Trans Med Im 1994;MI-13:687–701.
29. Erdogan H, Fessler JA. Accelerated monotonic algorithms for transmission tomography. In Proc IEEE Intl Conf on Image Processing 1998;2:6804.
30. Nuyts J, Michel C, Dupont P. Maximum-likelihood expectation maximization reconstruction of sinograms with arbitrary noise distribution using NEC transformations. IEEE Trans Med Imag 2001;MI-20:365–75.
31. Bai C, Kinahan P, Brasse D, Comtat C, Townsend D. Postinjection single photon transmission tomography with ordered-subset algorithms for whole-body PET. IEEE Trans Nucl Sci 2002;NS-49:74–81.
32. Lewitt RM. Alternatives to voxels for image representation in iterative reconstruction algorithms. Phys Med Biol 1992;37:705–16.
33. Matej S, Herman GT, Narayanan TK, Furie SS, Lewitt RM, Kinahan PE. Evaluation of task-oriented performance of several fully 3D PET reconstruction algorithms. Phys Med Biol 1994; 39:355–67.
34. Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Patt Anal Mach Intel 1984;PAMI-6:721–41.
35. Mumcuoglu EU et al. Bayesian reconstruction of PET images, methodology and performance analysis. Phys Med Biol 1996; 41:1777–807.
36. Leahy RM, Yan X. Incorporation of anatomical MR data for improved functional imaging with PET. In: Information Processing in Medical Imaging: 12th International Conference, Berlin: Springer-Verlag, pp 105–20, 1991.
37. Gindi G, Lee M, Rangarajan A, Zubal G. Bayesian reconstruction of functional images using anatomical information as priors. IEEE Trans Med Imag 1993;12:670–80.
38. Bowsher JE, Johnson VE, Turkington TG, Jaszczak RJ, Floyd Jr CE, Coleman RE. Bayesian reconstruction and use of anatomical a priori information for emission tomography. IEEE Trans Med Imag 1996;MI-15:673–86.
39. Comtat C, Kinahan P, Fessler F, Beyer T, Townsend D, Defrise M, Michel C. Clinically feasible reconstruction of whole-body PET/CT data using blurred anatomical labels. Phys Med Biol 2002;47:1–20.
40. Baete K, Nuyts J, Van Paesschen W, Suetens P, Dupont P. Anatomical based FDG-PET reconstruction for the detection of hypo-metabolic regions in epilepsy. To appear in IEEE Trans Med Imag 2004.
41. Obi T, Matej S, Lewitt RM, Herman GT. 2.5D simultaneous multislice reconstruction by iterative algorithms from Fourier-rebinned PET data. IEEE Trans Med Imag 2000;MI-19:474–84.
42. Nichols TE, Qi J, Leahy RM. Continuous time dynamic pet imaging using list mode data. In: Colchester A et al. (eds), Information Processing in Medical Imaging, 12th International Conference, Berlin: Springer, pp 98–111, 1999.
43. Asma E, Nichols TE, Qi J, Leahy RM. 4D PET image reconstruction from list-mode data. Records of the 2000 IEEE Medical Imaging Symposium, Lyon (France), paper 15–57.
44. Fessler J. Penalized weighted least-square image reconstruction for PET. IEEE Trans Med Imag 1994;MI-13:290–300.
45. Lalush D, Tsui B. A fast and stable maximum a posteriori conjugate gradient reconstruction algorithm. Med Phys 1995; 22:1273–1284.
46. Leahy R, Byrne C. Recent developments in iterative image reconstruction for PET and SPECT. IEEE Trans Med Imag 2000;19:257–60.
47. Lange K, Hunter D, Yang I. Optimization Transfer algorithms using surrogate objective functions. J Comp Graph Stat 2000; 9:1–59.
48. Daube-Witherspoon ME, Muehllehner G. An iterative image space reconstruction algorithm for volume ECT. IEEE Trans Med Imag 1986;MI-5:61–6.
49. De Pierro A. On the relation between the ISRA and the EM algorithm for positron emission tomography. IEEE Trans Med Imag 1993;MI-12:328–33.
50. Levitan E, Herman GT. A maximum a posteriori expectation maximization algorithm for image reconstruction in emission tomography. IEEE Trans Med Imag 1987;MI-6:185–92.
51. Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 1977;39:1–38.
52. Shepp L A, Vardi Y. Maximum likelihood reconstruction for emission tomography. IEEE Trans Med Imag 1982;MI-1:113–22.
53. Barrett HH, White T, Parra L. List-mode likelihood. J Opt Soc Am 1997;14:2914–23.
54. Reader A, Erlandsson K, Flower M, Ott R. Fast, accurate, iterative reconstruction for low-statistics positron volume imaging. Phys Med Biol 1998;43:835–46.
55. Levkovitz R, Falikman D, Zibulevsky M, Ben-Tal A, Nemirovski A. The design and implementation of COSEM, an iterative algorithm for fully 3D list-mode data. IEEE Trans Med Imag 2001;MI-20:633–42.
56. Huesman RM, Klein G, Moses W, Qi J, Reutter B, Virador P. List-mode maximum-likelihood reconstruction applied to positron emission mammography (PEM) with irregular sampling. IEEE Trans Med Imag 2000;MI-19:532–7.
57. Barrett HH, Wilson DW, Tsui BMW. Noise properties of the EM algorithm: I. Theory. Phys Med Biol 1994;39:833–46.
58. Wilson DW, Tsui BMW, Barrett HH. Noise properties of the EM algorithm: II. Monte Carlo simulations. Phys Med Biol 1994; 39:847–71.
59. De Pierro AR, Yamagishi MEB. Fast EM-like methods for maximum “a posteriori” estimates in emission tomography. IEEE Trans Med Imag 2001;MI-20:280–8.
60. Veklerov E, Llacer J. Stopping rule for the MLE algorithm based on statistical hypothesis testing. IEEE Trans Med Imag 1987;MI-6:313–19.
61. Selivanov V, Lapointe D, Bentourkia M, Lecomte R. Cross-validation stopping rule for ML-EM reconstruction of dynamic PET series: effect on image quality and quantitative accuracy. IEEE Trans Nucl Sc 2001;NS-48:883–9.
62. Hudson H, Larkin R. Accelerated image reconstruction using ordered subsets of projection data. IEEE Trans Med Imag 1994;MI-13:601–9.
63. Browne J, De Pierro AR. A row action alternative to the EM algorithm for maximizing likelihoods in emission tomography. IEEE Trans Med Imag 1996;MI-15:687–99.
64. Byrne CL. Convergent block-iterative algorithms for image reconstruction from inconsistent data. IEEE Trans Imag Proc 1997;IP-6:1296–304.
65. Hebert T, Leahy RM. Fast methods for including attenuation in the EM algorithm. IEEE Trans Nucl Sc 1990;NS-37:754–8.
66. Comtat C, Kinahan PE, Defrise M, Michel C, Townsend DW. Fast reconstruction of 3D PET with accurate statistical modeling. IEEE Trans Nucl Sc 1998;NS-45:1083–9.
67. Stayman J W, Fessler J A. Fast methods for approximating the resolution and covariance for SPECT, in IEEE Nuclear Science and Medical Imaging Conference; 2002, Norfolk, VA, paper M3-146.
68. Stayman JW, Fessler JA. Regularization for uniform spatial resolution properties in penalized-likelihood image reconstruction. IEEE Trans Med Imag 2000;MI-19:601–615.
69. Nuyts J, Fessler JA. A penalized-likelihood image reconstruction method for emission tomography, compared to post-smoothed

maximum-likelihood with matched spatial resolution. IEEE Trans Med Imag 2003;MI-22:1042–1052.

70. Orlov SS. Theory of three-dimensional reconstruction. 1. Conditions of a complete set of projections. Sov Phys Crystallography 1976;20:429–33.

71. Stearns CW, Chesler DA, Brownell GL. Accelerated image reconstruction for a cylindrical positron tomograph using Fourier domain methods. IEEE Trans Nucl Sci 1990;NS-37:773–7.

72. Colsher JG. Fully three-dimensional PET. Phys Med Biol 1980; 25:103–115.

73. Defrise M, Townsend DW, Clack R. Image reconstruction from truncated two-dimensional projections. Inverse Problems 1995;11:287–313.

74. Defrise M, Townsend DW, Deconinck F. Statistical noise in three-dimensional positron tomography. Phys Med Biol 1990; 35:131–138.

75. Stearns CW, Crawford CR, Hu H. Oversampled filters for quantitative volumetric PET reconstruction. Phys Med Biol 1994;39:381–8.

76. Kinahan PE, Rogers JG. Analytic three-dimensional image reconstruction using all detected events. IEEE Trans Nucl Sci 1990;NS-36:964–8.

77. Daube-Witherspoon ME, Muehllehner G. Treatment of axial data in three-dimensional PET. J Nucl Med 1987;28:1717–24.

78. Erlandsson K, Esser PD, Strand S-E, van Heertum RL. 3D reconstruction for a multi-ring PET scanner by single-slice rebinning and axial deconvolution. Phys Med Biol 1994;39:619–29.

79. Tanaka E, Amo Y. A Fourier rebinning algorithm incorporating spectral transfer efficiency for 3D PET. Phys Med Biol 1998; 43:739–46.

80. Lewitt RM, Muehllehner G, Karp JS. Three-dimensional reconstruction for PET by multi-slice rebinning and axial image filtering. Phys Med Biol 1994;39:321–40.

81. Defrise M, Kinahan PE, Townsend DW, Michel C, Sibomana M, Newport DF. Exact and approximate rebinning algorithms for 3D PET data. IEEE Trans Med Imag 1997;MI-16:145–58.

82. Defrise M, Liu X. A fast rebinning algorithm for 3D PET using John's equation, Inverse Problems 1999;15:1047–65.

83. Liu X, Defrise M, Michel C, Sibomana M, Comtat C, Kinahan PE, Townsend DW. Exact rebinning methods for three-dimensional positron tomography. IEEE Trans Med Imag 1999;MI-18:657–64.

84. Matej S, Karp JS, Lewitt RM, Becher AJ. Performance of the Fourier rebinning algorithm for 3D PET with large acceptance angles. Phys Med Biol 1998;43:787–97.

85. Michel C, Hamill J, Panin V, Conti M, Jones J, Kehren F, Casey M, Bendriem B, Byars L, Defrise M. FORE(J)+OSEM2D versus OSEM3D reconstruction for Large Aperture Rotating LSO, In: IEEE Nuclear Science and Medical Imaging Conference; 2002, Norfolk, VA, paper M7-93.

86. Kinahan PE, Fessler JA, Karp JS. Statistical image reconstruction in PET with compensation for missing data. IEEE Trans Nucl Sc 1997;NS-44:1552–7.

87. Stearns CW, Fessler JA. 3D PET reconstruction with FORE and WLS-OS-EM. In: IEEE Nuclear Science and Medical Imaging Conference; 2002, Norfolk, VA, paper M5-5.

88. Janeiro L, Comtat C, Lartizien C, Kinahan P, Defrise M, Michel C, Trébossen R, Almeida P. NEC-scaling applied to FORE+OSEM. In: IEEE Nuclear Science and Medical Imaging Conference; 2002, Norfolk, VA, paper M3-71.

89. Alessio A, Sauer K, Bouman C. PET statistical reconstruction with modeling of axial effects of FORE. In: IEEE Nuclear Science and Medical Imaging Conference; 2003, Portland, OR, paper M11-194.

90. Daube-Witherspoon ME, Matej S, Karp JS, Lewitt RM. Application of the Row Action Maximum Likelihood Algorithm with spherical basis functions to clinical PET imaging. IEEE Trans Nucl Sc 2001;NS-48:24–30.

91. Kinahan PE, Michel C, Defrise M, Townsend DW, Sibomana M, Lonneux M, Newport DF, Luketich JD. Fast iterative image reconstruction of 3D PET data. In: IEEE Nuclear Science and Medical Imaging Conference; 1996, Anaheim, CA, 1918–22.

92. Liu X, Comtat C, Michel C, Kinahan PE, Defrise M, Townsend DW. Comparison of 3D reconstruction with 3D OSEM and with FORE+OSEM for PET. IEEE Trans Med Imag 2001;MI-20:804–14.

93. Ollinger J. Maximum likelihood reconstruction in fully 3D PET via the SAGE algorithm. In Proc. 1996 IEEE Nucl. Sci. Symp. Medical Imaging Conf., Anaheim, CA, 1594–1598.

94. Johnson C, Seidel S, Carson R, Gandler W, Sofer A, Green M, Daube-Witherspoon M. Evaluation of 3-D reconstruction algorithms for a small animal PET camera. IEEE Trans Nucl Sci 1997;NS-44:1303–1308.

95. Qi J, Leahy RM. Resolution and noise properties of MAP reconstruction for fully 3D PET. IEEE Trans Med Imag MI-19:493–506.

96. Bai B, Li Q, Holdsworth C, Asma E, Tai Y, Chatziioannou A, Leahy R. Modelbased normalization for iterative 3D PET image reconstruction. Phys Med Biol 2002;47:2773–2784.

97. Llacer J and Meng J. Matrix-based image reconstruction methods for tomography. IEEE Trans Nucl Sci 1985; NS-32:855–864.

98. Jones W, Byars L, Casey M. Design of a super fast three-dimensional projection system for positron emission tomography. IEEE Trans Nucl Sci 1990;NS-37:800–804.

99. Barresi S, Bollini D, Del Guerra A. Use of a transputer system for fast 3-D image reconstruction in 3-D PET. IEEE Trans Nucl Sci 1990;NS-37:812–816.

100. Atkins S, Murray D, Harrop R. Use of transputers in a 3-D positron emission tomograph. IEEE Trans Med Imag 1991; MI-10:276–283.

101. Chen C.-M., Lee S.-Y., Cho Z. Parallelization of the EM algorithm for 3-D PET image reconstruction. IEEE Trans Med Imag 1991;MI-10:513–522.

102. Rajan K, Patnaik L, Ramakrishna J. High-speed computation of the EM algorithm for PET image reconstruction. IEEE Trans Nucl Sci 1994;NS-41:1721–1728.

103. Johnson C, Yan Y, Carson R, Martino R, Daube-Witherspoon M. A system for the 3-D reconstruction of retracted-septa PET data using the EM algorithm. IEEE Trans Nucl Sci 1995; NS-42:1223–1227.

104. Cruz-Rivera J, DiBella E, Wills D, Gaylord T, Glytsis E. Parallelized formulation of the maximum likelihood expectation maximization algorithm for fine-grain message-passing architectures. IEEE Trans Med Imag 1995;MI-14:758–762.

105. Rajan K, Patnaik L, Ramakrishna J. Linear array implementation of the EM reconstruction algorithm for PET image reconstruction. IEEE Trans Nucl Sci 1995;NS-42:1439–1444.

106. Shattuck D, Rapela J, Asma E, Chatzioannou A, Qi J, Leahy R. Internet2-based 3-D PET image reconstruction using a PC cluster. Phys Med Biol 2002;47:2785–2795.

107. Vollmar S, Michel C, Treffert J, Newport D, Casey M, Knoss C, Wienhard K, Liu X, Defrise M, Heiss W-D. HeinzelCluster: accelerated reconstruction for FORE and OSEM3D. Phys Med Biol 2002;47:2651–2658.

108. Jones J, Jones W, Kehren F, Newport D, Reed J, Lenox M, Baker K, Byars L, Michel C, Casey M. SPMD cluster-based parallel 3D OSEM. IEEE Trans Nucl Sci 2003;NS-50:1498–1502.

109. Jones J, Jones W, Kehren F, Burbar Z, Reed J, Lenox M, Baker K, Byars L, Michel C, Casey M. Clinical Time OSEM3D: Infrastructure Issues. In: IEEE Nuclear Science and Medical Imaging Conference; 2003, Portland, OR, paper M10-244.

110. The current home page of the PARAPET/STIR project is http://stir.sourceforge.net/homepage.shtm

111. Bettinardi V, Pagani E, Gilardi M-C, Alenius S, Thielemans K, Teras M, Fazio F. Implementation and evaluation of a 3D One Step Late reconstruction algorithm for 3D Positron Emission Tomography studies using Median Root Prior. Eur J Nucl Med 2002;29:7–18.

112. Jacobson M, Levkovitz R, Ben Tal A, Thielemans K, Spinks T, Belluzzo D, Pagani E, Bettinardi V, Gilardi MC, Zverovich A, Mitra G. Enhanced 3D PET OSEM reconstruction using interupdate Metz filtering. Phys Med Biol 2000;45:2417–2439.