Modelling data from single photon emission computed tomography

Iain S. Weir and Peter J. Green,
Department of Mathematics,
University of Bristol,
Bristol, BS8 1TW, UK.

Abstract

The completely random character of radioactive disintegration provides the basis of a strong justification for a Poisson linear model for SPECT data, that can be used to produce reconstructions of isotope densities, whether by maximum likelihood or Bayesian methods. However, such a model requires the construction of a matrix of weights, which represent mean rates of arrival at each detector of photons originating from each point within the body space. Two methods of constructing these weights are discussed, and resulting reconstructions from phantom and real data are presented.

Keywords: attenuation, EM algorithm, gamma camera, one step late, radiopharmaceutical, scatter.

1 Introduction

The *gamma camera* is an imaging device finding increasing use in modern medical interpretation. It is incapable of the high spatial resolution of anatomic detail possible with digital X–ray and magnetic resonance imaging, but it has an increasingly important role through its ability to study function rather than form.

A radioactively labelled substance is introduced into the patient, usually though injection or inhalation. The substance is chosen as one known to become concentrated in the organ of interest, in a manner related to the phenomenon under study, and of course the labelling allows the pattern of concentration to be observed non-invasively. Photon emission occurs in the organ, at a rate varying spatially according to the concentration, and indirect measurements of this concentration can thus be made by counting emitted photons in the gamma camera.

The physical and operational properties of this instrument are described by Larssen (1980). Incoming photons have to traverse the narrow parallel bores in a lead collimator; those unable to do so are absorbed. The photons successful in passing through impact on a fluorescing crystal, causing a fluorescence whose location and energy are measured by means of photo-multiplier tubes and electronic circuitry.

The array of photon counts forms a digital image, which can be displayed "raw" for visual interpretation, or post-processed to extract quantitative information. In its application to single photon emission computerised tomography (SPECT), the camera is swung around the patient to collect a sequence of two-dimensional projections, at equally spaced angles. The object of *reconstruction* is to extract a three-dimensional representation of the pattern of isotope concentration from this sequence of projections.

The data consist of an array of counts of detected photons, which we denote by $\{y_t, t \in T\}$. Note that t denotes a (projection, pixel) pair. The appropriate statistical model for these data is essentially that which applies to all emission data (see, for example, Shepp and Vardi (1982)). If $x(\mathbf{u})$ denotes the isotope concentration at spatial location $\mathbf{u} \in \mathcal{R}^3$, and $a_t(\mathbf{u})$ the mean rate of arrival at detector location t of photons emitted at $\mathbf{u}$, per unit concentration at $\mathbf{u}$, then we find

$$y_t \sim \text{Poisson}\left(\int a_t(\mathbf{u})x(\mathbf{u})\mathrm{d}\mathbf{u}\right) \tag{1.1}$$

independently for all $t \in T$. This model captures the principal physical effects of absorption and scattering in the body (due to nuclear interactions), and the collimation, all of which lead to the thinning expressed by $a_t(\mathbf{u})$, while the integral corresponds to the superposition of emissions from different point sources $\mathbf{u}$.

Although we intend in later work to explore the possibility of working with this model directly, here we follow the usual approach of discretising x and hence a to obtain

$$y_t \sim \text{Poisson}\left(\sum_s a_{ts}x_s\right). \tag{1.2}$$

Here, x_s is the true total radiopharmaceutical concentration within pixel s, and a_{ts} corresponds to $a_t(\mathbf{u})$ for $\mathbf{u}$ in the centre of pixel s. These weights a_{ts} depend on various factors: the geometry of the detection system, the strength of the isotope and exposure time, and the extent of attenuation and scattering between source and detector. Early numerical experiments on reconstruction algorithms typically used simulated data, and so the weights a_{ts} were rather arbitrarily assigned, and assumed known. In analysing real data, the weights must be constructed, and this must be done with some care. The purpose of this paper is to propose two methods to construct these weights. The first method utilises only knowledge of the geometric/physical properties of the SPECT system whilst the second method is empirical.

The problem of how best to estimate $\{x_s\}$ from $\{y_t\}$, given the model above, and prior contextual information about the probable true pattern of isotope concentration shall not be considered here. Shepp and Vardi (1982), in pioneering work aimed specifically at positron emission tomography (PET), proposed using the EM algorithm for the MLE solution of model (1.2). However, MLE solutions have the unattractive feature of being very noisy in appearance, because of the high dimensionality of the parameter vector x. Bayesian reconstruction from SPECT data was proposed by Geman and McClure (1985), following the general paradigm of Bayesian image analysis introduced by Geman and Geman (1984), and Besag (1986). This involves the construction of a second model which is a probability distribution on the space of true patterns of isotope concentration. Reconstruction can then be accomplished by considering the posterior distribution that follows from Bayes' theorem on the two model components and leads to far more satisfactory results. Geman and McClure used Iterated Conditional Modes (ICM) and Markov chain Monte Carlo (MCMC) methods in this work. Green (1990) proposed the one step late (OSL) algorithm as an adaptation of the EM algorithm to obtain such reconstructions. Weir (1993) has found the algorithm more usable than MCMC methods because of its comparative computational cheapness. For a more detailed discussion on the algorithm and the types of prior distributions used for the SPECT application, see Green (1990).

2 Geometric/physical modelling

In this section, the aim is to use simple physical and geometrical arguments to derive values for the weights $\{a_{ts}\}$ that are sufficiently accurate for reasonable reconstruction of x, without the need for auxiliary transmission experiments. The values will involve only known physical constants and easily measured dimensions, and do not depend on the data obtained during the scanning of the patient.

Recall that a_{ts} represents the mean number of photons recorded at bin t, per unit concentration of the isotope at pixel s. It is therefore influenced by a number of factors:

(a) strength of activity of isotope
(b) exposure time
(c) rate of radioactive decay
(d) rate of elimination of the radiopharmaceutical from the organ
(e) spatial correspondence between detector and pixel
(f) detector geometry
(g) absorption of photons
(h) photon scattering.

Note that (d), (g) and (h) depend on the patient, and so in the absence of auxiliary data can at best be approximated. In this section, we shall neglect (a), (d) and (h). The effect is to assume that elimination is negligible within the duration of the data acquisition, that scattering can be neglected, and finally means that our reconstruction has an arbitrary scale: some external calibration is needed to express the reconstructed x in physical units. The remaining factors are accounted for by assuming that each a_{ts} is the product of three factors:

(i) the proportion of radioactivity that has not decayed by the time at which photons are collected in detector t,
(ii) the proportion of emissions that survive attenuation,
(iii) the solid angle of view of the centre of pixel s into detector t, which is treated as a cylindrical tube of known length and radius.

These considerations apply for three-dimensional reconstruction from a sequence of two-dimensional projections. Our numerical examples will address the usual smaller problem obtained by ignoring the third dimension: the spatial component parallel to the axis of rotation. We thus reconstruct each section or slice of the patient, perpendicular to this axis, from the one-dimensional projections corresponding to that slice. We shall now derive expressions for each of these three factors.

2.1 Radioactive decay

The rate of emission declines appreciably during the period in which the data is collected so a decay term is included. Typically, for a twenty minute scan using Technetium ($^{99}\mathrm{Tc}^{m}$), the rate is reduced to 95% of its initial value. Since the half-life of the radioactive isotope is generally known, the adjustment required to compensate for the amount of decay reduces to a trivial calculation. Let N_0 be the number of nuclei at the start of the imaging session and assume that the exposure time for each of the n projections is the same. Since we are dealing with very large numbers of nuclei, we can neglect random variation, and so the number of nuclei, N_i, at the time of acquisition of projection i is given, approximately, by:

$$N_i \quad = \quad N_0 \exp(-\frac{(i-1)T}{n\tau} \log 2) \tag{2.1}$$

where τ is the half-life and T is the total acquisition time. If the exposure times of different projections are not constant then, providing the various times are known, the extra calculation is minor.

2.2 Attenuation

It is well known that attenuation operates multiplicatively (see Larssen, 1980), so that the proportion of photons surviving attenuation along a line L is of the form

$$\exp\left(-\int_L \mu(\mathbf{u})\mathrm{d}\mathbf{u}\right) \tag{2.2}$$

The linear attenuation rate μ varies markedly with the medium, and in particular has different values in bone, soft tissue, and air. The approximation we will make is that μ has a constant value within the patient's body, and zero elsewhere. Further, the body is approximated as a cylinder with elliptical cross-section, parallel with the tomographic axis. The dimensions of the ellipse can be easily measured from the patient, or estimated from a trial reconstruction from the data, presuming that the body outline can be readily distinguished. In our experience five EM iterations are sufficient to determine the approximate centre and length of axes of the ellipse. Let the ellipse lie in the X–Y plane with its axes aligned on the x and y axes. Denote the coordinates of the centre by (x_c, y_c) and the length of the axes by a and b. Then the attenuation distance from a point (x, y) to a detector at (x_D, y_D) can be found by examining where the line between the two cuts the ellipse and from this calculating the distance travelled through the attenuation region. If we parameterise the line as $(x + du, y + dv)$ where $u = x_D - x$, $v = y_D - y$ and $u^2 + v^2 = 1$, then we need to solve the quadratic in d given by the expression

$$\left(\frac{x + du - x_c}{a}\right)^2 + \left(\frac{y + dv - y_c}{b}\right)^2 \quad = \quad 1 \tag{2.3}$$

Referring to Figure 1, the signs of the two roots are relevant in determining the attenuation distance, d, namely

1. no roots: $d = 0$,
2. one positive and one negative: $d =$ positive root,
3. two positive: $d =$ absolute value of their difference,
4. two negative: $d = 0$.

The proportion of emissions that survive attenuation is then given by $\exp(-\mu d)$ where the value of μ is known.

2.3 Solid angle of view

The detectors are modelled as cylindrical tubes, admitting particles proportionally to the solid angle of view through the cylinder, as viewed from the source pixel. Referring to Figure 2, this depends only on:

(i) distance, t, of pixel from tube axis ,
(ii) distance, s, along axis between tube and perpendicular from pixel,
(iii) radius, q, and half length, p, of tube.

By elementary trigonometrical arguments, it can be shown that the area of overlap of two discs of radius α, with centres 2θ apart, as a fraction of the area of one of the discs, is $f(\theta/\alpha)$ where

$$f(x) = \begin{cases} (2/\pi)(\cos^{-1}(x) - x\sqrt{1-x^2}) & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

Now if q and t are small compared with p and s respectively, the two ends of the detector tube appear from the pixel approximately as equal circular discs, so the solid angle of view into the tube, as a fraction of the solid angle subtended by one end of the tube, is close to $f(\theta/\alpha)$ where $\theta = t/s$ and $\alpha = q/p$.

It is necessary to account for the third dimension, as emissions outside a slice may be registered within that slice. The approximation made here is that successive slices are sufficiently similar that they can be assumed equal: x_s therefore denotes the concentration throughout a narrow prism parallel to the axis of rotation. The solid angle above is then replaced by its integral over the third dimension:

$$\int_{-\infty}^{\infty} f(\sqrt{\theta^2+\phi^2}/\alpha)\mathrm{d}\phi = 2\int_0^{\sqrt{\alpha^2-\theta^2}} f(\sqrt{\theta^2+\phi^2}/\alpha)\mathrm{d}\phi \tag{2.5}$$

There is no analytic expression for this integral, but it can be evaluated numerically by Simpson's rule. The shape appeared to be gaussian so for convenience a scaled gaussian curve was fitted. The integral in (2.5) and the corresponding fitted gaussian curve, are shown in Figure 3. The approximation seemed sufficiently close not to cause concern, in view of the assumptions made thus far.

In use in an expression for the weights, this curve is scaled by s^{-2} to account for the dependence of solid angle on distance.

2.4 Storage of coefficients

Calculation of our approximate weights thus reduces to derivation of some simple but tedious trigonometrical formulae, and substitution into these of a dozen or so physical constants and measurements. Notice that since the patient will not typically be perfectly centred on the tomographic axis, there are no symmetries among the $\{a_{ts}\}$ that can be exploited to save storage or computing time, even under the simple model described above.

We generally reconstruct on a 48 x 48 grid from 64 projections each of 52 counts. With these dimensions, only about 10 percent of the weights $\{a_{ts}\}$ are nonzero, and so a sparse matrix representation is used: the weights are stored columnwise together with pointers to the detectors concerned.

2.5 Allowing for scatter

In this section we have effectively ignored the phenomenon of photon scatter. It is appropriate to make a few comments about this.

The phenomena of absorption and scatter are closely related. Every interaction experienced by a photon during its flight produces both a loss of energy and a deflection to its path. The energy loss is random, and determines the angle of deflection; these are quantified by the Klein-Nishina formula (Evans (1955), chapter 23). Since photons that have experienced a large deflection are much less informative about their source, and large deflections correspond to large loss of energy, it is standard practice to adjust the gamma camera electronics so as to record only photons arriving with energy (measured subject to error) above a threshold, chosen to admit only photons suffering negligible deflection. A more complete statistical model would allow for this scattering, and thus make more use of the (unthresholded) data. However, such a model would carry a heavier computational burden, as it would mean that *all* the weights $\{a_{ts}\}$ would be nonzero.

In future work, Green and Professor M. Hudson intend to develop models providing a good approximation for the scattering phenomenon that are computationally feasible.

3 Empirical modelling

An alternative to the analytic model for the weights of the previous section can be constructed empirically. This involves the design of an experiment to allow the depth-dependent and machine-specific effects of attenuation and scatter to be measured directly. Geman, Manbeck and McClure (1991b) propose a model that is consistent with the main physical and geometric effects and backed up by experimental evidence. A set of experiments are devised to enable the development of the model's form and subsequent estimation of its required parameters. In this section the model is reviewed and fitted to experimental data resulting from repetitions of the experiments on the General Electric machine at the Bristol Royal Infirmary (BRI). The model is extended in order to more accurately reflect the underlying physical/geometric process, and other forms of the model are quantitatively compared to see if the fit to the data could be improved.

3.1 Model of Geman, Manbeck and McClure

Geman *et al.* (1991b) model the a_{ts} through a detection density whose integral over the width of bin y_t is proportional to a_{ts}. The density, α, is modelled as a function of

1. perpendicular distance, s, from the centre of pixel x_s to the camera containing bin y_t,
2. distance, d, along the perpendicular from the centre of x_s to the boundary of the imaged object,
3. distance, z, along the camera surface from the perpendicular projection of x_s

(see Figure 4). Thus,

$$a_{ts} = k \int_{t_1}^{t_2} \alpha(d, s, z) \mathrm{d}z \tag{3.1}$$

where k is a scale parameter dependent on the choice of physical units. The model utilises only 2D information and ignores radioactive decay. The latter can simply be included by multiplying by a decay factor as in Section 2.1.

The density is assumed to be comprised of two additive parts relating to a *direct* contribution from detected photons that have not been scattered and a *scatter* contribution from those that have. Each of the terms is factored into depth dependent attenuation terms depending on d and a residual line spread function which integrates to one with respect to z. The attenuation region is assumed uniform with known attenuation coefficient, μ. The form of the model components is deduced from the set of experiments repeated in Section 3.2. The final chosen model utilises normal distributions for both the direct and scatter line spread functions. The standard deviations linearly increase with the distances s and d respectively. The weighting between the two is determined by the depth-dependent function $C\exp(-\nu d)$ involving the two parameters C and ν. Thus,

$$\begin{aligned} \alpha(d,s,z) &= e^{-\mu d}\frac{1}{\sqrt{2\pi\sigma_{di}^2(s)}}e^{-z^2/2\sigma_{di}^2(s)} + \\ & \quad (Ce^{-\nu d} - e^{-\mu d})\frac{1}{\sqrt{2\pi\sigma_{sc}^2(d)}}e^{-z^2/2\sigma_{sc}^2(d)}, \end{aligned} \tag{3.2}$$

where

$$\sigma_{di}(s) = m_{di}s + b_{di}, \tag{3.3}$$
$$\sigma_{sc}(d) = m_{sc}d + b_{sc}. \tag{3.4}$$

In the above, there are six parameters to be estimated from the experimental data (C, ν, m_{di}, b_{di}, m_{sc} and b_{sc}), together with the scale parameter, k, from equation (3.1).

3.2 Experiment

Independent experiments were conducted based on those of Geman *et al.* In these a line source is made from filling a pipette with $^{99}\mathrm{Tc}^{\mathrm{m}}$. A cylindrical tank is placed on its side with the line source stretching from the top at one end to the bottom at the other as in Figure 5. A gamma camera placed horizontally directly above and parallel to the cylinder's axis is used to record the photons released as the isotope decays for a set period of time. This is done first with air in the tank and then with water. The air medium is chosen since its low density means essentially that photons do not get scattered or absorbed. It is therefore possible to isolate the effects of the camera geometry. However, the denser medium of water will result in significant amounts of absorption and scatter. Water is used to approximate the assumed uniform attenuation region since it has similar attenuation characteristics to most of the human body.

The data were collected by a Starport 400AT camera in a 64x64 array and then trimmed to 35 rows and 60 columns (see Figure 6). Figure 7 shows the histograms of counts for row numbers 1, 17 and 35 (corresponding to the two end and the middle rows) for both media. The first column is for the air data where the fanning out of the distributions with distance can be seen. The second column is for the water data which clearly reveals the scatter component of the proposed model. Note the corresponding increase in spread of the data with distance for both the direct and scatter contribution. Also, it can be seen that the ratio of scattered to direct counts increases with attenuation distance.

The estimation method proposed by Geman *et al.* begins by using solely the air data to model the direct line spread function. As there is assumed to be no scattering, the row totals approximate the scaled integration with respect to z of the direct term of equation (3.2), namely $k\exp(-\mu d)$. It is observed that there is no apparent decrease in the row totals, thus confirming the assumption of negligible attenuation (i.e. $d \approx 0$). The row totals therefore can be used to estimate k. The shape of the counts within each row is deemed to be well fit by Gaussian curves whose standard deviations increase linearly with the distance s. Regression is used to obtain estimates of m_{di} and b_{di}.

The water data row totals are observed to decay exponentially with distance. The scaled integral of equation (3.2) is $kC\exp(-\nu d)$. Thus, an exponential curve fitted to the water data row totals together with the estimate of k, can be used to find estimates of C and ν.

The parameters estimated so far are used to correct the water data to isolate the effect of the scatter line spread distribution. After this correction, the rows are deemed well fit by Gaussian curves whose standard deviations increase linearly with the distance d. Again, regression is used to obtain the parameters m_{sc} and b_{sc}.

For the BRI data the ad hoc nature of this systematic estimation method is found to be unsatisfactory, see Weir (1993). In particular, after correction of the water data, some rows of data are left with negative counts. The resulting parameter estimates are displayed in Table 1.

3.3 Estimation of model parameters using maximum likelihood

In this section the method of maximum likelihood is used to obtain the model parameters. The method not only has the advantage of simultaneous estimation of all parameters, but also produces estimates with the desirable properties of maximum likelihood. The log-likelihood of the Poisson data is

$$L = \sum_t (y_t \log \mu_t - \mu_t) \tag{3.5}$$

where μ_t is the expected value of the datum y_t. If we set the concentration value of the isotope (arbitrarily) to one, then the expected value of the count in bin t is

$$\begin{aligned} \mu_t = k \int_{t_1}^{t_2} \Bigg[& e^{-\mu d} \frac{1}{\sqrt{2\pi\sigma_{di}^2(s)}} e^{-z^2/2\sigma_{di}^2(s)} + \\ & (Ce^{-\nu d} - e^{-\mu d}) \frac{1}{\sqrt{2\pi\sigma_{sc}^2(d)}} e^{-z^2/2\sigma_{sc}^2(d)} \Bigg] \mathrm{d}z. \end{aligned} \tag{3.6}$$

The maximisation of the likelihood is subject to several constraints on its parameters. Since the scatter contribution has to be positive this necessitates $C \geq 1$. If we assume that the ratio of scatter to direct counts increases with attenuation distance, then this leads to $0 < \nu < \mu$. The scale parameter, k is required to be positive and we also have the linear constraints, due to the standard deviations of the direct and scatter components, that for all s and d

$$m_{di}s + b_{di} > 0 \tag{3.7}$$

$$m_{sc}d + b_{sc} > 0. \tag{3.8}$$

The problem of choosing a method for maximising the log-likelihood needs to be addressed. The complication of the constraints mean that the conventional methods of Newton-Raphson or Fisher scoring will not be adequate.

The choice of maximisation algorithm is dictated by whether or not the first and second partial derivatives of the log-likelihood are used. Let θ_j and θ_k be two of the seven parameters to be estimated. The derivatives have the form

$$\frac{\partial L}{\partial \theta_j} = \sum_t \frac{\partial L}{\partial \mu_t}\frac{\partial \mu_t}{\partial \theta_j} \tag{3.9}$$

$$= \sum_t \left(\frac{y_t}{\mu_t} - 1\right)\frac{\partial \mu_t}{\partial \theta_j} \tag{3.10}$$

$$\frac{\partial^2 L}{\partial \theta_j \partial \theta_k} = \sum_t \left[\frac{\partial L}{\partial \mu_t}\frac{\partial^2 \mu_t}{\partial \theta_j \partial \theta_k} + \frac{\partial^2 L}{\partial \mu_t^2}\frac{\partial \mu_t}{\partial \theta_k}\frac{\partial \mu_t}{\partial \theta_j}\right] \tag{3.11}$$

$$= \sum_t \left[\left(\frac{y_t}{\mu_t} - 1\right)\frac{\partial^2 \mu_t}{\partial \theta_j \partial \theta_k} - \frac{y_t}{\mu_t^2}\frac{\partial \mu_t}{\partial \theta_k}\frac{\partial \mu_t}{\partial \theta_j}\right] \tag{3.12}$$

$$(\text{since } \partial^2 L/\partial \mu_t \partial \mu_s = 0 \text{ if } s \neq t)$$

Expressions for the derivatives of μ_t that are needed are straightforward but tedious to obtain, and will be omitted here. It is not immediately clear that the performance of a maximisation algorithm will be improved sufficiently to offset the extra computational burden caused by including partial derivatives. We therefore compare the following three algorithms

1. Quasi-Newton algorithm, no partial derivatives required,
2. modified-Newton algorithm, first partial derivatives required,
3. modified-Newton algorithm, first and second partial derivatives required.

Choice of initial values can be based on previous estimation or chosen from sensible guesses incorporating knowledge of the data. Initial values for the standard deviation parameters can be obtained by letting $m_{di} = m_{sc} = 0$ and setting b_{di} and b_{sc} equal to the calculated standard deviations from the air and water data respectively. The average row total of the air data can be used for the initial value for k. This leaves the choice of C and ν: two possibilities are $C = 1$, $\nu = \mu$ and $C = 2$, $\nu = \mu/2$.

Regardless of starting values with the BRI data, only the Quasi-Newton algorithm successfully finds a maximum each time. These maxima are consistent and higher than the values returned by the other two algorithms. If the other two algorithms are started with slightly perturbed values of the maximum, then they too successfully converge to the found maximum. The performance of the algorithms can be further explored using simulated data. If the data are set to the expected values, the true values are found by all three algorithms regardless of initial parameter values. However, with simulated Poisson data from the expected means, it is again found that the Quasi-Newton algorithm consistently finds a maximum, the other two algorithms being sensitive to the closeness of the starting value to the solution.

The maximum that is found gives an unsatisfactory parameter estimate, in that m_{sc} has a negative value, implying that the spread of scatter decreases with attenuation distance. The reason for this is that the attenuation distance for the air data is assumed to be zero. However, close inspection of the air data reveals that there is a small amount of depth dependent scatter caused by the sides of the acrylic tank. In order to use the air data, this phenomenon would need to be accounted for. However, as this process would be complex, we

	Geman's	MLE
deviance	43918.4	16580.7
k	21594.4	28590.4
m_d	0.01079	0.01599
b_d	0.4045	0.2533
m_s	0.0994	0.1126
b_s	2.1194	2.8031
C	1.00947	1.0000
ν	0.1082	0.1423

Table 1: Comparison of model parameter estimates

simply ignore the air data and calculate the MLE solely from the water data. This presents no problems as the air data are only required for the estimation procedure of Geman *et al.*

The MLEs calculated using only the water data are given in Table 1. The coefficient of m_{sc} is now positive as expected. The only concern now is that, as C is estimated as 1, the solution is on the boundary of the feasible region allowed by the constraints. This suggests that the form of the model may need to be reviewed. In the next section we shall explore modifications to the model.

Comparison of the estimates obtained from the two estimation methods reveals substantial differences in the parameter estimates. The discrepancy of the fits can be compared using the deviance function

$$\text{Deviance} = 2\sum_t \left[y_t \log(y_t/\mu_t) - (y_t - \mu_t)\right], \tag{3.13}$$

where, in order for the comparison to be valid, the summation is just over the water data. The substantially lower value for the maximum likelihood method indicates that it is a much better fit. This can be clearly seen in the resulting fitted count densities (Figure 8).

The histograms of the data (Figure 7) show that with respect to the distributions, the counts are coarsely grouped. With this in mind, an alternative set of estimates can be obtained by considering the counts to come from a multinomial distribution whose probabilities are given by equation 3.6. The MLE can then be calculated using iteratively reweighted least squares (see Green (1984)).

3.4 Modifying the model

In attempt to obtain estimates that are strictly within the feasible parameter region, Weir (1993) proposed modifications to the Geman *et al.* model which were sequentially accepted or rejected. In this section the modifications are reviewed and the effect of the introduction of each of them is examined.

3.4.1 Changes to the detection density

The histograms of the water data (Figure 7) indicate quite clearly that even with a mixture of two model components, there remain a substantial number of outliers that appear to be uniformly distributed. Possible sources for these counts are detector misreadings or background photon counts. Their presence may be very influential in the parameter estimation of the direct and scatter effects. For this reason the detection density may be extended by the addition of a third term that is uniform over the field of view. That is,

$$\alpha^*(d,s,z) \quad = \quad \alpha(d,s,z) + u. \tag{3.14}$$

So far we have only considered fitting the line spread functions that Geman *et al.* considered to fit their data well. In an earlier report (1991a), they suggested that a double exponential distribution

$$\frac{1}{2\sigma}\exp\left(-\frac{|z|}{\sigma}\right), \qquad -\infty < z < \infty, \tag{3.15}$$

was appropriate for the scatter component. The scale parameter σ again increases linearly with attenuation distance. As the distribution has longer tails than the Gaussian, it might cope more successfully with the apparent outliers that have already been noted.

3.4.2 Accounting for the third dimension/correcting the attenuation distance

We shall now extend the model to recognise the fact that counts in a row can come from neighbouring slices. In Figure 9 the distances for the detection density are demonstrated from a point on the line source to a detector. Previously, the integration of the variable, z, was simply over the bin's width. However, the amount of radiation received by a particular bin is now related to its orientation with respect to the direction defined by z. As a first approximation, we shall still integrate over the bin's width.

The attenuation distance used by Geman *et al.* is only correct for a detector that is centred around the perpendicular from the camera surface to x_s. This distance either over or under-estimates the true distance depending on the relative position of x_s and the attenuation region. As we have already seen in Section 2.2, we need to use the distance travelled through the attenuation region along a straight line from the centre of x_s to the centre of the detector y_t. Since the weights are calculated once and then stored, there seems to be no advantage in using the approximate distance.

The attenuation distance from a general point on the line source (x_l, y_l, z_l) to the detector at (x_D, y_D, z_D) can be deduced from firstly noting that as the cylinder is parallel to the z axis, the parameterisation of Section 2.2 can be used to find the x and y coordinates of the intersection of the line between the two with the cylinder. Having obtained this, the z coordinate can be found using the line parameterisation

$$\frac{x_i - x_D}{x_l - x_D} = \frac{y_i - y_D}{y_l - y_D} = \frac{z_i - z_D}{z_l - z_D} \tag{3.16}$$

where x_i, y_i, z_i are the respective intersection points. There is a possibility that z_i will be outside the range corresponding to the length of the cylinder. If this is the case, then the above parameterisation can be used to find x_i, y_i after having substituted the relevant z cylinder end coordinate for z_i. The attenuation distance is given by the Euclidean distance between (x_l, y_l, z_l) and (x_i, y_i, z_i).

The expected value of the count in bin t is given by the integral of the detection density over the length of the line source. Let q be the bin width and note that the distances d,s, and z are functions of the position on the line source. Then

$$\mu_t \quad = \quad k\int_{l_1}^{l_2}\int_{t_1}^{t_2} \alpha(d(l), s(l), z(l))\mathrm{d}z\mathrm{d}l, \tag{3.17}$$

direct	normal		dble exponential	
scatter	normal	dble exp	normal	dble exp
deviance	5925.0	6032.9	7731.7	9253.2

Table 2: Deviances for extended model line source combinations

where

$$t_1 = z(l) - q/2, \tag{3.18}$$
$$t_2 = z(l) + q/2. \tag{3.19}$$

and l_1, l_2 are the two end points of line source. The expected value of the count in bin t will be evaluated numerically using Simpson's rule.

3.4.3 Results

The extended model was fitted using combinations of the normal and double exponential line spread functions for the direct and scatter counts. The resulting deviances are displayed in Table 2. Since we have non-nested models, it would be usual to use Mallows' C_p or Akaike's information criterion to compare the fit. These take into account not only the deviance but also the number of parameters used. However, as each model has the same number of parameters, we can simply use the deviance to assess the comparative measure of fit. The normal/normal combination has the lowest deviance indicating the best fit of the four. The improvement of the fit compared to the previous model can be evaluated by noting we have nested models. Thus, a F-statistic for the reduction in deviance could be calculated. However, it is immediately apparent from the magnitude of the reductions that the changes to the model are very significant. The improvement in fit can be clearly seen by examining the fitted densities displayed in Figure 10.

The parameter estimates for the chosen model are displayed in Table 3. These have the pleasing property of being inside the feasible parameter space region as the estimated value of C is now greater than one. Thus the fully extended model no longer gives any reason to question the model's form. The estimates of error in the parameters can be calculated from the diagonal elements of the inverted Hessian matrix. These are given in Table 3. Note, if the solution is still on the parameter space boundary, then the calculation of error estimates is not so trivial (see Self and Liang (1987)).

3.5 Storage of coefficients

The third dimension is incorporated into the weight calculation using the same assumption as in Section 2.3. We therefore need to integrate the detection density numerically along a line through the centre of pixel x_s parallel to the axis of rotation. The weights, a_{ts}, are thus given by equation (3.17) where the limits of integration, l_1 and l_2, now represent the end points of the body along the line of integration. Note that $s(l)$ is constant along this line. The attenuation distance, $d(l)$, is calculated assuming the body has elliptical cross-section. Thus, in contrast to Section 2.3, the attenuation is considered to be a function of the third dimension. The weights are then multiplied by a decay factor.

parameter	estimate	std error
k	765514.1	2272.3
m_{di}	0.02542	3.495e-04
b_{di}	0.07113	6.140e-03
m_{sc}	0.1999	3.470e-03
b_{sc}	1.2941	3.556e-02
C	1.0049	1.208e-03
ν	0.1797	1.341e-04
u	2.043e-06	7.765e-08

Table 3: Extended normal/normal model parameter and standard error estimates

The geometrical/physical weights have a cut-off point that limits the number of non-zero coefficients. However, due to the infinite tail of the normal distribution and the uniform third term, all the empirical coefficients are non-zero. This, together with the dimension of the system, leads to a vast increase in the storage space required and the amount of computational work, which is in direct proportion to the number of coefficients. In order to make reconstructions practical, it is necessary to have an approximation to the full matrix that reduces the computational burden and, in addition, permits a sparse matrix representation. Let A^+ be the matrix whose entries are the direct and scatter contributions when $z(l)$ is less than a pre-determined multiple of their respective standard deviations and let A^- contain the remaining contributions. Also, let U^+ be the matrix whose elements are all equal to the uniform term, u, and U^- be the adjustment for decay of the matrix U^+. Then the matrix of coefficients, A, can be decomposed into the four additive parts

$$A = A^+ + A^- + U^+ - U^-. \tag{3.20}$$

If the contribution from A^- and U^- is assumed to be small then

$$A \approx A^+ + U^+. \tag{3.21}$$

Sparse matrix techniques can now be used since only the non-zero elements of A^+ need to be stored. Despite effectively still having the full number of non-zero coefficients, the inclusion of the uniform third term causes very little extra computation.

4 Reconstructions

In this section we present reconstructions obtained using coefficients calculated from both approaches. In addition, we include those obtained by *filtered back-projection* which is the standard reconstruction technique in use today. The latter does not use a statistical model but rather operates entirely linearly on the data. Two choices of cut-off points for the empirical coefficients are presented. The first is the geometrical/physical one which allows a comparison between the two methods when the number of coefficients is the same. The second truncates the normal distributions of the direct and scatter contributions by restricting their respective ranges to three and one standard deviations. This results in approximately three times as many coefficients. The data and filtered back-projection reconstructions were supplied by the BRI.

4.1 Phantom data

The phantom consists of a cylindrical acrylic tank into which different acrylic inserts may be placed.

1. The "hot" lesion insert consists of eight pairs of holes drilled through the block, each pair separated by a distance equal to the hole diameters. The diameter of each pair is 25% larger than that of the preceding pair. The eight pairs form a V-shape either side of which are drilled two larger holes. The 18 holes form hot spots of the radio-pharmaceutical which simulates brain and bone scans.

2. The "cold" lesion insert is made up of seven plastic rods where each rod is 25% larger in diameter than the preceding one. Above these are seven spheres with the same respective diameters, each positioned so that their centres lie in the same plane. Cold lesions in a hot radioactive field simulate lung and liver scans.

3. The "linearity" insert is a crossed grid of channels cut out in a square pattern.

When the tank is fitted with the inserts and filled with a radio-pharmaceutical, there is also a section above the inserts that forms a region of uniform activity. Since the tank was in fact the container used in the line source experiment, the reconstructions from the phantom data do not involve any extrapolation of the model.

In Figure 11 we compare the four different reconstructions for the hot lesion insert. In the filtered back-projection reconstruction, the larger of the hot spots is visible but these are not as well defined as in the reconstructions from the statistical model. Note the presence of spurious radial artifacts, one of the well-documented deficiencies of the method. When using the same number of coefficients, the empirical method is slightly better at isolating the hot spots than the geometrical/physical approach. The increase in the number of empirical coefficients improves the contrast between the hot and cold areas.

The difference between the statistical reconstructions is less pronounced for the cold lesion insert and the area of uniform activity. Figure 12 displays a slice through the rods where only the three largest are clearly identifiable. None of the spheres above the rods can be identified. The uniform region reconstructions are very similar. However, the improvement over filtered back-projection for each of these slices is evident.

Filtered back-projection does cope well with the linearity insert (Figure 13). This is probably due to the fact that the channels are naturally well aligned with the back-projections. However, the contrast is again not as good as in the statistical reconstructions.

4.2 Real data

The data relates to horizontal slices through the brain and pelvis. The reconstructions of the brain (Figure 14) clearly demonstrate the inadequacy of filtered back-projection. The statistical reconstructions reveal a clear isotope pattern that is very similar in each case. The pelvis reconstructions (Figure 15) again reveal little difference between the statistical reconstructions whilst offering a vast improvement over filtered back-projection.

5 Discussion

It is clear that the use of a statistical model improves the quality of SPECT reconstructions. The calculation of the model's weights by geometrical/physical reasoning only requires knowledge of easily obtained physical constants and machine specific measurements. The empirical coefficients, however, require auxiliary experimentation. Maximum likelihood has been shown to be a usable method with regard to the subsequent parameter estimation.

It appears that the empirical method can obtain better reconstructions for certain types of data, the quality of these improving with the number of coefficients used. However, this type of data is somewhat artificial in that its source has only two intensities, corresponding to presence and absence of the isotope. In real data the transition between hot and cold spots is gradual and reconstructions show very little perceptible difference. Therefore, we suggest that the geometrical/physical coefficients are generally adequate. If the data are expected to have come from a near dichotomous truth, then a larger number of empirical coefficients should be used. However, in this case, incorporating scatter into the geometrical/physical coefficients may produce comparable reconstructions.

6 Acknowledgements

We are most grateful to the Bristol Royal Infirmary for providing the data and in particular Liz Berry and Liz Pitcher for their advice and help. Financial support was provided by the Science and Engineering Research Council.

References

BESAG, J.E. (1986) On the statistical analysis of dirty pictures (with discussion). *J. Roy. Statist. Soc.*, B, 48, 259-302.

EVANS, R.D. (1955) *The atomic nucleus*, McGraw-Hill, New York.

GEMAN, S. & GEMAN, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.*, 6, 721-741.

GEMAN, S., MANBECK, K.M. & McCLURE, D.E. (1991a) A comprehensive statistical model for single photon emission tomography. Reports in pattern analysis, Brown University, 153.

GEMAN, S., MANBECK, K.M. & McCLURE, D.E. (1991b) Design and implementation of a statistical method for emission computed tomography. Reports in pattern analysis, Brown University, 154.

GEMAN, S. & McCLURE, D.E. (1985) Bayesian image analysis: an application to single photon emission tomography. *Proc. Amer. Statist. Assoc., Stat. Comp.* Sect., 12-18.

GREEN, P.J. (1984) Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *J. R. Statist. Soc.* B 46, 149-92.

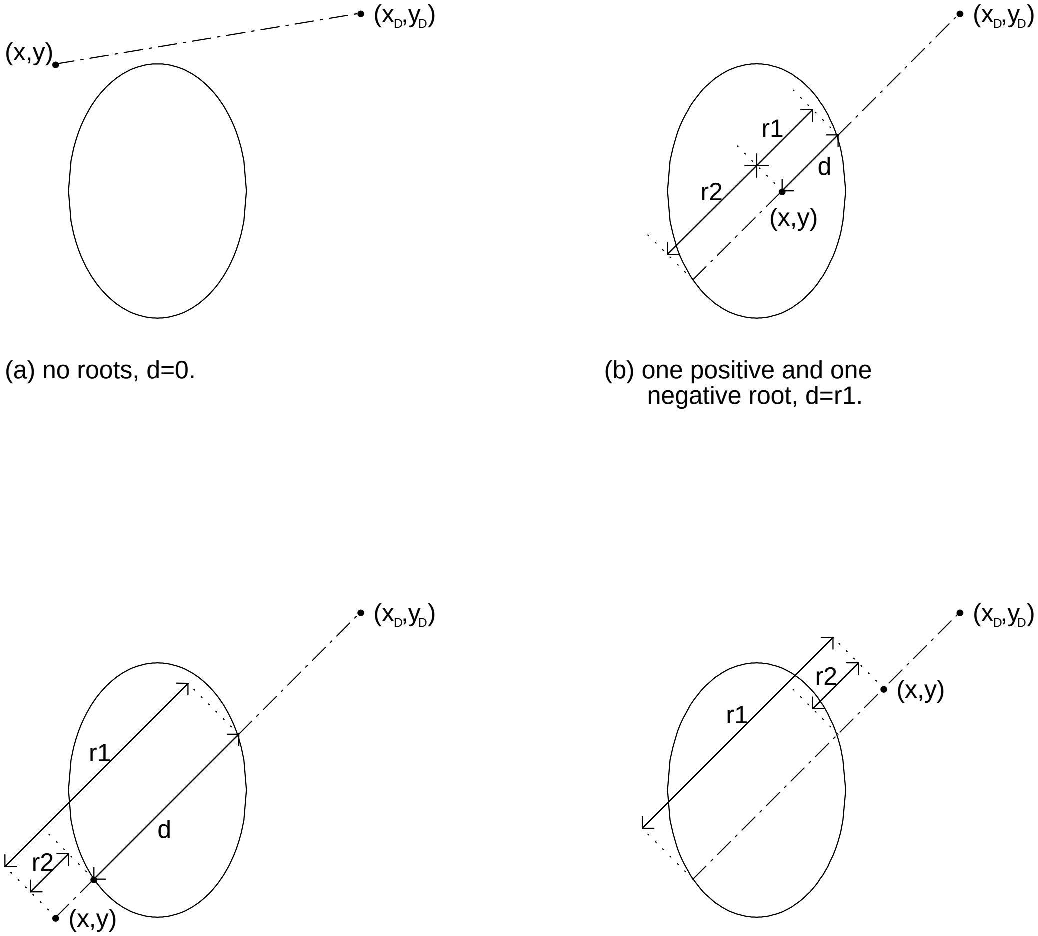
GREEN, P.J. (1990) Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE Trans. Medical Imaging*, 9, 84-93.

LARSSON, S.A. (1980) Gamma camera emission tomography. *Acta Radiologica*, Supplementum 363, Stockholm.

SELF, S.G. & LIANG, K (1987) Likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.*, 82, 605-610.

SHEPP, L.A. & VARDI, Y. (1982) Maximum likelihood reconstruction in positron emission tomography. *IEEE Trans. Medical Imaging*, 1, 113-122.

WEIR, I.S. (1993) Statistical modelling and reconstructions in single photon emission computed tomography. *Ph.D thesis*, Bristol University.



(a) no roots, d=0.

(b) one positive and one negative root, d=r1.

(c) two positive roots, d=r1-r2.

(d) two negative roots, d=0.

Figure 1: Attenuation distance through an ellipse

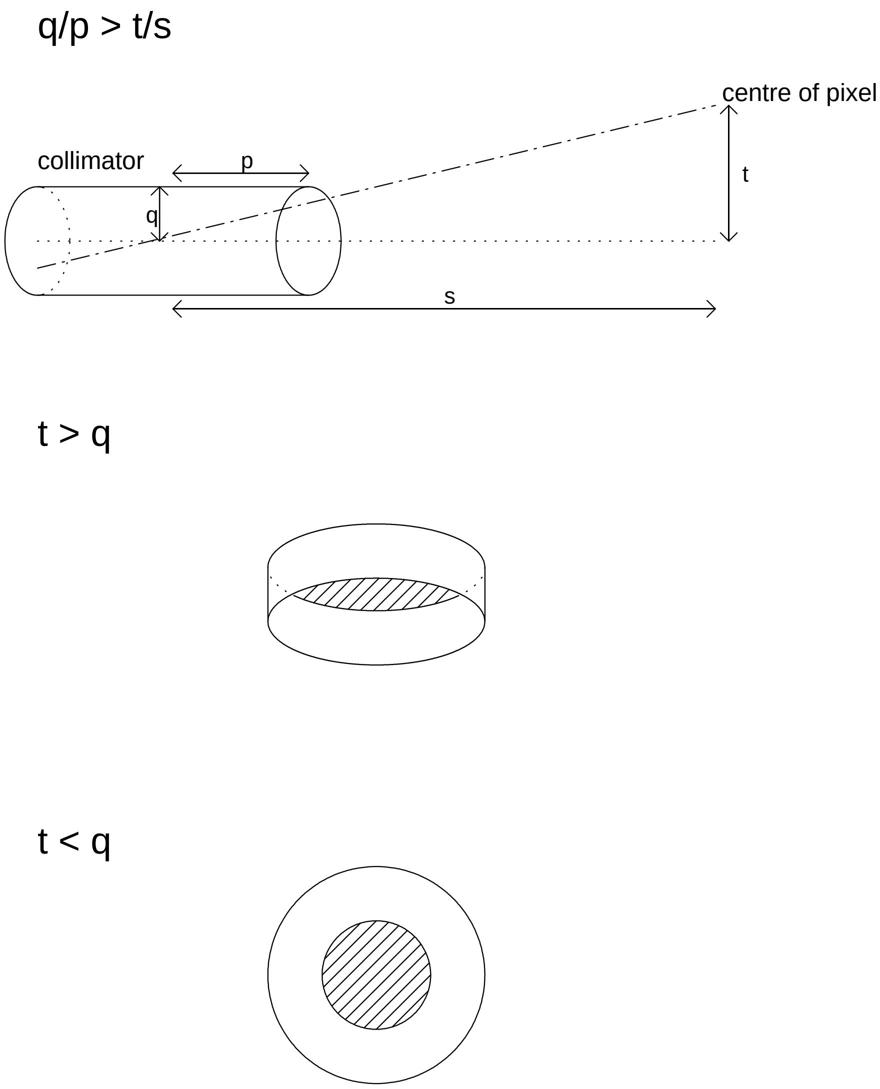


Figure 2: Solid angle of view into collimator

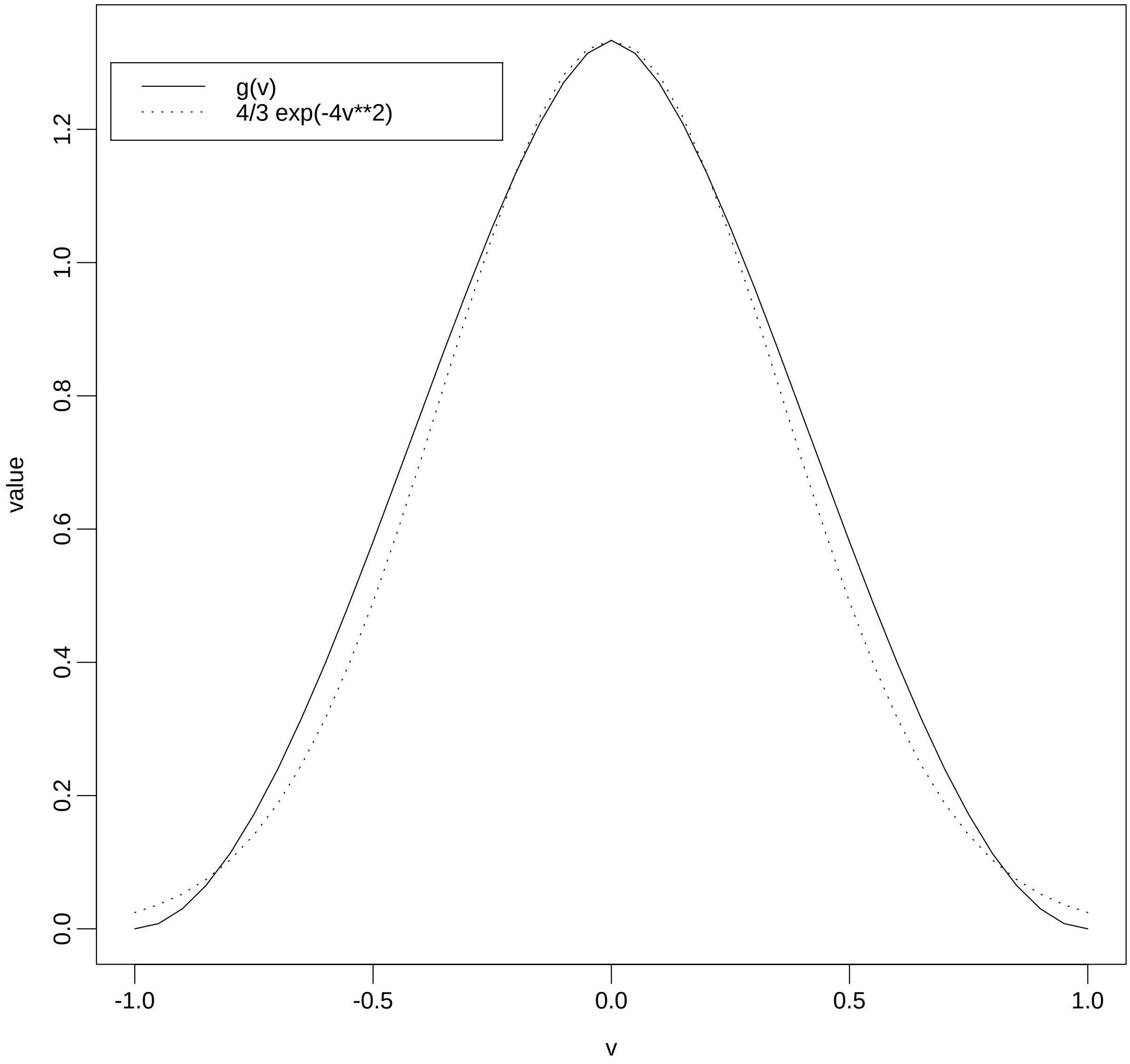


Figure 3: Curve fitted to solid angle of view function

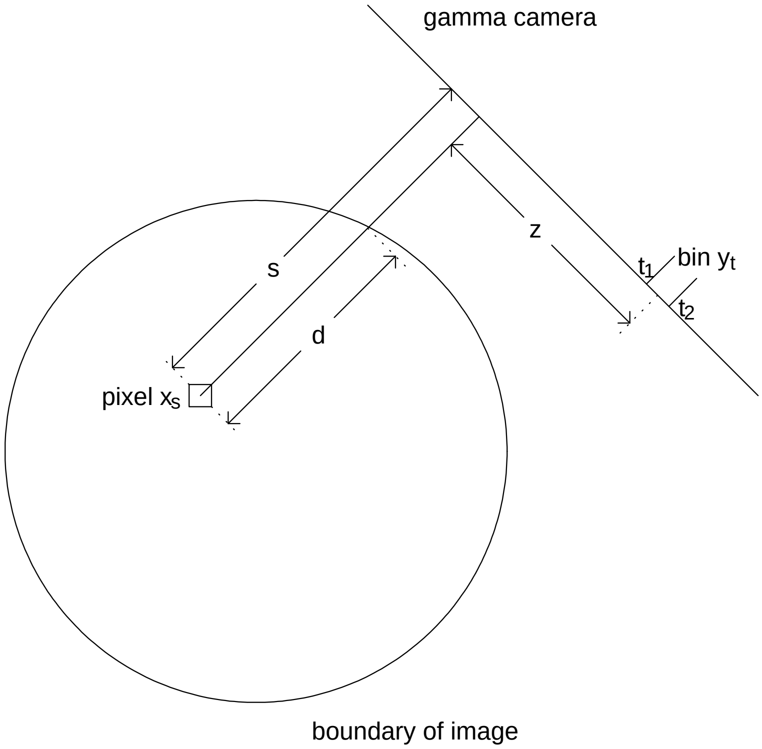
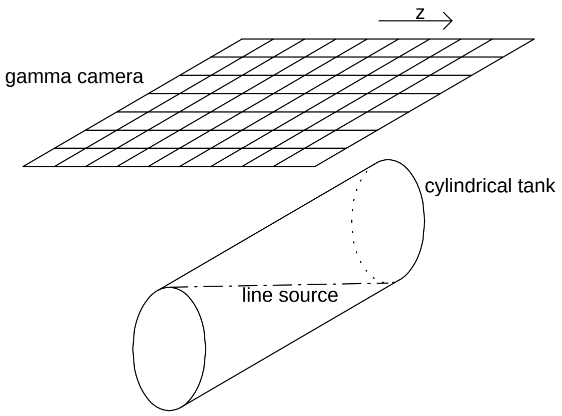


Figure 4: Distances of Geman's model

(a) Apparatus set up. Note horizontal direction of line spread function distance.



(b) Side view of equipment showing slices and vertical distances for a particular slice. Note each slice corresponds to a row in the data and has associated values of s and d.

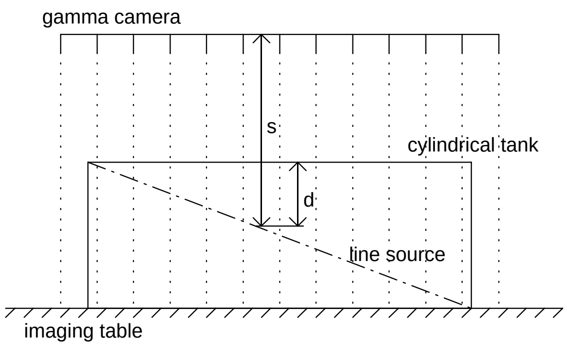
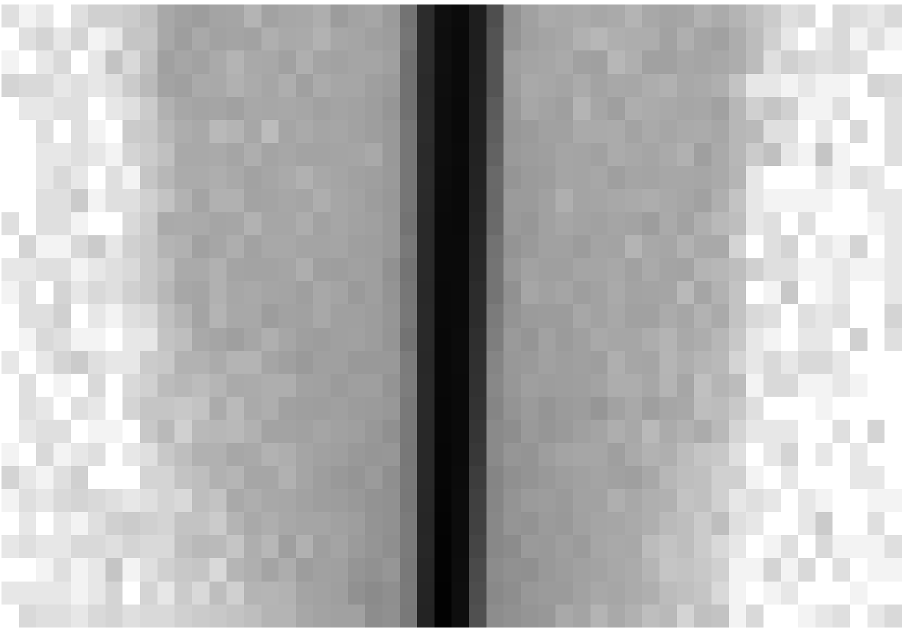
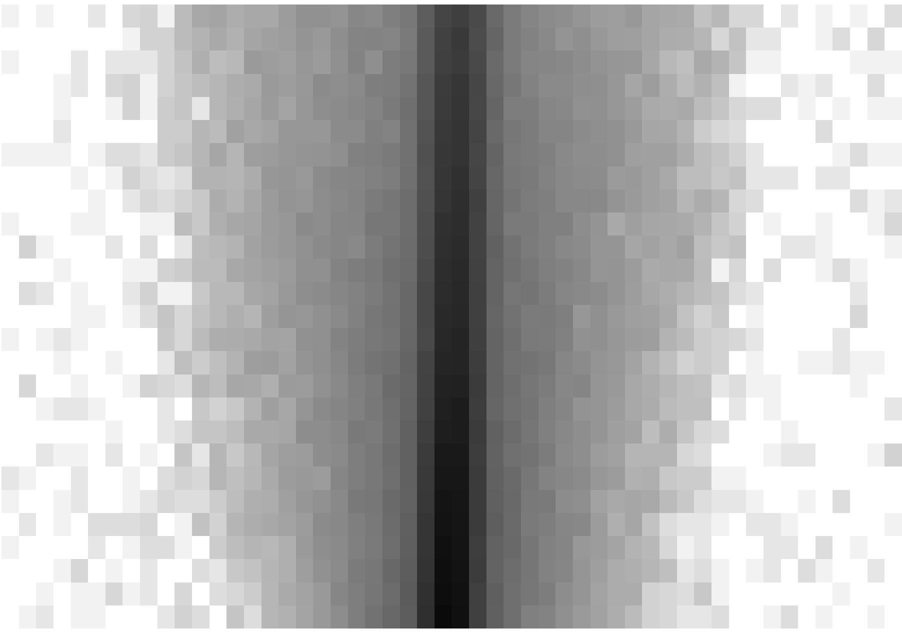


Figure 5: Experimental apparatus



(a) air data



(b) water data

Figure 6: Logged experimental data

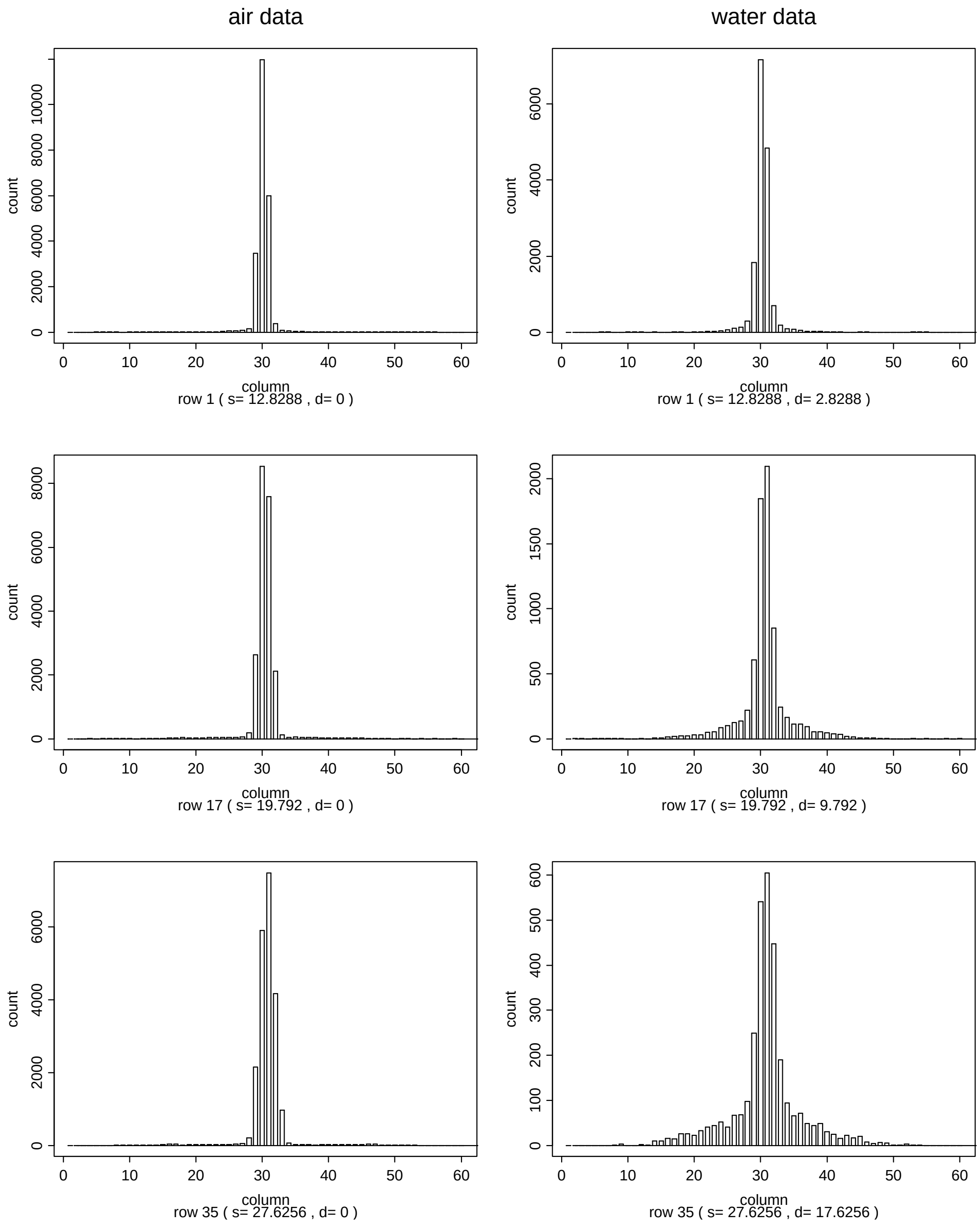


Figure 7: Histograms of rows of experimental data

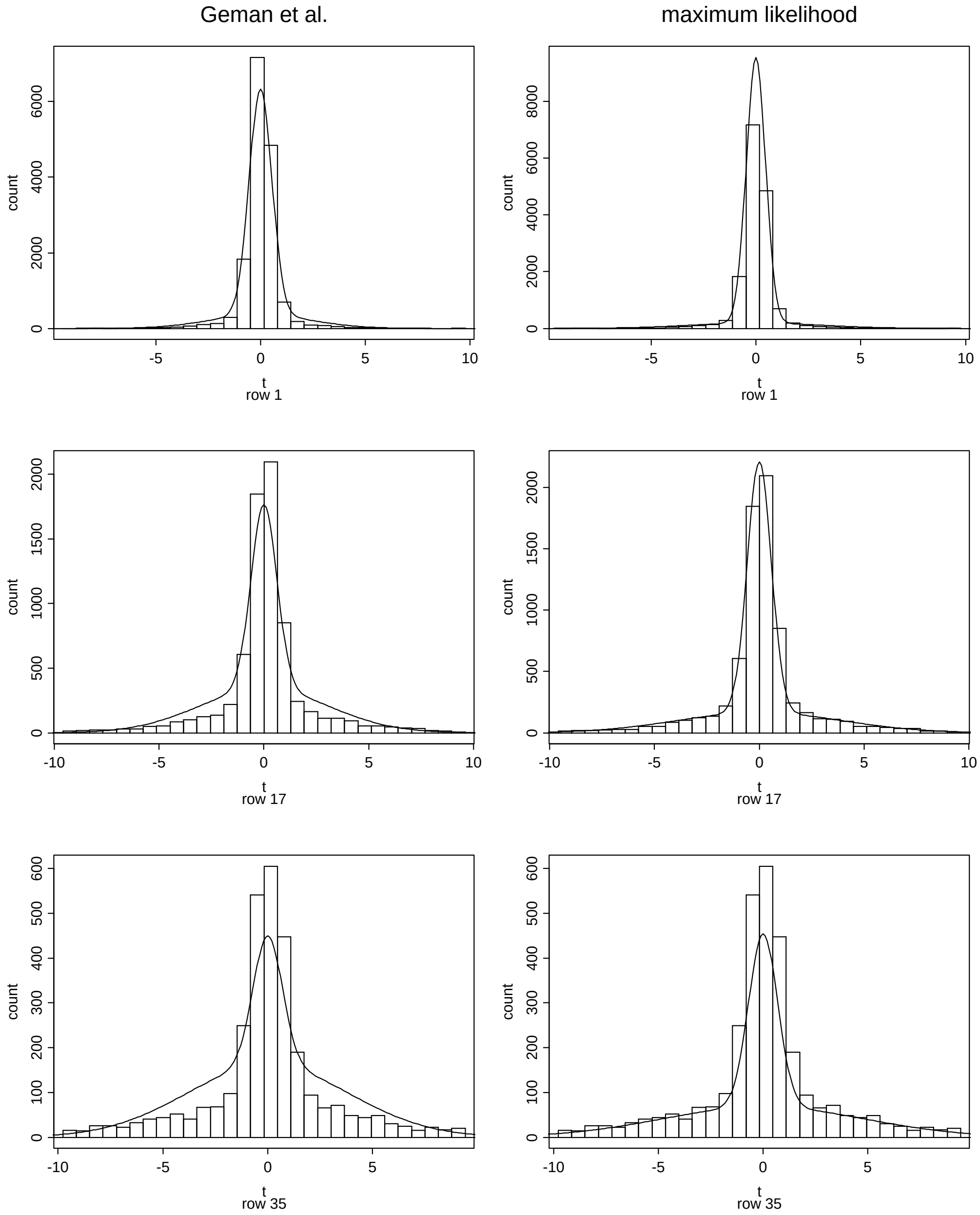


Figure 8: Comparison of Geman *et al.*'s method and maximum likelihood

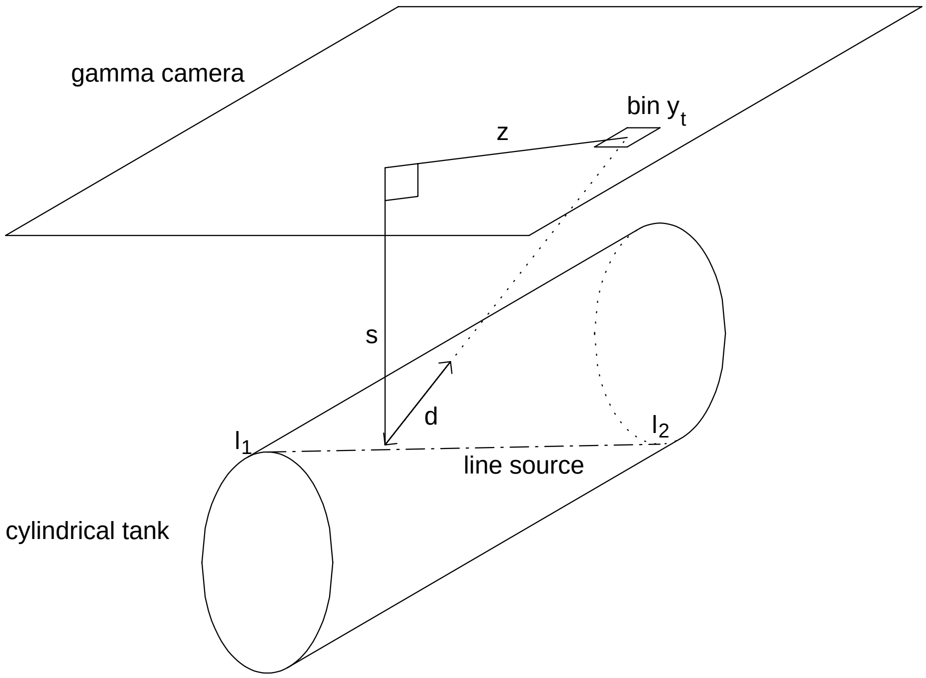


Figure 9: Incorporating the third dimension into the model

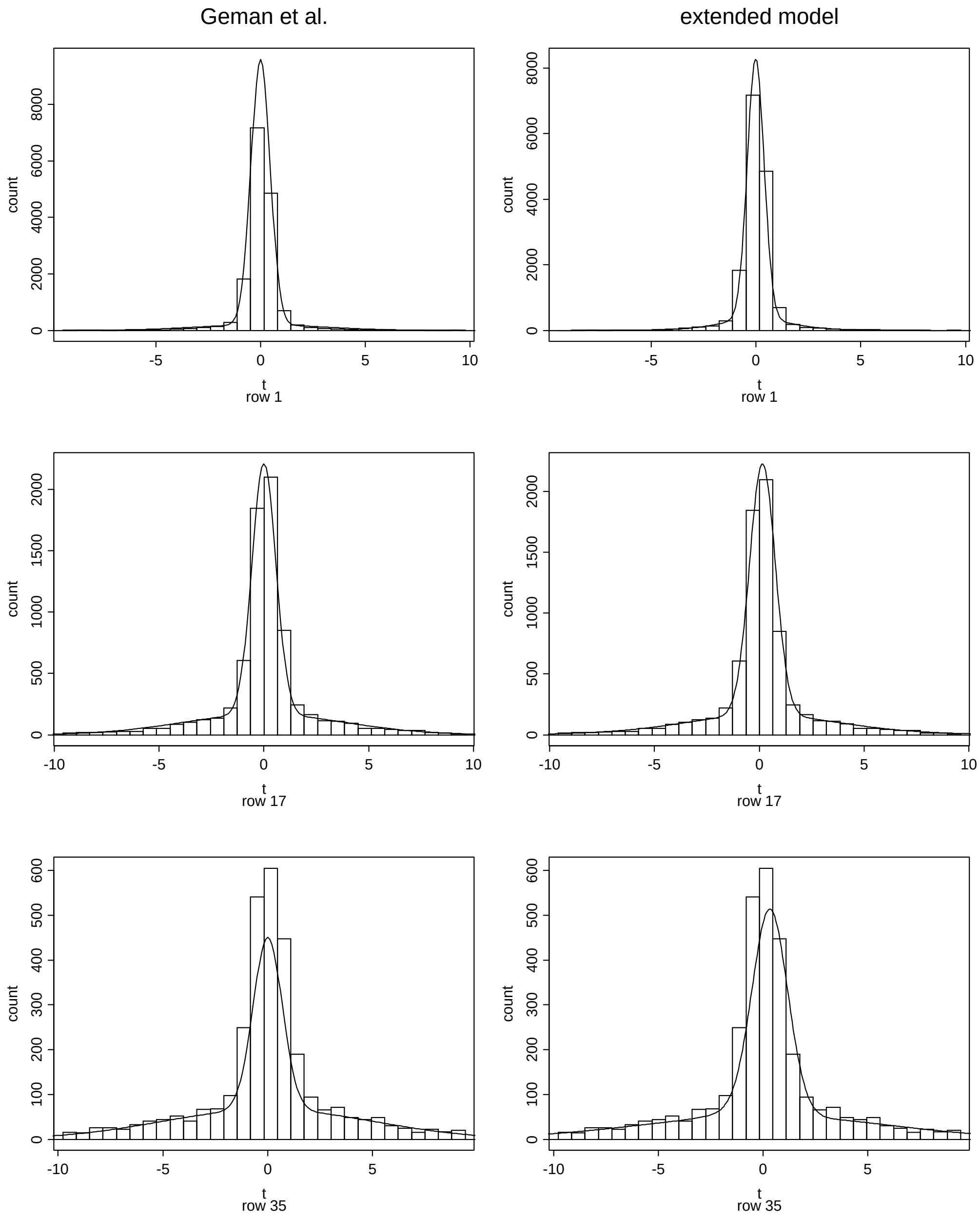
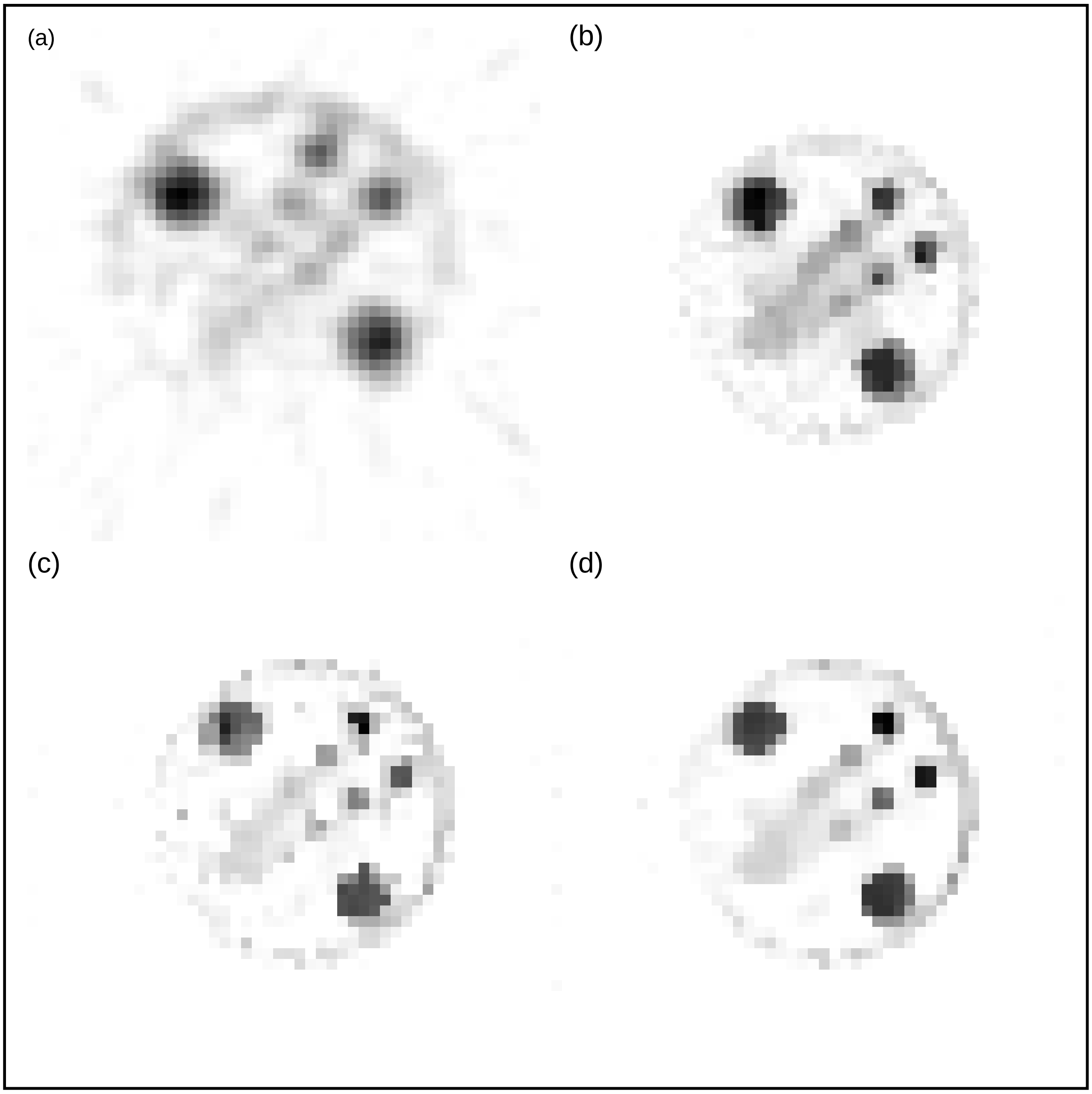
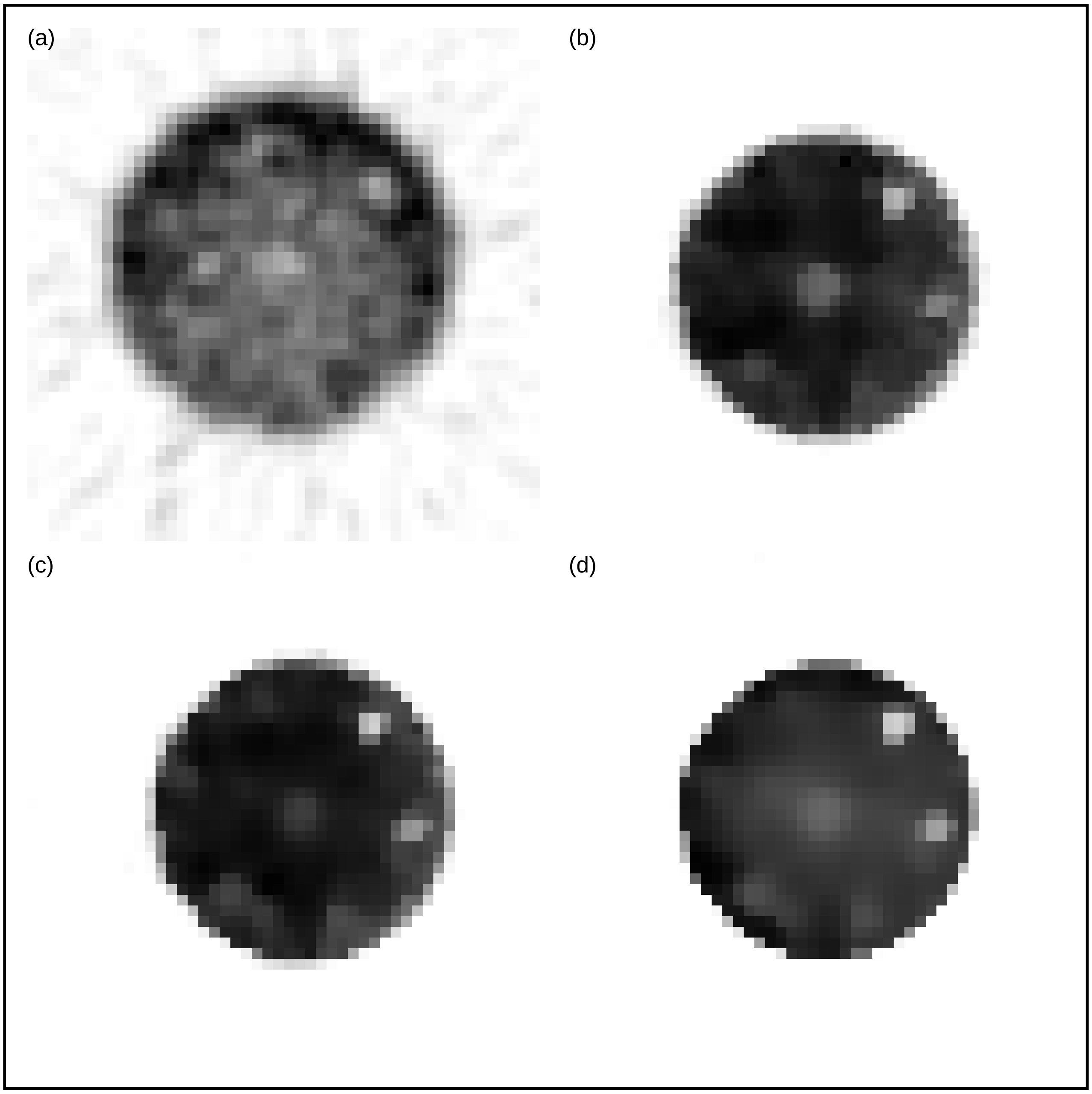


Figure 10: Comparison of Geman *et al.*'s model and fully extended model



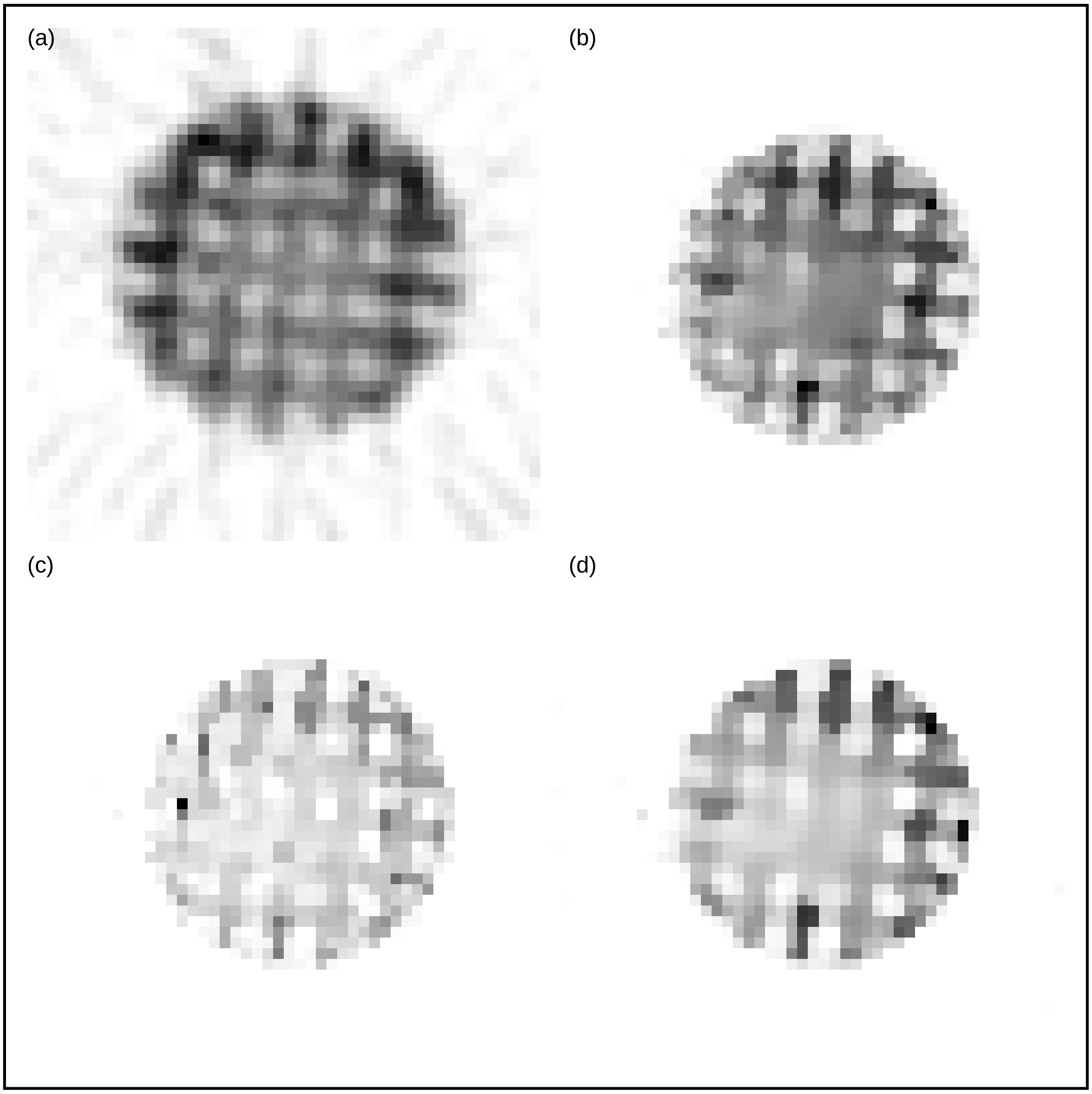
(a) filtered back-projection
(b) geometrical/physical
(c) numerical (geometrical/physical cut-off)
(d) numerical (3 sd direct, 1 sd scatter)

Figure 11: Reconstructions of hot insert



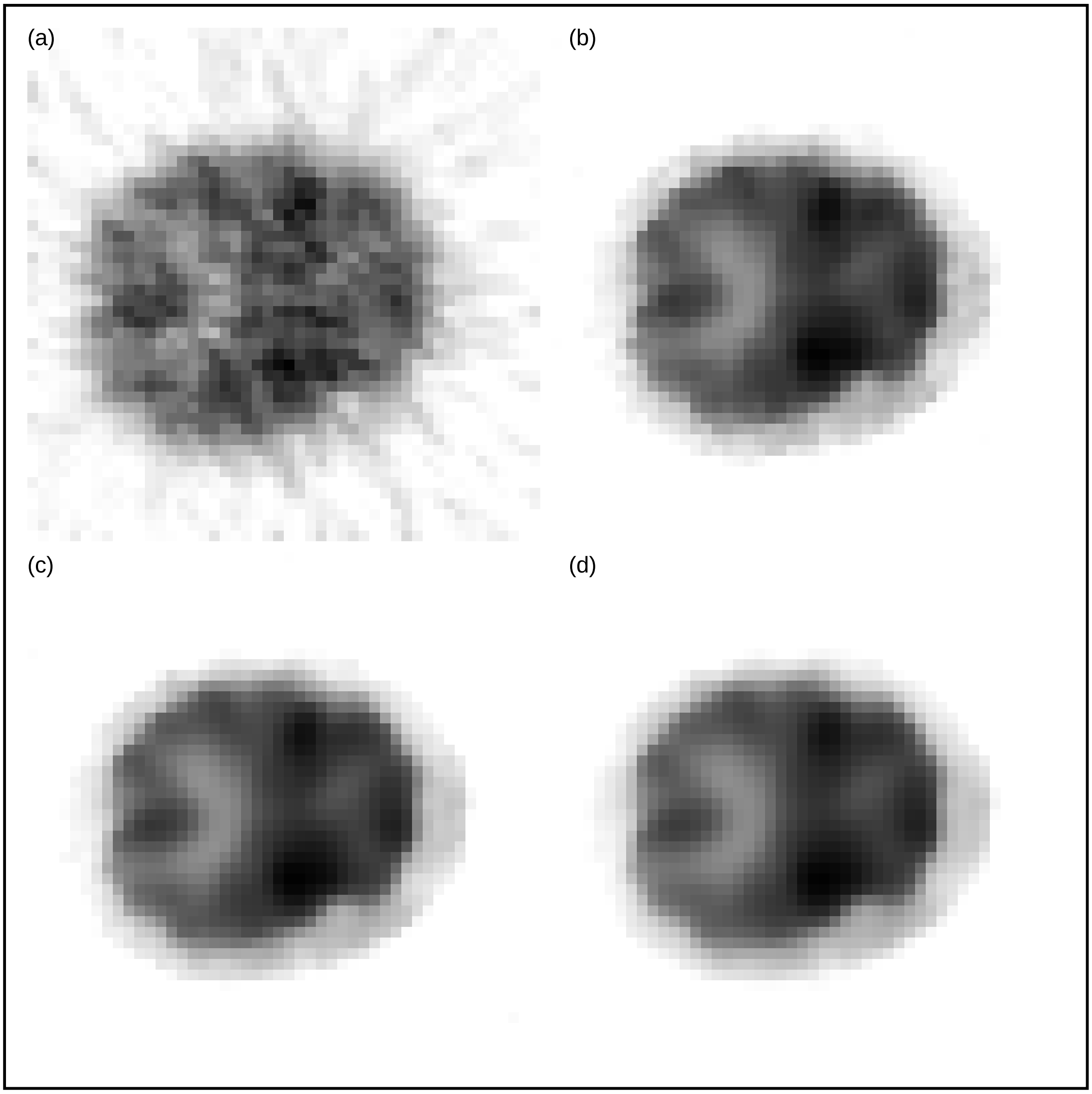
(a) filtered back-projection
(b) geometrical/physical
(c) numerical (geometrical/physical cut-off)
(d) numerical (3 sd direct, 1 sd scatter)

Figure 12: Reconstructions of cold insert



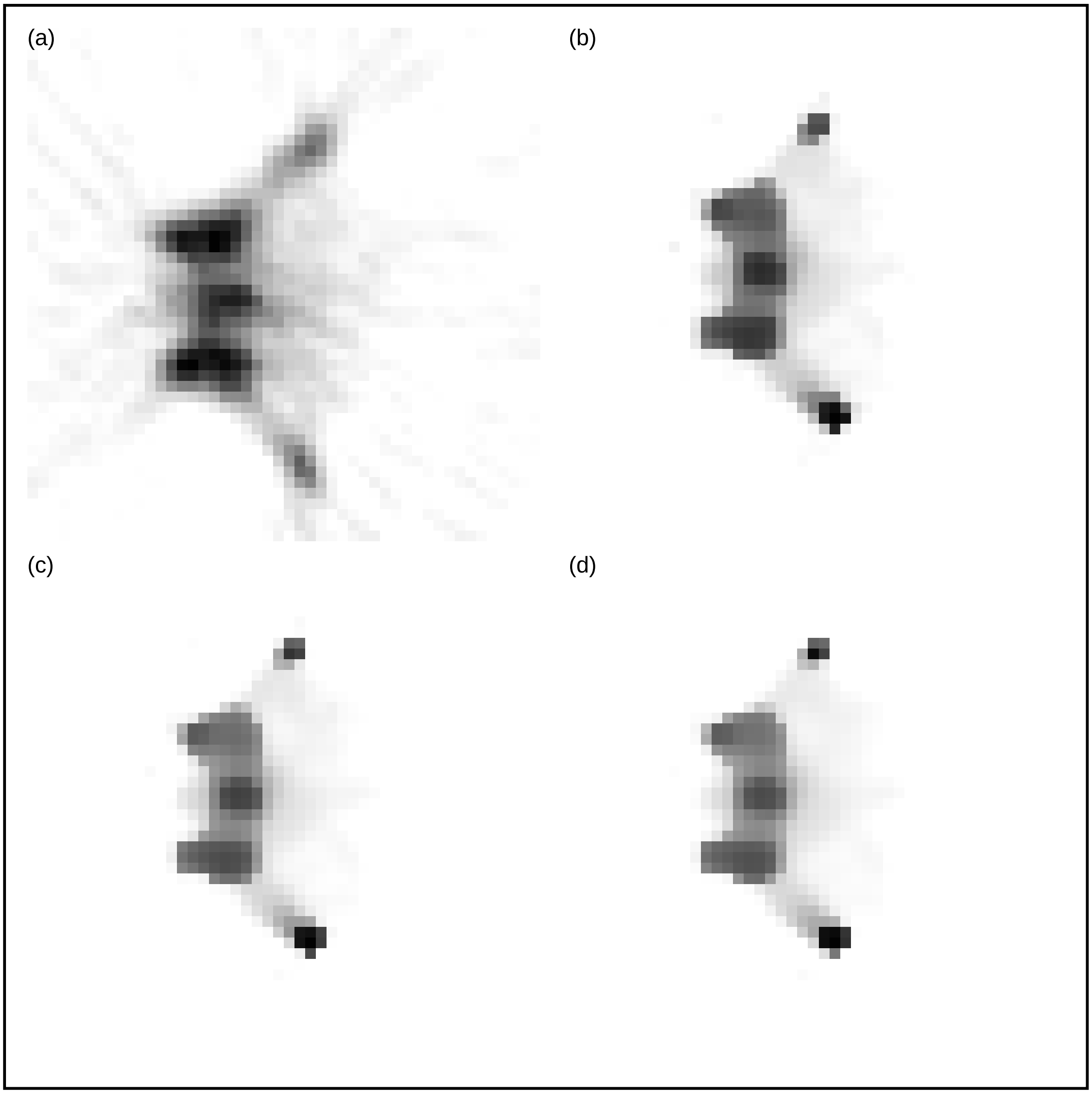
(a) filtered back-projection
(b) geometrical/physical
(c) numerical (geometrical/physical cut-off)
(d) numerical (3 sd direct, 1 sd scatter)

Figure 13: Reconstructions of linearity insert



(a) filtered back-projection
(b) geometrical/physical
(c) numerical (geometrical/physical cut-off)
(d) numerical (3 sd direct, 1 sd scatter)

Figure 14: Reconstructions of brain



(a) filtered back-projection
(b) geometrical/physical
(c) numerical (geometrical/physical cut-off)
(d) numerical (3 sd direct, 1 sd scatter)

Figure 15: Reconstructions of pelvis