Robust compartmental model fitting in direct emission tomography reconstruction

László Szirmay-Kalos¹ · Ágota Kacsó¹ · Milán Magdics¹ · Balázs Tóth¹

Accepted: 7 December 2020 / Published online: 6 February 2021
 © The Author(s) 2021

Abstract

Dynamic tomography reconstructs a time activity curve (TAC) for every voxel assuming that the algebraic form of the function is known a priori. The algebraic form derived from the analysis of compartmental models depends nonlinearly on the nonnegative parameters to be determined. Direct methods apply fitting in every iteration step. Because of the iterative nature of the maximum likelihood–expectation maximization (ML–EM) reconstruction, the fitting result of the previous step can serve as a good starting point in the current step; thus, after the first iteration we have a guess that is not far from the solution, which allows the use of gradient-based local optimization methods. However, finding good initial guesses for the first ML–EM iteration is a critical problem since gradient-based local optimization algorithms do not guarantee convergence to the global optimum if they are started at an inappropriate location. This paper examines the robust solution of the fitting problem both in the initial phase and during the ML–EM iteration. This solution is implemented on GPUs and is built into the 4D reconstruction module of the TeraTomo software.

Keywords Position emission tomography · ML–EM methods · GPU

1 Introduction

Dynamic positron emission tomography (PET) examines the dynamics of radiotracer accumulation in tissues [17,19,25]. The determined TAC of *voxel* V is searched in a predefined algebraic form $\mathcal{K}(\theta_V, t)$ of *time* t and *parameter vector* θ_V . The algebraic form is defined, for example, by compartmental models [5,8,28], i.e., as the solution of the differential equation of radiotracer exchange between compartments [29,31].

In direct methods the kinetic model is integrated into the reconstruction algorithm. The measurement time is partitioned into intervals (t_F, t_{F+1}) called *frames*. The *expected*

number of decays $\tilde{x}_F$ in frame F is

$$\tilde{x}_F(\theta_V) = \int_{t_F}^{t_{F+1}} \mathcal{K}(\theta_V, t) e^{-\lambda t} dt \quad (1)$$

where λ defines the *decay rate of the radiotracer*. The maximum likelihood estimator finds *voxel parameters* θ_V that maximize the likelihood of the measured events. At the location of the extremum, the derivatives of the likelihood are zero, which leads to the following nonlinear equations for every voxel V and *parameter* P :

$$\sum_F \frac{\partial \tilde{x}_F}{\partial \theta_{V,P}} \left(\frac{x_{V,F}}{\tilde{x}_F(\theta_V)} - 1 \right) = 0 \quad (2)$$

where $x_{V,F}$ is the *activity estimate of voxel* V in frame F after a pair of static forward and back projections, which couples the equations of all voxels. For a computationally straightforward implementation, the nested EM algorithm [14,25,27] decouples the equations of voxels by using $x_{V,F}$ from the previous iteration step. This method iterates two steps: The first executes forward and back projections in each frame to get updated voxel activity $x_{V,F}$, and the second fits param-

This research has been financed by OTKA K–124124.

✉ László Szirmay-Kalos
 szirmay@iit.bme.hu

¹ Department of Control Engineering and Information Technology, Budapest University of Technology and Economics, Budapest, Hungary

ters θ_V in each voxel V . The significant benefit of the nested EM algorithm is that the fitting step can be executed independently for each voxel. In this case, the objective functions of the local fitting establish a surrogate of the likelihood, which is a Kullback–Leibler-like term:

$$E_{\text{KL}}(\theta_V) = \sum_F \tilde{x}_F(\theta_V) - x_{V,F} \log(\tilde{x}_F(\theta_V)). \quad (3)$$

Instead of maximizing the global likelihood, this error term is minimized independently in every voxel V . The gradient of this term is zero at the solution of Eq. (2). As a similar problem is solved in every pixel, from now on, voxel subscript V is removed from the formulas.

Having selected the fitting criterion, algorithms for finding the best fit are needed. The Levenberg–Marquardt algorithm [15] is often used for least-square fitting, especially when there are no constraints. However, Eq. (2) is not equivalent to least-square fitting.

The main difference is that the least-square fitting would minimize the absolute error between the data points and the fitted function, while Eq. (2) can be interpreted as the minimization of the relative error. Such problems can be attacked by replacing the $1/\tilde{x}_F(\theta)$ term in Eq. (2) by its Taylor's expansion [22]:

$$\begin{aligned} \frac{1}{\tilde{x}_F(\theta^* + \mathbf{d})} &\approx \frac{1}{\tilde{x}_F(\theta^*)} + \sum_Q \frac{\partial 1/\tilde{x}_F}{\partial \theta_Q} \mathbf{d}_Q \\ &= \frac{1}{\tilde{x}_F(\theta^*)} - \frac{1}{\tilde{x}_F^2(\theta^*)} \sum_Q \frac{\partial \tilde{x}_F}{\partial \theta_Q} \mathbf{d}_Q \end{aligned} \quad (4)$$

where θ^* is the *current estimate* and $\mathbf{d} = \theta - \theta^*$ is the *step* to the new estimate. With this substitution, we get a linear system of equations for the unknown step

$$\mathbf{F} \cdot \mathbf{d} = \mathbf{r} \quad (5)$$

where the elements of matrix $\mathbf{F}$ and vector $\mathbf{d}$ are evaluated at current estimate θ^* :

$$\begin{aligned} \mathbf{F}_{P,Q} &= \sum_F \frac{\partial \tilde{x}_F}{\partial \theta_P} \frac{x_F}{\tilde{x}_F^2(\theta)} \frac{\partial \tilde{x}_F}{\partial \theta_Q}, \\ \mathbf{r}_P &= \sum_F \frac{\partial \tilde{x}_F}{\partial \theta_P} \left(\frac{x_F}{\tilde{x}_F(\theta)} - 1 \right). \end{aligned} \quad (6)$$

The organization of this paper is as follows. Section 2 reviews the related previous work, identifies the possibilities of improvement and presents the research objectives. Section 3 discusses the original and the modified algebraic descriptions of compartmental models. In Sect. 4, we present our analytic computation scheme for activity values and their

derivatives. Section 5 discusses our model-specific improvement of the Levenberg–Marquardt scheme, and Sect. 6 proposes the method for finding appropriate initial values for the iterative optimization. Sections 7 and 8 compare the proposed methods with the state of the art in 2D reconstruction of a mathematical phantom and in the reconstruction of 3D data. The paper closes with conclusions.

2 Previous work, problem statement and objectives

The state of the art of dynamic positron emission tomography is reviewed in survey papers [4,19,25,26]. Our paper belongs to the category of nonlinear fitting since the compartmental models are essentially nonlinear [29,31]. Such fitting methods have several difficulties. The measured data are contaminated by noise [23,24]. The actual guess of the kinetic model should be integrated in time and then derived with respect to the optimization parameters. The optimization parameters are nonnegative and upper bounded. The error term of Eq. 3 is not quadratic, i.e., its derivative is not linear. Thus, this is a nonlinear, constrained fitting, which needs numerical methods as there is no direct analytic solution. Integrals can be estimated with numerical quadrature [30], but they have high computational cost and high error if the function changes quickly. Applied general numerical optimization techniques include coordinate descent [8], Newton–Raphson iteration [20], preconditioned gradient descent [16] and Levenberg–Marquardt iteration [28], which is most popular choice because it needs only the first derivatives and is simple to implement. Gradient-based optimization methods converge quickly close to the solution, but do not guarantee convergence, and even if they converge, they can be stuck in a local optimum. Sophisticated global optimization techniques would have high computational cost since this fitting should be executed for every voxel and in each ML–EM iteration [11,12]. For gradient-based solvers, the specification of the initial state becomes crucial. Starting with constant initial values or initializing the parameters with random numbers may cause slow initial convergence. To attack this problem, we propose an initialization scheme that exploits the original measured data and is simple to compute; thus, its additional computational cost is amortized by the faster convergence.

Note that general optimization techniques do not take into account the particular properties of the PET problem. Problem-specific solutions include simplifications, e.g., the replacement of the nonlinear fitting by a much simpler least-square fitting [18]. Note that a direct solution would be available for the least-square fitting of a function that linearly depends on its parameters. However, modifying the optimization criteria would alter and slow down convergence. If only

a subset of parameters have such linear behavior, the fitting can be decomposed to a nonlinear fitting and then the least-square fitting of the linear parameters [2,7].

The main research goal of this paper is to propose algorithms that solve the fitting problem efficiently, i.e., more robustly and more accurately than the classical Levenberg–Marquardt method without significantly increasing the computation time. In particular, the main contributions of this paper are as follows:

- A modification of the algebraic form of the two-tissue model is proposed in Sect. 3, which is exploited by the refinement of the Levenberg–Marquardt scheme in Sect. 5. Unlike previous work, our simplified fitting schemes do not replace the complicated but accurate nonlinear solvers, but provide additional potential refinements. Their improvement is checked, and if they fail to improve the fitting, their proposal is ignored. In our GPU implementation, the additional improvement trial has negligible additional computational cost.
- We present the analytic computation of matrix $\mathbf{F}$ and vector $\mathbf{r}$ of the linear system of Eq. (5) as well as its automatic derivation in Sect. 4. The analytic computation not only leads to a straightforward implementation, but is also significantly faster than the numeric evaluation.
- A low computational cost initial estimation process is established to guess the parameters from where the ML–EM iteration can be started (Sect. 6).

Our goal is a method that is applicable in clinical and preclinical practice and can solve reconstruction problems involving about a billion lines of responses and voxels in reasonable time, and therefore is appropriate for parallel GPU implementation. As GPUs prefer single-precision arithmetics, the proposed techniques are to be stable and robust to lower-precision computations.

3 Algebraic forms of compartmental models

Based on the solution of differential equations expressing the tracer exchange between n compartments [5], the TAC has an algebraic form that is a nonnegative linear combination of convolutions of the *blood input function* $C_p(t)$ and the *impulse response* $w(t)$ of the tissue. The impulse response is the non-negatively weighted sum of exponentials of unknown, nonnegative *rate constants* α_i and constrained only for positive *time values* t by the *Heaviside step function* $\epsilon(t)$:

$$w(t) = \epsilon(t) \sum_{i=1}^n c_i e^{-\alpha_i t}$$

where c_i factors are the nonnegative weights. Taking into account that a voxel is a mixture of the tissue and blood, we obtain

$$\mathcal{K}(\theta, t) = f_v C_W(t) + (1 - f_v) w(t) * C_p(t) \quad (7)$$

where $f_v \in [0, 1]$ is the unknown *fraction of blood* and C_W is the known or separately measured *total blood concentration function* taking also into account that portion of the radio-tracer that cannot diffuse into the tissues.

Reconstruction means the identification of parameters $\theta = (f_v, c_1, \alpha_1, c_2, \alpha_2, \dots)$ for every voxel. However, these parameters are not optimal for fitting since the TAC depends on these parameters nonlinearly. Thus, we use an equivalent algebraic form, where $f_v, a_1, a_2, \dots$ are linear parameters whenever the rate constants are fixed:

$$\mathcal{K}(\theta, t) = f_v C_W(t) + \left(\epsilon(t) \sum_{i=1}^n a_i e^{-\alpha_i t} \right) * C_p(t) \quad (8)$$

where the correspondence between the new and the original parameters is $a_i = (1 - f_v) c_i$.

Blood input functions $C_p(t)$ and $C_W(t)$ are also subjects to fitting executed once at the beginning of the reconstruction. *Feng's model* [3] assumes them in the following algebraic form if the radiotracer is injected in $t = 0$:

$$C_p(t) = A_1 t e^{-\beta_1 t} + \sum_{j=2}^4 A_j (e^{-\beta_j t} - e^{-\beta_1 t}) \quad (9)$$

if $t > 0$ and zero otherwise. *Blood exponents* showing the blood activity dynamics $\beta_1, \dots, \beta_4$ and *linear blood parameters* $A_1, \dots, A_4$ are determined with simulated annealing.

4 Analytic integration and derivation of the activity

Voxel activity requires the integration of the convolutions of exponential functions. During the iterative fitting of the kinetic model, the derivatives of the convolutions are needed. This section presents an analytic method with an easy program implementation to obtain these integrals and their derivatives.

According to Eq. (1), the activity of a voxel in a frame is expressed by a time integral. The antiderivative of the integrand is

$$I(t) = \int \left(\left(\epsilon(t) \sum_{i=1}^n a_i e^{-\alpha_i t} \right) * C_p(t) \right) e^{-\lambda t} dt$$

$$= \sum_{i=1}^n a_i \left(A_1 H_i(t) + \sum_{j=2}^4 A_j (G_{ij}(t) - G_{i1}(t)) \right) \quad (10)$$

where there are two types of convolutions of terms from the blood input function and the impulse response. Using the $\alpha_i^* = \alpha_i + \lambda$ and $\beta_j^* = \beta_j + \lambda$ notations, the first type is:

$$G_{ij}(t) = \int (\epsilon(t) e^{-\alpha_i t} * \epsilon(t) e^{-\beta_j t}) e^{-\lambda t} dt$$

$$= \frac{e^{-\alpha_i^* t} / \alpha_i^* - e^{-\beta_j^* t} / \beta_j^*}{(\alpha_i^* - \beta_j^*)}. \quad (11)$$

The second type of convolutions is

$$H_i(t) = \int (\epsilon(t) e^{-\alpha_i t} * \epsilon(t) t e^{-\beta_1 t}) e^{-\lambda t} dt$$

$$= \frac{e^{-\beta_1^* t} / \beta_1^* - e^{-\alpha_i^* t} / \alpha_i^* - (\alpha_i^* - \beta_1^*)(t + 1/\beta_1^*) e^{-\beta_1^* t} / \beta_1^*}{(\alpha_i^* - \beta_1^*)^2}. \quad (12)$$

Finally, the activity in a frame is

$$\tilde{x}_F(\theta) = f_v \int_{t_F}^{t_{F+1}} C_W(t) dt + I(t_{F+1}) - I(t_F). \quad (13)$$

During the fitting process, we also need the derivatives of these integrals with respect to the kinetic parameters:

$$\frac{\partial I(t)}{\partial a_i} = A_1 H_i(t) + \sum_{j=2}^4 A_j (G_{ij}(t) - G_{i1}(t)),$$

$$\frac{\partial I(t)}{\partial \alpha_i} = a_i \left(A_1 \frac{\partial H_i(t)}{\partial \alpha_i} + \sum_{j=2}^4 A_j \left(\frac{\partial G_{ij}(t)}{\partial \alpha_i} - \frac{\partial G_{i1}(t)}{\partial \alpha_i} \right) \right) \quad (14)$$

where the derivatives of time integrals $G_{ij}(t)$ and $H_i(t)$ can still be analytically expressed, but become quite complicated.

Another problem is that the above formulas are invalid if $\alpha_i^* = \beta_j^*$ or $\alpha_i^* = 0$ or $\beta_j^* = 0$, and are numerically unstable when the values are close to these conditions. As β_j^* blood exponents represent the dynamics of the blood activity, while rate constants α_i^* represent the dynamics of the tissue activity, it can easily happen that during the numerical optimization these values get very close to each other. On the other hand, irreversibility can also cause α_i to go to zero.

Both the algebraic complexity of the derivatives and the need for special cases can be attacked by the application

of *dual numbers* [1] in the computer implementation. This means that a function f is represented by a dual number $f + \mathbf{i} f'$ where $\mathbf{i}$ is the imaginary unit with the $\mathbf{i}^2 = 0$ definition, the real part is the value of function f and the imaginary part is the value of the derivative at the same location f' . It is easy to see that the arithmetic rules of basic operations of addition, subtraction, multiplication and division for such dual numbers are the same as the original arithmetic rules in the real part and as the derivation rules in the imaginary part.

When we encounter a 0/0 type undefined division or the numerator and denominator are close to zero causing numerical instability, the l'Hospital rule can automatically be applied. If $f_1(x)$ and $f_2(x)$ are close to zero, then

$$\frac{f_1}{f_2} \approx \frac{f_1'}{f_2'}, \quad \left(\frac{f_1}{f_2} \right)' \approx \frac{f_1'' f_2' - f_1' f_2''}{2 (f_2')^2}. \quad (15)$$

Thus, we also need the second derivatives. To cope with this or with Newton–Raphson iteration, the dual numbers should be generalized for at least two imaginary units $f + \mathbf{i} f' + \mathbf{j} f''$. The arithmetic rules of the imaginary units that make the basic operations similar to derivation rules are

$$\mathbf{i}^2 = 2\mathbf{j}, \quad \mathbf{i}\mathbf{j} = 0, \quad \mathbf{j}^2 = 0. \quad (16)$$

The chain rule on the exponential has the following form:

$$e^{f + \mathbf{i} f' + \mathbf{j} f''} = e^f + \mathbf{i} e^f f' + \mathbf{j} e^f ((f')^2 + f''). \quad (17)$$

This works well when convolution G_{ij} of Eq. (11) is computed. However, convolution H_i of Eq. (12) has $(\alpha_i^* - \beta_1^*)^2$ in its denominator, i.e., its derivative has $(\alpha_i^* - \beta_1^*)^4$ in its denominator. Thus, l'Hospital rule should be applied four times to obtain a nonzero value in the denominator, making the computation of derivatives up to the second order insufficient. It would be possible to further extend dual numbers, but the performance penalty would be too high. Therefore, when α_i is very close to β_1 , it is perturbed and the derivative is computed a little farther.

The arithmetic rules of the dual numbers can be summarized by a simple C++ class exploiting operator overloading. With this, we can implement only the computation of the integrated values, while the derivatives and the 0/0 type divisions are automatically taken care of. Figure 1 shows the H_1 and G_{11} plotted as functions of α_1^* and compares our analytic approach to numerical integration and differentiation. The analytic approach is not only more accurate and robust, but is an order of magnitude faster to compute.

Fig. 1 Convolutions H_1 and G_{11} plotted as functions of rate constant α_1^* , setting $\beta_1^* = 0.5$, $t_F = 0$, $t_{F+1} = 6$. Note that these functions have singularities at $\alpha_1^* = 0$ and $\alpha_1^* = \beta_1^* = 0.5$. The upper figure shows the results of analytic integration and automatic derivation. The lower figure depicts the numerical results when integrals are computed with step $\Delta t = 0.02$ and derivatives are computed with $\Delta \alpha = 10^{-5}$. We used single-precision arithmetic (float) in all cases

5 Fitting during iterative reconstruction

The solution of Eq. (2) can be interpreted as a nonlinear constraint fitting that minimizes the relative error. However, the simple approach of Euclidean norm and linear models can at least partially be used for more complex models. On the one hand, in Sect. 3 an equivalent algebraic form was proposed, in which a larger linear subgroup of the parameters can be identified. A set of parameters is said to form a *linear subgroup* if the model depends on them linearly when all parameters outside this group are fixed. On the other hand, we find a weighting scheme for the Euclidean norm, which can be minimized with a direct method. With these tricks, we can easily refine the parameter values of the linear subgroup. If the extra step of the simple fitting for the linear parameters does not improve the error term of Eq. 3, then the result of this extra step can be ignored.

Having obtained an estimate for the parameters by general constraint fitting, we can further refine the linear parameters $\bar{\theta} = (f_v, a_1, a_2, \dots)$ of activity $\tilde{x}_F$:

$$\tilde{x}_F = \mathbf{b}_F^T \cdot \bar{\theta} \quad (18)$$

Fig. 2 2D phantom and tomograph model

where vector $\mathbf{b}_F^T = (b_F^{(0)}, b_F^{(1)}, b_F^{(2)}, \dots)$ contains the integrals of the basis functions in frame F :

$$b_F^{(0)} = \int_{t_F}^{t_{F+1}} C_W(t) e^{-\lambda t} dt,$$

$$b_F^{(i)} = \int_{t_F}^{t_{F+1}} [(\epsilon(t) e^{-\alpha_i t}) * C_p(t)] e^{-\lambda t} dt, \quad i = 1, 2, \dots$$

Note that these can be obtained analytically using the results of Sect. 4.

Fitting requires the solution of Eq. (2), which can be rewritten as:

$$\sum_F \frac{\partial \tilde{x}_F}{\partial \theta_P} \cdot \frac{x_F - \tilde{x}_F(\theta)}{\tilde{x}_F(\theta)} = 0. \quad (19)$$

As this phase starts with the estimates obtained with a nonlinear equation solver and then refines the linear parameters, we already have a guess $\hat{x}_F$ for $\tilde{x}_F(\theta)$ showing up in the denominator. Let us substitute $\tilde{x}_F(\theta)$ by $\mathbf{b}_F^T \cdot \bar{\theta}$ in the numerator and the partial derivative:

$$\sum_F \frac{\partial \mathbf{b}_F^T \cdot \bar{\theta}}{\partial \theta_P} \cdot \frac{x_F - \mathbf{b}_F^T \cdot \bar{\theta}}{\hat{x}_F} = 0. \quad (20)$$

Considering all linear parameters and rearranging the terms, a system of linear equations is obtained for the linear parameters:

$$\left(\sum_F \frac{\mathbf{b}_F \cdot \mathbf{b}_F^T}{\hat{x}_F} \right) \cdot \bar{\theta} = \sum_F \frac{\mathbf{b}_F}{\hat{x}_F} x_F. \quad (21)$$

This method is called the *weighted refinement*. If $\hat{x}_F$ is set to 1, which corresponds to a non-weighted least-square fitting, then the method is called *linear refinement*.

Fig. 3 Error map of the four regions in the 2D phantom. The two axes correspond to rate constants α_1 and α_2 , and colors depict the error that is obtained by fitting the remaining set of linear parameters

Solving the linear system of Eq. (21) may result in values that are outside of the allowed range, i.e., weights a_i may be negative and f_v outside of $[0, 1]$. Negative parameters could be removed by nonnegative least-square fitting, other violations by inequality constrained least squares. Here, we use a simpler technique. If some parameters are outside of the allowed range, their value is substituted by the boundary value that is overstepped and x_F is reduced by the product of the boundary value and the corresponding basis function. For the parameters inside the range, another linear system is constructed in the form of Eq. (21), but the elements of the fixed parameters are removed from basis vector $\mathbf{b}$ reducing the size of the linear system to the number of not-fixed parameters. This operation is repeated until all parameters are inside the allowed range or on its boundary.

6 Initial estimation

The initial estimation of the parameters can use just the measured data, which is the list of events that are binned in frames, and thus can be expressed by a matrix $y_{L,F}$ or a vector of LOR values $\mathbf{y}_F$ in each frame F .

To find a guess for the voxel activities, we use a direct method and not iterative techniques, because at this stage, we do not have information about the order of magnitude of the activity values. Starting an iteration from a many order of magnitude different value may cause numerical instability in the GPU favouring 32 bit precision arithmetics.

The initial direct estimation assumes that the volume can be partitioned into N_R homogeneous regions $\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_{N_R}$ where all voxels belonging to the same region R share the same parameters and thus have the same activity values $\tilde{x}_{R,F}$. Thus, in this phase, the dimension of the problem is reduced from the number of voxels to the number of regions, simplifying the *expected number of coincidences* $\tilde{y}_{L,F}$ in LOR L during frame F to

$$\tilde{y}_{L,F} = \sum_R \mathbf{B}_{L,R} \tilde{x}_{R,F} \quad (22)$$

Table 1 Ground truth and the resulting parameters after the initial guess and the refinement consisting of 20 region-based ML-EM iterations. Don't care reference values are indicated by symbol $\times$

	a_1	a_2	α_1	α_2	f_v
Air reference	0.00	0.00	$\times$	$\times$	0.00
Initial guess	0.00	0.00	1.76	0.1	0.00
Refinement	0.00	0.00	20	0.27	0.00
Gray matter reference	1.20	1.20	1.00	5.00	0.00
Initial guess	1.21	0.93	1.00	4.10	0.02
Refinement	1.21	1.19	0.99	4.82	0.00
White matter reference	0.20	0.10	0.10	2.00	0.00
Initial guess	0.19	0.11	0.09	1.66	0.00
Refinement	0.20	0.10	0.10	1.83	0.00
Blood reference	0.00	0.00	$\times$	$\times$	1.00
Initial guess	0.51	0.1	0.02	9.74	0.02
Refinement	0.01	0.38	0.15	20	1.00

where $\mathbf{B}_{L,R}$ is the *region-based system matrix*, storing the probabilities that a decay in region R causes a coincidence event in LOR L . Its row R can be obtained by forward projecting a volume of constant 1 values in voxels belonging to region R and zero elsewhere.

Because of considering regions rather than individual voxels, the Poisson model can be replaced by a Gaussian model since a region has much more decays than single voxels, and the Gaussian assumption is acceptable for high statistic measurements. Note that in the extreme case, we can consider the whole field of view as a single region. In case of the Gaussian model, the reconstruction is equivalent to the solution of the following linear system:

$$y_{L,F} = \sum_R \mathbf{B}_{L,R} \tilde{x}_{R,F}, \text{ in matrix form: } \mathbf{y}_F = \mathbf{B} \cdot \tilde{\mathbf{x}}_F, \quad (23)$$

where $y_{L,F}$ is the measured number of coincidences in LOR L during frame F . Comparing this equation to Eq. (22), we note that the expected number of the coincidences $\tilde{y}_{L,F}$ is replaced by the measured number of coincidences $y_{L,F}$,

Fig. 4 Reconstructions of macroparameters K_1 , V_D and K_i of the 120k coincidence measurement. The horizontal coordinate is the ground truth, and the vertical is the reconstructed value. Perfect reconstructions

would be on the diagonal line. Purple dots depict white matter data and green dots gray matter data. The average errors are also shown below the plots

Fig. 5 Reconstructions of macroparameters K_1 , V_D and K_i of the 12k coincidence measurement. The horizontal coordinate is the ground truth, and the vertical is the reconstructed value. Perfect reconstructions

would be on the diagonal line. Purple dots depict white matter data and green dots gray matter data. The average errors are also shown below the plots

Fig. 6 Reconstructed TACs and kinetic parameters using different refinements of the linear parameters together with the Levenberg–Marquardt algorithm

which is a maximum likelihood estimation in case of Gaussian distribution. This overdetermined linear equation can be solved using the Moore–Penrose pseudo-inverse:

$$\tilde{\mathbf{x}}_F = (\mathbf{B}^T \cdot \mathbf{B})^{-1} \cdot \mathbf{B} \cdot \mathbf{y}_F. \quad (24)$$

Executing this step for every frame, we have a discrete time activity function for every region. The next step is to obtain

the initial parameters. Thanks to the down-sampling from voxels to a significantly smaller number of regions, sophisticated global optimization methods can be applied at this point. Rate constants α_i are the essentially nonlinear part of the concentration function. If rate constants are known, the concentration depends on the other parameters linearly, which could be determined at least approximately with the solution of a linear system. For the initial estimation of the

Fig. 7 Higher statistic measurement: The relative L_2 error of the reconstructed TAC from 120k coincidences using different refinement strategies for the linear parameters and taking the Levenberg–Marquardt algorithm

Fig. 8 Lower statistic measurement: The relative L_2 error of the reconstructed TAC from 12k coincidences using different refinement strategies for the linear parameters and taking the Levenberg–Marquardt algorithm

rate constants, we use either a grid of the possible values or explore this space with simulated annealing. When we take a point in this space, the rate constants are given a coordinates and other linear parameters are obtained by the discussed method for the linear subgroup, and the Kullback–Leibler-like error term (Eq. (3)) is evaluated and assigned to the point of rate constants. The goal of visiting grid points or simulated annealing is to find that point of rate constant where this fitting error is minimal.

So far, the results are based on the Gaussian assumption, which is justified by working with larger regions and not individual voxels. The initial estimates can be further refined by considering the Poisson model. It means that a few region-based ML–EM iterations are executed. As the number of regions is significantly less than the number of voxels, the complete initial guess has negligible computation time with respect to the reconstruction.

7 Results of 2D reconstruction

The proposed methods are tested first with a 2D phantom where the number of LORs is $N_L = 10k$ and the number of voxels is $N_V = 1024$ (Fig. 2). The 2D phantom has four anatomic parts, namely the air, gray matter, white matter and the blood. Parameters of the two-tissue compartment model $[f_v, a_1, a_2, \alpha_1, \alpha_2]$ used in the test are listed in Table 1 as reference values. Figure 6 depicts the reference blood input function and the TAC in different tissues. The total number of coincidences during the 10 second long measurement time is 120k. The measurement time is decomposed to 20 frames. The average number of coincidences per LOR per frame is about 0.6; thus, it would be a low statistic data for static reconstruction. In order to have an even lower statistic measurement, we also investigate the case of 12k coincidences. In this 2D test, we applied the method of sieves as a regularization, i.e., executed an anatomy-aware smoothing after each iteration step.

7.1 Initial guess of parameters

The initial guess assigns initial parameters to regions, so such regions should be defined. Now we use the anatomic parts as regions. Thus, to get the first guess of the time activity functions, just a linear system of Eq. (24) needs to be solved in each frame. The initial parameter fitting generates 10^4 samples in the 2D domain of the two rate constants, and the remaining linear parameters are found by solving a linear system of Eq. (21) for each sample. Figure 3 shows the error maps of different regions, where the color of a point depicts the error of the best fit when the two rate constants are defined by the coordinates of the point. This figure demonstrates that the search space is indeed complicated and has many local minima and maxima; thus, a robust global optimization method is necessary to initialize the rate constants. Then, to move toward the Poisson model, 20 region-based ML–EM iterations are executed.

Table 1 lists the resulting parameters after the global optimization and the region-based ML–EM iterations involving weighted refinement. Note that even these are fairly accurate since the test case meets the assumption that regions are uniform.

In the second round of tests, we used the described region-based initial parameter guessing method for sets of reference K_1, k_2, k_3, k_4 kinetic parameters, which are generated randomly in the $[0.02, 2]$ interval. With the random reference parameters, we have simulated 100 measurements of total coincidences between 14k and 900k. Reconstructed macroparameters K_1, V_D and K_i are paired with the ground truth values and depicted as points in 2D plots of Fig. 4. This figure compares three options including also the average

errors of the reconstructed parameters. In case of “ML–EM iterations,” the proposed initial guess is executed assuming a single region that includes all voxels, and then, 20 ML–EM iterations are executed with strong anatomy-aware regularization. The results of “Initial guess” are obtained executing the region-based initial parameter estimation. Finally, the method of “Initial guess + ML–EM iterations” refines the results of the initial estimation with 20 region-based ML–EM iterations. Let us observe the macroparameter error reduction due to the initial estimation and the added region-based ML–EM iteration. The same test has also been executed having reduced the activity to its tenth. The results from which similar conclusions can be drawn are shown in Fig. 5.

7.2 Fitting during the iterative reconstruction, without region-based initialization

In these tests, we investigate the effect of the refinement of linear parameters. The initial estimation is turned off, i.e., the volume is handled as a single region during parameter initialization. The K_1 and V_d parametric images as well as the mean and the standard derivation of time activity functions are obtained after 50 iterations. In a single ML–EM iteration, two Levenberg–Marquardt sub-iterations are executed, which may be followed by the proposed linear or weighted refinement step. The added computation cost of the refinement is less than ten percent of a single Levenberg–Marquardt sub-iteration. Figure 6 shows TACs with reconstructed region activity average and standard deviation, and demonstrates the effect of the refinement of the linear parameters on the reconstructed time activity curves and macroparameter maps. The relative RMS error of the TAC curve of the Levenberg–Marquardt method is reduced from 13.9% to 9.1% and to 8.4% by the added linear refinement and weighted refinement, respectively. The difference is especially noticeable in the [0,1] second interval of the gray matter and white matter time activity functions in Fig. 6.

The reconstruction errors as functions of the number of iteration steps are depicted in Figs. 7 and 8, demonstrating that the refinement is worth executing as it further reduces the error. The weighted refinement is only slightly better than the linear refinement and is worth using in case of low statistic measurements.

8 Dynamic 3D reconstruction

The performance of the proposed method in 3D dynamic reconstruction is demonstrated with input data obtained by

simulating the measurement of the Zubal phantom [33] by GATE [6] assuming the Mediso NanoScan PET/CT scanner. The “measured data” are reconstructed with our implemented system at $128 \times 128 \times 64$ voxel resolution and compared to the ground truth. The 10 second long measurement time is partitioned into 20 frames of lengths inversely proportional to the activity. The brain phantom has nine different homogeneous regions, including gray matter, white matter, cerebellum, caudate nucleus, putamen, bone, skin, blood and air. In these tests, 20 full ML–EM iterations are executed, while we applied total variation (TV) regularization of the same parameter in all cases. Our method is orthogonal to the applied spatial regularization, so other regularization approaches like Bregman iteration [21], anisotropic diffusion [23] or sparse representation-based techniques could also be incorporated [9,10,32].

The proposed method including optional initial estimation and weighted refinement in iterations is compared to nonparametric reconstruction when frames are reconstructed independently and to the original nested ML–EM method. In our initial estimation, four major regions are distinguished: air, bone + skin, cerebellum and the composition of everything else. The TACs are shown in Fig. 9, and the reconstructed activities in frame 5 in Fig. 10. Direct, i.e., parametric reconstructions are superior to indirect, i.e., static reconstruction in all cases. The proposed initial estimation and refinement are also beneficial and improves the reconstruction with respect to the original nested ML–EM scheme. Note, for example, the more accurate TAC of the putamen and especially the caudate nucleus, and also the sharper boundary of the gray matter.

The execution times of different steps on an NVIDIA GeForce RTX 2080 GPU are shown in Table 2. The initial estimation time counts only once, the fitting time should be multiplied by the number of iterations, forward and back projection times by the number of frames and the number of iterations. The time of refinement is about 6% of the time of fitting. A full iteration with 20 frames takes 60 seconds, and the results are obtained by 20 iterations; thus, the 5 second long time required by the initial estimation is negligible.

Fig. 9 Reconstructed TACs of the Zubal phantom with 20 iterations

Fig. 10 Reconstructed activities in frame 5 of the Zubal phantom and the relative L_2 error of all voxels and all frames

Table 2 Running times

Initial estimation	5 sec
Forward projection	0.62 sec
Back projection	2.21 sec
ML-EM fitting	0.17 sec
Refinement	0.01 sec

9 Conclusions

This paper proposed improvements to the nested ML-EM algorithm to robustly and efficiently fit compartment models to noisy data. Both the initial estimation and the steps of the iterative solution have been addressed. We also discussed an analytic approach to compute the parameters of

the Levenberg–Marquardt solver, which is not only precise but also much faster than the numerical estimation of the integrals of convolutions. During the iterative solution, a refinement step is included which proposes a modified value after the Levenberg–Marquardt solution, which may or may not be accepted using the local fitness criterion. All steps of the method are appropriate for GPU implementation. We have shown results obtained with a 2D phantom and also 3D reconstructions of GATE simulated data. This solution is integrated into the TeraTomo system [13].

Funding Open Access funding provided by Budapest University of Technology and Economics

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Baydin, A.G., Pearlmutter, B., Radul, A.A., Siskind, J.: Automatic differentiation in machine learning: a survey. *J. Mach. Learn. Res.* **18**, 1–43 (2018)
2. Cheng, X., Li, Z., Liu, Z., Navab, N., Huang, S.C., Keller, U., Ziegler, S.I., Shi, K.: Direct parametric image reconstruction in reduced parameter space for rapid multi-tracer pet imaging. *IEEE Trans. Med. Imaging* **34**(7), 1498–1512 (2015)
3. Feng, D., Huang, S.-C., Wang, X.: Models for computer simulation studies of input functions for tracer kinetic modeling with positron emission tomography. *Int. J. Bio-Med. Comput.* **32**(2), 95–110 (1993)
4. Gallezot, J., Lu, Y., Naganawa, M., Carson, R.E.: Parametric imaging with pet and spect. *IEEE Trans. Radiat. Plasma Med. Sci.* **4**(1), 1–23 (2020)
5. Gunn, R.N., Gunn, S.R., Cunningham, V.J.: Positron emission tomography compartmental models. *J. Cereb. Blood Flow Metab.* **21**(6), 635–652 (2001). PMID: 11488533
6. Jan, S., et al.: GATE: A simulation toolkit for PET and SPECT. *Phys. Med. Biol.* **49**(19), 4543–4561 (2004)
7. Kadrmas, D.J., Oktay, M.B.: Generalized separable parameter space techniques for fitting 1k–5k serial compartment models. *Med. Phys.* **40**(7), 072502 (2013)
8. Kamasak, M.E., Bouman, C.A., Morris, E.D., Sauer, K.: Direct reconstruction of kinetic parameter images from dynamic pet data. *IEEE Trans. Med. Imaging* **24**(5), 636–650 (2005)
9. Li, X., Shen, H., Zhang, L., Zhang, H., Yuan, Q., Yang, G.: Recovering quantitative remote sensing products contaminated by thick clouds and shadows using multitemporal dictionary learning. *IEEE Trans. Geosci. Remote Sens.* **52**(11), 7086–7098 (2014)
10. Liao, H. Y., Sapiro, G.: Sparse representations for limited data tomography. In: 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 1375–1378, (2008)
11. Liao, W., Lange, K., Bergsneider, M., Huang, S.: Optimal design in dynamic pet data acquisition: a new approach using simulated annealing and component-wise metropolis updating. *IEEE Trans. Nucl. Sci.* **49**(5), 2291–2296 (2002)
12. Liu, X., Hou, F., Hao, A., Qin, H.: A parallelized 4d reconstruction algorithm for vascular structures and motions based on energy optimization. *Vis. Comput.* **31**(11), 1431–1446 (2015)
13. Magdics, M., et al.: TeraTomo project: a fully 3D GPU based reconstruction code for exploiting the imaging capability of the NanoPET/CT system. In: World Molecular Imaging Congress, (2010)
14. Matthews, J. C., Angelis, G. I., Kotasidis, F. A., Markiewicz, P. J., Reader, A. J.: Direct reconstruction of parametric images using any spatiotemporal 4d image based model and maximum likelihood expectation maximisation. In: IEEE Nuclear Science Symposium Medical Imaging Conference, pp. 2435–2441 (2010)
15. Moré, J.J.: The levenberg-marquardt algorithm: implementation and theory. In: Watson, G.A. (ed.) *Lecture Notes on Computer Science*, Springer, vol. 630, pp. 105–116 (1978)
16. Rahmim, A., Tang, J., Mohy-ud Din, H.: Direct 4d parametric imaging in dynamic myocardial perfusion pet. *Front. Biomed. Technol.* **1**(1), 4–13 (2014)
17. Rahmim, A., Tang, J., Zaidi, H.: Four-dimensional (4d) image reconstruction strategies in dynamic pet: beyond conventional independent frame reconstruction. *Med. Phys.* **36**(8), 3654–70 (2009)
18. Reader, A.J., Matthews, J.C., Sureau, F.C., Comtat, C., Trebussen, R., Buvat, I.: Iterative kinetic parameter estimation within fully 4d pet image reconstruction. In: IEEE Nuclear Science Symposium Conference Record vol. 3, pp. 1752–1756 (2006)
19. Reader, A.J., Verhaeghe, J.: 4d image reconstruction for emission tomography. *Phys. Med. Biol.* **59**(22), R371 (2014)
20. Reutter, B.W., Sang Oh, Gullberg, G.T., Huesman, R.H.: Improved quantitation of dynamic spect via fully 4-d joint estimation of compartmental models and blood input function directly from projections. In: IEEE Nuclear Science Symposium Conference Record, vol. 4, pp. 2337–2341, (2005)
21. Szirmay-Kalos, L., Tóth, B., Jakab, G.: Efficient Bregman iteration in fully 3d PET. In: IEEE Nuclear science symposium and medical imaging conference, MIC' 14, (2014)
22. Szirmay-Kalos, L., Kacsó, Á., Magdics, M., Tóth, B.: Dynamic pet reconstruction on the GPU. *Period. Polytech. Electr. Eng. Comput. Sci.* **62**(4), 134–143 (2018)
23. Szirmay-Kalos, L., Magdics, M., Tóth, B.: Volume enhancement with externally controlled anisotropic diffusion. *Vis. Comput.* **33**, 331–342 (2017)
24. Wang, C., Yang, J.: Poisson noise removal of images on graphs using tight wavelet frames. *Vis. Comput.* **34**(10), 1357–1369 (2018)
25. Wang, G., Qi, J.: Direct estimation of kinetic parametric images for dynamic pet. *Theranostics* **3**(10), 802–815 (2013)
26. Wang, G., Rahmim, A., Gunn, R.N.: Pet parametric imaging: Past, present, and future. *IEEE Trans. Radiat. Plasma Med. Sci.* **4**(6), 663–675 (2020)
27. Wang, G., Qi, J.: Acceleration of the direct reconstruction of linear parametric images using nested algorithms. *Phys. Med. Biol.* **55**(5), 1505–1517 (2010)

28. Wang, G., Qi, J.: An optimization transfer algorithm for nonlinear parametric image reconstruction from dynamic pet data. *IEEE Trans. Med. Imaging* **31**(10), 1977–1988 (2012)
29. Watabe, H., Ikoma, Y., Kimura, Y., Naganawa, M., Shidahara, M.: Pet kinetic analysis—compartmental model. *Ann. Nucl. Med.* **20**(9), 583 (2006)
30. Yan, J., Planeta-Wilson, B., Carson, R.E.: Direct 4-d pet list mode parametric reconstruction with a novel em algorithm. *IEEE Trans. Med. Imaging* **31**(12), 2213–2223 (2012)
31. Yan, J., Planeta-Wilson, B., and Carson, R.E.: Direct 4d list mode parametric reconstruction for pet with a novel em algorithm. In: *IEEE Nuclear Science Symposium Conference Record, NSS '08*, pp. 3625–3628, (2008)
32. Ye, J., Wang, H., Yang, W.: Image reconstruction for electrical capacitance tomography based on sparse representation. *IEEE Trans. Instrum. Meas.* **64**(1), 89–102 (2015)
33. Zubal, I.G., Harrell, C.R., Smith, E.O., Rattner, Z., Gindi, G., Hoffer, P.B.: Computerized three-dimensional segmented human anatomy. *Med. Phys.* **21**(2), 299–302 (1994)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

László Szirmay-Kalos was graduated from the Budapest University of Technology in 1987, received Ph.D. and Doctor of Science degree from the Hungarian Academy of Science in 1991 and in 2001, respectively. He is currently a full professor of computer graphics at Budapest University of Technology. His research interests include rendering, Monte Carlo methods and medical imaging. He is the fellow of Eurographics. His webpage is <https://www.iit.bme.hu/users/dr-szirmay-kalos-l%C3%A1szl%C3%B3?language=en>

Ágota Kacsó is a PhD student at the Budapest University of Technology. Her research areas include fluid simulation with medial applications, retina response identification and positron emission tomography. Web page: <https://www.iit.bme.hu/users/kacs%C3%B3-%C3%A1gota-enik%C5%91?language=en>

Milán Magdics was graduated from the Eötvös Lóránd Science University and received Ph.D. from the Budapest University of Technology in 2014 in medical imaging. He is now involved in Monte Carlo methods, medical imaging and computational physics and works as an associate professor. He is member of Eurographics. His web page is <https://www.iit.bme.hu/users/dr-magdics-mil%C3%A1n?language=en>

Balázs Tóth works at the Budapest University of Technology from where he obtained his Ph.D. degree in 2010 in computer graphics and game technology. He is currently an associate professor at the same institution. He is member of Eurographics. His web page is <https://www.iit.bme.hu/users/dr-t%C3%B3th-bal%C3%A1zs-gy%C3%B6rgy?language=en>