Reconstruction, analysis and interpretation of posterior probability distributions of PET images, using the posterior bootstrap

Marina Filipović, Thomas Dautremer, Claude Comtat, Simon Stute, Eric Barat

Physics in Medicine & Biology

PAPER • OPEN ACCESS

Reconstruction, analysis and interpretation of posterior probability distributions of PET images, using the posterior bootstrap

To cite this article: Marina Filipovi *et al* 2021 *Phys. Med. Biol.* **66** 125018

IOP Publishing Phys. Med. Biol. 66 (2021) 125018 https://doi.org/10.1088/1361-6560/ac06e1

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
19 January 2021

REVISED
4 May 2021

ACCEPTED FOR PUBLICATION
1 June 2021

PUBLISHED
17 June 2021

PAPER

Reconstruction, analysis and interpretation of posterior probability distributions of PET images, using the posterior bootstrap

Marina Filipović[1], **Thomas Dautremer**[2], **Claude Comtat**[1], **Simon Stute**[3,4] **and Éric Barat**[2]

[1] Université Paris-Saclay, CEA, CNRS, Inserm, BioMaps, Service Hospitalier Frédéric Joliot, Orsay, France
[2] CEA, LIST, Laboratory of Systems Modelling and Simulation, Gif-sur-Yvette, France
[3] Nuclear Medicine Department, University Hospital, Nantes, France
[4] CRCINA, INSERM, CNRS, Université d'Angers, Université de Nantes, Nantes, France

E-mail: marina.filipovic.work@gmail.com

Keywords: PET image reconstruction, PET/MRI, posterior probability distribution, posterior bootstrap, uncertainty quantification, Bayesian inference, multimodal image reconstruction

Supplementary material for this article is available online

Abstract

The uncertainty of reconstructed PET images remains difficult to assess and to interpret for the use in diagnostic and quantification tasks. Here we provide (1) an easy-to-use methodology for uncertainty assessment for almost any Bayesian model in PET reconstruction from single datasets and (2) a detailed analysis and interpretation of produced posterior image distributions. We apply a recent posterior bootstrap framework to the PET image reconstruction inverse problem and obtain simple parallelizable algorithms based on random weights and on existing maximum *a posteriori* (MAP) (posterior maximum) optimization-based algorithms. Posterior distributions are produced, analyzed and interpreted for several common Bayesian models. Their relationship with the distribution of the MAP image estimate over multiple dataset realizations is exposed. The coverage properties of posterior distributions are validated. More insight is obtained for the interpretation of posterior distributions in order to open the way for including uncertainty information into diagnostic and quantification tasks.

1. Introduction

In PET (Positron Emission Tomography) medical imaging, the raw data acquired by the scanner have a low SNR (Signal to Noise Ratio) and the noise is of Poisson type. In addition, the image reconstruction inverse problem is ill-posed. Standard PET image reconstruction methods model the noise and use some type of spatial regularization to mitigate the noise propagation from the data to the image (Qi and Leahy 2006). Usually, the reconstruction procedure produces a single image estimate, which depends on the reconstruction method used and on the tuning of method parameters. An image estimate, and thus any subsequent visual or quantitative analysis, depend on the given noisy dataset and on the characteristics of the chosen reconstruction method. Hence, there is a need for assessing the uncertainty of voxel values in reconstructed PET image estimates.

Two types of uncertainty have been explored in PET image reconstruction, providing answers to the following two different questions:

(i) Given an optimization-based iterative image reconstruction method that produces a single image estimate, if PET imaging is repeated many times on the same patient in the same state in the exact same conditions, what is the probability distribution of the reconstructed images over acquired datasets? This estimator distribution and its characteristics (e.g. estimator/ensemble bias, estimator/ensemble variance, confidence intervals) can be obtained from simulated data but not as easily from a single real dataset. Some characteristics (e.g. variance) have been previously approximated from a single dataset using either

analytical approximations, (Fessler 1996, Qi and Leahy 1999), or classical dataset bootstrap approaches, (Dahlbom 2001, Buvat 2002, Markiewicz *et al* 2014).

(ii) Given a single acquired dataset and a Bayesian model relating the unknown image to this dataset, what is the probability distribution of the image? This distribution, called a posterior image distribution, was previously produced for some models using either analytical approximations, (Zhang *et al* 2019), or Monte Carlo Markov chain (MCMC) samplers, (Sitek 2012, Filipović *et al* 2019) for PET and (Weir 1995, Higdon *et al* 1997) for SPECT.

Both these questions are relevant for most existing PET image reconstruction approaches.

The exact maximum likelihood (ML) PET image estimate, obtained for instance using the MLEM algorithm at convergence, is almost never used in practice because of high image noise, e.g. Jaskowiak *et al* (2005). Instead, various strategies are used to mitigate the image noise. For instance, MLEM iterations can be stopped before reaching convergence, as mostly lower spatial frequencies are reconstructed in the first iterations. Many optimization-based iterative methods either add a penalty/regularization term to the likelihood objective function to obtain a penalized ML solution or they add a prior image probability distribution, building a complete parametric Bayesian model of PET image reconstruction to obtain a posterior maximum (maximum *a posteriori* (MAP)) solution. These two approaches have a different theoretical interpretation but often result in equivalent algorithms and image estimates. They both have the purpose to introduce some assumptions about the smoothness, roughness and edges in the image. Images from other modalities having higher spatial resolution and lower noise (e.g. CT, MRI) have been used in the literature to aid this spatial regularization task (Bai *et al* 2013). The ML solution may be viewed as a MAP solution using a uniform image prior distribution, or equivalenty no prior information about the image. The work presented here belongs to the context of MAP approaches but can be extended to a larger context.

Here we attempt to answer the second uncertainty question defined above using a new statistical framework, called here for short the 'posterior bootstrap', which represents a synthesis from the following approaches: Newton *et al* (2020), Fong *et al* (2019), Lyddon *et al* (2019). It allows for drawing approximate samples from the posterior distribution of almost any Bayesian model by using random weights and existing MAP optimization algorithms. Hence, the border may appear blurred between optimization-based iterative reconstruction methods and sampling from posterior distributions, as well as between regularization/penalty and image priors.

Drawing samples from a posterior distribution is a daunting task in high-dimensional ill-posed inverse problems with correlated variables (Girolami and Calderhead 2011, Robert and Casella 2013), which is the case in PET image reconstruction. As an exemple, in our previous work, (Filipović *et al* 2019), we used a MCMC sampler, designed for a single though versatile type of spatially regularizing prior distribution (distance-dependent Chinese Restaurant Process), and experienced some MCMC convergence difficulties. Compared to MCMC, the posterior bootstrap avoids convergence issues, realization discarding, sequential computation, and the need for analytical reformulation for each Bayesian model. The posterior bootstrap is applicable to any Bayesian model, provided a corresponding MAP method is available, while remaining independent of the choice of the particular optimization algorithm. It is scalable with respect to the amount of data and to the complexity of the model and the realizations can be computed in parallel. However, the theoretical properties are different from MCMC samplers and the drawn samples remain approximate, as currently there are no proofs stating that they correspond exactly to the posterior distribution of the given Bayesian model when only a limited amount of data is available, see Fong *et al* (2019), Newton *et al* (2020) for the currently available proofs. The posterior bootstrap does not introduce additional analytical approximations. It is more general and has different theoretical interpretations compared to the classical data bootstrap, (Efron 1979).

The main aims of this work are to (1) provide a versatile methodology for uncertainty assessment for real data and for common approaches in PET reconstruction (2) provide a detailed analysis and interpretation of the obtained uncertainty. This is a necessary step before including the uncertainty information into diagnostic tasks. Redesigning the diagnostic tasks (e.g. lesion detection, lesion characterization, comparison of pathological and healthy tissues) and quantitative processing (e.g. kinetic modeling) in terms of uncertainty is a vast subject in itself and requires further exploration in close collaboration with physicians. As this work has strong explanatory purposes, the methods, the results and the discussion are interleaved throughout the paper.

2. Theory

Let us consider a list-mode dataset composed of K detected counts, where each count has some attributes, e.g. coordinates of the line-of-response (LOR) in which the count was detected, time-of-flight (TOF) measurement. The attribute values represent a realization of a random variable r, which is independent and identically distributed (iid) for each count, given the radiotracer emission concentration distribution in the patient (PET

image) λ. Let the space of attribute values be discrete, with i indexing available attribute values, j the voxels, k the counts. Let the list-mode dataset represent a realization from a Poisson point random process, see Barrett *et al* (1997) for a detailed definition. If detected counts are histogrammed into any kind of bins (e.g. LOR, sinogram, TOF), let a histogram dataset y represent a realization from a joint Poisson distribution with unknown parameters as in equation (1): the A system matrix contains probabilities that an annihilation occurred in voxel j is detected in the detection bin i and $\bar{q}_i$ is the expectation of the number of random and scattered counts

$$p(y|\lambda) = \prod_i \text{Poisson}\left(\sum_j A_{ij}\lambda_j + \bar{q}_i\right). \tag{1}$$

Let us build a Bayesian model that relates the unknown PET image λ to the acquired dataset (either list-mode r or histogram y): the probability distribution $p(data|\lambda)$ is the likelihood of the acquired dataset and a prior probability distribution of the image $p(\lambda)$ is specified in order to introduce assumptions about smoothness and roughness properties in the image. The posterior image distribution results from the prior and the likelihood distributions as $p(\lambda|data) \propto p(data|\lambda)p(\lambda)$. We thus update our prior beliefs about the image upon observing some actual acquired data.

It should be noted that most optimization-based iterative reconstruction methods in PET have a Bayesian model lurking inside. MAP methods produce an image estimate that maximizes the posterior image distribution. From now on we consider directly the natural logarithms of mentioned probability distributions. If $R(\lambda)$ is the log prior and $L(data|\lambda)$ the log likelihood, then the MAP image estimate $\hat{\lambda}$ is obtained as

$$\hat{\lambda} = \underset{\lambda}{\textbf{argmax}}\ L(data|\lambda) + R(\lambda). \tag{2}$$

The posterior bootstrap approach presented here results from several converging ideas and can be interpreted from several points of view (Newton and Raftery 1994, Rubin 1981, Newton *et al* 2020, Fong *et al* 2019, Lyddon *et al* 2019). Applying this framework to PET reconstruction requires to model the PET raw data using iid random variables, hence using the list-mode data format. The list-mode log likelihood, (Barrett *et al* 1997, Huesman *et al* 2000), takes into account all the detected counts at once, but can be transformed into a sum over counts as

$$L(r|\lambda) = \sum_k \left(\ln\left(\sum_j A_{i_k j}\lambda_j + \bar{q}_{i_k}\right) - \frac{1}{K}\sum_i\sum_j A_{ij}\lambda_j + \bar{q}_i\right). \tag{3}$$

To apply the posterior bootstrap, K positive random weights w_k (one per detected count, with $\sum_k w_k = 1$) are drawn from a chosen probability distribution and used to randomly perturb the contribution of each detected count to the log likelihood, to produce a randomly perturbed log likelihood L_w

$$L_w(r|\lambda) = \sum_k Kw_k\left(\ln\left(\sum_j A_{i_k j}\lambda_j + \bar{q}_{i_k}\right) - \frac{1}{K}\sum_i\sum_j A_{ij}\lambda_j + \bar{q}_i\right). \tag{4}$$

Maximizing the objective function $L_w(r|\lambda) + R(\lambda)$ over λ then produces $\tilde{\lambda}$, which represents an approximate realization from the posterior distribution of the image λ, and not a usual MAP image estimate $\hat{\lambda}$. By repeating this process of drawing weights and maximizing the obtained objective function B times, we produce a sample of B realizations from the posterior image distribution.

Now let us consider a histogram dataset y and its log likelihood

$$L(y|\lambda) = \sum_i y_i \ln\left(\sum_j A_{ij}\lambda_j + \bar{q}_i\right) - \sum_i\sum_j A_{ij}\lambda_j + \bar{q}_i. \tag{5}$$

The randomly perturbed list-mode log likelihood in equation (4) can be transformed into an expression for histogram log likelihood as

$$L_w(y^*|\lambda) = \sum_i \ln\left(\sum_j A_{ij}\lambda_j + \bar{q}_i\right)\sum_{k\in S_i} Kw_k - \sum_i\sum_j A_{ij}\lambda_j + \bar{q}_i. \tag{6}$$

This is equivalent to the expression for the log likelihood of a different histogram dataset y^*, where $y_i^* = \sum_{k\in S_i} Kw_k$, S_i being the set of counts detected in bin i.

Maximizing the objective function $L_w(y^*|\lambda) + R(\lambda)$ over λ also produces an approximate realization $\tilde{\lambda}$ from the posterior image distribution. Repeating this process B times generates a sample of size B from the posterior image distribution. A randomized histogram y^* is obtained by drawing a new number of counts for each histogram bin, by drawing a realization from a distribution derived from the chosen distribution for the weights

w and from the original histogram dataset y. The choice of the weights distribution and the demonstration of the resulting histogram resampling procedure is described and discussed in what follows.

2.1. Interpretation and implementation of the weights w

The weights w have several intuitive interpretations. They can be viewed as a way to randomly perturb the contribution of each data realization (count) to the overall likelihood, as in Newton and Raftery (1994). They can also be viewed as probabilities that we assign to each data realization (count) in the dataset, before we perform some Bayesian modeling as described above: the K weights themselves represent a probability distribution. This is similar to the idea behind the Bayesian bootstrap (Rubin 1981) and the classical bootstrap (Efron 1979). In the classical bootstrap for list-mode datasets, a bootstrapped dataset is obtained by drawing randomly K counts with replacement from the original dataset. This is equivalent to assigning the same probability $1/K$ to each data realization (count). In the Bayesian bootstrap, no longer a fixed but a random probability is assigned to each data realization. These assigned probabilities can be used either analytically to perform some modeling/inference/computation, or they can be used to bootstrap the original dataset before performing some further modeling/inference/computation.

The weights w can also be viewed as an expression of uncertainty about the Bayesian model itself, as formally defined in Fong *et al* (2019) and Lyddon *et al* (2018). Every model is misspecified to some degree or does not match the reality perfectly, so it is relevant to express and include our belief/uncertainty about the postulated model itself, and the weights serve this purpose. More discussion about this interpretation will be given in the section 3.6.

The choice of a probability distribution from which to draw the weights w_k depends on their interpretation. As these weights represent themselves a probability distribution, the simplest choice is the uniform Dirichlet distribution with K parameters, Dir(1, 1, …, 1): a realization $(w_1, w_2, \ldots, w_K)$ drawn from this Dirichlet distribution represents itself a discrete probability distribution over K possible outcomes (counts). In this work, we choose this distribution and show how to draw the random weights for list-mode and histogram data in what follows.

The sampling from a Dirichlet distribution can be easily implemented using Gamma distributions: first, each random weight is drawn from Gamma(1, 1), and then the weights are normalized to satisfy $\sum_k w_k = 1$. The normalization can be avoided by making a simple approximation (see the appendix), resulting in $Kw_k \sim$ Gamma(1, 1). The algorithm for drawing a sample of B realizations from the posterior image distribution from a list-mode dataset is given in algorithm 1. The objective function remains the same as for the original list-mode dataset except for the multiplicative weights. These weights do not modify the properties of the objective function (e.g. derivatives, convexity), so any usual numerical solution algorithm can be applied by taking into account the weights.

For histogram data, it follows (see the appendix) that the number of counts ($\in \mathbb{R}$) for each randomized histogram bin y^*_i is first drawn from Gamma(y_i, 1), and then the randomized histogram is normalized to contain exactly the same number of counts as the original histogram. Again, this normalization step can be avoided by making a simple approximation (see the appendix), resulting in $y_i^* \sim$ Gamma(y_i, 1). The expectation of such a Gamma distribution per histogram bin is equal to the actual acquired number of counts in the histogram bin. It should be noted that this resampling does not imply nor assume that the histogram data follow such Gamma distributions. Also, the randomized histogram y^* is not assumed to represent a realization of acquired PET data (i.e. a repeated acquisition dataset) and it does not follow a joint Poisson distribution. This is not an issue, because the Poisson likelihood assumption for the original acquired dataset remains valid. The implementation of the posterior bootstrap for histogram data amounts to repeatedly resampling the dataset and applying usual MAP reconstruction methods, as shown in algorithm 2. The objective function to maximize remains the same as for the original dataset, so there are no modifications regarding the choice and the properties of numerical solution algorithms. This implementation of the posterior bootstrap is used in this work.

2.2. Algorithms

Algorithm 1. List-mode PET data.

1: **for** $b = 1$ to B **do**
2: draw Kw_{1b}, Kw_{2b},…Kw_{Kb} from Gamma (1, 1)
3: $\tilde{\lambda}_b = \arg\max_\lambda L_{w_b}(r|\lambda) + R(\lambda)$
4: **end for**

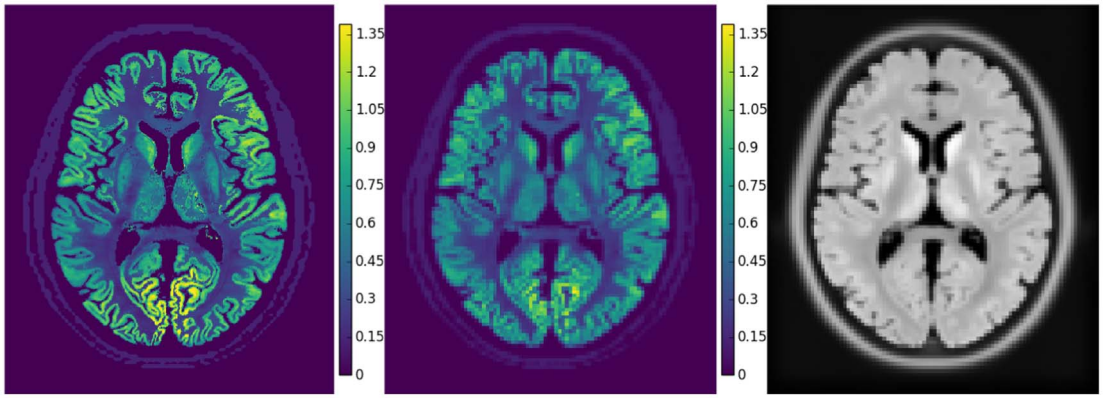

Figure 1. Left: high resolution PET phantom, Center and Right: PET phantom downsampled to the PET system resolution and the associated downsampled MRI image.

Algorithm 2. Histogram PET data.

1: **for** $b = 1$ to B **do**
2: draw dataset y_b, where each y_{ib} is drawn from Gamma (y_i, 1)
3: $\tilde{\lambda}_b = \arg\max_\lambda L_{w_b}(y_b^*|\lambda) + R(\lambda)$
4: **end for**

3. Methods and results

3.1. Phantom and simulation

We used a highly realistic ^{18}F-FDG PET/MRI brain phantom, (Belzunce and Reader 2020). The spatial resolution of the PET phantom (figure 1 left) is higher than the typical resolution of clinical PET scanners (≈800 μm compared to ≈4 mm FWHM) and the spatial distribution of the tracer emission concentration presents inhomogeneities instead of being piece-wise constant. The associated MRI image is real (acquired post-mortem, T1-weighted post-processed, as available from BigBrain, Amunts *et al* 2013), so that the matching of smooth areas and edges between PET and MRI images is imperfect, similarly to real exams. An in-house simulation library Stute *et al* (2015) was used to simulate dataset realizations (repeated dataset acquisitions or an ensemble of datasets) of PET histogram data, given the same high-resolution PET phantom. The Siemens Biograph 6 TruePoint PET/CT geometry (Jakoby *et al* 2009) was used and the bins in the histogram dataset represented the available LORs. The simulation included attenuation, random and scattered coincidences and resolution modeling using an image-domain point spread function (PSF). The simulation was 2D, the total number of counts (true, random and scattered) was 5*e*6, the number of noise equivalent counts was 1.5*e*6, approximating count rates ocurring in clinical brain ^{18}F-FDG PET exams for a 2D slice in the center of the axial field-of-view.

3.2. Models

The proposed posterior bootstrap approach was used to produce posterior image distributions for several Bayesian models commonly used in PET reconstruction. The likelihood being always the same, the difference between the models lies in the priors.

The choice of the prior is a widely discussed question in Bayesian approaches: the prior image probability distribution should convey the actual prior assumptions one might have about the PET image. The usual prior assumptions are related to the properties of smoothness/roughness in the image, based on local or nonlocal voxel neighbourhoods and possibly on additional data or images from other modalities (e.g. MRI, CT). Such prior image distributions represent a mathematical modelling of our (uncertain) prior assumptions about the smoothness/roughness in the PET image, and do not stand for an absolute truth with respect to the physical reality. The prior also conditions the interpretation of the posterior distribution, because the posterior distribution has to be understood and used while having in mind the chosen prior.

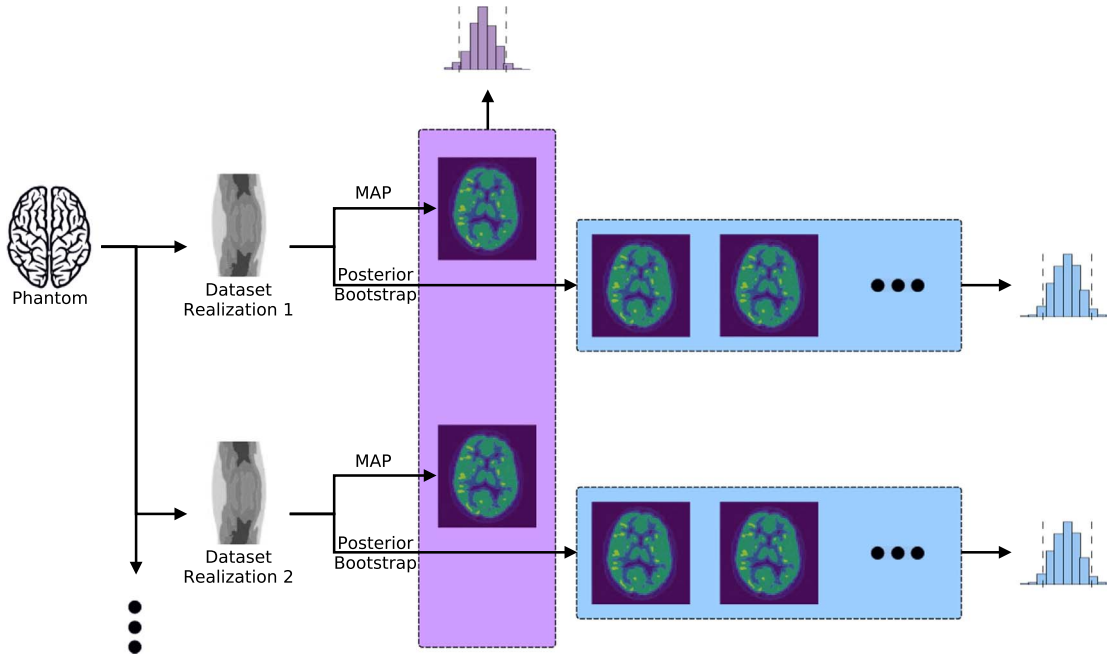

Figure 2. Illustration of the production of posterior distributions and of the distribution of the MAP image estimate over dataset realizations for a single Bayesian model, with simulated data.

- U: uniform prior, the optimization algorithm is MLEM.
- MRF-Q: Markov random field prior with the quadratic potential function, the MAP algorithm is MAP-EM, (De Pierro 1995).
- MRF-Q-MRI: the same with the addition of an associated MRI image, using the asymmetric Bowsher method, (Vunckx and Nuyts 2010).
- MRF-RD: Markov random field prior with the relative differences potential function, the MAP algorithm is the preconditioned gradient-based algorithm as in Nuyts *et al* (2002).
- MRF-RD-MRI: the same using an MRI image, using the asymmetric Bowsher method.

The posterior distributions in these models are convex and differentiable functions.

3.3. Reminder

A posterior image distribution differs by definition from the distribution of a MAP image estimate over dataset realizations. A posterior distribution expresses the uncertainty of the PET image, given the single acquired dataset and given the model (likelihood + prior, given the assumptions about the system matrix A and random/scattered coincidences). The distribution of a MAP image estimate over dataset realizations (estimator distribution) expresses the uncertainty of the MAP image estimator for a given model when data acquisition is repeated on the same patient in the exact same conditions. When several dataset realizations are available, it should be noted that for each Bayesian model there are several posterior distributions (each corresponding to a dataset realization), while there is only one distribution of the MAP image estimate over dataset realizations, see figure 2.

3.4. Implementation

All the MAP reconstruction methods were fully quantitative and contained image PSF resolution modelling and the corrections (random and scattered coincidences, attenuation, normalization). They were entirely implemented using the CASToR (Customizable and Advanced Software for Tomographic Reconstruction) platform in C++, (CAS 2017, Merlin *et al* 2018). The voxel size for reconstructed images was set to 2.2 mm $\times$ 2.2 mm $\times$ 2.8 mm to match the simulated PET scanner spatial resolution (4 mm FWHM). Figure 1 center shows the PET phantom downsampled to the PET scanner resolution. For MAP methods that use the associated MRI image, the MRI image was downsampled to the PET scanner resolution before being input into the reconstruction. The color scales for all the PET images related to the simulated data have the same unit as the PET phantom (relative uptake value, Belzunce and Reader 2020).

Table 1. Hyperparameter values.

	MRF-Q	MRF-RD	MRF-Q-MRI	MRF-RD-MRI
β	61	50	350	160
Neighbourhood sphere radius (mm)	3	3	8	8
Bowsher threshold	/	/	30%	30%
Relative differences γ	/	0	/	3

The number of posterior realizations (the size of the posterior sample) B for a single posterior distribution was set to 1000. The number of dataset realizations for the distribution of a MAP image over dataset realizations (estimator distribution) was also set to 1000. All the dataset realizations were reconstructed with all the different MAP reconstruction methods. As a posterior distribution is computed from a single dataset realization, posterior distributions were computed for 10 dataset realizations for each model, to appreciate how posterior distributions vary over dataset realizations. Figure 2 shows the overall simulation and reconstruction procedure and an illustration of posterior distributions versus the distribution of the MAP image over dataset realizations for a single model.

All the MAP (including ML) algorithms were run for 1000 iterations, both for MAP image reconstruction and inside the posterior bootstrap algorithm. The chosen number of iterations achieved an empirical convergence for MRF MAP algorithms, and resulted in a relatively early stopping for MLEM. MAP methods have in addition hyperparameters (parameters of the prior distribution) which tune the characteristics and the strength of prior smoothness/roughness assumptions, such as the general weight β in MRF priors (Qi and Leahy 2006), the tradeoff γ between allowing for edges and reducing noise for MRF-RD (Nuyts *et al* 2002), the size of the voxel neighbourhood (defined here as a sphere with a radius in mm (CAS 2017)) and the percentage of neighbourhood voxels selected by the Bowsher method (Bowsher *et al* 2004, CAS 2017). The hyperparameter values were fixed by running each MAP method on a single dataset for a wide range of hyperparameters values and choosing the values that minimize the RMSE with respect to the true phantom image downsampled to the PET system resolution. They are given in table 1.

3.5. Characteristics of posterior distributions

We explore and present the characteristics of posterior distributions, as well as their relationship with the distribution of the corresponding MAP solution over dataset realizations (estimator distribution).

3.5.1. Results overview

Figure 3 shows the voxel-wise posterior mean, posterior variance and posterior interval size for a dataset realization, for all the Bayesian models. Figure 4 shows the voxel-wise mean, variance and interval size corresponding to the distribution of the MAP image over dataset realizations (also called estimator mean and variance, and confidence intervals), for all the Bayesian models. The interval size for a voxel is computed as the difference between the maximum and minimum voxel intensity realization in the sample, see section 3.6 for more details.

3.5.2. Mean

The posterior mean image (obtained with the posterior bootstrap) is visually indistinguishable from the posterior maximum (MAP) image (obtained with the corresponding MAP algorithm, not shown), for the same dataset realization. The quantitative difference is low (root mean square difference < 0.01). This is due to the convexity and to a degree of symmetry of posterior distributions for the Bayesian models used here. Hence, the posterior mean shows here the same properties (e.g. image noise) as the MAP solution.

The estimator mean image (the average of MAP images over dataset realizations) presents lower image noise than a single MAP solution, which is expected because the estimator mean image approaches the MAP image reconstructed from noiseless data, as discussed in Fessler (1996). This estimator mean image is thus less noisy but shows the estimator bias. It could be argued that this image would be useful for diagnostic purposes, but it is not exactly obtainable from a single dataset.

The posterior mean shows the propagation of data noise due to the ill-posedness of the PET inverse problem, which depends on the given single noisy dataset and on the model. This implies that the posterior mean from one dataset realization may present visible differences compared to the posterior mean from a different dataset realization, as illustrated in figure 5. These differences are most obvious in the nuclei caudate area, possibly because of the proximity of high and low uptake (high contrast) and of a low system sensitivity mostly due to attenuation: for the first dataset realization, the nuclei have a similar intensity for all the models, and for the

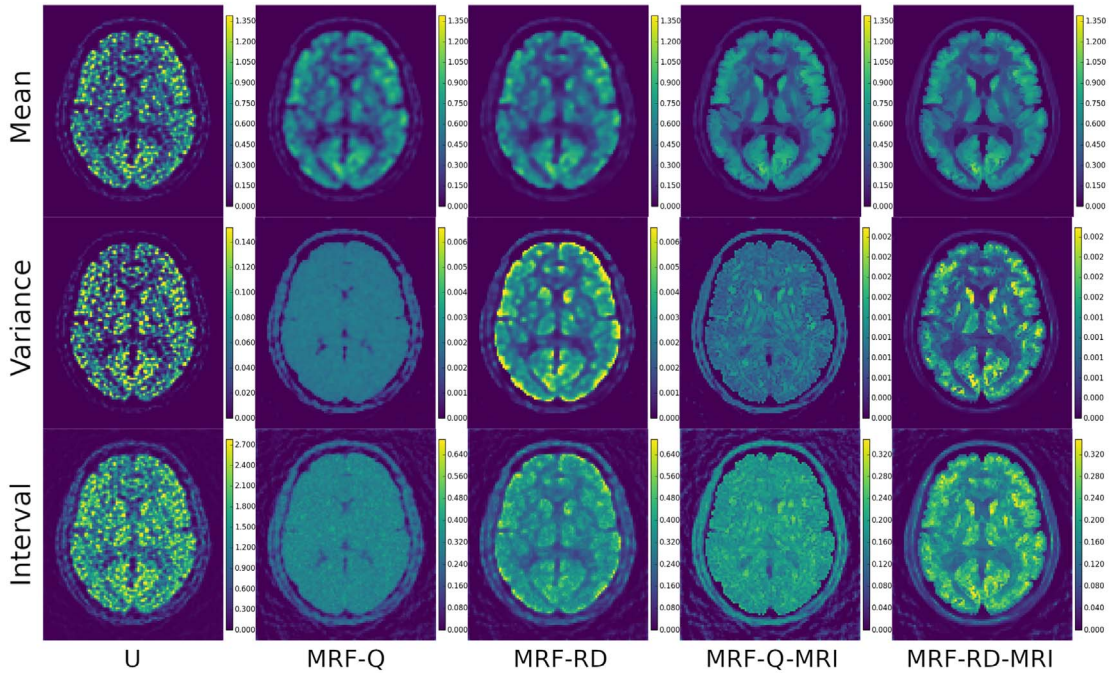

Figure 3. Posterior distribution: voxel-wise posterior mean, posterior variance, posterior interval size for a dataset realization, for different models. The colorscale maximum differs across models for the variance and the interval size.

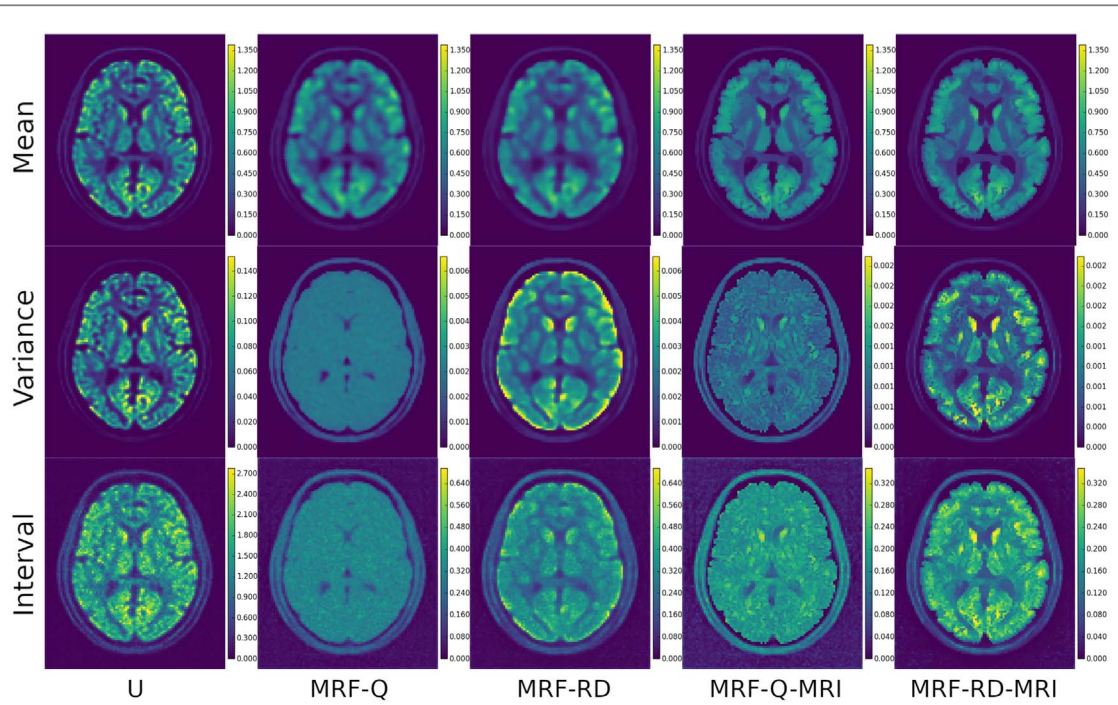

Figure 4. Distribution of the MAP image estimate over dataset realizations: voxel-wise estimator mean, estimator variance and confidence interval size for different models. The colorscale maximum differs across models for the variance and the interval size.

second dataset realization, the left nucleus has a visibly higher intensity than the right one. In other image regions, the differences between dataset realizations are less noticeable.

The differences between the posterior mean and the estimator mean should be kept in mind when interpreting other distribution characteristics, e.g. (co)variance, intervals.

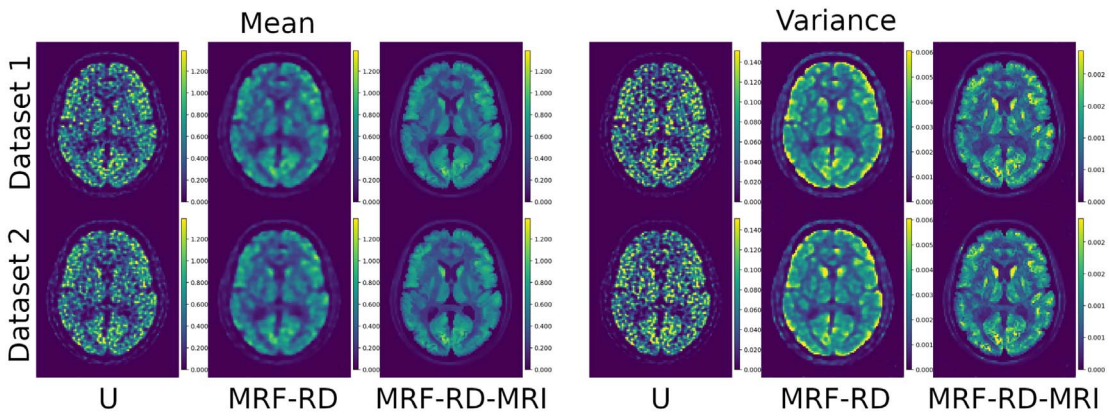

Figure 5. Posterior mean and variance for two different dataset realizations, for some models.

3.5.3. Variance

The voxel-wise posterior variance image is similar to the MAP estimator voxel-wise variance image in terms of structures, intensities and model-dependent characteristics. However, it presents more local variations and a 'noisier' appearance with respect to the estimator variance. The reason is that the posterior variance is meaningful with respect to the posterior mean, which shows some characteristics intrinsic to inverse problem noise propagation from the given single dataset realization, as illustrated in figure 5, while the estimator variance is meaningful with respect to an almost noiseless mean image. This effect is more visible for the U model than for models with spatial smoothness priors. Both types of variance decrease with stronger priors. This is expected because both types of variance are mostly due to the Poisson noise in the dataset and to its propagation in the inverse problem, while the spatial smoothness priors mitigate this noise. The overall variance intensity ranks from higher to lower for different models (having in mind the fixed hyperparameter values and the MLEM early stopping) as: U, MRF-RD, MRF-Q, MRF-RD-MRI, MRF-Q-MRI. The MRF-Q variance is rather flat, as already observed in Qi and Leahy (1999). The MRF-RD variance presents some local smoothness with amplification near some strong edges, e.g. the edge between the gray matter and the skull, which is reminiscent of the behaviour of total variation type regularization. The MRF-RD-MRI variance presents clearer edges, because of the influence of the MRI image, and higher intensities in some regions such as the nuclei caudate.

For the Bayesian models used here, and as obtained with the proposed posterior bootstrap approach, the posterior variance and more generally the uncertainty conveyed by the posterior distribution shows mostly the uncertainty related to the noise in the data and to its propagation through the chosen model. It conveys some spatial regularization properties of the models and so can be used to compare models between them. Higher variance may point out some areas in the image where the inversion or the spatial regularization struggle, for instance areas with strong contrast or sharp edges.

3.5.4. Covariance

The voxel-wise posterior covariance is similar to the estimator covariance for all the models in the same manner as the variance. It is more difficult to visualize because there are as many covariance images as there are voxels. An example of posterior covariance images is shown for a voxel in the gray matter in figure 6. It is consistent with the already known spatial regularization properties of the models used here. Without prior information (U), the covariance of a voxel with its neighbourhood is substantially lower than the covariance of the voxel with itself (its variance). The MRF-Q covariance is isotropically high within the voxel neighbourhood. The MRF-RD covariance is lower in edges and higher in smoother areas. The MRF-Q-MRI and MRF-RD-MRI covariance is high in a larger neighbourhood that appears smooth in the MRI image. The absolute covariance values decrease with stronger priors and some Gibbs ringing pattern can be seen around the voxel, especially for MRF-Q and MRF-RD.

3.5.5. Whole distribution

Examples of both posterior and estimator distributions for several gray matter voxels are given in figure 7 and for several nuclei caudate voxels in figure 8. For each voxel, the figures show two posterior distributions, obtained from two different dataset realizations, and the distribution of the MAP solution over dataset realizations, for different models. Both distributions for U are skewed and have long tails. The posterior distributions for U vary the most over dataset realizations, because of high image noise and dataset dependent noise propagation. The

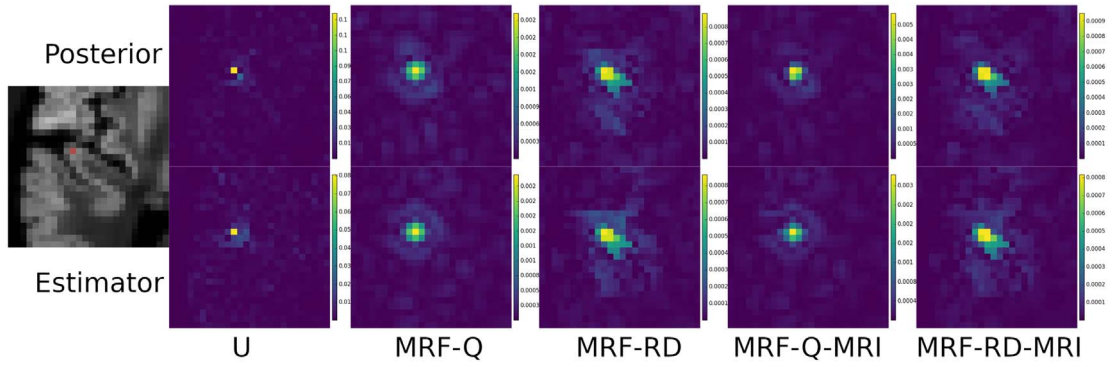

Figure 6. Posterior and MAP estimator covariance images for a gray matter voxel (shown in the left in red) for different models. The colorscale maximum differs across models.

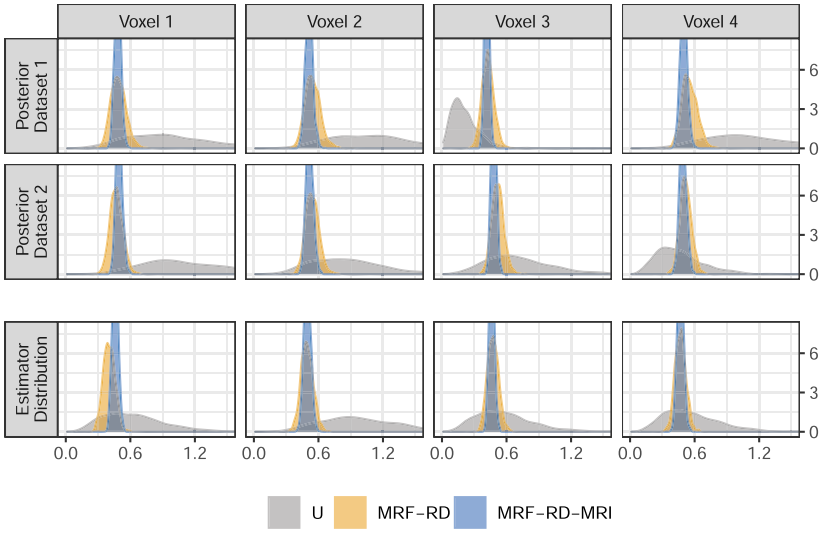

Figure 7. Examples of entire distributions for several gray matter voxels for 3 models: posterior distributions from two dataset realizations and the distribution of the MAP solution over dataset realizations.

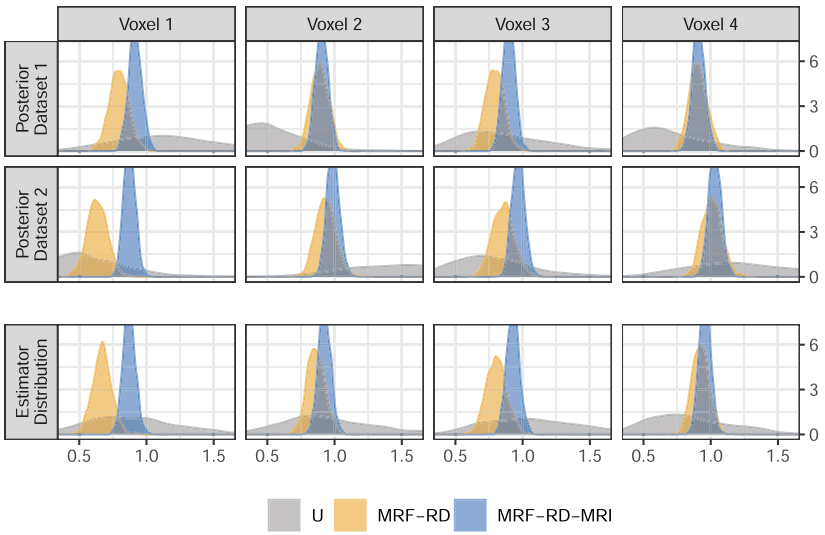

Figure 8. Examples of entire distributions for several nuclei caudate voxel for 3 models: posterior distributions from two dataset realizations and the distribution of the MAP solution over dataset realizations.

distributions tend to concentrate around their maximum with stronger spatial regularization and present slight skewness.

3.5.6. Discussion

Posterior distributions presented here convey mostly the uncertainty due to the noise in the data and to its propagation through the model. Hence, they cannot auto assess model misspecification (imperfect match to reality, e.g. system matrix A, mismatch of smooth areas in MRI and PET images), nor the model bias. However, the posterior bootstrap framework has an other interpretation in the context of misspecified models which can take into account the uncertainty of some modelling assumptions, but this is subject for future work, see section 3.6 question (ii) for more details. All the distributions in these models are convex, but it should be noted that the posterior bootstrap is applicable to non convex posteriors with possibly multiple maxima, provided a corresponding MAP optimization algorithm is available.

The posterior bootstrap framework implies that all the optimization algorithms are run to convergence. In practice, convergence is never strictly achieved due to limited computational resources and to the properties of optimization algorithms. MLEM and MAP-EM have strict convergence proofs, while the algorithm used for MRF-RD does not. MAP algorithms for MRF priors usually converge faster than MLEM. Also, U, MRF-Q and MRF-RD models are well defined Bayesian models, while the asymmetric Bowsher approach is an empirical modification of the original MRI Bowsher prior (Vunckx and Nuyts 2010). In this work, for models with MRF priors, an empirical convergence was achieved with the chosen number of iterations (1000). For the U model, the convergence would require much more iterations and would produce substantially noisier images, so the presented results correspond to a relatively early stopped MLEM. If MLEM were run to convergence, it could be argued that the uncertainty, both posterior and estimator, may not be useful for diagnostic and quantitative purposes, because it conveys an amount of uncertainty so high that it becomes useless (e.g. the variance becomes approximately proportional to the square of the mean image (Barrett *et al* 1994)). Some illustrative results using MLEM with much more (10 000) iterations are presented in the supplementary material available online at stacks.iop.org/PMB/66/125018/mmedia. The effect of choosing a lower number of iterations for MRF models depends on and can be anticipated according to the convergence behaviour of corresponding MAP algorithms for the given dataset. An example is shown in the supplementary material.

The choice for the posterior sample size B (the number of realizations drawn from the posterior distribution) depends on the shape of the posterior distribution and on which posterior characteristics are of interest (e.g. covariance, intervals, quantiles). For instance, if the posterior distribution has very long tails and we wish to characterize them accurately, a larger sample size will be needed. If we wish to estimate the variance, B can be lower. Also, convex posteriors put lower requirements on B than multimodal posteriors. In this work, we empirically observed that the posterior mean and variance stabilize beyond the chosen B (1000) for all the models. The effect of choosing a lower number of posterior realizations depends on and can be anticipated according to the overall shape of the posterior distribution and the posterior characteristic of interest for the given dataset. An example is shown in the supplementary material.

2D simulation was chosen over 3D for computational reasons: a large number of dataset realizations and of posterior realizations was required for a thorough analysis and it was of interest to show results for several different Bayesian models. The posterior bootstrap framework itself is independent of the dataset dimensions and of the influence of 2D versus 3D data on MAP solution and MAP algorithm performance.

Posterior distributions depend on the chosen values for the parameters of the image prior. The dependence of the MAP solution on hyperparameter values has already been studied in the literature and it applies also to the posterior mean. The (co)variance and the intervals tend to increase when the strength of prior smoothness/roughness assumptions descreases (lower β) and vice-versa. An example for the MRF-RD-MRI model is shown for lower and higher β values in the supplementary material. The general conclusions in this work do not depend on the choice of hyperparameter values.

3.6. Assessment

There are no standard methods for validating posterior distributions. Several questions can be addressed:

(i) Do the computed posterior realizations match the corresponding 'true' posterior distribution, where 'true' means with respect to the postulated Bayesian model and to the available dataset?

(ii) To what extent does the model match the reality?

(iii) Are the produced posterior distributions 'well calibrated', i.e. behave consistently in different cases (more explanations in the answer)?

(iv) How does the posterior bootstrap compare to other methods for generating posterior distributions?

It should be noted that there is no 'true' posterior distribution outside of the context of the given model and dataset. Bayesian inference can be viewed as updating our current prior assumptions about the PET image using the available acquired data, i.e. more as a method of reasoning and inference than a search for an absolute truth. Some answers and some discussion are provided in what follows.

(i) It should be noted that there are currently no established figures of merit for the performance of the posterior bootstrap and that there is no gold standard for posterior distribution estimation, especially in the context of high-dimensional ill-posed inverse problems such as PET. Some validation approaches consist in checking the relevance of posterior distributions by using them to predict new PET data, (Gelman *et al* 1996). The data predicted by the model can then be compared to the actual acquired dataset but it is not clear yet which comparison criteria would be most relevant and reliable for PET. This is material for future work.
Some theoretical proofs for the posterior bootstrap currently exist only for the asymptotic case, when the amount of data (the number of counts in the dataset) approaches infinity, (Fong *et al* 2019, Newton *et al* 2020).

(ii) A posterior distribution, as defined here, is based on the assumption that the underlying Bayesian model is true, in the sense that there exists an image λ, supported by the prior, for which the likelihood distribution generates the acquired data. However, the posterior bootstrap has an other theoretical interpretation in which the produced posterior distributions are exact (no longer approximate), but they do not have the same meaning. First, a vocabulary reminder. All the Bayesian models used in this and in previous works in PET image reconstruction are called parametric: the noisy acquired data are assumed to follow a probability distribution of known type (i.e. Poisson) and of unknown parameters. Another kind of Bayesian models is called nonparametric: the measured data follow an unknown type of probability distribution, which is itself a realization drawn from probability distributions capable of generating probability distributions (e.g. a Dirichlet process), so there are no direct notions of unknown parameters. The other interpretation of the posterior bootstrap is nonparametric and provides useful insights. The produced posterior distribution is exact (no longer an approximation) but refers to a different modelling context: the usual parametric Bayesian model is no longer viewed as true. It is instead viewed as misspecified to some degree, i.e. as an imperfect approximation of the reality: we assert openly that we are not sure about this parametric Bayesian model. We build a different independent nonparametric Bayesian model focused on the data distribution and use the parametric model only as an imperfect but convenient image estimator. The nonparametric Bayesian model contains a prior on the data distribution, which is here a uniform Dirichlet distribution (using weights w), which does not carry assumptions about the data probability distribution (except the i.i.d. assumption of list-mode counts). In this context, the produced posterior distributions represent posterior distributions of imperfect image estimators with respect to a noninformative prior (no prior assumptions) on the data distribution and with respect to the acquired dataset. See Fong *et al* (2019), Lyddon *et al* (2018, 2019) for more explanations. This interpretation can be viewed as a generalization of parametric posterior image distributions and of image estimators, and is material for further exploration of the uncertainty in PET reconstruction.

(iii) It is argued in Rubin (1984) and Bayarri and Berger (2004) that it is desirable in practice that posterior distributions be well calibrated, i.e. that the (Bayesian) posterior distributions meet (frequentist) estimator distributions in some aspects: for instance, that (e.g. 95%) posterior intervals contain the 'true' value in the same percentage (e.g. 95%) of 'cases', where a 'case' is a realization of the joint distribution of the acquired dataset and of the image. Some theoretical considerations and proofs about the meeting point between Bayesian and frequentist approaches for inverse problems such as in PET in the limit case of infinite amount of data are given in Bochkina and Green (2014).
First, a quick reminder about interval names:

- A 'posterior interval' refers to an interval on the posterior distribution of voxel intensity for each voxel, so each voxel has a posterior interval for each model and for each dataset realization.
- A 'confidence interval' refers to an interval on the MAP estimator distribution of voxel intensity over dataset realizations for each voxel, so each voxel has a confidence interval for each model.
Hence, for a given model, each voxel has a single confidence interval and several posterior intervals (each corresponding to a dataset realization), which can be confusing, see figure 2.
As here the number of realizations (1000) was relatively low from a statistical point of view and as the distribution tails are long due to the Poisson nature of the data, computing intervals with a precise percentage contents (e.g. 95%) would not be reliable. Hence, we settle for approximate intervals: we

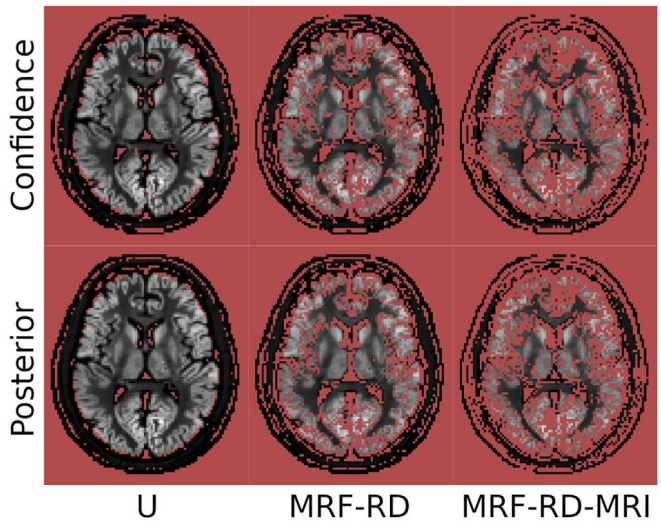

Figure 9. In red are shown the voxels for which posterior and confidence intervals fail to cover the true value, for different models.

Table 2. Average % of coverage of the true value by posterior and confidence intervals for different models.

	U	MRF-Q	MRF-RD	MRF-Q-MRI	MRF-RD-MRI
Posterior	90	70	67	58	51
Confidence	93	74	70	61	50

compute intervals for each voxel by taking the difference between the maximum and the minimum values in the sample and assume that these intervals correspond to approximate high percentage intervals (90%–100%), with some possible positive bias.

As the MAP estimator presents some estimator bias for all the models, the confidence intervals by definition contain the biased true value (the MAP estimator mean) in a high percentage of dataset realizations, but contain the actual true value (from the undersampled phantom) in a lower percentage of dataset realizations. The good calibration of posterior distributions implies that the coverage of the true and of the biased true value should match between confidence and posterior intervals. Figure 9 shows in red the voxels whose true value is not contained in confidence or posterior intervals. These maps are similar for confidence and posterior intervals and show mostly areas with 0 or very low uptake values (e.g. CSF, background), which can be explained by some positive bias in low uptake areas due to the positivity constraint inherent in the MAP algorithms.

We checked the following calibration properties of the posterior intervals:

- The posterior intervals should contain the biased true value in a high percentage of dataset realizations.
- The posterior intervals should contain the true value in the same percentage of dataset realizations as the confidence intervals.
- The posterior intervals should contain the biased true value in a high percentage of brain voxels.
- The posterior intervals should contain the true value in the same percentage of brain voxels as the confidence intervals.

Given the approximate intervals computation, we regard these properties as reasonably fulfilled. The coverage of the biased value is in average above 90%. The table 2 shows the average percentage of coverage of the true value by posterior and confidence intervals (for voxels with true value >0): the posterior intervals' coverage is a couple of percents lower than the confidence interval percentage. In terms of coverage, MRF-Q and MRF-RD models behave similarly, as well as MRF-Q-MRI and MRF-RD-MRI.

(iv) Relevant comparisons would consist in implementing other methods that produce posterior image distributions for the exact same Bayesian models. The widely used MCMC methods suffer from convergence issues, especially for high-dimensional complex models (Girolami and Calderhead 2011,

Robert and Casella 2013). In addition, direct comparison with previous work that used MCMC samplers is difficult. The previously proposed MCMC method by our group, (Filipović *et al* 2019), was designed for a different prior (ddCRP), which cannot be used in the posterior bootstrap framework because there are no corresponding MAP optimization algorithms to our knowledge. Some previous MCMC methods (Weir 1995, Higdon *et al* 1997) developed for SPECT reconstruction used complex samplers and similar priors though not identical to the ones used in this work. The origin ensemble method (Sitek 2011) does not produce directly posterior distributions, though additional steps were provided for producing posterior distributions for some specific priors in Sitek (2012), which are different from the common priors used in this work.

3.7. Preliminary discussion for using posterior image distributions

A posterior distribution may be closer to an intuitive understanding than an estimator distribution. It can be used to compute directly the probabilities of pathological features. It can also be obtained as a whole distribution, whereas one can currently obtain only the maximum and sometimes an analytically approximated co(variance) for an estimator distribution.

A posterior image distribution provides a posterior distribution for each voxel, as well as a posterior distribution of any voxel summary. A voxel summary refers to a function of several voxels (e.g. characteristics of regions of interest (ROIs)). A posterior distribution can be used to extract directly some relevant probabilities of interest, such as the probability that a voxel or ROI emission concentration is above a certain level or is higher/lower than some other voxel or ROI emission concentration. For such uses of posterior distributions, a clinically relevant task needs to be clearly formulated in terms of probabilities in close collaboration with physicians. Different Bayesian models may provide different answers to the same diagnostic question. A posterior image distribution can also be incorporated into kinetic modelling, by redefining kinetic models as Bayesian models, as for instance in Sitek *et al* (2016). We expect that posterior distributions will be useful in the cases of doubtful diagnosis (e.g. distinguishing a lesion from image noise) and in the case of any quantitative analysis. When redesigning diagnostic questions in terms of uncertainty, it should be kept in mind that the voxels are correlated and that their covariance depends on the model used, as shown in figure 6. It should also be kept in mind that these models contain some bias, whose value depends on model components and on data amount and which is difficult to deal with.

In this work, the simulated PET phantom image was produced using a real epilepsy brain ^{18}F-FDG exam with no associated specific diagnostic information, (Belzunce and Reader 2020), which allows for a general analysis on realistic data. Applications of posterior distributions in different cases of pathology and tracers is closely related to the redesign of diagnostic tasks in terms of uncertainty and is subject for future work. For instance, in view of applying posterior distributions on low-contrast lesion detection tasks, a different phantom could be built using the same methodology (Belzunce and Reader 2020) and real exams with confirmed lesions.

The choice of voxel summaries is not straightforward in the context of posterior distributions. A common voxel summary used with single estimate reconstruction methods is the mean of a ROI. In what follows, we discuss the use of the ROI mean in the context of posterior distributions.

3.7.1. ROI mean

The ROI mean has the usually desirable property of being less sensitive to the noise in the image than individual voxel intensities, though it remains sensitive to the estimator bias. This property may actually not be desirable when the aim is to take into account the uncertainty, because any type of variance is reduced artificially. It should be noted that models with stronger priors tend to produce smoother voxel intensities inside ROIs, so the ROI mean tends to be similar to individual voxel intensities in the ROI.

To illustrate the behaviour of posterior distributions of ROI means, two pairs of contralateral ROIs were drawn on the high resolution PET phantom and then downsampled to the PET system resolution, see figure 10. The spatial histograms for high-resolution ROIs show an aspect of the 'true' difference between the contralateral ROIs and presumably represent real situations better than strictly uniform ROIs. The spatial histograms overlap slightly for the gray matter ROI pair and strongly for the nuclei caudate ROI pair, representing respectively a rather different and a rather similar pair of ROIs. The spatial histograms for the ROIs downsampled to PET system resolution are not shown because they contain few voxels ($\sim$10).

Figure 11 shows the posterior ROI mean distribution for 2 dataset realizations and the distribution of the MAP ROI mean over dataset realizations, for several models. For each model, the distributions are rather similar: the posterior distributions do not vary substantially across dataset realizations and are similar to the estimator distribution. The overlap between the contralateral ROI distributions is low for U and almost non existent for models with spatially regularizing prior information. These observations can be explained by the

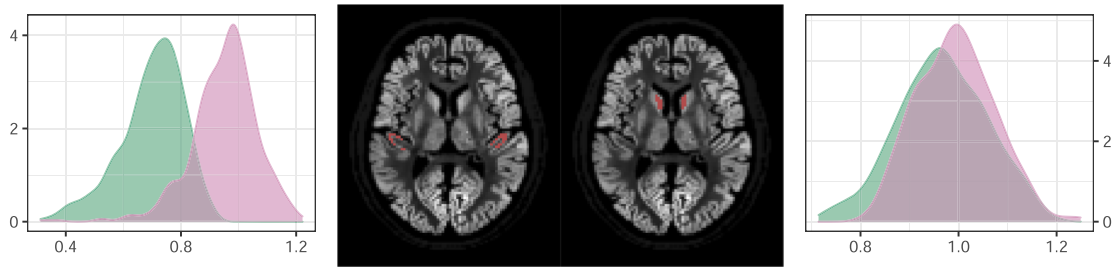

Figure 10. Contralateral ROI pairs and their spatial histograms: gray matter (left) and nuclei caudate (right).

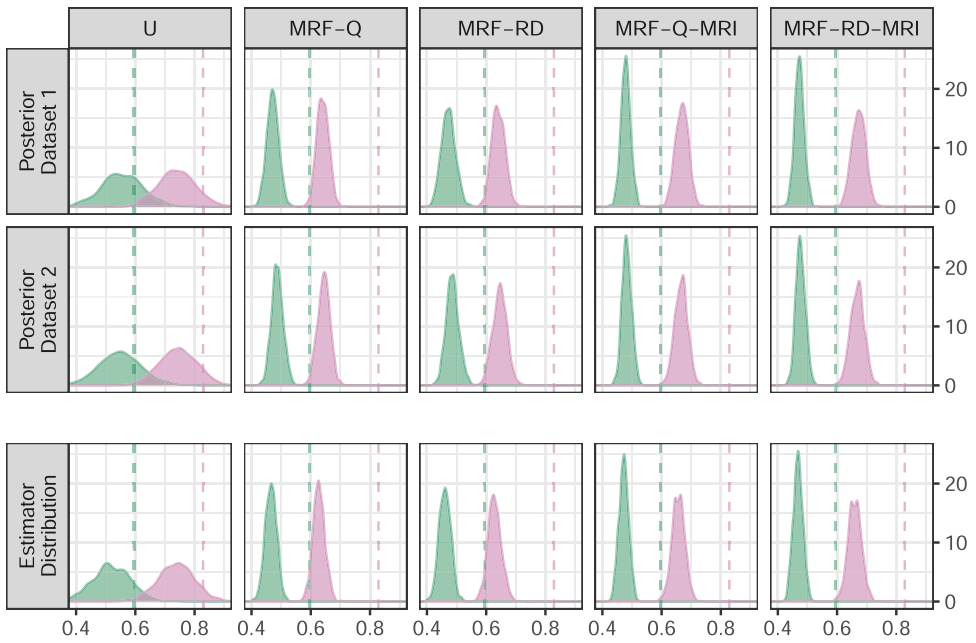

Figure 11. Posterior distribution of the ROI mean for the gray matter ROI pair, for 2 dataset realizations, and the distribution of the MAP ROI mean over data realizations, for different models. The dashed lines represent the true values (from the downsampled phantom).

tendency of the ROI mean summary to lower the variance. This effect is more visible for U, having a high image noise, than for the other models which produce rather smooth voxel intensities in the ROIs. The vertical dashed lines show the true values, as computed from the PET phantom downsampled to the PET system resolution. Models with MRF priors tend to be more biased than U for these ROIs.

Figure 12 shows ROI mean distributions for the nuclei caudate ROI pair. The overlap of posterior distributions varies across models and across data realizations. This is due to the dependency on the specific data realization and to a high variability in the nuclei caudate area, as explained in detail in section 3.5.2. The models with MRI tend to have a lower MAP estimator bias for these ROIs. Other voxel summaries could be explored or designed for taking full advantage of uncertainty information, in conjuction with redesigning various diagnostic tasks in terms of uncertainty.

3.8. Real data

Two real clinical exams were obtained from a GE Signa PET/MR scanner in the histogram data format, where the bins represented all the available LORs and TOF bins. The implementation of the reconstructions was identical to the one presented for simulated data (fully quantitative, including corrections and image PSF resolution modeling), except for the following: only the MRF-RD-MRI model was used, the reconstruction voxel size was 1.56 mm $\times$ 1.56 mm $\times$ 2.78 mm and the hyperparameters were set empirically to a subjective compromise between spatial regularization and conservation of PET specific features. Making a compromise between computation time and method accuracy (MAP algorithm convergence and posterior distribution characterization) resulted in a lower number of MAP iterations, the use of subsets, and a lower number of posterior realizations compared to the simulated data: the number of MAP iterations was 16, the number of

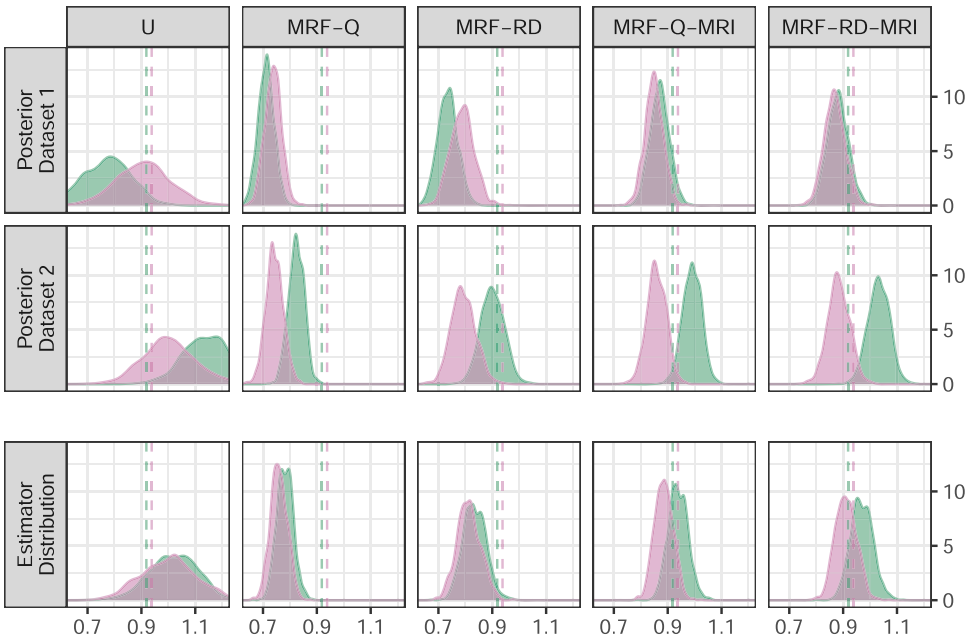

Figure 12. Posterior distribution of the ROI mean for the nuclei caudate ROI pair, for 2 dataset realizations, and the distribution of the MAP ROI mean over data realizations, for different models. The dashed lines represent the true values.

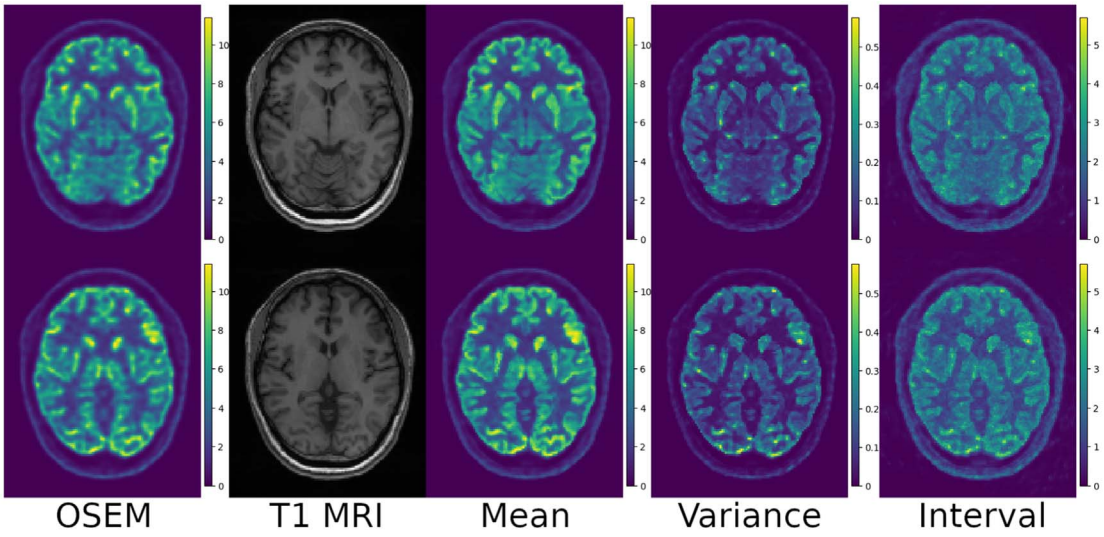

Figure 13. Neurodegenerative disease exam: (from left to right) clinical reconstruction, MRI image, and posterior mean, posterior variance and posterior intervals size for MRF-RD-MRI, for two axial slices.

subsets 28, the number of posterior realizations 400. A clinical reconstruction is also presented, using OSEM with 28 subsets and 8 iterations, without post-smoothing, using the same corrections and resolution modeling. The unit for all the reconstructed images is the standard uptake value (SUV).

Exam characteristics are given below:

- ^{18}F-FDG neurodegenerative disease exam (1.18*e*8 noise equivalent counts), showing no signs of pathology, with an associated 3D T1 weighted MRI image, and the following hyperparameter values: neighbourhood sphere radius = 6 mm, $\gamma = 3$, $\beta = 0.002$, Bowsher percentage = 30%.
- Brain bed step of a whole body ^{18}F-FDG oncological exam (4.5*e*7 noise equivalent counts), showing a metastatic lesion in the brain stem, with an associated 3D T1 weighted MRI image acquired after Gd contrast agent injection, and the following hyperparameter values: neighbourhood sphere radius = 6 mm, $\gamma = 3$, $\beta = 0.005$, Bowsher percentage = 30%.

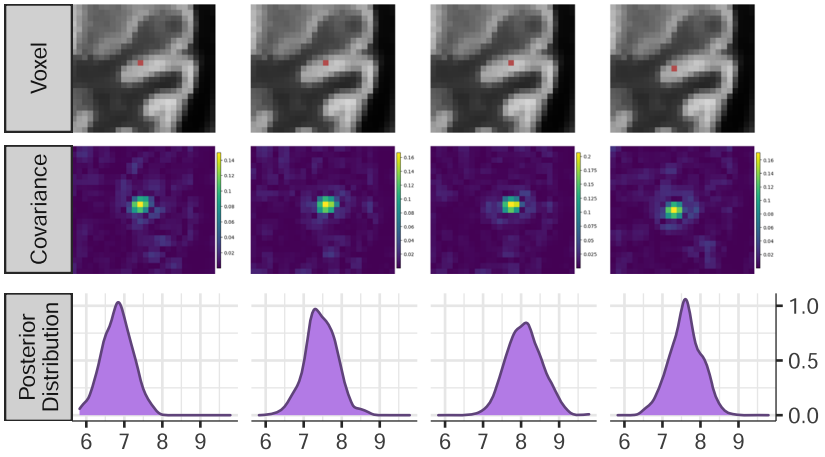

Figure 14. Neurodegenerative disease exam: posterior covariance and entire posterior distributions for several gray matter voxels.

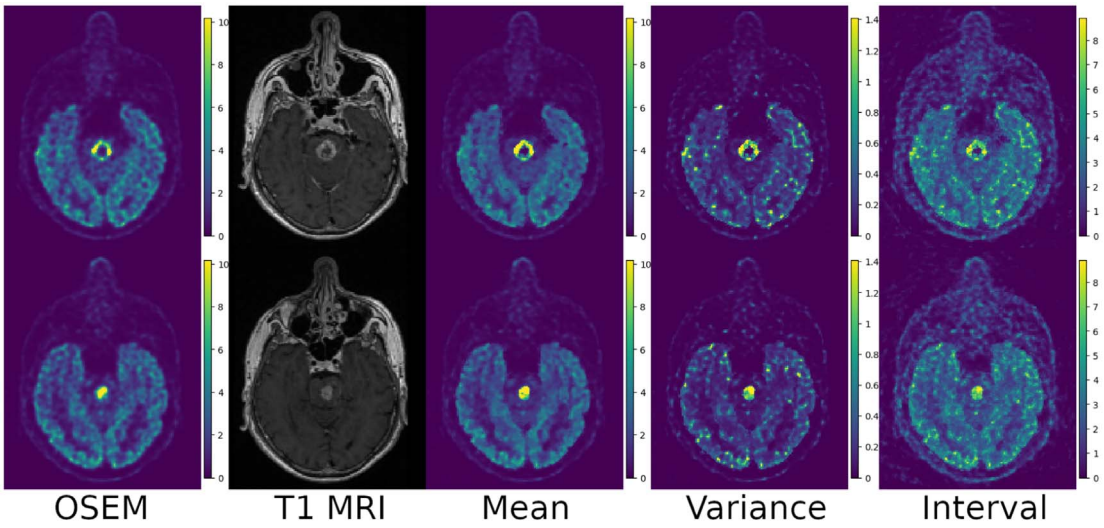

Figure 15. Oncological exam: (from left to right) clinical reconstruction, MRI image, and posterior mean, posterior variance and posterior intervals size for MRF-RD-MRI, for two axial slices in the lesion area.

For the neurodegenerative disease exam, figure 13 shows a standard clinical reconstruction, the MRI image, and some characteristics of posterior MRF-RD-MRI distribution (mean, variance, intervals size) for two example axial slices located in an anatomical region similar to the phantom used for simulated data, while figure 14 shows covariance images and entire posterior distributions for several example gray matter voxels. For the oncological exam, figure 15 shows a standard clinical reconstruction, the MRI image and some characteristics of posterior MRF-RD-MRI distribution (mean, variance, intervals size) for two example axial slices located in the lesion area, while figure 16 shows covariance images and entire posterior distributions for several example lesion voxels.

For both exams, the posterior mean image was visually indistinguishable from the corresponding MAP estimate (not shown), with quantitative differences being low (the root mean square difference < 0.03). The posterior mean shows clearer edges than OSEM because of the MRI-influenced spatial regularization. The variance is higher in areas with higher uptake and near some edges. The covariance is highest in the nearest voxel neighbourhood. Entire posterior distributions tend to concentrate around their maximum and present some skewness, similarly to the simulated data. For the oncological exam, the posterior mean, variance, and intervals size images have a noisier appearance than for the neurodegenerative disease exam, because the noise equivalent count rate is lower.

For a further interpretation and analysis of posterior distribution characteristics from real exams, close collaboration with physicians using diagnostic information is required. There are several perspectives for future work, e.g. building more elaborate priors that include some prior clinical knowledge about the tracer and the

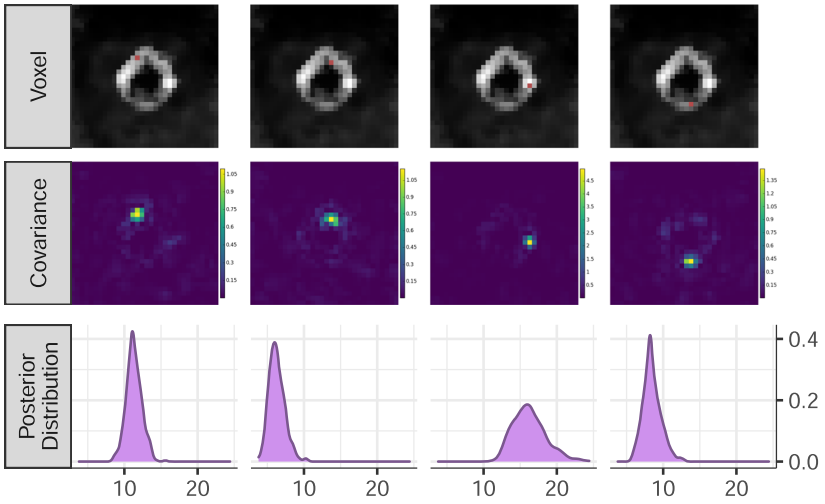

Figure 16. Oncological exam: posterior covariance and entire posterior distributions for several voxels in the lesion.

pathology, or propagating the uncertainty information into quantitative image processing (kinetic modelling, texture features, biomarkers, machine learning).

4. Conclusion

We show that the posterior bootstrap framework can be easily applied to PET image reconstruction to produce approximate posterior image distributions for any usual Bayesian model for a single patient dataset. Posterior distributions were obtained for several Bayesian models with spatially regularizing image priors for simulated data. They were assessed, analyzed and described in detail in terms of the mean, (co)variance, intervals. Their relationship with the corresponding distributions of the MAP image estimate over dataset realizations was exposed. The methodology was applied on two real datasets from a PET/MRI scanner. Posterior image uncertainties provide information about the propagation of the data noise, as dependent on the modelling assumptions, the chosen image prior, parameter values, and on the available dataset. Pathway is opened for the use of posterior uncertainties in diagnostic and quantification tasks.

Acknowledgments

This work was performed on a platform of France Life Imaging network partly funded by the grant ANR-11-INBS-0006. This work is partly supported by the 'MMIPROB' project funded by ITMO Cancer (France). Thanks to Antoine Pierucci for the diagrams. Thanks to Florent Sureau for proofreading and for insightful comments.

Appendix. Random weights

As discussed in section 2.1, the random weights w required in equation (4) are drawn from a uniform Dirichlet distribution with K unitary parameters:

$$(w_1,\ w_2\ ,...,\ w_K)\ \sim\ \text{Dirichlet}(1,\ 1,\ldots,1). \tag{A.1}$$

According to the properties of Dirichlet distributions, drawing a realization from this K-dimensional Dirichlet distribution can be implemented using K Gamma distributions, with their shape parameters equal to the parameters of the Dirichlet distribution, as:

$$p_k\ \sim\ \text{Gamma}(1,\ 1) \tag{A.2}$$

$$w_k = p_k \Big/ \sum_m p_m \tag{A.3}$$

$$Kw_k = \frac{K}{\sum_m p_m} p_k. \tag{A.4}$$

As $\sum_m p_m \sim \text{Gamma}(K, 1)$, and $E(\text{Gamma}(K, 1)) = K$, and as $K \gg 1$, the following approximation can be made for simplifying the implementation:

$$\frac{K}{\sum_m p_m} \approx 1 \tag{A.5}$$

resulting in

$$Kw_k \sim \text{Gamma}(1,\ 1) \tag{A.6}$$

as in algorithm 1.

As was already shown in equation (6), the posterior bootstrap can be applied on histogram data by drawing randomized histograms y_b, where the number of counts in each randomized histogram bin is

$$y_{ib} = \sum_{k \in S_i} Kw_{kb} \tag{A.7}$$

$$y_{ib} = \frac{K}{\sum_m p_m} \sum_{k \in S_i} p_k. \tag{A.8}$$

Following the properties of Gamma distributions, $\sum_{k \in S\_i} p_k \sim \text{Gamma}(y_i, 1)$, where y_i is the number of counts in the original histogram bin i. Then, using the same approximation as above, the number of counts in each randomized histogram bin can be obtained as:

$$y_{ib} \sim \text{Gamma}(y_i,\ 1) \tag{A.9}$$

as in algorithm 2.

ORCID iDs

Marina Filipović https://orcid.org/0000-0002-8560-8888

References

Amunts K *et al* 2013 Bigbrain: an ultrahigh-resolution 3D human brain model *Science* **340** 1472–5

Bai B, Li Q and Leahy R M 2013 Magnetic resonance-guided positron emission tomography image reconstruction *Semin. Nucl. Med.* **43** 30–44

Barrett H H, White T and Parra L C 1997 List-mode likelihood *J. Opt. Soc. Am.* A **14** 2914–23

Barrett H H, Wilson D W and Tsui B M W 1994 Noise properties of the em algorithm: I. Theory *Phys. Med. Biol.* **39** 833–46

Bayarri M J and Berger J O 2004 The interplay of bayesian and frequentist analysis *Stat. Sci.* **19** 58–80

Belzunce M A and Reader A J 2020 Technical note: ultra high-resolution radiotracer-specific digital pet brain phantoms based on the bigbrain atlas *Med. Phys.* **47** 3356–62

Bochkina N A and Green P J 2014 The bernstein-von mises theorem and nonregular models *Ann. Stat.* **42** 1850–78

Bowsher J E *et al* 2004 Utilizing mri information to estimate f18-fdg distributions in rat flank tumors *IEEE Symp. Conf. Record Nuclear Science 2004* vol 4 *(Rome, Italy, 16–22 October 2004)* pp 2488–92

Buvat I 2002 A non-parametric bootstrap approach for analysing the statistical properties of SPECT and PET images *Phys. Med. Biol.* **47** 1761–75

CAS 2017 Castor—customizable and advanced software for tomographic reconstruction (http://castor-project.org/)

Dahlbom M 2001 Estimation of image noise in pet using the bootstrap method *2001 IEEE Nucl. Sci. Symp. Conf. Record* vol 4 *(San Diego, CA, 4–10 November 2001)* (Piscataway, NJ: IEEE) pp 2075–9

De Pierro A R 1995 A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography *IEEE Trans. Med. Imaging* **14** 132–7

Efron B 1979 Bootstrap methods: another look at the jackknife *Ann. Stat.* **7** 1–26

Fessler J A 1996 Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): applications to tomography *IEEE Trans. Image Process.* **5** 493–506

Filipović M, Barat E, Dautremer T, Comtat C and Stute S 2019 Pet reconstruction of the posterior image probability, including multimodal images *IEEE Trans. Med. Imaging* **38** 1643–54

Fong E, Lyddon S and Holmes C 2019 Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap *J. Mach. Learn. Res.* **97** 1952–62

Gelman A, Meng X-L and Stern H 1996 Posterior predictive assessment of model fitness via realized discrepancies *Stat. Sin.* **6** 733–60

Girolami M and Calderhead B 2011 Riemann manifold langevin and hamiltonian monte carlo methods *J. R. Stat. Soc.* B **73** 123–214

Higdon D M, Bowsher J E, Johnson V E, Turkington T G, Gilland D R and Jaszczak R J 1997 Fully bayesian estimation of gibbs hyperparameters for emission computed tomography data *IEEE Trans. Med. Imaging* **16** 516–26

Huesman R H, Klein G J, Moses W W, Qi J, Reutter B W and Virador P R G 2000 List-mode maximum-likelihood reconstruction applied to positron emission mammography (pem) with irregular sampling *IEEE Trans. Med. Imaging* **19** 532–7

Jakoby B W, Bercier Y, Watson C C, Bendriem B and Townsend D W 2009 Performance characteristics of a new lso pet/ct scanner with extended axial field-of-view and psf reconstruction *IEEE Trans. Nucl. Sci.* **56** 633–9

Jaskowiak C J, Bianco J A, Perlman S B and Fine J P 2005 Influence of reconstruction iterations on 18f-fdg pet/ct standardized uptake values *J. Nucl. Med.* **46** 424–8

Lyddon S P, Holmes C C and Walker S G 2019 General Bayesian updating and the loss-likelihood bootstrap *Biometrika* **106** 465–78

Lyddon S, Walker S and Holmes C C 2018 Nonparametric learning from bayesian models with randomized objective functionsarXiv:1806.11544v2

Markiewicz P J, Reader A J and Matthews J C 2014 Assessment of bootstrap resampling performance for PET data *Phys. Med. Biol.* **60** 279–99

Merlin T, Stute S, Benoit D, Bert J, Carlier T, Comtat C, Filipovic M, Lamare F and Visvikis D 2018 Castor: a generic data organization and processing code framework for multi-modal and multi-dimensional tomographic reconstruction *Phys. Med. Biol.* **63** 185005
Newton M A, Polson N G and Xu J 2020 Weighted Bayesian bootstrap for scalable posterior distributions *Can. J. Stat.* **49** 421–37
Newton M A and Raftery A E 1994 Approximate bayesian inference with the weighted likelihood bootstrap *J. R. Stat. Soc.* B **56** 3–26
Nuyts J, Beque D, Dupont P and Mortelmans L 2002 A concave prior penalizing relative differences for maximum-a-posteriori reconstruction in emission tomography *IEEE Trans. Nucl. Sci.* **49** 56–60
Qi J and Leahy R M 1999 A theoretical study of the contrast recovery and variance of map reconstructions from pet data *IEEE Trans. Med. Imaging* **18** 293–305
Qi J and Leahy R M 2006 Iterative reconstruction techniques in emission computed tomography *Phys. Med. Biol.* **51** 541–78
Robert C and Casella G 2013 *Monte Carlo Statistical Methods, Springer Texts in Statistics* (Berlin: Springer) (https://doi.org/10.1007/978-1-4757-4145-2)
Rubin D B 1981 The bayesian bootstrap *Ann. Stat.* **9** 130–4
Rubin D B 1984 Bayesianly justifiable and relevant frequency calculations for the applied statistician *Ann. Stat.* **12** 1151–72
Sitek A 2011 Reconstruction of emission tomography data using origin ensembles *IEEE Trans. Med. Imaging* **30** 946–56
Sitek A 2012 Data analysis in emission tomography using emission-count posteriors *Phys. Med. Biol.* **57** 6779–95
Sitek A, Li Q, Fakhri G E and Alpert N M 2016 Validation of bayesian analysis of compartmental kinetic models in medical imaging *Phys. Med.* **32** 1252–8
Stute S, Tauber C, Leroy C, Bottlaender M, Brulon V and Comtat C 2015 Analytical simulations of dynamic pet scans with realistic count rates properties *IEEE Nucl. Sci. Symp. and Med. Imaging Conf. (NSS/MIC) (San Diego, CA, 31 October–7 November 2015)* (Piscataway, NJ: IEEE) pp 1–3
Vunckx K and Nuyts J 2010 Heuristic modification of an anatomical markov prior improves its performance *IEEE Nucl. Sci. Symp. Med. Imaging Conf. (Knoxville, TN, 30 October–6 November 2010)* (Piscataway, NJ: IEEE) pp 3262–6
Weir I S 1995 Fully Bayesian reconstructions from single-photon emission computed tomography data *J. Am. Stat. Assoc.* **92** 49–60
Zhang C, Arridge S and Jin B 2019 Expectation propagation for poisson data *Inverse Problems* **35** 085006