Nonparametric posterior learning for emission tomography

Fedor Goncharov, Eric Barat, Thomas Dautremer

Nonparametric posterior learning for emission tomography with multimodal data*

Fedor Goncharov†, Éric Barat†, and Thomas Dautremer†

Abstract. We continue studies of the uncertainty quantification problem in emission tomographies such as PET or SPECT when additional multimodal data (anatomical MRI images) are available. To solve the aforementioned problem we adapt the recently proposed nonparametric posterior learning technique to the context of Poisson-type data in emission tomography. Using this approach we derive sampling algorithms which are trivially parallelizable, scalable and very easy to implement. In addition, we prove conditional consistency and tightness for the distribution of produced samples in the small noise limit (i.e., when the acquisition time tends to infinity) and derive new geometrical and necessary condition on how MRI images must be used. This condition arises naturally in the context of identifiability problem for misspecified generalized Poisson models with wrong design. We also contrast our approach with Bayesian Markov Chain Monte Carlo sampling based on one data augmentation scheme which is very popular in the context of Expectation-Maximization algorithms for PET or SPECT. We show theoretically and also numerically that such data augmentation significantly increases mixing times for the Markov chain. In view of this, our algorithms seem to give a reasonable trade-off between design complexity, scalability, numerical load and assessment for the uncertainty.

Key words. tomography, inverse problems, MCMC, Bayesian inference, bootstrap

AMS subject classifications. 62-04, 62F15, 62C10

1. Introduction.

Emission tomographies (further referred as ET) such as Positron Emission Tomography (PET) or Single Photon Emission Computed Tomography (SPECT) are functional imaging modalities of nuclear medicine which are used to image activity processes and, in particular, metabolism in soft tissues. The level of metabolism and uptake of specific biomarkers provide crucial information for diagnostics and treatment of cancers; see e.g., [53], [36] and references therein. Therefore, quality of images in ET and their respective resolution are critical for the diagnostics-treatment pipeline. In this work we continue studies on the two following problems:

Problem 1. Quantify the uncertainty of reconstructions in ET.

Problem 2. Regularize the inverse problem using the multimodal data (e.g., images from CT or MRI).

Problem 1 is not new and several approaches have been established already which in turn can be grouped according to the statistical view of the problem – frequentist ([12], [1], [30]), Bayesian ([23], [54], [10], [46], [3], [14]) and bootstrap ([20], [8], [28], [15]). Note that given list is far from being complete and it should include references therein.

Problem 2 can be splitted further depending on which type of exterior data are used - CT or MRI. More generally, main reasons to use multimodal data in ET are the ill-posedness of corresponding inverse problems (in PET/SPECT forward operators are ill-conditioned; see

*Submitted to the editors October 26, 2022.

Funding: This work is partly supported by the 'MMIPROB' project funded by ITMO Cancer (France).

†Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France (fedor.goncharov@cea.fr).

e.g., [24]) and very low signal-to-noise ratio in the raw measured data. All this together results in loss of resolution in reconstructed images and consequently in oversmoothing, e.g., when applying standard methods such as spatially invariant filters for post-smoothing. The common way of using CT and MRI images consists in extracting boundaries of anatomical features and embedding them into regularization schemes via special penalties and/or non-invariant filters; see e.g., [13], [6], [22], [7], [52]. The foundation of the above approaches is that there are correlations between PET and MRI signals starting from simple anatomical up to biological ones (e.g., PET-MRI investigation on tumor imaging in [5]). Therefore, potentially MRI data can be used to regularize accurately the inverse problem, however, it still requires construction of fine models to describe such correlations. Finally, from very practical point of view Problem 2 with additional MRI data is of interest due to availability of commercially available models of PET-MRI scanners [34], [26] which allow simultaneous registrations of both signals. In this work for multimodal data we use series of presegmented anatomical MRI images which are used differently than it was explained before. In section 2 we explain in detail how we use the MRI data and compare it with existing approaches.

Already the definition of uncertainty in Problem 1 is not obvious: for exposure period $[0, t)$ raw data Y^t (sinogram) is generated by unknown (binned) point process PP^t (typically it is assumed to be Poisson with unknown intensity parameter $\lambda_* \in \mathbb{R}^p_+$ and known design $A \in \mathrm{Mat}(d, p)$, i.e., $PP^t = PP^t_{A\lambda_*} = \mathrm{Po}(t \cdot A\lambda_*)$). Therfore, for any estimator $\widehat{\lambda}^t$ the uncertainty propagates directly from Y^t. This is known as aleatoric uncertainty which corresponds to frequentist approach, and for ET it often leads to estimation of confidence bounds for the maximum likelihood estimator (MLE) or the penalized maximum log-likelihood estimator (pMLE or MAP; both are M-estimators [49]); see e.g., [12]. Frequentist approach has an advantage of being relatively robust to model misspecification (i.e., when $PP^t \neq PP^t_{A\lambda}$). In this case, for large t consistent estimator $\widehat{\lambda}^t$ will tend a.s. to a projection of PP^t onto $PP^t_{A\lambda}$ with respect to some chosen distance between probability distributions (e.g., for Kullback-Liebler divergence). Under additional assumptions on PP^t even in misspecified case it is still possible to establish asymptotic distribution of $\widehat{\lambda}^t$ such as asymptotic normality, from which the confidence intervals can be retrieved. However, practical use of asymptotic estimates for ET seems doubtful since very little data are available in a single scan.

Epistemic uncertainty is another type of uncertainty which corresponds to Bayesian or bootstrap approaches in statistics. For the Bayesian case the initial uncertainty on the parameter of interest is encoded in some prior measure (using anatomical information from side images, assumptions on support and smoothness) which is updated using model $PP^t_{A\lambda}$ and conditioning on Y^t to define the posterior distribution via the well-known Bayes' formula. Sampling from such complicated posteriors is usually done via Markov Chain Monte Carlo (MCMC) techniques [54], [23], [10], [14]. However, there are common bottlenecks: complicated design of samplers and their implementations, high numerical load per iteration, lack of scalability and most importantly – poor mixing in constructed chains; see e.g., [14], [50], [41]. Additional methodological issue is the misspecification of the model (e.g., incorrect design) which cannot be included in the classical Bayesian framework and for robust inference it leads to the recently proposed general Bayesian updating and bootstrap-type sampling; see [42], Section 1.

As noted before bootstrap is another attractive technique to assess the uncertainty which can be also seen as some probabilistic sensitivity analysis or as approximate/exact sampling from nonparametric Bayesian posteriors; see e.g., [38], [35], [16]. Nontrivial questions for bootstrapping ET are the following ones: (1) how to define the procedure for Poisson-type raw data in ET and also include side information (2) provide guarantees (theoretical and numerical) on the coverage of the true signal by new credible intervals. A common approach to answer question (1) is to use resampling; see e.g., [20], [8]. For ET this one targets to resample photon counts and then propagate the uncertainty by using any reconstruction algorithm (e.g., FBP (Filtered backprojection), MLE or MAP (maximum a posteriori)). Question (2) is resolved theoretically often by demonstrating asymptotic equivalence between bootstrap, Bayesian and frequentist approaches via Bernstein von-Mises type theorems (see e.g., [49], [35], [39] or equivalence of Edgeworth's expansions for higher orders (see [42]) and numerically via calibration (e.g., using Q-Q plots).

In view of the above discussion, we note that for practice it seems that it is not of great importance which kind of uncertainty model is used – frequentist, Bayesian or bootstrap. The most important is to make usable the resulting framework and algorithms by practitioners, hence, they should be simple to implement, desirably with tractable parameters and numerically efficient (scalability is crucial for high-dimensional models in ET).

Being inspired with nonparametric posterior learning (further referred as NPL) originating from [35], [16], we propose sampling algorithms for ET of bootstrap type with and without MRI data at hand. Therefore, our main contribution is that we extend the NPL originally proposed for regular statistical models and i.i.d data to the non-regular generalized Poisson model of ET (see [3]), where the raw data are not i.i.d but a sample from a point process. The initial motivation for this work was the problem of poor mixing for the Gibbs-type sampler in [14] which was designed for posterior sampling in the PET-MRI context. Below we give a detailed analysis of this phenomenon and conclude with a few generic advices on design of MCMC-samplers for ill-posed inverse problems such as PET or SPECT. Our new algorithms solve the above problem since sampled images are automatically i.i.d, moreover, the scheme is trivially parallelizable, scalable and very easy to implement because it relies on the well-known EM-type reconstruction methods from [44], [11]. Our samplers are tested numerically on a synthetic dataset by demonstrating the regularization effect of MRI as well on calibration of the posterior. We also conduct a theoretical study for when large dataset is available (for ET this is equivalent to $t \to +\infty$) and establish consistency and tightness of the posterior for almost any trajectory Y^t, $t \in [0, +\infty)$. As a byproduct of our study, for the misspecified scenario with incorrect design matrix (which is always true in practice) we discover an intuitive sufficient condition for identifiability to persist. The latter can be of interest for further theoretical studies of ET model under misspecification.

This paper is organized as follows. In section 2 we give notations and all necessary preliminaries on the statistical model of ET and on use of multimodal data. In section 3 we give a very informative example for the problem of poor mixing for MCMC. In section 4 we adapt the NPL for ET context and derive our sampling algorithms. In section 5 we present results of the numerical experiment on a synthetic dataset. In section 6 we study theoretically the asymptotic properties of our algorithms. In section 7 we discuss our results and possibilities for future work.

2. Preliminaries.

2.1. Notations. By $\mathbb{N}_0$ we denote the set of non-negative integers, $\mathbb{R}^n_+$ denotes the positive cone of $\mathbb{R}^n$, by $x \succeq y$, $x \in \mathbb{R}^n$, $y \in \mathbb{R}^n$, we denote the property that $x_j \geq y_j$ for all $j = 1, \ldots, n$, $x \succ y$ denotes the same but with strict inequalities, both $\langle x, y\rangle$ or $x^T y$ stand for the scalar product, $R_+(A)$ denotes the image of $\mathbb{R}^p_+$ under action of operator $A \in \mathrm{Mat}(d, p)$, by $X \sim F$ we denote the property that r.v. X has distribution F, $\mathrm{Po}(\lambda)$ denotes the Poisson distribution with intensity λ, $\lambda \geq 0$, $\Gamma(\alpha, \beta)$ denotes the gamma distribution with shape α, and scale β ($\xi \sim \Gamma(\alpha, \beta)$, $\mathbb{E}\xi = \alpha\beta$, $\mathrm{var}(\xi) = \alpha\beta^2$). Let $A \in \mathrm{Mat}(d, p)$, then $\mathrm{cond}(A)$ denotes the condition number of A, A_I, $I \subset \{1, \ldots, d\}$ denotes the submatrix of A with rows indexed by elements in I, $\mathrm{Span}(A^T)$ denotes the span of the rows of A being considered as vectors in $\mathbb{R}^p$. Let Z be a complete separable metric space equipped with metric $\rho_Z(\cdot, \cdot)$ and boundedly finite non-negative measure dz, $B(Z)$ denotes the sigma algebra of borel sets in Z. By $\mathcal{PP}$ we denote a (spatio-temporal) point process on $Z \times \mathbb{R}_+$ and $\mathcal{PP}_\Lambda$ denotes the Poisson point process on $Z \times \mathbb{R}_+$ with intensity $\Lambda(z)\, dz\, dt$, where Λ is the nonnegative function $\Lambda = \Lambda(z)$, $z \in Z$, Λ is integrable w.r.t dz. Weighted gamma process on Z is denoted by $GP(\alpha, \beta) = G_{\alpha,\beta}$, where α is the shape measure on Z and β is the scale which is a non-negative function Z and also α-integrable; see, e.g., [33] for construction. Finally, by $\mathcal{KL}(P, Q)$ we denote the standard Kullback-Leibler divergence between probability distributions P, Q.

2.2. Mathematical model for ET. Raw data in ET are described by the so-called sinogram $Y^t = (Y^t_1, \ldots, Y^t_d) \in (\mathbb{N}_0)^d$ which stands for the photon counts recorded during exposure period $[0, t)$ along d lines of response (LORs). It is assumed that

$$\begin{aligned} &Y^t_i \sim \mathrm{Po}(t\Lambda_i),\ \Lambda_i = a_i^T \lambda, \\ &Y^t_i \text{ are mutually independent for } i \in \{1, \ldots, d\}, \end{aligned} \tag{2.1}$$

where $\lambda \in \mathbb{R}^p_+$ is the parameter of interest on which we aim to perform inference. In practice, vector λ denotes the spatial emission concentration of the isotope measured in $[\mathrm{Bq/mm}^3]$, that is λ_j is the concentration at pixel $j \in \{1, \ldots, p\}$. Vector $\Lambda = (\Lambda_1, \ldots, \Lambda_d)$ denotes the observed photon intensities along LORs $\{1, \ldots, d\}$, respectively. To separate the LORs with strictly positive intensities from those ones with zeros we introduce following notations:

$$I_0(\Lambda) = \{i : \Lambda_i = 0\},\ I_1(\Lambda) = \{i : \Lambda_i > 0\},\ I_0 \sqcup I_1 = \{1, \ldots, d\}. \tag{2.2}$$

Collection of $a_i \in \mathbb{R}^p$ in (2.1) constitute matrix $A = [a^T_1, \ldots, a^T_d]^T$, $A \in \mathrm{Mat}(d, p)$ which is called by projector or system matrix in applied literature on ET and by design (or design matrix in statistical literature). Each element a_{ij} in A denotes the probability to observe a pair of photons along LOR $i \in \{1, \ldots, d\}$ if both they were emitted from pixel $j \in \{1, \ldots, p\}$. In view of such interpretation, for design A we assume the following:

$$a_{ij} \geq 0 \text{ for all pairs } (i, j), \tag{2.3}$$

$$A_j = \sum_{i=1}^{d} a_{ij},\ 0 < A_j \leq 1 \text{ for all } j \in \{1, \ldots, p\}, \tag{2.4}$$

$$\sum_{j=1}^{p} a_{ij} > 0 \text{ for all } i \in \{1, \ldots, d\}. \tag{2.5}$$

If any of formulas (2.4), (2.5) would not be satisfied, then, in practice it would mean that either some pixel is not detectable at all (hence it can be completely removed from the model) or some detector pair is broken and cannot detect any of incoming photons. These scenarios are outside of our scope.

It is well-known that the inverse problems for PET and SPECT are mildly ill-posed (see e.g., [24], [37]), which in practice means that

$$\ker A \neq \{0\}. \tag{2.6}$$

Remark 2.1. Matrix A represents a discretization of weighted Radon transform operator R_a for ET with complete angle data on the plane (see [37], Chapter 2). Since A approximates R_a in strong operator norm we know that

$$\sigma_k \asymp k^{-1/2},\ k = 1, \dots, p, \tag{2.7}$$

where σ_k are the singular values of A. In particular, even if A is injective for p large enough, due to (2.7), it may happen that $\mathrm{cond}(A) > \varepsilon_F^{-1}$, where ε_F is the floating-point precision. In the latter case, due to the cancelling effect singular values of A numerically will be equivalent to machine zeros which means then exactly the existence of a nontrivial kernel for A.

Likelihood and negative log-likelihood functions for model in (2.1) are given by the formulas:

$$PP^t_{A,\lambda}(Y^t) = \mathrm{p}(Y^t \mid A, \lambda, t) = \prod_{i=1}^{d} \frac{(ta_i^T\lambda)^{Y_i^t}}{Y_i^t!} e^{-ta_i^T\lambda},\ \lambda \in \mathbb{R}^p_+,\ t \geq 0, \tag{2.8}$$

$$L(\lambda \mid Y^t, A, t) = \sum_{i=1}^{d} -Y_i^t \log(t\Lambda_i) + t\Lambda_i,\ \Lambda_i = a_i^T\lambda. \tag{2.9}$$

Note that for A satisfying (2.6) and for any Y^t function $L(\lambda \mid Y^t, A, t)$ is not strictly convex even at the point of the global minima since $L(\lambda + u \mid Y^t, A, t) = L(\lambda \mid Y^t, A, t)$ for any $\lambda \in \mathbb{R}^p_+$ and $u \in \ker A$. To avoid numerical instabilities due to this phenomenon a convex penalty $\varphi(\lambda)$ is added to $L(\lambda \mid Y^t, A, t)$, so we also consider the penalized negative log-likelihood:

$$L_p(\lambda \mid Y^t, A, t, \beta^t) = L(\lambda \mid Y^t, A, t) + \beta^t\varphi(\lambda),\ \lambda \in \mathbb{R}^p_+, \tag{2.10}$$

where $\beta^t \geq 0$ is the regularization coefficient. Parameter β^t may increase with t at a certain rate which is important for practice in order to increase the signal-to-noise ratio in reconstructed images.

2.3. Regularization penalty. The role of penalty $\varphi(\lambda)$ in (2.10) is to decrease the numerical instability in the underlying inverse problem and to make function $L_p(\lambda \mid Y^t, A, t, \beta^t)$ more convex, especially in directions close to $\ker A$.

In view of this we assume that

$$\varphi \text{ is continuous and convex on } \mathbb{R}^p, \tag{2.11}$$

$$g_u(w) = \varphi(u + w) \text{ is strictly convex in } w \in \ker A \text{ for any } u \in \mathrm{Span}(A^T). \tag{2.12}$$

For numerical tests in section 5 we choose φ to be the sum of two pairwise-difference functions for neighboring pixels: first is of log-cosh type which is standard for ET (see [3], [54]), and second is the pure ℓ_2-squared norm to add more smoothness to sparse images reconstructed with log-cosh type regularization.

Since A is not injective, even for infinite amount of data ($Y^t \sim PP^t_{A,\lambda_*}$, $t \to +\infty$), one is able to find λ_* at most up to its projection $\ker A$ (modulo extra information due to constraint $\lambda_* \in \mathbb{R}^p_+$). With regularization the projection of λ_* onto $\ker A$ will be defined uniquely by φ and positivity constraints. To describe this effect we define the following function:

$$w_{A,\lambda}(u) = \underset{\substack{\lambda+u+w\succeq 0\\ w\in\ker(A)}}{\arg\min}\ \varphi(\lambda+u+w),\ u \in \mathrm{Span}(A^T),\ \lambda \succeq 0. \tag{2.13}$$

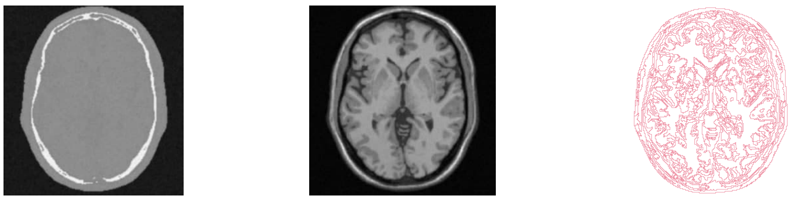
Then, intuitively (this is made rigorous in section 6), the best one can hope to reconstruct using MAP-estimator in (2.10) (or, equivalently, the penalized KL-projection) when $t \to +\infty$ and $\beta^t/t \to 0$, will be

$$\lambda_{*opt} = \lambda_* + w_{A,\lambda_*}(0) = \lim_{\beta\to+0} \underset{\lambda\succeq 0}{\arg\min}\, L_p(\lambda|A\lambda_*, A, 1, \beta) \tag{2.14}$$

Thus, in what follows, the numerical quality of reconstructions, calibration etc., is tested against λ_{*opt} rather than λ_* which is inaccessible no matter the amount of data.

2.4. Multimodal data for ET.

In order to increase the SNR in reconstructed images and not to loose a lot in resolution one can regularize the inverse problem using multimodal data – scans from CT or MRI. We choose MRI since it provides anatomical information with high contrast in soft tissues in comparison to CT (see Figure 1 (a), (b)).



(a) CT (b) MRI (c) $M \in \mathcal{M}$

Figure 1: (a), (b) Multimodal data for ET of the brain; (c) segmented MRI-image in (b)

MRI-guided reconstructions in PET is an active topic of research (see the discussion in [15] and references therein), however, still a lot of work is needed to describe precisely correlations between ET and MRI signals (especially from biological point of view). Because of the latter current use of MRI data is essentially image-based: spatially regularizing penalties are constructed using MRI data in [4], [5], [52] (PET signals are penalized stronger when being constant across edges in MRI images), models built upon MRI-segmented data for locally-constant tracer distribution are used in [14] and also in our work.

In this work we assume that our side data consists of r presegmented MRI images $\mathcal{M} = \{M_1, \dots, M_r\}$ (see Figure 1 (c); segmentations of MRI images are precomputed using the dd-CRP algorithm from [17]), where segments are being disjoint and connected subsets of pixels. First, using $\mathcal{M}$ we construct a lower-dimensional model $Y^t \sim \mathrm{Po}(t\Lambda_{\mathcal{M}})$, $\Lambda_{\mathcal{M}} = A_{\mathcal{M}}\lambda_{\mathcal{M}}$, $A \in \mathrm{Mat}(d, p_{\mathcal{M}})$, $\lambda_{\mathcal{M}} \in \mathbb{R}_+^{p_{\mathcal{M}}}$ ($p_{\mathcal{M}} \ll p$); see also (2.1). Second, randomized pseudo-observations(-sinograms) from this model are mixed with observed Y^t into new sinograms. Subsequent reconstructions from the latter constitute our samples being regularized by $\mathcal{M}$.

Now we explain the construction of $A_{\mathcal{M}}$ and $\lambda_{\mathcal{M}}$ and the actual sampling will be given further in subsection 4.4. Let p_k be the number of segments in $M_k \in \mathcal{M}$, $S(M_k)$ be their collection. For each M_k we define new projector by the formulas:

$$A_k = (a^k_{ij}) \in \mathrm{Mat}(d, p_k), \tag{2.15}$$

$$a^k_{is} = \sum_{j=1}^{p} a_{ij}\mathbb{1}\{\text{pixel } j \text{ belongs to segment } s \in S(M_k)\},\ k \in \{1, \dots, r\}, \tag{2.16}$$

where $A = (a_{ij})$ is the projector for the full model from subection 2.2. Finally, we stack all segments and projectors into one model:

$$A_{\mathcal{M}} = (A_1, \dots, A_r),\ p_{\mathcal{M}} = \sum_{k=1}^{r} p_k, \tag{2.17}$$

$$\lambda_{\mathcal{M}} = (\lambda^1_1, \dots, \lambda^1_{p_1}, \dots, \lambda^r_1, \dots, \lambda^r_{p_r}),\ \Lambda_{\mathcal{M}} = A_{\mathcal{M}}\lambda_{\mathcal{M}},\ \Lambda_{\mathcal{M}} = (\Lambda_{\mathcal{M},1}, \dots, \Lambda_{\mathcal{M},d}). \tag{2.18}$$

Therefore, $\lambda_{\mathcal{M}}$ is a positive linear combination of all segments from all images in $\mathcal{M}$ with constant signal in each segment, and $A_{\mathcal{M}}$ being respective projector derived from A. For $A_{\mathcal{M}}$ we assume that it is injective and well-conditioned, that is

$$\ker A_{\mathcal{M}} = \{0\},\ \mathrm{cond}(A_{\mathcal{M}}) < c_{\mathcal{M}}, \tag{2.19}$$

where $c_{\mathcal{M}}$ is some moderate constant. The latter assumption reflects the idea that images in $\mathcal{M}$ consist of low number of large segments.

3. A motivating example for NPL in ET.

Recently a Gibbs-type sampler was proposed in [14] for Bayesian inference for PET-MRI. Despite a number of positive practical features (spatial regularization, use of multimodal data) the problem of slow mixing for the corresponding Markov chain was observed. Below we consider a simplified version which shares the same mixing problem and explain the phenomenon numerically and theoretically.

In algorithms for ETs it is common to augment data Y^t by $n^t = \{n^t_{ij}\}$, where n^t_{ij} is the number of photons being emitted from pixel j and detected in LOR i, $n^t_{ij} \sim \mathrm{Po}(ta_{ij}\lambda_j)$, n^t_{ij} are mutually independent for all (i, j); see e.g., [44]. In view of this physical interpretation, for pair (n^t, Y^t) the following coherence condition must be satisfied:

$$\sum_{j=1}^{p} n^t_{ij} = Y^t_i \text{ for all } i \in \{1, \dots, d\}. \tag{3.1}$$

By (3.1) one sees Y^t is a function of n^t, so (Y^t, n^t) is indeed a data augmentation of Y^t. Note that n^t are not observed in a real experiment but they greatly simplify design of samplers (see e.g., [25], [14]), because conditional distributions $p(n^t \mid Y^t, A, \lambda, t)$, $p(\lambda \mid n^t, A, t)$ admit simple analytical forms even for nontrivial priors involving multimodal data. For our example below we use only a simple pixel-wise positivity gamma-prior:

$$\pi(\lambda) = \prod_{j=1}^{p} \pi_j(\lambda_j),\ \pi_j = \Gamma(\alpha, \beta^{-1}),\ \alpha > 0,\ \beta > 0, \tag{3.2}$$

where α, β are some fixed constants. For the prior in (3.2) and model (2.1) distributions $p(n^t \mid Y^t, A, \lambda, t)$, $p(\lambda \mid n^t, A, t)$ are as follows:

$$\begin{gathered} p(n_{ij}^t \mid Y^t, A, \lambda, t) = \mathrm{Multinomial}(Y_i^t, p_{i1}(\lambda), \ldots, p_{ip}(\lambda)), \\ p_{ij}(\lambda) = \frac{a_{ij}\lambda_j}{\sum_k a_{ik}\lambda_k},\ i \in \{1, \ldots, d\}, \end{gathered} \tag{3.3}$$

$$p(\lambda_j^t \mid n^t, Y^t, A, t) = \Gamma\left(\sum_{i=1}^{d} n_{ij}^t + \alpha, (tA_j + \beta)^{-1}\right),\ j \in \{1, \ldots, p\}, \tag{3.4}$$

where A_j is defined in (2.4).

Using (3.3), (3.4) the construction a Gibbs sampler for Bayesian posterior sampling from $p(\lambda \mid Y^t, A, t)$ is straightforward.

Algorithm 1 Gibbs sampler for $p(\lambda \mid Y^t, A, t)$

data : Y^t
input : $\lambda_0 \in \mathbb{R}_+^p$, $\pi(\lambda_j) = \Gamma(\alpha, \beta^{-1})$,
B – number of samples
for $k \leftarrow 1$ to B **do**
  $n_k^t \sim p(n^t \mid Y^t, A, \lambda_{k-1}, t)$
  $\lambda_k^t \sim p(\lambda \mid n_k^t, Y^t, A, t)$
end for
return $\{\lambda_k^t\}_{k=1}^B$,
Folklore: empirical distribution of $\{\lambda_k^t\}_{k=1}^B$ approximates posterior $p(\lambda \mid Y^t, A, t)$

Remark 3.1. One may argue that prior in (3.2) is a very bad choice from practical point of view, especially in view of ill-posedness of the inverse problem since it does not bring any spatial regularization. However, the mixing rate for the Markov chain in Algorithm 1 asymptotically (i.e., when $t \to +\infty$) will not depend on the choice of $\pi(\lambda)$ in the small noise limit due to Bernstein von-Mises phenomenon (see e.g., [3] and formulas (3.6), (3.5)). At the same time, below we show that mixing is affected primarily by the choice of augmentation scheme and the decision to sample n^t.

We consider the correlations between values of $h(\lambda) = h^T\lambda$, $h \in \mathbb{R}^p$, for subsequent samples from the Markov chain in Algorithm 1:

$$\gamma^t(h) = \mathrm{corr}(h(\lambda_{k+1}^t), h(\lambda_k^t) \mid Y^t, t). \tag{3.5}$$

In formula (3.5) we assumed that the chain is in stationary state, i.e. k can be any.

Markov chain for the sampler in Algorithm 1 coincides with data augmentation schemes from [31], [32], where the latter are exactly Gibbs samplers with only one layer of latent variables. In Bayesian framework $\gamma^t(h)$ is also known as fraction of missing information; see [31]. In particular, in [31] authors gave an exact formula for $\gamma^t(h)$ which can be written for our example as follows:

$$\gamma^t(h) = 1 - \frac{\mathbb{E}[\mathrm{var}(h(\lambda) \mid n^t, Y^t, t) \mid Y^t, t]}{\mathrm{var}(h(\lambda) \mid Y^t, t)}. \tag{3.6}$$

Exact formula for (3.6) for arbitrary t seem difficult (if possible) to obtain, however, in the asymptotic regime $t \to +\infty$ one can apply the Bernstein von-Mises type theorem from [3] and arrive to the following simple expression:

$$\gamma(h) = \lim_{t\to+\infty} \gamma^t(h) = 1 - \frac{h^T F_{aug}^{-1}(\lambda_*) h}{h^T F_{obs}^{-1}(\lambda_*) h},\ h \in \mathbb{R}^p,\ \text{ a.s. } Y^t, t \in (0, +\infty). \tag{3.7}$$

where

$$.\lambda_* \text{ is the true parameter, } \lambda_* \succ 0, \tag{3.8}$$

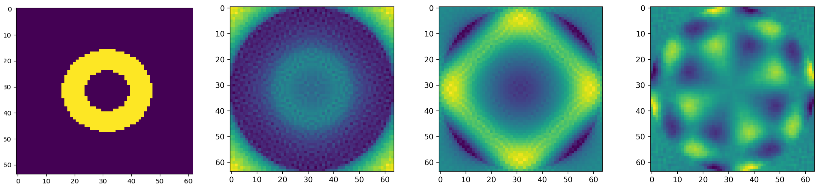
$$F_{obs}(\lambda_*) = \sum_{i=1}^{d} \frac{a_i a_i^T}{\Lambda_i^*} = A^T D_{\Lambda^*}^{-1} A,\ D_{\Lambda^*} = \mathrm{diag}(\dots, \Lambda_i^*, \dots),\ \Lambda_i^* = a_i^T \lambda_*, \tag{3.9}$$

$$F_{aug}(\lambda_*) = \mathrm{diag}(\dots, c_j, \dots),\ c_j = A_j / \lambda_{*j}. \tag{3.10}$$

Note that from (2.5) and (3.8) it follows that $\Lambda_i^* > 0$ for all i, therefore division by Λ_i^* in (3.9) is well-defined. Matrices $F_{obs}(\lambda_*)$, $F_{aug}(\lambda_*)$ are the Fisher information matrices at λ_* for Poisson models with observables Y^t and n^t, respectively. Note also that F_{obs} is not invertible in the usual sense, so in (3.7) its pseudo-inversion in the sense of Moore-Penrose is considered.

Remark 3.2. Strict positivity assumption in (3.8) is not practical and a precise analytic formula which extends (3.7) for $\lambda_* \succeq 0$ can be established using the results from [3]. The point is that model (2.1) is non-regular since the parameter of interest belongs to a domain with a boundary, so a separate result for Bernstein von-Mises phenomenon is needed in this case. For our toy example it is sufficient to consider the case in (3.8) as if we were interested in mixing times of the chain in areas with positive tracer concentration.

Let $h_1, \dots, h_p$ be the orthonormal basis of eigenvectors of $F_{obs}(\lambda_*)$ with corresponding eigenvalues $s_1 \geq s_2 \cdots \geq s_p \geq 0$. Intuitively, in $\{h_m\}_{m=1}^p$ higher indices m correspond to higher frequencies on images (see Figure 2 (a)-(d)).



(a) λ_* (b) h_1 (c) h_2 (d) h_{50}

Figure 2: eigenvectors h_m for $F_{obs}(\lambda_*)$

From (3.7) it follows that

$$\gamma(h_m) = 1 - s_m h_m^T F_{aug}^{-1} h_m. \tag{3.11}$$

Matrix F_{aug} is well-conditioned, continuously invertible and the quadratic term in (3.11) admits the following bound:

$$F_{aug}^{-1}(\lambda_*) = \mathrm{diag}(\dots, \frac{\lambda_{*j}}{A_j}, \dots) \Rightarrow h_m^T F_{aug}^{-1}(\lambda_*) h_m \leq \frac{\max_j(\lambda_{*j})}{\min_j(A_j)}. \tag{3.12}$$

Regular behavior of F_{aug}^{-1} is not surprising because this is the Fisher information matrix for latent variables n^t for which the inverse problem is not ill-posed at all. From (3.9) and the ill-conditioning nature of A it follows that $F_{obs}(\lambda_*)$ is also ill-conditioned (see [18]), moreover, $s_m \asymp m^{-1}$ for large m (see Remark 2.1). From this and (3.11), (3.12) we conclude that

$$\gamma(h_m) \approx 1 \text{ for large } m. \tag{3.13}$$

Formulas (3.5) and (3.13) constitute a clear evidence of poor mixing in the Markov chain in Algorithm 1. Though (3.7)–(3.13) were derived for $t \to +\infty$, they reflect well the behavior of the chain for moderate t which is seen from the numerical experiment below (see Supplementary Materials, section SM5 for details).

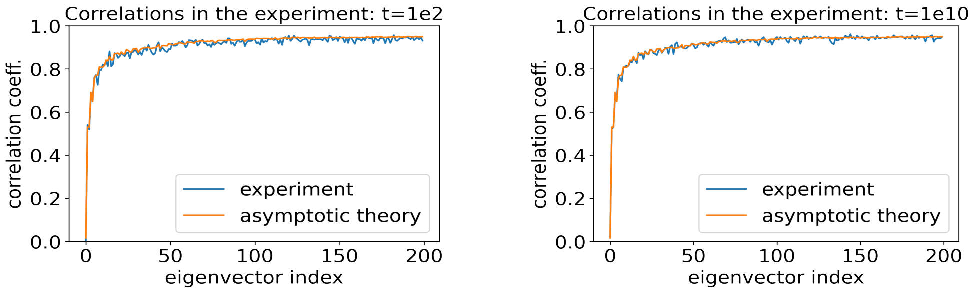


Figure 3: $\mathrm{corr}(h^T\lambda_k^t, h^T\lambda_{k+1}^t \mid Y^t)$ for $t = 10^2, 10^{10}$ for $h = h_m$; blue curve – empirical correlations computed from 2000 samples, orange curve – values for $\gamma(h_m)$ for $m = 1, \dots, 200$ by formula (3.7).

Here one concludes that mixing is much slower for high-frequency parts of images. Therefore, to estimate reliably, say mean $h^T\lambda$ for some mask $h \in \mathbb{R}^p$, one needs almost infinite number of samples if h contains a high-frequency component in terms of $\{h_m\}_{m=1}^p$ (see Supplementary Material, section SM4 for details). This also can be seen as a recommendation for choosing h in practice: h should belong to $\mathrm{Span}(A^T)$ and $|h^T h_m|$ should be as small as possible for large m.

Note that such behavior of the sampler is not due to the choice of pixel-wise prior but due to sampling of n_{ij}^t which correspond to observations for the well-posed inverse problem. A practical advice would be to avoid sampling of missing data in the Markov chain or to use a strong smoothing prior/regularizer (for example by greatly increasing regularization coefficients so that asymptotic arguments in (3.7) will no longer hold but the posterior consistency is still preserved). The latter approach will accelerate mixing at cost of oversmoothing in sampled images.

By this negative but informative example we support the message in [50] saying that design of a data augmentation scheme while preserving good mixing in the Markov chain is an “Art”, especially in the case of ill-posed inverse problems. In view of poor mixing, complexity of the design and implementation, lack of scalability and high numerical load while using MCMC ([54], [23], [10], [41], [14]) we turn to NPL as a practical relaxation of Bayesian sampling for the problem of ETs.

4. Nonparametric posterior learning for emission tomography.

To derive the NPL for ET we prefer to start from the completely nonparametric setting as it was originally done in [35]. This allows us to concentrate on essential ideas behind and, moreover, all practical algorithms can be directly obtained by binning nonparametric objects to finite dimensions. A reader interested mainly in practical outcomes may skip this and go directly to subsection 4.5.

4.1. Nonparametric model for ET.

Nonparametric framework for ET can be seen as a classical scanning scenario with a machine having infinite number of infinitely small detectors. Let Z be the space of all detector positions in the acquisition geometry of a scanner (e.g., for one slice Z consists of all non-oriented straight lines in $\mathbb{R}^2$). We also assume that Z is equipped with a boundedly-finite measure dz and with a metric ρ_Z (describing distances between the lines). Then, for exposure period $[0, t)$ the raw data are given by random measure Z^t generated by a point process:

$$Z^t = \sum_{j=1}^{N^t} \delta_{(z_j, t_j)},\ (z_j, t_j) \in Z \times \mathbb{R}_+,\ t_j < t_{j+1},\ t_j \le t, \tag{4.1}$$

where

$$N^t \text{ is total number of registered photons,} \tag{4.2}$$

$$\{z_j\}_{j=1}^{N^t},\ \{t_j\}_{j=1}^{N^t} \text{ are the LORs and arrival times of events, respectively.} \tag{4.3}$$

In practical literature on ET sample Z^t is known as list-mode data, whereas Y^t (sinogram) is the version of Z^t integrated withing $[0, t)$. Under the assumption of temporal stationarity of the generating process, Y^t contains the same amount of information as Z^t since the first one is then a sufficient statistic.

For statistical model of Z^t, one takes the family of temporally stationary Poisson point processes $\mathcal{PP}_{A\lambda}$ on $Z \times \mathbb{R}_+$, where A, λ stand for nonparametric versions of the projector and the tracer concentration from section section 2. For intuition, in such model the intensity parameter of the process in LOR $z \in Z$ at time t is $\Lambda(z)\,dz\,dt = [A\lambda](z)dz\,dt$, therefore $\Lambda(z)\,dt\,dz$ is the density function for the intensity measure of the Poisson process.

The negative log-likelihood for $\mathcal{PP}_{A\lambda}$ with observation Z^t is defined via the following formula (see, e.g., [24], Section 2; [9], Section 2.1):

$$
\begin{aligned}
L(\lambda \mid Z^t, A, t) &= -\sum_{j=1}^{N^t} \log(\Lambda(z_j)) + \int_{Z\times[0,t)} \Lambda(z)\,dz\,dt \\
&= -\int_{Z\times[0,t)} \log(\Lambda(z)) Z^t(dz\,dt) + t\int_Z \Lambda(z)\,dz,\ \Lambda(z) = A\lambda(z).
\end{aligned} \tag{4.4}
$$

4.2. Misspecification and the KL-projection. In reality our model assumption is always incorrect and $Z^t \sim \mathcal{PP}^t = \mathcal{PP}|_{Z\times[0,t)}$ (marginal for interval $[0,t)$) for some point process $\mathcal{PP}$ on $Z\times\mathbb{R}_+$, where $\mathcal{PP}$, $\mathcal{PP} \neq \mathcal{PP}_{A\lambda}$ for any $\lambda \succeq 0$. Since the (penalized) maximum log-likelihood estimates are the most popular in ET, we say that the best one can hope to reconstruct using measurements from $\mathcal{PP}_{A\lambda}$ on $[0,t)$ is the projection of $\mathcal{PP}^t$ onto $\mathcal{PP}^t_{A\lambda} = \mathcal{PP}_{A\lambda}|_{Z\times[0,t)}$ in the sense of Kullback-Leibler divergence:

$$
\lambda_*(\mathcal{PP}, [0,t)) = \arg\min_{\lambda\succeq 0} \mathcal{KL}(\mathcal{PP}^t, \mathcal{PP}^t_{A\lambda}). \tag{4.5}
$$

Since A is ill-conditioned, in general, λ_* in (4.5) may not be defined uniquely. For this we consider the penalized KL-projection defined by the formula:

$$
\lambda_*(\mathcal{PP}, [0,t), \beta^t) = \arg\min_{\lambda\succeq 0} [\mathcal{KL}(\mathcal{PP}^t, \mathcal{PP}^t_{A\lambda}) + \beta^t \varphi(\lambda)], \tag{4.6}
$$

where β^t is the regularization coefficient and $\varphi(\lambda)$ is a nonparametric version of penalty from section 2. From (4.4) and the definition of Kullback-Leibler divergence it follows that (up to terms independent of λ):

$$
\mathcal{KL}(\mathcal{PP}^t, \mathcal{PP}^t_{A\lambda}) = -\int_{Z\times[0,t)} \log(\Lambda(z)) \mathbb{E}_{\mathcal{PP}^t}[Z^t(dz\,dt)] + t\int_Z \Lambda(z)dz, \tag{4.7}
$$

where $\mathbb{E}_{\mathcal{PP}^t}$ is the expectation on Z^t with respect to $\mathcal{PP}^t$. Putting together (4.6), (4.7), for the penalized KL-projection we get the following formulas:

$$
\lambda_*(\mathcal{PP}, [0,t), \beta^t) = \arg\min_{\lambda\succeq 0} \mathbb{L}_p(\lambda \mid \mathcal{PP},\ A,\ t,\ \beta^t), \tag{4.8}
$$

$$
\begin{aligned}
\mathbb{L}_p(\lambda \mid \mathcal{PP}^t,\ A,\ t,\ \beta^t) &= -\int_{Z\times[0,t)} \log(\Lambda(z)) \mathbb{E}_{\mathcal{PP}^t}[Z^t(dz\,dt)] + t\int_Z \Lambda(z)dz + \beta^t\varphi(\lambda), \\
\Lambda(z) &= A\lambda(z).
\end{aligned} \tag{4.9}
$$

4.3. Propagation of uncertainty and the generic algorithm. Following the idea from [35], we say that uncertainty on λ propagates from the one on $\mathcal{PP}$ via (4.8), (4.9). Let $\pi_{\mathcal{M}}$ be a prior in which we encode our beliefs over a set of possible $\mathcal{PP}$'s, that is $\pi_{\mathcal{M}}$ is a nonparametric prior on spatio-temporal point processes on $Z \times \mathbb{R}_+$ and it is constructed using $\mathcal{M}$. Let data be list-mode Z^t or sinogram Y^t, then our prior beliefs can be updated in form of posterior distribution $\pi_{\mathcal{M}}(\cdot \mid Z^t \vee Y^t, t)$.

Algorithm 2 Generic NPL for ET

1: **data** : Z^t or Y^t, $\mathcal{M}$
2: **input** : B – number of samples
3: **for** $b \leftarrow 1$ to B **do**
4: $\quad \widetilde{\mathcal{PP}} \sim \pi_{\mathcal{M}}(\cdot \mid Z^t \vee Y^t, t)$
5: $\quad \widetilde{\lambda}_b^t \leftarrow \arg\min_{\lambda \succeq 0} \mathbb{L}_p(\lambda \mid \widetilde{\mathcal{PP}}, A, t, \beta^t)$ for $\mathbb{L}_p(\cdot)$ defined in (4.9)
6: **end for**
7: **return** $\{\widetilde{\lambda}_b^t\}_{b=1}^B$.

As it has already been outlined before and in [35], [16], the above scheme produces i.i.d samples and is trivially parallelizable which is a strong numerical advantage in front of MCMC sampling from pure Bayesian posteriors. In what follows 'tilde' will denote samples produced by NPL in ET (either nonparametric or binned).

4.4. Constructions of $\pi_{\mathcal{M}}(\cdot)$ and $\pi_{\mathcal{M}}(\cdot \mid Z^t \vee Y^t, t)$. Binning. In view of the physical model of ET we assume that $\mathcal{PP}$ belongs to the family of temporally stationary Poisson processes, that is

$$\mathcal{PP} = \mathcal{PP}_\Lambda \text{ with some density } \Lambda(z)\,dz\,dt \text{ on } Z \times \mathbb{R}_+, \tag{4.10}$$
$$\Lambda(z) \geq 0 \text{ a.s. and integrable on } Z \text{ w.r.t. } dz.$$

Hence, to build $\pi_{\mathcal{M}}$ we construct a prior on Λ using $\mathcal{M}$, and consequently, the posterior will also defined on Λ while propagating the uncertainty via (4.10) on $\mathcal{PP}$. For the sake of accessibility, discussion of the above assumption (restrictivity and generalizations) with detailed theoretical constructions of nonparametric $\pi_{\mathcal{M}}(\cdot)$ and $\pi_{\mathcal{M}}(\cdot|Z^t \vee Y^t, t)$ are put in Supplementary Materials, section SM6. Below we present finite-dimensional versions which are also used in our numerical experiments.

In finite dimensions (after binning) process $\mathcal{PP}_\Lambda$ boils down to d independent stationary Poisson processes on $\mathbb{R}_+$ with intensities $\Lambda_1, \ldots, \Lambda_d$. For the prior on $\Lambda = (\Lambda_1, \ldots, \Lambda_d)$ we choose the mixture of independent gamma distributions (further denoted by MGP – mixture of gamma processes (due to its nonparametric origin)):

$$\Lambda_{\mathcal{M}} = (\Lambda_{\mathcal{M},1}, \ldots, \Lambda_{\mathcal{M},d}) \sim P_{\mathcal{M}}(\cdot),\ \Lambda_i \mid \Lambda_{\mathcal{M},i} \sim \Gamma(\theta^t \Lambda_{\mathcal{M},i}, (\theta^t)^{-1}),\ i = 1, \ldots, d, \tag{4.11}$$

where $\Lambda_{\mathcal{M}}$ is the mixing parameter which also corresponds to the mean intensity in the MRI-based model from subsection 2.4, $P_{\mathcal{M}}(\cdot)$ is the mixing distribution (hyperprior), θ^t is a positive scalar. The choice of such specific parametrization by θ^t in (4.11) allows to center the gamma distribution on $\Lambda_{\mathcal{M}}$ ($\mathbb{E}[\Lambda|\Lambda_{\mathcal{M}}] = \Lambda_{\mathcal{M},i}$), so θ^t controls only the spread – $\theta^t = 0$ corresponds to improper uniform distribution on $\mathbb{R}_+^d$, $\theta^t = +\infty$ is equal to $\Lambda = \Lambda_{\mathcal{M}} \sim P_{\mathcal{M}}$. In short, for the prior in (4.11) we will use the following notation

$$\pi_{\mathcal{M}}(\cdot) = \mathrm{MGP}(t, P_{\mathcal{M}}(\Lambda_{\mathcal{M}}), \theta^t \Lambda_{\mathcal{M}}, (\theta^t)^{-1}). \tag{4.12}$$

Conjugacy between Poisson distribution of Y^t and Gamma distributions of $\Lambda|\Lambda_{\mathcal{M}}$ implies that

$$\pi_{\mathcal{M}}(\cdot \mid Z^t \vee Y^t, t) = \mathrm{MGP}(t, P_{\mathcal{M}}(\widetilde{\Lambda}^t_{\mathcal{M}} \mid Z^t \vee Y^t, t), Y^t + \theta^t \widetilde{\Lambda}^t_{\mathcal{M}}, (\theta^t + t)^{-1}), \tag{4.13}$$

where $P_{\mathcal{M}}(\widetilde{\Lambda}^t_{\mathcal{M}} \mid Z^t \vee Y^t, t)$ is the posterior for $P_{\mathcal{M}}$ which we specify now. Distribution of $P_{\mathcal{M}}(\Lambda_{\mathcal{M}})$ is defined directly by sampling:

$$\lambda_{\mathcal{M}} = (\lambda^1_1, \ldots, \lambda^1_{p_1}, \ldots, \lambda^r_1, \ldots, \lambda^r_{p_r}) : \lambda^k_s \sim \Gamma(1, \infty),\ \Lambda_{\mathcal{M}} = A_{\mathcal{M}} \lambda_{\mathcal{M}}. \tag{4.14}$$

where $\lambda_{\mathcal{M}}$, $A_{\mathcal{M}}$ are constructed in subsection 2.4, $\Gamma(1, \infty)$ is the uniform (improper) distribution on $\mathbb{R}_+$. Then, posterior $P_{\mathcal{M}}(\widetilde{\Lambda}^t_{\mathcal{M}} \mid Z^t \vee Y^t, t)$ is defined by the classical Bayes formula for model $Y^t \sim \mathrm{Po}(t\Lambda_{\mathcal{M}})$ and the prior in (4.14). In principle, due to moderate size of $A_{\mathcal{M}}$ and good conditioning it is possible to use MCMC-approach (e.g., a Gibbs sampler) to sample from $P_{\mathcal{M}}(\widetilde{\Lambda}^t_{\mathcal{M}} \mid Z^t \vee Y^t, t)$, however, in order to keep the overall implementation as simple as possible we turn to WLB from [38] for approximate posterior sampling.

Algorithm 3 Approximate sampling from $P_{\mathcal{M}}(\widetilde{\Lambda}^t_{\mathcal{M}} \mid Z^t \vee Y^t, t)$ via WLB

1: **data** : Y^t
2: **input** : $A_{\mathcal{M}} \in \mathrm{Mat}(d, p_{\mathcal{M}})$ from (2.15) and (2.17)
3: $\widetilde{\Lambda}^t \leftarrow (\widetilde{\Lambda}^t_1, \ldots, \widetilde{\Lambda}^t_d)$, where independently $\widetilde{\Lambda}^t_i \sim \Gamma(Y^t_i, t^{-1})$
4: $\widetilde{\lambda}^t_{\mathcal{M}} \leftarrow \arg\min\limits_{\lambda_{\mathcal{M}} \succeq 0} L(\lambda_{\mathcal{M}} \mid \widetilde{\Lambda}^t, A_{\mathcal{M}}, 1)$
5: $\widetilde{\Lambda}^t_{\mathcal{M}} \leftarrow A_{\mathcal{M}} \widetilde{\lambda}^t_{\mathcal{M}}$
6: **return** $\widetilde{\Lambda}^t_{\mathcal{M}}$

Remark 4.1. Since we assume that $A_{\mathcal{M}}$ is well-conditioned, minimizer $\widetilde{\lambda}^t_{\mathcal{M}}$ in Step 4 of Algorithm 3 can be efficiently computed via the classical EM-algorithm from [44].

From (4.13) and construction of $P_{\mathcal{M}}(\widetilde{\Lambda}^t_{\mathcal{M}} \mid Z^t \vee Y^t, t)$ one can see that overall MGP posterior acts as (doubly randomized) linear combination of the raw sinogram Y^t and pseudo-sinogram $t\widetilde{\Lambda}^t_{\mathcal{M}}$ proposed by the MRI-based model; see also Figure 4.

4.5. Final algorithm.

Algorithm 4 NPL for ET

1: **data** : Y^t
2: **input** : B – number of samples, θ^t, A, β^t, $\varphi(\lambda)$
3: **for** $b \leftarrow 1$ to B **do**
4: $\widetilde{\Lambda}^t_{\mathcal{M}} \leftarrow (\widetilde{\Lambda}^t_{\mathcal{M},1}, \ldots, \widetilde{\Lambda}^t_{\mathcal{M},d}) \sim P_{\mathcal{M}}(\widetilde{\Lambda}^t_{\mathcal{M}} \mid Z^t \vee Y^t, t)$ via Algorithm 3
5: $\widetilde{\Lambda}^t_b \leftarrow (\widetilde{\Lambda}^t_{b,1}, \ldots \widetilde{\Lambda}^t_{b,d})$, where independently $\widetilde{\Lambda}^t_{b,i} \sim \Gamma(Y^t_i + \theta^t \widetilde{\Lambda}^t_{\mathcal{M},i}, (\theta^t + t)^{-1})$
6: $\widetilde{\lambda}^t_b \leftarrow \arg\min\limits_{\lambda \succeq 0} L_p(\lambda \mid \widetilde{\Lambda}^t_b, A, t, \beta^t/t)$ for $L_p(\cdot)$ defined in (2.10)
7: **end for**
8: **return** $\{\widetilde{\lambda}^t_b\}^B_{b=1}$

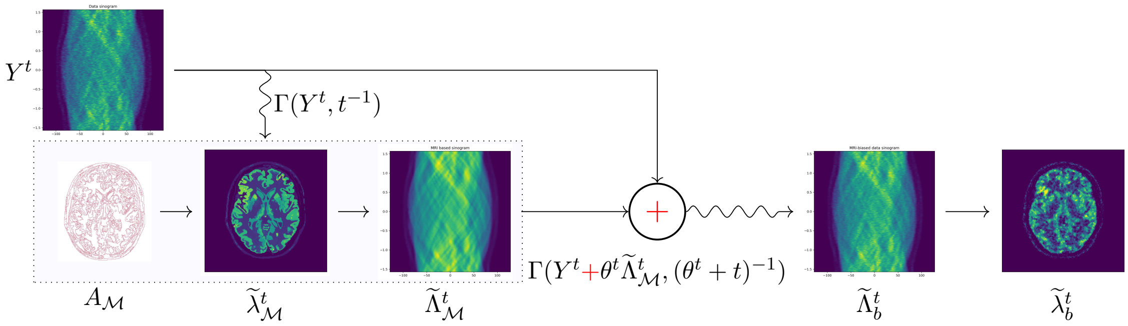


Figure 4: NPL-ET pipeline for one sample in Algorithm 4: wave-like arrows denote randomization of inputs, transparent blue region denotes steps within Algorithm 3.

Remark 4.2. In Step 6 of Algorithm 4 we have used the fact that binned version of $\mathbb{L}_p(\cdot)$ from (4.9) coincides with $L_p(\cdot)$ from (2.10). Moreover,

$$L_p(\lambda \mid t\widetilde{\Lambda}^t_b, A, t, \beta^t) = tL_p(\lambda \mid \widetilde{\Lambda}^t_b, A, 1, \beta^t/t) + R, \tag{4.15}$$

where R is independent of λ, hence, the minimization is directly applied to $L_p(\lambda \mid \widetilde{\Lambda}^t_b, A, 1, \beta^t/t)$ instead of $L_p(\lambda \mid t\widetilde{\Lambda}^t_b, A, t, \beta^t)$. If the numerical complexity of Step 4 is controlled by our choice of $P_{\mathcal{M}}(\cdot)$, Step 6 is inevitable in the paradigm of NPL, hence, it must be numerically feasible via some scalable optimization algorithm. This is the case for us in view of the well-known the Generalized Expectation-Maximization(GEM)-type algorithm from [11] which is specially designed for ET with Poisson-type log-likelihood $L_p(\cdot)$, where $\varphi(\cdot)$ must be a C^2-smooth convex pairwise difference penalty; see Supplementary Materials, section SM7 for details on design of the algorithm.

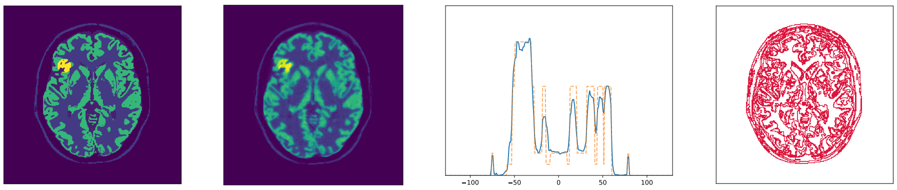
Remark 4.3. Parameter θ^t in Algorithm 4 admits the following interpretation: it is the rate of creation of "pseudo-photons" in the model constructed from MRI data and being conditioned with Y^t. By choosing $\theta^t = \rho t$, $\rho \geq 0$ in Step 5 we sum up sinograms Y^t and $t\widetilde{\Lambda}^t_{\mathcal{M}}$ in proportions $1/(1+\rho)$ and $\rho/(1+\rho)$, respectively. For $\theta^t = 0$ side information $\mathcal{M}$ is not used at all and we see Algorithm 4 as a version of WLB from [38] being adapted for the ET context; see also [35], [16], [42] for connections between the WLB and NPL in the iid setting.

5. Numerical experiment. [1]

5.1. Design. We illustrate Algorithm 4 on synthetic PET data based on a realistic phantom from the BrainWeb database [52]. Typical activity concentrations have been assigned to annotated tissues (gray matter, white matter, skin, etc.) and we delineated a tumor lesion area, not present in the initial phantom with an uptake of 50% compared to the gray matter activity; see Figure 5(a). We consider the worst case scenario for the prior, where the anatomical MRI (T1) phantom (see Figure 5(b)) does not contain any information relative to the lesion. Therefore, model $Y^t \sim \mathrm{Po}(tA_{\mathcal{M}}\lambda_{\mathcal{M}})$ in subsection 2.4 is strongly misspecified

[1] Source code in Python can be found at https://gitlab.com/eric.barat/npl-pet

(with increased bias) in the lesion area. For segmentation of MRI-images we used ddCRP [2] with a concentration parameter equals 10^{-5} leading to a few hundreds of random segments for a 2D brain slice.



(a) λ_* (b) λ_{*opt} (c) λ_* vs. λ_{*opt} (d) $M_1 \in \mathcal{M}$

Figure 5: emission map with lesion hot spot at (a), optimal achievable reconstruction λ_{*opt} at (b), profile through lesion λ_* – orange dotted, λ_{*opt} – in blue at (c) segmented MRI at (b)

The reconstruction grid for images is of size 256×256 ($p = 2^{16}$) being identical to the phantom's one. Acquisition geometry consists of LORs derived from a ring of 512 detectors spaced uniformly on a circle. Design A was computed via classical Siddon's algorithm [45] and $A_{\mathcal{M}}$ was computed from A using formulas (2.15), (2.17). Intensity λ_* was set so that $\sum_{j=1}^{p} \lambda_{*j} = 5 \cdot 10^5$ and for the experiment two sinograms were generated via formula (2.1) for $t \in \{t_1, t_2\}$, $t_1 = 1$, $t_2 = 100$. Case with t_1 corresponds to realistic setting, whereas $t_2 = 100$ is used to describe nearly asymptotic regime of the sampler. Below we present results for $t = t_1$ (for t_2 and additional experiments see Supplementary Materials, subsection SM7.1). To compute λ_{*opt}, we have used (2.14) with $\beta = \beta_{min} = 10^{-3}$, where β_{min} was chosen subjectively such that λ_{*opt} does not contain strong visible numerical artifacts related to the implementation of projector A (see also Remark 4.2). For $\varphi(\lambda)$ convex pairwise-difference penalty from (SM8.1) with hyperparameters (ζ, ν), where the latter were chosen to be always fixed ($\zeta = 0.05$, $\nu = 0.15$) including $\beta^t/t = 2 \times 10^{-3}$.

For $t_1 = 1$ we present results for $\rho = \theta^t/t \in \{0, 0.5, 1.0, 2.0, 4.0\}$; see Remark 4.3. Using Algorithm 4, for each combination of (t, ρ) we generated $B = 1000$ bootstrap draws from which further statistics (empirical mean, variance) as well calibration curves and plots were computed. Main results are presented in Figure 6 and Table Table 1. First, we check visually the effect of ρ on bias and variance (columns (a), (b); no need for λ_{*opt} to compute), and second, calibration of the overall posterior (columns (c), (d), (e); requires λ_{*opt}). For calibration we employ the approach in [51], [21], which says that a model is well-calibrated if for any level $\alpha \in [0, 1]$ (target coverage), the corresponding posterior α-level HPD-intervals (highest probability density) computed pixel-wise will contain λ_{*opt} for $\alpha \cdot 100\%$ of all pixels (achieved coverage – fraction of j's for which $\lambda_{*opt,j} \in [\widehat{q}^{L}_{j,\alpha}, \widehat{q}^{U}_{j,\alpha}]$, where $[\widehat{q}^{L}_{j,\alpha}, \widehat{q}^{U}_{j,\alpha}]$ being the shortest interval such that $P(\widetilde{\lambda}^{t}_{b,j} \in [\widehat{q}^{L}_{j,\alpha}, \widehat{q}^{U}_{j,\alpha}] | Y^t) = \alpha$) (column (c) – reliability curve). Thus, if the achieved coverage is smaller than the target one, then the model is considered to be overconfident and for vice versa – under-confident (or conservative). Note that for practice it is preferable to have slightly conservative model than overconfident one, especially in such domain as medical imaging; see the discussion in [21].

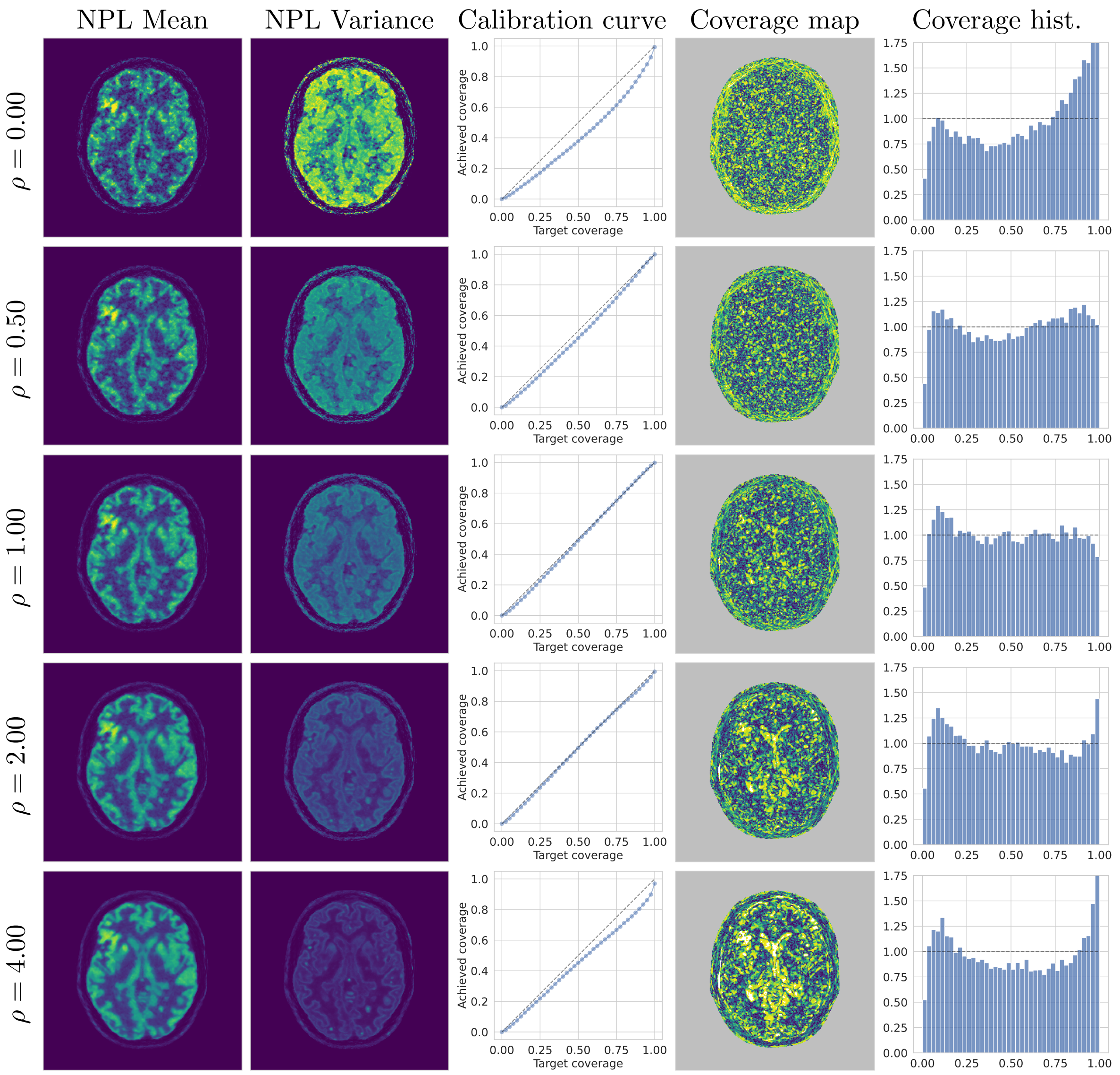


Figure 6: Columns : (a) NPL-mean, (b) NPL-variance (same color scale as mean), (c) calibration curve, (d) coverage probability map (mask in gray), (e) coverage histogram.

ρ	0.00	0.50	1.00	2.00	4.00
PSNR	21.42	24.15	25.29	**25.84**	25.66
MSWD	$8.74\,10^{5}$	1.16	**0.83**	1.04	1.78
ECE	$9.41\,10^{-2}$	$3.29\,10^{-2}$	$1.35\,10^{-2}$	$\mathbf{1.03\,10^{-2}}$	$5.29\,10^{-2}$
KLC	$5.52\,10^{-2}$	$1.20\,10^{-2}$	$\mathbf{9.76\,10^{-3}}$	$1.14\,10^{-2}$	$3.85\,10^{-2}$

Table 1: Performance metrics

Since the definition of calibration does not take into account correlations between pixels, columns (d), (e) are used for diagnostic of the latter. Coverage map (d) shows for each pixel the smallest probability so that the HPD-interval contains λ_{*opt} while the normalized coverage histogram in (e) corresponds to the (empirical) probability density function for the coverage curve in (c) being viewed as c.d.f. (note that for perfect calibration the c.d.f. in (c) and p.d.f. in (e) correspond to the uniform distribution on $[0, 1]$). In Table 1 we compute PSNR for the NPL-mean, ECE – expected calibration error (ℓ_1-norm between the calibration curve in (c) and diagonal $x = y$ on $[0, 1]$), MSWD – mean-squared weighted deviation between λ_{*opt} and the NPL-mean, KLC – Kullback-Leibler divergence between uniform distiribution and coverage histogram in (e); see Supplementary Materials, subsection SM8.2 for precise definitions and connections to other metrics.

5.2. Interpretation. The increase of ρ reduces the noise, but on the other hand, it increases bias in the lesion area; see Figure 6(a) and PSNR in Table 1. The latter is due to the aforementioned misspecification, therefore the high signal is being spread over the larger segment in $\mathcal{M}$ containing the lesion. Being subjective, for us the most visually appealing results for the trade-off between noise and preservation of contours of the lesion were obtained for $\rho \in \{0.50, 1.00\}$. Note also that pixel-wise variance in (b) decreases. However, for $\rho \to +\infty$ the limit is not zero but the posterior variance in the MRI-based model[2] which is much smaller, for example, than for $\rho = 0$ (because Y^t contains much more information for low-dimensional $\lambda_{\mathcal{M}} \in \mathbb{R}^{p_{\mathcal{M}}}$ than for $\lambda \in \mathbb{R}^p$). Spikes for variance in (b) (e.g., for $\rho = 4.0$) correspond to smallest segments in $\mathcal{M}$ where the signal is more sensitive to perturbations in $\widetilde{\Lambda}_b^t$ due to ill-conditioning nature of A. With calibration results in (c), (d), (e), and in Table 1 we can choose objectively one optimal ρ by arguing on guarantees of covering λ_{*opt} by the posterior. First, note that for $\rho = 0$ the posterior is essentially overconfident (columns (c), (e)) – this is due to large amount of pixels in the slab between the cranium and soft tissues (exterior yellow ring on images in (d)) where in fact the isotope concentration is zero. Coverage map (d) and histogram (e) reveal that these pixels require very large credible levels to cover λ_{*opt} meaning that the posterior in this region is overcontracted. We explain the overcontraction by the fact that for many LORs crossing such pixels and nearly tangential to the brain the intensities Λ_i^* are so small (though positive) that for $t = 1$ (mild regime) it happens that $Y_i^t = 0$. Then, in Step 5 one can see that $\widetilde{\Lambda}_{b,i}^t \sim \Gamma(0, t^{-1}) = \delta_0$ for $\rho = 0$, so $\widetilde{\Lambda}_{b,i}^t \equiv 0$ c.a.s. and no uncertainty can propagate from such LOR in Step 6 which results in overcontraction. Moreover, in subsection 6.3 we show that for Poisson model the event $Y_i^t = 0$ make the posterior contract much stronger to zeros in pixels intersected by LOR i (effect of positivity constraints in Step 6) which is another argument for overcontraction. Finally, overcontraction was already reported for (non-Poisson) WLB in [40] with a proposal to fix it different from NPL; see also Remark 4.3. An additional numerical experiment supporting our explanation is given in the Supplementary Materials, subsection SM8.3. For $\rho \in \{0.5, 1.00\}$, since the aforementioned empty slab is splitted into larger segments for which $\widetilde{\Lambda}_{\mathcal{M},i}^t > 0$, the overcontraction is corrected while improving the overall calibration and reaching the optimum for KLC and MSWD at $\rho = 1.0$ (see (c), (e) and ECE, KLC in Table 1). Further increase $\rho \in \{2.00, 4.00\}$

[2] $\mathrm{var}[\widetilde{\Lambda}_b^t|Y^t] = (Y^t/t + \rho\mathbb{E}[\widetilde{\Lambda}_{\mathcal{M}}^t|Y^t])/t^2(1+\rho)^2 + \rho^2\mathrm{var}[\widetilde{\Lambda}_{\mathcal{M}}^t|Y^t]/t^2(1+\rho)^2$, $\lim_{\rho\to+\infty} \mathrm{var}[\widetilde{\Lambda}_b^t|Y^t] = \mathrm{var}[\widetilde{\Lambda}_{\mathcal{M}}^t|Y^t]/t^2$

results in increased bias in the lesion area and since the posterior intervals are being more contracted, the posterior again turns to be overconfident (see column (c) supported by sharp increase for high confidence levels in (e) and also large yellow structures in (d) in the lesion and central segments). In conclusion, calibration with ρ is simple and tractable, seemingly with one optimum w.r.t bias (in the lesion) and (global-)variance trade-off.

6. Asymptotic analysis of the algorithm. Statistical model (2.1) is non-regular since domain $\mathbb{R}^p_+$ contains a boundary and, often it is the case that $\lambda_* \in \partial\mathbb{R}^p_+$. The results of [3] for the classical Bayesian framework show that for the well-specified case and large class of priors the posterior is consistent at λ_* and the asymptotic distribution is complex because it splits in three modes due to the effect of positivity constraints (exponential, Gaussian and half-Gaussian; two latter have the same standard contraction rates but the first one). Consistency at λ_* and a very similar splitting are also present in NPL with the asymptotic distribution being tight around strongly consistent estimator $\widehat{\lambda}^t_{sc}$ satisfying some contraction properties in observation (sinogram) space. Interestingly, the aforementioned splitting depends not on λ_* (as it was in [3]) but again on $\widehat{\lambda}^t_{sc}$ because of which yet we fail to demonstrate fully the asymptotic normality since it requires additional results on behavior of strongly consistent estimators with constraints on the domain (detailed discussion is given in Supplementary Materials, section SM9).

The problem of misspecification for the generalized Poisson model with wrong design arises twice our setting: first, in Algorithm 3 when sampling $\widetilde{\Lambda}^t_{\mathcal{M}}$ (because we assume that $Y^t \sim P^t_{A_{\mathcal{M}},\lambda_{\mathcal{M}}}$ whereas $Y^t \sim P^t_{A,\lambda_*}$) and, second, when we assume that model (2.1) is wrong, in general. Suprisingly, in this simple case the identifiability of λ_* can be lost even for injective designs which we show by an explicit example below. We propose an intuitive sufficient condition on observed intensities along LORs and design A to retrieve it back.

6.1. Convergence for conditional probabilities. Let $(\Omega, \mathcal{F}, P)$ be the common probability space on which process Y^t, $t \in [0, +\infty)$ and MGP prior in (4.12) are defined (see Supplementary Materials, section SM1 for details). By $U \,|\, Y^t$ we denote the distribution of U conditionally on the sigma algebra generated by Y^τ, $\tau \in [0, t)$.

Definition 6.1. *We say that U^t converges in conditional probability to U almost surely Y^t if for every $\varepsilon > 0$ the following holds:*

$$P(\|U^t - U\| > \varepsilon \mid Y^t) \to 0 \text{ when } t \to +\infty, \text{ a.s. } Y^t, t \in [0, +\infty). \tag{6.1}$$

This type of convergence will be denoted as follows:

$$U^t \xrightarrow{c.p.} U. \tag{6.2}$$

In our proofs for $U^t \xrightarrow{c.p.} 0$ we also write

$$U^t = o_{cp}(1). \tag{6.3}$$

Definition 6.2. *We say that U^t is conditionally tight almost surely Y^t if for any $\varepsilon > 0$ and almost any trajectory Y^t, $t \in [0, +\infty)$ there exists $M = M(\varepsilon, \{Y^t\}_{t \in (0,+\infty)})$ such that*

$$\sup_{t \in [0,+\infty)} P(\|U^t\| > M \mid Y^t) < \varepsilon. \tag{6.4}$$

In short, in the definitions above almost surely Y^t means that statements in (6.1), (6.4) hold for almost every trajectory Y^t, $t \in [0, +\infty)$.

6.2. Consistency.

Assumption 6.3. Model (2.1) is well-specified, that is

$$Y^t \sim PP^t_{A,\lambda_*}, \text{ for some } \lambda_* \in \mathbb{R}^p_+ \text{ and all } t \in [0, +\infty), \tag{6.5}$$

where A satisfies (2.3)–(2.6), $PP^t_{A,\lambda}$ is defined in (2.8).

Theorem 6.4. *Let Assumption 6.3 and conditions (2.11), (2.12) for φ be satisfied. Let also β^t, θ^t be such that*

$$\beta^t/t \to 0, \ \theta^t/t \to 0 \ \text{ when } t \to +\infty. \tag{6.6}$$

Then,

$$\widetilde{\lambda}^t_b \xrightarrow{c.p.} \lambda_{*opt}, \tag{6.7}$$

where $\widetilde{\lambda}^t_b$ is sampled in Algorithm 4, λ_{*opt} is defined in (2.14).

The above result is merely a consequence a more general statement for any bootstrap-type procedure which is given below.

Theorem 6.5. *Let conditions of Theorem 6.4 be satisfied but Assumption 6.3. Assume also that*

$$\widetilde{\Lambda}^t_b \xrightarrow{c.p.} \Lambda^* = A\lambda_* \text{ for some } \lambda_* \in \mathbb{R}^p_+. \tag{6.8}$$

Then, formula (6.7) remains valid.

Thus the conditional distribution of $\widetilde{\lambda}^t_b$ asymptotically concentrates at λ_* in the subspace where parameter λ is identifiable through design A and also regarding the positivity constraints. Projection of λ_* onto $\ker(A)$ which not "visible" by positivity constraints is not identifiable in model (2.1) and it is defined solely by $w_{A,\lambda_*}(0)$; see formula (2.13).

6.3. Tightness.

Assumption 6.6. $A_{\mathcal{M}} \in \mathrm{Mat}(d, p_{\mathcal{M}})$ is injective.

Assumption 6.7 (non-expansiveness condition). Let $\Lambda^* \in \mathbb{R}^d_+$, $A_{\mathcal{M}} \in \mathrm{Mat}(d, p_{\mathcal{M}})$, $A_{\mathcal{M}}$ has only positive entries and analog of (2.4) for $A_{\mathcal{M}}$ holds (i.e., $A_{\mathcal{M},j} = \sum_{i=1}^{d} a_{\mathcal{M},ij} > 0$). Define set

$$\lambda_{\mathcal{M},*} = \operatorname*{arg\,min}_{\lambda_{\mathcal{M}} \succeq 0} L(\lambda_{\mathcal{M}} \mid \Lambda^*, A_{\mathcal{M}}, 1), \tag{6.9}$$

where $L(\lambda_{\mathcal{M}} \mid \Lambda^*, A_{\mathcal{M}}, 1)$ is given in (2.8). There exists at least one point in $\lambda_{\mathcal{M},*}$ for which the following holds:

$$I_0(\Lambda^*_{\mathcal{M}}) = I_0(\Lambda^*), \ \Lambda^*_{\mathcal{M}} = A_{\mathcal{M}}\lambda_{\mathcal{M},*}, \tag{6.10}$$

where $I_0(\cdot)$ is defined in (2.2).

The proposition below states that the non-expansiveness condition is always meaningful and not very restrictive (for more details see Supplementary Materials, section SM11).

Proposition 6.8. *Let $\Lambda^* \in \mathbb{R}^d_+$, $A_{\mathcal{M}} \in \mathrm{Mat}(d, p_{\mathcal{M}})$, $A_{\mathcal{M}}$ has only positive entries and the analog of (2.4) for $A_{\mathcal{M}}$ holds (i.e., $A_{\mathcal{M},j} = \sum_{i=1}^{d} a_{\mathcal{M},ij} > 0$). Then, the set of minimizers in (6.9) is non-empty and constitutes an affine subset of $(p_{\mathcal{M}}-1)$-dimensional simplex $\Delta^p_{A_{\mathcal{M}}}(\Lambda^*)$ defined by the formula:*

$$\Delta^{p_{\mathcal{M}}}_{A_{\mathcal{M}}}(\Lambda^*) = \{\lambda_{\mathcal{M}} \in \mathbb{R}^p_+ \mid \sum_{j=1}^{p_{\mathcal{M}}} A_{\mathcal{M},j}\lambda_{\mathcal{M},j} = \sum_{i=1}^{d} \Lambda^*_i \geq 0\}. \tag{6.11}$$

Moreover, it always holds that

$$I_1(\Lambda^*) \subset I_1(\Lambda^*_{\mathcal{M}}) \text{ or equivalently } I_0(\Lambda^*_{\mathcal{M}}) \subset I_0(\Lambda^*), \text{ where } \Lambda^*_{\mathcal{M}} = A_{\mathcal{M}}\lambda_{\mathcal{M},*}. \tag{6.12}$$

The aim of the non-expansiveness condition is to have a unique and stable KL-minimizer $\lambda_{\mathcal{M},*}$ so that the the prior effect of $\mathcal{M}$ on $\widetilde{\lambda}^t_b$ via $\widetilde{\Lambda}^t_{\mathcal{M}}$ (which concentrates near $\Lambda_{\mathcal{M},*} = A_{\mathcal{M}}\lambda_{\mathcal{M},*}$) is not spread ambiguously among different (but equivalent in terms of observations) combinations of signals in segments of $\mathcal{M}$. This is provided by the theorem below.

Theorem 6.9 (identifiability in the prior model). *Let Assumptions 6.6 and 6.7 be satisfied. Then, $\lambda_{\mathcal{M},*}$ defined in (6.9) has only one point and the following approximation holds:*

$$\begin{aligned} L(\lambda_{\mathcal{M}} \mid \Lambda^*, A_{\mathcal{M}}, 1) - L(\lambda_{\mathcal{M},*} \mid \Lambda^*, A_{\mathcal{M}}, 1) &= \mu^T_{\mathcal{M},*}\lambda_{\mathcal{M}} + \frac{1}{2}\sum_{i\in I_1(\Lambda^*)} \Lambda^*_i \frac{(\Lambda_{\mathcal{M},i} - \Lambda^*_{\mathcal{M},i})^2}{(\Lambda^*_{\mathcal{M},i})^2} \\ &\quad + o(\|\Pi_{A^T_{\mathcal{M},I_1(\Lambda^*)}}(\lambda_{\mathcal{M}} - \lambda_{\mathcal{M},*})\|^2), \end{aligned} \tag{6.13}$$

where $\Pi_{A^T_{\mathcal{M},I_1(\Lambda^*)}}$ denotes the orthogonal projector onto $\mathrm{Span}(A^T_{\mathcal{M},I_1(\Lambda^*)})$,

$$\begin{aligned} \mu_{\mathcal{M},*} &= \sum_{i\in I_1(\Lambda^*)} -\Lambda^*_i \frac{a_{\mathcal{M},i}}{\Lambda^*_{\mathcal{M},i}} + \sum_{i=1}^{d} a_{\mathcal{M},i}, \\ \mu_{\mathcal{M},*} &\succeq 0,\ \mu_{\mathcal{M},*,j}\lambda_{\mathcal{M},*,j} = 0 \text{ for all } j \in \{1,\ldots,p_{\mathcal{M}}\}. \end{aligned} \tag{6.14}$$

In particular, $L(\lambda_{\mathcal{M}} \mid \Lambda^*, A, 1)$ is strongly convex at $\lambda_{\mathcal{M},*}$, so, there exists an open ball $B_* = B(\lambda_{\mathcal{M},*}, \delta_*)$, $\delta_* = \delta_*(A_{\mathcal{M}}, \Lambda_*) > 0$ and constant $C_* = C_*(A_{\mathcal{M}}, \Lambda_*) > 0$ such that

$$L(\lambda_{\mathcal{M}} \mid \Lambda^*, A_{\mathcal{M}}, 1) - L(\lambda_{\mathcal{M},*} \mid \Lambda^*, A_{\mathcal{M}}, 1) \geq C_*\|\lambda_{\mathcal{M}} - \lambda_{\mathcal{M},*}\|^2,\ \lambda \in B_* \cap \mathbb{R}^{p_{\mathcal{M}}}_+. \tag{6.15}$$

Result of Theorem 6.9 is also a positive answer to the general identification problem when model (2.1) is misspecified in the sense of wrong design. In subsection 6.4 we show that the non-expansiveness condition is essential and counterexamples are possible if it is removed.

Now we can turn to our main result on the tightness of the NPL-posterior.

Let $\{e_j\}_{j=1}^p$ be the standard basis in $\mathbb{R}^p$ and define the following spaces:

$$\mathcal{V} = \operatorname{Span}\{e_j \mid \exists\, i \in I_0(\Lambda^*) \text{ s.t. } a_{ij} > 0\}, \tag{6.16}$$

$$\mathcal{U} = \mathcal{V}^\perp \cap \operatorname{Span}\{A^T_{I_1(\Lambda^*)}\}, \tag{6.17}$$

$$\mathcal{W} = (\mathcal{V} \oplus \mathcal{U})^\perp \cap \ker A. \tag{6.18}$$

Let also

$$\Pi_\mathcal{V}, \Pi_\mathcal{V}, \Pi_\mathcal{W} \text{ be the orthogonal projectors on } \mathcal{V}, \mathcal{V}, \mathcal{W}, \text{ respectively.} \tag{6.19}$$

Theorem 6.10. *Let Assumptions 6.3 and 6.7 be satisfied and assume also that*

$$\varphi \text{ satisfies (2.11), (2.12) and it is locally Lipschitz continous.} \tag{6.20}$$

Let $\widetilde{\lambda}^t_b$ be defined as in Algorithm 4, $\theta^t = o(\sqrt{t/\log\log t})$, $\beta^t = o(\sqrt{t})$ and assume that there exists a strongly consistent estimator $\widehat{\lambda}^t_{sc}$ of λ_* on $\mathcal{V} \oplus \mathcal{U}$ (i.e., $\Pi_{\mathcal{U}\oplus\mathcal{V}}\widehat{\lambda}^t_{sc} \xrightarrow{a.s.} \Pi_{\mathcal{U}\oplus\mathcal{V}}\lambda_*$) such that

$$\widehat{\lambda}^t_{sc} \succeq 0, \tag{6.21}$$

$$\limsup_{t\to+\infty} \left| \sum_{i \in I_1(\Lambda^*)} \sqrt{t} \frac{Y_i^t/t - \widehat{\Lambda}^t_{sc,i}}{\widehat{\Lambda}^t_{sc,i}} a_i \right| < +\infty \text{ a.s. } Y^t, \tag{6.22}$$

$$t\widehat{\Lambda}^t_{sc,i} \xrightarrow{a.s.} 0 \text{ for } i \in I_0(\Lambda^*), \tag{6.23}$$

where $\widehat{\Lambda}^t_{sc} = A\widehat{\lambda}^t_{sc}$. Then,

(i)

$$t\Pi_\mathcal{V}(\widetilde{\lambda}^t_b - \widehat{\lambda}^t_{sc}) \xrightarrow{c.p.} 0. \tag{6.24}$$

(ii)

$$\sqrt{t}\Pi_\mathcal{U}(\widetilde{\lambda}^t_b - \widehat{\lambda}^t_{sc}) \text{ is conditionally tight a.s. } Y^t. \tag{6.25}$$

Statement in (i) claims that for pixels which are interested by LORs with $\Lambda^*_i = 0$, the posterior distribution contracts to zero with faster rate than for the ones intersected by LORs with positive intensities. Indeed, pixels in subspace $\mathcal{V}$ are strongly forced to be zeros by the positivity constraints (i.e., if $\Lambda^*_i = 0$ and λ_*, $a_i \in \mathbb{R}^p_+$, then necessarily $\lambda_{*,j} = 0$ where $a_{ij} > 0$). Statement in (ii) claims that, in general, the posterior concentrates around $\widehat{\lambda}^t_{sc}$ in subspace $\mathcal{U}$ with standard scaling rate $\sqrt{t}$. This is not surprising since $\mathcal{U}$ is orthogonal to $\mathcal{V}$, so the positivity constraints do not give extra information to achieve the faster contraction rate. Finally, requiring the non-expansiveness condition for the prior (Assumption 6.7) may seem surprising at first sight. The intuition behind is that it forbids our sampler to create "too many" pseudo-photons in LORs where intensity is zero a.s. ($\Lambda^*_i = 0$ implies $Y_i^t \equiv 0$) and significantly simplifies the theoretical analysis.

For $\widehat{\lambda}^t_{sc}$ we propose to take the MAP-estimate which is defined by the formula:

$$\widehat{\lambda}^t_{pMLE} = \arg\min_{\lambda \succeq 0} L_p(\lambda \mid Y^t, A, t, \beta^t), \tag{6.26}$$

where $L_p(\cdot)$ is defined in (2.10).

Conjecture 6.11. Let assumptions of Theorem 6.10 be satisfied and $\widehat{\lambda}^t_{sc} = \widehat{\lambda}^t_{pMLE}$, where the latter is defined by (6.26). Then, $\widehat{\lambda}^t_{sc}$ is a strongly consistent estimator of λ_* on $\mathcal{V} \oplus \mathcal{V}$ and formulas (6.21)–(6.23) hold.

The requirement for existence of a strongly consistent estimator is not new and already appears for WLB in [39]. However, in that case the sampling is performed via unconstrained optimization of quadratic functionals with ℓ_1-penalties for which existence of such estimators is trivial by taking the standard OLS estimator or LASSO estimator; see the discussion after Theorem 3.3 in [39]. In our case, according to Kolmogorov's 0-1 Law the statements in (6.22) and (6.23) either hold with probability one (i.e., almost surely Y^t, $t \in [0, +\infty)$) or zero, and the case of zero probability would mean a very exotic and unexpected behavior of the constrained MLE estimate for such model because they are trivially satisfied, for example, if A is diagonal. Another plausible argument in favour of existence of required $\widehat{\lambda}^t_{sc}$ comes from [3] where the asymptotic posterior mean is strongly consistent and satisfies (6.21)-(6.23) (for details see Supplementary Materials, section SM9).

Finally, establishing tightness of the posterior is the first step towards the proof of asymptotic normality (see Bernstein von-Mises type theorems in [49], [39], [42]) which, in particular, implies that for large dataset the posterior distribution, in general (but not always if misspecified; see e.g., [27]; an interesting case of posterior inconsistency was found in [19]), is correctly calibrated against frequentist distribution of some strongly consistent estimator.

6.4. Misspecification in design and identifiability. Assumption 6.3 in subsection 6.2 reflects our belief that model (2.1) is correct. At the same time, for any practitioner in ET it is known that such model is by far approximate: the tracer inside the human body surely does not respect locally constant behavior, design A is known only approximately (with non-negligible errors, since it contains patient's attenuation map which is reconstructed via a separate MRI or CT scan; see e.g., [48]), non-stationarity of the process due to kinetics of the tracer, scattered photons, errors from multiple events etc.; see e.g., [29], [43].

Assume that exposure period is $[0, t)$ and PP^t is the unknown (binned) process that generates Y^t:

$$\begin{aligned} &Y^t \sim PP^t,\, Y^t \in (\mathbb{N}_0)^d, \\ &\mathbb{E}_{PP^t}[Y^t] = \mathrm{var}_{PP^t}[Y^t] = \Lambda^*(t) \text{ for some } \Lambda^*(t) = (\Lambda^*_1(t), \ldots, \Lambda^*_d(t)) \in \mathbb{R}^d_+. \end{aligned} \tag{6.27}$$

Formulas in (6.27) reflect our belief that Y^t has Poisson-type behavior (e.g., non-stationary Poisson process) at least for its two first moments which is not far from truth in practice [47]. Most importantly, we do not assume that $\Lambda^*(t) \in R_+(A)$. The main question now is the identifiability of λ which translated via (2.9) and (6.27) to the question of uniqueness in the following minimization problem:

$$\lambda_*(PP, [0,t)) = \arg\min_{\lambda \succeq 0} \mathcal{KL}(PP^t, PP^t_{A\lambda}) = \arg\min_{\lambda \succeq 0} L(\lambda \mid \Lambda^*(t)/t, A, 1), \tag{6.28}$$

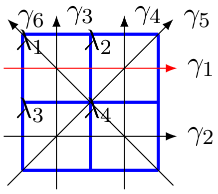
where $PP^t_{A\lambda}$ is defined in (2.8). It appears that, in general, the answer is negative even for very meaningful choices of A and $\Lambda^*(t)$.

Theorem 6.12. *Let $t = 1$. There exist $\Lambda^* = (\Lambda^*_1, \dots, \Lambda^*_d) \in \mathbb{R}^d_+$, $\Lambda^* \neq 0$, $A \in \mathrm{Mat}(d, p)$ which has only nonnegative entries, it is stochastic column-wise and injective such that solutions of the optimization problem (6.28) constitute a non-empty polytope of positive dimension of the $(p-1)$-simplex $\Delta_p(\Lambda^*) = \left\{\lambda \in \mathbb{R}^p_+ : \sum_{j=1}^p \lambda_j = \sum_{i=1}^d \Lambda^*_i\right\}$.*

Proof. We construct Λ^* and A for $p = 4$, $d = 6$. Let $\mathcal{I}$ be the square of four pixels each with side length 1 as shown below, i.e., $\lambda = (\lambda_1, \dots, \lambda_4) \in \mathbb{R}^4_+$, and $\Gamma = \{\gamma_1, \dots, \gamma_6\}$ be the set of rays. Let A' be the classical Radon transform on $\mathcal{I}$ for geometry Γ (i.e., a'_{ij} being the length of intersection of ray γ_i with pixel j):

$$A' = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & \sqrt{2} & \sqrt{2} & 0 \\ \sqrt{2} & 0 & 0 & \sqrt{2} \end{pmatrix}$$

$$\det(A'^T A') = 128 \neq 0.$$



$\mathcal{I}$

Let A be a column-wise normalization of A', i.e., $a_{ij} = a'_{ij}/(\sum_i a'_{ij})$ (this obviously does not break the injectivity of A'). Let $\Lambda^* = (1, 0, 0, 0, 0, 0)$. Then, for (6.28) we get

$$\lambda_* = \arg\min_{\lambda \succeq 0} -\log\left(\frac{\lambda_1 + \lambda_2}{2 + \sqrt{2}}\right) + \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4. \tag{6.29}$$

Note that in (6.29) we have used the fact that $\sum_i a_{ij} = 1$ for all $j \in \{1, \dots, 4\}$. It is obvious that the set of minimizers in (6.29) is an affine set of the following form:

$$\lambda_{*3} = \lambda_{*4} = 0,\ \lambda_{*1} + \lambda_{*2} = 1 \tag{6.30}$$

which gives the desired non-uniqueness. Theorem is proved. ■

Finally, note that Theorem 6.9 provides identifiability under the non-expansiveness condition and injectivity of A.

7. Discussion.

Algorithm 4 solves Problems 1 and 2 simultaneously and efficiently: generated samples are automatically iid, algorithm is scalable because the crucial Step 6 is performed via the classical GEM-type algorithm and, finally, our main calibration parameter ρ ($\theta^t = t\rho$, $\rho \geq 0$; see Remark 4.3) can be interpreted as amount of pseudo-data (pseudo-photons) generated from the MRI-based posterior. Due to the latter the numerical calibration of the posterior is tractable. Moreover, in our experiment on the synthetic dataset for the worst case scenario (when MRI has no information on the lesion) we have observed that moderate values of ρ, indeed, improve calibration error as well PSNR and MSWD. Our principal theoretical results (posterior consistency and tightness) are complicated by the non-standard form of ET but show a great number of connections to existing works ([35], [16], [3]). The new non-expansiveness condition (Assumption 6.7) is of independent geometric interest and is a key to extend all previous results to the fully misspecified case. Among possible extensions, one most interesting for us is to relax the independence of increments of the Gamma process in the prior and consider ones with correlations (for example, scaled Polyà-tree priors for $\Lambda_{\mathcal{M}}$). These correlations can be used to smooth out sinogram Y^t by projecting it (non-linearly) on the stable part of Span(A^T) using an MRI-based model and, in addition, remove completely the need for regularizer φ (high frequencies are still regularized by φ whereas $\mathcal{M}$ is used for low-frequencies). Our preliminary numerical results show that it improves resolution while retaining the interpretability of calibration parameters as before. Another improvement could be to replace the (random) segmentations of MRI-images via ddCRP with other machine learning techniques (such as DNNs) that on input will take MRI-scans with sinograms and output possible low-dimensional models $A_{\mathcal{M}}, \lambda_{\mathcal{M}}$ (possibly corrected by medical experts). This has a chance to reduce bias in the lesion while non-increasing the calibration error and variance. Finally, an experiment on real PET-MRI data is of great importance and will be given elsewhere.

Supplementary materials. Supplementary materials include discussion of the assumption in (4.10) and remarks on nonparametric constructions in subsection 4.4, all details of numerical experiments in sections 3 and 5 (with additional numerical experiments for large t), proofs of all theoretical results in section 6, a separate discussion of results on ET from [3] with connections to Conjecture 6.11, a remark on the geometric intuition behind the non-expansiveness condition (Assumption 6.7) and a remark on the choice of centering term in Theorem 6.10.

Acknowledgments. We are grateful to Zacharie Naulet from Université d'Orsay for many valuable comments on statistical side of the paper, to our colleagues from Service Hospitalier Frédéric Joliot (SHFJ) – Marina Filipović, Claude Comtat and Simon Stute for many practical insights on the topic of PET-MRI reconstructions and also to anonymous referees for remarks that have helped to improve greatly the presentation of this work, especially in the numerical part.

REFERENCES

[1] H. H. Barrett, D. W. Wilson, and B. M. Tsui, *Noise properties of the em-algorithm. i. theory*, Phys. Med. Biol., 39 (1994), p. 833.

[2] D. Blei and P. Frazier, *Distance dependent chinese restaurant processes*, Journal of Machine Learning Research, 12 (2011).

[3] N. A. Bochkina and P. J. Green, *The bernstein–von mises theorem and nonregular models*, The Annals of Statistics, 42 (2014), pp. 1850–1878.

[4] J. Bowsher, V. Johnson, T. Turkington, R. Jaszczak, C. Floyd, and R. Coleman, *Bayesian reconstruction and use of anatomical a priori information for emission tomography*, IEEE Transactions on Medical Imaging, 15 (1996), pp. 673–686.

[5] J. Bowsher, H. Yuan, L. Hedlund, T. Turkington, G. Akabani, A. Badea, W. Kurylo, C. Wheeler, G. Cofer, M. Dewhirst, and G. Johnson, *Utilizing mri information to estimate f18-fdg distributions in rat flank tumors*, in IEEE Symposium Conference Record Nuclear Science, vol. 4, IEEE, 2004.

[6] S. Y. Chun, J. A. Fessler, and Y. K. Dewaraja, *Post-reconstruction non-local means filtering methods using ct side information for quantitative spect*, Physics in Medicine & Biology, 58 (2013), p. 6225.

[7] C. Comtat, P. E. Kinahan, J. A. Fessler, T. Beyer, D. W. Townsend, M. Defrise, and C. Michel, *Clinically feasible reconstruction of 3d whole-body pet/ct data using blurred anatomical labels*, Physics in Medicine & Biology, 47 (2001), p. 1.

[8] M. Dahlbom, *Estimation of image noise in pet using the bootstrap method*, in IEEE Nuclear Science Symposium Conference Record, vol. 4, IEEE, 2001.
[9] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes: volume I: elementary theory and methods*, Springer Science & Business Media, 2005.
[10] A. R. Ferreira and K. H. Lee, *Single photon emission computed tomography example*, in Multiscale Modeling, Springer Series in Statistics, 2007.
[11] J. Fessler and A. Hero, *Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms*, IEEE Transactions on Image Processing, 4 (1995), pp. 1417–1429.
[12] J. A. Fessler, *Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): Applications to tomography*, IEEE Transactions on Image Processing, 5 (1996), pp. 493–506.
[13] J. A. Fessler, N. H. Clinthorne, and W. L. Rogers, *Regularized emission image reconstruction using imperfect side information*, IEEE Transactions on Nuclear Science, 39 (1992), pp. 1464–1471.
[14] M. Filipović, E. Barat, T. Dautremer, C. Comtat, and S. Stute, *Pet reconstruction of the posterior image probability, including multimodal images.*, IEEE transactions on medical imaging, 38 (2018), pp. 1643–1654.
[15] M. Filipović, T. Dautremer, C. Comtat, S. Stute, and E. Barat, *Reconstruction, analysis and interpretation of posterior probability distributions of pet images, using the posterior bootstrap*, Physics in Medicine & Biology, 66 (2021), p. 125018.
[16] E. Fong, S. Lyddon, and C. Holmes, *Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap*, in Proceedings of the 36th International Conference on Machine Learning, vol. 97, PMLR, 09–15 Jun 2019, pp. 1952–1962.
[17] S. Ghosh, A. B. Ungureanu, E. B. Sudderth, and D. M. Blei, *Spatial distance dependent chinese restaurant processes for image segmentation*, in Advances in Neural Information Processing Systems 24, Curran Associates, Inc., 2011, pp. 1476–1484.
[18] P. J. Green, *Bayesian reconstructions from emission tomography data using a modified em algorithm*, IEEE Trans. Med. Imag., 9 (1990), pp. 84–93.
[19] P. Grünwald and T. van Ommen, *Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It*, Bayesian Analysis, 12 (2017), pp. 1069 – 1103.
[20] D. R. Haynor and S. D. Woods, *Resampling estimates of precision in emission tomography*, IEEE Transactions on Medical Imaging, 8 (1989), pp. 337–343.
[21] J. Hermans, A. Delaunoy, F. Rozet, A. Wehenkel, and G. Louppe, *Averting a crisis in simulation-based inference*, arXiv preprint arXiv:2110.06581v2, (2021).
[22] A. O. Hero, R. Piramuthu, J. A. Fessler, and S. R. Titus, *Minimax emission computed tomography using high-resolution anatomical side information and b-spline models*, IEEE Transactions on Information Theory, 45 (1999), pp. 920–938.
[23] D. Higdon, J. Bowsher, V. Johnson, T. Turkington, D. Gilland, and R. Jaszczak, *Fully bayesian estimation of gibbs hyperparameters for emission computed tomography data*, IEEE Transactions on Medical Imaging, 16 (1997), p. 516.
[24] T. Hohage and F. Werner, *Inverse problems with poisson data: statistical regularization theory, applications and algorithms*, Inverse Problems, 32 (2016), p. 093001.
[25] L. F. James, *Bayesian calculus for gamma processes with applications to semiparametric intensity models*, Sankhyā: The Indian Journal of Statistics, (2003), pp. 179–206.
[26] M. Judenhofer, H. Wehrl, D. Newport, C. Catana, S. Siegel, M. Becker, A. Thielscher, M. Kneilling, M. Lichy, M. Eichner, K. Klingel, G. Reischl, S. Widmaier, M. Röcken, R. Nutt, H. Machulla, K. Uluda, S. Cherry, C. Claussen, and B. Pichler, *Simultaneous pet-mri: a new approach for functional and morphological imaging*, Nature medicine, 14 (2008), pp. 459–465.
[27] B. J. K. Kleijn and A. W. Van der Vaart, *The bernstein-von-mises theorem under misspecification*, Electronic Journal of Statistics, 6 (2012), pp. 354–381.
[28] C. Lartizien, J.-B. Aubin, and I. Buvat, *Comparison of bootstrap resampling methods for 3-d pet imaging*, IEEE Transactions on Medical Imaging, 29 (2010), pp. 1442–1454.
[29] C. S. Levin, M. Dahlbom, and E. J. Hoffman, *A monte carlo correction for the effect of compton scattering in 3-d pet brain imaging*, IEEE Transactions on Nuclear Science, 42 (1995), pp. 1181–1185.

[30] Y. Li, *Noise propagation for iterative penalized-likelihood image reconstruction based of fisher information*, Phys. Med. Biol., 56 (2011), p. 1083.

[31] J. S. Liu, *The fraction of missing information and convergence rate for data augmentation*, Computing Science and Statistics, (1994), pp. 490–497.

[32] J. S. Liu, W. H. Wong, and A. Kong, *Covariance structure of the gibbs sampler with applications to the comparisons of estimators and augmentation schemes*, Biometrika, 81 (1994), pp. 27–40.

[33] A. Y. Lo, *Bayesian nonparametric statistical inference for poisson point processes*, Zeitschrift fur Wahrscheinlichkeitsteorie und verwandte Gebiete, 59 (1982), pp. 55–66.

[34] A. Luna, J. C. Vilanova, L. C. Hygino da Cruz Jr, and S. E. Rossi, *Functional imaging in oncology: biophysical basis and technical approaches - Vol. 1*, Springer Science & Business Media, 2013.

[35] S. Lyddon, S. Walker, and C. Holmes, *Nonparametric learning from bayesian models with randomized objective functions*, Advances in Neural Information Processing Systems, (2018).

[36] L. G. Marcu, L. Moghaddasi, and E. Bezak, *Imaging of tumor characteristics and molecular pathways with pet: developments over the last decade toward personalized cancer therapy*, International Journal of Radiation Oncology Biology Physics, 102 (2018), pp. 1165–1182.

[37] F. Natterer, *The mathematics of computerized tomography*, Society for Industrial and Applied Mathematics, 2001.

[38] M. A. Newton and A. E. Raftery, *Approximate bayesian inference with the weighted likelihood bootstrap*, Journal of the Royal Statistical Society: Series B (Methodology), 56 (1994), pp. 3–26.

[39] T. L. Ng and M. A. Newton, *Random weighting in LASSO regression*, Electronic Journal of Statistics, 16 (2022), pp. 3430 – 3481.

[40] L. Nie and V. Ročková, *Bayesian bootstrap spike-and-slab lasso*, Journal of the American Statistical Association, (2022), pp. 1–16.

[41] O. Papaspiliopoulos, G. O. Roberts, and M. Sköld, *A General Framework for the Parametrization of Hierarchical Models*, Statistical Science, 22 (2007), pp. 59 – 73.

[42] E. Pompe, *Introducing prior information in weighted likelihood bootstrap with applications to model misspecification*, arXiv preprint arXiv:2103.14445, (2021).

[43] A. Rahmim, J. Tang, and H. Zaidi, *Four-dimensional (4d) image reconstruction strategies in dynamic pet: Beyond conventional independent frame reconstruction*, Medical Physics, 36 (2009), pp. 3654–3670.

[44] L. A. Shepp and Y. Vardi, *Maximum likelihood reconstruction for emission tomography*, IEEE transactions on medical imaging, 1 (1982), pp. 113–122.

[45] R. Siddon, *Fast calculation of the exact radiological path for a three-dimensional ct array.*, Medical physics, 12 2 (1985), pp. 252–5.

[46] A. Sitek, *Data analysis in emission tomography using emission count posteriors*, Physics in Medicine & Biology, 52 (2012), p. 6779.

[47] A. Sitek and M. A. Celler, *Limitations of poisson statistics in describing radioactive decay*, Physica Medica, 31 (2015), pp. 1105–1107.

[48] S. Stute and C. Comtat, *Practical considerations for image-based psf and blobs reconstruction in pet*, Physics in Medicine & Biology, 58 (2013), p. 3849.

[49] A. W. Van der Vaart, *Asymptotic statistics*, vol. 3, Cambridge university press, 2000.

[50] D. A. Van Dyk and X.-L. Meng, *The art of data augmentation*, Journal of Computational and Graphical Statistics, 10 (2001), pp. 1–50.

[51] F. Vasconcelos, B. He, N. Singh, and Y. W. Teh, *Uncertainr: Uncertainty quantification of end-to-end implicit neural representations for computed tomography*, arXiv preprint arXiv:2202.10847v2, (2022).

[52] K. Vunckx, A. Atre, K. Baete, A. Reilhac, C. M. Deroose, K. Van Laere, and J. Nuyts, *Evaluation of three mri-based anatomical priors for quantitative pet brain imaging*, IEEE transactions on medical imaging, 31 (2011), pp. 599–612.

[53] W. A. Weber, *Use of pet for monitoring cancer therapy and for predicting outcome.*, Journal of Nuclear Medicine, 46 (2005), pp. 983–995.

[54] I. S. Weir, *Fully bayesian reconstructions from single-photon emission computed tomography data*, Journal of the American Statistical Association, 92 (1997), pp. 49–60.