Sparsity promoting reconstructions via hierarchical prior models in diffuse optical tomography

Anssi Manninen[1], Meghdoot Mozumder[2], Tanja Tarvainen[2,3], Andreas Hauptmann[1,3]

[1]Research Unit of Mathematical Sciences, University of Oulu, Oulu 90014, Finland

[2]Department of Applied Physics, University of Eastern Finland, Kuopio 70211, Finland

[3]Department of Computer Science, University College London, London WC1E 6BT, United Kingdom

September 20, 2022

Abstract

Diffuse optical tomography (DOT) is a severely ill-posed nonlinear inverse problem that seeks to estimate optical parameters from boundary measurements. In the Bayesian framework, the ill-posedness is diminished by incorporating *a priori* information of the optical parameters via the prior distribution. In case the target is sparse or sharp-edged, the common choice as the prior model are non-differentiable total variation and ℓ^1 priors. Alternatively, one can hierarchically extend the variances of a Gaussian prior to obtain differentiable sparsity promoting priors. By doing this, the variances are treated as unknowns allowing the estimation to locate the discontinuities.

In this work, we formulate hierarchical prior models for the nonlinear DOT inverse problem using exponential, standard gamma and inverse-gamma hyperpriors. Depending on the hyperprior and the hyperparameters, the hierarchical models promote different levels of sparsity and smoothness. To compute the MAP estimates, the previously proposed alternating algorithm is adapted to work with the nonlinear model. We then propose an approach based on the cumulative distribution function of the hyperpriors to select the hyperparameters. We evaluate the performance of the hyperpriors with numerical simulations and show that the hierarchical models can improve the localization, contrast and edge sharpness of the reconstructions.

1 Introduction

Diffuse optical tomography (DOT) is a highly ill-posed nonlinear problem which utilizes boundary measurements of near-infrared light to estimate spatially distributed absorption and reduced scattering parameters in biological tissue [2, 18]. Due to the highly ill-posed nature of DOT, advanced inversion techniques are required to acquire feasible estimates [31, 32, 45]. This can be realized in a variational approach, where a regularizer based on certain assumptions is used, or through a Bayesian approach. The regularizer can, for example, utilize assumptions on smoothness or sparsity of the solution [4, 39], or sparsity of its derivative, i.e., total variation (TV) [34] to obtain stable inversion. On the other hand, Bayesian estimation utilizes prior probability distributions of the unknowns, based on previously available knowledge, to compute the posterior probability distribution as a

solution to the inverse problem [26]. Typically the solution is computed as a single point estimate such as the *maximum a posteriori* (MAP) estimate. For nonlinear problems such as the DOT, the MAP estimate lacks analytical form, therefore requiring the use of iterative methods.

In Bayesian inversion, choosing a plausible prior distribution has a crucial role in mitigating the ill-posedness and determining the type of the prior information. A popular choice as prior is the Gaussian distribution, which encapsulates prior information on the mean, variance and correlations of the unknown parameters. Additionally, the Gaussian prior provides a closed-form for the posterior. In order for a Gaussian prior to provide the optimal information of a target with discontinuities, i.e., sudden jumps or sharp edges, one would need to know the locations of the discontinuities. Then, setting large variances for these locations and diminishing other variances would enable the Gaussian prior to favour discontinuities in correct locations. Evidently, knowing the locations beforehand is not the case for tomographic problems where the premise is to locate the changes in the target.

The remedy to avoid lack of prior information and incorrect prior parameters such as the variances (of uncorrelated unknowns) is to include an uncertainty at a hierarchically higher level. By assuming the uncertain variances as unknowns, we can express the uncertainty with the hyperprior distributions. Most of the studied hierarchical models are built around the Gaussian priors, which we also investigate in this work.

Two popular Gaussian priors used for the hierarchical models in tomographic inverse problems are the uncorrelated Gaussian prior [14], that is, a multivariate Gaussian distribution without correlations and the structural difference prior [11]. For both priors, the hierarchical models are commonly formed by assuming the variances (for difference prior also referred to as weights) to be the unknown parameters that follow the chosen hyperprior, as in [12, 14]. Alternatively, the other parameters of Gaussian priors could also be hierarchically extended, i.e., the mean [20], characteristic correlation length of Gaussian smoothness prior [36], or all entries of the covariance [1]. In this work, we limit the investigation of hierarchical models on the uncorrelated Gaussian prior and structural difference prior due to their desirable sparsity and discontinuity promoting properties as described later.

For the two considered Gaussian priors, several different types of hyperpriors have been utilized, such as the uniform [10], exponential [11], standard gamma [14] and inverse-gamma [12] distributions. The exponential and uniform distributions can be considered non-informative hyperpriors, not incorporating (almost) any assumption on the unknowns. Whereas the standard and inverse-gamma hyperpriors can be harnessed to promote sparsity or sharp-edges [8, 12]. As it turns out, the MAP estimates with the standard and inverse-gamma hyperpriors are related to sparsity regularization, i.e., a TV or ^{1} regularized problem. Conveniently, using the hyperprior approach, regularized solutions that are conventionally computed from non-smooth problems can now be computed approximately from a differentiable problem [14]. The differentiability arises from the fact that the MAP estimation problems from the hierarchical models lie in the differentiable ^{2} regularized framework.

In tomographic inverse problems the hierarchical models have been used for instance, in electrical impedance tomography (EIT) [5,28], magnetoencephalography (MEG) [7,29,33] and linearized DOT [20, 30, 40, 41, 44]. A lot of the previous work concentrates on using either the structural difference prior or the uncorrelated Gaussian prior, combined with the standard gamma hyperpriors for the variances aiming to enhance the spatial accuracy of the estimates [6, 33, 40]. The majority of the previously studied hierarchical models applied in

tomographic problems have focused on using linear or linearized forward models, such as the Rytov approximation to linearize the forward model in DOT, see, e.g., [40, 41, 44]. Although some work with nonlinear forward models can be found. For instance, in [36], the authors demonstrated the use of hierarchical Whittle-Maté priors or in [23] the mixture Gaussian prior model with hyperpriors was applied in a nonlinear DOT problem.

In this work, we study the effect of the exponential, standard gamma, and inverse-gamma hyperprior hierarchical models on the nonlinear two-dimensional (2D) frequency-domain DOT problem. The hierarchical models are built around the uncorrelated Gaussian prior and difference prior, for which we assume the unknown variances follow hyperpriors. The hierarchical models are evaluated with simulated piecewise linear targets that would require finding the discontinuities for optimal results, which we try to obtain via the unknown variances. The role of the hyperparameters is discussed, and a simple selection method based on the cumulative distribution function of the hyperpriors is considered. This is in contrast to previous work [8, 14], where a rigorous automatized hyperparameter selection method was proposed, but only for problems with a linear model. We study the MAP estimates with different hyperparameters, which are computed from the hierarchical models by adapting the previously proposed iterative alternating sequence (IAS) algorithm [8, 12] that is modified here to work with the nonlinear forward operators. Additionally, we give new empirical insight into the convergence of the nonlinear IAS algorithm.

The remaining sections of the paper are organized as follows. In Section 2, we formulate the inverse problem in the Bayesian regime and introduce the used priors and hyperpriors. Then we formulate the nonlinear IAS algorithm for each of the hierarchical models. In Section 3, we formulate the 2D DOT inverse problem. In Section 4, numerical implementations are described. In Section 5, we provide the numerical results of the hierarchical models from the simulated DOT problems and discuss the empirical convergence of the nonlinear IAS. In Section 6, conclusions are given.

2 Hierarchical Bayesian models with nonlinear forward model

Let us consider a discrete observation model of the form

$$y = A(x) + e, \tag{1}$$

where $y \in \mathbb{R}^m$ is a discrete set of measurements, $x \in \mathbb{R}^n$ are the parameters to be estimated, $A : \mathbb{R}^n \to \mathbb{R}^m$ is a nonlinear forward operator mapping the parameters x to the data space, and $e \in \mathbb{R}^m$ is the additive measurement noise. The inverse problem is then to recover the set of unknown parameters x from the measured noisy data y.

In the Bayesian approach to inverse problems, all the parameters are considered as random variables, and the uncertainties of their values are encoded into probability distribution models [26]. The solution to the inverse problem is the posterior probability distribution which is obtained through Bayes' theorem and can be written as

$$\pi(x|y) = \frac{\pi(y|x)\pi_{\text{prior}}(x)}{\pi(y)} \propto \pi(y|x)\pi_{\text{prior}}(x), \tag{2}$$

where $\pi(y|x)$ is the likelihood density and $\pi(x)$ is the prior density. Since the measurements are known, we will neglect the normalization factor $\pi(y)$. As diffuse tomographic

problems such as DOT and EIT appear as severely ill-posed, choosing appropriate prior distribution plays a principal part in overcoming the ill-posedness.

For inverse problems with expensive forward operators as well as high dimensional problems, computing the entire posterior is computationally infeasible. Therefore it is common to compute a point estimate from the posterior. A commonly used point estimate is the MAP estimate, which is computed as

$$x_{\mathrm{MAP}} = \arg\max_x \{\pi(x|y)\}.$$

To formulate the likelihood, assume mutually independent unknowns x and noise e that is Gaussian distributed

$$e \quad N(\mu_e, C_e),$$

where $\mu_e \quad \mathsf{R}^m$ is the mean, and $C_e \quad \mathsf{R}^{m\times m}$ is the covariance matrix. The likelihood is given by [26]

$$\pi(y|x) = \pi_e(y - A(x)) = \frac{1}{\sqrt{(2\pi)^m |C_e|}} \exp\left(-\frac{1}{2} \left\| L_e(y - A(x) - \mu_e) \right\|_2^2\right), \tag{3}$$

where $|C_e|$ is the determinant of C_e, and L_e is given through the Cholesky decomposition $C_e^{-1} = L_e^{\mathrm{T}} L_e$. To simplify the notations we assume here origin centered noise ($\mu_e = 0$). The MAP estimate is now the minimizer of the functional

$$F(x) = \frac{1}{2} \left\| L_e(y - A(x)) \right\|_2^2 - \log(\pi_{\mathrm{prior}}(x)). \tag{4}$$

This problem resembles a variational approach to regularization, where the regularization term is the negative logarithm of the prior density [13].

For the choice of prior in Eq. (2) we consider two different Gaussian priors which we later extend hierarchically. These prior types have already been established to produce computationally feasible MAP estimates with the hierarchical posterior model with linear forward operators [8, 12].

Let us first consider a situation where the entries x_j of the unknown vector x are mutually independent and Gaussian distributed,

$$x_j \quad N\left(\mu, \sqrt{\theta_j}\right),$$

where $\mu \quad \mathsf{R}$ is the mean and $\theta_j \quad \mathsf{R}_+$ the variance. Note, that all of the unknowns x_j are assumed to have same mean. Then an uncorrelated Gaussian prior is given by

$$\pi_{\mathrm{prior}}(x) = \frac{1}{(2\pi)^{n/2} \sum_{j=1}^n \sqrt{\theta_j}} \exp\left(-\frac{1}{2} \sum_{j=1}^n \frac{(x_i - \mu)^2}{\theta_i}\right). \tag{5}$$

Alternatively, we introduce also a difference prior of form

$$\pi_{\mathrm{prior}}(x) = C(\theta) \exp\left(-\frac{1}{2} \sum_{t\ D} \frac{d_t}{\theta_t}^2\right) = C(\theta) \exp\left(-\frac{1}{2} \left\| \Lambda^{1/2} B x \right\|_2^2\right), \tag{6}$$

where C is the normalization factor depending on the variances, $D \quad \mathsf{R}^q$ is a set that contains q indice pairs (i, j) determining the differences $d_{(i,j)} = x_i - x_j$, Λ is a diagonal matrix $\Lambda = \mathrm{diag}(d_{(1,0)}, \ldots, d_{(n,n-1)})$ where x_0 is defined as zero and $B \quad \mathsf{R}^{q\times n}$ is the difference

matrix defining the differences d between the adjancent unknowns, such that, $d = Bx$, $d = (d_1, \ldots, d_q)$.

For the uncorrelated Gaussian prior, the MAP estimate is the minimizer of the functional

$$F(x) = \frac{1}{2} \left\| L_e(y - A(x)) \right\|_2^2 + \frac{1}{2} \sum_{j=1}^{n} \frac{(x_j - \mu)^2}{\theta_j}. \tag{7}$$

Correspondingly, for the difference prior, the MAP estimate is obtained as a minimizer of

$$F(x) = \frac{1}{2} \left\| L_e(y - A(x)) \right\|_2^2 + \frac{1}{2} \sum_{t \in D} \frac{d_j^2}{\theta_t}. \tag{8}$$

Note, that for fixed variances θ_j the normalization terms of the uncorrelated Gaussian prior (5) and difference prior (6) can be omitted from the minimizations (7) and (8). As it can be seen, both prior models yield a quadratic penalty term, the first (7) for the unknowns x and the latter (8) for the differences d. The quadratic penalty tends to smooth sudden jumps, that is, the outliers or the edges of a piecewise linear target. For imaging modalities, this smoothness on x can appear as blurriness and loss of contrast in the images, possibly making the quality of the image undesired.

2.1 Extending the priors hierarchically

To overcome the incomplete knowledge of the discontinuity locations, we assume the variances as random variables and aim to estimate these locations. We extend Bayes' theorem in Eq. (2), by introducing a conditional probability on the variances

$$\pi(x|y, \theta) \quad \pi(y|x)\pi_{\text{prior}}(x|\theta)\pi_{\text{hyper}}(\theta),$$

where $\pi_{\text{hyper}}(\theta)$ is now a hyperprior of the variances. Therefore the MAP estimate of the hierarchically extended posterior is given by

$$(x_{\text{MAP}}, \theta_{\text{MAP}}) = \arg\max_{x,\theta} \{\pi(x|y, \theta)\}. \tag{9}$$

Choosing a hyperprior for the variances allows us to express our assumption on how rare or pronounced the discontinuities are. On the other hand, assuming a uniform hyperprior provides no information on how many or extreme the discontinuities are.

2.1.1 Exponential hyperprior

In this work, we only consider the case of independent variances leading to independent hyperpriors. Letting the variances be mutually independent means that the discontinuities are assumed to be sudden. For the uncorrelated Gaussian prior (5) the suddenness means promoting point-like outliers. Whereas the difference prior (6) promotes solutions with sudden jumps, i.e., sharp-edges.

To begin with, we describe a non-informative hyperprior for cases where we do not want to limit the amount of large variances, that is, the amount of outliers. Let us consider the exponential hyperprior for variance θ_j

$$\pi_{\text{hyper}}(\theta_j) \quad \exp\left(-\frac{1}{2}\frac{\gamma}{\theta_j}\right), \tag{10}$$

where $\gamma > 0$ is a hyperparameter called the rate parameter. If variances θ_j of the uncorrelated Gaussian prior (5) follow the exponential hyperprior with the same hyperparameter γ, computing the MAP estimate is equal to minimizing a functional

$$F(x, \theta) = \frac{1}{2} \left\| L_e(y - A(x)) \right\|_2^2 + \frac{1}{2} \sum_{j=1}^{n} \frac{(x_j - \mu)^2}{\theta_j} + \frac{\gamma}{2} \sum_{j=1}^{n} \frac{1}{\theta_j} - \frac{1}{2} \sum_{j=1}^{n} \log \theta_j. \tag{11}$$

Notice, that the last term of Eq. (11) emerges from the normalization factor of the prior (5) which cannot be omitted due to variances θ being part of the estimation. Assuming the difference prior (6) and variances with exponential hyperprior leads minimizing

$$F(x, \theta) = \frac{1}{2} \left\| L_e(y - A(x)) \right\|_2^2 + \frac{1}{2} \sum_{t \in D} \frac{d_t^2}{\theta_t} + \frac{\gamma}{2} \sum_{t \in D} \frac{1}{\theta_t} - \frac{1}{2} \sum_{t \in D} \log \theta_t. \tag{12}$$

While increasing the hyperparameter γ, the exponential hyperprior (10) starts favouring large variances over small ones, hence allowing more outliers or sharp edges to occur. On the other hand, when $\gamma \to 0$, the exponential hyperprior approaches a uniform distribution. Thus, with small γ we can get a behaviour close to a uniform hyperprior and favour small and large variances almost equally. Note, that if $\gamma = 0$, the hyperprior is uniform and the minimizer of (11) corresponds to the mean $x = \mu$ (or $d = 0$ for (12)). This can be observed from (11) and (12), by setting $x_j = \mu$ (or $d_j = 0$) for all j and taking limit $\theta \to 0$, which has no lower bound. As described later in Section 2.2, the beneficial aspect of using the exponential hyperprior over the uniform hyperprior is its inherent way of incorporating the (logarithmic) positivity constraint for the variances.

2.1.2 Standard gamma and inverse-gamma hyperpriors

The exponential hyperprior is easy to apply due to it having only a single hyperparameter. If the target is sparse, a possible drawback of using a non-informative hyperprior, such as the exponential (with small γ), is the tendency to promote many outliers leading to noisy background. In order to promote sparse solutions, we need to consider hyperprior densities decreasing towards the large variances, hence favouring less outliers. As discussed in [12], if we wish to have only few prominent outliers, the established hyperpriors are the standard gamma distribution

$$\pi_{\text{hyper}}(\theta_j | \beta, \vartheta) \quad \theta_j^{\beta - 1} \exp\left(- \frac{\theta_j}{\vartheta} \right), \tag{13}$$

and the inverse-gamma distribution

$$\pi_{\text{hyper}}(\theta_j | \beta, \vartheta) \quad \theta_j^{-\beta - 1} \exp\left(- \frac{\vartheta}{\theta_j} \right), \tag{14}$$

where $\beta > 0$ and $\vartheta > 0$ are hyperparameters called shape and scaling parameter, respectively. As demonstrated in [12], the samples drawn from the standard gamma distribution are more likely to be outliers than samples drawn from the inverse-gamma distribution. On the other hand, the outliers drawn from the standard gamma distribution are less extreme compared to the inverse-gamma distribution. This indicates that the inverse-gamma hyperprior is ideal for promoting sparse solutions with pronounced outliers. In contrast, the standard gamma hyperprior works best with less sparse solutions with only moderate outliers.

In order to understand the effect of these hyperpriors, let us consider the uncorrelated Gaussian prior (5), with unknown variances θ_j following standard gamma hyperprior (13)

$$\theta_j \quad \text{Gamma}\,(\beta, \vartheta_j)\,.$$

The MAP estimate is then the minimizer of a functional

$$F(x,\theta) = \frac{1}{2}\ L_e(y - A(x))\ _2^2 + \frac{1}{2}\sum_{j=1}^{n}\frac{(x_j - \mu)^2}{\theta_j} + \sum_{j=1}^{n}\left[\frac{\theta_j}{\vartheta_j} - \eta\log\left(\frac{\theta_j}{\vartheta_j}\right)\right], \qquad (15)$$

where $\eta = \beta - \frac{3}{2}$. In the previous work [8], functional (15) was reformulated solely in terms of x, i.e., by finding the function f so that $\theta = f(x)$. Further, it was shown in [7] that for fixed x and $\eta > 0$ the functional has the limit

$$\lim_{\eta\to 0} F(x, f(x)) = \frac{1}{2}\ L_e(y - A(x))\ _2^2 + \sqrt{2}\sum_{j=1}^{n}\frac{|x_j - \mu|}{\sqrt{\vartheta_j}}, \qquad (16)$$

that is the ^{1} constraint on x. In other words, by using small η, the MAP estimate approximately yields the solution of the ^{1} constrained problem. The hyperparameters ϑ_j weight the penalty term, similar to the sensitivity-weighting, used to compensate sensitivity differences of the forward model (see e.g. [27]). Computing the non-smooth ^{1} constrained solution directly from the right hand side of (16) requires computationally expensive methods (see e.g. [42]). The benefit of acquiring an approximation of the ^{1} constrained solution via the functional (15) is that now the term emerging from the prior and the hyperprior (the last sum of Eq. (15)) is differentiable.

In case the variances of the uncorrelated Gaussian prior (5) follow the inverse-gamma hyperprior (14),

$$\theta_j \quad \text{InvGamma}\,(\beta, \vartheta_j)\,,$$

then the minimized functional becomes

$$F(x,\theta) = \frac{1}{2}\ L_e(y - A(x))\ _2^2 + \frac{1}{2}\sum_{j=1}^{n}\frac{(x_j - \mu)^2}{\theta_j} + \sum_{j=1}^{n}\left[\frac{\vartheta_j}{\theta_j} + (\beta + \frac{3}{2})\log\left(\frac{\theta_j}{\vartheta_j}\right)\right]. \quad (17)$$

Similarly as with Eq. (15), we can write the functional (17) merely in terms of x [8]

$$F(x, f(x)) = \frac{1}{2}\ L_e(y - A(x))\ _2^2 + \sum_{j=1}^{n}\log\left(\frac{(x_j - \mu)^2}{\vartheta_j} + 2\right)^{\beta+\frac{3}{2}}. \qquad (18)$$

As pointed out in [8], when $\vartheta_j = 1$ for all j and $\beta \to 0$, the logarithmic penalty term in (18) approaches a penalty equivalent to one produced by n invidual student distributions

$$\text{Student}(x_j \mid \nu) \quad \frac{1}{\left(1 + \frac{(x_j-\mu)^2}{\nu}\right)^{(\nu+1)/2}},$$

when $\nu = 2$. That is, a distribution favouring outliers. We note that, while using a linear forward operator A and assuming the minimizer to be unique, the MAP minimization problem (15) with the standard hyperpriors was shown to be globally convex [7]. Similarly, by observing the positive definiteness of the Hessian, the MAP minimization problem (11) with the exponential hyperprior could also be shown to be strictly convex. Whereas, with

the inverse-gamma hyperprior, the minimization problem (17) is only locally convex, where the convexity radius depends on the hyperparameter β [8].

With the difference prior (6), the standard gamma and inverse-gamma hyperpriors have similar effect as in Eqs. (16) and (18), but now in terms of the differences d. If the variances of the difference prior follow the standard gamma hyperpriors, the MAP estimate has a limit

$$\lim_{\eta \to 0} F(x, f(x)) = \frac{1}{2} \left\| L_e(y - A(x)) \right\|_2^2 + \sqrt{2} \sum_{t \in D} \frac{|d_t|}{\sqrt{\vartheta_t}}. \tag{19}$$

Therefore, for small η the difference prior with variances following standard gamma hyperpriors yields a penalty term equivalent to the weighted TV penalty. Correspondingly to Eq. (16), for the inverse-gamma hyperpriors, the minimized functional is [12]

$$F(x, f(x)) = \frac{1}{2} \left\| L_e(y - A(x)) \right\|_2^2 + \sum_{t \in D} \log \left(\frac{d_t^2}{\vartheta_t} + 2 \right)^{\beta + \frac{3}{2}}. \tag{20}$$

The penalty term in Eq. (20) is similar to the Perona-Malik functional, used for edge weighted diffusion in image processing [35]. By applying the standard and the inverse-gamma hyperpriors with the difference prior, we get two slightly different optimization problems, trying to establish the edges via the variances.

2.2 Iterative alternating sequence

For the efficient computation of the MAP estimates from models with many additional unknowns, such as the variances, it is common to consider alternating algorithms. The alternating algorithms tackle the minimization problem by alternatingly updating the actual unknowns and the model (prior) parameters. The framework of alternating algorithms for hierarchical models in tomography problems is well established, with various different variations (see e.g [1, 21, 23, 28, 46]). In this paper, we extend the iterative alternating algorithm (IAS) described in [11, 14], to minimize the functionals (11),(15) and (17), as well as their corresponding form for the difference prior, with a nonlinear forward operator. To our knowledge, this is the first time this algorithm has been modified for a nonlinear forward model.

First, consider the uncorrelated Gaussian prior (5) and inspect the corresponding energy functional in two parts as

$$F(x, \theta) = \left(\overbrace{\frac{1}{2} \left\| L_e(y - A(x)) \right\|_2^2}^{\text{a)}} + \underbrace{\frac{1}{2} \sum_{j=1}^{n} \frac{(x_j - \mu)^2}{\theta_j} + \frac{1}{2} \sum_{j=1}^{n} \log \theta_j + \sum_{j=1}^{n} g_j(\theta_j)}_{\text{b)}} \right), \tag{21}$$

where the functions g_j depend on the selected hyperprior type. Now the a) part contains the terms depending on x and the b) part depends on the variances θ. For fixed θ, minimizing Eq. (21) corresponds to minimizing the part a)

$$\hat{x} = \arg\min_x F(x, \theta) = \arg\min_x \frac{1}{2} \left\| L_e(y - A(x)) \right\|_2^2 + \frac{1}{2} \sum_{j=1}^{n} \frac{(x_j - \mu)^2}{\theta_j}. \tag{22}$$

Correspondingly, for fixed x we only need to minimize the part b). The IAS algorithm for nonlinear forward operators is outlined in Algorithm 1. On each iteration of the IAS, the

Algorithm 1: Generic iterative alternating sequential (Generic IAS)

Determine hyperparameters
Set initial (x^0, θ^0) and $t = 0$
while solution (x^t, θ^t) not converged
  For fixed θ^t compute minimizer $\hat{x}$ of (21) and set $x^{t+1} = \hat{x}$
  For fixed x^{t+1} compute minimizer $\widehat{\theta}$ of (21) and set $\theta^{t+1} = \widehat{\theta}$
  $t = t + 1$

variances are updated based on the results of the previous iteration. This idea is closely related to the bootstrap priors, where the prior is updated based on the previous recontructions [9].

For the considered hyperpriors, the IAS is efficient since by applying it with the exponential, standard gamma or inverse-gamma hyperpriors, there exist analytical minimizer with respect to θ. For instance, consider the variances following standard gamma hyperpriors. Then, for fixed x, minimizing functional (15) equals minimizing b)

$$\widehat{\theta} = \arg\min_{\theta} F(x, \theta) = \arg\min_{\theta} \frac{1}{2} \sum_{j=1}^{n} \frac{(x_j - \mu)^2}{\theta_j} + \sum_{j=1}^{n} \left[\frac{\theta_j}{\vartheta_j} - \alpha \log \left(\frac{\theta_j}{\vartheta_j} \right) \right] \tag{23}$$

which has the analytical minimizer [12]

$$\widehat{\theta}_j = \vartheta_j \left(\frac{\eta}{2} + \sqrt{\frac{\eta^2}{4} + \frac{(x_j - \mu)^2}{2\vartheta_j}} \right). \tag{24}$$

Correspondingly, for the inverse-gamma hyperpriors, with fixed x the functional (18) has the minimizer

$$\widehat{\theta}_j = \frac{1}{\beta + 3/2} \left(\vartheta_j + \frac{1}{2}(x_j - \mu)^2 \right). \tag{25}$$

Moreover, the functional (11) with exponential hyperpriors also has an analytical minimizer with respect to the variances, that is [11]

$$\widehat{\theta}_j = (x_j - \mu)^2 + \gamma. \tag{26}$$

As it can be seen from equations (24)-(26), for each hyperprior type, the hyperparameters determine (positive) lower bounds for the variances. For the standard gamma hyperprior (24) the shape parameter η provides a small relaxation parameter so that the variances will not vanish, and ϑ_j controls how small the relaxation parameter is and how much the term $(x - \mu)^2$ is amplified. Whereas, for the inverse gamma hyperprior (25), the scale parameter ϑ_j provides a relaxation parameter, and controlling β can be used to amplify or dampen the variance values. The exponential hyperprior (26) has only γ hyperparameter, which acts as the relaxation parameter while updating the variances.

In order to apply the IAS algorithm to the MAP estimation problem with the difference prior (8), we would first need to determine the form of the normalization factor C. Unfortunately, the closed-form of the normalization factor remains unknown in terms of x. Therefore, one option is to use Bayesian variational methods to approximate the posterior distribution, which is already established for the hierarchical models [40, 41, 44]. Alternatively, as described in [11], to avoid cumbersome estimation of the normalization factor,

we reformulate the minimization problem in terms of the differences as

$$F(d,\theta) = \left(\overbrace{\frac{1}{2} \; L_e(y - A(Pd)) \; _2^2 + \frac{1}{2}\sum_{j=1}^{n} \frac{d^2}{\theta_j}}^{\text{a)}} + \underbrace{\frac{1}{2}\sum_{j=1}^{n} \log\theta_j + \sum_{j=1}^{n} g_j(\theta_j)}_{\text{b)}} \right), \tag{27}$$

where $P \quad \mathrm{R}^{q\times q}$ is a matrix such that $x = Pd$. Then we can use Algorithm 1 directly to minimize (27). Due to fact that the difference matrix B in (6) is not invertible the matrix P cannot be formulated directly from the relation $d = Bx$. Instead, the inverse mapping must be inferred from the fact that summing differences d_i of any closed loop in the mesh needs to be zero. This constraint can be expressed in a matrix form as

$$Md + \quad = 0, \tag{28}$$

where $M \quad \mathrm{R}^{p\times q}$ is a matrix representing each loop constraint in its rows, p is the number of loops around the mesh elements, and $\quad > 0$ is a small constant allowing us to compute approximated solutions of $Md = 0$. The detailed formulation of the constraint (28) is shown in [11]. The constraint (28) can be simultaneously solved with the part a) of (27) to make the differences d satisfy the constraint. Compared to the hierarchical model with white noise prior, minimizing part $a)$ is now drastically more expensive due to a high number of unknown differences. For instance, a triangular mesh has $3(n-1)-b$ differences, where b denotes the number of nodes on the hull.

2.3 Selecting the hyperparameters

Selecting the hyperparameters is especially important for the more informative hyperpriors, which aim to favor certain solution types such as the sparsity promoting standard and inverse-gamma hyperpriors. We will focus on the selection of the hyperparameter ϑ due to its easily interpreted role on controlling the significance of the constraints in (15) and (18). Let us first review a hyperparameter ϑ selection method for models with a linear forward operator as described in [14].

2.3.1 Automatizing the scaling hyperparameters for linear forward models

In [14] it was shown that for a linear forward operator $A(x) = Ax$ satisfying the exhangeability condition, the hyperparameters ϑ_j should be set as

$$\vartheta_j = \frac{D}{Ae_j \; _2^2}, \tag{29}$$

where term D depends on the estimated signal-to-noise ratio, hyperparameter β, and assumed sparsity of the x (see [14] for details).

As discussed in [14], the idea of setting the hyperparameters ϑ according to the forward model is closely related to the idea of sensitivity-weighting used in the regularization scheme. In sensitivity-weighting the regularizing norm such as the ^{1} regularizer is scaled for each component x_j with respect to the sensitivity

$$\left\| \frac{\partial(Ax)}{\partial x_j} \right\|_2 = \; Ae_j \; _2,$$

where e_j is the jth canonical basis vector. Similarly, in the automatized selection (29), the sensitivity term $\|Ae_j\|_2$ appears. The sensitivity-weighting is established to improve the spatial quality of recontructions in applications, such as MEG [6], where the sensitivity differences of the linear forward model are prominent. On the other hand, if we choose the same value for all θ_j, no spatial preferences are incorporated among the unknowns x.

2.3.2 Choosing the hyperparameters based on the confidence interval

For a nonlinear model, we cannot deduce similar selection method utilizing the sensitivity of the model (i.e. $\|Ae_j\|$), since the sensitivity of $A(x)$ varies for different x. In practical applications it is common to know coarse upper bound for the unknown physical quantities. Therefore, we can incorporate this information via the confidence intervals of the probability distributions. For instance, let us assume that $|x_j - \mu| < M$, $M > 0$ with 95% probability. Then the variance θ_j is set as $U = (M/2)^2$.

On the other hand, if θ_j follows the hyperprior, we have 5% probability for x_j being an outlier, for which θ_j should be larger than $(M/2)^2$. Thereby we can determine ϑ_j from the cumulative distribution function (CDF) of the hyperprior, i.e., we choose ϑ_j s.t.

$$\mathrm{CDF}(U, \vartheta_j) = 0.95. \tag{30}$$

For the exponential hyperprior, the CDF is not bounded and thus cannot be used to determine γ.

3 Diffuse optical tomography

Diffuse optical tomography (DOT) is a technique for imaging spatially varying optical parameters in biological tissue [2, 17, 18]. The distribution of these optical parameters provides tissue biochemical and structural information with applications, for example, in early-diagnosis and imaging of breast cancer [19], monitoring neonatal brain health [22], functional brain imaging of adults [43], and preclinical imaging of small animals [15]. In general, DOT is non-ionizing and non-invasive, and its instrumentation is relatively simple, low-cost, and portable compared to conventional medical tomographic techniques [43]. For more information on image reconstruction problem of DOT and various methodologies, see e.g. [3, 16, 17, 24, 39] and the references therein.

3.1 Forward model

In a typical DOT measurement setup, near-infrared light is introduced into an object from its boundary. Let $\Omega \subset \mathbb{R}^d$, $(d = 2 \text{ or } 3)$ denote the domain with boundary $\partial\Omega$ where d is the (spatial) dimension of the domain. In a diffusive medium, like soft biological tissue, the commonly used light transport model for DOT is the diffusion approximation to the radiative transfer equation [25]. Here, we consider the frequency-domain version of the diffusion approximation [2]

$$\left(-\nabla \cdot \frac{1}{d(\mu_\mathrm{a}(r) + \mu_\mathrm{s}(r))} \nabla + \mu_\mathrm{a}(r) + \frac{\mathrm{j}\omega}{c} \right) \Phi(r) = 0, \quad r \in \Omega, \tag{31}$$

$$\Phi(r) + \frac{1}{2\zeta} \frac{1}{d(\mu_\mathrm{a}(r) + \mu_\mathrm{s}(r))} \alpha \frac{\partial \Phi(r)}{\partial \hat{n}} = \begin{cases} \frac{q}{\zeta}, & r \in s \\ 0, & r \in \partial\Omega \setminus s \end{cases}, \tag{32}$$

where $\Phi(r)$ is the photon density, $\mu_{\mathrm{a}}(r)$ is the absorption coefficient, $\mu_{\mathrm{s}}(r)$ is the reduced scattering coefficient, j is the imaginary unit and c is the speed of light in the medium. The parameter q is the strength of the light source at location $s \quad \partial\Omega$, operating at angular modulation frequency ω. Further, the parameter ζ is a dimension-dependent constant ($\zeta = 1/\pi$ when $\Omega \quad \mathrm{R}^2$, $\zeta = 1/2$ when $\Omega \quad \mathrm{R}^3$) and α is a parameter governing the internal reflection at the boundary $\partial\Omega$, and $\hat{n}$ is an outward unit vector normal to the boundary. The measurable data on the boundary of the object, exitance $\Gamma(r)$, is given by

$$\Gamma(r) = -\frac{1}{d(\mu_{\mathrm{a}}(r) + \mu_{\mathrm{s}}(r))}\frac{\partial\Phi(r)}{\partial\hat{n}} = \frac{2\zeta}{\alpha}\Phi(r). \tag{33}$$

The numerical approximation of the forward model (31)-(33) is often based on a finite element (FE) approximation [2]. In this work, we use a finite dimensional approximations in piecewise linear basis for absorption, reduced scattering and fluence as described in [2].

3.2 Inverse problem of DOT

Let us consider observation model (1), where y is the vector of measurable data that typically in frequency domain DOT are logarithm of amplitude and phase delay of exitance, $x = (\mu_{\mathrm{a},1}, \ldots, \mu_{\mathrm{a},n/2}, \mu_{\mathrm{s},1}, \ldots, \mu_{\mathrm{s},n/2}) \quad \mathrm{R}^n$ is vector containing the absorption μ_{a} and reduced scattering μ_{s}, $A(x)$ is the discretised forward operator, i.e. the FE-approximation of the forward model (31)-(32), and e is the additive noise. For this nonlinear observation model, we can compute the MAP estimates of the hierarchical models described in Section 2.1 with the introduced nonlinear IAS algorithm (Algorithm 1).

In order to use the IAS iterations, we first need to minimize part a) of the functional (21). Since the minimizer has no closed-form solution, iterative methods, such as Gauss-Newton method are used to approximate the solution. The Gauss-Newton iterations can be written as

$$x^{i+1} = x^i + s_i\delta x^i \tag{34}$$

with step-length parameter s_i. The update direction δx^i is given by

$$\delta x^i = \left(J_A^{\mathrm{T}}C_e^{-1}J_A^{\mathrm{T}} + C_x^{-1}\right)\left(J_A^{\mathrm{T}}C_e^{-1}(y - A(x)) + C_x^{-1}(x - \mu)\right), \tag{35}$$

where the prior covariance

$$C_x = \begin{pmatrix} C_{\mu_{\mathrm{a}}} & 0 \\ 0 & C_{\mu_{\mathrm{s}}'} \end{pmatrix}$$

contains separate prior covariances for the absorption $C_{\mu_{\mathrm{a}}}$ and reduced scattering $C_{\mu_{\mathrm{s}}'}$. Here Jacobian J_A is the discrete representation of the Fréchet derivative of the nonlinear operator $A(x)$ at the point x^i. For the difference prior model we need to estimate the differences d by substituting $d = Px$. After this the chain rule is used to obtain the Jacobian of $A(Pd)$ with respect to the difference d. The following minimization step (part b) to update the variances is then performed by computing the analytical minimizers, using either (24), (25) or (26) depending on the used type of hyperprior.

4 Simulations

The DOT simulation domain was set to a circle with a radius of 25 mm. The setup consisted of 32 sources and 32 detectors modeled as 2 mm wide surface patches located at

equispaced angular intervals on the boundary. Hence the total number of data was 2048, that is, all source-detector pairs of the logarithm of amplitude and phase. For data simulation, the domain was discretized using 10062 triangular elements and $n_s = 5149$ nodes. The solution of the DA (31)-(32) was numerically approximated using the finite element method via the the Toast++ software [37]. The simulated data was corrupted with additive white noise that was drawn from Gaussian distribution $\pi(e) \quad N(0, C_e)$, where standard deviations ($C_e = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$) of the noise were selected as 0.4% of the maximum value from (complex and real parts) of the noise-free measurement data. For all simulated targets, the signal-to-noise ratio was approximately (in decibels) 45.

In the inverse problem, the FE-approximation of the DA implemented with Toast++ was used to approximate the model for light propagation. The FE-mesh of the photon density contained 9102 triangular elements and $n = 4663$ nodes. For the image space, we used the same FE-mesh. Since a reduced number of nodes was used for the inversion, a small modeling error was included in the simulations.

The MAP estimates of the optical parameters were computed by using the IAS algorithm described in Section 2.2. The inner iterations were performed by the Gauss-Newton method with bisection style line-search (see [38]) for optimal step-length. The stopping criteria for both linear and nonlinear IAS was met when the relative difference of the estimates x was less than ε, i.e.,

$$\frac{x^t - x^{t+1}}{x^t} < \varepsilon, \tag{36}$$

over three consecutive iterations or when the minimized functional F stopped decreasing, i.e.,

$$F(x^t, \theta^t) - F(x^{t+1}, \theta^{t+1}) < 0 \tag{37}$$

over three iterations. For all simulations, we used $\varepsilon = 10^{-5}$. For the Gauss-Newton algorithm, the iterations were set to stop after the difference $F(x^t, \theta^t) - F(x^{t+1}, \theta^{t+1})$ was smaller than 10^{-12} guaranteeing strict convergence of the inner iterations. The initial variances θ_j were set as 0.25^2 and 0.0025^2 for reduced scattering and absoprtion.

In addition to visual inspection, the computed MAP estimates $x_{\mathrm{MAP}} \quad \mathrm{R}^n$ were compared by computing relative errors as

$$\mathrm{RE} = \frac{x_{\mathrm{true}} - x_{\mathrm{MAP}}}{x_{\mathrm{true}}},$$

where $x_{\mathrm{true}} \quad \mathrm{R}^{n_s}$ is the simulated true value. Since different parameter spaces of x_{true} and $x_{\mathrm{MAP}} \quad \mathrm{R}^n$, the relative error is computed after interpolating both values to a basis with equal amount of nodes.

The computations were performed with a laptop computer, equipped with Intel Core i7-11800H @ 2.30GHz processor and Nvidia T1200 Laptop GPU. The FE-solution was computed by utilizing the Toast++ software used via MATLAB (R2021b, Mathworks, Natick, MA).

5 Results and discussion

5.1 Reconstruction utilizing the uncorrelated Gaussian prior model

First, we tested the performance of the uncorrelated Gaussian prior with the standard gamma, inverse-gamma, and exponential hyperpriors. As the test phantom, we used a

Table 1: Used hyperparameter (ϑ and γ) values for the hierarchical uncorrelated Gaussian prior (unc.) and structural difference prior (dif.). The hyperparameter ϑ values were computed from the CDF (30). The hyperparameter values for reduced scattering (scat.) were computed with $M \in \{0.3, 1, 10\}$ (unc.) and $M \in \{1, 5, 10\}$ (dif.) for standard gamma and $M \in \{0.3, 1, 5\}$ (unc.) and $M \in \{0.25, 1, 4\}$ (dif.) for inverse-gamma. For absorption (abs.), the assumed M was 0.01 times the corresponding reduced scattering values.

Hyperprior \ Hyperparam.		Low		Intermediate		High	
		Unc.	Dif.	Unc.	Dif.	Unc.	Dif.
Exponential (γ)	Scat.	10^{-10}	$9 \cdot 10^{-4}$	$2.5{\cdot}10^{-3}$	$3.6{\cdot}10^{-3}$	0.25	0.14
	Abs.	10^{-14}	$9 \cdot 10^{-8}$	$2.5{\cdot}10^{-7}$	$3.6{\cdot}10^{-7}$	$2.5{\cdot}10^{-3}$	$1.4{\cdot}10^{-3}$
Standard gamma (ϑ)	Scat.	$5.8{\cdot}10^{-3}$	$6.4{\cdot}10^{-2}$	$6.4{\cdot}10^{-2}$	1.6	6.4	6.4
	Abs.	$5.8{\cdot}10^{-7}$	$6.4{\cdot}10^{-6}$	$6.4{\cdot}10^{-6}$	$1.6 \cdot 10^{-4}$	$6.4 \cdot 10^{-4}$	$6.4 \cdot 10^{-4}$
Inverse-gamma (ϑ)	Scat.	$4 \cdot 10^{-3}$	$5.5{\cdot}10^{-3}$	$1.1{\cdot}10^{-2}$	$8.8{\cdot}10^{-2}$	0.4	1.4
	Abs.	$4 \cdot 10^{-7}$	$5.5{\cdot}10^{-7}$	$1.1{\cdot}10^{-6}$	$8.8{\cdot}10^{-6}$	$4 \cdot 10^{-3}$	$1.4 \cdot 10^{-4}$

piecewise linear target with two inclusions of different sizes, which should be an ideal target for the sparsity promoting standard gamma and inverse-gamma hyperpriors.

First of all, we note that changing η and β parameters of the standard and inverse-gamma hyperpriors was not observed to have a significant effect on the reconstructions as long as the values were sufficiently small, i.e., $\eta < 10^{-2}$ and $\beta < 5$. For larger values, the effect of the hyperpriors diminished. Thereby we only provide results with hyperparameter values $\eta = 10^{-4}$ and $\beta = 1.5$ and alter the hyperparameters ϑ and γ instead.

For each of the hyperprior type, three different magnitudes of hyperparameters (ϑ and γ) were used: *(i)* A *low* value that provided only slight effect to the optimization problem, *(ii)* an *intermediate* value close to optimal, and *(iii) high* value strongly affecting on the optimization. The used hyperparameters ϑ and γ are given in Table 1. The utilized ϑ values were chosen by using the cumulative distribution function (30) by assuming $|x_j - \mu| < M$ with $M \in \{0.3, 1, 10\}$ for standard gamma and $M \in \{0.3, 1, 5\}$ for inverse-gamma. The smallest M value correspond to the *low*, the middle to the *intermediate* and the largest to the *high*. The mean μ of the uncorrelated Gaussian prior was set to be 0.01 for the absorption and 1 for the reduced scattering. These values were also the true values of the background. Since the magnitude of the absorption values was assumed to be 100 times smaller than the reduced scattering values, M values used for the absorption were the reduced scattering values M divided by 100. For the same reason, the hyperparameters γ (which sets the minimum variance in IAS iterations (26)) used for the absorption were 0.01% of γ used for the reduced scattering.

Figure 1 shows the simulated target and the MAP estimates computed for all three cases by using the IAS method (Algorithm 1). For comparison, the reconstruction was also computed using fixed variance values shown at the top of Figure 1. The fixed variances were set as 0.25^2 and 0.0025^2 for reduced scattering and absorption, respectively, since these values produced admissible reconstruction. Note, that the reconstruction with the fixed variances corresponds to the first iterations of the IAS. The colorbars in Figure 1 exclude some of the largest and smallest values. Relative errors of the estimates are shown in Table 2

From Figure 1, we can observe that the reconstructions with smaller hyperparameters, i.e., with stronger prior information, produce a smooth background for both optical pa-

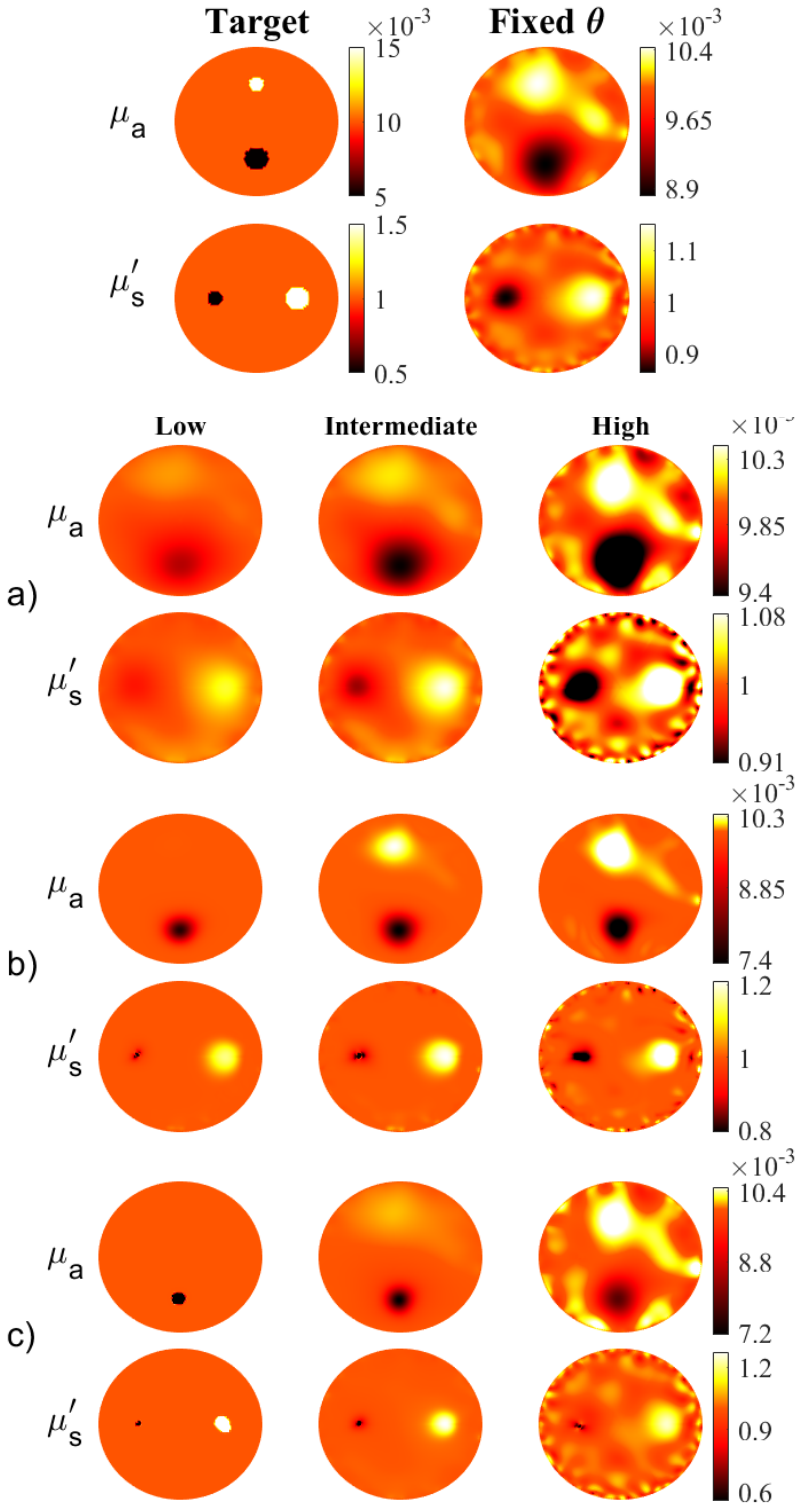

Figure 1: Computed MAP estimates. The two top-most rows show the true values and the estimates when the uncorrelated Gaussian prior (5) with fixed variances of 0.25^2 and 0.0025^2 for reduced scattering and absorption were used. Other rows show the estimates when variances were a) inverse-gamma, b) standard gamma, and c) exponentially distributed. The columns show the estimates with small (left), intermediate (mid) and large (right) hyperparameter ϑ or γ values as listed in Table 1. The colorbars exclude some of the highest and smallest values.

Table 2: Relative errors (%) of the MAP estimates when the uncorrelated Gaussian prior (5) was used with low, intermediate and high hyperparameter ϑ (or γ) values. The left-most column indicates the type of the hyperprior used for the unknown variances. The non-hierarchical model with the fixed variances is denoted as "no hyperprior". For reduced scattering and absorption, the smallest relative error is bold.

Hyperparam. / Hyperprior	Low		Intermediate		High	
	RE (μ_a)	RE (μ_s)	RE (μ_a)	RE (μ_s)	RE (μ_a)	RE (μ_s)
No hyperprior	7.35	6.80	7.35	6.80	7.35	6.80
Exponential	7.87	16.46	6.55	5.42	6.81	7.01
Standard gamma	6.63	6.51	6.29	5.81	**5.57**	**5.28**
Inverse-gamma	7.94	7.69	7.73	7.34	7.06	6.74

rameters. Whereas the reconstructions with the larger hyperparameters and weaker prior information produced inclusions with increased contrast with a noisier background. On the other hand, the strong sparsity assumption (left-most column) causes the smaller inclusions to vanish. Notably, none of the MAP estimates produced cross-talk artefacts between scattering and absorbtion, which commonly occur in DOT. For some of the reconstructions in Figure 1, the single smallest reduced scattering value was drastically smaller than the others.

Comparing the qualitative performance between the hyperprior types shows that the standard gamma hyperprior (b) yields superior absorption estimates compared to the exponential (c) and inverse-gamma (a) hyperprior, which cannot properly localize the smaller absorption inclusion. Additionally, a well-working set of hyperparameters was observed to be broad for the model with the standard gamma hyperpriors. Table 2 shows that the standard gamma hyperprior also yields the best estimates regarding the relative errors.

5.2 Reconstruction utilizing the difference prior model

The second prior model considered in this work was the difference prior (6). To test the performance of the hierarchical difference prior models, we used a target with two inclusions with sharp edges that should be ideal for investigating the edge-preserving properties. Compared to the inclusions used for the uncorrelated Gaussian prior Figure 1, the inclusions were now slightly larger. This was mainly because the difference prior model was observed to vanish too small inclusions.

Similarly, as with the uncorrelated Gaussian prior, the hierarchical difference prior was tested with three different values of hyperparameters ϑ and γ. To compute the hyperparameters ϑ from the CDF (30) we used three different M values. For the reduced scattering we set $M \quad \{1, 5, 10\}$ with standard gamma and $M \quad \{0.25, 1, 4\}$ with inverse-gamma hyperpriors. Since absorption values were known to be approximately 1% of the reduced scattering values, the differences d of absorption are approximately 1% of the reduced scattering differences as well. Therefore we used 100 times smaller M values for the absorption. For the non-hierarchical model, the variances were set as 0.1^2 for reduced scattering and 0.001^2 for absorption yielding an acceptable reconstruction quality. For η, we used 10^{-4} and for $\beta = 3/2$, i.e., the same values as with the uncorrelated Gaussian prior.

Figure 2 shows the MAP estimates and the used target. Some of the highest and lowest values are excluded from the colorbars. Table 3 shows the corresponding relative errors.

Table 3: Relative errors (%) of the MAP estimates when the difference prior (6) was used with low, intermediate and high hyperparameter ϑ (or γ) values. The left-most column indicates the type of the hyperprior used for the unknown variances. The non-hierarchical model with the fixed variances is denoted as "no hyperprior". For reduced scattering and absorption, the smallest relative error is bold.

Hyperparam. / Hyperprior	Low		Intermediate		High	
	RE (μ_a)	RE (μ_s)	RE (μ_a)	RE (μ_s)	RE (μ_a)	RE (μ_s)
No hyperprior	9.78	7.17	9.78	7.17	9.78	7.17
Exponential	11.14	9.00	9.60	6.89	9.44	6.67
Standard gamma	10.55	6.54	8.74	**4.28**	**7.80**	4.40
Inverse-gamma	11.08	6.14	10.26	7.67	9.12	6.34

As can be seen from the figure, the hierarchical models with the smaller hyperparameters produce sharper-edged inclusions. The disadvantage of using too small hyperparameter values is the falsely estimated edges of the absorption inclusion. With the large hyperparameters, the hierarchical models tend to increase the background noise, excluding the standard gamma hyperprior model (b), which still enhances the sharpness of the inclusions compared to the fixed variances estimates. Again the standard gamma hyperprior is observed to produce the most feasible estimates when ϑ is sufficiently large. The exponential hyperprior (c) is seemingly performing well with the reduced scattering, but the absorption inclusions remain as blurry as with the fixed variances. With the inverse-gamma hyperprior (a), no significant improvement is observed. Similarly as with the uncorrelated Gaussian noise, no cross-talk artefacts appeared in any of the reconstructions shown in Figure 2.

Figures 1 and 2 demonstrate the unstable performance of the exponential and inverse-gamma hyperpriors. With smaller hyperparameters, the exponential and inverse-gamma hyperpriors diminished the background noise but also significantly decreased the contrast of the inclusions. While with the larger hyperparameters, the contrast of the inclusions was better, but now the background noise was more substantial. On the other hand, the effect of these hyperpriors can be more plausible with a different kind of test target. For instance, in Figure 3, we show one alternative target, with only symmetric and positive inclusions with eight times higher contrast. The MAP estimates shown in Figure 3 were computed using the hierarchical difference prior with the same intermediate hyperprior values, used to obtain the MAP estimates in the middle column of Figure 2. The estimates with the fixed variances were computed with variances of 0.5^2 and 0.005^2 for reduced scattering and absorption, respectively.

In contrast to the previous MAP estimates, by observing Figure 3, we can see that the inverse-gamma and exponential hyperpriors have a substantial effect on the noise artefacts and sharpness of inclusions. Now, the inverse-gamma hypermodel produces sharp inclusions for both reduced scattering and absorption, while the exponential hyperprior greatly increases the contrast of the inclusions. On the other hand, the visibility of the background noise is also increased, which was expected due to the exponential hyperpriors feature to promote more common outliers. The standard gamma hyperprior produces sharp-edged inclusions but, as a drawback, yields two false cross-talk inclusions for the reduced scattering. The cross-talk artefacts are also appearing in the other reconstructions, but with reduced contrast.

In general, we observed that the exponential and inverse-gamma hyperprior worked

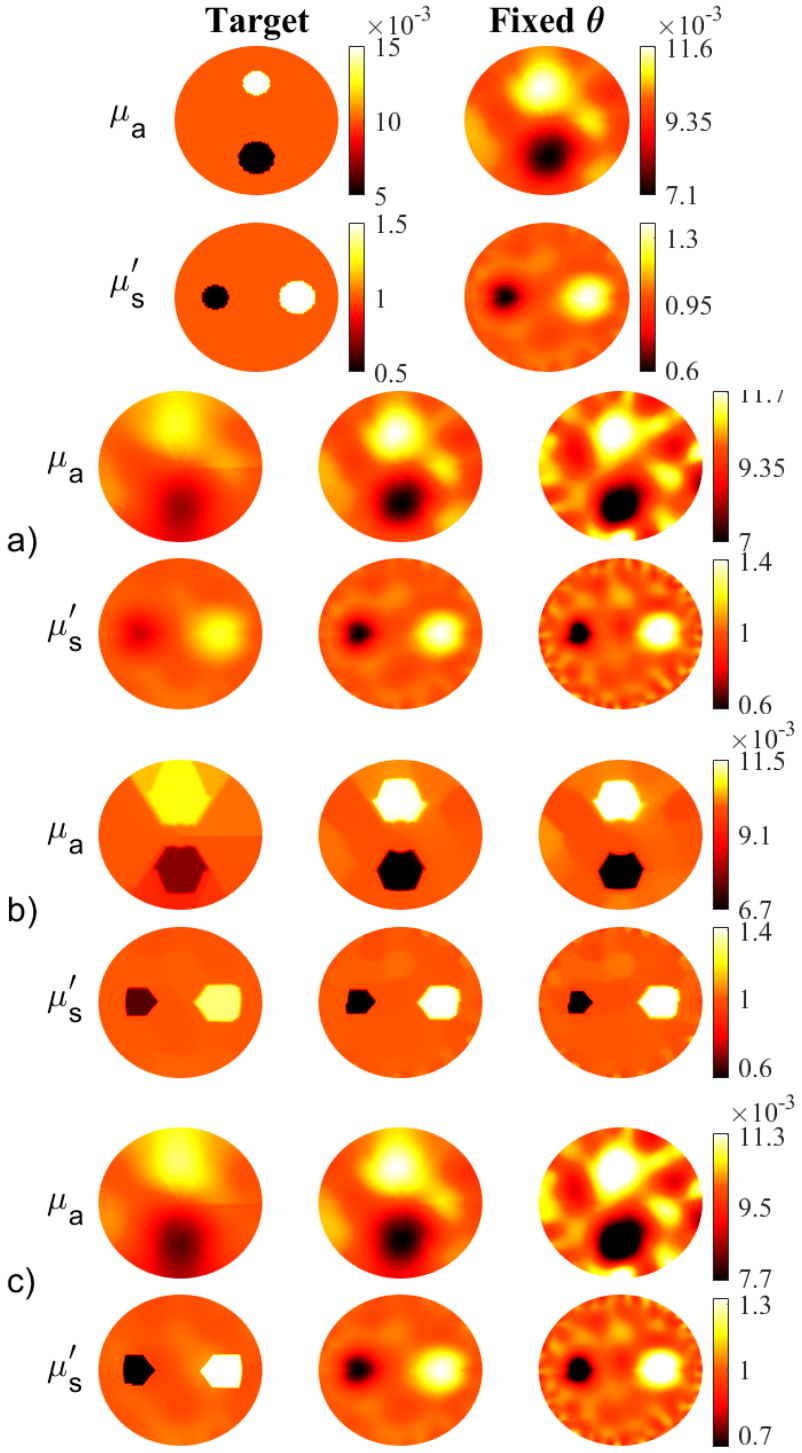

Figure 2: Computed MAP estimates. The two top-most rows show the true values and the estimates when the uncorrelated Gaussian prior (5) with fixed variances of 0.1^2 and 0.001^2 for reduced scattering and absorption were used. Other rows show the estimates, when variances were a) inverse-gamma, b) standard gamma and c) exponentially distributed. The columns show the estimates with small (left), intermediate (mid) and large (right) hyperparameter ϑ or γ values as listed in Table 1. The colorbars exclude some of the highest and smallest values.

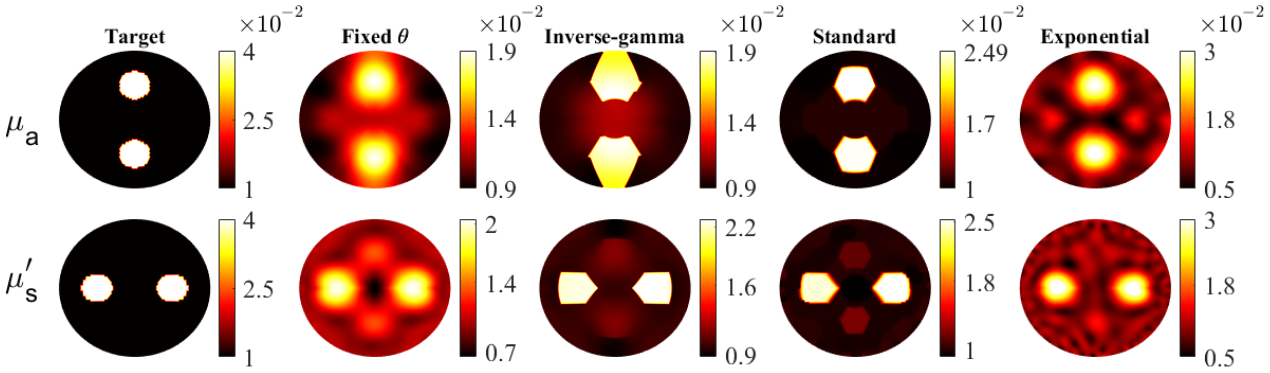

Figure 3: Computed MAP estimates of alternative target type with only positive inclusions. The left-most column show the target. The second column shows the estimates when the difference prior (6) with fixed variances of 0.1^2 and 0.001^2 for reduced scattering and absorption was used. Other columns show the estimates, when variances were inverse-gamma, standard gamma and exponentially distributed. The hyperparameters ϑ were set as the intermediate values, also used for recontructions in Figure 2.

well in cases with only positive (or negative) inclusions with a relatively large contrast. On the other hand, the estimates were more prone to produce cross-talk artefacts with the high contrast targets.

5.3 Convergence of the nonlinear IAS

Besides the effect of the hierarchical models on the MAP estimates, we also wanted to investigate the empirical convergence of the nonlinear IAS. The convergence was only inspected iteration-wise due to the changing amount of the (Gauss-Newton) inner iterations. Generally, the first few IAS iterations took more inner iterations, while this was quickly reduced to just a few. A single Gauss-Newton iteration too 10-30 seconds for the models with the uncorrelated Gaussian prior and 30-50 seconds with the difference prior. Updating part b), i.e., the variances during the IAS iterations took less than 0.1 seconds, which was neglectable compared to the total running time of a single IAS iteration. The total amount of inner iterations before reaching the stopping criteria varied between the used hyperparameters and hyperpriors. The models with the difference prior were observed to need more IAS and inner iterations than those with the uncorrelated Gaussian prior. Similarly, from hyperpriors, models with the inverse-gamma hyperprior reached the convergence fastest while with exponential hyperprior had the slowest convergence. The minimum amount of total inner iterations was 60, obtained with uncorrelated Gaussian prior utilizing inverse-gamma hyperpriors. Respectively, the maximum was 552 when the difference prior was used with exponential hyperpriors.

The convergence was observed for each MAP estimate with the intermediate hyperparameter values. That is, Figure 4 shows the convergence of the IAS, related to the MAP estimates shown in the middle columns of Figures 1 and 2. We also plotted a
linear convergence (blue dashed line) for reference in Figure 4. By looking at the convergence slopes, the convergence of the nonlinear IAS can be observed to be comparable with the linear convergence. The linear convergence has a convergence constant, i.e., $\|x^{i+1} - x_{\mathrm{MAP}}\|_2 \leq \mu \|x^i - x_{\mathrm{MAP}}\|_2$ with μ of 0.6. Thereby, the convergence of the nonlin-

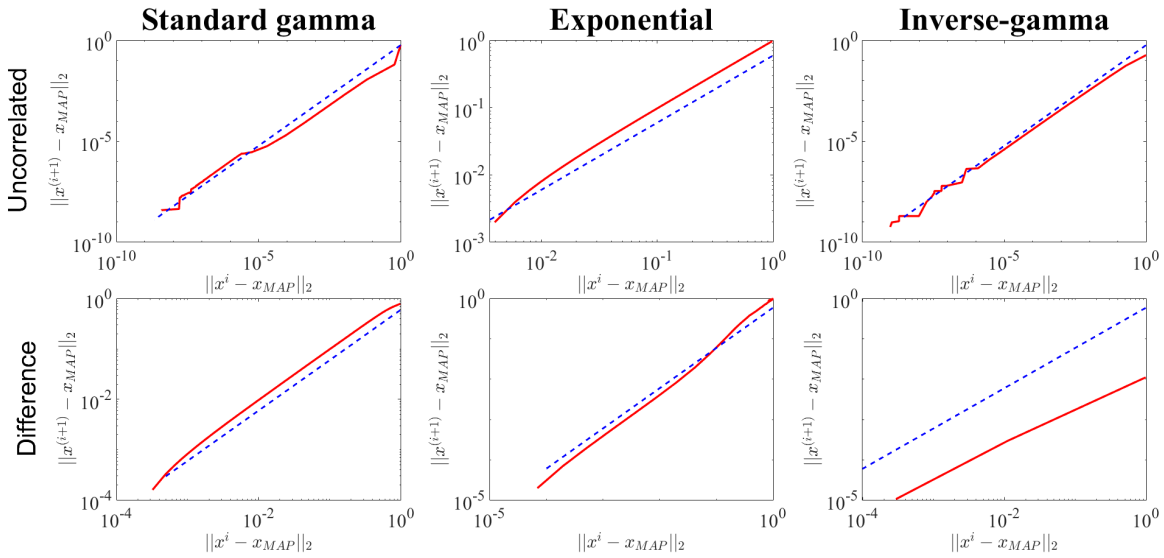

Figure 4: Convergence of the MAP estimates of the hierarchical models. The vertical axis shows the error $||x^i - x_{MAP}||_2$ of the ith iteration and the horizontal axis the error of the $(i+1)$th iteration. The MAP estimates x_{MAP} were computed by accurately solving the corresponding optimization problems. The bottom row shows the convergence of the difference prior model and the top row convergence of the uncorrelated Gaussian prior. The left, middle and right columns show the convergence for standard gamma, exponential and inverse-gamma hyperpriors, respectively. The red lines show the convergence of the used nonlinear IAS algorithm and the dashed blue line shows a linear convergence with convergence constant of $\mu = 0.6$.

ear IAS was observed to be relatively fast in the linear sense. These empirical results are also in line with the previously proven [14] at-least linear convergence for the uncorrelated Gaussian noise with standard gamma hyperpriors when the forward model is linear.

Two important factors affecting the time-wise convergence are the required accuracy of the inner iterations (of Gauss-Newton) and the stopping criteria of the outer iterations. In this work, we used strict stopping conditions for the inner iterations to achieve an accurate minimizer of part a). On the other hand, a less accurate optimization of part a) could produce a sufficient approximation of the minimizer, allowing faster computation of the IAS iterations. For instance, with the linear forward model in MEG [7], the Krylov-subspace approximation of the least squares solution has been shown to provide sufficient approximation leading to significant speedup.

During the experimented IAS runs, it was observed that the first few iterations already encapsulate the main effect of the hierarchical models, especially with the uncorrelated Gaussian prior. After the early iterations, the IAS iterations only seemed to increase the contrast of the discontinuity points. Using early stopping criteria can drastically reduce the run time of the IAS algorithm, yet one needs to develop a systematic way to determine the criteria. To demonstrate the possibility of using early stopping, Figure 5 shows the estimates of the optical parameters just after three iterations when the intermediate hyperparameter values were used as in Figure 1. Finding a systematic way to set the early stopping criteria was beyond the scope of this work but could provide substantial speedup. For the nonlinear forward model, this would likely require further assumptions on the model.

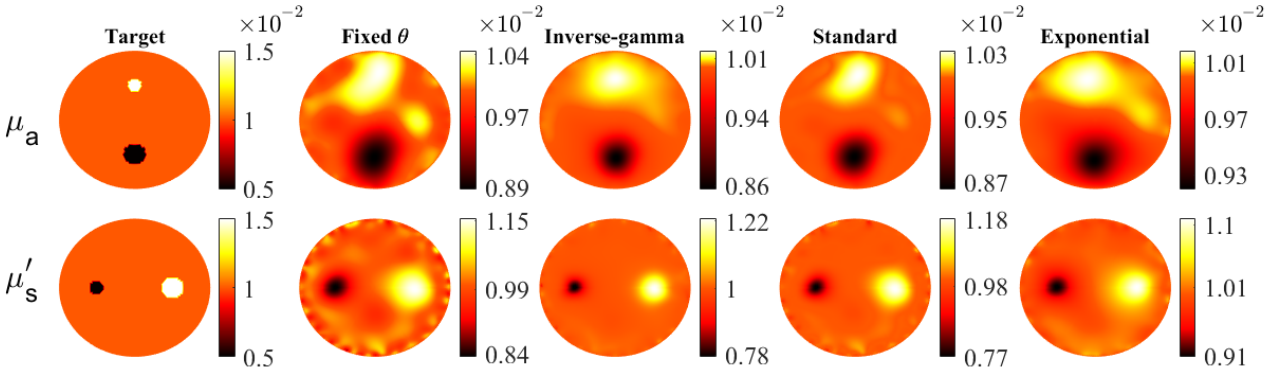

Figure 5: The absorption and reduced scattering estimates after three IAS iterations, when uncorrelated Gaussian prior was used with the intermediate hyperparameter values as in Figure 1.

6 Conclusions

In this work, hierarchical priors were formulated in the Bayesian framework for the highly ill-posed and nonlinear problem of diffuse optical tomography. The studied hyperpriors included exponential, standard, and inverse-gamma hyperpriors that were used with the uncorrelated Gaussian and difference prior. The simulated DOT problem demonstrated the hyperpriors having a considerable effect on the reconstructions of the piecewise linear targets. While the plausibility of the reconstructed characteristics depended on the chosen prior, hyperprior, and hyperparameters.

Results with the uncorrelated Gaussian prior showed the hierarchical models improving the localization of the inclusions by diminishing most of the background noise. The standard gamma hyperprior was observed to perform most robustly, while the exponential and inverse-gamma hyperpriors could not localize absorption inclusion well. When the difference prior was utilized, the hierarchical models were shown to enhance the sharpness of the edges. Again the standard gamma hyperprior was observed to work most robustly, with the broadest set of hyperparameters. The exponential and inverse-gamma hyperpriors could only enhance the edges of the simpler targets. Notably, the hierarchical priors showed excellent performance in supressing cross-talk artefacts.

For all studied hierarchical models, the convergence of the nonlinear IAS was empirically investigated and was observed to be linear. These observations hold only when the inner (Gauss-Newton) iterations have strict stopping criteria, which can lead to slow computation times of a single IAS iteration. A computational speedup could be achieved by solving the inner iterations less accurately.

In this work, we utilized the cumulative distribution approach to obtain suitable hyperparameter values for the hierarchical models with nonlinear forward operators. This method worked sufficiently well, but finding close to optimal hyperparameters still needed manual adjustment. Additionally, the cumulative distribution function approach excludes the possibility of considering spatial sensitivity differences, that is, a sensitivity-weighting. To justify more rigorous use of the hierarchical models with a nonlinear forward model, future work needs to focus on developing hyperparameter selection rules.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 101001417 – QUANTOM). This work was supported by the Finnish Cultural Foundation, the Academy of Finland (Center of Excellence in Inverse Modeling and Imaging project 336799, 336796, the Flagship Program Photonics Research and Innovation grant 320166, and the Academy Research Fellow project 338408, 346574)

References

[1] C. Aguerrebere, A. Almansa, J. Delon, Y. Gousseau, and P. Musé. A Bayesian hyper-prior approach for joint image denoising and interpolation, with an application to hdr imaging. *IEEE Transactions on Computational Imaging*, 3(4):633–646, 2017.

[2] S. R. Arridge. Optical tomography in medical imaging. *Inverse Problems*, 15(2):R41, 1999.

[3] S. R. Arridge and J. Schotland. Optical tomography: forward and inverse problems. *Inverse Problems*, 25:123010, 2009.

[4] S. R. Arridge, M. Schweiger, M. Hiraoka, and D. Delpy. Performance of an iterative reconstruction algorithm for near-infrared absorption and scatter imaging. In *Photon Migration and Imaging in Random Media and Tissues*, volume 1888, pages 360–371. International Society for Optics and Photonics, 1993.

[5] J. M. Bardsley, A. Seppanen, A. Solonen, H. Haario, and J. P. Kaipio. Randomize-then-optimize for sampling and uncertainty quantification in electrical impedance tomography. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):1136–1158, 2015.

[6] D. Calvetti, H. Hakula, S. Pursiainen, and E. Somersalo. Conditionally Gaussian hypermodels for cerebral source localization. *SIAM Journal on Imaging Sciences*, 2(3), 2009.

[7] D. Calvetti, A. Pascarella, F. Pitolli, E. Somersalo, and B. Vantaggi. A hierarchical Krylov-Bayes iterative inverse solver for MEG with physiological preconditioning. *Inverse Problems*, 31(12):125005, 2015.

[8] D. Calvetti, M. Pragliola, E. Somersalo, and A. Strang. Sparse reconstructions from few noisy data: analysis of hierarchical Bayesian models with generalized gamma hyperpriors. *Inverse Problems*, 36(2):025010, 2020.

[9] D. Calvetti, F. Sgallari, and E. Somersalo. Image inpainting with structural bootstrap priors. *Image and Vision Computing*, 24(7):782–793, 2006.

[10] D. Calvetti and E. Somersalo. Local regularization and Bayesian hypermodels. *Proc. SPIE 5910, Advanced Signal Processing Algorithms, Architectures, and Implementations XV*, 5910, 2005.

[11] D. Calvetti and E. Somersalo. A Gaussian hypermodel to recover blocky objects. *Inverse problems*, 23(2):733, 2007.

[12] D. Calvetti and E. Somersalo. Hypermodels in the Bayesian imaging framework. *Inverse Problems*, 24(3):034013, 2008.

[13] D. Calvetti and E. Somersalo. Inverse problems: From regularization to Bayesian inference. *WIREs Computatinal Statistics*, 10(3), 2018.

[14] D. Calvetti, E. Somersalo, and A. Strang. Hierachical Bayesian models and sparsity: ^{2}-magic. *Inverse Problems*, 35(3):035003, 2019.

[15] J. Chen. Optical tomography in small animals with time-resolved Monte Carlo methods. *Dissertation Abstracts International*, 74(04), 2012.

[16] S. G. Diamond, T. J. Huppert, V. Kolehmainen, M. A. Franceschini, J. P. Kaipio, S. R. Arridge, and D. A. Boas. Dynamic physiological modeling for functional diffuse optical tomography. *NeuroImage*, 30:88–101, 2006.

[17] T. Durduran, R. Choe, W. B. Baker, and A. G. Yodh. Diffuse optics for tissue monitoring and tomography. *Reports on Progress in Physics*, 73:076701, 2015.

[18] A. Gibson, J. Hebden, and S. R. Arridge. Recent advances in diffuse optical imaging. *Physics in Medicine & Biology*, 50(4):R1, 2005.

[19] D. Grosenick, H. Rinneberg, R Cubeddu, and P. Taroni. Review of optical breast imaging and spectroscopy. *Journal of Biomedical Optics*, 21(10):091311, 2016.

[20] M. Guven, B. Yazici, X. Intes, and B. Chance. Diffuse optical tomography with a priori anatomical information. *Physics in Medicine & Biology*, 50(12):2837, 2005.

[21] M. Guven, B. Yazici, X. Intes, and B. Chance. Hierarchical Bayesian algorithm for diffuse optical tomography. In *34th Applied Imagery and Pattern Recognition Workshop (AIPR'05)*, pages 6–pp. IEEE, 2005.

[22] J. C. Hebden, A. Gibson, R. Md. Yusof, N. Everdell, E. M. C. Hillman, D. T. Delpy, S. R. Arridge, T. Austin, J. H. Meek, and J. S. Wyatt. Three-dimensional optical tomography of the premature infant brain. *Physics in Medicine & Biology*, 47(23):4155, 2002.

[23] P. Hiltunen, S. Prince, and S. R. Arridge. A combined reconstruction-classification method for diffuse optical tomography. *Physics in Medicine & Biology*, 54(21):6457, 2009.

[24] Y. Hoshi and Y. Yamada. Overview of diffuse optical tomography and its clinical applications. *Journal of Biomedical Optics*, 21(9):091312, 2016.

[25] A. Ishimaru. *Wave Propagation and Scattering in Random Media*. Academic Press, 1978.

[26] J. P. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*. Springer, Newyork, 2005.

[27] F. Lin, T. Witzel, S. P. Ahlfors, S. M. Stufflebeam, J. W. Belliveau, and M. S. Hämäläinen. Assessing and improving the spatial accuracy in MEG source localization by depth-weighted minimum-norm estimates. *NeuroImage*, 31(1):160–171, 2006.

[28] S. Liu, J. Jia, D. Yimi Y. Zhang, and Y. Yang. Image reconstruction in electrical impedance tomography based on structure-aware sparse Bayesian learning. *IEEE transactions on medical imaging*, 37(9):2090–2102, 2018.

[29] J. Mattout, C. Phillips, W. Penny, M. Rugg, and K. Friston. MEG source localization under multiple constraints: an extended Bayesian framework. *NeuroImage*, 30(3):753–767, 2006.

[30] A. Miyamoto, K. Watanabe, K. Ikeda, and M. Sato. Phase diagrams of a variational Bayesian approach with ARD prior in NIRS-DOT. In *The 2011 International Joint Conference on Neural Networks*, pages 1230–1236. IEEE, 2011.

[31] M. Mozumder, A. Hauptmann, I. Nissilä, S. R. Arridge, and T. Tarvainen. A model-based iterative learning approach for diffuse optical tomography. *IEEE Transactions on Medical Imaging*, 41(5):1289–1299, 2021.

[32] M. Mozumder, T. Tarvainen, J.P. Kaipio, S. R. Arridge, and V. Kolehmainen. Compensation of modeling errors due to unknown domain boundary in diffuse optical tomography. *JOSA A*, 31(8):1847–1855, 2014.

[33] A. Nummenmaa, T. Auranen, M. Hämäläinen, I. Jääskeläinen, J. Lampinen, M. Sams, and A. Vehtari. Hierarchical Bayesian estimates of distributed MEG sources: theoretical aspects and comparison of variational and MCMC methods. *NeuroImage*, 35(2):669–685, 2007.

[34] K. Paulsen and H. Jiang. Enhanced frequency-domain optical image reconstruction in tissues through total-variation minimization. *Applied optics*, 35(19):3447–3458, 1996.

[35] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990.

[36] L. Roininen, M. Girolami, S. Lasanen, and M. Markkanen. Hyperpriors for matérn fields with applications in Bayesian inversion. *arXiv preprint arXiv:1612.02989*, 2016.

[37] M. Schweiger and S. R. Arridge. The Toast++ software suite for forward and inverse modeling in optical tomography. *Journal of Biomedical Optics*, 19(4):040801, 2014.

[38] M. Schweiger, S. R. Arridge, and I. Nissilä. Gauss–Newton method for image reconstruction in diffuse optical tomography. *Physics in Medicine & Biology*, 50(10):2365, 2005.

[39] C. B. Shaw and P. K. Yalavarthy. Performance evaluation of typical approximation algorithms for nonconvex $_p$-minimization in diffuse optical tomography. *Journal of the Optical Society of America A*, 31(4):852–862, 2014.

[40] T. Shimokawa, T. Kosaka, O. Yamashita, N. Hiroe, T. Amita, Y. Inoue, and M. Sato. Hierarchical Bayesian estimation improves depth accuracy and spatial resolution of diffuse optical tomography. *Optics express*, 20(18):20427–20446, 2012.

[41] T. Shimokawa, T. Kosaka, O. Yamashita, N. Hiroe, T. Amita, Y. Inoue, and M. Sato. Extended hierarchical Bayesian diffuse optical tomography for removing scalp artifact. *Biomedical Optics Express*, 4(11):2411–2432, 2013.

[42] D. Vidaurre, C. Bielza, and P. Larra naga. A survey of L1 regression. *International Statistical Review*, 81(3):361–387, 2013.

[43] M. Wheelock, J. Culver, and A. Eggebrecht. High-density diffuse optical tomography for imaging human brain function. *Review of Scientific Instruments*, 90(5):051101, 2019.

[44] O. Yamashita, T. Shimokawa, R. Aisu, T. Amita, Y. Inoue, and M. Sato. Multi-subject and multi-task experimental validation of the hierarchical Bayesian diffuse optical tomography algorithm. *NeuroImage*, 135:287–299, 2016.

[45] Jaejun Yoo, Sohail Sabir, Duchang Heo, Kee Hyun Kim, Abdul Wahab, Yoonseok Choi, Seul-I Lee, Eun Young Chae, Hak Hee Kim, Young Min Bae, et al. Deep learning diffuse optical tomography. *IEEE transactions on medical imaging*, 39(4):877–887, 2019.

[46] G. Zhang, X. Cao, B. Zhang, F. Liu, J. Luo, and J. Bai. MAP estimation with structural priors for fluorescence molecular tomography. *Physics in Medicine & Biology*, 58(2):351, 2012.